

Towards Reliable Semantic Vision

by

Tejas Gokhale

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2023 by the
Graduate Supervisory Committee:

Yezhou Yang, Co-Chair
Chitta Baral, Co-Chair
Heni Ben Amor
Rushil Anirudh

ARIZONA STATE UNIVERSITY

May 2023

© Tejas Gokhale.

All Rights Reserved.

ABSTRACT

Models that learn from data are widely and rapidly being deployed today for real-world use, and have become an integral and embedded part of human lives. While these technological advances are exciting and impactful, such data-driven computer vision systems often fail in inscrutable ways. This dissertation seeks to study and improve the reliability of machine learning models from several perspectives including the development of robust training algorithms to mitigate the risks of such failures, construction of new datasets that provide a new perspective on capabilities of vision models, and the design of evaluation metrics for re-calibrating the perception of performance improvements. I will first address distribution shift in image classification with the following contributions: (1) two methods for improving the robustness of image classifiers to distribution shift by leveraging the classifier's failures into an adversarial data transformation pipeline guided by domain knowledge, (2) an interpolation-based technique for flagging out-of-distribution samples, and (3) an intriguing trade-off between distributional and adversarial robustness resulting from data modification strategies. I will then explore reliability considerations for *semantic vision* models that learn from both visual and natural language data; I will discuss how logical and semantic sentence transformations affect the performance of vision–language models and my contributions towards developing knowledge-guided learning algorithms to mitigate these failures. Finally, I will describe the effort towards building and evaluating complex reasoning capabilities of vision–language models towards the long-term goal of robust and reliable computer vision models that can communicate, collaborate, and reason with humans.

ACKNOWLEDGMENTS

ॐ सहनाववतु ॥ सहनौभुनक्तु ॥

सह वीर्यं करवावहै ॥

तेजस्विनावधीतमस्तु मा विद्विषावहै ॥

om saha nāvavatu . saha nau bhunaktu .

saha vīryam̄ karavāvahai .

tejasvi nāvadhitamastu mā vidvisāvahai .

Om, may it (Knowledge), protect us both (the teacher and the student).

May it nourish us both. May we generate great energy.

May our study lead to Brilliance (of Understanding, leading to Knowledge).

May it not give rise to hostility (due to lack of Understanding).

Taittiriya Upanishad, Anandavalli Invocation

There are many who have directly and personally shaped my thought process as a scientist. Foremost among them are my advisors Yezhou “YZ” Yang and Chitta Baral.

If Ph.D. is an entrepreneurial venture, Yezhou Yang is my angel investor. Yezhou is an amazing mentor and the nicest person you will ever meet – he created a healthy work environment that allowed me to explore, stumble, and to learn to pick myself up and offered me unflinching support to keep going. He not only shaped the way I think about research problems, but he has had a deep impact on my approach to the managerial aspect of research – it wouldn’t be an understatement to say that he has been a critical factor in my decision to pursue an academic career.

I am grateful to Chitta Baral for co-advising me at ASU. Discussions with him made me realize how several problems raised by decades of work on knowledge representation and reasoning remain unaddressed in computer vision – Chitta’s insights brought me closer to the domain of multimodal learning. Chitta has been instrumental in helping me think big and to think about ideas that will have long-term impact. It has been an absolute privilege to learn from and collaborate with him.

I am grateful for the mentorship of Rushil Anirudh at Lawrence Livermore National Laboratory. The terrific trio of Rushil, Jay Thiagarajan, and Bhavya Kailkhura introduced me to many ideas in robustness and generalization, making my two summers at LLNL a game-changer for my Ph.D. thesis. I would like to thank Hamid Palangi, Besa Nushi, Vibhav Vineet, and Eric Horvitz who hosted me at Microsoft Research – their insights has led to a very important research agenda on benchmarking abilities of large vision-language models. Eric’s energy, erudition, and enthusiasm for technological advances has been truly inspirational. I am grateful to Heni Ben Amor, a crucial member of my supervisory committee – discussions with him about the role of language in robotics have shaped my future research goals.

I would like to thank Aswin Sankaranarayanan for advising me at Carnegie Mellon University and Kuldeep Kulkarni for mentoring my first research project. Aswin and Kuldeep were catalysts in my decision to pursue a Ph.D. Thanks also to Guru Krishnan and Shree Nayar for mentoring me at Snap Research.

A special word of thanks to Pratyay Banerjee, Zhiyuan Fang, and Man Luo for being partners in creating an atmosphere of ambition and excellence, and for our initiatives such as our ODRUM workshop and V&L seminar series. I have been fortunate to be part of two incredible labs with an amazing set of collaborators: Pratyay, Zhiyuan, Man, Swaroop Mishra, Neeraj Varshney, Joshua Feinglass, Maitreya Patel, Abhishek Chaudhary, Ethan Wisdom, Shailaja Sampat, Sheng Cheng, and labmates: Changhoon Kim, Xin Ye, Mohammad Farhadi, Rudra Saha, Aadhavan Sadasivam, Nilay Yilmaz, Arindam Mitra, Blake Harrison, Arpit Sharma, Kazuaki Kashihara, Sam Rawal, Aurgho Bhattacharjee, Ishan Shrivastava, Ming Shen, Himanshu Gupta, Yiran Luo, Agneet Chatterjee, and Pavel Dolin. Thanks to Serita Sulzman, Lu Cheng, Meijia Wang, Huiliang Shao, Varun Jammula, Sumedh Godbole, Mihir Parmar, Kuntal Pal, John Janiczek, and Kowshik Thopalli for all fun times in Tempe and beyond.

I am grateful to Yoojung Choi, Nakul Gopalan, Eduardo Blanco, Pavan Turaga, Suren Jayasuriya, Yapeng Tian, Sarath Sreedharan, Lu Cheng, Rowan Zellers, Vicente Ordóñez, Chaowei Xiao, Gautam Dasarathy, Elizabeth Bondi-Kelly, Ted Pavlic, and the Future Faculty Forum (created by Supreeth Shastri and Vijay Chidambaram) for their insights and feedback during my faculty job search, and to Kaize Ding and Raha Moraffah for their feedback on my jobtalk. I am indebted to the unsung heroes of ASU – Pamela Dunn, Monica Dugan, and Theresa Chai in SCAI Business Office; Jaya Krishnamurthy, Arzuhan Kavak, Christina Sebring, and Araxi Hovhannessian in SCAI Advising; Lincoln Slade and Brint MacMillan in SCAI IT; ASU Research Computing, GPSA, and ISSC for their support with logistics, infrastructure, payroll, and travel.

I am thankful to my band of brothers *a.k.a. TFG++*, who have been family away from home for more than ten years. A huge thank you to my Apte family for their love, care, support, and lots of pampering.

I have been extremely fortunate and blessed to have grown up in an atmosphere of love, care, integrity, and excellence, created by my parents Ravindra Gokhale and Mugdha Gokhale who made me who I am today, and my sister Meghana Gokhale for leading by example. I am thankful to my grandparents Dr. Manohar Joshi and Dr. Sulekha Joshi who introduced me to the joy of learning.

Finally, I am thankful to generations of courageous Bharatiyas who sustained the civilization and liberated my motherland from a thousand long and dark years of terrorism, subjugation, and colonization. Without their sacrifices, I would not have had the opportunity of writing these words.

“Knowledge is in the end based on acknowledgement.”

- Ludwig Wittgenstein

TABLE OF CONTENTS

	Page
LIST OF TABLES	xi
LIST OF FIGURES	xvii
CHAPTER	
1 INTRODUCTION	1
1.1 Understanding What Cameras See	2
1.2 Recent Success In Visual Understanding	4
1.3 Robustness and Generalization	7
1.4 Background	10
1.4.1 Robust Machine Learning	10
1.4.2 Robust Natural Language Understanding	13
1.4.3 Robustness of Vision–Language Models.	14
1.5 Overview: Towards Reliable Visual Understanding	16
2 ROBUSTNESS UNDER ATTRIBUTE SHIFT	22
2.1 Related Work	24
2.2 Setup	25
2.3 Attribute Guided Adversarial Training	27
2.3.1 Proposed Approach	28
2.4 CLEVR-Singles Dataset Creation	32
2.5 Experiments.....	34
2.5.1 Semantic Object-Level Perturbations	35
2.5.2 Geometric Transformations	39
2.5.3 Common Image Corruptions	44
2.6 AGAT-Generated Images.....	47
2.7 Discussion.....	47

CHAPTER	Page
3 ADVERSARILY LEARNED TRANSFORMATIONS FOR SINGLE-SOURCE DOMAIN GENERALIZATION	51
3.1 Introduction	51
3.2 Related Work	55
3.3 Proposed Approach	58
3.3.1 ALT: Adversarially Learned Transformations	59
3.4 Experiments	63
3.4.1 PACS	65
3.4.2 Office-Home	67
3.4.3 Digits	68
3.5 Analysis of ALT	69
3.5.1 ALT is better than naïve diversity	71
3.5.2 Effect of Varying ALT Hyperparameters	72
3.5.3 Complexity of Adversary Network	72
3.6 Conclusion	74
4 COVARIATE SHIFT DETECTION VIA DOMAIN INTERPOLATION SENSITIVITY	75
4.1 Introduction	75
4.2 Covariate Shift Detection	78
4.3 Domain Interpolation Sensitivity	80
4.4 Experiments	81
4.5 Related Work	84
4.6 Outlook	85

CHAPTER	Page
5 COMPARING THE EFFECTS OF DATA MODIFICATION METHODS ON OUT-OF-DOMAIN GENERALIZATION AND ADVERSARIAL ROBUSTNESS	86
5.1 Introduction	87
5.2 Categorization of Domain Generalization Methods	90
5.3 Toy Example: Concentric Circles	93
5.4 Experiments	95
5.5 Analysis	98
5.5.1 Correlation between Adversarial Robustness and OOD Gen- eralization	100
5.6 Related Work	101
5.7 Discussion	102
5.8 Broader Impact	103
6 VISION, LANGUAGE, AND LOGIC: ROBUSTNESS TO LOGICAL COMPOSITIONS	105
6.1 Introduction	105
6.2 Related Work	109
6.3 The Lens of Logic	111
6.3.1 Composite Questions	113
6.3.2 Dataset Creation Process	113
6.3.3 Analytical Setup	114
6.4 Method	116
6.4.1 Cross-Modal Feature Encoder	116
6.4.2 Our Model: Lens of Logic (LOL)	117

CHAPTER	Page
6.4.3 Loss Functions	118
6.4.4 Implementation Details	119
6.5 Experiments	119
6.5.1 Can't We Just Parse the Question into Components?	120
6.5.2 Explicit Training with Logically Composed Questions	123
6.5.3 Analysis	123
6.5.4 Evaluation on VQA v2.0 Test Data	126
6.6 Discussion	127
7 SEMANTICALLY DISTRIBUTED ROBUST OPTIMIZATION FOR VISION AND LANGUAGE INFERENCE	129
7.1 Method	132
7.1.1 Preliminaries	132
7.1.2 SDRO	134
7.2 SISP Sentence Transformations	136
7.2.1 Data Generation Pipeline	138
7.2.2 Data Analysis	138
7.2.3 Statistics	138
7.3 Experiments	147
7.3.1 Results	149
7.3.2 Fine-Grained Results	151
7.4 Analysis	153
7.4.1 Visualization of Perturbations	153
7.4.2 Comparison of Model Calibration	156
7.4.3 Size of Training Dataset	159

CHAPTER	Page
7.4.4 Ablation Studies	160
7.4.5 Robustness to Text-Attacks	164
7.5 Related Work	164
7.6 Discussion	166
8 BEYOND VISION-LANGUAGE ALIGNMENT: ENHANCING VIDEO CAPTIONING VIA COMMONSENSE DESCRIPTIONS	168
8.1 Introduction	168
8.2 Video to Commonsense (V2C)	171
8.2.1 V2C-Transformer	172
8.3 The V2C Dataset	174
8.3.1 Querying from ATOMIC and Re-ranking	176
8.3.2 Detailed Human Annotation	177
8.4 Experiments	183
8.4.1 Results	185
8.4.2 Qualitative Generation Results	189
8.5 V2C-QA	189
8.6 Related Work	193
8.7 Outlook	196
8.8 Conclusion	197
9 BENCHMARKING SPATIAL RELATIONSHIPS IN TEXT-TO-IMAGE GENERATION	198
9.1 Introduction	199
9.2 Related Work	202
9.3 Spatial Relationships Challenge Dataset	204

CHAPTER	Page
9.4 VISOR Metric	205
9.5 Experiments.....	209
9.5.1 Ineffectiveness of Existing Metrics	210
9.5.2 Benchmarking Results.....	211
9.5.3 Human Study	212
9.6 Analysis	214
9.7 Discussion and Conclusion.....	227
10 CONCLUSIONS.....	228
10.1 Summary of Contributions	228
10.2 Impact	230
10.3 Future Research Agenda	232
REFERENCES	236
APPENDIX	
A STATEMENT ON PREVIOUSLY PUBLISHED ARTICLES	276

LIST OF TABLES

Table	Page
2.1 List of Properties that Can Be Assigned to an Object to Render the Images in CLEVR-Singles.....	33
2.2 The Train and Test Splits for Our Experiments with Semantic Object-Level Perturbations for CLEVR-Singles	35
2.3 Classification Accuracy for Color-Classification on CLEVR-Singles. Source and Target Sets Are Split on <i>Size+position</i> Attribute for the Third Column, and <i>Material+Position</i> for the Fourth Column.	39
2.4 Results on the MNIST-RTS Robustness Benchmark for Rotation (R), Translation (T), Scaling (S), and Random Combination (RTS).....	41
2.5 The Effect of Augmentation Interval at Different Percentages of Augmented Samples.....	43
2.6 The Effect of Classification Loss Function at Different Percentages of Augmented Samples. GT Denotes the First Term and CR Is Consistency Regularization.	44
2.7 Comparison of Classification Accuracies on CIFAR-10-C Corruption Categories (Weather, Blur, Noise, Digital). Our Best Scores Are Underlined; Overall Best Are Bold.	46
3.1 Training Settings and Hyper-Parameters for Experiments on Each Benchmark.	64
3.2 Single-Source Domain Generalization Accuracy (%) on PACS. $X \rightarrow Y$ Implies X Is the Source Dataset and Y Is the Target Dataset. <i>P: Photo; A: Art-Painting; C: Cartoon; S: Sketch.</i> Performance Is Reported as Mean of 5 Repetitions	65

Table	Page
3.3 Single-Source Domain Generalization Accuracy (%) on Office-Home Over Five Repetitions. $X \rightarrow Y$ Implies X Is the Source Dataset and Y Is the Target Dataset. R : Real; A : Art; C : Clipart; P : Product. Performance Is Reported as Mean of 5 Repetitions	67
3.4 Single-Source Domain Generalization Accuracy (%) on Digit Classification, with MNIST-10K as Source and MNIST-M, SVHN, USPS, and SYNTH as Target Domains. Performance Reported Over Five Repetitions. Note: ADA and M-ADA Do Not Report Standard Deviation.	68
3.5 Effect of the Depth (Number of Convolutional Layers) of the Adversity Network g on Average Domain Generalization on All Three Benchmarks.	73
4.1 Covariate Shift Detection Performance on the OfficeHome Benchmark. All Methods Use the Same ResNet Classifier Trained on the “Real” and Results Are Shown as AUROC \uparrow / FPR95 \downarrow	83
4.2 Covariate Shift Detection Performance on the PACS Benchmark. All Methods Use the Same ResNet Classifier Trained on the “Photos” Domain. Results Are Shown as AUROC \uparrow / FPR95 \downarrow	83
4.3 Covariate Shift Detection Performance on the ColoredMNIST Benchmark. All Methods Use the Same CNN Classifier and Results Are Shown as AUROC / FPR95	83
5.1 List of Method Categories and Specific Methods that We Use Under Each Task Setting in Our Experiments.....	95
5.2 Source (In-Domain) Accuracy and Domain Generalization (OOD Accuracy) on the Digits Benchmark with MNIST-10k as Source Dataset.	96

Table	Page
6.1 Illustration of Question Composition in VQA-Compose. QF: Question Formula, AF: Answer Formula	111
6.2 Comparison of LXMERT and LOL Trained on VQA Data, Combinations with Compose, Supplement, and Our Frechet-Compatibility (FC) Loss	121
6.3 Validation Accuracies (%) for Compositional Generalization. Note that 50% Is Random Performance	122
6.4 Validation Accuracies for Compositional Generalization and Commutative Property. Note that 50% Is Random Performance	122
6.5 Performance on ‘Test-Standard’ Set of VQA-V2 and Validation Set of Our Datasets. LOL Performance Is Close to SOTA on VQA-V2, but Significantly Better at Logical Robustness.....	127
7.1 Illustrative Examples for the Effect of Each SISP Transformation on Input Sentences.....	133
7.2 Number of SISP-Transformed Samples Generated Per Category for the NLVR2 Dataset.....	142
7.3 Number of SISP-Transformed Samples Generated Per Category for the VIOLIN Dataset.....	143
7.4 Number of SISP-Transformed Samples Generated Per Category for the VQA Yes-No Dataset.....	144
7.5 Text-Only Evaluation of Biases Due to SISP Transformations. 50% Indicates No Bias.....	145
7.6 Human Validation of Our SISP Transforms Split According to the Category of Transformation.....	147

Table	Page
7.7 Human Validation of Our SISP Transforms Split According to the GT Label of the Original Sample.	147
7.8 Human Validation of Our SISP Transforms Split According to the GT Label of the Original Sample.	148
7.9 Results on the NLVR2 Public Test Set.	150
7.10 Results on VIOLIN Test Set.	151
7.11 Results on the VQA Yes/no Subset. Not to Be Compared with VQA-V2 Leaderboard Since We Use a Smaller Training Set.	152
7.12 Evaluation of NLVR2 Baselines on SISP Test Samples.	153
7.13 Evaluation of VIOLIN Baselines on SISP Test Samples.	154
7.14 Evaluation of VQA Yes/No baselines on SISP test samples.	156
7.15 Evaluation of SDRO Models NLVR ² SISP Test Samples.	157
7.16 Evaluation of SDRO Models on VIOLIN SISP Data	158
7.17 Evaluation of SDRO models on VQA Yes/No SISP Data.	159
7.18 Comparison of Performance when Only SP, Only SI, or Both Types of Transformations Are Performed.	163
7.19 Comparison of Performance if Only Positive Samples Are Used as Inputs for SISP Transformations.	163
7.20 Performance on “Text-Attack” of NLI Test Set.	164
8.1 Examples of Commonsense Annotations (Intentions, Attributes and Effects) Retrieved from ATOMIC for Captions in MSR-VTT.	177

Table	Page
8.2 Evaluation of V2C Completion Task Using CIDEr, BLEU, Rouge, and Meteor Metrics. We Use Only BLEU-1 to Evaluate the Attribute Generation Since the Average Length of the Ground Truth Is Just Less Than 2	178
8.3 Human Evaluation Scores for V2C. Captions Are an Input for the V2C-Completion Task, and Generated for the V2C-Generation Task. The Best Model Is Given in Bold, while the Overall Best Is Underlined.....	180
8.4 Illustrative Samples Generated by Our V2C-Transformer Model on V2C-Completion Task.	190
8.5 Examples of Open-Ended V2C-QA Samples.....	192
8.6 Precision (p) and Recall (r) for V2C-QA for Each Type of Question....	194
9.1 Examples Text Inputs from the SR2D Dataset for a Pair of Objects (A, B) and Relationship R Between Them.....	202
9.2 s/Δ_s Scores for T2I Metrics Shown in the 0 to 1 Range. All Previous Metrics Have Low Δ_s (Magenta) Whereas VISOR Has High Δ_s (Green), Showing They Are Ineffective in Quantifying and Benchmarking Spatial Understanding. s/Δ_s for All VISOR Variants Are in Supp.Mat.	209
9.3 Comparison of the Performance of All Models in Terms of Object Accuracy (OA) and Each Version of VISOR.	210
9.4 Majority / Unanimous Inter-Worker Agreement (%) for Each Question in Our Human Study.	214
9.5 Agreement(%) of Human Responses with Automated Metrics	214
9.6 Comparison of Visor and OA Split by Relationship Type.	221

Table	Page
9.7 Consistency of Generated Spatial Relationships for Equivalent Inputs. Bold: Highest, Underline: Lowest Consistency.	225
9.8 Effect of Prompt Variations on OA and VISOR Scores. All Three Versions Use the Same Stable Diffusion (SD) Model	226

LIST OF FIGURES

Figure	Page
1.1 Cave Paintings from All Around the World Depict Human Life in Prehistoric Times -- Interactions with Animals, Hunting and Collaborative Scenes, and a Community of People. These Cave Paintings Are an Ancient Example of Humans Storing Their Visual Memories as Images.	2
1.2 A Pyramidal Hierarchy of Visual Understanding. This Thesis Largely Deals with Semantic Vision, Especially Designative and Communicative Aspects of Semantic Vision.....	3
1.3 Different Tasks Within the Umbrella of Computer Vision Seek to Understand the Same Image in Different Ways.	5
1.4 Illustration of the Discrepancies Between Training Data and Real-World Test Data. A <i>Robust</i> Image Classifier Is Expected to Perform Reliably on a Wide Range of Image Styles and Sources.	7
1.5 Noise, Blur, Weather, or Digital Artifacts Can Also Impact Classifier Performance at Test-Time.	8
1.6 Digit Classifiers Trained on MNIST Images Suffer Significant Performance Degradation when They Are Evaluated on Real World Digit Images Obtained from the Post Office, Street Signs, Car Plates and Housing Number Plates, or when Digits Have Different Colors and Backgrounds, or Varying Amounts of Rotation, Translation and Scaling.	9
1.7 Aspects of Generalization in VQA.	14
2.1 Overview of the Problem Setup and Our Attribute-Guided Adversarial Training Method.	26

Figure	Page
2.2 Sample Images from the Training and Test Splits for Robustness Experiments on the CLEVR-Singles Dataset. The First Row Shows the Train-Test Split on the Attributes <i>Size+position</i> , and the Second Row for <i>Material+position</i>	34
2.3 Examples of Images from CLEVR-Singles and the Color Labels and (Size, Shape, Material, Position) Attributes.	35
2.4 RTS-Perturbed MNIST Images	36
2.5 Images Generated by AttGAN for the Images in Column 1, Conditioned on Attributes.	37
2.6 Visualization of the Effect of Weight β of the Constraint Loss ℓ_{const} on the Generated Images. Row 2 Has Higher β Than Row 1. Illustration Also Shows that AttGAN Is Able to Generate Multiple Objects (of Same Color for Row 1 and Different Colors for Row 2), though Absent in Training Data.	40
2.7 Comparison of Random RTS Accuracies when Controlling Each Parameter to a Max. Value. Left: R, Center: T, Right: S	42
2.8 Examples of Images Generated by AGAT for the CLEVR-Singles Dataset. The AttGAN Generator Is Able to Explore the Attribute Space and Generate Images with Different Attributes, as Well as Novel Scenes Such as those Containing Multiple Objects.	48
2.9 Examples of Images Generated by AGAT for MNIST. AGAT Is Able to Explore the Attribute Space and Generate Images with Different Rotation, Translations, and Skews.	49

Figure	Page
2.10 Examples of Images Generated by AGAT for the CIFAR-10 Dataset. AGAT Is Able to Explore the Attribute Space and Generate Images Having Varying Degrees of Noise and Blur, Which Helps Improve Robustness on All 15 Categories of Corruptions.	49
3.1 ALT Consists of a <i>Diversity</i> Module (Data Augmentation Functions Such as Augmix or RandConv and an <i>Adversary</i> Network (to Learn Image Transformations that Fool the Classifier). We Show an Ex- ample from the PACS Benchmark Under the Single-Source Domain Generalization Setting, with Real Photos (P) as the Source Domain and Art Paintings (A), Cartoons (C), and Sketches (S) as the Target Domains. The Plot Summarizes Our Results -- while Diversity Alone Improves Performance Over the Naive ERM Baseline, Adapting this Diversity Using Adversarially Learned Transformations (ALT) Provides a Significant Boost for Domain Generalization on Multiple Benchmarks.	52
3.2 (<i>Left</i>) TSNE Plot Showing the Discrepancy Between the Source Distri- bution (MNIST) and the Out-Of-Distribution Datasets for the “Digits” Benchmark. The Diversity Introduced by ALT Is Much Larger and Wide-Spread Than Data Augmentation Techniques Such as RandConv. (<i>Right</i>) Qualitative Comparison of PACS Images Transformed by Rand- Conv Data Augmentation Vs. ALT ($ALT_{RandConv}$), Illustrating the Wide Range of Transformations Learned by ALT.	70

Figure	Page
3.3 Analysis: We Study the Effect of Each Hyper-Parameter in ALT on the Average Accuracy Using the Digits Benchmark (Shown as 1 Standard Deviation Around the Mean Over 5 Runs). We Observe that the Consistency (<i>Left</i>) Is Generally Important Until a Certain Point, after Which It Becomes Harmful; (<i>Middle</i>) Taking More Adversarial Steps Improves Performance; (<i>Right</i>) Surprisingly, We Find that the Trade-Off Between Diversity and Adversity Is Non- Trivial and Dataset Dependent. In Our Benchmarking Experiments We Do Not Perform Any Hyper-Parameter Tuning, and Set $w_r = 1$, I.e. Equal Weight to Adversity and Diversity	70
4.1 Domain Generalization (DG) and Covariate Shift Detection (CSD) Are Both Important, but Orthogonal Aspects of Robustness Evaluation. While the Aim of DG Is to Predict the Correct Label for Inputs from Unseen Domains, the Aim of CSD Is to Detect Unseen Domains -- I.e. Detect Covariate Shift in Test Inputs.	77
4.2 Summary of Results on Covariate Shift Detection Benchmarks for Image Classification, in Terms of AUROC (<i>Higher Value Is Better</i>) and FPR95 (<i>Lower Value Is Better</i>). Detailed Results Can Be Found in the Appendix. The Key Observation Is that Recent OOD Methods Perform Worse Than the Baseline MSP, in Terms of Both AUROC and FPR95. Our DIS Method Improves CSD Detection Performance on All Three Benchmarks.	81

Figure	Page
5.1 Our Toy Experimental Setting Consists of Points in \curvearrowright^2 Belonging to Two Classes (0/1). This Illustration Shows the Discrepancy Between the Source Dataset (SS) and the Out-Of-Domain Dataset (OOD).	93
5.2 This Figure Illustrates the Effect of Data Modification Techniques on the Training Distribution. The Leftmost Figure Shows the Training Distribution in the Single-Source Setting. The Introduction of a Second Dataset or Data-Augmentation (Done Using Small Perturbations of Source Samples with Gaussian Noise) Makes the Distribution More Diverse in the Multi-Source (MS) and Data Augmentation (DA) Setting Respectively. On the Other Hand, Data Filtering, in Order to Remove Spurious Correlations from the Dataset, Removes Points from Certain Sectors of the Distribution. The Effect of Each Strategy on OOD Generalization and Robustness Is Shown Below Each Plot.	94
5.3 Evaluation of Adversarial Robustness (Using 10 Attack Methods) for MNIST10k.	96
5.4 Comparison Between SS and DF Models Trained with Different Percentages of MNIST10k.	98
5.5 Pearson Correlation Between OOD Accuracy and Robustness for SS and DF Models on MNIST10k.	99
6.1 State-Of-The-Art Models Answer Questions from the VQA Dataset (Q_1, Q_2) Correctly, but Struggle when Asked a Logical Composition Including Negation, Conjunction, Disjunction, and Antonyms. We Develop a Model that Improves on this Metric Substantially, while Retaining VQA Performance.	106

Figure	Page
6.2 Some Questions in VQA-Supplement Created with Adversarial Antonyms.	112
6.3 LOL Model Architecture Showing a Cross-Modal Feature Encoder Followed by Our Question-Attention (q_{ATT}) and Logic Attention (ℓ_{ATT}) Modules. The Concatenated Output of Is Used by the Answering Module to Predict the Answer.....	116
6.4 Learning Curve Comparison for Models (Red: LXMERT, Blue: LOL) Trained on Our Datasets (Solid Lines: VQA + Comp, Dotted Lines: VQA + Comp + Supp)	120
6.5 Accuracy for Each Type of Question in (a) VQA-Compose, (b) VQA- Supplement and for Questions with Number of Operands Greater Than 2.....	125
7.1 VLI Models Predict Whether a Sentence Is True or False About the Visual Input. (<i>Top</i>) Sample from NLVR ² with Two Images as Input; (<i>Bottom</i>) Sample from VIOLIN with Video and Subtitles as Input.	130
7.2 Comparison between (<i>left</i>) ϵ -bounded image perturbations and (<i>right</i>) linguistics-based <i>Semantics-preserving</i> (blue) as well as <i>Semantics-</i> <i>Inverting</i> (red) transformations for sentences.	131
7.3 Illustration of the Work-Flow for Generating SISP-Transformed Versions of Input Sentences. A Semantics-Preserving (SP) Transformation Is Shown Above.	139
7.4 Illustration of the Work-Flow for Generating SISP-Transformed Versions of Input Sentences. A Semantics-Inverting (SI) Transformation Is Shown Above.	140

Figure	Page
7.5 Snapshot of a SISP Example Being Evaluated by Human Subjects. Columns from Left to Right: Sample-ID, SISP-Tag, Left Image, Right Image, Original Sentence, Original Label, New Sentence, New Label . . .	141
7.6 Comparison of Original Sentences (Black) with (<i>Left</i>) SISP-Transformed Sentences (Blue) and (<i>Right</i>) ϵ -Bounded Perturbations as a TSNE Plot.	155
7.7 Original Test Inputs for NLVR ² with Their Respective SP (Green) and SI (Yellow) Test Samples and the Prediction and Confidence of Models with VILLA Backbone. Wrong Predictions Are Highlighted in Red.	155
7.8 Comparison of Reliability Curves on the Clean Test Set (<i>Left</i>) and SISP Test Set (<i>Right</i>).	160
7.9 Effect of Size of Training Data (<i>Left</i>) NLVR ² , (<i>Right</i>) VIOLIN. SDRO Models Are Consistently Better Than Baselines, Even in Low-Data Settings.	161
7.10 Plots Showing the Effect of the Percentage of Augmented Samples on Clean, SP, and SI Accuracies on NLVR ² , when Using Data-Augmentation, and SDRO.	161
7.11 Plots Showing the Effect of the Percentage of Augmented Samples on Clean, SP, and SI Accuracies on VIOLIN, when Using Naive Data- Augmentation, SW-SDRO, and GW-SDRO.	162
7.12 Plots Showing the Effect of the Percentage of Augmented Samples on Clean, SP, and SI Accuracies on VQA Yes/No, when Using Naive Data-Augmentation, SW-SDRO, and GW-SDRO.	162

Figure	Page
8.1 Comparison of Conventional Video Captioning with Our Commonsense-Enriched Captioning. Our Captions Describe Intention Behind the Action (Red), Attribute of the Agent (Blue), and Effect of the Action on the Agent (Green).....	169
8.2 The V2C-Transformer Model Architecture Contains: (a) Video Encoder Designed to Take Video Frames as Input and Encode Them Into Frame-Wise Representations, (b) Decoder Module Consisting of a Caption Decoder and a Commonsense Decoder, and (c) Transformer Decoder Module Containing a Stack of N Consecutive Transformer Blocks (Shown Inside the Dashed Area).....	172
8.3 The Overall Three-Step Pipeline (Retrieval from ATOMIC, BERT Re-Ranking, and Human Labeling) to Construct Our V2C Dataset.	175
8.4 The Data Creation Flow for V2C. We Use the Retrieved Videos and Captions from MSR-VTT and Use the BERT Re-Ranking Module to Obtain a List of Top-3 Intentions (I), Effects (E), and Attributes (A). These Are then Further Improved by Human Labeling. A Subset of Annotations Is Also Converted to Full Sentences by Human Annotators.	175
8.5 Our Human Labeling Interface. We Ask Human Workers to Select Relevant Commonsense Descriptions as Well Provide Additional Texts in Their Own Words.....	179
8.6 Word Cloud Figure of the Intention Commonsense Annotations from Our V2C Dataset.	180
8.7 Top-100 Most Frequent Words in Our V2C Dataset (Stop Words Are Ignored).	181

Figure	Page
8.8 Qualitative Examples of Our V2C Dataset.	182
8.9 Examples of Outputs of Our Model for the V2C Completion and Generation Tasks Along with the Ground-Truth (GT) Caption. A Failure Example Shown in the Bottom Red Box.	185
8.10 Snapshot of Our AMT Human Evaluation Interface for V2C-Completion Task.....	187
8.11 Snapshot of Our AMT Human Evaluation Interface for V2C-Generation Task.....	188
8.12 Example Questions from V2C-QA Compared with Conventional Video Question Answering.	191
9.1 We Benchmark T2I Models on Their Competency with Generating Appropriate Spatial Relationships in Their Visual Renderings. Although Text Inputs May Explicitly Mention these Spatial Relationships, T2I Models Lack Such Spatial Understanding.	199
9.2 Examples Illustrating the Intuition Behind OA, VISOR, VISOR_cond, and VISOR_1/2/3/4. Purple Box: Cases where One or Both Objects Are Not Generated; Red Box: Both Objects Are Generated but with a Wrong Spatial Relationship; Green Box: Successful Cases.	206
9.3 For Text t and Corresponding Generated Image $x = g(T)$, Object Centroids Are Located and Converted Into Predicates Indicating the Spatial Relationship Between Them. These Predicates Are Compared with the Ground Truth Relationship R to Obtain the VISOR Score.	207
9.4 The Human Study Interface with an Image on the Left and Seven Multiple Choice Questions About It.	212

Figure	Page
9.5 Summary of Responses to Each Question in the Human Study, Compared Across All Four Models.	213
9.6 Illustrative Examples of Text Prompts from Our SR2D Dataset and Corresponding Images Generated by Each T2I Model.	215
9.7 Illustrative Examples of Images Generated by Each of the 7 Benchmark Models Using Text Prompts (Top Row) from the SR2D Dataset.	216
9.8 Illustrative Examples of Images Generated by Each of the 7 Benchmark Models Using Text Prompts (Top Row) from the SR2D Dataset.	217
9.9 Illustrative Examples of Images Generated by Each of the 7 Benchmark Models Using Text Prompts (Top Row) from the SR2D Dataset.	217
9.10 Illustrative Examples of Images Generated by Each of the 7 Benchmark Models Using Text Prompts (Top Row) from the SR2D Dataset.	218
9.11 Illustrative Examples of Images Generated by Each of the 7 Benchmark Models Using Text Prompts (Top Row) from the SR2D Dataset.	218
9.12 Illustrative Examples of Images Generated by Each of the 7 Benchmark Models Using Text Prompts (Top Row) from the SR2D Dataset.	219
9.13 Illustrative Examples of Images Generated by Each of the 7 Benchmark Models Using Text Prompts (Top Row) from the SR2D Dataset.	219
9.14 Illustrative Examples of Images Generated by Each of the 7 Benchmark Models Using Text Prompts (Top Row) from the SR2D Dataset.	220
9.15 Illustrative Examples where the Two Objects from the Text Input Appear to Be Merged. From Left to Right: A,b,c,d.	220
9.16 VISOR Scores for Each Supercategory Pair.....	221

Figure	Page
9.17 Correlation of Our Metrics with P_COCO, the Object Co-Occurrence Probability in MS-COCO.	223
9.18 Comparison of Object Accuracy for Text with Single and Multiple Objects Reveals a Bias Towards Single Objects.	223
9.19 Comparison of Object Accuracy for Object A and B Reveals a Bias Towards A, the First Object Appearing in the Prompt).	224
9.20 Comparing VISOR Performance with Different Combinations of Attributes. “Z, ZC” Indicates a Prompt Describing Object A with a Size Attribute and Object B with Both Size and Color.	226

Chapter 1

INTRODUCTION

Humans, from time immemorial, have looked at the earth and seen things interacting with each other; they have looked at the skies and wondered what the stars and planets were; they have looked at each other and invented communication, co-operation, society, and culture. Humans, from a very long time ago, have expressed themselves by creating images. They have communicated with other people and people from the future – i.e. us, through these images. For instance, the oldest known figurative painting in Fig. 1.1a, dates back to at least 40,000 years ago, and shows a bovine animal; modern computer vision datasets have continued this tradition of iconic images of animals, vegetation, and other common objects that we interact with. One cave paintings from Libya (Fig. 1.1b) depicts what may be interpreted as a hunting scene, while another from Argentina (Fig. 1.1c) is made up of hands prints of real people who lived 9000 years ago – telling us a story of a human community. These cave paintings are indeed *images* that have allowed our ancestors to communicate what they saw — the environment, other creatures, other humans, and their interactions with them. These images, these paintings, these sculptures, are human memories in visual form. They maybe stories. They may be mythology. They may be history. But in essence, they are visual knowledge.

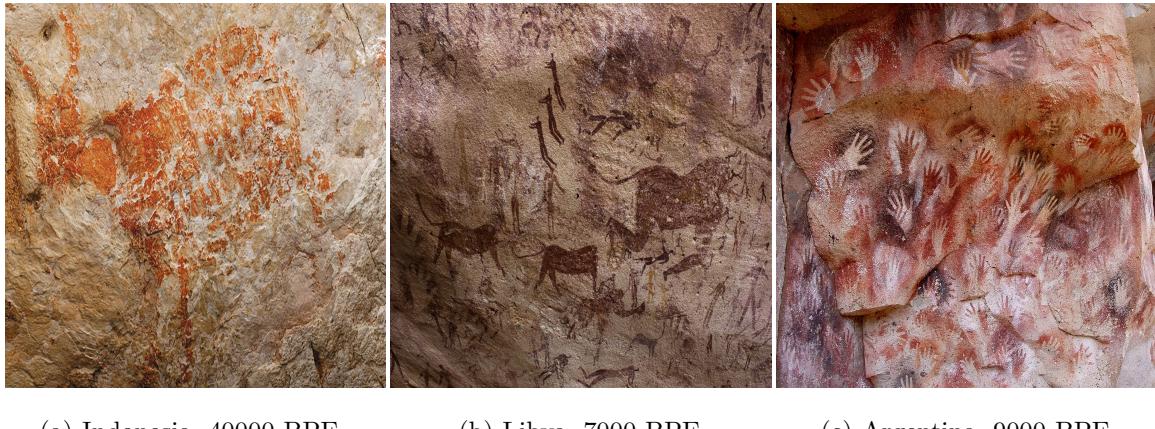
The process of observing the world and understanding these observations is arguably still the way we reason about the world. Ever since the democratization of cameras, the process of reasoning about visual observations has become computational. The pursuit of developing such computations *is* computer vision. When the outputs of the reasoning process convey meaningful concepts to humans, we shall call it semantic

visual understanding.

This introductory chapter is designed to be a bird’s-eye-view of semantic vision. First, I will introduce a hierarchy of computer vision tasks that allows us to separate semantics from physics. Second, I will briefly revisit the progress made in prototypical computer vision tasks and the benchmarks and datasets that have been established by researchers over the years for these tasks. Third, we will discuss the implications of recent successes in computer vision and motivate the core ideas of this thesis along with an outline. This will lay the groundwork for the rest of this dissertation.

1.1 Understanding What Cameras See

While computer vision traces its roots and shares many of the goals of digital image processing ([Rosenfeld, 1976](#)), in the 1970s, computer vision underwent a paradigm shift. While image processing was more focused on developing algorithms for acquisition, storage and information extraction from images, computer vision researchers are



(a) Indonesia, 40000 BPE (b) Libya, 7000 BPE (c) Argentina, 9000 BPE

Figure 1.1: Cave paintings from all around the world depict human life in prehistoric times – interactions with animals, hunting and collaborative scenes, and a community of people. These cave paintings are an ancient example of humans storing their visual memories as images.

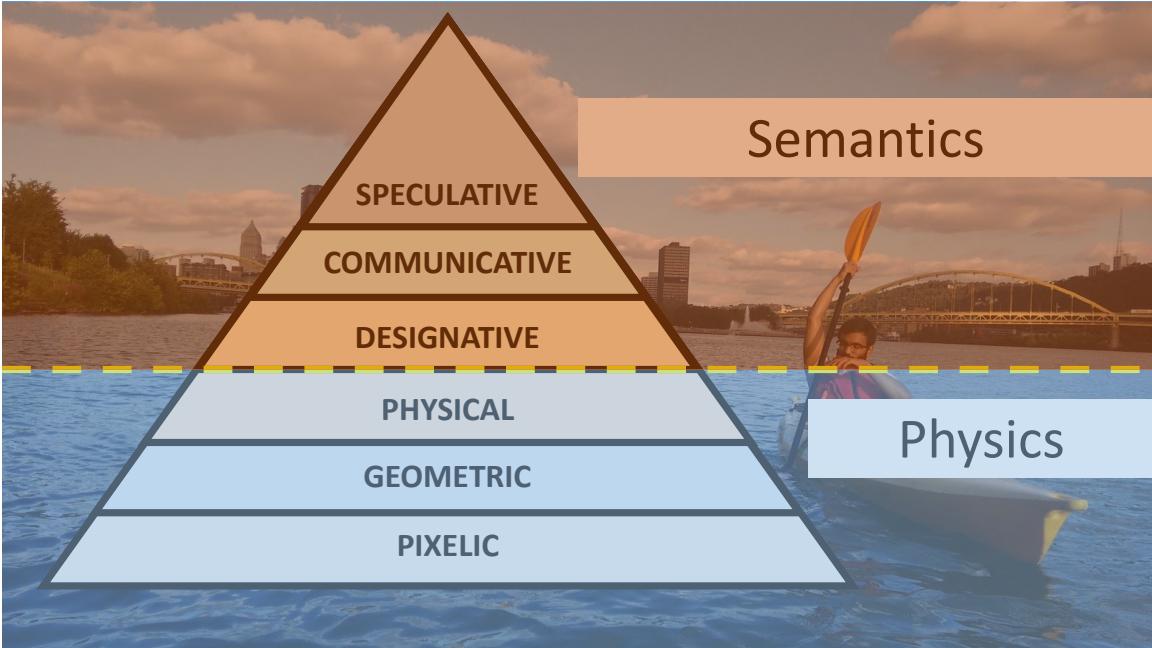


Figure 1.2: A pyramidal hierarchy of visual understanding. This thesis largely deals with semantic vision, especially designative and communicative aspects of semantic vision.

motivated by a larger goal – to *recover* information about scenes that is lost when an image is captured by a camera. This includes “*pixelic*” analysis such as edge detection and color detection; “*geometric*” properties of the scene, for instance, three-dimensional properties such as depth, surface normals, orientation, and shape; and “*physical*” properties of the scene such as estimating material reflectances, albedo, texture, etc. Computer vision also seeks to understand what objects (such as people, boats, trees, bridges, buildings) are present in an image, and where (localization). I call this category of vision, “*designative*” vision since the aim is to designate or assign names and categorical labels to the observed visual input.

In the past decade, computer vision is undergoing another paradigm shift – this time towards semantics and the explicit use of language to understand and communicate the meaning of images. Computer vision are now learning not only from datasets of images, but datasets of images with text descriptions. These models have also begun

to serve in a “*communicative*” role, for instance, the task of learning captions for images (Yang *et al.*, 2011; Karpathy and Li, 2015) allows computer vision algorithms to express the scene in a human-understandable format. Similarly, visual question answering (Antol *et al.*, 2015) enable users to ask questions about the image and get answers from the computer vision system. There has also been interest in “*speculative*” vision, sometimes also controversially referred to as “commonsense” (Zellers *et al.*, 2019; Park *et al.*, 2020) in literature; the aim is to speculate about intentions of people, their relationship with each other or with other objects, to speculate about their past or future actions and the effect of these actions, or to reason about why those actions are being performed. Another form of speculation (and perhaps reasoning) is the task of learning to generate (or render) images from a text description (Zhang *et al.*, 2017).

Figure 1.2 summarizes this hierarchy of visual understanding into six broad categories, and divides this spectrum into two main parts – semantic vision and physics-based vision. Figure 1.3 shows a few example tasks with the same input image – this is designed to help the reader understand how different levels of the pyramid focus on different aspects of image understanding.

This thesis focuses on semantic vision.

1.2 Recent Success In Visual Understanding

The fabled memo-meme from the 1950s to “solve” computer vision as a summer project (Papert, 1966) has grown as one of the biggest and most active research programs of scientific research in recent years (Su and Crandall, 2021). Computer vision occupies a large role in the popularity and applicability of machine learning algorithms. Several standards benchmarks have been established, adopted, and celebrated by the community for many vision tasks. Below, I will briefly review progress made on some of these datasets which lie in the category of “*designative*” and

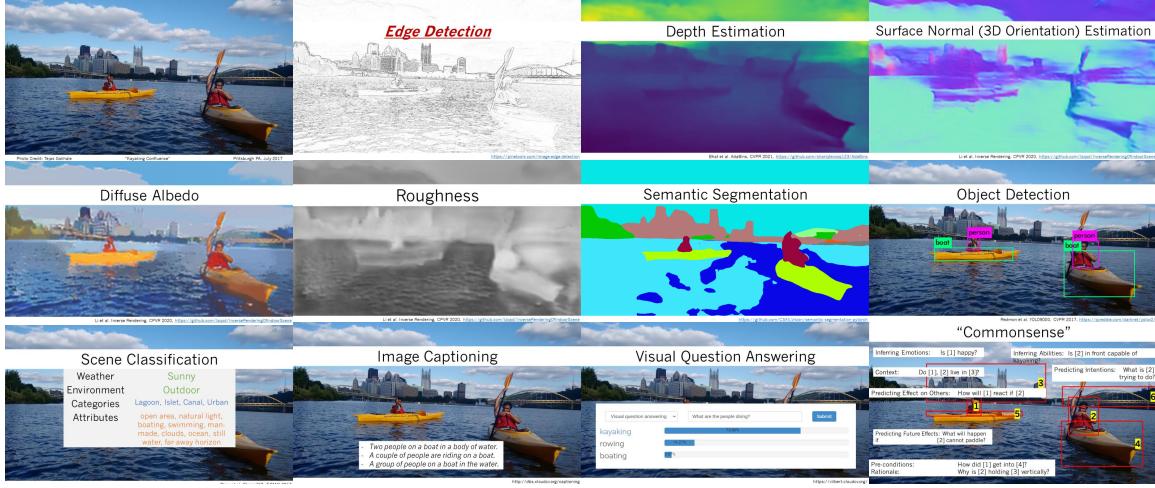


Figure 1.3: Different tasks within the umbrella of computer vision seek to understand the same image in different ways.

“communicative” vision tasks such as image classification, object detection, semantic segmentation, image captioning, and visual question answering.

Image Classification. *MNIST* (LeCun *et al.*, 1998), introduced in 1998, is a handwritten digit classification; the efficacy of the convolutional neural network architecture and learning via backpropagation (LeCun *et al.*, 1989) was demonstrated on MNIST. Today, machine learning systems are used by many businesses including the US Post Office, to flawlessly convert handwritten digits, into digital formats. *ImageNet* (Deng *et al.*, 2009), introduced in 2009, proved to be a catalyst for the emergence of neural networks as the de-facto solution for many problems in perception. Quick progress was made via AlexNet (Krizhevsky *et al.*, 2012) and ResNet (He *et al.*, 2016) reached 96.4% on the same metric. ResNet features, i.e. the features of a ResNet model pretrained on ImageNet, became a standard choice for initialization of neural networks. In 2021, the top-5 accuracy is $\sim 99\%$ ¹. *CIFAR-10* and *CIFAR-100* (Krizhevsky, 2009) were

¹performance metrics obtained from <https://paperswithcode.com/sota/>

introduced in 2009; at that time, RBM-based methods () achieved \sim 60% accuracy on CIFAR-10, while in 2021, the Vision Transformer () reports an accuracy of 99.50%.

Object Detection. *PASCAL-VOC* (Everingham *et al.*, 2010) (2010) and *MS-COCO* (Lin *et al.*, 2014) (2014) are popular benchmarks for the task of object detection – predicting the categories and locations of multiple objects in an image. While traditional approaches based on deformable parts model (Fidler *et al.*, 2013) had reached a mean average precision of \sim 40% and 29.6% on PASCAL-VOC and 19.1% on MS-COCO in 2014, in 2021, neural network object detectors report 62.4% and 89.30% on the same metric.

Semantic Segmentation. ADE20K (Zhou *et al.*, 2017) and Cityscapes (Cordts *et al.*, 2016) introduced in 2017 and 2016 respectively are popular benchmarks for semantic segmentation – the task of assigning a categorical label for each pixel in the image. Performance on ADE20K has improved from a mean-IoU of 44.94 in 2017 to 59.90 in 2021, and performance on Cityscapes has improved from 73.60 in 2017 to 84.40 in 2021.

Image Captioning. *MS-COCO* (Lin *et al.*, 2014) and *Flickr-30k* (Young *et al.*, 2014), both introduced in 2014 are popular benchmarks for the task of image captioning, i.e. generating a natural language description for an input image. In 2015, the performance in terms of BLEU-4 score (Papineni *et al.*, 2002) was 23.0 and 15.7 respectively, while in 2021, this has increased to 37.4 (Li *et al.*, 2020c) and 30.10 (Zhou *et al.*, 2020b), respectively.

Visual Question Answering. Several datasets based on images from MS-COCO, VisualGenome, etc. have been used for visual question answering, i.e., predicting

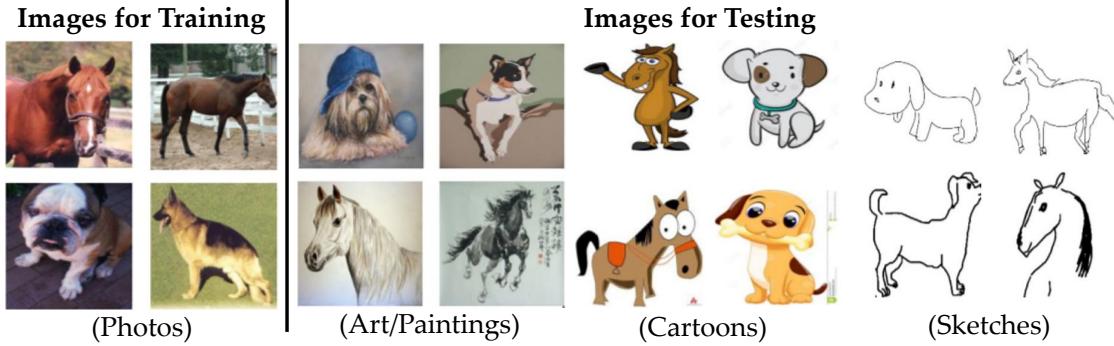


Figure 1.4: Illustration of the discrepancies between training data and real-world test data. A *robust* image classifier is expected to perform reliably on a wide range of image styles and sources. Image adapted from [Li et al. \(2017\)](#).

answers for questions about an image. VQA-v2 ([Goyal et al., 2017](#)) is a popular benchmark for this task, with accuracy improving from 62.7% in 2016 ([Fukui et al., 2016](#)) to \sim 80% in 2021 ([Zhang et al., 2021](#); [Li et al., 2020c](#)).

1.3 Robustness and Generalization

These results clearly indicate the semantic vision has seen substantial improvements over the last decade. This progress, powered largely by innovations in neural network architectures, optimization techniques, and availability of copious amounts of training data, is definitely good news. These results corroborate prior theoretical work on learnability in statistics and machine learning ([Valiant, 1984](#); [Hoeffding, 1994](#); [Hornik et al., 1989](#)). Does the performance on standard metrics and benchmarks mean that the problem of *understanding what cameras see* is close to being “solved”? Has semantic vision indeed reached (or surpassed human-level abilities) according to the above results, as some news articles claim?

The answer to these questions, unfortunately is in the negative. Although highly capable of learning from training data, recent studies show that neural networks

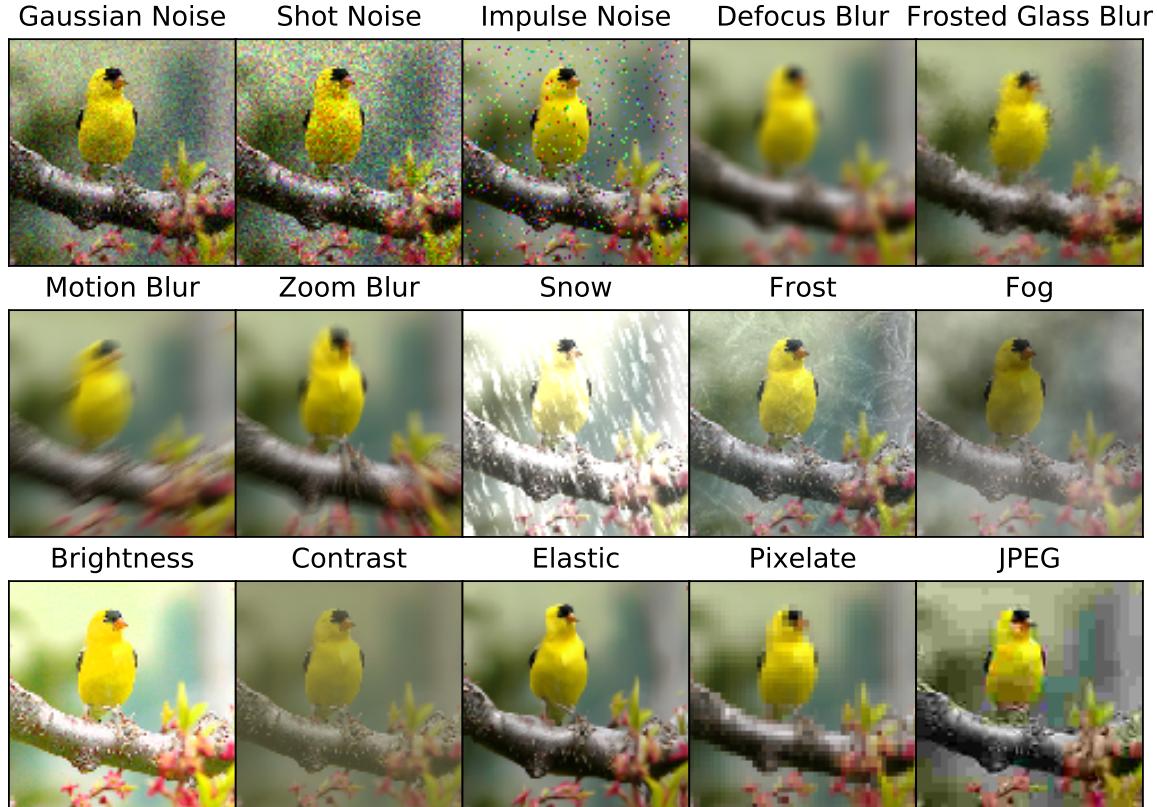


Figure 1.5: Noise, blur, weather, or digital artifacts can also impact classifier performance at test-time. Image from [Hendrycks and Dietterich \(2019\)](#).

are prone to failure on new test sets or under distribution shift ([Taori *et al.*, 2020](#)), natural corruptions ([Hendrycks and Dietterich, 2019](#)), adversarial attacks ([Goodfellow *et al.*, 2015](#)), spurious correlations ([Beery *et al.*, 2018](#)), and many other types of “unseen” changes in test samples. This shortcoming stems from the *i.i.d.* assumption in statistical machine learning which guarantees good performance only on test samples that are drawn from an underlying distribution that is identical to the training dataset.

Consider the case of digit classification – considered by some to be a “solved” task. Digit recognition models trained on the black-and-white MNIST training images are almost perfect ($> 99.5\%$ accuracy) on the corresponding *i.i.d.* test set, yet their performance on colored digits and real-world digits from street number plates is only

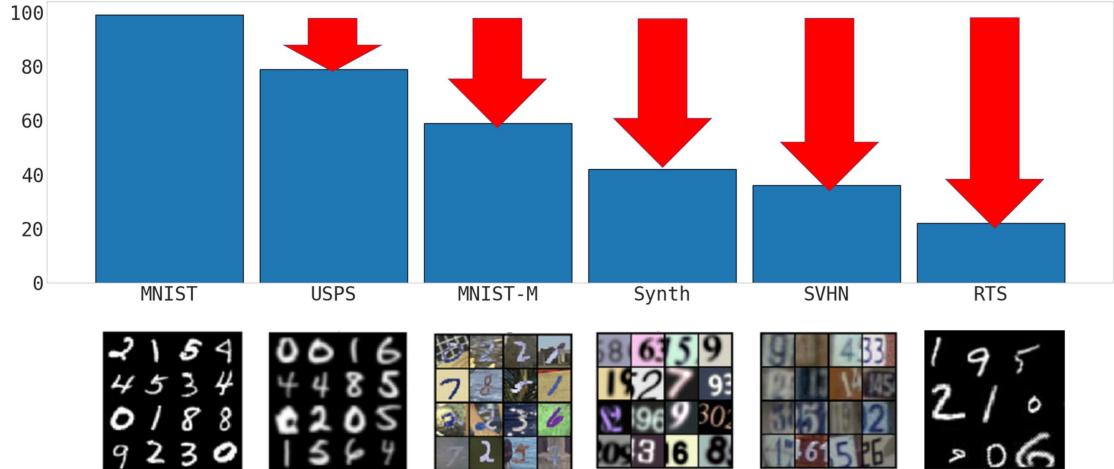


Figure 1.6: Digit classifiers trained on MNIST images suffer significant performance degradation when they are evaluated on real world digit images obtained from the post office, street signs, car plates and housing number plates, or when digits have different colors and backgrounds, or varying amounts of rotation, translation and scaling.

around 70% ([Xu et al., 2020b](#)). This catastrophic performance drop is illustrated in Fig. 1.6.

Although the accuracy of image classifiers on multiple datasets of real-world photographs such as ImageNet and CIFAR is above 90%, when these models are used for classifying other styles of images (such as cartoons, sketches, paintings, etc.) ([Venkateswara et al., 2017](#); [Li et al., 2017](#)) belonging to the same classes, a performance degradation is observed – the accuracies are as low as 24% for cartoons, 29% for sketches, 64% for paintings ([Nam et al., 2021](#)), as shown in Figure 1.4. There is also a large accuracy drop when models are tested on natural corruptions ([Hendrycks and Dietterich, 2019](#)) such as fog, snow, rain, noise, blur etc. or geometric perturbations such as rotation, translation, scaling ([Wong and Kolter, 2020](#)), as illustrated in Figure 1.5. In short, when models that have been trained in “lab settings” are tested in “real-world settings”, the levee breaks, and we start getting a glimpse into the

robustness of computer vision models. *This thesis is about identifying such modes of failure, analyzing them, and providing solutions for improving model robustness.*

1.4 Background

Engineered systems often come with conditions for operation – televisions and phones are rated for use only in a range of temperature values, and voltage standards, beyond which they can malfunction. Machine learning models are no different – they operate under a set of assumptions about the inputs. These assumptions are critical when machine learning models encounter variations in inputs at test time or during deployment. This chapter contains an overview of existing techniques that have been designed to overcome challenges in machine learning that are encountered as a result of deviations between training and test distributions. In this chapter, we will review literature on robust machine learning, and relevant work in image classification and visual question answering, which will serve as a starting point for subsequent chapters.

1.4.1 Robust Machine Learning

Robustness is an overloaded word in machine learning and is often used with inconsistent connotations by different researchers. Before we get into the math, it makes sense to develop a simple intuitive definition that we will stick with in this dissertation. This simple definition is as follows: machine learning models are trained on a certain dataset but may encounter different types of changes at test time and with varying degrees – a robust machine learning model is able to reliably produce outputs irrespective of such changes in inputs. When these changes are designed (either by humans or by algorithms) to fool the model into making incorrect predictions, we will denote such inputs as adversarial examples; the property of a model which is robust under such circumstances will be denoted as “*adversarial robustness*”. When

the entire input distribution shifts in meaningful ways (for instance, grayscale images vs. color images, or photos vs. sketches, or even new categories or classes), we will denote this as distribution shift, and the property of models that are robust to such shifts as “*distributional robustness*”.

Consider a training distribution P_{tr} consisting of inputs \mathbf{x} and labels \mathbf{y} . Under the empirical risk minimization (ERM), the following risk is minimized:

$$\mathcal{R}_{ERM} = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim P_{tr}} \ell(f(\mathbf{x}; \theta), \mathbf{y}). \quad (1.1)$$

ERM provides generalization guarantees (Vapnik, 1991) for i.i.d. test samples, but not for out-of-distribution or adversarial examples (Biggio *et al.*, 2013; Szegedy *et al.*, 2014).

Distributed Robust Optimization (DRO) (Hu *et al.*, 2018a; Sagawa *et al.*, 2020) searches for loss-maximizing perturbations of the input within an ϵ -divergence ball around P_{tr} and minimize the risk over such perturbed distributions.

$$\mathcal{R}_{DRO} = \sup_{P: D(P, P_{tr}) < \epsilon} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim P} \ell(f(\mathbf{x}; \theta), \mathbf{y}). \quad (1.2)$$

The solution to Equation 1.2 guarantees robustness inside such ϵ -bounded distributions P . The inner maximization is typically solved using gradient-based methods (Madry *et al.*, 2018a) over additive perturbations δ such that $\mathbf{x} + \delta$ fools the classifier.

Adversarial Examples. To formally define adversarial examples, first an *attack model* is specified – this provides a precise definition of the input perturbations that we want robustness against. Given an input x , we define this set of allowed perturbations $\mathcal{S} \subseteq \mathcal{X}^d$. For image classification, \mathcal{S} is typically chosen to be the ℓ_∞ -ball around x (Goodfellow *et al.*, 2015). Defense against is formulated by Madry *et al.* (2018a) as a min-max optimization, where the inner maximization allows the attack model to

maximally fool the classifier, and the outer minimization of the classifier loss uses the resulting adversarial examples to update model parameters.

$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\delta \in \mathcal{S}} \ell(\theta, x + \delta, y) \right]. \quad (1.3)$$

Domain Generalization has been explored under both multi-source (MSDG) and single-source (SSDG) settings. Techniques designed for MSDG seek to utilize the multiple domains to perform feature fusion ([Shen et al., 2019](#)), learning domain-invariant features ([Ganin et al., 2016](#)), meta-learning ([Li et al., 2018a](#)), invariant risk minimization ([Arjovsky et al., 2019](#)), learning mappings between multiple training domains ([Robey et al., 2021](#)), and style randomization ([Nam et al., 2021](#)). [Gulrajani et al.](#) ([Gulrajani and Lopez-Paz, 2021](#)) provide an extensive comparative study of these approaches and report that simply performing ERM on the combination of source domains leads to the best performance. Many benchmarks have been proposed to evaluate MSDG performance such as PACS ([Li et al., 2017](#)), OfficeHome ([Venkateswara et al., 2017](#)), Digits ([Volpi et al., 2018](#)), and WILDS ([Koh et al., 2021](#)) which is a compendium of MSDG datasets for various tasks such as image classification, text sentiment classification, text toxicity prediction, etc. In the context of multi-source DG, ([Zhou et al., 2020a](#)) propose to synthesize novel domains using a conditional generator trained on multiple domains using cycle consistency – whereas we are primarily interested in the single source setting where such a method may not be feasible. Moreover, we strictly synthesize novel domains as functions of the source domain, and place emphasis on the nature of functions that are learnable during training with a convolutional network with objectives such as an adversarial cost and consistency measures.

SSDG is a harder setting due to the lack of multiple datasets for using the above methods; most work in SSDG has therefore focused on data augmentation. Notable

among these methods is the idea of adversarial data augmentation – ADA ([Volpi *et al.*, 2018](#)) and M-ADA ([Qiao *et al.*, 2020](#)) apply pixel-level additive perturbations to the image in order to fool the classifier. Resulting images are used as augmented data to train the classifier. RandConv ([Xu *et al.*, 2020b](#)) shows that shape-preserving transformations in the form of random convolutions of images lead to impressive performance gains on Digits.

Robustness to Image Corruptions. There has also been interest in training classifiers that are robust to corruptions that occur in the real world, such as different types of noise and blur, artifacts due to compression techniques, and weather-related environments such as fog, rain, and snow. ([Vasiljevic *et al.*, 2016](#); [Geirhos *et al.*, 2018](#)) show that training models with particular types of corruption augmentations does not guarantee robustness to other unseen types of corruptions or even different severities of corruptions. Hendrycks *et al.* ([Hendrycks and Dietterich, 2019](#)) curate benchmarks (ImageNet-C and CIFAR-C) to test robustness along a fixed set of corruptions. They also provide a benchmark called ImageNet-P which tests robustness against other corruption types such as small tilts and changes in brightness. A similar benchmark for corruptions of handwritten digit images, MNIST-C ([Mu and Gilmer, 2019](#)) has also been introduced.

1.4.2 Robust Natural Language Understanding

Generation of semantics-preserving adversarial examples ([Jia and Liang, 2017a](#); [Ribeiro *et al.*, 2018a](#); [Iyyer *et al.*, 2018a](#); [Alzantot *et al.*, 2018a](#)), and approaches to defend against word substitution ([Jia *et al.*, 2019a](#)) have been explored. Evaluation datasets have also been proposed for textual entailment that are manually crafted ([Gardner *et al.*, 2020a](#)) or template-based ([McCoy *et al.*, 2019a](#); [Glockner *et al.*, 2018a](#);

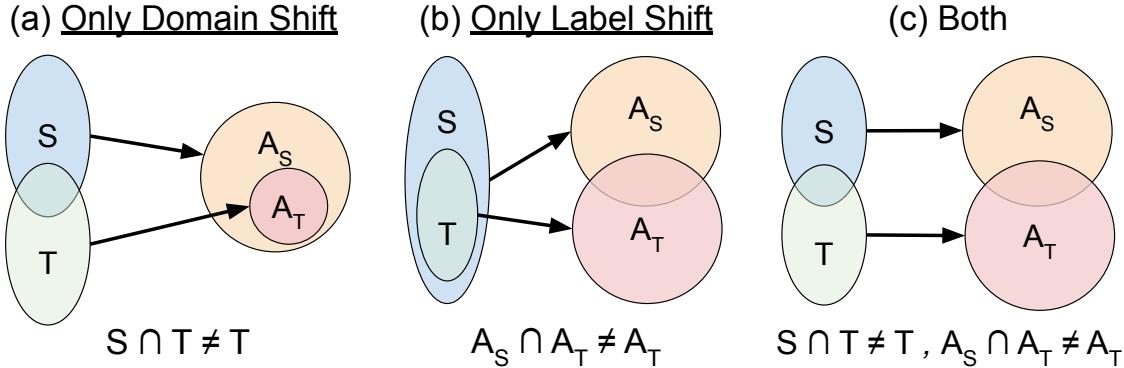


Figure 1.7: Aspects of generalization in VQA.

Naik *et al.*, 2018a). Belinkov and Bisk (2018); Ebrahimi *et al.* (2018); Jones *et al.* (2020) investigate the practical problem of typographical errors in NLP systems. While humans seem to understand most spelling or grammar errors when presented with context, this is not true for existing NLP systems. Zhao *et al.* (2017); Hendricks *et al.* (2018); Rudinger *et al.* (2018) show that NLP systems can propagate and amplify socio-cultural biases. Jia *et al.* (2019b) have shown how interval bound propagation (Dvijotham *et al.*, 2018) can be used to defend text classification models from adversarial word substitutions. Inspired by the work on adversarial training in computer vision, several optimization strategies have been developed for fine-tuning pre-trained language models (Miyato *et al.*, 2017; Oren *et al.*, 2019; Jiang *et al.*, 2020; Zhu *et al.*, 2020).

1.4.3 Robustness of Vision-Language Models.

The presence of two modalities in V&L tasks implies that we need to talk about robustness and generalization of V&L modalities from the perspective of vision, and from the perspective of language. Most of the robustness literature in V&L has been for the task of visual question answering. We review it below.

Robustness in VQA can be defined as shown in Figure 1.7 under two situations:

domain shift and label shift. Under domain shift, generalization to a new input domain (such as different styles of questions or novel scenes) is desired, characterized by $S \cap T \neq T$ where S and T denote the train and test input domains. Under label shift, generalization to novel answers is desired (predicting answers not seen during training), characterized by $A_S \cap A_T \neq A_T$, where A_S and A_T are the set of answers seen during training and test-time.

Performance under **domain shift** has been evaluated for new domains of test questions with unseen words and objects (Teney and Hengel, 2016; Ramakrishnan *et al.*, 2017), novel compositions (Johnson *et al.*, 2017; Agrawal *et al.*, 2017), logical connectives (Gokhale *et al.*, 2020c), as well as questions that are implied (Ribeiro *et al.*, 2019a), entailed (Ray *et al.*, 2019) or sub-questions (Selvaraju *et al.*, 2020); or for datasets with varying linguistic styles (Chao *et al.*, 2018; Xu *et al.*, 2020a; Shrestha *et al.*, 2019) and different reasoning capabilities (Kafle and Kanan, 2017). Other work seeks to answer target questions that are sub-questions (Selvaraju *et al.*, 2020), or are implied (Ribeiro *et al.*, 2019a) or entailed (Ray *et al.*, 2019) by source questions. More recently, a new benchmark has been proposed that combines many of the papers mentioned above into a unified evaluation dataset for testing robustness of VQA models (Li *et al.*, 2020b).

Label shift or Prior Probability Shift (Storkey, 2009) has been implicitly explored in VQA-CP (Agrawal *et al.*, 2018), where the conditional probabilities of answers given the question type deviate at test-time. A similar benchmark for the spatial visual question answering task has been introduced via GQA-OOD (Kervadec *et al.*, 2021), which focuses on rare question-answer pairs, and showed that existing VQA models overfitted to dataset biases and underperformed on such OOD questions. Similar to GQA-OOD, the VQA-CE dataset (Dancette *et al.*, 2021) generated a new evaluation

set by using rules to mine questions that can fool existing models. As for GQA-OOD, this dataset demonstrated that SOTA models do not perform well when they can't rely on shortcuts, even models which use bias-reduction techniques. A comprehensive survey of methods that tackle such dataset biases and priors is given by [Shrestha *et al.* \(2022\)](#) and [Teney *et al.* \(2020b\)](#).

1.5 Overview: Towards Reliable Visual Understanding

The recent findings about the brittleness and high sensitivity of models on real-world data pose a significant challenge to the practical adoption of computer vision models and their reliability in the real-world, especially when dealing with sensitive data such as biomedical and satellite imagery, personal information, private records, etc. In this dissertation, I will address these shortcomings in the domains of image classification as well as multi-modal visual understanding *a.k.a.* vision-and-language (V&L). My work complements the new innovations of the past decade, by studying their robustness and generalization capabilities. This involves:

1. **identifying failure modes**, i.e., situations under which systems may fail, for perceptual tasks (such as image classification) and semantic tasks (such as visual question answering, visual reasoning, and image/video captioning),
2. creating **evaluation and analysis tools** to diagnose failures, achieved by creating datasets for targeted evaluation, probing models with specific types of input perturbations, evaluating generalization under domain shift or across datasets, etc.
3. **developing techniques** that provide greater robustness to mitigate the risks posed by such situations.

In Chapter 2, we consider a setup where robustness is expected over an unseen test

domain that is not *i.i.d.* but deviates from the training domain. While this deviation may not be exactly known, its broad characterization is specified *a priori*, in terms of attributes. We propose an adversarial training approach which learns to generate new samples so as to maximize exposure of the classifier to the attributes-space, without having access to the data from the test domain. Our approach enables deep neural networks to be robust against a wide range of naturally occurring perturbations. We demonstrate the usefulness of the proposed approach by showing the robustness gains of deep neural networks trained using our adversarial training on MNIST, CIFAR-10, and a new variant of the CLEVR dataset. This work was published as a conference paper in AAAI 2021 ([Gokhale et al., 2021](#)).

To be successful in single source domain generalization, maximizing diversity of synthesized domains has emerged as one of the most effective strategies. Many of the recent successes have come from methods that pre-specify the types of diversity a model is exposed to during training so that it can ultimately generalize well to new domains. However, naïve diversity based augmentations do not work effectively for domain generalization either because they cannot model large semantic shifts, or the span of transforms that are pre-specified, do not cover the semantic shifts commonly occurring in domain generalization.

To address this issue, in Chapter 3, we present a novel framework that uses adversarially learned transformations (ALT) using a neural network to model plausible, yet hard image transformations that fool the classifier. This network is randomly initialized for each batch and trained for a fixed number of steps to maximize classification error. With extensive empirical analysis, we find that this new form of adversarial transformations achieve both objectives of diversity and hardness together, outperforming all existing techniques on competitive benchmarks for single source domain generalization. We also show that ALT can naturally work with existing

diversity modules to produce highly distinct, and large transformations of the source domain leading to state-of-the-art performance. This work was published at WACV 2023 ([Gokhale et al., 2023](#))

In AGAT and ALT, we developed methods to improve robustness to domain shift. However, for transparent decision making, it is equally important to develop methods that can indicate whether or not a test image belongs to an unseen domain. Domain generalization algorithms are trained with the aim of making accurate predictions for seen as well as unseen domains. However in cases where accuracy under domain shift is low, it is equally important to detect, or flag, cases where there may be a domain shift for safe and reliable use of classifiers. These types of flagging mechanisms will be explored in Chapter 4 where we propose an interpolation-based method to detect domain shift and flag such potentially OOD samples. This work was presented at the Neurips 2022 Workshop on Interpolation ([Gokhale et al., 2022b](#)).

We will conclude our discussion on robust image classification with an intriguing result that connects the concept of data modification with its impact on distributional and adversarial robustness. In Chapter 5 we found that some data modification techniques like data filtering may improve distributional robustness but negatively impact adversarial robustness. This negative result opens up many potential avenues for research on the different aspects of real-world reliability and potential trade-offs that may exist fundamentally in statistical machine learning and may express in various manifestations in domains such as computer vision and natural language processing. This work, was published in ACL Findings 2022 ([Gokhale et al., 2022c](#)).

In Chapter 6 we start our foray into multi-modal vision-and-language tasks. Multi-modal tasks involving both vision and language (V&L) inputs, such as visual question answering (VQA), open up many more types of domain discrepancies that can affect model performance of test time. For the VQA task, given an image and a question

about it, models are trained to predict the answers to those questions. In VQA-LOL, we discovered that existing VQA models fail when logical transformations such as negation, conjunction, and disjunction are introduced in the questions. This surprising finding led us to develop a data augmentation tool that allows us to produce logical combinations of multiple questions in the source dataset, and a training objective that is based on Frechet inequalities to guide the predicted probabilities of answers to questions with negation, conjunction, and disjunction. Thus, given a known transformation between source and target domains, we developed a method that can leverage data augmentation for improving robustness of VQA models. This work was published as a conference paper in ECCV 2020 ([Gokhale et al., 2020c](#)).

The problem of logical and linguistic brittleness is not limited to VQA. In Chapter 7, we consider the task of vision-and-language inference (VLI), i.e. predicting whether an input sentence is **True** or **False** for a given image or video. We define a set of linguistic transformations called “SISP” that contain semantics-inverting as well as semantics-preserving text transformations, such as negation, synonyms, antonyms, paraphrasing. Analysis of VLI models using SISP transformations reveals their brittleness, especially under semantics-inverting phenomena. While data augmentation techniques have been designed to mitigate against these failure modes, methods that can integrate this knowledge into the training pipeline remain under-explored. We present **SDRO**, a model-agnostic method that utilizes a set linguistic transformations in a distributed robust optimization setting, along with an ensembling technique to leverage these transformations during inference. SDRO also allows us to learn in low-resource settings, serving as a smart data augmentation tool – SDRO models trained only with 80% of the original dataset outperform existing state-of-the-art which utilizes the entire dataset. This work was published in ACL 2022 ([Gokhale et al., 2022a](#)). Shortly after, Neeraj Varshney constructed a larger set of transformations for unsupervised learning

for the text-only task of natural language inference. This was published at ACL Findings 2022 ([Varshney et al., 2022](#)).

We will then expand into the realm of visual reasoning – reasoning beyond simple image–text alignment. In Chapter 8 we will explore how standard video captioning techniques can be enhanced by learning to speculate about commonsense aspects behind agents in the videos: this includes speculating about the intentions of agents, their typical attributes, and effects of their actions. This chapter will also describe the large effort behind creating the Video2Commonsense (V2C) dataset. This work was published at EMNLP 2020 ([Fang et al., 2020](#)) with equal contribution from Zhiyuan Fang. We recently extended reasoning about agency from the perspective of physical characteristics of objects – in our EMNLP 2022 paper ([Patel et al., 2022](#)) we introduced a new video question-answering dataset CRIPP-VQA for reasoning about the implicit physical properties of objects in a scene. CRIPP-VQA contains videos of objects in motion, annotated with questions that involve counterfactual reasoning about the effect of actions, questions about planning in order to reach a goal, and descriptive questions about visible properties of objects. Details of this work can be found in Maitreya Patel’s MS thesis ([Patel, 2022](#)).

Finally, in Chapter 9 we will study spatial reasoning abilities of vision–language models, particularly the recent wave of diffusion models that generate images given text prompts. Spatial reasoning is also a fundamental ability that VL models should have if they are to be used in robotics and human collaboration. But we found that vision–language models lack this ability – in fact they fail to distinguish between simple relationships like “left”, “right”, “above”, “below” etc. We created an automated metric and large-scale dataset for evaluating spatial reasoning abilities of text-to-image generators. This work is currently under review and can be accessed as a pre-print ([Gokhale et al., 2022d](#)).

Disclosure of Funding Sources. Work done at ASU has been funded through grants from National Science Foundation (grants #1816039, #1750082, #2132724) Defence Advanced Research Projects Agency (SAIL-ON program #W911NF2020006 and the KAIROS program), and Office of Naval Research Research grant #00014-20-1-2332. My work as an intern at Lawrence Livermore National Laboratory was performed under the auspices of the U.S. Department of Energy under contract DE-AC52-07NA27344, Lawrence Livermore National Security, LLC. and was supported by the LDRD Program under project 22-ERD-006 with IM release number LLNL-JRNL836221 and project 20-ER-014 released with LLNL tracking number LLNL-JRNL-814425. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the funding agencies and employers.

Chapter 2

ROBUSTNESS UNDER ATTRIBUTE SHIFT

The goal of *robust* machine learning models for tasks such as image classification is to make accurate predictions on *unseen* samples. The i.i.d. assumption is the simplest case in which unseen samples come from the same distribution as the training dataset. However, in most real-world situations, this assumption breaks down and so do models trained under the i.i.d. paradigm ([Recht *et al.*, 2018](#); [Bulusu *et al.*, 2020](#)).

As discussed in the previous chapter, most prior work on non-*i.i.d.* robustness has been focused on adversarial robustness, i.e. the ability to maintain good performance when the image undergoes pixel-level ℓ_p norm-bounded perturbations such as additive noise ([Goodfellow *et al.*, 2015](#); [Sinha *et al.*, 2018a](#); [Madry *et al.*, 2018a](#); [Raghunathan *et al.*, 2018](#)). While such perturbations allow the use of tractable mathematical formulations, in practice, they are not the only perturbations that might be encountered at test time. For example, geometric transforms such as rotation, translation, or scaling of images, that are commonly encountered in the real world are not accounted for by pixel-wise ℓ_p bounded perturbations.

Images are parameterized by several unique attributes ranging from low-level information responsible for image formation like lighting, camera angle and resolution; to high-level semantic information like changes in background, size, shape, or color of objects in a scene. Perturbations along many of these attributes are irrelevant to tasks like image classification and are thus “semantics-preserving” perturbations. For instance, translating a digit inside an image in a digit classification task, or manipulating the shape of an object in a color classification task, will not result in a change in the true class-label. Yet, perturbations along these attributes are

likely to cause models to fail when they are changed intentionally or otherwise (Xiao *et al.*, 2021; Joshi *et al.*, 2019; Liu *et al.*, 2019a). Shifts in such “nuisance attributes” typically result in large ℓ_p perturbations, posing significant challenges for existing pixel-level perturbation models. On the other hand, it is impractical to sample the entire attribute space effectively in order to guard against potential failures at test time. In this work, we shall build robust image classifiers that can deal with such types of attribute shift.

The approach that we will discuss in this chapter, is a robust modeling technique that we call **AGAT: Attribute Guided Adversarial Training**, which learns to generate new samples so as to maximize the exposure of the classifier to variations in the attribute space. Our approach falls under the broad category of adversarial training (Madry *et al.*, 2018b), and utilizes a min-max optimization setup, wherein the inner maximization step generates adversarial attribute perturbations while the outer minimization step identifies model parameters that reduce the task-specific loss (e.g., categorical cross entropy) under these perturbations. We find that the attribute-based specification produces models that can more effectively handle challenging real-world distribution shifts than standard ℓ_p norm-bounded perturbations (Qiao *et al.*, 2020). Furthermore, our proposed approach is flexible to support a wide-range of attribute specifications, which we demonstrate with three different use-cases:

1. Object-level shifts from a conditional GAN for adversarial training on a new variant of the CLEVR dataset;
2. Geometric transformations implemented using a spatial transformer for MNIST data; and
3. Synthetic image corruptions on CIFAR-10 data.

Our contributions can be summarized as follows:

- We consider the problem of robustness under a set of specified attributes, that go beyond typically considered ℓ_p robustness in the pixel space.
- We present Attribute-Guided Adversarial Training (AGAT), a robust modeling technique that solves a min-max optimization problem and learns to explore the attribute space and to manipulate images in novel ways without access to any test samples.
- We create a new benchmark called “CLEVR-Singles” to evaluate robustness to semantic shifts. The dataset consists of images with a single block having variable colors, shapes, sizes, materials, and position.
- We demonstrate the efficacy of our method on three classes of semantics-preserving perturbations: object-level shifts, geometric transformations, and common image corruptions.
- Our method outperforms competitive baselines on three robustness benchmarks: CLEVR-Singles, MNIST-RTS, and CIFAR10-C.

2.1 Related Work

Most existing work on robustness deals with the problem of finding ℓ_p perturbations which focus on additive noise, with tractable mathematical guarantees of performance when test data falls within an ϵ -ball of the training distribution (Goodfellow *et al.*, 2015; Sinha *et al.*, 2018a; Madry *et al.*, 2018a; Raghunathan *et al.*, 2018). Such perturbation are typically *imperceptible* to the human eye. As a result, there is an increasing interest in addressing challenges that arise from natural corruptions or perturbations (Hendrycks and Dietterich, 2019) that are *perceptible* shifts in the data, more likely to be encountered in the real world. For example, (Liu *et al.*, 2019a)

use a differentiable renderer to design adversarial perturbations sensitive to semantic concepts like lighting and geometry in a scene; ([Joshi et al., 2019](#)) design perturbations only along certain pre-specified attributes by optimizing over the range-space of a conditional generator. Our work focuses on building robust models against semantic, or more generally attribute guided concepts that may or may not exist in the training distribution, using a surrogate function.

ℓ_p -norm based robustness methods make no assumptions about the test distribution, except that the methods are guaranteed to be robust only inside the ϵ -ball of the training distribution ([Volpi et al., 2018](#); [Qiao et al., 2020](#)). Some recent approaches extend this notion to assume some access to data from the test distribution such as TTT ([Sun et al., 2020](#)) that achieves robustness for a test example by minimizing the cost of an auxiliary task for each test sample; and ([Wong and Kolter, 2020](#)) learn a CVAE using possible corruptions one might encounter, to then guarantee robustness of a classifier within the learned perturbation set. For comparison, our method assumes access to no data from the test distribution, but only knowledge of a specification, which is the intended functionality of the system specified in human-understandable attributes. Under this challenging set-up, we show our method still outperforms existing robustness techniques on popular and standard benchmarks.

2.2 Setup

We begin by defining the classifier parameterized by a set of neural network weights, θ , as $H_\theta : \mathcal{X}_s \mapsto \mathcal{Y}$, where \mathcal{X}_s denotes the space of the observed image data (or source) and \mathcal{Y} denotes the label space for the task of interest.

Robustness to natural perturbations. Our goal is to train an H_θ that is robust to *natural* perturbations, which are typically larger in magnitude than the *imperceptible*

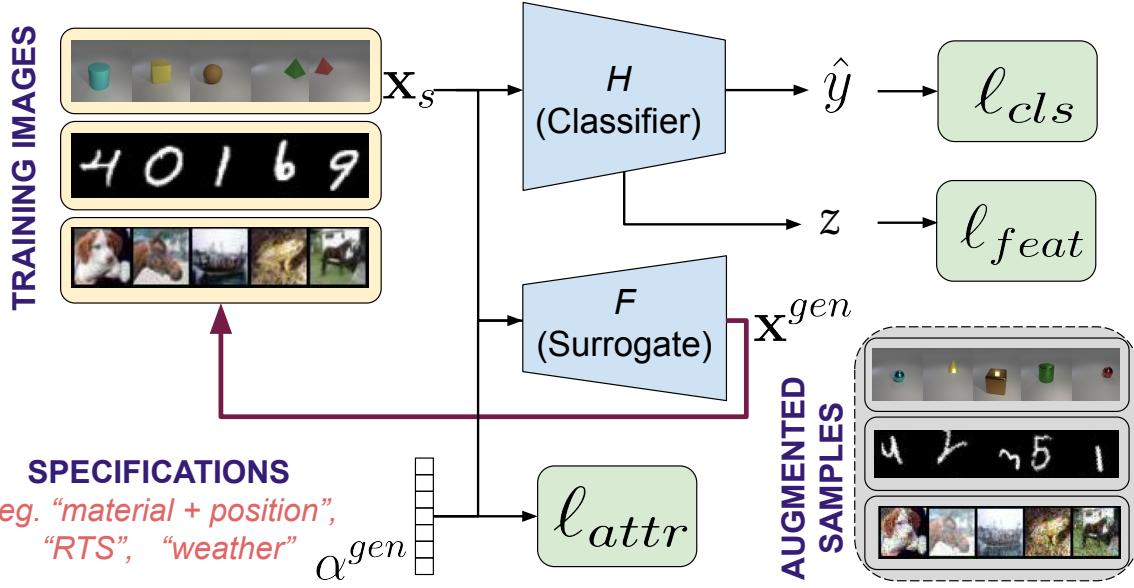


Figure 2.1: Overview of the problem setup and our attribute-guided adversarial training method.

ℓ_p -bounded pixel-space perturbations, considered in the literature. We will consider semantic shifts in attributes such as shape, size, texture, and position of objects; geometric transformations of varying intensities; and common image corruptions such as noise, blur, weather, and digital artifacts. Most of these perturbations do not naturally fall within small ℓ_p -norm ball deviations ($|\mathbf{x} - \tilde{\mathbf{x}}|_p \leq \epsilon$), for which most existing robustness methods are designed, and are bound to fail when the classifier encounters such data in the wild. However, making ϵ arbitrarily large in robustness formulations does not work in practice, since the image quality degrades significantly. Hence, we propose a new framework to design models that are robust to such natural perturbations.

2.3 Attribute Guided Adversarial Training

Let us denote an image by \mathbf{x}_α parameterized by a set of attributes α related to image formation (lighting, viewing angle, position) as well as abstract semantic information (color, shape, size, etc.). Manipulating images with new combinations of attributes that are not seen in the training set, requires access to the underlying physical generative processes, which is unrealistic. We do not assume direct access to such deterministic mechanisms.

Our goal is to train classifiers robust to natural perturbations along attributes in α that are specified *a priori*. Inspired by recent developments in robust optimization and adversarial training ([Madry et al., 2018a](#)), we consider the following worst-case problem around N attributes of the training data:

$$\min_{\theta \in \Theta} \sum_{i=1}^N \max_{|\hat{\alpha}_i - \alpha_i| \leq \epsilon} \ell(\theta; (\mathbf{x}_{\hat{\alpha}_i}, y_i)), \quad (2.1)$$

where $\ell(\cdot)$ is the cross-entropy loss.

The solution to worst-case optimization in Equation 2.1 guarantees good performance against test data that is distance ϵ away from the training data in the attribute space. In other words, we expect the model learned using Equation 2.1 to be robust against ϵ -bounded natural perturbations. Interestingly, as we will empirically show later, models learnt using Equation 2.1 perform better than existing pixel-level techniques even on ℓ_p -bounded imperceptible perturbations.

Although the structure of the attribute-guided adversarial training problem may look similar to standard adversarial training, we explain next why solving Equation 2.1 is significantly more challenging and requires us to make several algorithmic innovations. Note that Equation 2.1 solves a min-max optimization problem, with the inner maximization generating natural perturbations by maximizing the classification loss over attribute space, and the outer minimization finding model parameters by

minimizing the loss on natural perturbations of the training data generated from the inner maximization. The success of this method crucially relies on solving the inner optimization problem. Motivated by the standard adversarial training, one might be tempted to approximately solve the re-parameterized inner optimization problem

$$\max_{\|\delta\|_p \leq \epsilon} l(\theta; (\mathbf{x}_{\alpha_i+\delta}, y_i)), \quad (2.2)$$

and generate the natural perturbations $\mathbf{x}_{\hat{\alpha}_i}^*$ using projected gradient descent (PGD) as:

$$\delta_i^* := \mathcal{P}_\epsilon(\delta_i - \lambda \nabla_\delta l(\theta; (\mathbf{x}_{\alpha_i+\delta}, y_i))), \quad (2.3)$$

where λ is the gradient step and \mathcal{P}_ϵ is projection on l_p ball of radius ϵ . However, there are two fundamental issues with this approach making it infeasible in practice: first, we cannot compute gradients as we do not have access to the attribute space; and second, we do not have access to the true generative mechanism conditioned on the attributes.

2.3.1 Proposed Approach

Surrogate Functions. We propose to use differentiable surrogate functions parameterized by attributes to overcome the limitation described above. In other words, we have $\mathbf{x}_{\alpha+\delta} \approx F_\delta(\mathbf{x}_\alpha)$, where F_δ is differentiable. Typically, exact perturbations $\mathbf{x}_{\alpha+\delta} = F_\delta(\mathbf{x}_\alpha)$ can be performed for PGD attacks or other l_p norm bounded attacks. However, in our case accessing the true generative process to manipulate images along α is not feasible. For example, we cannot rely on deterministic functions to manipulate semantic features in the image like size, shape or texture of an object. As a result, we resort to using *approximate* image manipulators in the form of surrogate functions which act as proxies to the true generative process. Depending on the type of attributes against which we wish to train for robustness, the surrogate function can

take different forms:

- generative editing models for semantic perturbation that is learned from the training data itself,
- analytical functions for geometric transformations in the form of spatial transformers (STNs), or
- an analytical approximation (or tractable upper bound) of the natural perturbation space.

For example, if we want a classifier robust to unknown affine transforms then F is the spatial transformer layer parameterized by α which now represents 6 parameters controlling rotation, scale, and shift of the image.

Note that we do not assume access to any additional data other than the clean training dataset \mathcal{X}_s , and specification of the class of functions against which robustness is desired. While such surrogate functions only approximate the natural perturbations, we show that they are sufficient for enabling us to make classifiers more robust to natural perturbations.

Iterative Training Procedure. Having access to attribute parameterized surrogate function, we aim to solve Equation 2.1. Note, the success of the adversarial training is dependent on the quality of the generated perturbations. Thus, we aim to generate natural perturbations that have a larger coverage over the specified attribute space than the training samples images \mathbf{x}_s . Consider the classifier H_θ which outputs the predicted class \hat{y} and intermediate features z , let the surrogate function F be parameterized by the attribute vector α . We propose an iterative training procedure called Attribute-Guided Adversarial Training (**AGAT**) detailed in Algorithm 1. Our algorithm has two objectives: to minimize the classification loss over input images and

to maximize the divergence between the training samples and generated perturbations. Thus, the key idea here is to explore novel and hard images that only *vary along the specified attributes*. To achieve this, we impose a constraint that maximizes the distance between features of dataset images and perturbed images. Additionally, since we would also like to explore new regions in the attribute space, we impose a similar constraint on the attributes of perturbed images.

We express this constraint as the loss function given by:

$$\ell_{const} = \lambda_1 \ell_{feat} + \lambda_2 \ell_{attr}, \quad \lambda_1, \lambda_2 \in (0, 1) \quad (2.4)$$

where $\ell_{feat} = \|\mathbf{z} - \mathbf{z}^{\text{gen}}\|_2^2$, and $\ell_{attr} = \|\alpha - \alpha^{\text{gen}}\|_2^2$

To ensure that the generated images belong to the same class as the input image, we combine classification loss with respect to the ground truth label \mathbf{y} with consistency regularization with respect to the predicted label $\hat{\mathbf{y}}$ of \mathbf{x} .

$$\ell_{cls} = \ell_{BCE}(\mathbf{y}, \mathbf{y}^{\text{gen}}) + \ell_{BCE}(\hat{\mathbf{y}}, \mathbf{y}^{\text{gen}}) \quad (2.5)$$

The overall loss function is computed as the Lagrangian:

$$\ell_{AGAT} = \ell_{cls} - \beta \cdot \ell_{const} \quad (2.6)$$

Intuitively, ℓ_{cls} encourages the augmented images to belong to the same class-label as the input image, while the constraint ℓ_{const} encourages the adversarial learning algorithm to perturb the image features as well as the attributes away from the input features and attributes. We first pre-train the classifier only on the source samples \mathbf{x}_s for N_{pre} epochs. Then, we initiate our augmentation process. To generate new samples, we minimize Equation 2.6 and update the attribute vector for M update steps as:

$$\alpha^{\text{gen}} \leftarrow \alpha^{\text{gen}} - \mu \nabla \ell_{AGAT}. \quad (2.7)$$

Algorithm 1 Attribute-Guided Adversarial Training

```
1: Initialize:  $\theta \leftarrow \theta_0, \mathcal{D}_S^{aug} \leftarrow \mathcal{D}_S$ 
2: for each  $n = 1 \dots N_{epochs}$  do
3:   if  $n < N_{pre}$  then
4:     for each  $t = 1 : T$  do
5:        $\theta \leftarrow \theta - \eta \nabla \ell_{cls}(\theta; (\mathbf{x}_t, y_t))$ 
6:     end for each
7:   else
8:     if  $n \bmod N_{aug} = 0$  then
9:       for each  $t = 1 \dots T_{aug}$  do
10:        sample  $(\mathbf{x}_t, y_t)_{t=1}^{T_{aug}}$  from  $\mathcal{D}_S$ 
11:         $z_t, \hat{y}_t = H(\mathbf{x}_t)$ 
12:        Initialize:  $\alpha_t^{gen}$ 
13:        for each  $i = 1 \dots M$  do
14:           $z_t^{gen}, \hat{y}_t^{gen} = H(\mathbf{x}_t, \alpha_t^{gen})$ 
15:           $\mathbf{x}_t^{gen} \leftarrow f(\mathbf{x}_t, \alpha_t^{gen})$ 
16:           $\alpha_t^{gen} \leftarrow \alpha_t^{gen} - \mu \nabla (\cdot \ell_{cls} - \beta \cdot \ell_{cons})$ 
17:        end for each
18:         $\mathcal{D}_S^{aug} \leftarrow \mathcal{D}_S^{aug} \cup \mathbf{x}_t^{gen}$ 
19:      end for each
20:    else
21:      for each  $(\mathbf{x}_t, y_t) \in \mathcal{D}_S^{aug}$  do
22:         $\theta \leftarrow \theta - \eta \nabla \ell(\theta; (\mathbf{x}_t, y_t))$ 
23:      end for each
24:    end if
25:  end if
26: end for each
```

Finally, synthetic images are generated using the surrogate function

$$\mathbf{x}^{gen} \leftarrow F(\mathbf{x}, \alpha^{gen}), \quad (2.8)$$

These generated images are then appended to the training data. This adversarial data augmentation is performed after every N_{aug} epochs during which T_{aug} images are generated. The total number of augmented samples is expressed as a percentage of the number of training samples so as to allow fair comparison across datasets and types of perturbations. The pseudocode for AGAT is shown in Algorithm 1.

The distinguishing factor for **AGAT** is that we perturb the attribute space and use surrogate functions to synthesize images, while previous adversarial augmentation protocols such as M-ADA ([Qiao et al., 2020](#)) and GUD ([Volpi et al., 2018](#)) perturb only in the pixel-space, thus being restricted to ℓ_p perturbations. It is important to note that our method is agnostic to the choice of surrogate functions, which can take the form of additive noise, affine transformation in pixel-space, or conditional generative adversarial networks ([Mirza and Osindero, 2014](#)) trained to transform an input image according to an input attribute vector.

2.4 CLEVR-Singles Dataset Creation

The CLEVR-Singles dataset has been created to allow a diagnostic setup with discrete and controllable set of attributes for our robustness experiments for semantic object-level perturbations. In this section we delineate the dataset creation process and provide more examples from the dataset.

We use Blender ([Blender Online Community, 2018](#)) to render images in CLEVR-Singles. Every scene contains a single object, have a randomly chosen size, shape, material, position, and color. Our code for rendering these images is a modification of the data creation process for the CLEVR dataset ([Johnson et al., 2017](#)). The choices

Property	Choices
Color	gray, red, blue, green, brown, purple, cyan, yellow
Size	large (1.2), medium (0.9), small (0.6)
Shape	Cube, Sphere, Cylinder, Pyramid
Material	Rubber, Metal
Position	NW, NE, SW, SE
Rotation	$\theta \in [-180, 180]$

Table 2.1: List of properties that can be assigned to an object to render the images in CLEVR-Singles.

for each property of the object are shown in Table 2.1.

Size refers to the height or diameter of the shapes which are symmetric along x and y axes. In terms of 3D coordinates, the objects are placed at the position (X, Y, Z) such that:

$$X \in [-3, 3], Y \in [-3, 3], Z = 0.5 * size.$$

For our experiments, we define position to be one of the four quadrants of the xy plane and denote these by NW, NE, SW, SE, which allows us to easily split the dataset for robustness experiments.

The positions of the lighting and camera and are slightly jittered to introduce variation. The objects are rotated at a random angle while placing. We generate images of size 256×256 using a tile size of 128 for rendering. The CLEVR-Singles dataset contains 50000 images for training, 10000 for validation, and 10000 test samples. Each image takes approximately 5 seconds to render using a single GPU, and can also be rendered on CPU at 8 seconds per image. Along with the images, scene graphs are also generated which contain labels for each attribute, color, and other properties.

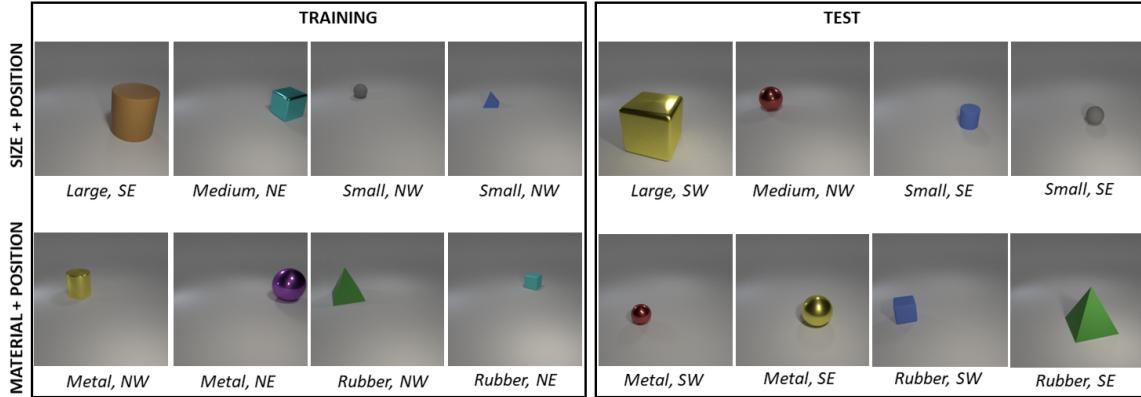


Figure 2.2: Sample images from the training and test splits for robustness experiments on the CLEVR-Singles dataset. The first row shows the train-test split on the attributes *size+position*, and the second row for *material+position*.

Illustrative examples are shown in Figure 2.2 with the first row showing samples from training and test splits for *size+position*, and the second row for *material+position*. The data-generation code and train-test splits for robustness experiments is available at <https://github.com/tejas-gokhale/CLEVR-Singles>.

2.5 Experiments

In this section, we introduce the three types of robustness specifications that we experiment with, along with details about the datasets, baselines, and metrics used for each.

Model Architectures We use the same classifier for baselines as well as our model for fair comparison. For experiments on CLEVR-Singles, we use a color-classification neural network with 4 convolutional layers with stride 2 and kernel-size 3, followed by 2 fully-connected layers. For experiments on MNIST, we use a classifier with two convolutional layers with kernel-size 5 and stride 1 and 3 fully-connected layers. For CIFAR-10-C experiments we use the classifier architecture used by (Sun *et al.*, 2020)

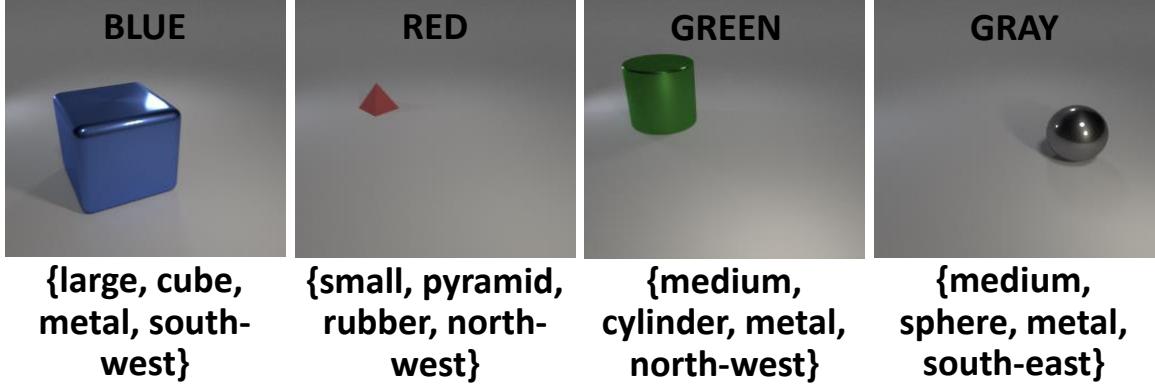


Figure 2.3: Examples of images from CLEVR-Singles and the color labels and (size, shape, material, position) attributes.

Attribute	Train	Test
Size and Position	(small, NW), (medium, NE), (large, SE)	(small, SE), (medium, NW), (large, SW)
Material and Position	(metal, NW), (rubber, NW), (metal, NE), (rubber, NE)	(rubber, SW), (metal, SW), (rubber, SE), (metal, SE)

Table 2.2: The train and test splits for our experiments with semantic object-level perturbations for CLEVR-Singles.

for fair comparison. The classifier is a ResNet (He *et al.*, 2016) constructed specially for CIFAR-10 images and has 26 layers, group norm (Wu and He, 2018) with 8 layers.

2.5.1 Semantic Object-Level Perturbations

One class of real-world perturbations is when properties or attributes of images or objects in images change at test time. These changes do not affect the classification label, but significantly change the appearance of the image. For instance, consider the task of color classification in objects of varied shapes and textures. Here, *red metallic spheres* and *red rubber cubes* both belong to the class label “red”, however may appear

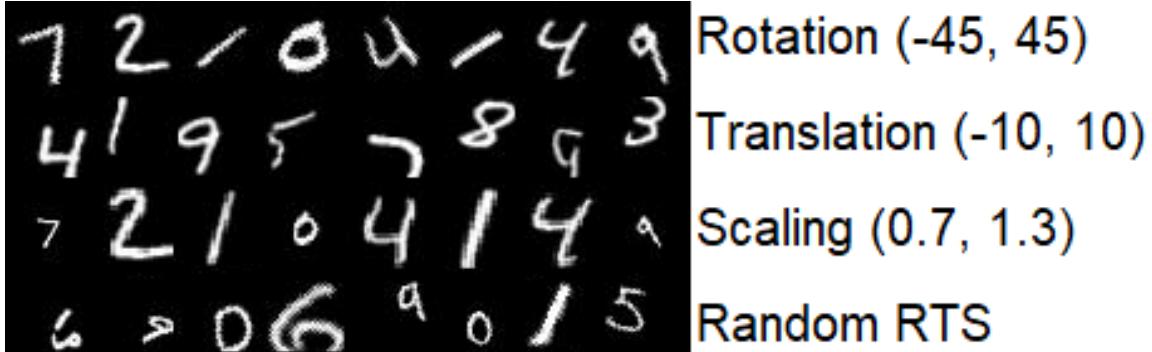


Figure 2.4: RTS-perturbed MNIST images.

very different in their shapes and textures. Thus, if only *red metallic cubes* are seen during training, conventional classifier predictions for test images consisting of *red rubber cubes* can fail to generalize. Thus, although the class label is invariant to such semantic factors, robustness to perturbations along these factors is desirable.

Dataset: To study the problem of such object-level shifts along semantic factors of an image in a controlled fashion, we create a new benchmark called CLEVR-Singles¹ by modifying the data generation process from CLEVR (Johnson *et al.*, 2017). We create images of single objects having one of eight colors, and use color classification as our task in this paper. Each object has four variable attributes that do not affect the color class of the image; these are: *shape* (cube, sphere, pyramid, or cylinder), *size* (small, medium, or large), *material* (rubber or metal), and *position* (northwest, southwest, northeast, southeast). While the objects are generated at continuous (X, Y, Z) world coordinates, we assign them a discrete position class for our experiments. Object-level perturbations can be made over these four attributes for our robustness experiments. In other words, it is known that one or more of $\{\textit{shape}, \textit{size}, \textit{material}, \textit{position}\}$ of the image may change at test-time without knowing the magnitude or combinations

¹Dataset: <https://github.com/tejas-gokhale/CLEVR-Singles>

INPUT	ATTRIBUTE		Size		Shape			Material	
	S	M	L	Sph.	Cylin.	Pyra.	Cube	Rubber	Metal

Figure 2.5: Images generated by AttGAN for the images in column 1, conditioned on attributes.

of the change. We split the dataset based on a combination of attributes as shown in Table 2.2; for instance only certain combinations of size and position are observed in the training set, but robustness is expected from the color classifier on unknown combinations.

AttGAN as the Surrogate Function: Conditional generative adversarial networks (cGANs) have been shown to perform exceptionally well on image-to-image translation in various domains ([Isola et al., 2017](#); [Zhang et al., 2017](#); [Karras et al., 2019](#)). AttGAN ([He et al., 2019](#)) is one such conditional GAN which is trained to manipulate attributes of input face images. Thus given an image and a vector of desired attributes, AttGAN can manipulate the face image along the desired attribute dimensions. We leverage this powerful image manipulation technique as our surrogate function $\mathbf{x}^{gen} = F_{GAN}(\mathbf{x}, \alpha)$. Formally, we define the attribute vector to be a 13-dimensional binary hash-code with 1 and 0 indicating presence or absence of an

attribute. For each experiment, we train the AttGAN on the training dataset outlined in Table 2.2 to generate 128x128 images, with a learning rate of 2e-4 for 100 epochs on a single 16GB GPU. Examples of images generated by AttGAN are shown in Figure 2.5, when manipulating certain attributes such as size, shape, and material of the object.

Baselines: For the color classification task on CLEVR-Singles images, we compare against two pixel-level domain augmentation baselines: GUD ([Volpi et al., 2018](#)) which performs adversarial data augmentation to generate fictitious target domains, and M-ADA ([Qiao et al., 2020](#)) which uses a meta-learning framework to generate multiple domains of samples. We also report the performance of a classifier directly without any adversarial training as a naive baseline. The same classifier architecture is used for each baseline for fair comparison. All models are trained for 15 epochs including pre-training epochs $N_{pre} = 5$, batch-size 64, and $M = 15$ update steps for adversarial augmentation. The number of augmented samples T_{aug} is 30% of the original source data, and augmentation interval N_{aug} is fixed at 2 epochs. For our model the coefficients in Equations 2.4, and 2.5 are: $\lambda_1 = 0.5$, $\lambda_2 = 0.5$, $\beta = 0.25$. The learning rates η, μ for the classifier and adversarial augmentation are both 5e-5.

Results: The test classification accuracies for different splits are reported in Table 2.3. We observe that our model is better than all baselines considered here, with a boost of 5 percentage points in accuracy on the harder experiment along *Material+Position*.

Analysis: In Figure 2.6, we show 8 different examples generated by AttGAN during adversarial training. We can see the effect of the coefficient β , from the constraint loss ℓ_{const} in Equation 2.6, in exploring the attribute space. An appropriately chosen value for β encourages useful perturbations without violating the class-label consistency cost

Method	Source	Size+Pos.	Mat.+Pos.
B	99.81	89.92	59.90
GUD (Volpi et al., 2018)	99.94	93.69	65.03
M-ADA (Qiao et al., 2020)	99.96	94.52	65.50
Ours	99.97	95.22	69.49

Table 2.3: Classification accuracy for color-classification on CLEVR-Singles. Source and target sets are split on *size+position* attribute for the third column, and *Material+Position* for the fourth column.

ℓ_{cls} as seen in the top row of Figure 2.6. On the other hand, a higher β would mean a higher weight for exploring the regions (or combinations) in attribute space not seen in training. In the bottom row we see that a high β encourages novel attribute exploration at the cost of higher classification error as a result of generating objects with different colors within the same image. It is noteworthy that AttGAN is able to generate images with multiple objects, even when it trained on images with only a single object, thus demonstrating its suitability to explore novel attributes using the proposed AGAT training.

2.5.2 Geometric Transformations

Another common class of perturbations is geometric transformations, i.e. a composition of rotation, translation, and scaling of an image. These perturbations are common since cameras may capture a scene from different orientations, distances, and inclinations. It is well known that standard image classifiers are not robust to these common perturbations ([Cohen and Welling, 2015](#)).

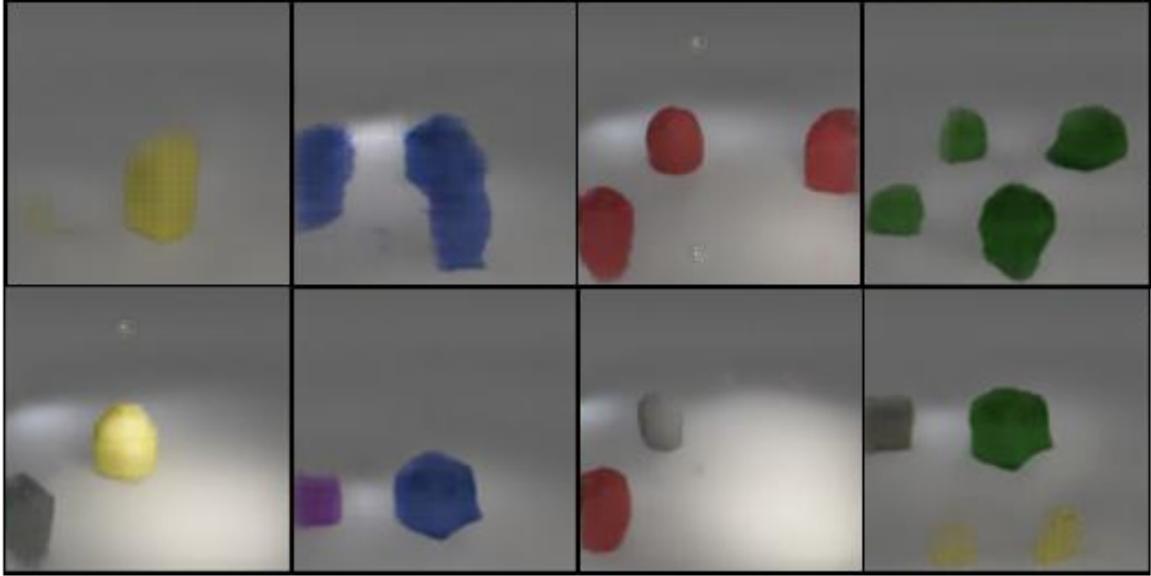


Figure 2.6: Visualization of the effect of weight β of the constraint loss ℓ_{const} on the generated images. Row 2 has higher β than Row 1. Illustration also shows that AttGAN is able to generate multiple objects (of same color for Row 1 and different colors for Row 2), though absent in training data.

Dataset: We address this problem in the digit classification setting, with the training images from MNIST ([LeCun et al., 1998](#)), and the test images that are perturbed along rotation-translation-scale (RTS), as shown in Figure 2.4. We use the standard RTS setup ([Jaderberg et al., 2015](#)) with angle of rotation in $(-45, 45)^\circ$, translation in $(-10, 10)$ pixels in both directions, and a scale factor in the range $(0.7, 1.3)$.

Surrogate Function: The attributes of interest, α , consist of a 2×3 affine matrix that controls rotation, translation, and scale. To perform affine transformations on the image with a perturbed α , we use Spatial Transformer Networks (STN) ([Jaderberg et al., 2015](#)) which allow differentiable spatial manipulation of input images in a convolutional neural network, such as RTS and or general warping. The perturbed images are generated as: $\mathbf{x}^{gen} = F_{STN}(\mathbf{x}_s, \alpha)$.

Method	R	T	S	RTS
B	84.44	27.67	95.76	21.91
GUD (Volpi <i>et al.</i> , 2018)	86.08	29.09	97.89	23.10
MADA (Qiao <i>et al.</i> , 2020)	87.37	29.25	98.32	22.68
PS (Wong and Kolter, 2020)	87.86	45.36	96.00	39.38
Ours ($T_{aug}=30\%$)	<u>84.93</u>	52.95	<u>96.11</u>	<u>41.43</u>

Table 2.4: Results on the MNIST-RTS robustness benchmark for rotation (R), translation (T), scaling (S), and random combination (RTS).

Baselines: We compare the robustness performance to RTS perturbations with a naive baseline, denoted by (B), that is only trained on the standard MNIST dataset, and pixel-level perturbation methods MADA (Qiao *et al.*, 2020) and GUD (Volpi *et al.*, 2018). Additionally, we also use the RTS perturbation sets generated by (Wong and Kolter, 2020) (PS) and use them as augmented training samples. All models are trained for 12 epochs including pre-training epochs $N_{pre} = 5$, with a batch-size 64, and $M = 10$ update steps for adversarial augmentation. The number of augmented samples T_{aug} is 30% of the original source data, and augmentation interval N_{aug} is fixed at 10 epochs. Our model the coefficients in Equations 2.4, and 2.5 are: $\lambda_1 = 1$, $\lambda_2 = 1$, $\beta = 5$. The learning rate η for the classifier is 1e–4 and μ for the adversarial augmentation is 0.1.

Results: We report digit classification accuracies on the target test set containing only rotations (R), only translations (T), only skew (S), as well as a random combination of RTS. Our model performs well on all four metrics, and beats the perturbation sets (PS) even though their augmentation model has access to RTS perturbations during training. In particular, we observe a significant improvement compared with

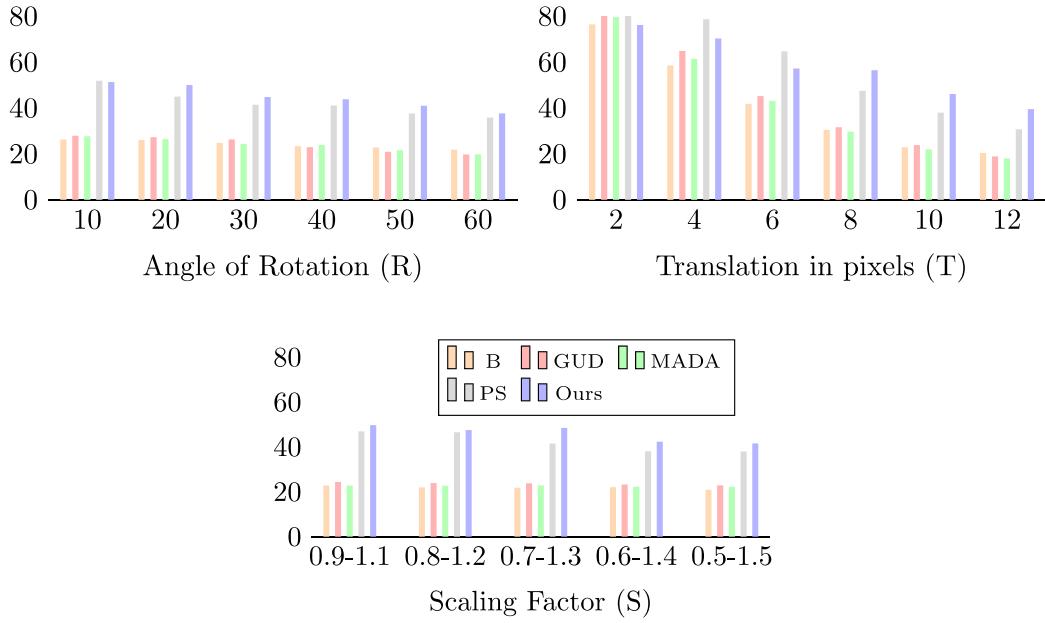


Figure 2.7: Comparison of random RTS accuracies when controlling each parameter to a max. value. Left: R, Center: T, Right: S

MADA and GUD, in the robustness on the translation experiment, which is the hardest task among the three.

Analysis: The pixel-level perturbation methods still perform reasonably well on rotation and scale experiments in Table 2.4 because in each case the rotations/translations/scale are randomly sampled, resulting in several test examples that are very close to the training examples (with no RTS). In order to resolve this further, we study the performance by controlling the magnitude of R, T, and S in the test set. Figure 2.4 shows the bar-plots when the range of rotation is varied from $(-10, 10)$ to $(-60, 60)$, translation from $(-2, 2)$ to $(-12, 12)$ pixels, and scaling factor from $(0.9, 1.1)$ to $(0.5, 1.5)$. It can be observed that at higher severity of perturbation, our model (in blue) significantly outperforms all baselines. The model trained with Perturbations Sets ([Wong and](#)

N_{aug}	T_{aug}	R	T	S	RTS
1	10	84.12	43.33	96.65	34.12
	30	83.80	54.17	95.89	40.46
	50	84.49	59.97	96.29	47.62
	70	84.35	62.76	96.24	51.13
2	10	84.97	47.21	96.41	36.84
	30	84.93	52.95	96.11	41.43
	50	86.35	61.07	95.76	47.59
	70	84.59	62.75	95.79	50.28

Table 2.5: The effect of augmentation interval (N_{aug}) at different percentages of augmented samples (T_{aug}).

Kolter, 2020) (in gray) is competitive at lower severities.

We also analyze the effect of the number of augmented samples (T_{aug}) expressed as a percentage of the size of the training data, while controlling for the augmentation interval N_{aug} . As expected, larger number of augmented samples improve robustness even higher than in table 2.4 (which fixes number of additional augmented examples at 30% for all baselines). Larger augmentation intervals contribute positively at lower percentages of augmented samples.

Finally, we perform an ablation study with and without the consistency regularization defined in Equation 2.5 and show that the regularization indeed helps improve performance.

Loss	T_{aug}	R	T	S	RTS
GT	10	85.41	29.48	96.74	23.57
	30	84.82	48.46	96.75	37.80
	50	84.17	52.44	95.86	41.82
	70	84.07	55.14	95.70	44.20
GT + CR	10	84.97	47.21	96.41	36.84
	30	84.93	52.95	96.11	41.43
	50	86.35	61.07	95.76	47.59
	70	84.59	62.75	95.79	50.28

Table 2.6: The effect of classification loss function at different percentages of augmented samples. GT denotes the first term and CR is consistency regularization in Eq (2.5).

2.5.3 Common Image Corruptions

Image corruptions are another common class of perturbations. These can occur due to image digitization artifacts, weather, camera calibration, and other sources of noise.

Dataset: The CIFAR10 dataset ([Krizhevsky, 2009](#)) contains $50k$ training images belonging to 10 classes. Recently, CIFAR10-C ([Hendrycks and Dietterich, 2019](#)) which contains image corruptions for CIFAR10 images, was proposed to benchmark robustness of image classifiers, with 4 major and 15 fine-grained categories of corruption: *Weather* (fog, snow, frost), *Blur* (zoom, defocus, glass, motion), *Noise* (shot, impulse, Gaussian), and *Digital* (JPEG, pixelation, elastic transform, brightness, contrast). There are five levels of severity of corruptions; we focus on the highest severity.

Surrogate Function: We use a general surrogate function – a composition of additive Gaussian noise and Gaussian blur filter parameterized by $\alpha = \{\alpha_1, \alpha_2\}$:

$$\mathbf{x}^{gen} = \frac{1}{\sqrt{2\pi\alpha_1^2}} e^{-\frac{\mathbf{x}^2}{2\alpha_1^2}} + n, \text{ where } n \sim \mathcal{N}(0, \alpha_2). \quad (2.9)$$

We evaluate the performance gains using this surrogate function with the proposed AGAT training on the challenging CIFAR-10-C dataset.

Baselines: Test-Time Training (TTT) ([Sun et al., 2020](#)) is a recent approach in which a classifier is trained only on source data, but the test sample is utilized to update the classifier during inference. Adversarial Logit Pairing (ALP) ([Kannan et al., 2018](#)), a technique for defending against adversarial attacks, and pixel-wise domain augmentation techniques MADA ([Qiao et al., 2020](#)) and GUD ([Volpi et al., 2018](#)) are also considered as baselines. We use ResNet-26 ([He et al., 2016](#)) specially designed for CIFAR-10 ([Russakovsky et al., 2015](#)), with group normalization ([Wu and He, 2018](#)) which is stable with different batch sizes. This acts as the naive classifier-only baseline (B). We also consider the classifier trained with an auxiliary self-supervised task of angle prediction ([Gidaris et al., 2018](#)) (B+SS). Our joint-training (JT) baseline is from TTT based on ([Hendrycks and Dietterich, 2019](#)).

We compare three versions of our model: with additive noise only, with Gaussian filtering, and with a composition of Gaussian filter and noise. Our models are trained for 150 epochs including pre-training epochs $N_{pre}=100$, batch-size 128, and $M=15$ update steps for adversarial augmentation. The number of augmented samples is 30% of the original source data, and augmentation interval N_{aug} is fixed at 2 epochs. For our model the coefficients in Equations 2.4, and 2.5 are: $\lambda_1 = 0.5, \lambda_2 = 0.5, \beta = 0.25$. The learning rates η, μ for the classifier and adversarial augmentation are both 5e-5.

Method	Src.	W	B	N	D	Avg.
B	90.6	70.6	69.0	45.5	71.6	66.4
B+SS	91.1	70.6	68.5	48.7	69.7	67.0
GUD	-	71.7	59.2	30.5	64.7	58.3
MADA	-	75.6	63.8	54.2	65.1	65.6
JT	91.9	71.7	69.0	50.6	71.6	68.3
ALP	83.5	60.9	74.7	75.4	68.5	70.0
TTT	92.1	73.7	71.3	54.2	73.4	70.5
Ours (b. only)	<u>91.3</u>	78.4	75.1	49.3	75.4	70.0
Ours (n. only)	90.3	75.0	73.3	62.4	73.1	71.3
Ours	89.3	77.8	74.1	<u>65.8</u>	71.6	72.3

Table 2.7: Comparison of classification accuracies on CIFAR-10-C corruption categories (Weather, Blur, Noise, Digital). Our best scores are underlined; overall best are bold.

Results: In Table 2.7 we show the classification accuracies on CIFAR10-C. It can be seen that our method consistently outperforms all baselines overall, and also on three of the four categories of corruptions (weather, blur, and digital). It is interesting to note that the ALP performance on the Noise category is distinctly greater than all previous methods, potentially because it is designed to defend against projected gradient descent adversarial attacks (Madry *et al.*, 2018b). ALP uses a similar loss function as Equation 2.5 to train the classifier, but still operates in pixel-space and does not perturb the attribute space. In Table 2.7 we also demonstrate that our models which uses only blur or only noise as surrogate are also better than previous state-of-the art. Note that the “noise only” model is in essence a pixel-level perturbation achieved by only perturbing along the variance parameter using AGAT training, and yet we

see a significant boost in performance over all other pixel-level additive noise methods. Similarly, the “blur only” model also gives performance boosts on weather and digital categories, further indicating the general applicability of our AGAT training approach.

2.6 AGAT-Generated Images

We provide additional samples generated by attribute-guided adversarial training algorithm (AGAT) for each of our experiments.

Semantic Object-Level Perturbations: Figure 2.8 shows the images generated by AttGAN ([He et al., 2019](#)) during AGAT. It can be seen that AttGAN is able to explore various sizes, materials during the augmentation, as well generate novel scenes with multiple objects which are not present in the training dataset.

Geometric Transformations: Figure 2.9 shows the images generated by STN ([Jaderberg et al., 2015](#)) during AGAT training. It can be seen that rotations, translations, skew, and combinations of these are generated.

Common Image Corruptions: Figure 2.10 shows the images during AGAT training on the CIFAR-10 dataset using additive Gaussian noise and blur as the surrogate function.

2.7 Discussion

In this paper, we propose a new adversarial training strategy for robustness against large perturbations that are common in practical settings. Our adversarial training algorithm perturbs the attribute space to synthesize new images instead of pixel-level perturbations which are common to the robustness literature. The new CLEVR-Singles dataset that we have created can be used in future work for studying robustness to

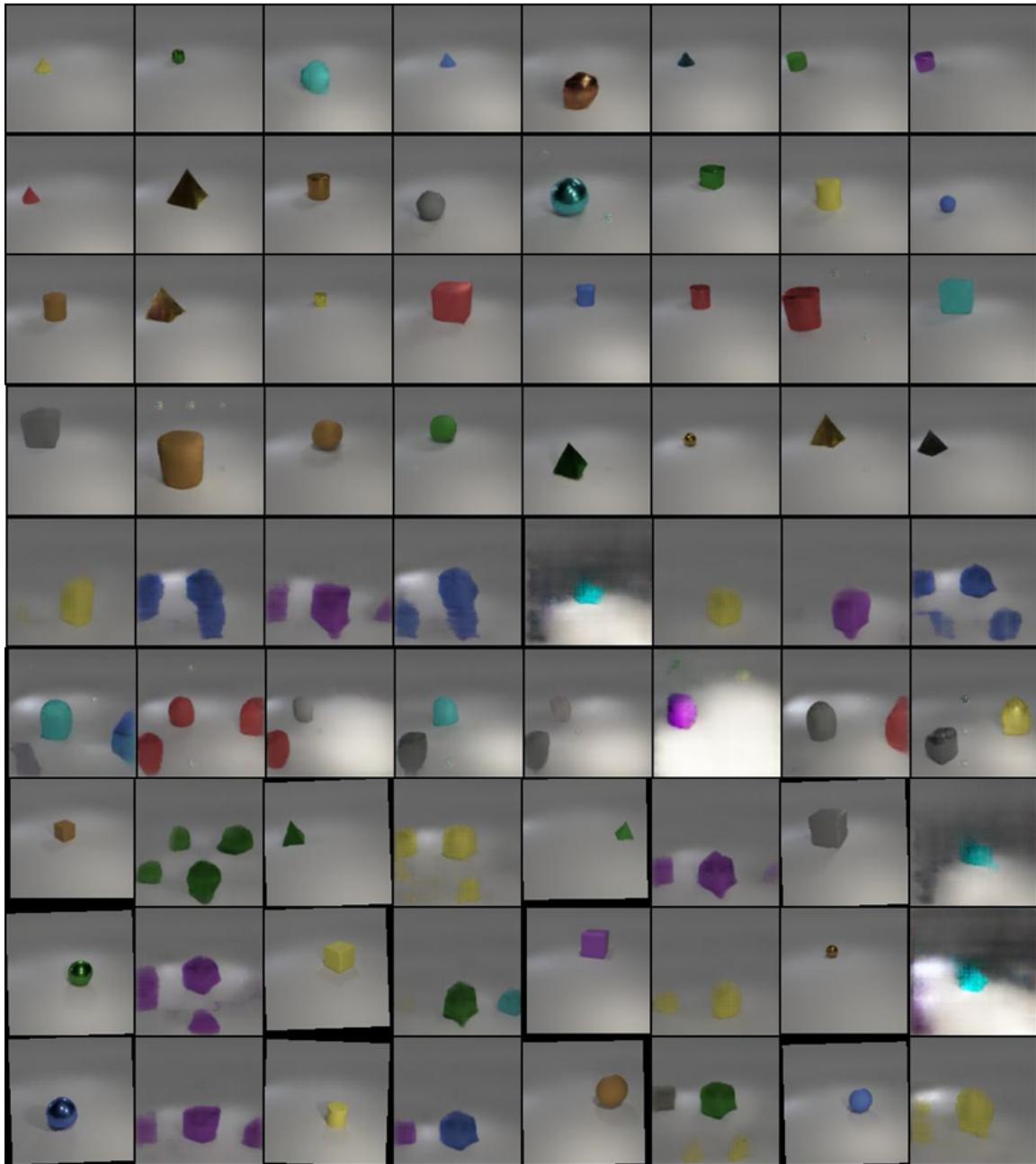


Figure 2.8: Examples of images generated by AGAT for the CLEVR-Singles dataset. The AttGAN generator is able to explore the attribute space and generate images with different attributes, as well as novel scenes such as those containing multiple objects.

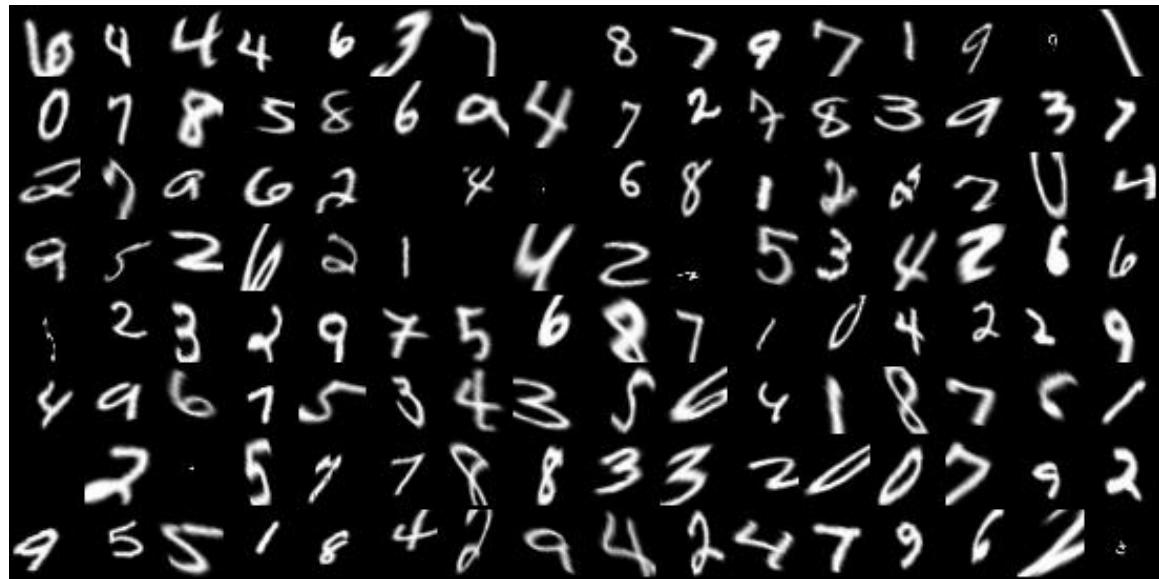


Figure 2.9: Examples of images generated by AGAT for MNIST. AGAT is able to explore the attribute space and generate images with different rotation, translations, and skews.



Figure 2.10: Examples of images generated by AGAT for the CIFAR-10 dataset. AGAT is able to explore the attribute space and generate images having varying degrees of noise and blur, which helps improve robustness on all 15 categories of corruptions as seen in Table 2.7.

semantic shifts. We extensively evaluate AGAT training on three benchmarks and achieve state-of-the-art performance. We empirically show that AGAT is applicable to a three types of naturally occurring perturbations, and can be used with different classes of surrogate functions. AGAT can potentially be applied to a broad range of robustness problems not limited to classification.

The concept of robustness is critical when it comes to deploying machine learning systems in practical settings where input signals may undergo perturbations due to weather (such as fog, smog, rain), digital corruptions in transmission, or changes in camera inclinations causing geometric transformations or artifacts such as defocusing, or motion blur. Our method for developing robust classifiers is broadly applicable if such classes of perturbations are known *a priori*. Robustness research is also crucial for avoiding or removing unintended biases that may percolate from the training data into the classification model. Recent studies ([Bolukbasi et al., 2016](#); [Zhao et al., 2017](#); [Hendricks et al., 2018](#)) have shown that models trained on biased data can in fact amplify this bias when performing inference on test samples. We believe that work in the lines of AGAT could be potentially used for mitigating social biases due to biased training data, such as gender or racial biases.

Chapter 3

ADVERSARIALLY LEARNED TRANSFORMATIONS FOR SINGLE-SOURCE DOMAIN GENERALIZATION

In Chapter 2, we looked at a relaxed setting of the generalization problem – in that setting, information about the target domain was available in terms of a set of attributes that are known to differ at test time. In this chapter we will go beyond attributes and address the harder and broader problem of domain generalization. This chapter is based upon [Gokhale *et al.* \(2023\)](#).

3.1 Introduction

Domain generalization is the problem of making accurate predictions on previously unseen domains, especially when these domains are very different from the data distribution on which the model was trained. This is a challenging problem that has seen steady progress over the last few years ([Carlucci *et al.*, 2019](#); [Volpi *et al.*, 2018](#); [Qiao *et al.*, 2020](#); [Xu *et al.*, 2020b](#); [Nam *et al.*, 2021](#)). This paper focuses on the special case – single source domain generalization (SSDG) – where the model has access only to a single training domain, and is expected to generalize to multiple different testing domains. This is especially hard because of the limited information available to train the model with just a single source.

When multiple source domains are available (MSDG), recent analysis ([Gulrajani and Lopez-Paz, 2021](#)) shows that even simple methods like minimizing empirical risk jointly on all domains, performs better than most existing sophisticated formulations. A corollary to this finding is that success in DG is dependent on *diversity* – i.e., exposing the model to as many potential training domains as possible. As the SSDG

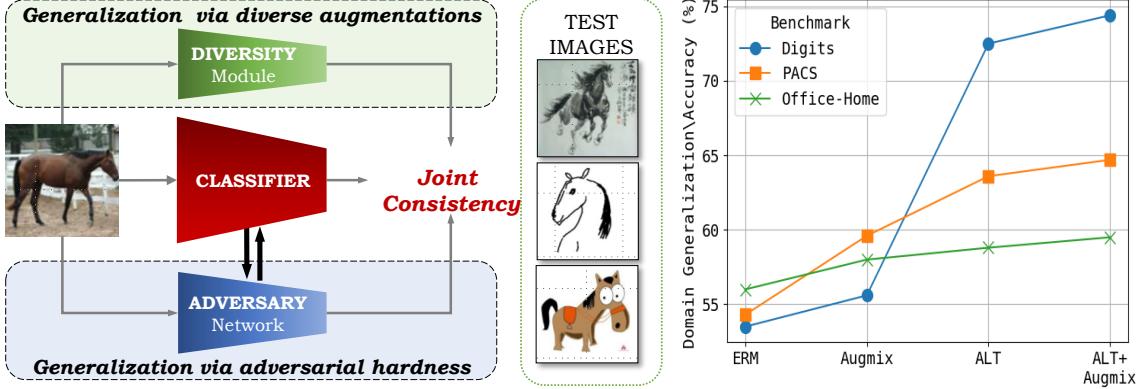


Figure 3.1: ALT consists of a *diversity* module (data augmentation functions such as Augmix (Hendrycks *et al.*, 2020c) or RandConv (Xu *et al.*, 2020b) and an *adversary* network (to learn image transformations that fool the classifier). We show an example from the PACS benchmark under the single-source domain generalization setting, with real photos (P) as the source domain and art paintings (A), cartoons (C), and sketches (S) as the target domains. The plot summarizes our results – while diversity alone improves performance over the naive ERM baseline, adapting this diversity using adversarially learned transformations (ALT) provides a significant boost for domain generalization on multiple benchmarks.

problem allows access only to a single training domain, such an exposure must come in the form of diverse transformations of the source domain that can simulate the presence of multiple domains, ultimately leading to low generalization error.

The idea of using diversity to train models has been sufficiently explored – (Hendrycks *et al.*, 2020c; Yun *et al.*, 2019; Zhang *et al.*, 2018; Cubuk *et al.*, 2020) show that a diverse set of augmentations during training improves a model’s robustness under distribution shifts. Specific augmentations can be used if the type of diversity encountered at test time is known; for example, if it is known that the test set contains random combinations of rotation, translation, and scaling, using augmentations correlated with this domain shift would lead to good performance (Benton *et al.*, 2020; Wong and Kolter, 2020; Gokhale *et al.*, 2021). However, since we cannot assume

knowledge of the test domain under the SSDG problem statement, the extent to which the model needs to be exposed to specific augmentations remains unclear. Augmentation methods impose a strong prior in terms of the types of diversity that the model is exposed to, which may not match with desirable test-time transformations. As we will show in this paper, data augmentation methods that produce good results on one dataset, do not necessarily work on other datasets – in some cases, they may even hurt performance!

In addition to such a knowledge gap, unfortunately, such augmentation methods can only achieve invariance under small distribution shifts like unknown corruptions, noise, or adversarial perturbations, but do not work effectively when the distribution shift is large and of a semantic nature, as in the case of domain generalization. On the other hand, some recent methods have directly used randomized convolutions to synthesize diverse image manipulations ([Xu et al., 2020b](#)), motivated by the large space of potentially realizable functions induced by a convolutional layer, which cannot be easily emulated using simple analytical functions.

In this paper we hypothesize that, while diversity is necessary for single-source domain generalization, diversity alone is insufficient – blindly exposing a model to a wide range of transformations may not guarantee greater generalization. Instead, we argue that carefully designed forms of diversity are needed – specifically those that can expose the model to unique and task-dependent transformations with large semantic changes that are otherwise unrealizable with plug-and-play augmentations as before. To this end, we introduce an adversary network whose objective is to *find* plausible image transformations that maximize classification error. This adversary network enables access to a much richer family of image transformations as compared to prior work on data augmentation. By randomly initializing the adversary network in each iteration, we ensure the adversarial transformations are unique and diverse themselves.

We enforce a consistency between a **diversity module** and the **adversary network** during training along with the classifier’s predictions, so that together they expose the model to learn from both diverse and challenging domains.

Our method, dubbed ALT (adversarially learned transformations), offers an interplay between diversity and adversity. Over time, a synergistic partnership between the diversity and adversary networks emerges, exposing the model to increasingly unique, challenging and semantically diverse examples that are ideally suited for single source domain generalization. The adversary network benefits from the classifier being exposed to the diversity module, and as such avoid trivial adversarial samples with appropriate checks. This allows the adversarial maximization to explore a wider space of adversarial transformations that cannot be covered by prior work on pixel-level additive perturbations.

We demonstrate this advantage of our method empirically on multiple benchmarks – PACS ([Li et al., 2017](#)), Office-Home ([Venkateswara et al., 2017](#)), and Digits ([Volpi et al., 2018](#)). On each benchmark, we outperform the state-of-the-art single source domain generalization methods by a significant margin. Moreover, since our framework disentangles diversity and adversarial modules, we can combine it with various diversity enforcing techniques – we identify two such state-of-the-art methods with AugMix ([Hendrycks et al., 2020c](#)), and RandConv ([Xu et al., 2020b](#)), and show that placing them inside our framework leads to significantly improved generalization performance over their vanilla counterparts. We illustrate this idea in Figure 3.1 where we show an image of a horse from the ‘photo’ training distribution in PACS and the different styles of cartoon/sketch/art painting horses that may be encountered at test time.

Contributions: We summarize our contributions below.

- We introduce a method, dubbed ALT, which produces adversarially learned image

transformations that expose a classifier to a large space of image transformations for superior domain generalization performance. ALT performs adversarial training in the parameter space of an adversary network as opposed to pixel-level adversarial training.

- We show how ALT integrates diversity-inducing data augmentation and hardness-inducing adversarial training in a synergistic pipeline, leading to diverse transformations that cannot be realized by blind augmentation strategies or adversarial training methods on their own.
- We validate our methods empirically on three benchmarks (PACS, Office-Home, and Digits) demonstrating state-of-the-art performance and provide analysis of our approach.

3.2 Related Work

Multi-Source Domain Generalization. Domain generalization has been explored under both multi-source (MSDG) and single-source (SSDG) settings. For the MSDG task, multiple source domains are available for training and performance is evaluated on other unseen target domains. Techniques designed for MSDG seek to utilize these multiple domains to perform feature fusion ([Shen et al., 2019](#)), learning domain-invariant features ([Ganin et al., 2016](#)), meta-learning ([Li et al., 2018a](#)), invariant risk minimization ([Arjovsky et al., 2019](#)), learning mappings between multiple training domains ([Robey et al., 2021](#)), style randomization ([Nam et al., 2021](#)), and learning a conditional generator to synthesize novel domains using cycle-consistency ([Zhou et al., 2020a](#)) [Gulrajani et al.](#) ([Gulrajani and Lopez-Paz, 2021](#)) provide an extensive comparative study of these approaches and report that simply performing ERM on the combination of source domains leads to the best performance. Many

benchmarks have been proposed to evaluate MSDG performance such as PACS (Li *et al.*, 2017), OfficeHome (Venkateswara *et al.*, 2017), Digits (Volpi *et al.*, 2018), and WILDS (Koh *et al.*, 2021) which is a compendium of MSDG datasets.

In the **Single-Source Domain Generalization** setting, only one domain is available for training, and as SSDG is harder as MSDG methods are infeasible; most work has therefore focused on data augmentation. Notable among these methods is the idea of adversarial data augmentation – ADA(Volpi *et al.*, 2018) and M-ADA (Qiao *et al.*, 2020) apply pixel-level additive perturbations to the image in order to fool the classifier. Resulting images are used as augmented data to train the classifier. RandConv (Xu *et al.*, 2020b) shows that shape-preserving transformations in the form of random convolutions of images lead to impressive performance gains on Digits.

Adversarial Attack and Defense. Adversarial attack algorithms have been developed to successfully fool image classifiers via pixelwise perturbations (Goodfellow *et al.*, 2015; Moosavi-Dezfooli *et al.*, 2016; Carlini and Wagner, 2017; Dong *et al.*, 2018). Algorithms have been developed to defend against such adversarial attacks (Moosavi-Dezfooli *et al.*, 2016; Dhillon *et al.*, 2018; Yuan *et al.*, 2019; Jang *et al.*, 2019). The scope of this paper is not to perform adversarial attack and defense, but to develop a framework to obtain adversarially generated samples that improve domain generalization performance.

Adversarial Training. In ALT, we emphasize on the nature of the diversity that could be acquired during training, which is crucial in the single-source setting. ALT learns adversarial perturbations in the function space of *neural network weights*. This allows us access to a wider and richer space of augmentations compared to pixel-wise perturbations such as ADA and M-ADA, or combinatorial augmentation search methods such as ESDA (Volpi and Murino, 2019). The adversarial component in ALT allows the network to seek newer and harder transformations for every batch

as training progresses, which cannot be achieved with static augmentations such as AugMix or RandConv, or by utilizing normalization layer statistics for style debiasing (Nam *et al.*, 2021).

Robustness to Image Corruptions. There has also been interest in training classifiers that are robust to corruptions that occur in the real world, such as different types of noise and blur, artifacts due to compression techniques, and weather-related environments such as fog, rain, and snow. (Vasiljevic *et al.*, 2016; Geirhos *et al.*, 2018) show that training models with particular types of corruption augmentations does not guarantee robustness to other unseen types of corruptions or different levels of corruption severity. Hendrycks *et al.* (Hendrycks and Dietterich, 2019) curate benchmarks (ImageNet-C and CIFAR-C) to test robustness along a fixed set of corruptions. They also provide a benchmark called ImageNet-P which tests robustness against other corruption types such as small tilts and changes in brightness. A similar benchmark for corruptions of handwritten digit images, MNIST-C (Mu and Gilmer, 2019) has also been introduced.

Data Augmentation has been an effective strategy for improving in-domain generalization using simple techniques such as random cropping, horizontal flipping (He *et al.*, 2016), occlusion or removal of patches (DeVries and Taylor, 2017; Zhong *et al.*, 2020). Data augmentation techniques have been shown to improve robustness against adversarial attacks and natural image corruptions (Zhang *et al.*, 2018; Yun *et al.*, 2019; Cubuk *et al.*, 2020). Learning to augment data has been explored in the context of object detection (Zoph *et al.*, 2020) and image classification (Ratner *et al.*, 2017; Cubuk *et al.*, 2019; Zhang *et al.*, 2020).

3.3 Proposed Approach

Under the single-source domain generalization setting, consider the training dataset \mathcal{D} containing N image-label pairs $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, and a classifier f parameterized by neural network weights θ . The standard expected risk minimization (ERM) approach seeks to learn θ by minimizing the in-domain risk measured by a suitable loss function such as the cross-entropy loss.

$$\mathcal{R}_{ERM} = \mathbb{E}_{\mathbf{x} \in \mathcal{D}} \mathcal{L}_{CE}(f(\mathbf{x}; \theta), \mathbf{y}). \quad (3.1)$$

For SSDG, we are interested in a classifier that has the least risk on several *unseen* target domains \mathcal{D}' that are not observed during training. We consider SSDG under covariate shift, i.e. when $P(X)$ changes but $P(Y|X)$ remains the same. Our approach builds on diversity based and adversarial augmentation approaches which we outline next.

Generalization via Maximizing Diversity. A successful strategy to improve generalization on unseen domains is to utilize a set of pre-defined data augmentations \mathcal{F}_{div} , to emphasize the invariance properties that are important for $f(\theta)$ to learn. Such methods modify Equation 3.1 as:

$$\mathcal{R}_{div} = \mathbb{E}_{\mathbf{x} \in \mathcal{D}} \mathcal{L}_{CE}(f(\mathbf{x}; \theta), \mathbf{y}) + \lambda_{KL} D_{KL}, \quad (3.2)$$

where D_{KL} is a consistency term, typically a divergence, such as KL-Divergence, between the softmax probabilities of the classifier obtained with the clean and transformed data, respectively, *e.g.*, $D_{KL} = KL(f(\mathbf{x}) || f(\mathcal{F}_{div}(\mathbf{x})))$. The choice of \mathcal{F}_{div} leads to different types of augmentations; for instance, AugMix ([Hendrycks et al., 2020c](#)) utilizes a combination of pre-defined transformations such as shear, rotate, color jitter, An approach proposed by [Xu et al. \(2020b\)](#) is to apply a randomly initialized convolutional layer to the input image. Methods such as these are effective strategies to

enforce diversity-based consistencies for generalization. Although these methods have the advantage of being simple pre-defined transformations that are dataset agnostic, they suffer from drawbacks under the SSDG setting. When executed on their own, they may not capture sufficient diversity in terms of **large** semantic shifts, such as when expecting generalization on sketches from a model trained on photos.

Generalization via Adversarial Hardness. An alternative domain generalization approach is via adversarial augmentation which exposes a classifier to ‘hard’ samples during training – defined broadly as examples that are carefully designed to cause the model to fail. Such samples are augmented to the training set, with the expectation that exposure to such adversarial examples can improve the model’s generalization performance on unseen domains ([Volpi et al., 2018](#); [Qiao et al., 2020](#)). This is commonly enforced by learning an additive noise vector which when added, maximizes classifier cost. Unfortunately in the case of domain generalization, these methods have failed to match the performance of diversity-only methods optimizing for the cost outlined in Equation 3.2. This is in part because they lack sufficient diversity, and by design they can only guarantee robustness to small perturbations from the training domain, as opposed to large semantic and stylistic shifts, which are crucial for domain generalization.

3.3.1 ALT: Adversarially Learned Transformations

While diversity-only methods have shown promise, they are limited in their ability to generalize to domains with large shifts. On the other hand, techniques based purely on adversarial hardness are theoretically well-motivated but do not match the performance of diversity-based methods. In this paper, we propose a new approach that takes the best of these two approaches using an adversary network that is trained

to create *semantically consistent* image transformations that fool the classifier. These manipulated images are then used during training as examples on which the image must learn invariance. Since these perturbations are parameterized as learnable weights of a neural network, the network is free to choose large, complex transformations without being restricted to additive noise as done in previous work (Volpi *et al.*, 2018). Further, this network is randomly initialized for each batch, making the types of adversarial transformations discovered unique and diverse over the course of training. Formally, the adversary network g transforms the input image as

$$\mathbf{x}_g = g(\mathbf{x}), \quad \text{where } g: \curvearrowright^{C \times H \times W} \rightarrow \curvearrowright^{C \times H \times W} \quad (3.3)$$

where C, H, W are the number of channels, height, and width of input images. g is parameterized by weights ϕ . This network, dubbed ALT, forms the backbone of our method.

To train ALT, we setup an adversarial optimization problem with the goal of producing transformations, which when applied to the source domain, can fool the classifier f . While existing efforts dealing with robustness to small corruptions use ℓ_p norm-bounded pixel-level perturbations to fool the model, we find that this is not sufficient for domain generalization as such methods do not allow searching for adversarial samples with semantic changes. Instead, we directly perform adversarial training in the space of ϕ , i.e., the neural network weights of ALT. Given input images \mathbf{x} , parameters ϕ are randomly initialized, and the corresponding adversarial samples \mathbf{x}_g are found as:

$$\mathbf{x}_g = \max_{\phi} \mathcal{L}_{CE}(f(g(\mathbf{x}; \phi); \theta), y) - \mathcal{L}_{TV}(g(\mathbf{x}; \phi)). \quad (3.4)$$

The first term seeks to update ϕ to maximize the classifier loss, while \mathcal{L}_{TV} (total variation) (Rudin *et al.*, 1992) acts as a smoothness regularization for the generated

Algorithm 2 Adaptive Diversity via ALT

Input: Source dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$

Output: Network Parameters θ^*

```

1: Initialize:  $\theta \leftarrow \theta_0$                                      ▷ weights of  $f()$ 
2: for each  $t \in \{1 \dots T\}$  do
3:    $x_t, y_t \sim \mathcal{D}$                                          ▷ sample input batch
4:   if  $t < T_{pre}$  then
5:      $\theta \leftarrow \theta - \eta \nabla \mathcal{L}_{CE}(f(x_t; \theta), y_t)$ 
6:   else
7:      $\rho \leftarrow \rho_0, \phi \leftarrow \phi_0$                                ▷ weights of  $r(), g()$ 
8:     for each  $i \in 1 \dots m_{adv}$  do
9:        $\hat{y}_g \leftarrow f(g(x; \phi); \theta)$ 
10:       $\phi \leftarrow \phi + \nabla(\mathcal{L}_{cls}(\hat{y}_g, y) - \mathcal{L}_{TV}(x_g))$ 
11:    end for each
12:     $\theta \leftarrow \theta - \eta_{adv} \nabla \mathcal{L}_{ALT}$                       ▷ see Equation 3.5, 3.7
13:  end if
14: end for each
15: return  $\theta$ 

```

image $x_g = g(x; \phi)$. The maximization in Eq. 3.4 is solved by performing m_{adv} steps of gradient descent with learning rate η_{adv} . We note a few important aspects of ALT – unlike existing methods that explicitly place an ℓ_p -norm constraint on the adversarial perturbations, we control the strength of the adversarial examples by limiting the number of optimization steps taken by g to maximize classification error. Next, since we randomly initialize g for each batch, the network is reset to a random function. In fact, when the number of adversarial steps is set to 0, g behaves similar to RandConv (Xu *et al.*, 2020b) since it is only a set of convolutional layers, with additional non-linearity. Finally, in addition to limiting the number of adversarial steps, we place a simple total variation loss on the generated image to force smoothness in the output.

This naturally suppresses high frequency noise-like artifacts and encourages realistic image transformations. It also prevents the optimization from resorting to learning trivial transformations in order to maximize classifier loss, such as noise addition or entirely removing or obfuscating the semantic content of the image.

Improving Diversity. The samples \mathbf{x}_g obtained by solving Equation 3.4 represent hard adversarial images that can be leveraged by the model to generalize to domain shift. But it also lends itself to exploit other forms of naïve diversity achieved by methods like RandConv and AugMix. We represent these “diversity modules” as r , which produce outputs $\mathbf{x}_r = r(\mathbf{x})$. Our method utilizes these samples in the training process by enforcing a consistency between the predictions of the classifier on the source image and its transformations from r and g . By including the diversity module into the optimization process, the invariances inferred by the classifier lead to stronger and more diverse adversarial examples in future epochs. Eventually, a synergistic partnership emerges between the diversity module and the adversary network to produce a wide range of image transformations that are significantly different from the source domain.

Let p_c , p_r , and p_g denote the softmax prediction probabilities of classifier f on \mathbf{x} , \mathbf{x}_r , and \mathbf{x}_g , respectively. Then the consistency between these predictions can be computed using Kullback-Leibler divergence ([Kullback et al., 1951](#)) as:

$$\begin{aligned} \mathcal{L}_{KL} &= D_{KL}(p_{mix} || p_c) + w_r D_{KL}(p_{mix} || p_r) \\ &\quad + (2 - w_r) D_{KL}(p_{mix} || p_g), \end{aligned} \tag{3.5}$$

where p_{mix} denotes the mixed prediction:

$$p_{mix} = \frac{p_c + w_r p_r + (2 - w_r) p_g}{3}. \tag{3.6}$$

The weight $w_r \in [0, 2]$ controls the relative contribution of diversity and adversity to

the consistency loss; $w_r > 1$ implies more weight on consistency with the diversity module; $w_r < 1$ implies more weight on consistency with the adversary network. In our experiments, we use $w_r = 1$, i.e., both diversity and adversary are given equal importance.

Our final loss function for training the classifier is given as the convex combination of the consistency \mathcal{L}_{KL} and the classifier loss $\mathcal{L}_{cls} = \mathcal{L}_{CE}(f(g(\mathbf{x}); \theta), \mathbf{y})$, as shown below:

$$\mathcal{L}_{ALT} = (1 - \lambda_{KL})\mathcal{L}_{cls} + \lambda_{KL}\mathcal{L}_{KL}. \quad (3.7)$$

Implementation. Algorithm 2 shows how ALT is implemented. In our experiments, we use RandConv or AugMix as the diversity module r and a fully-convolutional image-to-image network as the adversary network g . g has 5 convolutional layers with kernel size 3 and LeakyReLU activation. We train the classifier for a total of T batch iterations of which T_{pre} iterations are used for pre-training the classifier using standard ERM on only the source domain (with only \mathcal{L}_{cls}). During each batch iterations $t > T_{pre}$, we randomly initialize the weights of both r and g with the ‘‘Kaiming Normal’’ strategy (He *et al.*, 2016) as our starting point for producing diverse perturbations, and update g using the adversarial cost in Equation 3.4. After g is adversarially updated for the given batch, we use the combination of classifier loss and consistency in Equation 3.7 to update model parameters θ .

3.4 Experiments

We validate our approach with extensive empirical analysis of ALT and its constituent parts using three popularly used domain generalization benchmarks.

Variable	Digits	PACS	Office-Home
f architecture	DigitNet (Volpi <i>et al.</i> , 2018)	ResNet18 (He <i>et al.</i> , 2016)	ResNet50 (He <i>et al.</i> , 2016)
g architecture		$\{\text{conv}_{kernel=3, stride=1, padding=1, leakyReLU_{p=0.2}}\} \times 4$	
ρ_0, ϕ_0		Kaiming Normal Initialization (He <i>et al.</i> , 2015)	
T	10000	2000	2000
T_{pre}	1250	400	400
η	1e-4	0.004	0.004
m_{adv}	10	10	10
η_{adv}	5e-6	5e-5	5e-5
w_r	1.0	1.0	1.0
λ_{KL}	0.75	0.75	0.75

Table 3.1: Training settings and hyper-parameters for experiments on each benchmark.

Datasets. The SSDG setup is as follows: we train on a single source domain, and evaluate its performance on unobserved target (or test) domains with no access to any data from them during training. We demonstrate the effectiveness of our approach using three popular domain generalization benchmark datasets: **(a) *PACS*** (Li *et al.*, 2017) consists of images belonging to 7 classes from 4 domains (photo, art painting, cartoon, sketch); we choose one domain as the source and the rest as target domains. **(b) *Office-Home*** (Venkateswara *et al.*, 2017) consists of images belonging to 65 classes from 4 domains (art, clipart, real, product); we choose one domain as the source and the rest as target domains. **(c) *Digits***: we follow the setting from Volpi *et al.* (Volpi *et al.*, 2018) and use 1000 images from MNIST (LeCun *et al.*, 1998) as the source dataset, and USPS (Denker *et al.*, 1988), SVHN (Netzer *et al.*, 2011), MNIST-M and SYNTH (Ganin and Lempitsky, 2015) as the target datasets.

Training Settings Table 3.1 shows the training settings and hyperparameters used for experiments on each benchmark.

Method	A→C	A→S	A→P	C→A	C→S	C→P	S→A	S→C	S→P	P→A	P→C	P→S	Average
ERM	62.3	49.0	95.2	65.7	60.7	83.6	28.0	54.5	35.6	64.1	23.6	29.1	54.3
JiGen (Carlucci et al., 2019)	57.0	50.0	96.1	65.3	65.9	85.5	26.6	41.1	42.8	62.4	27.2	35.5	54.6
ADA (Volpi et al., 2018)	64.3	58.5	94.5	66.7	65.6	83.6	37.0	58.6	41.6	65.3	32.7	35.9	58.7
AugMix (Hendrycks et al., 2020c)	68.4	54.6	95.2	74.3	66.7	87.3	40.0	57.4	46.8	67.3	26.8	41.4	59.6
RandConv (Xu et al., 2020b)	61.1	60.5	87.3	57.1	72.9	73.7	52.2	63.9	46.1	61.3	37.6	50.5	60.3
SagNet (Nam et al., 2021)	67.1	56.8	95.7	72.1	69.2	85.7	41.1	62.9	46.2	69.8	35.1	40.7	61.9
ALT _{g-only}	63.5	63.8	94.9	68.9	74.4	84.6	39.7	61.1	49.3	68.8	43.4	50.8	63.6
ALT _{RandConv}	63.6	65.8	92.5	69.1	75.1	84.5	40.1	61.7	50.8	68.4	43.4	55.2	64.2
ALT _{AugMix}	65.7	68.2	93.2	71.9	74.2	86.0	40.2	62.9	49.1	68.5	43.5	53.3	64.7

Table 3.2: Single-source domain generalization accuracy (%) on PACS ([Csurka, 2017](#)). $X \rightarrow Y$ implies X is the source dataset and Y is the target dataset. P : photo; A : art-painting; C : cartoon; S : sketch. Performance is reported as mean of 5 repetitions¹.

Evaluation. For all datasets, we train models on each individual domain, and test on the remaining domains. We provide fine-grained results on each test set as well as the average domain generalization performance. We compare with several state-of-the-art techniques on SSDG and compare three variants of our methods: ALT_{g-only} refers to the simplest form of our method that only uses the adversary network during training without an explicit diversity module r . ALT_{RandConv} and ALT_{AugMix} utilize RandConv and AugMix, respectively, as the diversity module, where the consistency is now placed as explained in Equation 3.5.

3.4.1 PACS

Baselines. Our baselines are JiGen ([Carlucci et al., 2019](#)), ADA ([Volpi et al., 2018](#)), AugMix ([Hendrycks et al., 2020c](#)), RandConv ([Xu et al., 2020b](#)), and SagNet ([Nam et al., 2021](#)) – designed to reduce style bias using normalization techniques. We also implement a combination of RandConv and AugMix – i.e. instead of the ALT formulation of using a diversity module and our adversary network, we use two diversity

modules (RandConv and AugMix) and enforce the same consistency as Equation 3.5. This allows us to compare how effective the adversary network is, compare to using two sources of diversity. We use ResNet18 ([He et al., 2016](#)) pre-trained on ImageNet as our model architecture and train all models for 2000 iterations with batch-size of 32, learning rate 0.004, SGD optimizer with cosine annealing learning rate scheduler, weight decay of 0.0001, and momentum 0.9. For ALT, we set consistency coefficient $\lambda_{KL}=0.75$, adversarial learning rate $\eta_{adv}=5e-5$, number of adversarial steps $m_{adv}=10$ and $w_r=1.0$.

Results. Results are shown in Table 3.2. We observe that ALT without a diversity module (ALT_{g-only}) surpasses generalization performance of all prior methods including diversity methods RandConv and AugMix and the previous best SagNet ([Nam et al., 2021](#)). ALT with adaptive diversity further improves the results and ALT_{AugMix} establishes a new state-of-the-art accuracy of 64.7%. All three variants of ALT are better than the combination of RandConv+AugMix, providing further evidence that adversarially learned transformations are more effective than combinations of diversity-based augmentations. The *Sketch* (S) target domain (human drawn black-and-white sketches of real objects) has been the most difficult for previous methods; the difficulty can be observed in terms of performance in columns $A \rightarrow S$, $C \rightarrow S$, and $P \rightarrow S$. ALT significantly improves the performance on the sketch target domain. Generalizing from photos as source to C, S, A as targets is a very realistic setting, since large-scale natural image datasets such as ImageNet ([Deng et al., 2009](#)) are widely used and publicly available, while data for sketches, cartoons, and paintings are limited. ALT is the best model under this realistic setting.

Method	A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Average
ERM	42.61	59.18	69.45	48.37	56.09	59.38	46.07	40.18	68.19	63.12	45.13	74.34	56.00
AugMix (Hendrycks et al., 2020c)	45.31	61.88	71.88	49.30	58.93	62.24	50.04	42.59	71.51	64.10	47.56	75.95	58.44
RandConv (Xu et al., 2020b)	43.98	55.28	67.31	45.49	56.58	59.03	43.80	43.19	66.50	57.62	48.26	72.97	55.00
SagNet (Nam et al., 2021)	42.18	56.03	67.34	46.68	53.89	57.88	45.49	40.09	67.11	61.39	48.32	72.79	54.93
ALT _{g-only}	47.26	61.14	71.21	48.88	57.81	60.99	48.15	46.70	69.30	64.85	52.84	76.28	58.78
ALT _{RandConv}	48.33	61.19	71.75	50.13	58.82	62.26	49.21	47.03	70.53	64.88	53.10	76.07	59.44
ALT _{AugMix}	48.06	61.16	71.12	50.43	58.84	61.84	49.32	47.55	70.64	64.86	53.27	76.29	59.45

Table 3.3: Single-source domain generalization accuracy (%) on Office-Home ([Venkateswara et al., 2017](#)) over five repetitions. $X \rightarrow Y$ implies X is the source dataset and Y is the target dataset. R: *real*; A: *art*; C: *clipart*; P: *product*. Performance is reported as mean of 5 repetitions¹.

3.4.2 Office-Home

Baselines. For OfficeHome, we follow the protocol from the previous state-of-the-art Sagnet ([Nam et al., 2021](#)) and use ResNet50 as the model architecture. Note that we do not perform any hyperparameter tuning for OfficeHome and directly apply identical training settings and hyperparameters from PACS.

Results. Table 3.3 shows the results on Office-Home. We observe that RandConv (previous best on Digits) and SagNet (previous best on PACS) perform worse than ERM on OfficeHome, while AugMix is better by 2.44%. The combination of RandConv+AugMix is also worse than the ERM baseline. All three variants of ALT surpass prior results, with ALT_{AugMix} resulting in the best accuracy of 59.45%. The most difficult target domain for previous methods is *Clipart (C)*, possibly because most clip-art images have white backgrounds, while real world photos (R) and product images are naturally occurring. ALT improves performance in each case with C as the target domain. An observation similar to PACS can also be made here – ALT is

Method	MNIST-10K	MNIST-M	SVHN	USPS	SYNTH	Target Avg.
ERM	98.40 ± 0.84	58.87 ± 3.73	33.41 ± 5.28	79.27 ± 2.70	42.43 ± 5.46	53.50 ± 4.23
ADA (Volpi et al., 2018)	N/A	60.41	35.51	77.26	45.32	54.62
M-ADA (Qiao et al., 2020)	99.30	67.94	42.55	78.53	48.95	59.49
AugMix (Hendrycks et al., 2020c)	98.53 ± 0.18	53.36 ± 1.59	25.96 ± 0.80	96.12 ± 0.72	42.90 ± 0.60	54.59 ± 0.50
RandConv (Xu et al., 2020b)	98.85 ± 0.04	87.76 ± 0.83	57.62 ± 2.09	83.36 ± 0.96	62.88 ± 0.78	72.88 ± 0.58
ALT _{<i>g-only</i>}	98.46 ± 0.27	74.28 ± 1.36	52.25 ± 1.54	94.99 ± 0.68	68.44 ± 0.98	72.49 ± 0.87
ALT _{<i>RandConv</i>}	98.46 ± 0.25	76.90 ± 1.42	53.78 ± 1.97	95.40 ± 0.72	69.40 ± 1.07	73.87 ± 1.03
ALT _{<i>AugMix</i>}	98.55 ± 0.11	75.98 ± 0.59	55.01 ± 1.34	96.17 ± 0.45	69.93 ± 2.17	74.38 ± 0.86

Table 3.4: Single-source domain generalization accuracy (%) on digit classification, with MNIST-10K as source and MNIST-M, SVHN, USPS, and SYNTH as target domains. Performance reported over five repetitions. Note: ADA and M-ADA do not report standard deviation.

the best model under the realistic setting of generalizing from widely available real photos (R) to other domains.

3.4.3 Digits

Baselines. Our baselines include a naïve “source-only” model trained using expected risk minimization (ERM) on the source dataset, M-ADA ([Qiao et al., 2020](#)) – an adversarial data augmentation method, and AugMix ([Hendrycks et al., 2020c](#)) and RandConv ([Xu et al., 2020b](#)) which exploit diversity through consistency constraints. We also compare with ESDA ([Volpi and Murino, 2019](#)), an evolution-based search procedure over a pre-defined set of augmentations ([Cubuk et al., 2019](#)). We use DigitNet ([Volpi et al., 2018](#)) as the model architecture for all models for a fair comparison. All models are trained for $T=10000$ iterations, with batch-size of 32, learning rate of 0.0001, using the Adam optimizer. For ALT, we set the consistency

coefficient $\lambda_{KL}=0.75$, adversarial learning rate $\eta_{adv}=5e-6$, number of adversarial steps $m_{adv}=10$, and equal weight $w_r=1.0$ for diversity and adversary networks.

Results. Table 3.4 shows that pixel-level adversarial training approaches (ADA and M-ADA) offer only marginal improvements over the naïve ERM baseline. The results for diversity-promoting data augmentation methods are mixed – while AugMix is only 1.09% better than ERM, RandConv provides a significant boost. Interestingly, the base version of our approach, ALT_{g-only} , which is exclusively based on adversarial training, is significantly better than pixel-level adversarial training. More importantly, it is also better than diversity method AugMix, while performing lower than RandConv by a small margin 0.39%. When we trained ALT with adaptive diversity ($ALT_{RandConv}$ and ALT_{AugMix}), we achieved the best performance, beating previous state-of-the-art. SVHN and SYNTH are the hardest target domains as they contain real-world images of street signs or house number signs, whereas USPS is closely correlated with MNIST, both being black-and-white centered images of handwritten digits, and MNIST-M is derived from MNIST but with different backgrounds. AugMix fares poorly on both real-world datasets, but is able to generalize well to MNIST-M and USPS. Although AugMix results in an average accuracy of 54.59% on the target domains, when used in conjunction with ALT, the ALT_{AugMix} leads to a large gain of 19.79%, highlighting the significance of the adversary network.

3.5 Analysis of ALT

In this section we study the various components of ALT, and provide insights into their impact on generalization.

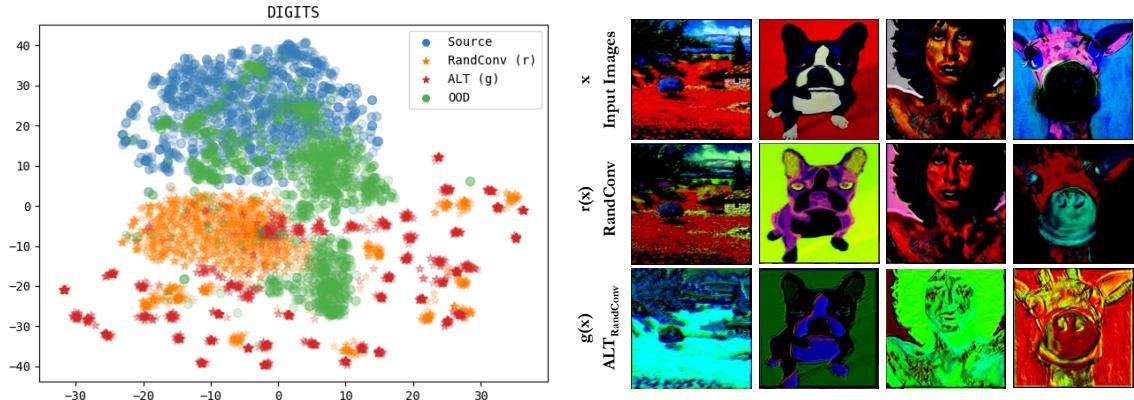


Figure 3.2: (*Left*) tSNE plot showing the discrepancy between the source distribution (MNIST) and the out-of-distribution datasets for the “Digits” benchmark. The diversity introduced by ALT is much larger and wide-spread than data augmentation techniques such as RandConv. (*Right*) Qualitative Comparison of PACS images transformed by RandConv data augmentation vs. ALT ($ALT_{RandConv}$), illustrating the wide range of transformations learned by ALT.

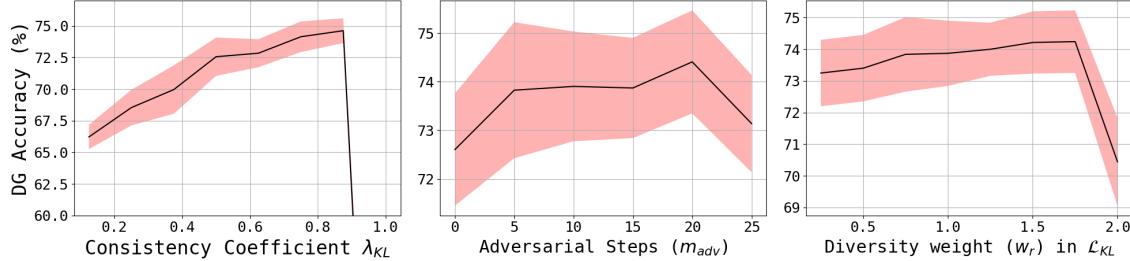


Figure 3.3: **Analysis:** We study the effect of each hyper-parameter in ALT on the average accuracy using the Digits benchmark (shown as 1 standard deviation around the mean over 5 runs). We observe that the consistency (*left*) is generally important until a certain point, after which it becomes harmful; (*middle*) taking more adversarial steps improves performance; (*right*) surprisingly, we find that the trade-off between diversity and adversity is non-trivial and dataset dependent. In our benchmarking experiments (Tables 3.2, 3.3 , 3.4) we do not perform any hyper-parameter tuning, and set $w_r=1$, i.e. equal weight to adversity and diversity.

3.5.1 ALT is better than naïve diversity.

Our first big insight is that ALT without an explicit diversity ($\text{ALT}_{g\text{-only}}$) module still outperforms all the top performing methods across the benchmarks we evaluated on, indicating that learned adversarial transformations are a powerful way to train classifiers for generalization.

Our next observation is that ALT makes the choice of diversity module fairly arbitrary. We see this effect on multiple benchmarks – for example, on the Digits benchmark shown in Table 3.4, AugMix has a relatively poor generalization performance when compared with the baseline ERM whereas $\text{ALT}_{\text{Augmix}}$ achieves state of the art. This is again seen in the Office-Home benchmark shown in Table 3.3, where RandConv is worse than ERM, but $\text{ALT}_{\text{RandConv}}$ is the best performing method. Thus, irrespective of the choice of diversity module, the adversarially learned transformations benefit generalization on all benchmarks.

In Figure 3.2 (*left panel*) we analyze the diversity introduced by ALT on the Digits benchmark, in comparison to the source distribution the target (OOD) distribution and the distribution of RandConv augmentations. While RandConv does simulate a domain shift compared to the source, most RandConv points are clustered close to each other. However, the diversity due to ALT is considerably larger and ALT samples are spread widely across the tSNE space. We believe this is because data augmentation functions have a fixed types of diversity (random convolution filter in the case of RandConv), while ALT *searches* for adversarial transformations for each batch – this leads to novel types of diversity for each batch of training samples. We also show qualitative examples of the image transformations learned with ALT in Figure 3.2, and it is clear that ALT achieves far more diverse and larger transformations of the input images than previous data augmentation techniques.

3.5.2 Effect of Varying ALT Hyperparameters.

The three main hyper-parameters that control ALT are: **(1)** λ_{KL} – the coefficient in Eq. 3.5 which decides the weight for the KL-divergence consistency in the total loss, **(2)** m_{adv} – the number of adversarial steps in the adversarial maximization of Eq. 3.4, and **(3)** w_r – the diversity weight which controls the interaction between the diversity module $r()$ and the adversary network $g()$ in Eq. 3.6. We investigate the effect of each of these on domain generalization accuracy in Figure 3.3. The first plot shows that the consistency coefficient λ_{KL} is impactful and a higher value leads to better generalization. However at $\lambda_{KL} = 1.0$ the accuracy degenerates to random performance; this is expected as the classifier loss gets $1 - \lambda_{KL}=0$ weight. From the second plot, we observe that the optimal number of adversarial steps is around 20. Note that performance at all non-zero values of m_{adv} that we tried (5, 10, 15, 20, 25) is greater than previous state-of-the-art. The importance of the adversarial module is evident from the third plot – performance at $w_r = 0$ (adversarial module only) is higher than performance of $w_r = 2$ (diversity module only), and the combination of both modules yields the best result. Clearly, the adversarial component is a critical factor that causes improvements in generalization.

3.5.3 Complexity of Adversary Network.

One limitation (and therefore scope for future work) is that we have considered one family of architecture for our adversary network g – fully convolutional image-to-image translation networks. We conduct additional analysis to understand how this choice affects generalization performance, and compare performance when using between 2 and 6 convolutional layers. We reuse all other training settings from our benchmark model $ALT_{RandConv}$ on both Digits and PACS. Results are shown in Table 3.5.

Benchmark	FCN Number of Layers				
	2	3	4	5	6
Digits	72.75	73.74	74.10	73.87	74.15
PACS	63.40	63.92	64.41	64.20	63.78
OfficeHome	59.67	59.56	59.79	59.42	59.45

Table 3.5: Effect of the depth (number of convolutional layers) of the adversary network g on average domain generalization on all three benchmarks.

For PACS and OfficHome, we observe that all ALT models compared are better than previous baselines including AugMix and RandConv. For Digits, we observe that performance of ALT with a 2-layer g is close to RandConv, and is greater than all previous baselines for higher depth of the network. We do not see a clear correlation across datasets between the number of layers and the domain generalization performance. Investigating the dynamics of model capacity of the adversary network and how it may affect domain generalization, is an interesting direction for future work.

It may be possible that more complex generative architectures (i.e. greater complexity of transformations) may be needed for larger domain shift is larger, to model diversity and adversity for a given source domain. Thus the choice of architecture for g is an interesting direction; nevertheless, in this paper we show that the simple fully convolutional architecture gives us performance boosts in all three datasets.

We believe that ideas presented in this paper, although evaluated on image classification, have the potential of being widely applicable to many other vision tasks for domain generalization. They may also be applied to other application areas such as audio or text, where the transformation function g may take different forms.

3.6 Conclusion

In this paper, we address the problem of single source domain generalization. Our approach, Adversarially Learned Transformations (ALT) uses a randomly initialized convolutional network to learn plausible image transformations of the source domain that can fool the classifier. These images are used to enforce a consistency with the predictions on clean images. We showed that this strategy outperforms all existing techniques on multiple benchmarks because it is able to generate a diverse set of large transformations of the source domain. We also find that ALT can be naturally combined with existing diversity modules like RandConv or AugMix to improve their performance (sometimes significantly). We also studied the different parts of ALT through extensive ablations and analysis to obtain insights into its performance gains. Our studies indicate that naïve diversity alone is insufficient, but needs to be combined adversarial transformations to maximize generalization performance.

As a follow-up work to ALT, we developed Adversarial Bayesian Augmentation (ABA) draws on the strengths of adversarial learning and Bayesian neural networks to guide the generation of diverse data augmentations. The introduction of weight uncertainty by the Bayesian neural network further enhances the strength of data augmentation, leading to state of the art results on domain shift in standard image classification datasets as well as medical image classification and subpopulation shift datasets. More details on ABA can be found in [Cheng *et al.* \(2023\)](#).

Chapter 4

COVARIATE SHIFT DETECTION VIA DOMAIN INTERPOLATION SENSITIVITY

Covariate shift is a major roadblock in the reliability of image classifiers in the real world. Work on covariate shift has been focused on training classifiers to adapt or generalize to unseen domains. However, for transparent decision making, it is equally desirable to develop *covariate shift detection* methods that can indicate whether or not a test image belongs to an unseen domain. In this paper, we introduce a benchmark for covariate shift detection (CSD), that builds upon and complements previous work on domain generalization. We use state-of-the-art OOD detection¹ methods as baselines and find them to be worse than simple confidence-based methods on our CSD benchmark. We propose an interpolation-based technique, Domain Interpolation Sensitivity (DIS), based on the simple hypothesis that interpolation between the test input and randomly sampled inputs from the training domain, offers sufficient information to distinguish between the training domain and unseen domains under covariate shift. DIS surpasses all OOD detection baselines for CSD on multiple domain generalization benchmarks.

4.1 Introduction

Machine learning models such as image classifiers are being increasingly deployed in real-world settings. Covariate shift is a commonly occurring phenomena, where test images are from the same categories as the training data, but undergo a shift in terms

¹Recent literature uses the term “OOD detection” to refer to novel category detection.

of style. For instance, the training data may contain images taken during the day under sunny conditions, but the classifier may encounter nighttime images or foggy or rainy images. Models trained under the empirical risk minimization (Vapnik and Chervonenkis, 1991) paradigm can only offer performance guarantees under the *i.i.d.* setting, and are known to fail under various types of covariate shift (Taori *et al.*, 2020; Hendrycks and Dietterich, 2019; Beery *et al.*, 2018).

To mitigate the risks associated with covariate shift, domain generalization algorithms have been developed (Arjovsky *et al.*, 2019; Volpi *et al.*, 2018; Xu *et al.*, 2020b; Gokhale *et al.*, 2023). However, improving the accuracy of classifiers on unseen domains cannot be the only criteria for reliable decision making – for transparency, methods that *detect* covariate shift should also be investigated. Unfortunately, this aspect of reliable decision making has not been previously explored. In this paper, we investigate covariate shift detection (CSD) for image classifiers.

Methods for detecting “out-of-distribution” (OOD) test examples have been previously developed (Hendrycks and Gimpel, 2017; Liu *et al.*, 2020b; Liang *et al.*, 2018; Huang *et al.*, 2021; Bhattacharjee *et al.*, 2020). However it is important to note that OOD detection algorithms are designed to detect *novel categories* at test time. In this paper, we are interested in detecting covariate shift (i.e. detecting test inputs that belong to a previously unseen domain, but one of the classes that the classifier is trained on). Towards this end, we develop a new benchmark for covariate shift detection. We utilize common domain generalization benchmarks and train the classifier on one of the domains, and benchmark CSD methods’ performance in detecting images that belong to other domains. An example is shown in Figure 4.1 – we compare the two dimensions of reliable predictions under covariate shift. Domain generalization algorithms are trained with the aim of making accurate predictions for seen as well as unseen domains. However in cases where accuracy under domain shift is low, it is

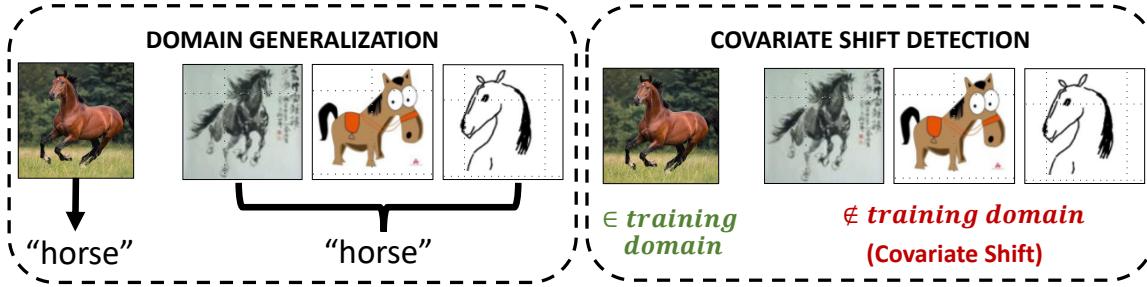


Figure 4.1: Domain generalization (DG) and covariate shift detection (CSD) are both important, but orthogonal aspects of robustness evaluation. While the aim of DG is to predict the correct label for inputs from unseen domains, the aim of CSD is to detect unseen domains – i.e. detect covariate shift in test inputs.

equally important to detect, or flag, cases where there may be a covariate shift for safe and reliable use of classifiers.

Interpolation between training examples has been shown to provide unique information suitable for model regularization (Nam *et al.*, 2021; Krueger *et al.*, 2021). Motivated by this, we hypothesize that interpolating test inputs with training inputs offers sufficient information to distinguish between in-domain and out-of-domain examples. We use test-time interpolations as a powerful tool for understanding the behavior of pre-trained classifiers and detecting covariate shift. Our motivation is as follows: if interpolations during training benefit model generalization, then interpolating test inputs with training inputs could help us understand how the model might perform on unseen test distributions.

We develop a method, named “Domain Interpolation Sensitivity” (DIS), that achieves state-of-the-art results on covariate shift detection for image classification and text classification. We find that methods developed for OOD detection (novel category detection) underperform on the CSD benchmark. Surprisingly we find that methods that perform better on OOD detection benchmarks than the maximum

softmax probability (MSP) baseline by Hendrycks et al. ([Hendrycks and Gimpel, 2017](#)), perform much worse than MSP on our CSD benchmarks.

Our contributions are summarized below:

- We study covariate shift detection (CSD) as a mechanism for improving the reliability of classifier predictions; CSD is designed to complement domain generalization as a robustness metric.
- We propose CSD benchmarks that are derived from existing benchmarks for domain generalization.
- We develop a interpolation-based technique that outperforms existing outlier and OOD detection methods on four CSD benchmarks (three for image classification, and one for text classification).

4.2 Covariate Shift Detection

We will consider classification tasks, for which a neural network f is trained on a dataset \mathcal{D}_{in} containing labeled input–output pairs (x, y) , with inputs $x \in \mathcal{X}_{in}$ and outputs $y \in \mathcal{Y}_{in}$. Let \mathcal{D}_{out} denote previously unseen data. The nature of the shift between \mathcal{D}_{in} and \mathcal{D}_{out} can take multiple forms. One such type of distribution shift is the presence of novel categories in \mathcal{D}_{out} , i.e. if $(x, y) \in \mathcal{D}_{out}$, then the categorical label of x , $y \notin \mathcal{Y}_{out}$. An example of this phenomena of novel categories is if \mathcal{D}_{in} is a dataset for cat-dog classification, whereas \mathcal{D}_{out} contains images of handwritten digits. Another type of distribution shift can occur when the categorical space remains the same, but domain \mathcal{X}_{out} undergoes a covariate shift, i.e. $p_{in}(x) \neq p_{out}(x)$. For example, a covariate shift exists between \mathcal{D}_{in} and \mathcal{D}_{out} if \mathcal{D}_{in} is a set of real cat-dog images while \mathcal{D}_{out} contains cartoons or sketches of cats and dogs. In this paper, we will consider covariate shift.

Scoring Function for Covariate Shift Detection. Covariate shift detection can be formulated as a binary classification task. Given a classifier f trained on distribution \mathcal{D}_{in} , the goal is to design a estimator g that estimates whether or not a test input lies within the training domain.

$$g(x) = \begin{cases} 1 & \text{if } S(x) \geq \gamma \quad (\text{in-domain}) \\ 0 & \text{if } S(x) < \gamma \quad (\text{covariate shift}) \end{cases} \quad (4.1)$$

The threshold γ is chosen such that 95% of in-domain data is correctly classified by Eq. 4.1. The choice of the scoring function S is the key to improving covariate shift detection. Previous approaches for OOD detection have utilized the model’s outputs (for eg. the maximum softmax probability (Hendrycks and Gimpel, 2017), or energy of the softmax output (Liu *et al.*, 2020b)), or model’s gradient space (Huang *et al.*, 2021). In this work, we develop a scoring function $S(x)$ by leveraging the interpolation of the test input with training inputs.

Benchmarking Covariate Shift Detection. We leverage existing domain generalization datasets for benchmarking CSD. Specifically, we operate under the single-source domain genralization setting (Volpi *et al.*, 2018), where the classifier is trained only on one domain and tested on all domains within the dataset. This is illustrated in Figure 4.1 which shows domain generalization on the PACS (Li *et al.*, 2017) dataset – the classifier is trained on real-world photos (source domain) and tested on all domains including photos, art-paintings, cartoons, and sketches. Given a classifier trained on the source domain \mathcal{D}_{in} , the goal of a covariate shift detection algorithm is to use a scoring function $S(x)$ to estimate whether or not x belongs to the source domain or not. Thus, for any domain generalization dataset, we can compare performance of CSD algorithms on the corresponding CSD benchmark.

4.3 Domain Interpolation Sensitivity

In this section, we describe our method for covariate shift detection using test-time input interpolation.

Test-Time Input Interpolation. Consider a randomly sampled training input $x_i \in \mathcal{X}_{in}$ and a test input x . We define the interpolation of x_i and x to be $\hat{x} = h(x_i, x, \epsilon)$ for a mixing coefficient $\epsilon \in [0, 1]$. Note that $h(x_i, x, 0) = x$ and $h(x_i, x, 1) = x_i$. In practice, h can be implemented in multiple ways; for image classification tasks, we use a simple pixel-wise convex combination. For text classification, we use a token-wise swapping.

$$h_{pixelwise}(x_i, x, \epsilon) = \epsilon x_i + (1 - \epsilon)x . \quad (4.2)$$

Domain Interpolation Sensitivity. Let $[0, \epsilon, 2\epsilon \dots T\epsilon]$ be an increasing sequence of mixing coefficients such that $0 \leq \epsilon \leq T\epsilon \leq 1$. We generate a sequence of interpolated images

$$X_i = [h(x_i, x, 0), h(x_i, x, \epsilon), \dots, h(x_i, x, T\epsilon)]. \quad (4.3)$$

We obtain a corresponding sequence of softmax predictions probabilities from model f as:

$$Y_i = f(X_i) = [f(h(x_i, x, 0)), f(h(x_i, x, \epsilon_1)), \dots, f(h(x_i, x, \epsilon_T))]. \quad (4.4)$$

Note that we can generate Y_i for each choice of training exemplar X_i . By using n training exemplars, we can obtain an average prediction sequence \bar{Y} :

$$\bar{Y} = \left[\frac{1}{n} \sum_{i=1}^n f(h(x_i, x, 0)), \frac{1}{n} \sum_{i=1}^n f(h(x_i, x, \epsilon_1)), \dots, \frac{1}{n} \sum_{i=1}^n f(h(x_i, x, \epsilon_T)) \right]. \quad (4.5)$$

Let $c = \underset{\mathcal{Y}_{in}}{\operatorname{argmax}} f(x)$ be the predicted category for the test input. Then, the domain interpolation sensitivity curve is defined as the softmax probability of c in each element of \bar{Y} .

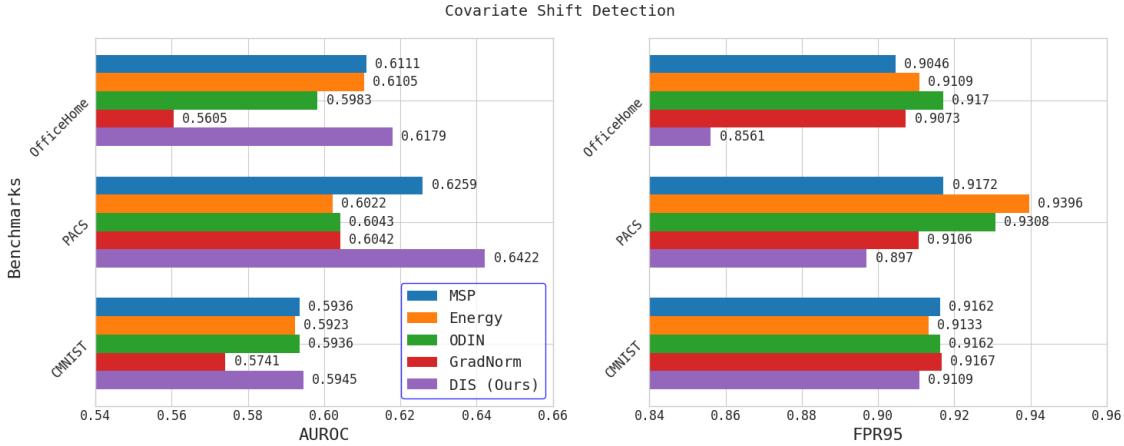


Figure 4.2: Summary of results on covariate shift detection benchmarks for image classification, in terms of AUROC (*higher value is better*) and FPR95 (*lower value is better*). Detailed results can be found in the appendix. The key observation is that recent OOD methods perform worse than the baseline MSP (Hendrycks and Gimpel, 2017), in terms of both AUROC and FPR95. Our DIS method improves CSD detection performance on all three benchmarks.

Once we've obtained the above DIS curves, we use the area under the DIS curve as the scoring function $S(x)$ for covariate shift detection, i.e. $S(x) = AUC(\bar{Y})$.

4.4 Experiments

Baselines. We use widely adopted OOD detection approaches MSP (Hendrycks and Gimpel, 2017), Energy (Liu *et al.*, 2020b), ODIN (Liang *et al.*, 2018), and GradNorm (Huang *et al.*, 2021) as our baselines. These methods do not rely on any additional training or other modifications to the classifier used for detection and are thus comparable to our own approach.

Metrics. For comparative evaluation of our method against the baselines, we use two standard metrics (Hendrycks and Gimpel, 2017); these are: (i) **AUROC**: area

under the ROC curve (Davis and Goadrich, 2006), (ii) **FPR95**: false positive rate on the OOD set when the true positive rate on the ID set is 95%.

Datasets. We consider three image classification benchmarks: PACS (Li *et al.*, 2017), OfficeHome (Venkateswara *et al.*, 2017), and ColoredMNIST (Arjovsky *et al.*, 2019), as well as a proposed text review classification benchmark inspired by (Hendrycks *et al.*, 2020a). PACS contains four domains: photos, art-paintings, cartoons, and sketches. OfficeHome contains four domains: real images, art, clipart, and product images. ColoredMNIST contains three domains of digit images with varying degrees of spurious correlations (+90, +80, -90) between the digit and color.

For the image classification benchmarks, we follow the training protocol from DomainBed (Gulrajani and Lopez-Paz, 2021) and train a ResNet-18 model (He *et al.*, 2016) on ‘Photos’ for PACS, ‘Real’ for OfficeHome and the domain with “+90” spurious correlation for ColoredMNIST.

Hyperparameters. For image classification experiments, we use $n=16$ exemplars, step size $\epsilon=0.05$, and number of interpolation steps $T=4$.

Results. Our results on the image classification and text classification benchmarks reveal the strength of the interpolation-based method, with consistent improvements both in terms of AUROC and FPR95 on both tasks. Figure 4.2 summarizes our results on the image classification CSD benchmarks. Interestingly, we observe that Energy, ODIN, and GradNorm, which have been shown to be better than MSP in the OOD detection literature, are in fact, much worse than MSP on all three benchmarks (OfficeHome, PACS, ColoredMNIST). DIS outperforms all baselines.

The tables below show detailed results on each image classification benchmark, split by domain. These correspond to the summarized results in Figure 4.2.

Method	Art	Clipart	Product	Average
MSP (Hendrycks and Gimpel, 2017)	0.6863 / 0.8763	0.6379 / 0.9200	0.5534 / 0.9552	0.6259 / 0.9172
Energy (Liu <i>et al.</i> , 2020b)	0.6716 / 0.9084	0.6016 / 0.9448	0.5334 / 0.9656	0.6022 / 0.9396
ODIN (Liang <i>et al.</i> , 2018)	0.6753 / 0.8946	0.6074 / 0.9414	0.5301 / 0.9564	0.6043 / 0.9308
GN (Huang <i>et al.</i> , 2021)	0.6479 / 0.8694	0.6106 / 0.9256	0.5542 / 0.9369	0.6042 / 0.9106
Ours	0.7093 / 0.8339	0.6529 / 0.9087	0.5645 / 0.9483	0.6422 / 0.8970

Table 4.1: Covariate Shift Detection performance on the OfficeHome benchmark. All methods use the same ResNet classifier trained on the “Real” domain. Results are shown as AUROC \uparrow / FPR95 \downarrow .

Method	Art	Cartoon	Sketch	Average
MSP (Hendrycks and Gimpel, 2017)	0.6652 / 0.8868	0.4159 / 0.9760	0.7522 / 0.8510	0.6111 / 0.9046
Energy (Liu <i>et al.</i> , 2020b)	0.6708 / 0.8782	0.4252 / 0.9641	0.7356 / 0.8904	0.6105 / 0.9109
ODIN (Liang <i>et al.</i> , 2018)	0.6571 / 0.9060	0.4098 / 0.9431	0.7279 / 0.9019	0.5983 / 0.9170
GN (Huang <i>et al.</i> , 2021)	0.6298 / 0.8782	0.3980 / 0.9521	0.6536 / 0.8917	0.5605 / 0.9073
Ours	0.6693 / 0.8227	0.4278 / 0.9341	0.7567 / 0.8115	0.6179 / 0.8561

Table 4.2: Covariate Shift Detection performance on the PACS benchmark. All methods use the same ResNet classifier trained on the “Photos” domain. Results are shown as AUROC \uparrow / FPR95 \downarrow

Method	+80	-90	Average
MSP (Hendrycks and Gimpel, 2017)	0.5225 / 0.9378	0.6647 / 0.8946	0.5936 / 0.9162
Energy (Liu <i>et al.</i> , 2020b)	0.5225 / 0.9361	0.6620 / 0.8905	0.5923 / 0.9133
ODIN (Liang <i>et al.</i> , 2018)	0.5225 / 0.9378	0.6647 / 0.8946	0.5936 / 0.9162
GN (Huang <i>et al.</i> , 2021)	0.5213 / 0.9381	0.6269 / 0.8954	0.5741 / 0.9167
Ours	0.5242 / 0.9379	0.6649 / 0.8841	0.5945 / 0.9109

Table 4.3: Covariate Shift Detection performance on the ColoredMNIST benchmark. All methods use the same CNN classifier and results are shown as AUROC \uparrow / FPR95 \downarrow

4.5 Related Work

Distributional Robustness. Several dimensions of distribution robustness have been studied, which can be broadly classified into adversarial robustness (Goodfellow *et al.*, 2015; Madry *et al.*, 2018a), natural distributional robustness, and spurious correlations. Work on natural distributional robustness includes conditions such as common corruptions (Hendrycks and Dietterich, 2019), variations along style (Hendrycks *et al.*, 2021), geometric (Wong and Kolter, 2020) and attribute-level shift (Gokhale *et al.*, 2021), different dataset sources (Venkateswara *et al.*, 2017; Li *et al.*, 2017). Spurious correlations of features such as background (Beery *et al.*, 2018; Sagawa *et al.*, 2020) or texture (Geirhos *et al.*, 2019) with label space have been studied. In terms of label shift, anomaly/outlier detection, novelty detection, open-set recognition, and OOD detection have been studied (Yang *et al.*, 2021).

Correlation between ID and OOD performance. It is well known that models tested on data with covariate shift suffer a drop in performance compared to in-domain (ID) accuracy. Recently, there have been several studies that find a positive correlation between ID and OOD performance for tasks in both computer vision (Miller *et al.*, 2021) and natural language processing (Miller *et al.*, 2020). However, Teney et al.(Teney *et al.*, 2022) show that under certain real-world conditions, a negative correlation might exist, i.e. a decrease in ID accuracy may benefit OOD performance. Moayeri et al.(Moayeri *et al.*, 2022) show that there are trade-offs between adversarial and natural distributional robustness. There is also empirical evidence (Gokhale *et al.*, 2022c) that suggests that data modification techniques (for instance, data augmentation or data filtering (Bras *et al.*, 2020)) may have a negative impact on adversarial robustness. With the context of these findings, there is a large gap in our understanding of different robustness settings – characterizing distribution

shift in different ways is therefore crucial as a model selection criteria. Our work aims to aid investigations in this direction.

Interpolation for Model Selection. Interpolations have been extensively used for representation learning (Zhang *et al.*, 2018; Yun *et al.*, 2019; Nam *et al.*, 2021) for training robust classifiers. Bhattacharjee et al.(Bhattacharjee *et al.*, 2020) used interpolation of inputs during training for novel category detection. SMURF (Feinglass and Yang, 2021) used token interpolation between input texts and a noise process to estimate the robustness of language models based on the average monotonicity of a perplexity measure.

4.6 Outlook

In this work we introduced covariate shift detection benchmarks to study a complementary measure of model robustness. Our experiments reveal that for the CSD task, sophisticated OOD detection methods are worse than even the simple MSP baseline. We present a simple interpolation-based detection technique that surpasses all baselines on multiple CSD benchmarks on both image classification and text classification tasks. The results are promising and suggest that interpolation between training and test inputs can be a powerful tool for understanding and interpreting classification decisions as well as detecting outliers and covariate shift. We believe that test-time interpolation could also be useful for uncertainty quantification – recent results (Thiagarajan *et al.*, 2022) show how anchoring (a variant of interpolation) can be used during training for this purpose. In the future, we expect theoretical insights to emerge to complement our empirical findings with DIS.

Chapter 5

COMPARING THE EFFECTS OF DATA MODIFICATION METHODS ON OUT-OF-DOMAIN GENERALIZATION AND ADVERSARIAL ROBUSTNESS

Data modification, either via additional training datasets, data augmentation, debiasing, and dataset filtering, has been proposed as an effective solution for generalizing to out-of-domain (OOD) inputs, in both natural language processing and computer vision literature. However, the effect of data modification on adversarial robustness remains unclear. In this work, we conduct a comprehensive study of common data modification strategies and evaluate not only their in-domain and OOD performance, but also their adversarial robustness (AR). We also present results on a two-dimensional synthetic dataset to visualize the effect of each method on the training distribution. This work serves as an empirical study towards understanding the relationship between generalizing to unseen domains and defending against adversarial perturbations. Our findings suggest that more data (either via additional datasets or data augmentation) benefits both OOD accuracy and AR. However, data filtering (previously shown to improve OOD accuracy on natural language inference) hurts OOD accuracy on other tasks such as question answering and image classification. We provide insights from our experiments to inform future work in this direction.

This was a joint work with equal contribution from Swaroop Mishra and Man Luo, published in ACL Findings 2022 ([Gokhale *et al.*, 2022c](#)). In this chapter, I have discussed the results in computer vision tasks. For details on similar findings in NLP tasks, please refer to Man Luo’s PhD thesis.

5.1 Introduction

Deep neural networks have emerged as a widely popular architectural choice for modeling tasks in multiple domains such as (but not limited to) computer vision (Yuille and Liu, 2021), natural language processing (Hochreiter and Schmidhuber, 1997; Vaswani *et al.*, 2017), and audio (Hannun *et al.*, 2014). While these models are highly capable of learning from training data, recent studies show that they are quite prone to failure on new test sets or under distribution shift (Taori *et al.*, 2020), natural corruptions (Hendrycks and Dietterich, 2019), adversarial attacks (Goodfellow *et al.*, 2015), spurious correlations (Beery *et al.*, 2018), and many other types of “unseen” changes that may be encountered after training. This shortcoming stems from the *i.i.d.* assumption in statistical machine learning which guarantees good performance only on test samples that are drawn from an underlying distribution that is identical to the training dataset. For instance, digit recognition models trained on the black-and-white MNIST training images are almost perfect ($> 99\%$ accuracy) on the corresponding test set, yet their performance on colored digits and real-world digits from street number plates is less than 75%. Similarly, state-of-the-art NLP models have been shown to fail when negation is introduced in the input (Kassner and Schütze, 2020a). These findings pose a significant challenge to the practical adoption of these models and their reliability in the real-world.

To test model performance beyond the traditional notion of in-domain (ID) generalization, two prominent ideas have emerged: out-of-domain (OOD generalization) *a.k.a.* domain generalization¹, and adversarial robustness. The OOD generalization objective expects a model which is trained on distribution \mathcal{D} to perform reliably on

¹In this paper we use these two terms interchangeably.

unseen distributions $\mathcal{D}_e, e \in \{1, \dots, n\}$, that differ from \mathcal{D} . For a trained classifier f^* , OOD accuracy on previously unseen distribution \mathcal{D}_e is defined as:

$$\text{acc}_{\text{OOD}}^e = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}_e} [\mathbb{I}(f^*(\mathbf{x}) = \mathbf{y})] \quad (5.1)$$

To define adversarial robustness, consider an input \mathbf{x} and a true label \mathbf{y} . For a classifier loss function ℓ , a loss-maximizing perturbation δ^* within Δ_ϵ (an ϵ -bounded neighborhood of \mathbf{x}) is defined as:

$$\delta_{\mathbf{x}}^* = \max_{\delta \in \Delta_\epsilon} \ell(f^*(\mathbf{x} + \delta), \mathbf{y}). \quad (5.2)$$

The second idea is that of adversarial robustness. Recent work on adversarial examples has revealed the vulnerability of deep neural networks against small perturbations of the original data. Adversarial robustness in such under this setting is defined as the accuracy of the classifier on adversarial samples $\mathbf{x} + \delta_{\mathbf{x}}$, where the perturbation lies within an ℓ_p norm bound: $\|\delta_{\mathbf{x}}\|_p < \epsilon$.

$$\text{acc}_{\text{rob}} = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}} \mathbb{I}(f^*(\mathbf{x} + \delta_{\mathbf{x}}) = \mathbf{y}). \quad (5.3)$$

In the context of text classification, the norm-bound can also be in the form of small character-level or word-level perturbations such as swapping, inserting, or deleting characters or words. In essence, adversarial robustness measures the invariance of the classifier to small perturbations of the input.

Various methods have been developed that either improve OOD generalization or improve adversarial robustness. Notable among these are techniques that modify the distribution of the training dataset. In this paper, we focus on three major data modification techniques – the use of additional datasets (or multi-source training), data augmentation, and data filtering; in addition we also consider model-based debiasing techniques which do not alter the data distribution explicitly. We study

the performance of these methods on three representative tasks – natural language inference (NLI), extractive question answering (QA), and image classification (IC).

Our first aim in this paper is to understand whether the increase or decrease in OOD generalization by each method over the naive baseline (standard training on the source dataset) is consistent across tasks. To further conduct fine-grained analysis, we also analyze the effect of these methods on in-domain (ID) accuracy on the test set for each task, since in the ideal case improvement in OOD performance should not come at the cost of in-domain accuracy.

Recent work seeks to understand the relationships between in-domain and out-of-domain performance: for instance, [Miller et al. \(2021\)](#) empirically show that ID and OOD performance are strongly correlated, [Raghunathan et al. \(2020\)](#); [Yang et al. \(2020\)](#) show a trade-off between robustness and accuracy for adversarially trained models. However it is not clear how methods *designed for OOD generalization* affect robustness. This is largely because work on domain generalization reports only IID and OOD metrics, and work on robustness reports only ID and robustness metrics. Our second aim is to understand the effect of these generalization methods on adversarial robustness.

In addition to our experiments on NLP and vision tasks, we also provide an experiment on a synthetic binary classification dataset where points lie in a 2-dimensional feature space and are separated by concentric circles into class labels. This setting allows us to visualize the effect of data modification techniques on the training distribution and the resulting performance.

Our findings can be summarized as follows:

- More data benefits OOD generalization,
- Data filtering hurts OOD generalization, and
- Data filtering significantly hurts adversarial robustness on all benchmarks.

These findings and our additional analysis raise new questions for robustness and domain generalization research. Significant among these are the importance of both diversity and number of training samples for inductive bias and generalization guarantees, the problems associated with data filtering in terms of robustness, and the importance of a comprehensive set of evaluation metrics that could be adopted for future work.

5.2 Categorization of Domain Generalization Methods

In this section, we provide a categorization of methods that are typically used as baselines for domain generalization. We briefly explain the method and provide relevant related work in which these ideas are used as methods for domain generalization. Throughout this paper, we will refer to the original training distribution as the “*source*” and the out-of-distribution datasets as the “*targets*”.

Single-Source Training (SS) refers to the “vanilla” baseline which is trained only on the source dataset, without any dataset modification. SS utilizes no other information apart from the single source dataset \mathcal{D} and updates parameters θ of classifier f to minimize the risk on the source using approaches such as ERM ([Vapnik and Chervonenkis, 1991](#)).

$$\underset{\theta}{\text{minimize}} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}} \ell(f(\mathbf{x}; \theta), \mathbf{y}). \quad (5.4)$$

Multi-Source Training (MS). This method is identical to SS except that additional training datasets \mathcal{D}' are used for risk minimization.

$$\underset{\theta}{\text{minimize}} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D} \cup \mathcal{D}'} \ell(f(\mathbf{x}; \theta), \mathbf{y}). \quad (5.5)$$

Usually \mathcal{D}' are designed for the same task as \mathcal{D} but may have different styles, characteristics, or sources of collection. For instance, while both SNLI ([Bowman *et al.*, 2016](#),

2015a) and MNLI (Williams *et al.*, 2018) are datasets for natural language inference with identical class labels, SNLI was collected from image captions, while MNLI was collected from Open American National Corpus².

Gulrajani and Lopez-Paz (2021) provide an extensive comparative study of models trained for multi-source domain generalization for image classification and surprisingly find that if multiple source domains are available, ERM is empirically the best approach as compared to specially designed DG methods such as meta-learning (Li *et al.*, 2018a), learning domain-invariant features (Ganin *et al.*, 2016), invariant risk minimization (Arjovsky *et al.*, 2019), etc. These findings have also been observed on text classification experiments in (Koh *et al.*, 2021). Hendrycks *et al.* (2020a) show that pre-training transformer architectures on diverse data leads to higher OOD accuracies on multiple tasks such as semantic textual similarity, sentiment classification, reading comprehension and natural language inference.

Data Augmentation (DA). When additional training distributions are not directly available, transformations of samples in \mathcal{D} using pre-defined augmentation functions can be used to create \mathcal{D}' and train the model. Such data augmentation functions are typically derived from existing knowledge about the invariance of the task w.r.t. certain transformations. For instance, for image classification, addition of small noise, small translations, scaling, etc. are common data augmentation functions, since they do not change the true label for the image. Similarly, for text inputs, synonyms of words are commonly used since they do not change the semantics of the sentence. NLP data augmentation techniques include UDA (Xie *et al.*, 2020), EDA (Wei and Zou, 2019), and back-translation for question answering (Longpre *et al.*, 2019).

²<https://www.anc.org/>

Data Filtering (DF). Dataset filtering has been previously explored for quality control, such as, removing noise and artifacts to curate and improve publicly sourced datasets. However, there has been recent interest in considering DF as a method for bias reduction and generalization. This idea can be traced back to work by [Zellers et al. \(2018, 2019\)](#), that proposed DF as an algorithmic method to avoid annotation artifacts and spurious correlations during dataset construction. AFLite ([Bras et al., 2020](#)) extended this idea to a generic filtering methodology that can work without any pre-defined rules or strategies. Instead, AFLite operates by utilizing several weak learners (such as support-vector machines) trained over small subsets to identify samples that are easy to classify. It is argued that such samples are more likely to carry biases, and as such, could be removed. AFLite suggests that reduction of a dataset to even 10% of the original size can boost OOD accuracy on NLI. In the vision domain, similar ideas have been proposed concurrently, including REPAIR ([Li and Vasconcelos, 2019](#)) and RESOUND ([Li et al., 2018b](#)), in which instead of completely removing samples, biased samples are assigned smaller weights. However these methods require a prior knowledge of the bias variable. [Liu et al. \(2021\)](#) have recently proposed a simple approach which upweights samples which have higher loss – this is shown to improve worst-group accuracy without having access to the bias variable.

Model De-biasing (DB). Methods under this category do not directly alter the training dataset, but instead resort to changes in the modeling technique – these changes can be in terms of the optimization function, regularization, additional auxiliary costs, etc. The main idea in DB is to utilize known biases (or identify unknown biases) in the data distribution, model these biases in the training pipeline, and use this knowledge to train robust classifiers ([Clark et al., 2019; Wu et al., 2020](#);

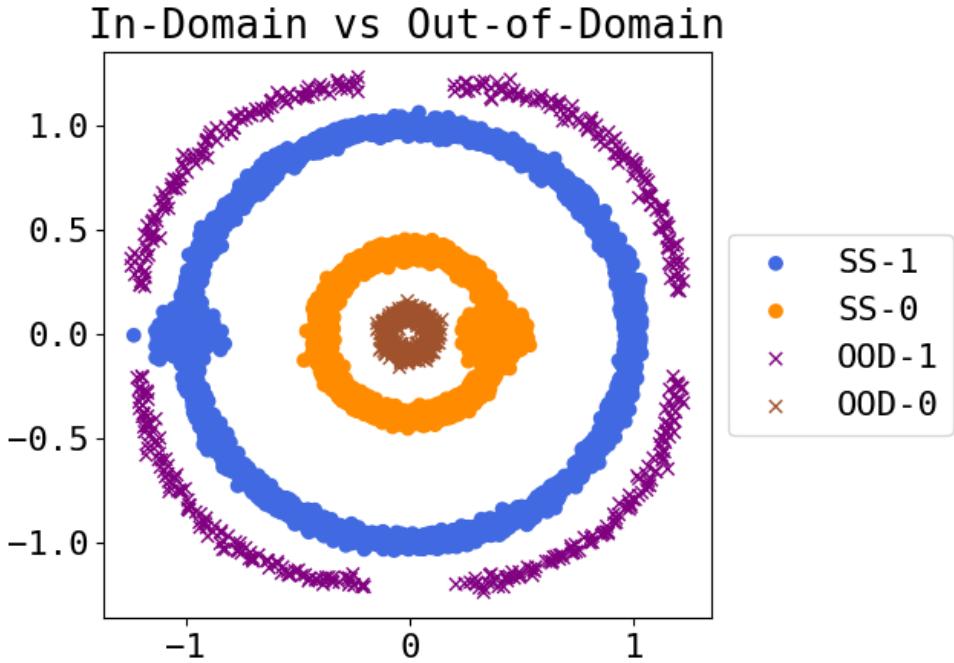


Figure 5.1: Our toy experimental setting consists of points in \mathbb{R}^2 belonging to two classes (0/1). This illustration shows the discrepancy between the source dataset (SS) and the out-of-domain dataset (OOD).

Bhargava *et al.*, 2021). In the image classification literature, there is growing consensus on enforcing a consistency on different views (or augmentations) of an image in order to achieve debiasing (Hendrycks *et al.*, 2020c; Xu *et al.*, 2020b; Chai *et al.*, 2021; Nam *et al.*, 2021). Unlike DF, model de-biasing does not directly alter the training distribution, but instead allows the model to learn which biases to ignore.

5.3 Toy Example: Concentric Circles

We begin with a simple two-dimensional example to illustrate our experimental setting and to show how each method affects the distribution of the training set. Consider the set of points shown in Figure 5.1 where the points belong to two class

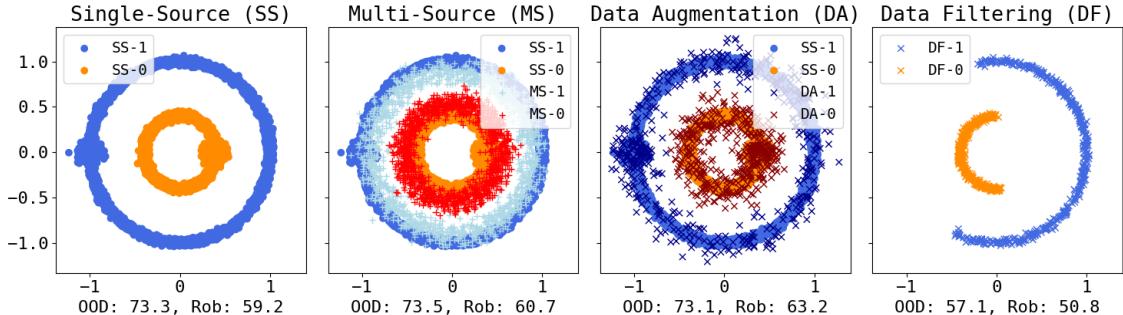


Figure 5.2: This figure illustrates the effect of data modification techniques on the training distribution. The leftmost figure shows the training distribution in the single-source setting. The introduction of a second dataset or Data-augmentation (done using small perturbations of source samples with Gaussian noise) makes the distribution more diverse in the multi-source (MS) and data augmentation (DA) setting respectively. On the other hand, data filtering, in order to remove spurious correlations from the dataset, removes points from certain sectors of the distribution. The effect of each strategy on OOD generalization and robustness is shown below each plot.

labels (either 0 or 1) and are seen to lie on concentric circles. Points with label 0 are closer to the origin, while points with label 1 are closer to a distance of 1 from the origin. Our aim is to start with the single source dataset and train the model to generalize on the out-of-domain (OOD) dataset. An important thing to note here is that the source dataset contains a subset of points with label 0 (orange) clustered around $(0.4, 0.0)$ and a subset with label 1 clustered around $(-1, 0.0)$. This implies that class-0 is biased towards $x > 0$, while class-1 is biased towards $x < 0$. In total, our SS dataset consists of 10000 samples, of which 20% are biased.

We apply three data modifications: additional source (MS), gaussian data augmentation (DA) $\sim \mathcal{N}(0, 0.1)$, and data filtering (AFLite) which reduces the dataset size to 10%. Note that we do not show model debiasing (DB) here, since it does not alter the data distribution. Figure 5.2 shows the effect on the data distribution. The most

Method Category	Tasks		
	Natural Language Inference	Question Answering	Image Classification
SS (Single-Source ERM)	SNLI	NQ (Kwiatkowski et al., 2019)	MNIST
MS (Multi-Source ERM)	SNLI + MNLI	NQ + SQuAD+NQA+HQA+SQA+TQA	MNIST + USPS
DA (Data Augmentation)	EDA (Wei and Zou, 2019)	QG (Chan and Fan, 2019)	M-ADA (Qiao et al., 2020)
DB (Model De-biasing)	LMH (Clark et al., 2019)	Mb-CR(Wu et al., 2020)	RandConv (Xu et al., 2020b)
DF (Data Filtering)	AFLite (Bras et al., 2020)	AFLite (adapted for QA)	AFLite

Table 5.1: List of method categories and specific methods that we use under each task setting in our experiments. Details for each can be found in Section 5.4 for the corresponding task.

striking is the effect of DF which removes all samples previously in the biased clusters near $(0.4, 0.0)$ and $(-1.0, 0.0)$.

Equipped with these resulting datasets, we train a linear SGD classifier with log-loss and evaluate the robustness of each model in terms of in-domain and OOD accuracies. We also evaluate adversarial robustness by using standard PGD attacks. Results are shown in the textboxes in Figure 5.2. It can be seen that data filtering significantly hurts both OOD generalization and robustness. This finding motivates our experiments to understand the effect of each method for NLP and vision tasks.

5.4 Experiments

In [Gokhale et al. \(2022c\)](#), we present three tasks and their corresponding experimental setup, evaluation protocol and our findings. A summary of methods belong to each category is provided in Table 5.1 and the abbreviations SS, MS, DA, DB, DF are used henceforth. In this section, I will discuss the results of the image classification experiments.

Experimental Setting. We conduct our experiments on the standard domain generalization benchmark “Digits”, which is a collection of handwritten digit classification

Method	In-Domain Acc. (%)	OOD Acc. (%)			
		MNIST-M	SVHN	SYNTH	Avg
SS	98.40	58.09	33.85	45.94	45.96
MS	98.54	59.79	33.87	48.42	47.36
DA	99.30	67.94	42.55	48.95	53.15
DB	98.86	87.67	54.95	63.37	68.66
DF	95.27	51.04	22.07	27.83	33.65

Table 5.2: Source (in-domain) accuracy and domain generalization (OOD accuracy) on the Digits benchmark with MNIST-10k as source dataset.

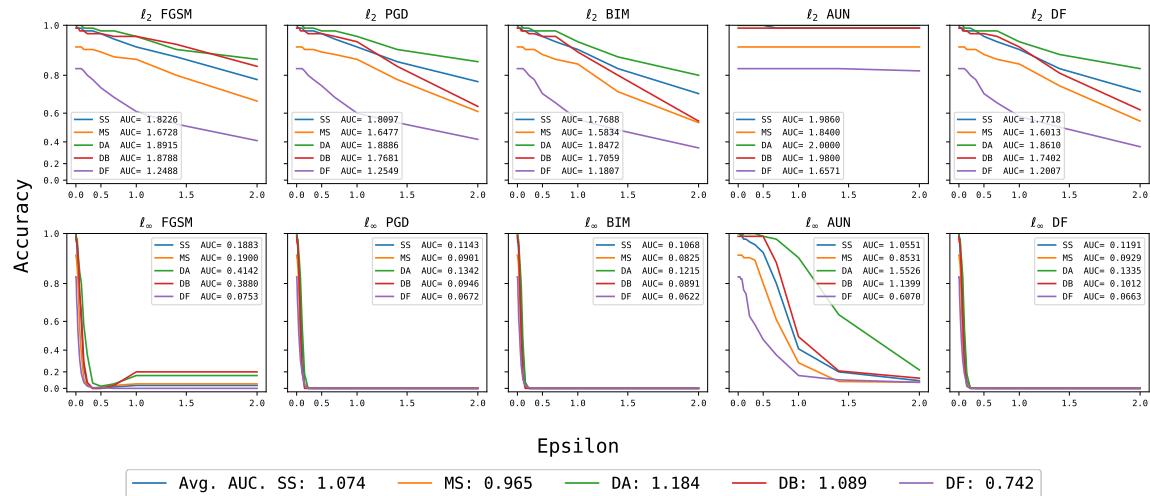


Figure 5.3: Evaluation of adversarial robustness (using 10 attack methods) for MNIST10k.

datasets belonging to 10 classes (digits 0–9). Following standard practice (Volpi *et al.*, 2018), we train models on 10000 images from MNIST (LeCun *et al.*, 1998) as the source, and use SVHN (Netzer *et al.*, 2011), SYN and MNIST-M (Ganin and Lempitsky, 2015) as the OOD datasets.

Methods. We use DigitNet (Volpi *et al.*, 2018) as our backbone image classifier architecture. Our **SS** baseline uses MNIST for training; **MS** uses MNIST and USPS (Denker *et al.*, 1988). For data augmentation we rely on M-ADA (Qiao *et al.*, 2020) which is a perturbation-based min-max algorithm to create augmented data. Our debiasing method is RandConv (Xu *et al.*, 2020b) which utilizes a random convolutional layer to generate novel views of each input image, and a KL-divergence based loss function that encourages the classifier to predict consistent predictions for each version of the image. This leads to the model being debiased on spurious features like background, texture, or color of digits. We use AFLite as our **DF** method.

Evaluation Protocol. We report IID accuracy on the MNIST test set and generalization as the accuracy on our OOD datasets. For evaluating adversarial robustness we use Foolbox (Rauber *et al.*, 2017) and use 10 attack methods (both ℓ_2 and ℓ_∞ versions of FGSM, PGD, BIM, AUN, and DeepFool). Robustness is calculated as the accuracy for 20 values of ϵ between [0, 2], and is plotted as robustness curves for visualization, along with the average values for area under the curve (AUC).

Results. Table 5.2 shows the performance of each method in terms of in-domain and OOD accuracy. **MS**, **DA** and **DB**, improve the generalization performance on each OOD dataset and also improve the in-domain performance, where **DB** displays best generalization capacity. **DF** dramatically reduces the OOD performance with significant reduction across all datasets; the in-domain accuracy also decreases. Figure 5.3 shows

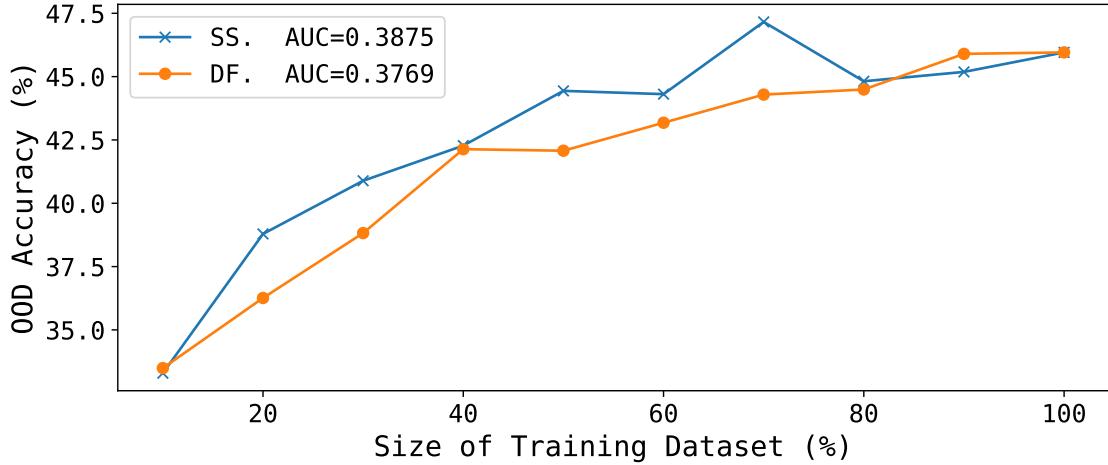


Figure 5.4: Comparison between SS and DF models trained with different percentages of MNIST10k.

robustness (accuracy) and area under the curve (AUC) for each plot. It can be observed that DF is worse than SS for all 10 attack variants. We observe that DA and DB are better than SS, and the drop for DF is the largest.

5.5 Analysis

Based on the results of three tasks, we have the following observations about the performance of each method compared to the SS baseline:

- MS increases OOD accuracy on all three tasks and robustness on two tasks (NLI and QA).
- DA increases OOD on two tasks (NLI and IC) and robustness on all three tasks.
- DB increases OOD on three tasks and robustness on two tasks (NLI and QA).
- DF decreases OOD on two tasks (QA and IC) and robustness on all three tasks.

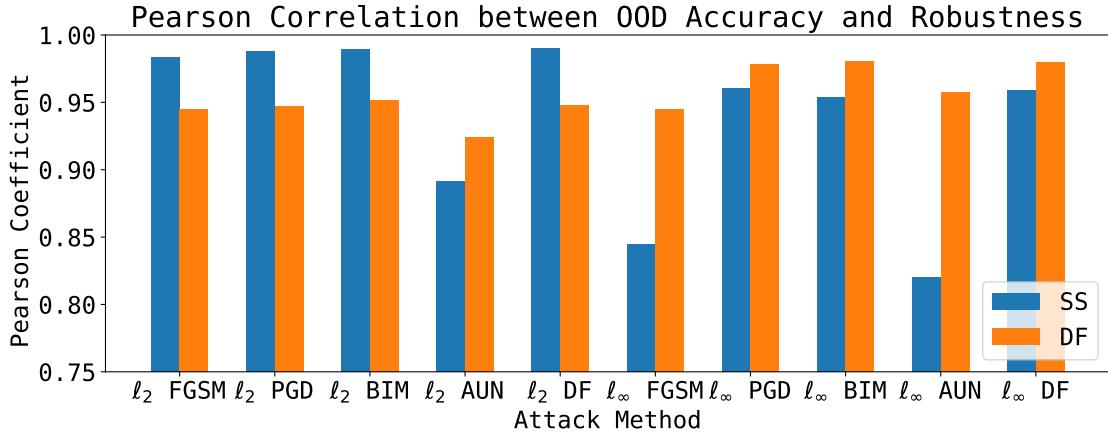


Figure 5.5: Pearson Correlation between OOD accuracy and robustness for SS and DF models on MNIST10k.

Decrease in NLI in-domain accuracy is seen for all methods, even though these lead to increase in OOD accuracy. This suggests that the training dataset (SNLI) has a large shift w.r.t. OOD datasets.

More data implies more OOD generalization: While this trend is observed for both MS and DA, there is one anomaly – DA for the QA task leads to marginal decrease compared to SS (a difference of 0.56%). This finding is aligned with [Longpre et al. \(2019\)](#), who report no significant effect of data augmentation (back translation) on OOD performance for question answering. This points to the need for improving data augmentation techniques in QA. On the other hand, the performance drop due to DF is significantly large for QA (11.53%).

Decrease in MNIST robustness: For MNIST, the DA method (M-ADA ([Qiao et al., 2020](#))) is the best in terms of robustness and also improves OOD accuracy. M-ADA is an “adversarial data augmentation” method, i.e., it uses a min-max objective to find loss-maximizing perturbations and uses these perturbations as augmented data.

It is therefore intuitive that such a method would do well on the adversarial robustness metric (although robustness evaluation was not reported by [Qiao et al. \(2020\)](#)).

Marginal Improvement on Robustness: From the results, it is easy to see that the improvement on OOD is more noticeable than robustness, for example, **MS** improves OOD performance by $\sim 10\%$, but improves only by $\sim 1\%$ under model-free evaluation. While this observation is reasonable since each method is designed to improve the generalization, new methods that improve both generalization and robustness should be encouraged.

5.5.1 Correlation between Adversarial Robustness and OOD Generalization

Our experiments reveal the alarming finding that across the board, **DF** reduces adversarial robustness. To investigate further, we conduct an analysis on the Digits benchmark and compare **SS** and **DF** when trained with equal amounts of data ($\{10\%, 20\%, \dots, 100\%\}$). Note that for **SS** the data are sampled randomly, while for **DF** the data are obtained via AFLite data filtering. Results are shown in Figure 5.4. It can be observed that the OOD accuracy increases as the size of the dataset increases, and is greater for **SS** than **DF**. To understand how an increase in OOD accuracy affects robustness, we also compute the robustness values at each size of training data, and compute the Pearson correlation coefficient for each attack method – positive correlation implies that as OOD accuracy increases, robustness also increases. Figure 5.5 shows clear evidence in favor of positive correlation; interestingly, **SS** has higher correlation for ℓ_2 attacks, while **DF** is higher for ℓ_∞ attacks. The evidence is clear: OOD generalization increases with the size of the dataset and adversarial robustness is positively correlated with OOD generalization.

Our experiments show that the size of the training set directly affects both

robustness and generalization. While removing 90% data increased OOD accuracy in NLI, the effect was the exact opposite for QA and MNIST. The key idea in domain generalization is that the test distributions are unknown and little information about them is available apart from the fact that there is no task shift. Without this prior knowledge, deciding whether (or how much) to filter a dataset is a challenging task.

5.6 Related Work

In Section 5.2 we have provided relevant work that falls into one of our five modeling categories. Here, we discuss additional literature on robustness and generalization and new efforts towards dataset creation, benchmarks, and evaluation.

Generalization Benchmarks. [Hendrycks *et al.* \(2020b\)](#) have constructed a robustness benchmark for multiple language understanding tasks by splitting training sets from existing benchmarks according to topics, styles, and vocabulary; this has been subsequently used to study robustness of model rankings ([Mishra and Arunkumar, 2021](#)). Benchmarks have also been constructed to study dataset artifacts and generalization capabilities of models ([Mishra *et al.*, 2020a,b](#); [Mishra and Sachdeva, 2020](#)). MRQA ([Fisch *et al.*, 2019](#)) is a benchmark for evaluating domain generalization of question answering (reading comprehensive) models. MRQA contains 6 datasets each for training, development, and evaluation. For image classification, many benchmarks have been proposed to evaluate domain generalization, such as PACS ([Li *et al.*, 2017](#)), OfficeHome ([Venkateswara *et al.*, 2017](#)), Digits ([Volpi *et al.*, 2018](#)), and WILDS ([Koh *et al.*, 2021](#)) which is a compendium of domain generalization benchmarks for tasks such as image classification, text sentiment and toxicity prediction.

Corruption Robustness. Hendrycks and Dietterich (2019) introduced ImageNet-C and CIFAR-C to test robustness along corruptions such as weather, noise, blur, and digital artifacts, and ImageNet-P which tests robustness against small tilts and changes in brightness. MNIST-C was introduced by Mu and Gilmer (2019) for similar corruptions of handwritten digit images.

Adversarial and Contrastive Sets. Generation of adversarial examples (Jia and Liang, 2017b; Ribeiro *et al.*, 2018b; Iyyer *et al.*, 2018b; Alzantot *et al.*, 2018b) and approaches to defend against word substitution (Jia *et al.*, 2019b) have been explored. Contrastive examples have been introduced as a means for evaluation, for example, manually crafted contrast sets for textual entailment (Gardner *et al.*, 2020b) or template-based (McCoy *et al.*, 2019b; Glockner *et al.*, 2018b; Naik *et al.*, 2018b). Model-in-the-loop dataset creation methods have also been proposed for various NLP tasks (Nie *et al.*, 2020; Arunkumar *et al.*, 2020; Kiela *et al.*, 2021) and visual question answering (Sheng *et al.*, 2021; Li *et al.*, 2021b).

5.7 Discussion

Recently, Miller *et al.* (2021) have empirically shown linear trends between in-distribution and out-of-distribution performance on multiple image classification tasks, across various model architectures, hyper-parameters, training set size, and duration of training. They also show that there are certain settings of domain shift under which the linear trend does not hold. Our work empirically shows that while data filtering may benefit OOD generalization on the NLI benchmark, this does not hold for other tasks such as image classification and question answering. This suggests that data filtering may benefit generalization in certain types of domain shift, but not on others. Concurrently, Yi *et al.* (2021) have theoretically shown that models robust to input

perturbations generalize well on OOD distribution within a Wasserstein radius around the training distribution. Our empirical observations in this paper in both vision and language domains, agree with the theory of [Yi et al. \(2021\)](#).

In this work, we conduct a comprehensive study of methods which are designed for OOD generalization on three tasks: NLI, QA, and IC. We evaluate each method on in-domain, OOD, and adversarial robustness. Our findings suggest that more data typically benefits both OOD and robustness. Data filtering hurts OOD accuracy on two out of three tasks, and also hurts robustness on all three tasks. In context of our findings and work by [Miller et al. \(2021\)](#); [Yi et al. \(2021\)](#), we recommend that methods designed either for robustness or generalization should be evaluated on multiple aspects and not on the single metric that they are optimized for.

5.8 Broader Impact

One underlying assumption behind using large datasets for training (or pre-training) vision and language models is that larger datasets increase the likelihood of obtaining a diverse set of samples to reduce overfitting. However, recent studies ([Bender et al., 2021](#); [Stanovsky et al., 2019](#)) serve as cautionary tales when employing uncurated internet data to train large language models, and discuss how large data does not necessarily imply that models will learn the dievrse distribution. At the same time, the inverse (small data aids diversity) is also not true (as shown by this paper) and comes with its own problems – for instance, Figure 5.2 shows that dataset filtering can lead to much larger changes in the data distribution beyond notions of proportionality and fairness. As such, the decision on how many and what samples to remove can also introduce its own set of biases. Data curation is a challenging problem and needs further task-specific study since the concepts of bias and fairness often depend on the task definition and specifications of ideal outcomes. Insights from this paper could

help researchers and practitioners in choosing appropriate approaches for improving generalization and robustness.

Chapter 6

VISION, LANGUAGE, AND LOGIC: ROBUSTNESS TO LOGICAL COMPOSITIONS

Logical connectives and their implications on the meaning of a natural language sentence are a fundamental aspect of understanding. In this work, we investigate whether visual question answering (VQA) systems trained to answer a question about an image, are able to answer the logical composition of multiple such questions. When put under this *Lens of Logic*, state-of-the-art VQA models have difficulty in correctly answering these logically composed questions. We construct an augmentation of the VQA dataset as a benchmark, with questions containing logical compositions and linguistic transformations (negation, disjunction, conjunction, and antonyms). We propose our Lens of Logic (LOL) model which uses question-attention and logic-attention to understand logical connectives in the question, and a novel Fréchet-Compatibility Loss, which ensures that the answers of the component questions and the composed question are consistent with the inferred logical operation. Our model shows substantial improvement in learning logical compositions while retaining performance on VQA. We suggest this work as a move towards robustness by embedding logical connectives in visual understanding.

6.1 Introduction

Theories about logic in human understanding have a long history. In modern times, Piaget and Fodor ([Piattelli-Palmarini, 1980](#)) studied the representation of logical hypotheses in the human mind. George Boole ([Boole, 1854](#)) formalized conjunction,

Image	Question	Predicted Answer	Accuracy (%)
	VQA Q_1 : Is there beer? YES (0.96) Q_2 : Is the man wearing shoes? NO (0.90)		SOTA 88.20 LOL 86.55
	VQA-Compose $\neg Q_2$: Is the man <i>not</i> wearing shoes? NO (0.80) $\neg Q_2 \wedge Q_1$: Is the man <i>not</i> wearing shoes <i>and</i> is there beer? NO (0.62) $Q_1 \wedge C$: Is there beer and does this seem like a man bending over to look inside of a fridge? NO (1.00)		50.69 82.39  
	VQA-Supplement $\neg Q_2 \vee B$: Is the man not wearing shoes or is there a clock? NO (1.00) $Q_1 \wedge \text{ant}(B)$: Is there beer and is there a wine glass? YES (0.84)		50.61 87.80  

Figure 6.1: State-of-the-art models answer questions from the VQA dataset (Q_1, Q_2) correctly, but struggle when asked a logical composition including negation, conjunction, disjunction, and antonyms. We develop a model that improves on this metric substantially, while retaining VQA performance.

disjunction, and negation into an “algebra of thought” as a way to improve, systemize, and mathematize Aristotle’s Logic (Corcoran, 1972). Horn regarded negation to be a fundamental and defining characteristic of human communication (Horn and Kato, 2000), following the traditions of Sankara (Raju, 1954), Spinoza (Spinoza, 1934), and Hegel (Hegel, 1929). Recent studies (Cesana-Arlotti *et al.*, 2018) have suggested that infants can formulate intuitive and stable logical structures to interpret dynamic scenes and to entertain and rationally modify hypotheses about the scenes. As such we argue that understanding logical structures in questions, is a fundamental requirement for any question-answering system.

If a question can be put at all, then it can be answered.

Wittgenstein (1921)

In the above proposition, Wittgenstein linked the process of asking a question with the existence of an answer. While we do not comment on the existence of an answer, we suggest the following softer proposition -

If questions $Q_1 \dots Q_n$ can be answered, then so should all composite questions created from $Q_1 \dots Q_n$

Visual question answering (VQA) ([Antol et al., 2015](#)) is an intuitive, yet challenging task that lies at a crucial intersection of vision and language. Given an image and a question about it, the goal of a VQA system is to provide a free-form or open-ended answer. Consider the image in Figure 6.1 which shows a person in front of an open fridge. When asked the questions Q_1 (*Is there beer?*) and Q_2 (*Is the man wearing shoes?*) independently, the state-of-the-art model LXMERT ([Tan and Bansal, 2019](#)) answers both correctly. However when we insert a negation in Q_2 (*Is the man not wearing shoes?*) or for a conjunction of two questions $\neg Q_2 \wedge Q_1$ (*Is the man not wearing shoes and is there beer?*), the system makes wrong predictions. Our motivation is to reliably answer such logically composed questions. In this paper, we analyze VQA systems under this *Lens of Logic (LOL)* and develop a model that can answer such questions reflecting human logical inference. We offer our work as the first investigation into the logical structure of questions in visual question-answering and provide a solution that *learns* to interpret logical connectives in questions.

The first question is: can models pre-trained on the VQA dataset answer logically composed questions? It turns out that these models are unable to do so, as illustrated in Figure 6.1 and Table 6.2. An obvious next experiment is to *split the question* into its component questions, predict the answer to each, and combine the answers logically. However language parsers (either oracle or trained parsers) are not accurate at understanding negation, and as such this approach does not yield correct answers for logically composed questions. The question then arises: can the model answer such questions, if we explicitly train it with data that also contains logically composed questions? For this investigation, we construct two datasets, **VQA-Compose** and **VQA-Supplement**, by utilizing annotations from the VQA dataset, as well as object

and caption annotations from COCO (Lin *et al.*, 2014). We use these datasets to train the state-of-the-art model LXMERT (Tan and Bansal, 2019) and perform multiple experiments to test for robustness towards logically composed questions.

After this investigation, we develop our LOL model architecture that jointly learns to answer questions while understanding the type of question and which logical connective exists in the question, through our attention modules, as shown in Figure 6.3. We further train our model with a novel Fréchet-Compatibility loss that ensures compatibility between the answers to the component questions and the answer of the logically composed question. One key finding is that our models are better than existing models trained on logical questions, with a small deviation from state-of-the-art on VQA test set. Our models also exhibit better *Compositional Generalization* i.e. models trained to answer questions with a single logical connective are able to answer those with multiple connectives.

Our contributions are summarized below:

1. We conduct a detailed analysis of the performance of the state-of-the-art VQA model with respect to logically composed questions,
2. We curate two large scale datasets VQA-Compose and VQA-Supplement that contain logically composed binary questions.
3. We propose *LOL* – our end-to-end model with dedicated attention modules that answer questions by understanding the logical connectives in questions.
4. We show a capability of answering logically composed questions, while retaining VQA performance.

6.2 Related Work

Logic in Human Expression: Is logical thinking a natural feature of human thought and expression? Evidence in psychological studies (Carey, 1985; Gopnik *et al.*, 1999; Cesana-Arlotti *et al.*, 2018) suggests that infants are capable of logical reasoning, toddlers understand logical operations in natural language and are able to compositionally compute meanings even in complex sentences containing multiple logical operators. Children are also able to use these meanings to assign truth values to complex experimental tasks. Given this, question-answering systems also need to answer compositional questions, and be robust to the manifestation of logical operators in natural language.

Logic in Natural Language Understanding: The task of understanding compositionality in question-answering (QA) can also be interpreted as understanding logical connectives in text. While question compositionality is largely unstudied, approaches in natural language understanding seek to transform sentences into symbolic formats such as first-order logic (FOL) or relational tables (Mintz *et al.*, 2009; Zettlemoyer and Collins, 2005; Lewis and Steedman, 2013). While such methods benefit from interpretability, they suffer from practical limitations like intractability, reliance on background knowledge, and failure to process noise and uncertainty. (Bordes *et al.*, 2013; Riedel *et al.*, 2013; Socher *et al.*, 2013) suggest that better generalization can be achieved by learning embeddings to reason about semantic relations, and to simulate FOL behavior (Rocktäschel *et al.*, 2014). Recursive neural networks have been shown to learn logical semantics on synthetic English-like sentences by using embeddings (Bowman *et al.*, 2015b; Neelakantan *et al.*, 2015).

Detection of negation in text has been studied for information extraction and

sentiment analysis (Morante and Sporleder, 2012). (Kassner and Schütze, 2020b) have shown that BERT-based models (Devlin *et al.*, 2019; Liu *et al.*, 2019c) are incapable of differentiating between sentences and their negations. Concurrent to our work, (Asai and Hajishirzi, 2020) show the efficacy of FOL-guided data augmentation for performance improvements on natural language QA tasks that require reasoning. Since our work deals with both vision and language modalities, it encounters a greater degree of ambiguity, thus calling for robust VQA systems that can deal with logical transformations.

Visual Question Answering (VQA) (Antol *et al.*, 2015) is a large-scale, human-annotated dataset for open-ended question-answering on images. VQA-v2(Goyal *et al.*, 2017) reduces the language bias in the dataset by collecting complementary images for each question-image pair. This ensures that the number of questions in the VQA dataset with the answer “YES” is equal to those with the answer “NO”. This dataset contains 204k images from MS-COCO (Lin *et al.*, 2014), and 1.1M questions.

Cross-modal pre-trained models (Tan and Bansal, 2019; Lu *et al.*, 2019; Zhou *et al.*, 2020b) have proved to be highly effective in vision-and-language tasks such as VQA, referring expression comprehension, and image retrieval. While neuro-symbolic approaches (Mao *et al.*, 2019) have been proposed for VQA tasks which require reasoning on synthetic images, their performance on natural images is lacking. Recent work seeks to incorporate reasoning in VQA, such as visual commonsense reasoning (Zellers *et al.*, 2019; Fang *et al.*, 2020), spatial reasoning (Hudson and Manning, 2019; Johnson *et al.*, 2017), and by integrating knowledge for end-to-end reasoning (Aditya *et al.*, 2019).

We take a step back and extensively analyze the pivotal task of VQA with respect

QF	Question	AF	Answer
Q_1	Is there beer?	A_1	Yes
Q_2	Is the man wearing shoes?	A_2	No
$\neg Q_1$	Is there no beer?	$\neg A_1$	No
$\neg Q_2$	Is the man not wearing shoes?	$\neg A_2$	Yes
$Q_1 \wedge Q_2$	Is there beer and is the man wearing shoes?	$A_1 \wedge A_2$	No
$Q_1 \vee Q_2$	Is there beer or is the man wearing shoes?	$A_1 \vee A_2$	Yes
$Q_1 \wedge \neg Q_2$	Is there beer and is the man not wearing shoes?	$A_1 \wedge \neg A_2$	Yes
$Q_1 \vee \neg Q_2$	Is there beer or is the man not wearing shoes?	$A_1 \vee \neg A_2$	Yes
$\neg Q_1 \wedge Q_2$	Is there no beer and is the man wearing shoes?	$\neg A_1 \wedge A_2$	No
$\neg Q_1 \vee Q_2$	Is there no beer or is the man wearing shoes?	$\neg A_1 \vee A_2$	No
$\neg Q_1 \wedge \neg Q_2$	Is there no beer and is the man not wearing shoes?	$\neg A_1 \wedge \neg A_2$	No
$\neg Q_1 \vee \neg Q_2$	Is there no beer or is the man not wearing shoes?	$\neg A_1 \vee \neg A_2$	Yes

Table 6.1: Illustration of question composition in VQA-Compose, for the same example as in Figure 6.1. QF: Question Formula, AF: Answer Formula

to various aspects of generalization. We consider a rigorous investigation of a task, dataset, and models to be equally important as proposing new challenges that are arguably harder. In this paper we analyse existing state-of-the-art VQA models with respect to their robustness to logical transformations of questions.

6.3 The Lens of Logic

A lens magnifies objects under investigation, by allowing us to zoom and focus on desired contents or processes. Our lens of logical composition of questions, allows us



Figure 6.2: Some questions in VQA-Supplement created with adversarial antonyms.

to magnify, identify, and analyze the problems in VQA models.

Consider Figure 6.2(a), where we transform the first question “*Is the lady holding the baby*” by first replacing “*lady*” with an adversarial antonym “*man*” and observe that the system provides a wrong answer with very high probability. Swapping “*man*” with “*baby*” results in a wrong answer as well. In 6.2(b) a conjunction of two questions containing antonyms (*girls* vs *boys*) yields a wrong answer. We identify that the ability to answer composite questions created by negation, conjunction and disjunction of questions is crucial for VQA.

We use “closed questions” as defined in (Bobrow, 1964) to construct logically composed questions. Under this definition, if a closed question has a negative (“NO”) answer then its negation must have an affirmative (“YES”) answer. Of the three types of questions in the VQA dataset (yes/no, numeric, other), ‘yes-no’ questions satisfy this requirement. Although, visual questions in the VQA dataset can have multiple correct answers (Bhattacharya *et al.*, 2019), 20.91% of the questions (around 160k) in the VQA dataset are closed questions, i.e. questions with a single unambiguous yes-or-no answer, unanimously annotated by multiple human workers. This allows us to treat these questions as propositions and create a truth table for answers to compose logical questions as shown in Table 6.1.

6.3.1 Composite Questions

Let \mathcal{D} be the VQA dataset. For closed questions Q_1 and Q_2 about image $I \in \mathcal{D}$, we define the composite question Q^* composed using connective $\circ \in \{\vee, \wedge\}$, as:

$$Q^* = \widehat{Q}_1 \circ \widehat{Q}_2, \quad \text{where } \widehat{Q}_1 \in \{Q_1, \neg Q_1\}, \quad \widehat{Q}_2 \in \{Q_2, \neg Q_2\}. \quad (6.1)$$

6.3.2 Dataset Creation Process

Using the above definition we create two new datasets by utilizing multiple questions about the same image (**VQA-Compose**) and external object and caption annotations about the image from COCO to create more questions (**VQA-Supplement**). The seed questions for creating these datasets are all closed binary questions from VQA-v2 (Goyal *et al.*, 2017). These datasets serve as test-beds, and enable experiments that analyze performance of models when answering such questions.

VQA-Compose: Consider the first two rows in Table 6.1. Q_1 and Q_2 are two questions about the image in Figure 6.1 taken from the VQA dataset. Additional questions are composed from Q_1 and Q_2 by using the formulas in Table 6.1. Thus for each pair of closed questions in the VQA dataset, we get 10 logically composed questions. Using the same train-val-test split as the VQA-v2 dataset (Goyal *et al.*, 2017), we get *1.25 million samples* for our **VQA-Compose** dataset. The dataset is balanced in terms of the number of questions with affirmative and negative answers.

VQA-Supplement: Images in VQA-v2 follow identical train-val-test splits as their source MS-COCO (Lin *et al.*, 2014). Therefore, we use the object annotations from COCO to create additional closed binary questions, such as “*Is there a bottle*” for the

example in Figure 6.1. We also create “adversarial” questions about objects, like “*Is there a wine-glass?*” by using an object that is not present in the image (wine-glass), but is *semantically close* to an object in the image (bottle). We use Glove vectors (Pennington *et al.*, 2014) to find the adversarial object with the closest embedding. Following a similar strategy, we also convert captions provided in COCO to closed binary questions, for example “*Does this seem like a man bending over to look inside the fridge?*”. Since we know what objects are present in the image, and the captions describe a “true” scene, we are able to obtain the ground-truth answers for questions created from objects and captions. Similar methods for creation of question-answer pairs have previously been used in (Ren *et al.*, 2015a; Malinowski and Fritz, 2014).

Thus for every question, we obtain several questions from objects and captions, and use these to compose additional questions by following a process similar to the one for VQA-Compose. For each closed question in the VQA dataset, we get 20 additional logically composed questions by utilizing questions created from objects and captions, yielding a total of *2.55 million samples* as VQA-Supplement.

6.3.3 Analytical Setup

In order to test the robustness of our models to logically composed questions, we devise five key experiments to analyse baseline models and our methods. These experiments help us gain insights into the nuances of the VQA dataset, and allow us to develop strategies for promoting robustness.

Effect of Data Augmentation: In this experiment, we compare the performance of models on VQA-Compose and VQA-Supplement with or without logically composed training data. This experiment allows us to test our hypotheses about the robustness

of any VQA model to logically composed questions. We first use models trained on VQA data to answer questions in our new datasets and record performance. We then explicitly train the same models with our new datasets, and make a comparison of performance with the pre-trained baseline.

Learning Curve: We train our models with an increasing number of logically composed questions and compare performance. This serves as an analysis of the number of logical samples needed by the model to understand logic in questions.

Training only with Closed Questions: In this ablation study, we restrict the training data to only closed questions i.e. “Yes-No” VQA questions, **VQA-Compose** and **VQA-Supplement**, allowing our model to focus solely on closed questions.

Compositional Generalization: We address whether training on closed questions containing single logical operation ($\neg Q_1$, $Q_1 \vee Q_2$) can generalize to multiple operations ($Q_1 \wedge \neg Q_2$, $\neg Q_1 \vee Q_2$). For instance, rows 1 through 6 in Table 6.1 are *single operation questions*, while rows 7 through 12 are *multi-operation questions*. Our aim is to have models that exhibit such compositional generalization.

Inductive Generalization: We investigate if training on compositions of two questions ($\neg Q_1 \vee Q_2$) can generalize to compositions of more than two questions ($Q_1 \wedge \neg Q_2 \wedge Q_3 \dots$). This studies whether our models develop an understanding of logical connectives, as opposed to simply learning patterns from large data.

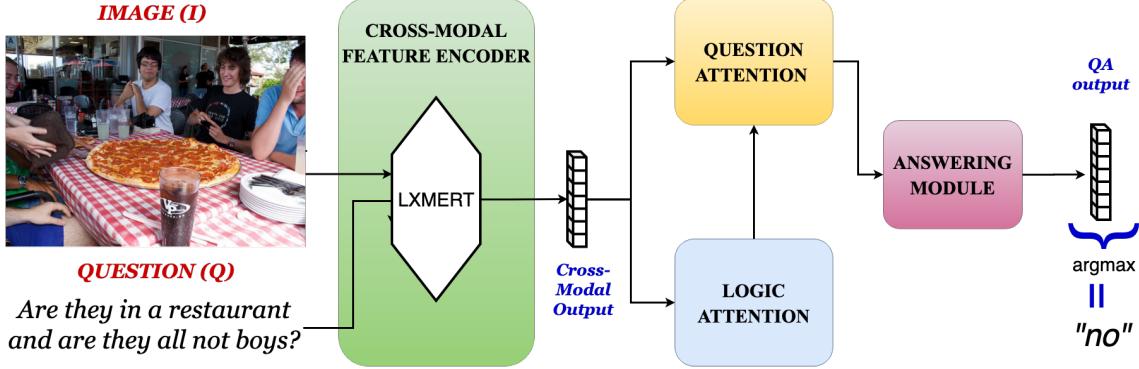


Figure 6.3: LOL model architecture showing a cross-modal feature encoder followed by our Question-Attention (q_{ATT}) and Logic Attention (ℓ_{ATT}) modules. The concatenated output of is used by the Answering Module to predict the answer.

6.4 Method

In this section, we describe LXMERT (Tan and Bansal, 2019) (a state-of-the-art VQA model), our Lens of Logic (LOL) model, attention modules which learn the question-type and logical connectives in the question, and the Fréchet-Compatibility (FC) Loss. This section refers to a composition of two questions, but applies to $n \geq 2$ questions.

6.4.1 Cross-Modal Feature Encoder

LXMERT (Learning Cross-Modality Encoder Representations from Transformers) (Tan and Bansal, 2019) is one of the first cross-modal pre-trained frameworks for vision-and-language tasks, that combines a strong visual feature extractor (Ren *et al.*, 2015b) with a strong language model (BERT)(Devlin *et al.*, 2019). LXMERT is pre-trained for key vision-and-language tasks, on a large corpus of $\sim 9M$ image-sentence pairs, making it a powerful cross-modal encoder for vision+language tasks such as visual question answering, as compared to other models such as MCAN (Yu *et al.*,

2019) and UpDn (Anderson *et al.*, 2018a), and strong representative baseline for our experiments.

6.4.2 Our Model: Lens of Logic (LOL)

The design for our LOL model is driven by three key insights:

1. As logically composed questions are closed questions, understanding the type of question will guide the model to answer them correctly.
2. Predicted answers must be compatible with the predicted question type. For instance, a closed question can have an answer that is either “Yes” or “No”.
3. The model must learn to identify the logical connectives in a question.

Given these insights, we develop the Question Attention module that encodes the type of question (*Yes-No*, *Number*, or *Other*), and the Logic Attention module that predicts the connectives (*AND*, *OR*, *NOT*, *no connective*) present in the question, and use these to learn representations. The overall model architecture is shown in Figure 6.3. For every question Q and corresponding image I , we obtain embeddings z_Q and z_I respectively, as well as a cross-modal embedding z_X .

Question Attention Module (q_{ATT}) takes cross-modal embedding z_x from LXMERT as input, and outputs vector $P_{type} = softmax(\mathbf{q}_{ATT}(z_x))$, representing the probabilities of each question-type. These probabilities are used to get a final representation \mathbf{z}^{type} which combines the features for each question-type.¹

Logic Attention Module (ℓ_{ATT}) takes the cross-modal embedding z_X from LXMERT as input, and outputs vector $P^{conn} = \sigma(\ell_{ATT}(z_X))$ which represents the probabilities of each type of connective. We use sigmoid (σ) instead of a softmax, since a question can have multiple connectives. These probabilities are used to combine the

features for each type of connective into a final representation \mathbf{z}^{conn} which encodes information about the connectives in the question.

6.4.3 Loss Functions

We train our models jointly with the loss function given by:

$$\mathcal{L} = (1 - \alpha_1 - \alpha_2) \cdot \mathcal{L}_{\text{ans}} + \alpha_1 \cdot \mathcal{L}_{\text{type}} + \alpha_2 \cdot \mathcal{L}_{\text{conn}} + \beta \cdot \mathcal{L}_{\text{FC}}. \quad (6.2)$$

Answering Loss ℓ_{ans} is conditioned on the type of question. We multiply the final prediction vector with the probability and the mask M_i for question-type i . M_i is a binary vector with 1 for every answer-index of type-i and 0 elsewhere:

$$\mathcal{L}_{\text{ans}} = \mathcal{L}_{\text{BCE}}\left(\sum_{i=1}^3 \hat{y} \odot M_i \cdot P_i^{\text{type}}, y_{\text{ans}}\right). \quad (6.3)$$

Attention Losses: q_{ATT} is trained to minimize a Negative Log Likelihood (NLL) classification loss, ensuring a shrinkage of probabilities of the answer choices of the wrong type. ℓ_{ATT} is trained to minimize a multi-label classification loss, using Binary Cross-Entropy (BCE) given by:

$$\mathcal{L}_{\text{type}} = \mathcal{L}_{\text{NLL}}(\text{softmax}(z^{\text{type}}), y_{\text{type}}), \quad (6.4)$$

$$\mathcal{L}_{\text{conn}} = \mathcal{L}_{\text{BCE}}(\sigma(z^{\text{conn}}), y_{\text{conn}}), \quad (6.5)$$

where $y_{\text{ans}}, y_{\text{type}}, y_{\text{conn}}$ are labels for answer, question-type and connective.

Fréchet-Compatibility Loss: We introduce a new loss function that ensures compatibility between the answers predicted by the model for the component questions Q_1 and Q_2 and the composed question Q . Let A, A_1, A_2 be the respective answers

predicted by the model for Q , Q_1 , and Q_2 . Q_i can have negation. Then Fréchet inequalities (Boole, 1854; Fréchet, 1935) provide us with bounds for the probabilities of the answers of the conjunction and disjunction of the two questions:

$$\max(0, p(A_1) + p(A_2) - 1) \leq p(A_1 \wedge A_2) \leq \min(p(A_1), p(A_2)). \quad (6.6)$$

$$\max(p(A_1), p(A_2)) \leq p(A_1 \vee A_2) \leq \min(1, p(A_1) + p(A_2)). \quad (6.7)$$

We define “Fréchet bounds” b_L and b_R to be the left and right bounds for the triplet A, A_1, A_2 , and the “Fréchet Mean” m_A to be the average of the Fréchet bounds; $m_A = (b_L + b_R)/2$. Then, the Fréchet-Compatibility Loss given by:

$$\mathcal{L}_{FC} = (p(A) - \mathbb{1}(m_A > 0.5))^2, \quad (6.8)$$

ensures that the predicted answer and that determined by m_A match.

6.4.4 Implementation Details

The LXMERT feature encoder produces a vector z of length 768 which is used by our attention modules, each having sub-networks $\mathbf{f}_i, \mathbf{g}_i$ with 2 feed-forward layers. We first train our models without FC loss. Then we select the best models with a checkpoint of 10 epochs and finetune these further for 3 epochs with FC loss, since the FC loss is designed to work for a model whose predictions are not random. Thus our improvements in accuracy are attributable to the FC Loss and not more training epochs. We utilize the Adam optimizer (Kingma and Ba, 2015) with a learning rate of $5e-5$, batch size of 32 and train for 20 epochs. Our models are trained on 4 NVIDIA V100 GPUs, and take approximately 24 hours for training 20 epochs.

6.5 Experiments

We first conduct analytical experiments to test for logical robustness and transfer learning capability. We use three datasets for our experiment: the VQA v2.0 (Antol

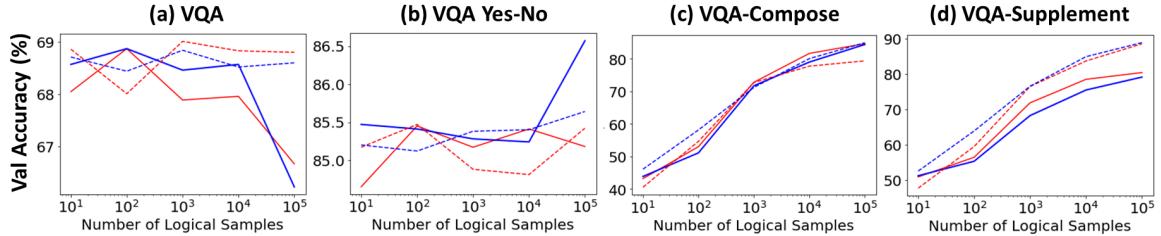


Figure 6.4: Learning Curve comparison for models (Red: LXMERT, Blue: LOL) trained on our datasets (solid lines: VQA + Comp, dotted lines: VQA + Comp + Supp)

et al., 2015) dataset, a combination of VQA and our VQA-Compose dataset, and a combination of VQA, VQA-Compose and VQA-Supplement. The size of the training dataset and the distribution of yes-no, number and other questions is kept the same as the original VQA dataset ($\sim 443k$) for fair comparison. Since VQA-Supplement uses captions and objects from MS-COCO, we use it to analyze the ability of our models to generalize to a new source of data (MS-COCO) as well as questions containing adversarial objects. After training, our attention modules (q_{ATT} and ℓ_{ATT}) achieve an accuracy of 99.9% on average, showing almost perfect performance when it comes to learning the type of question and the logical connectives present in the question.

6.5.1 Can't We Just Parse the Question into Components?

Since our questions are a composition of multiple questions, an obvious approach is to split the question into its components, and to discern the logical formula for composition. The answers to these component questions (predicted by VQA models) can be *re-combined* with the predicted logical formula to obtain the final answer. We use parsers to map components and logical operations to predefined slots in a logical function. The oracle parser uses the ground truth component questions and combines predicted answers using the true formula. However, at test time we do not have

⁰In all tables, best overall scores are bold, our best scores underlined.

Model	Trained on	Validation Accuracy (%) ↑			
		VQA	YN	Comp	Supp
LXMERT	VQA	68.94	86.65	50.79	50.51
	VQA + Comp	67.85	85.32	85.03	80.85
	VQA + Comp + Supp	68.83	84.83	70.28	85.17
	<i>with FC Loss</i> VQA + Comp + Supp	67.84	84.92	75.31	85.25
LOL (qATT)	VQA	69.08	<u>85.32</u>	48.99	50.54
	VQA + Comp	67.51	84.82	84.85	79.62
	VQA + Comp + Supp	68.72	84.99	79.88	87.12
	VQA + Comp	68.94	85.15	<u>85.13</u>	79.02
LOL (Full)	VQA + Comp + Supp	68.86	84.87	81.07	87.54
	<i>with FC Loss</i> VQA + Comp + Supp	68.10	84.75	82.39	<u>87.80</u>
LXMERT	YN + Comp	-	84.13	84.44	79.39
	YN + Comp + Supp	-	84.09	82.63	88.15
LOL (ℓ ATT)	YN + Comp	-	85.22	<u>85.31</u>	79.87
	YN + Comp + Supp	-	85.26	84.37	<u>89.00</u>

Table 6.2: Comparison of LXMERT and LOL trained on VQA data, combinations with **Compose**, **Supplement**, and our Frechet-Compatibility (FC) Loss

Model	YN	VQA-Compose		VQA-Supplement	
		Single	Multiple	Single	Multiple
LXMERT	85.07	83.95	61.99	86.65	60.00
LOL	85.12	84.60	66.03	87.42	66.05

Table 6.3: Validation accuracies (%) for Compositional Generalization. Note that 50% is random performance.

Model	VQA-Compose		VQA-Supplement	
	$Q_1 \circ Q_2$	$Q_2 \circ Q_1$	$Q_1 \circ Q_2$	$Q_2 \circ Q_1$
LXMERT	82.34	80.44	85.57	81.78
LOL	84.91	83.64	85.62	83.41

Table 6.4: Validation accuracies (%) for Compositional Generalization and Commutative Property. Note that 50% is random performance.

access to the true mapping and components. So we train a RoBERTa-Base (Liu *et al.*, 2019c) parser using B-I-O tagging (Ramshaw and Marcus, 1995) for a Named-Entity Recognition task with constituent questions as entities.¹

The performance of the oracle parser serves as the upper bound as we have a perfect mapping, with the QA system being the only source of error. The trained parser has an exact-match accuracy of 85%, but only a 72% accuracy in determining the number of operands. The parser has an accuracy of 89% for questions with 3 or less operands, but only 78% for longer compositions. End-to-end (E2E) models do not need to parse questions and hence overcome these hurdles, but do require an understanding of logical operations. Table 6.5 shows that both oracle and trained

parsers when used with LOL outperform parsers with LXMERT, by 6.82% and 5.60% respectively. The LOL model without using any parsers is better than both LXMERT and LOL with the trained parser by 7.55% and 1.95% respectively.

6.5.2 *Explicit Training with Logically Composed Questions*

Can models trained on the VQA-v2 dataset answer logically composed questions? The first section of Table 6.2 shows that LXMERT, when trained only on questions from VQA-v2 has near random accuracy ($\sim 50\%$) on our logically composed datasets, thus exhibiting little robustness to such questions.

Can baseline model improve if trained explicitly with logically composed questions questions? We train the models with data containing a combination of samples from VQA-v2, VQA-Compose, and VQA-Supplement. The accuracy on VQA-Compose and VQA-Supplement improves, but there is a drop in performance on yes-no questions from VQA. Our models with our attention modules (q_{ATT} and ℓ_{ATT}) are able to retain performance on VQA-v2 while achieving improvements on all validation datasets.

6.5.3 *Analysis*

Training with Closed Questions only: We analyse the performance of models when trained only with closed questions from VQA, VQA + Comp and VQA + Comp + Supp and see that our model achieves the best accuracy on logically composed questions, as shown in sections 3 and 4 in Table 6.2. Since we train only closed questions, we do not use our question attention module for this experiment.

Effect of Logically Composed Questions: We increase the number of logical samples in the training data on a log scale from 10 to 100k. As can be seen from the learning curves in Figure 6.4(a), models trained on VQA + Comp + Supp are able to retain performance on VQA validation data, while those trained only on VQA + Comp data deteriorate. Figure 6.4(b) shows that our models improve on VQA Yes-No performance after being trained on more logically composed samples, exhibiting transfer learning capabilities. In (c) both our models are comparable to the baseline, but our model shows improvements over the baseline when trained on VQA + Comp + Supp. In (d) for all levels of additional logical questions, our model trained on VQA + Comp + Supp is the best performing. From (c) and (d), we observe that a large number of logical questions are needed during training for the models to learn to answer them during inference. We also see that our model yields the best performance on VQA-Supplement.

Compositional Generalization: To test for compositional generalization, we train models on questions with a maximum of one connective (single) and test on those with multiple connectives. It can be seen from Table 6.4 that our models are better equipped than the baseline to generalize to multiple connectives and also to be able to generalize from VQA-Compose to Supplement.

Inductive Generalization: We test our models on questions composed with more than two components. Parser-based models have this property by default. As shown by Figure 6.3 our E2E models outperform the baseline LXMERT.

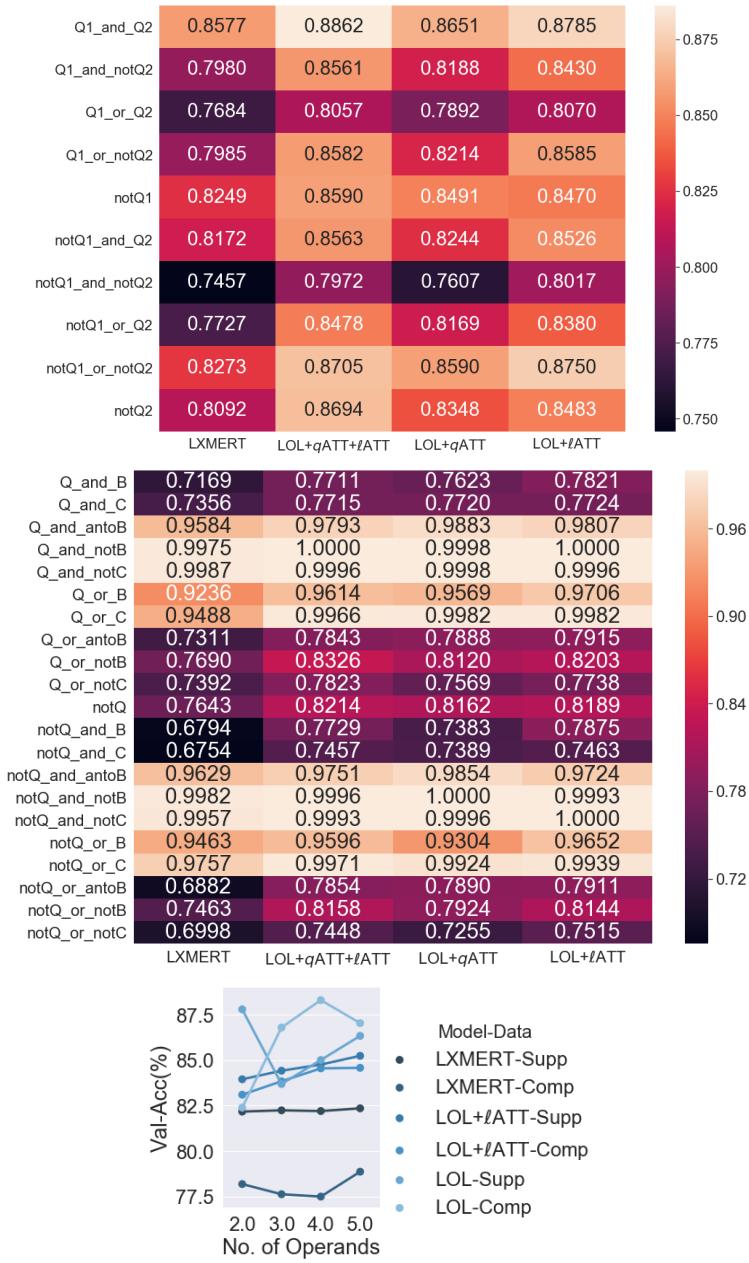


Figure 6.5: Accuracy for each type of question in (a) VQA-Compose, (b) VQA-Supplement and for questions with number of operands greater than 2.

Commutative Property: Our models have identical answers when the question is composed either as $Q_1 \circ Q_2$ or $Q_2 \circ Q_1$, for logical operation \circ , as shown in Table 6.4. The parser-based models are agnostic to the order of components if the parsing is accurate, while our E2E models are robust to the order.

Accuracy per Category of Question Composition: In Figure 6.5 we show a plot of accuracy versus question type for each model. Q, Q_1, Q_2 are questions from VQA, B, C are object-based and caption-based questions from COCO respectively. From the results, we interpret that questions such as $Q \wedge \text{antonym}(B), Q \wedge \neg B, Q \wedge \neg C$ are easy because the model is able to understand absence of objects, therefore can always answer these questions with a “NO”. Similarly, $Q \vee B, Q \vee C$ are easily answered since presence of the object makes the answer always “YES”. By simply understanding object presence many such questions can be answered. Figure 6.5 shows the model has the same accuracy for logically equivalent operations.

6.5.4 Evaluation on VQA v2.0 Test Data

Table 6.5 shows the performance the VQA Test-Standard datset. Our models maintain overall performance on the VQA test dataset, and at the same time substantially improve from random performance ($\sim 50\%$) on logically composed questions to 82.39% on VQA-Compose and 87.80% on VQA-Supplement. This shows that logical connectives in questions can be learned while not degrading the overall performance on the original VQA test set (our models are within $\sim 1.5\%$ of the state-of-the-art on all three types of questions on the VQA test-set).

Model	Parser	Training Data	Test-Std. Accuracy (%) ↑				Val. Accuracy (%) ↑		
			Yes-No	Number	Other	Overall	Compose	Supplement	Overall
MCAN	None	VQA (Yu et al., 2019)	86.82#	53.26#	60.72#	70.90	52.42	*	*
LXMERT	None	VQA (Tan and Bansal, 2019)	88.20	54.20	63.10	72.50	50.79	50.51	50.65
LOL (<i>q</i> ATT)	None	VQA	<u>87.33</u>	<u>54.03</u>	<u>62.40</u>	<u>72.03</u>	48.99	50.54	49.77
LXMERT	Oracle	VQA	88.20	54.20	63.10	72.50	86.38	74.29	80.33
LXMERT	Trained	VQA	88.20	54.20	63.10	72.50	86.35	68.75	77.55
LOL (full)	Oracle	VQA+Ours	86.55	53.42	61.58	71.04	85.79	88.51	87.15
LOL (full)	Trained	VQA+Ours	86.55	53.42	61.58	71.04	82.13	84.17	83.15
LXMERT	None	VQA+Ours	85.23	51.25	60.58	69.78	75.31	85.25	80.28
LOL (<i>q</i> ATT)	None	VQA+Ours	86.79	52.66	61.85	71.19	79.88	87.12	83.50
LOL (full)	None	VQA+Ours	86.55	53.42	61.58	71.04	82.39	87.80	85.10

Table 6.5: Performance on ‘test-standard’ set of VQA-v2 and validation set of our datasets.

LOL performance is close to SOTA on VQA-v2, but significantly better at logical robustness.

*MCAN uses a fixed vocabulary that prohibits evaluation on **VQA-Supplement** which has questions created from COCO captions. #Test-dev scores, since MCAN does not report test-std single-model scores²

6.6 Discussion

Consider the example, “*Is every boy who is holding an apple or a banana, not wearing a hat?*”, humans are able to answer it to be true if and only if each boy who is holding *at least one* of an apple or a banana is not wearing a hat ([Cesana-Arlotti et al., 2018](#)). Natural language contains such complex logical compositions, not to mention ambiguities and the influence of context. In this paper, we focus on the simplest – negation, conjunction, and disjunction. We have shown that existing VQA models are not robust to questions composed with these logical connectives, even when we train parsers to split the question into its components. When humans are faced with such questions, they may refrain from giving binary (Yes/No) answers.

For instance, logically, the question “*Did you eat the pizza and did you like it?*” has a negative answer if either of the two component questions has a negative answer. However, humans might answer the same question with the answer “*Yes, but I did not like it*”. While human question-answering is indeed elaborate, explanatory, and clarifying, that is the scope of our future work; here we focus only on predicting a single binary answer.

We have shown how connectives in a question can be identified by enhancing LXMERT encoders with dedicated attention modules and loss functions. We would like to stress on the fact that we do not use knowledge of the connectives during inference, but instead train the network to be aware of it based on cross-modal features, instead of predicting purely based on language model embeddings which fail to capture these nuances. We believe this work has potential implications on logic-guided data augmentation, logically robust question answering, and for conversational agents (with or without images). Similar strategies and learning mechanisms will be used in the next chapter to operate “logically” in the image-space at the level of object classes and their attributes.

The work on VQA-LOL spawned off VQA-MUTANT – a data augmentation strategy to address changing priors between train and test datasets. In this paper, we make use of simple image transformations which remove objects or change their colors, in addition to the logical transformations developed in VQA-LOL. Empirical results on the VQA-CP challenge ([Agrawal et al., 2018](#)) show that this method achieves robustness under changing question-answer priors (i.e. when the conditional probability of answers given a question type varies between train and test domains). This work was published as a conference paper in EMNLP 2020 ([Gokhale et al., 2020b](#)) and was a joint work with Pratyay Banerjee ([Banerjee, 2022](#)).

Chapter 7

SEMANTICALLY DISTRIBUTED ROBUST OPTIMIZATION FOR VISION AND LANGUAGE INFERENCE

“Does the text match the image?” – this simple question represents the Vision-and-Language Inference (VLI) task, as shown in Figure 7.1. Image-text matching forms the backbone for V&L pre-training (Sun *et al.*, 2019; Tan and Bansal, 2019; Lu *et al.*, 2019; Chen *et al.*, 2020b) and has resulted in improvements in downstream tasks such as visual question answering, image retrieval, referring expressions, and visual commonsense reasoning. While Natural Language Inference (without visual inputs) has been extensively studied (Bowman *et al.*, 2015a; Williams *et al.*, 2018; Khot *et al.*, 2018; Demszky *et al.*, 2018), VLI demands the additional capability of being grounded in the scene while understanding semantics. Although pre-trained language models (PLMs) (Vaswani *et al.*, 2017; Devlin *et al.*, 2019; Raffel *et al.*, 2020) have been useful for encoding text into vector embeddings, recent findings point to undesirably high cosine similarity of two random words (Ethayarajh, 2019), the struggle with negation (Kassner and Schütze, 2020a; Ettinger, 2020), and semantically equivalent adversarial examples (Ribeiro *et al.*, 2018a). These findings call for robust training protocols to avoid propagation of these findings into VLI models.

Adversarial training (AT) and distributed robust optimization (DRO) (Madry *et al.*, 2018a; Hu *et al.*, 2018b; Sinha *et al.*, 2018b) have emerged as effective solutions to related problems in robust image classification, such as adversarial defense and domain generalization (Volpi *et al.*, 2018). DRO assumes a perturbation set (typically an ℓ_p norm ball) around the training distribution, and minimizes the worst-case performance over this perturbation set. AT and DRO are popular for computer vision tasks, since

NLVR2



At least one sail boat is parked on the **True** left side of the dock.

The left and right image contains the same number of sailboats sailing on the **False** water.

VIOLIN



- *Gavin Mitchell's office. Rachel Green's office.*
- *Give me that phone*
- *Hello this is Rachel Green. How can I help you?*
- *Uh-huh. Okay then. I'll pass you back to your son.*
- *Hey, Mom. No that's just my secretary.*

The phone rings, a man picks it up, and a woman slams her hand on the desk and **True** demands the man give her the phone.

The man realizes it is the woman's mother who is calling and he passes the **False** phone back to the woman

Figure 7.1: VLI models predict whether a sentence is **True** or **False** about the visual input. (*Top*) sample from NLVR² with two images as input; (*bottom*) sample from VIOLIN with video and subtitles as input.

the small perturbations of pixel intensities do not change the categorical meaning of the image.

However, in the case of text inputs, even small perturbations of their vector embeddings may result in absurd sentences or vectors that do not map to any word-token in vocabulary. The topology of the PLM embedding space is not well understood, especially with regard to what kind (and magnitude) of perturbations result in specific changes in semantics, such as similar meanings (speak → talk) or opposite meanings (Heaven → Hell) without resulting in random or absurd words. We therefore argue that vector-based additive perturbations present limitations in the case of text inputs. However, in the domain of natural language, we are blessed with semantic and logical transformations as shown in Table 7.1. This enables linguistically-informed perturbations with control over the semantics of the resulting sentence and label, as

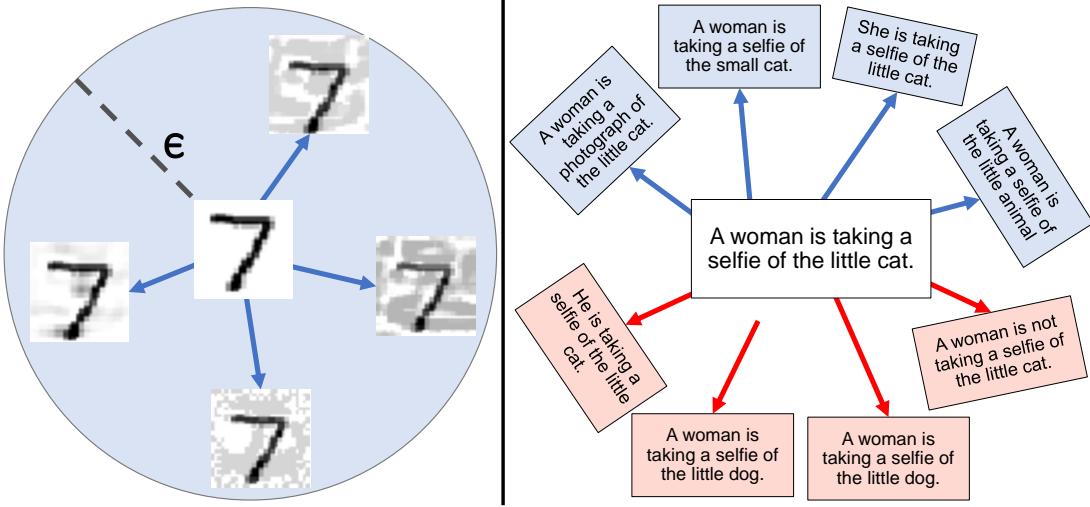


Figure 7.2: Comparison between (*left*) ϵ -bounded image perturbations and (*right*) linguistics-based *Semantics-preserving* (blue) as well as *Semantics-Inverting* (red) transformations for sentences.

shown in Figure 7.2.

We present a technique that modifies robust optimization by incorporating linguistically-informed transformations. Our approach: **Semantically Distributed Robust Optimization** (SDRO) utilizes a pre-defined set of linguistic transformations (such as negation, word substitution, and paraphrasing) as the perturbation set instead of optimizing over the vector-space. We dub this set of transformations “SISP” i.e., semantics-inverting (SI) and semantics-preserving (SP) transformations. SDRO is *model-agnostic* since it can be applied to text inputs of any existing VLI model and *dataset agnostic* since it uses automated transformations without explicit knowledge of the text domain.

We apply SDRO to two VLI benchmark datasets: image-based NLVR² ([Suhr et al., 2019](#)) as well as video-based VIOLIN ([Liu et al., 2020a](#)). To demonstrate the generalizability of SDRO to other V&L tasks, we also report results on the “yes/no” subset of VQA-v2 ([Goyal et al., 2017](#)). Our experiments show model-agnostic improvements in accuracy for all three benchmarks. While models trained with naive

data augmentation using SISP suffer from a trade-off between robustness and accuracy, models that utilize SDRO improve along both metrics. SDRO also allows us to learn in low-resource settings, serving as a smart data augmentation tool – SDRO models trained only with 80% of the original dataset outperform existing state-of-the-art which utilizes the entire dataset.

Since SISP transforms do not require the true label to either produce an SP or SI transformed sentence, we can apply them at test-time. Given a test input sentence, we generate its SISP versions and ensemble the predictions made by the model, giving equal weight to the prediction for the original sentence and the average predictions for all transformed sentences. We find that this ensembling of predictions of the SDRO model at test-time pushes the state-of-the-art further, thereby demonstrating the usefulness of semantic sentence transformations, both during training and testing.

7.1 Method

7.1.1 Preliminaries

Consider a training distribution P_{tr} consisting of inputs \mathbf{x} and labels \mathbf{y} . For VLI, input \mathbf{x} is multi-modal (visuals and text), with labels $\mathbf{y} \in \{\text{True}, \text{False}\}$. Under the empirical risk minimization (ERM), the following risk is minimized:

$$\mathcal{R}_{ERM} = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim P_{tr}} \ell(f(\mathbf{x}; \theta), \mathbf{y}). \quad (7.1)$$

ERM provides generalization guarantees (Vapnik, 1991) for i.i.d. test samples, but not for out-of-distribution or adversarial examples.

Distributed Robust Optimization (DRO) (Hu *et al.*, 2018a; Sagawa *et al.*, 2020) searches for loss-maximizing perturbations of the input within an ϵ -divergence

Category	Original	Transformed	
SI	Noun-Antonym	The two women are driving on the street with the convertible top down.	The two men are driving on the street with the convertible top down.
	Verb-Antonym	There are children standing by the door.	There are children sitting by the door.
	Comparative-Antonym	There are more monitors in the image on the right than on the left.	There are few monitors in the image on the right than on the left.
	Number-Substitution	There are three bowls of dough with only one spatula.	There are eleven bowls of dough with only one spatula.
	Pronoun-Substitution	In one of the images, a woman is taking a selfie.	In one of the images, he is taking a selfie.
	Subject-Object Swap	The two women are driving on the street with the convertible top down.	The two top are driving on the street with the convertible women down.
SP	Negation	The closet doors on the right are mirrored.	The closet doors on the right are not mirrored
	Noun-Synonym	The right image shows three bottles of beer lined up.	The right picture shows three bottles of beer lined up.
	Verb-Synonym	Someone is using a kitchen utensil	Someone is utilizing a kitchen utensil.
	Comparative-Synonym	The bottle on the right is larger than the bottle on the left.	The bottle on the right is bigger than the bottle on the left.
	Number-Substitution	The two white swans are swimming in the canal gracefully.	The less than seven white swans are swimming in the canal gracefully.
	Pronoun-Substitution	In one of the images, a woman is taking a selfie.	In one of the images, she is taking a selfie.
	Paraphrasing	A man in a green shirt came on the porch and knocked on the door.	A man in a green shirt came up to the porch and knocked on the door.

Table 7.1: Illustrative examples for the effect of each SISP transformation on input sentences.

ball around P_{tr} and minimize the risk over such perturbed distributions.

$$\mathcal{R}_{DRO} = \sup_{P: D(P, P_{tr}) < \epsilon} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim P} \ell(f(\mathbf{x}; \theta), \mathbf{y}). \quad (7.2)$$

The solution to Equation 7.2 guarantees robustness inside such ϵ -bounded distributions P . The inner maximization is typically solved using gradient-based methods (Madry *et al.*, 2018a) over additive perturbations δ such that $\mathbf{x} + \delta$ fools the classifier.

7.1.2 SDRO

For sentence inputs, additive perturbations are intangible and may result in ambiguity. An alternative approach is to consider *groups* or perturbations sets \mathcal{G} representing certain subpopulations or semantic categories within the data distribution. For text inputs, we consider perturbation sets that can be created using semantic sentence transformations such as those shown in Table 7.1, as the *groups* (or equivalently, *transformations*). These transformations $g(x, y) = (\mathbf{x}_g, \mathbf{y}_g)$ are of two types: semantics-preserving (SP) if $\mathbf{y}_g = \mathbf{y}$, or semantics-inverting (SI) if $\mathbf{y}_g \neq \mathbf{y}$. The ability of generating adversarial samples with inverted meanings is a key distinction between adversarial training (AT) and SDRO. While AT is restricted to SP perturbations inside an ϵ norm-ball, SDRO can impart larger linguistic perturbations (both SI and SP) beyond the norm-ball, by minimizing the worst-case expected risk over these groups:

$$\mathcal{R}_{SDRO} = \sup_{g \in \mathcal{G}} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim g} \ell(f(\mathbf{x}; \theta), \mathbf{y}). \quad (7.3)$$

Implementation. As a first step of SDRO, we randomly sample a subset \mathcal{C} of the training dataset \mathcal{D} s.t. $|\mathcal{C}|/|\mathcal{D}| = T$. We find adversarial samples after every epoch and create an augmented dataset \mathcal{D}_{aug} which contains $(1 - T)|\mathcal{D}|$ original samples and $T|\mathcal{D}|$ adversarial samples, thus retaining the size of the training dataset. We define ℓ_g

as the classification loss for a transformed sample $(\mathbf{x}_g, \mathbf{y}_g)$:

$$\ell_g(\mathbf{x}, \mathbf{y}) \triangleq \ell(f(\mathbf{x}_g), \mathbf{y}_g), \quad \forall g \in \mathcal{G}. \quad (7.4)$$

We design two variants of SDRO: Sample-Wise (SW) and Group-Wise (GW).

Sample-Wise SDRO: In this greedy version of SDRO, for every input \mathbf{x} , a transformation that maximally fools the classifier: $g^* = \text{argmax}_{g \in \mathcal{G}} \ell_g(\mathbf{x}, \mathbf{y})$, is added to the set of adversarial examples \mathcal{D}_{adv} . The model is then fine-tuned on the augmented dataset.

$$\mathcal{D}_{adv} = \{g^*(\mathbf{x}, \mathbf{y}) : (\mathbf{x}, \mathbf{y}) \in \mathcal{C}\}, \quad (7.5)$$

$$\mathcal{D}_{aug} = \mathcal{D}_{1:(1-T)|\mathcal{D}|} \cup \mathcal{D}_{adv} \quad (7.6)$$

However, this greedy approach is susceptible to the model’s biases towards certain transformations. For instance, if negation and verb-antonym are universally hard for most sentences, i.e., result in the maximum classifier loss amongst all transformations g , then \mathcal{D}_{adv} will be dominated by these groups, resulting in an unbalanced training set.

Group-Wise SDRO is devised to mitigate against the model becoming biased towards the “hardest” transformations. Using Equation 7.4, we calculate the transformation losses for each transformation of each sample in a training batch, yielding a set of classifier losses per “group” g :

$$L_g : \mathcal{C} \rightarrow \curvearrowright; \quad L_g = \{\ell_g(\mathbf{x}, \mathbf{y}) : (\mathbf{x}, \mathbf{y}) \in \mathcal{C}\}. \quad (7.7)$$

We obtain the top-k losses per group g as:

$$L_G^k = \underset{\Lambda \subset L_G, |\Lambda|=k}{\text{argmax}} \sum_{\lambda \in \Lambda} \lambda, \quad \text{where } k = \left\lfloor \frac{|\mathcal{C}|}{|\mathcal{G}|} \right\rfloor. \quad (7.8)$$

Then \mathcal{D}_{adv} is compiled as the union of per-group adversaries using Equation 7.8, and augmented to the training dataset using Equation 7.6.

Test-Time Ensembling of Predictions. Semantic transformations g allow us to obtain multiple “views” $\mathbf{x}_g = g(\mathbf{x})$ of the input, and the corresponding predictions $\hat{\mathbf{y}}_g = f(\mathbf{x}_g)$. We ensemble these predictions and the original prediction $\hat{\mathbf{y}} = f(\mathbf{x})$ with a simple weighted-average. Note that \mathcal{G} contains both SP and SI transformations, \mathcal{G}_{SP} and \mathcal{G}_{SI} . Since the expected label for \mathcal{G}_{SI} is flipped, during ensembling we use the flipped probabilities $1 - f(\mathbf{x}_g)$. The ensembled prediction is:

$$\hat{\mathbf{y}}_e = \alpha f(\mathbf{x}) + \frac{1-\alpha}{2} \sum_{g \in \mathcal{G}_{SP}} \frac{f(\mathbf{x}_g)}{|\mathcal{G}_{SP}|} + \frac{1-\alpha}{2} \sum_{g \in \mathcal{G}_{SI}} \frac{1-f(\mathbf{x}_g)}{|\mathcal{G}_{SI}|}. \quad (7.9)$$

This ensembling is in principle similar to [Chai et al. \(2021\)](#) who train a generative model g to output different views of an image, and tune α over a validation set. In our work, g are semantic sentence transformations, and a simple intuitive choice of $\alpha=0.5$ gives equal weight to the original sample and the SISP versions. We find that:

1. training models with SDRO using SISP transformations improves results on VLI tasks, and
2. ensembling predictions of SDRO at test-time using Equation 7.9 further improves results.

7.2 SISP Sentence Transformations

This section describes the generation of semantics-preserving (SP) and semantics-inverting (SI) statements. **SISP** transforms are implemented using Spacy ([Honnibal and Montani, 2017](#)).

Noun Synonym/Antonym: We extract nouns (subjects and objects) with dependency parsing, and find two nearest (synonyms) or farthest (antonyms) neighbors in the GloVe space ([Pennington et al., 2014](#)) using a threshold of 0.55.

Verb Synonym/Antonym: We extract verbs using POS tagging and obtain their synonyms or antonyms. Verbs are lemmatized and inflected to the correct form using Lemminflect ([Jascob, 2020](#)).

Comparative Synonym/Antonym: Adjectival complements and modifiers are replaced with synonyms (*large* → *big*) or antonyms (*large* → *small*).

Number Substitution: Numerals are replaced by number-words (2 → *two*) or vice versa for SP transformations, or by their lower or upper bounds, (SP: 3 → *more than two*, SI: *two* → *less than two*).

Pronoun Substitution: Human-related nouns (such as *woman*, *boy*, *people*) are substituted by pronouns, while pronouns are substituted by generic descriptors (*something*, *someone*, *somebody*, *they*).

Negation: We use template-based negation ([Gokhale et al., 2020c](#)) with Subject-Verb Agreement ([Wren and Martin, 2000](#)). We add ‘*did not*’ before a past-tense verb, ‘*do not*’, ‘*does not*’, or ‘*not*’ before a base-form verb, gerund, or participle, or a ‘*not*’ before an adposition or adjective.

Subject-Object Swap: Nominal or clausal subjects and direct or prepositional objects from the sentence are swapped for inverting semantics.

Paraphrasing: Input sentences are translated to Russian and then back-translated to English using neural machine translation ([Ott et al., 2019](#)).

7.2.1 Data Generation Pipeline

Figures 7.3 and 7.4 show flowcharts for our SISP transformation process for Semantics Preserving (SP) and Semantics Inverting (SI) respectively. For each image-sentence pair, the sentence is parsed using Spacy ([Honnibal and Montani, 2017](#)) into tokens, dependencies, POS-tags, and noun chunks. Using this, each SISP function (for instance “Noun Synonym”) generates insertions, deletions, substitutions, or paraphrasing as shown.

7.2.2 Data Analysis

7.2.3 Statistics

In Tables 7.2, 7.3, 7.4, we show the number of SISP-transformed samples generated for the test sets of NLVR², VIOLIN and VQA Yes/No respectively. While we generate samples exhaustively for each category of transformation, during training these are sampled according to the proportion of augmented samples T , using three sampling strategies – naive data augmentation, SW-SDRO or GW-SDRO. On average, we obtain 5.69 SI samples and 5.65 SP samples per original sample for the NLVR² dataset, 11.14 SI samples and 10.83 SP samples for VIOLIN, and 2.75 SI samples and 3.5 SP sample for the VQA-Yes/No subset.

Quantification of Bias: Since SISP transforms are based on templates, they can potentially introduce spurious linguistic correlations in the dataset. For example, in NLVR² and VIOLIN datasets, negations and indefinite pronouns are infrequent. To quantify how this could impact models, we mask out the entire image and evaluate models (with VILLA as the backbone for NLVR² and HERO for VIOLIN). This acts as a ‘text-only’ evaluation, with accuracies $\sim 50\%$ implying lesser bias since models

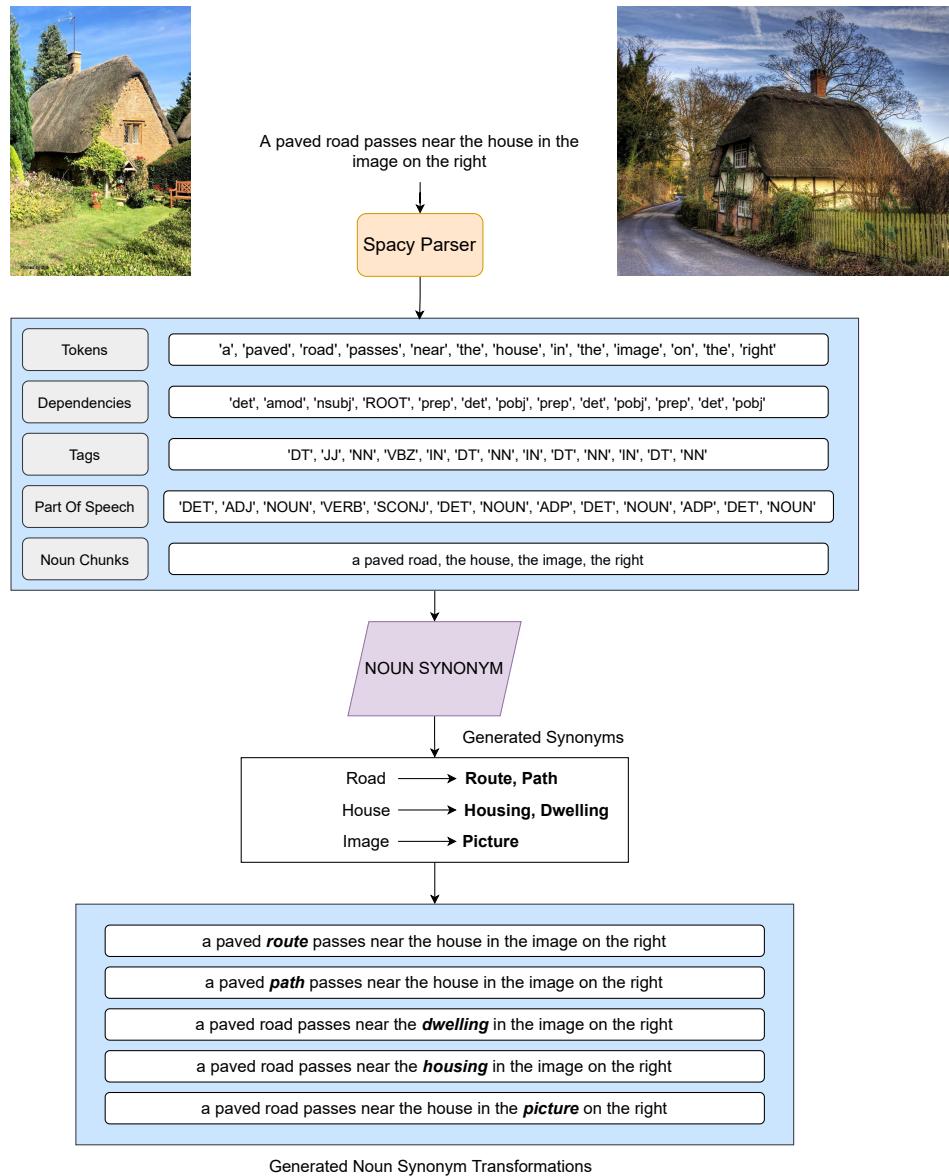


Figure 7.3: Illustration of the work-flow for generating SISP-transformed versions of input sentences. A Semantics-Preserving (SP) transformation is shown above.

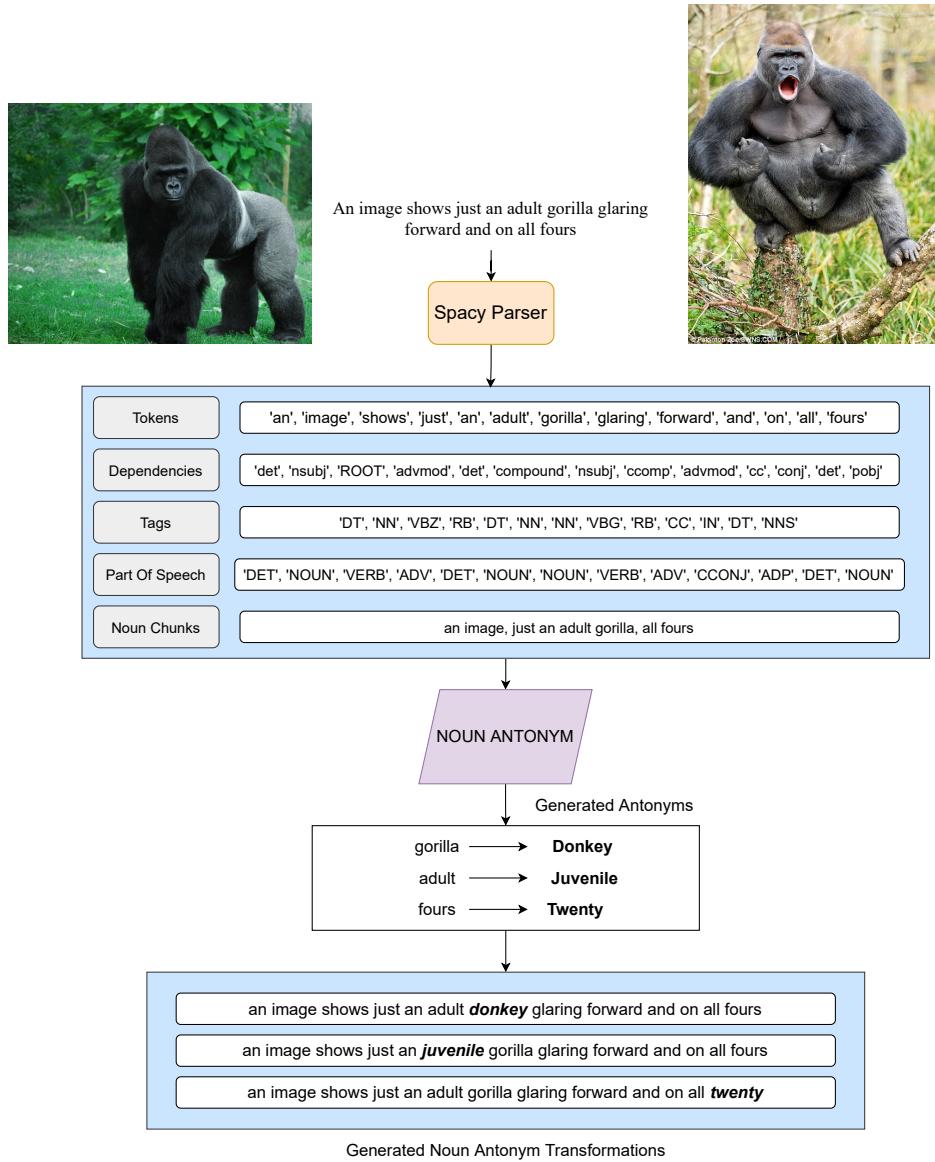


Figure 7.4: Illustration of the work-flow for generating SISP-transformed versions of input sentences. A Semantics-Inverting (SI) transformation is shown above.

3	sp_noun_synonym			In the right image, one horse is pulling a four-wheeled cart with two passengers to the right.	False	in the right image , one steed is pulling a four - wheeled cart with two passengers to the right .	False
4	si_negation			Three wine bottles with gold foil tops are stacked on a red mat.	True	three wine bottles with gold foil tops are not stacked on a red mat .	False
19	sp_noun_synonym			There are at least 4 gorillas sitting in the greenery.	True	there are at least 4 gorillas sitting in the green .	True
47	sp_noun_synonym			All of the pelicans are on shore and none of them are extending their wings.	True	all of the pelicans are on seaboard and none of them are extending their wings .	True
61	si_noun_anonym			One of the monitors is silver in color.	True	one of the monitors is purple in color .	False
62	sp_comparative_synonym			All paper rolls are upright, and one image shows a paper towel roll on an upright stand.	True	all paper rolls are vertical , and one image shows a paper towel roll on an upright stand .	True

Figure 7.5: Snapshot of a SISP example being evaluated by human subjects. Columns from left to right: sample-ID, SISP-tag, Left Image, Right Image, Original Sentence, Original Label, New Sentence, New Label.

	Category	Training	Test-P	Validation
	Original	86,373	6,967	6,982
SI	Comparative Antonym	14,177	1,244	1,172
	Negation	150,610	12,838	12,635
	Noun Antonym	148,959	12,719	12,635
	Number Substitution	83,080	7,468	7,113
	Pronoun Substitution	34,145	3,210	2,997
	Subject-Object Swap	30,533	2,944	2,787
	Verb Antonym	24,711	2,258	2,258
	Total SI	486215	42681	41714
SP	Comparative Synonym	13,302	1,066	1,163
	Paraphrasing	86,373	6,967	6,982
	Noun Synonym	212,904	18,570	18,968
	Number Substitution	60,582	5,194	4,994
	Pronoun Substitution	32,508	2,869	2,852
	Verb Synonym	78,103	7,314	6,919
	Total SP	483772	41980	41878

Table 7.2: Number of SISP-transformed samples generated per category for the NLVR2 dataset.

	Category	Training	Testing	Validation
SI	Original	76122	9600	9600
	Comparative Antonym	66,300	8,754	8,893
	Negation	249,836	31,634	31,923
	Noun Antonym	193,964	24,484	24,251
	Number Substitution	9,592	1,212	1,130
	Pronoun Substitution	156,466	19,785	20,100
	Subject-Object Swap	122,510	15,500	15,337
SP	Verb Antonym	49,802	6,358	6,356
	Total SI	848470	107727	107990
	Comparative Synonym	38,312	4,955	4,940
	Paraphrasing	76,122	9,600	9600
	Noun Synonym	418,285	52,857	52002
	Number Substitution	4,482	574	544
	Pronoun Substitution	91,125	11,464	11539
	Verb Synonym	196,826	25,044	25576
	Total SP	825152	104494	104201

Table 7.3: Number of SISP-transformed samples generated per category for the VIOLIN dataset.

	Category	Train	Trainval	Devval
SI	Original	92,761	38,374	5,323
	Comparative Antonym	18,839	8,044	1,139
	Negation	100,302	41,676	5,738
	Noun Antonym	82,885	34,543	4,835
	Number Substitution	1,505	730	95
	Pronoun Substitution	26,804	11,462	1,597
	Subject-Object Swap	11,793	4,999	683
SP	Verb Antonym	13,262	5,707	786
	Total SI	255390	107161	14873
	Comparative Synonym	21,259	9,037	1,271
	Paraphrasing	92,761	38,374	5,323
	Noun Synonym	119,301	49,977	6,850
	Number Substitution	1,443	678	95
	Pronoun Substitution	44,435	19,025	2,620
	Verb Synonym	45,612	19,384	2,622
	Total SP	324811	136475	18781

Table 7.4: Number of SISP-transformed samples generated per category for the VQA Yes-No dataset.

Method	NLVR2			VIOLIN		
	Clean	SP	SI	Clean	SP	SI
Data-Aug	51.07	50.92	40.74	61.12	62.78	62.15
SW-SDRO	51.14	50.97	40.75	62.78	58.13	64.78
GW-SDRO	51.07	50.92	40.73	62.15	52.79	74.98

Table 7.5: Text-only evaluation of biases due to SISP transformations. 50% indicates no bias.

do not have access to visual information. Table 7.5 shows that SP transforms inflict lesser bias on models than SI transforms. The effect of bias is dataset-specific; SI makes the prediction of NLVR² samples harder than random (less than 50% accuracy) but easier for VIOLIN.

Transformation Fidelity: We employ human subjects to evaluate the quality of SISP-transformed sentences on (1) correctness of labels, (2) grammar, (3) semantics, and (4) visual grounding. We report a unified average ‘transformation fidelity’. Fidelity is higher for SP samples than SI (90.50% v/s 79.51%), which resonates with the complexities of inversion of meaning ([Russell, 1905](#)) and leaves room for improvement in SI transformation. Although some level of ambiguity exists in SISP transforms, our results show that SDRO models benefit from transformed data.

For each of the 13 SISP categories, we sampled 100 SISP-transformed examples from NLVR², thus giving us a total of 1300 samples. We employed 10 human subjects to evaluate the quality of SISP-transformed sentences. These human subjects were all proficient in English and at the time of the study were enrolled in graduate programs in an English-speaking country. The subjects were shown samples with the original

images, sentences, and labels, as well as the new sentence and new label as shown in Figure 7.5. These subjects evaluated each sample with a binary (0/1) score, according to 4 metrics described below, along with an average “Transformation Fidelity”:

1. Label Correctness (LC) – *Is the new label correct for the new sentence?*
2. Grammatical Correctness (GC) – *Does the sentence appear to be grammatically correct?*
3. Visual Grounding (VG) – *Does the sentence refer to at-least one visual entity from the image?*
4. Semantic Correctness (SC) – *Is the sentence semantically sound and not absurd?*

The subjects were asked to view each sample and rate the new sentence and label on a binary scale for each of the four metrics. A snapshot of the interface used for the study as viewed by the human subjects is shown in Figure 7.5. Results are shown in Table 7.6 – split by the category of SISP transformation and in Table 7.7 – split by the ground-truth label of the original sample. Overall, our SISP transformed test set for NLVR² was rated at an average fidelity of 75.10%. It can be observed that on average, SP samples were rated to have higher average fidelity than SI samples, and False samples higher than True samples.

We also split the ratings (2 SISP categories and 2 labels: 2×2) and show results in Table 7.8. Overall, *SP(False)* has the highest average fidelity, and *SI(False)* has the lowest. LC (label correctness) for SI transformations of False statements is only 50%, probably because the inversion of a False statement using template-based methods may not always result in a True statement. On the other hand an SP transformation of a False statement remains False and got 100% LC. It is surprising to observe that LC for SP transformations of True statements is low. *SP(True)* received the highest

Category	Fidelity Metrics				
	LC	GC	VG	SC	Avg.
SP	73.33	80.00	96.67	70.00	80.00
SI	71.15	67.31	96.15	57.69	73.08
All	71.95	71.95	96.34	62.20	75.10

Table 7.6: Human validation of our SISP transforms split according to the category of transformation.

GT Label	Fidelity Metrics				
	LC	GC	VG	SC	Avg.
True	70.69	72.41	98.28	56.90	74.59
False	75.00	70.83	91.67	75.00	78.13
All	71.95	71.95	96.34	62.20	75.10

Table 7.7: Human validation of our SISP transforms split according to the GT label of the original sample.

GC and VG ratings, but low SC and LC ratings. VG ratings for all categories were consistently high.

7.3 Experiments

Datasets. For all datasets, given images/videos and natural language text as input, the system is expected to predict a binary class label. NLVR² ([Suhr et al., 2019](#)) contains $\sim 86K, 7K, 7K$ samples for training, development, and testing respectively. Each sample in NLVR² consists of a pair of images (from search engines) and a

GT Label	Fidelity Metrics				
	LC	GC	VG	SC	Avg.
SP(True)	55.55	83.33	100.0	66.67	76.39
SI(True)	77.50	67.50	97.50	52.50	73.75
SP(False)	100.0	75.00	91.67	75.00	85.42
SI(False)	50.00	66.67	91.67	75.00	70.83
All	71.95	71.95	96.34	62.20	75.10

Table 7.8: Human validation of our SISP transforms split according to the GT label of the original sample.

sentence (crowd-sourced). VIOLIN ([Liu et al., 2020a](#)) contains video clips from popular TV shows and movies along with subtitles and crowd-sourced statements. VIOLIN contains $76K$, $9.5K$, $9.5K$ samples for training, validation and testing. VQA Yes/No consists of image-question-answer triplets from VQA-v2 dataset ([Goyal et al., 2017](#)). While VQA-v2 consists of multiple question and answer types, we focus on the subset of questions with binary *yes/no* answers ($\sim 38\%$ of VQA-v2).

Evaluation Metrics. We use two evaluation metrics: (1) **Clean Accuracy**: accuracy on the i.i.d. benchmark test set, and (2) **SISP Accuracy**: average performance on SISP transformations of the test set. Since SISP transformations are automated and can be noisy (Sec 7.2.2), evaluation on the SISP test set can be considered a proxy for robustness.

³Notation: **bold**: higher than SOTA; shaded: higher than respective backbone; underlined: best SI/SP accuracies.

7.3.1 Results

We compare SDRO with backbone models that use standard training data (*BASE*) and data-augmentation (*+data-aug*). We train SDRO and backbones with the same hyperparameters. We apply test-time ensembling to the best SDRO model.

NLVR²: We use Transformer-based models LXMERT (Tan and Bansal, 2019), UNITER (Chen *et al.*, 2020b), and VILLA (Gan *et al.*, 2020) as backbones for SDRO. VILLA (the current state-of-the-art for NLVR²) uses standard adversarial training. The percentage of SISP-transformed samples is fixed at $T=20\%$. Table 7.9 shows results on the NLVR² test set, with consistent model-agnostic improvements in clean accuracy over each baseline model and improved robustness on average. Both variants of SDRO improve over VILLA_{BASE} by 0.84% and 1.02%, respectively. Test-time ensembling using Equation 7.9 leads to further gains, resulting in a new state-of-the-art accuracy of 82.22%, an improvement of 3.83% over VILLA_{BASE}. GW-SDRO results in the highest SI accuracy when used with each backbone model.

VIOLIN: We consider VIOLIN_{BASE} (Liu *et al.*, 2020a) and HERO (Li *et al.*, 2020a), the current state of the art, as baselines. VIOLIN_{BASE} separately computes visual features using Faster-RCNN (Ren *et al.*, 2015b) and textual features using BERT (Devlin *et al.*, 2019), and fuses them to be used as input to a classifier model. On the other-hand, HERO is a large-scale transformer-based pre-trained model which uses various V&L pre-training tasks to compute cross-modal features. We set $T=40\%$. The results can be seen in Table 7.10. SW-SDRO model with the HERO backbone improves the state-of-the-art to 68.83%, and test-time ensembling further improves it to 69.90%. Interestingly, similar improvements in clean accuracy are not observed for VIOLIN_{BASE}, potentially because it does not use cross-modal pre-trained features.

Model	Clean Acc.	SISP Acc.		
		SP	SI	Avg.
LXMERT _{BASE}	74.37	69.20	37.35	53.28
+ VILLA	75.98	69.94	39.09	56.15
+ data-aug	71.83	70.13	66.34	68.23
+ SW-SDRO	71.19	67.41	66.32	66.86
+ GW-SDRO	74.55	69.06	69.34	69.20
+ Ensemble	74.75	—”—	—”—	—”—
UNITER _{BASE}	77.85	72.73	34.86	53.80
+ data-aug	76.65	70.34	81.04	75.69
+ SW-SDRO	78.43	69.71	67.50	68.61
+ GW-SDRO	77.55	67.93	81.66	74.79
+ Ensemble	80.00	—”—	—”—	—”—
VILLA _{BASE}	78.39	<u>73.15</u>	34.15	53.65
+ data-aug	78.34	72.11	84.44	<u>77.77</u>
+ SW-SDRO	79.23	69.23	67.35	68.29
+ GW-SDRO	79.41	68.67	<u>84.54</u>	76.60
+ Ensemble	82.22	—”—	—”—	—”—

Table 7.9: Results on the NLVR² public test set.¹

Model	Clean Acc.	SISP Acc.		
		SP	SI	Avg.
VIOLIN _{BASE}	68.07	57.17	57.20	57.18
+ data-aug	61.58	<u>67.64</u>	67.70	67.67
+ SW-SDRO	62.81	62.84	62.68	62.76
+ GW-SDRO	63.71	64.58	63.16	63.87
+ Ensemble	66.56	—	—	—
HERO _{BASE}	68.55	65.59	32.00	48.80
+ data-aug	65.21	59.20	81.81	<u>70.51</u>
+ SW-SDRO	68.83	58.97	77.83	68.41
+ GW-SDRO	68.19	56.20	<u>82.92</u>	69.57
+ Ensemble	69.90	—	—	—

Table 7.10: Results on VIOLIN.¹

VQA Yes/No: We use UNITER and VILLA as the backbone models, with $T=20\%$. The motivation behind VQA experiments is to show that SISP transforms and SDRO can be extended to other V&L tasks. Table 7.11 shows that GW-SDRO is the best performing model in terms of clean accuracy, and is further improved by test-time ensembling.

7.3.2 Fine-Grained Results

Baseline Performance on SISP. In Tables 7.12, 7.13, 7.14 we compare the performance of baseline models on all 13 categories of SISP transforms. All baseline models are below random performance on all three datasets for all SI categories, except for VIOLIN_{BASE} (Liu *et al.*, 2020a). This is an interesting finding since VIOLIN_{BASE}

Model	Clean Acc.	SISP Acc.		
		SP	SI	Avg.
UNITER _{BASE}	83.49	72.04	38.90	55.47
+ data-aug	82.53	77.03	93.70	85.36
+ SW-SDRO	83.92	75.82	88.92	81.48
+ GW-SDRO	84.05	76.95	93.41	85.18
+ Ensemble	84.22	—	—	—
VILLA _{BASE}	84.82	74.15	37.40	55.77
+ data-aug	83.54	78.33	94.55	86.45
+ SW-SDRO	84.54	74.02	88.32	81.17
+ GW-SDRO	85.12	77.92	93.42	85.67
+ Ensemble	85.37	—	—	—

Table 7.11: Results on the VQA yes/no subset.¹ Not to be compared with VQA-v2 leaderboard since we use a smaller training set of *yes/no* questions.

is the only model that is not a pretrained transformer-based model, but uses simple fusion of visual and textual modalities. In this paper, we've considered 3 benchmarks, and $3 + 2 + 3 = 8$ backbone models in total. Of these, only VIOLIN_{BASE}— a non-transformer model, retains above-random performance on SISP samples. Performance on SP categories is the best for VILLA ([Gan et al., 2020](#)) for NLVR² and VQA Yes/No, and HERO ([Li et al., 2020a](#)) for VIOLIN.

SDRO Performance on SISP. In Tables 7.15, 7.16, 7.17 we compare performance for the state-of-the-art model VILLA, as well as models trained with naive data augmentation and our SDRO methods.

Category	LXMERT	UNITER	VILLA
Original	74.37	77.85	78.39
SI	Comparative Antonym	49.19	40.11
	Negation	35.19	36.92
	Noun Antonym	29.94	35.35
	Number Substitution	45.26	39.53
	Pronoun Substitution	47.76	34.79
	Subject-Object Swap	20.26	27.65
SP	Verb Antonym	27.86	29.72
	Comparative Synonym	61.35	65.58
	Paraphrasing	71.33	73.62
	Noun Synonym	71.24	75.32
	Number Substitution	70.68	74.33
	Pronoun Substitution	69.36	73.36
SV	Verb Synonym	71.26	74.16
	Comparative Antonym	49.19	40.11
	Negation	35.19	36.92
	Noun Antonym	29.94	35.35
	Number Substitution	45.26	39.53
	Pronoun Substitution	47.76	34.79

Table 7.12: Evaluation of NLVR2 baselines on SISP test samples.

7.4 Analysis

7.4.1 Visualization of Perturbations

In order to quantify the diverse and larger semantic transformations compared to additive perturbations, we study the tSNE (Van der Maaten and Hinton, 2008) embeddings of (i) original samples from NLVR² (P), (ii) their SISP-transformed versions (P_{SISP}), and (iii) their adversarially perturbed versions (P_{adv}). Input sentences are encoded using the UNITER text encoder for (i) and (ii), and the adversarial

Category	VIOLIN _{BASE}	HERO
Original	68.07	68.55
SI	Comparative Antonym	58.33
	Negation	57.75
	Noun Antonym	57.21
	Number Substitution	54.21
	Pronoun Substitution	57.66
	Subject-Object Swap	57.59
SP	Verb Antonym	57.68
	Comparative Synonym	57.92
	Paraphrasing	57.32
	Noun Synonym	57.67
	Number Substitution	54.87
	Pronoun Substitution	57.68
	Verb Synonym	57.53
		67.09

Table 7.13: Evaluation of VIOLIN baselines on SISP test samples.

perturbation mechanism ([Gan et al., 2020](#)) for (iii). 3D tSNE embeddings are visualized in Figure 7.6; SISP transformed sentences (blue) are farther away than the perturbed versions. This shift is quantified by the KL-divergence ([Kullback et al., 1951](#)) between the distributions, with $D_{KL}(P_{SISP}||P) > D_{KL}(P_{adv}||P)$ implying that the diversity of SISP transformations is higher.

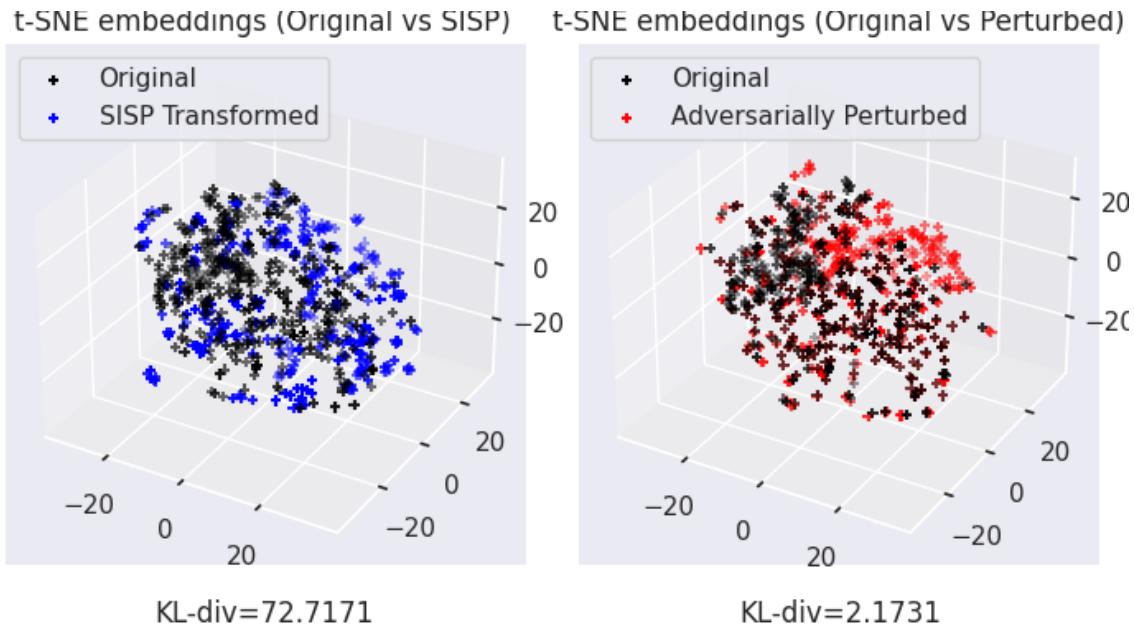


Figure 7.6: Comparison of original sentences (black) with (*left*) SISP-transformed sentences (blue) and (*right*) ϵ -bounded perturbations as a tSNE plot.

		An adult gorilla is opening its mouth wide, <u>uncovering</u> a mouthful of teeth in one image	An adult gorilla is <u>not</u> opening its mouth wide revealing a mouthful of teeth in one image.		
BASELINE	FALSE	0.9999	BASELINE	FALSE	0.9999
DATA-AUG	FALSE	0.9996	DATA-AUG	FALSE	0.7797
SW-SDRO	FALSE	0.9941	SW-SDRO	TRUE	0.9858
GW-SDRO	FALSE	0.9918	GW-SDRO	TRUE	0.9596
		At least one dog is next to <u>a cage</u> .	<u>Something</u> is next to a <u>not</u> caged area.		
BASELINE	FALSE	0.9999	BASELINE	FALSE	0.9997
DATA-AUG	FALSE	0.9999	DATA-AUG	TRUE	0.9977
SW-SDRO	FALSE	0.9985	SW-SDRO	TRUE	0.9601
GW-SDRO	FALSE	0.9909	GW-SDRO	TRUE	0.9885

Figure 7.7: Original test inputs for NLVR² with their respective SP (green) and SI (yellow) test samples and the prediction and confidence of models with VILLA backbone. Wrong predictions are highlighted in red.

Category	LXMERT	UNITER	VILLA
Original	83.13	83.655	84.82
SI	Comparative Antonym	36.7	39.07
	Negation	29.59	31.93
	Noun Antonym	48.36	53.21
	Number Substitution	26.32	42.11
	Pronoun Substitution	21.28	24.05
	Subject-Object Swap	24.68	31.33
SP	Verb Antonym	35.88	50.63
	Comparative Synonym	67.72	71.28
	Paraphrasing	79.63	79.37
	Noun Synonym	74.09	73.37
	Number Substitution	72.32	57.89
	Pronoun Substitution	74.82	76.11
SS	Verb Synonym	73.76	74.22
	Comparative Antonym	36.7	39.07
	Negation	29.59	31.93
	Noun Antonym	48.36	53.21
	Number Substitution	26.32	42.11
	Pronoun Substitution	21.28	24.05

Table 7.14: Evaluation of VQA Yes/No baselines on SISP test samples.

7.4.2 Comparison of Model Calibration

Figure 7.7 contains qualitative examples from NLVR² to compare output probabilities. We observe that SDRO models have higher clean accuracy, but lower confidence in the predictions than baseline and *data-aug* methods.

Reliability Diagrams. To validate this observation at scale, we use reliability diagrams to visualize model calibration ([Niculescu-Mizil and Caruana, 2005](#)), and

Category	VILLA	+ Dataaug	+ SW-SDRO	+ GW-SDRO
Original	78.39	78.34	79.23	79.41
SI	Noun Antonym	39.05	85.79	63.13
	Negation	35.39	65.75	72.78
	Subject-Object Swap	30.41	87.13	60.19
	Verb Antonym	34.89	72.58	55.18
	Number Substitution	35.24	95.79	75.79
	Pronoun Substitution	29.78	98.44	81.31
	Comparative Antonym	34.32	78.62	63.11
SP	Pronoun Substitution	73.16	72.81	64.91
	Number Substitution	74.37	81.27	77.63
	Comparative Synonym	66.88	64.63	64.16
	Verb Synonym	75.24	69.88	65.78
	Paraphrasing	73.46	74.89	75.74
	Noun Synonym	75.78	69.15	67.67
				61.64

Table 7.15: Evaluation of SDRO models NLVR² SISP test samples.

plot model accuracy as a function of confidence. We use the softmax probability \hat{p} of the predicted class as model confidence, split the range of probabilities into $M = 20$ equal-sized bins, and calculate bin accuracy $acc(B_m)$ and bin confidence $conf(B_m)$ (Guo *et al.*, 2017). If B_m is the set of all samples that fall in the m^{th} bin,

$$acc(B_m) \triangleq \frac{1}{|B_m|} \sum_{X_i \in B_m} \mathbb{1}(\hat{y}_i = y_i), \quad (7.10)$$

$$conf(B_m) \triangleq \frac{1}{|B_m|} \sum_{X_i \in B_m} \hat{p}_i. \quad (7.11)$$

Category	HERO	+ Dataaug	+ SW-SDRO	+ GW-SDRO
Original	68.55	65.21	68.83	68.19
SI	Noun Antonym	37.06	86.38	76.61
	Negation	34.73	53.18	58.31
	Subject-Object Swap	31.13	94.28	74.98
	Verb Antonym	38.77	81.96	77.05
	Number Substitution	26.07	76.32	80.03
	Pronoun Substitution	24.64	99.41	92.89
	Comparative Antonym	31.66	81.12	84.44
SP	Pronoun Substitution	66.74	62.29	60.76
	Number Substitution	58.88	56.62	57.14
	Comparative Synonym	67.87	58.31	57.64
	Verb Synonym	67.09	56.42	56.49
	Paraphrasing	65.81	63.59	65.22
	Noun Synonym	67.15	57.99	56.67
				50.47

Table 7.16: Evaluation of SDRO models on VIOLIN SISP Data

A model with perfect calibration should have a reliability diagram such that

$$acc(B_m) = conf(B_m)$$

. We also report Expected Calibration Error (Naeini *et al.*, 2015) over all n test samples:

$$ECE = \sum_{m=1}^M \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|. \quad (7.12)$$

Reliability diagrams and corresponding ECE values for the baseline VILLA trained with naive data augmentation and SDRO methods for NLVR² are shown in Figure 7.8. On both the clean test set and SISP test set, SDRO models have the lowest ECE.

Category	VILLA	+ Dataaug	+ SW-SDRO	+ GW-SDRO
Original	84.82	83.54	84.88	85.19
SI	Noun Antonym	50.88	97.85	92.04
	Negation	29.59	80.81	82.36
	Subject-Object Swap	26.06	98.83	96.19
	Verb Antonym	41.86	97.71	88.17
	Number Substitution	49.47	94.74	78.95
	Pronoun Substitution	24.36	95.36	90.86
	Comparative Antonym	39.59	96.58	89.64
SP	Pronoun Substitution	77.48	77.41	75.88
	Number Substitution	62.11	77.89	56.84
	Comparative Synonym	74.11	80.72	78.63
	Verb Synonym	75.82	76.85	76.13
	Paraphrasing	80.74	80.57	81.31
	Noun Synonym	74.61	76.55	75.27
				72.23

Table 7.17: Evaluation of SDRO models on VQA Yes/No SISP Data

While the ECE for SDRO is marginally better than data augmentation for the clean test set, SDRO is much better calibrated for the SISP test set, with SW-SDRO closest to ideal calibration.

7.4.3 Size of Training Dataset

We evaluate models trained on small subsets of the original dataset, and compare their performance in Figure 7.9. SDRO models are significantly better at all sizes of training datasets as shown by accuracy and AUC (area under the curve). Notably, SDRO models trained with only 10% ($\sim 8.6K$) samples have performances similar to

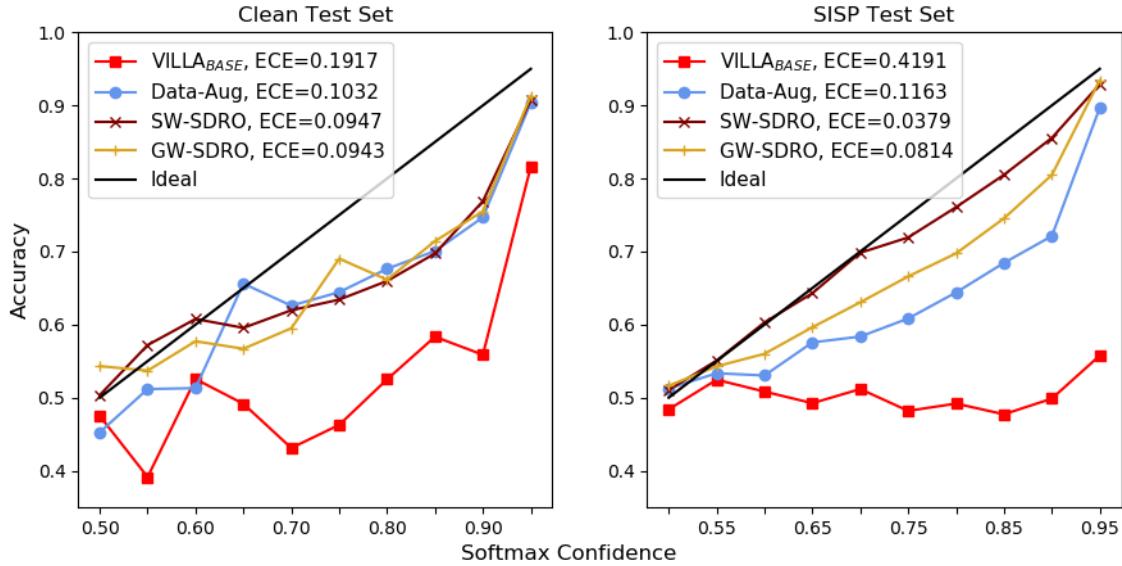


Figure 7.8: Comparison of reliability curves on the clean test set (*left*) and SISP test set (*right*).

the baseline trained with 30% samples; SDRO models with 20% data are better than the baseline model with 40% data. While models trained with naive augmentation saturate below SOTA, at $\sim 80\%$ data size, SDRO models cross the existing SOTA of 78.39%.

7.4.4 Ablation Studies

Proportion of Augmented Samples. The final dataset has the same size as the original training set, but with $T\%$ transformed samples and $(100-T)\%$ original samples. The effect of this hyperparameter T is reported in Figure 7.10 as a percentage improvement of accuracy w.r.t. VILLA_{BASE}. An optimal value of $T=20\%$ leads to improvements in clean accuracy, but a larger proportion of augmented samples degrades performance. Similarly, higher T leads to higher robust accuracy, pointing to a trade-off between clean accuracy and robust accuracy at values of T higher than the optimal. This conforms with similar findings from [Tsipras *et al.* \(2019\)](#). While

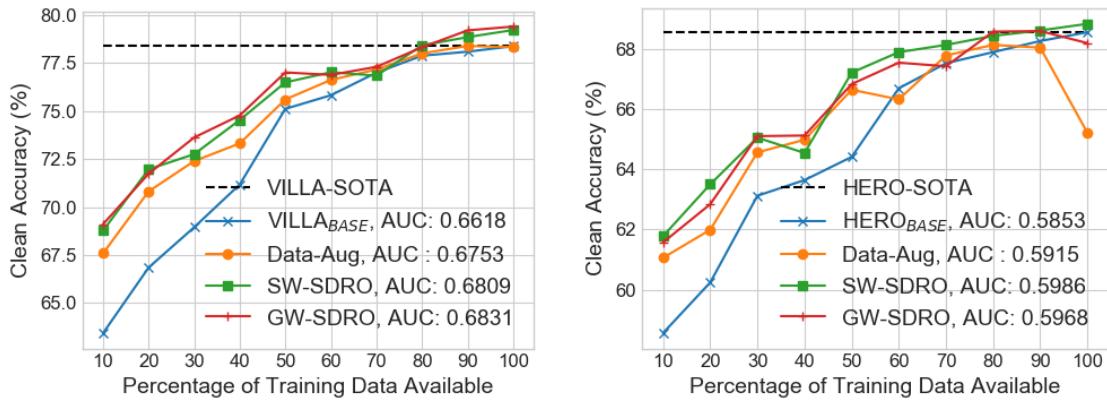


Figure 7.9: Effect of size of training data (*left*) NLVR², (*right*) VIOLIN. SDRO models are consistently better than baselines, even in low-data settings.

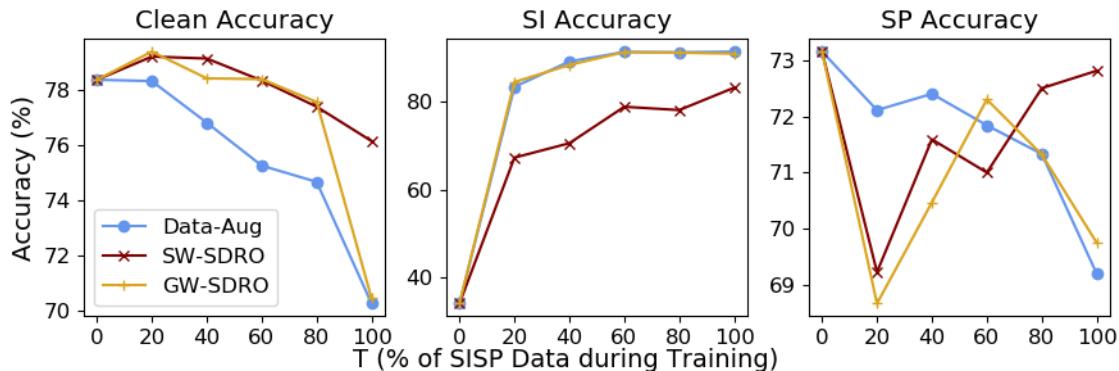


Figure 7.10: Plots showing the effect of the percentage of augmented samples on Clean, SP, and SI accuracies on NLVR², when using data-augmentation, and SDRO.

models trained with naive data-augmentation have better SISP accuracy than SDRO models as in Table 7.9, they do so by sacrificing clean accuracy, while SDRO models improve along both dimensions compared to the baselines.

Contributions of SI and SP independently: We analyze which of the two categories (semantics-inverting (SI) or semantics-preserving (SP)) is the most effective by performing SDRO with only SI transforms, or with only SP transforms, and when

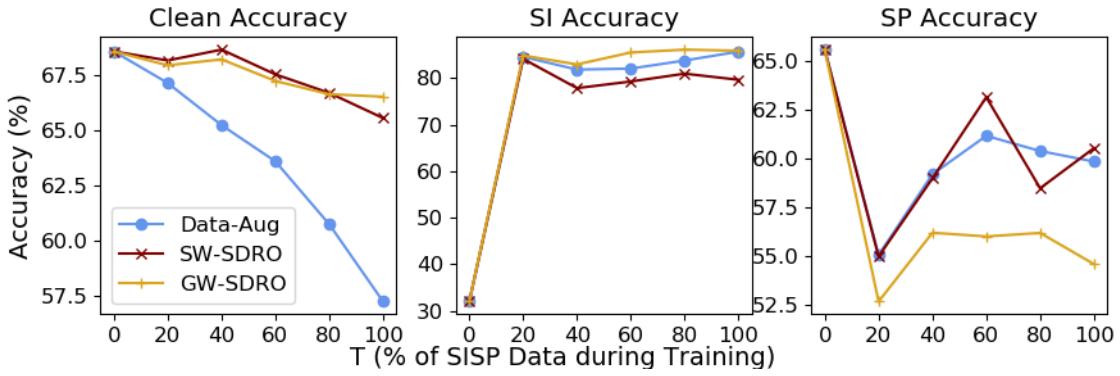


Figure 7.11: Plots showing the effect of the percentage of augmented samples on Clean, SP, and SI accuracies on VIOLIN, when using naive data-augmentation, SW-SDRO, and GW-SDRO.

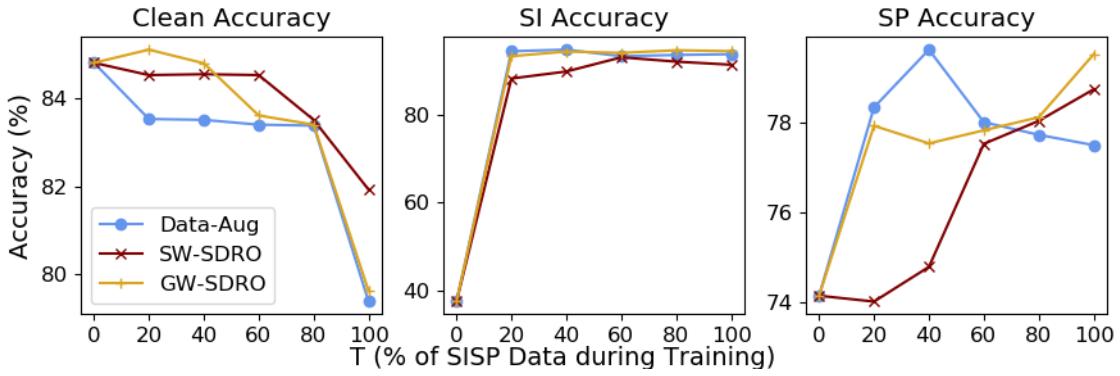


Figure 7.12: Plots showing the effect of the percentage of augmented samples on Clean, SP, and SI accuracies on VQA Yes/No, when using naive data-augmentation, SW-SDRO, and GW-SDRO.

Model	SP only			SI Only			Both		
	Clean	SP	SI	Clean	SP	SI	Clean	SP	SI
Data-Aug	76.07	74.89	35.77	69.51	53.68	94.89	78.34	72.11	84.44
SW-SDRO	79.79	76.93	30.72	79.27	55.53	88.76	79.23	69.23	67.35
GW-SDRO	79.46	75.72	33.04	79.13	54.31	93.25	79.41	68.67	84.54

Table 7.18: Comparison of performance when only SP, only SI, or both types of transformations are performed.

Model	SISP (Pos)			SISP (All)		
	Clean	SP	SI	Clean	SP	SI
Data-Aug	78.23	68.02	57.48	78.34	72.11	84.44
SW-SDRO	78.81	62.06	66.07	79.23	69.23	67.35
GW-SDRO	79.10	63.47	62.29	79.41	68.67	84.54

Table 7.19: Comparison of performance if only positive samples are used as inputs for SISP transformations.

using both. Table 7.18 shows that SDRO models trained only with SI suffer in terms of SP robustness and vice versa. However, there is still an increase in clean accuracy in both cases. This indicates that both SI and SP contribute towards improvements in robustness and clean accuracy.

Transformations of only True statements: Transforming *False* (negative) statements can lead to ambiguous and subjective meanings ([Russell, 1905](#)). We investigate if transforming only *True* (positive) statements is better than transforming both *True* and *False* statements. Table 7.19 shows that SISP transformations of both types of

	Model	CR	CS	CL	EDA	Emb	WN	Avg.
NLVR ²	VILLA	77.5	74.4	<u>74.4</u>	69.6	75.5	75.9	74.5
	+ SDRO	<u>78.5</u>	<u>77.2</u>	72.1	<u>71.1</u>	75.8	<u>76.4</u>	75.2
VIOLIN	HERO	66.1	63.0	68.6	60.9	63.8	63.4	64.3
	+ SDRO	68.7	65.0	69.0	61.3	65.5	64.6	65.7
VQA Yes/No	VILLA	80.5	75.7	84.9	74.6	78.6	76.4	78.5
	+ SDRO	86.0	84.5	84.1	87.0	84.3	84.0	85.00

Table 7.20: Performance on “text-attack” of NLI test set.

statements lead to higher clean accuracy and robustness.

7.4.5 Robustness to Text-Attacks

We utilize automated adversarial attack recipes (Morris *et al.*, 2020): CLARE (CR) (Li *et al.*, 2021a), character-swap (CS) (Pruthi *et al.*, 2019), Checklist (CL) (Ribeiro *et al.*, 2020), EDA (Wei and Zou, 2019), counter-fitted embeddings (Emb) (Alzantot *et al.*, 2018a), and WordNet-based swap (Ren *et al.*, 2019). Table 7.20 shows results using the best backbone and our SDRO model. On NLVR², VILLA+SDRO is better than VILLA for 4 out of 6 attack categories, and 0.65% on average. On VIOLIN, HERO+SDRO outperforms the baseline on all attack categories, leading to an average gain of 1.39%. On VQA-Yes/No, VILLA+SDRO outperforms the baseline on all attack categories, and 6.54% on average.

7.5 Related Work

Adversarial Training (AT) has been studied under a game-theoretic (Dalvi *et al.*, 2004) and min-max setup (Madry *et al.*, 2018a). Volpi *et al.* (2018) use AT

to adversarially augment image classification datasets and show improved domain generalization for digit classification. Wong and Kolter (2020); Gokhale *et al.* (2020a) modify AT for real-world adversaries beyond norm-bounded perturbations. AT has been used for text classification with LSTMs (Miyato *et al.*, 2017) and for pretraining transformer-based models by adding label-preserving adversarial perturbations to embeddings of word tokens (Zhu *et al.*, 2020; Jiang *et al.*, 2020; Gan *et al.*, 2020). Contrastive examples have been explored, collected from humans (Agrawal *et al.*, 2018), negative mining (Shi *et al.*, 2018), or synthetic generation (Agarwal *et al.*, 2020; Chen *et al.*, 2020a; Gokhale *et al.*, 2020b; Teney *et al.*, 2020a).

Robustness in V&L has been explored for VQA, such as performance under prior probability shift (Agrawal *et al.*, 2018) and domain adaptation (Chao *et al.*, 2018; Xu *et al.*, 2020a), along with robustness for implied questions (Ribeiro *et al.*, 2019b) and novel compositions (Johnson *et al.*, 2017; Agrawal *et al.*, 2017), and robustness to logical connectives (including negation) Gokhale *et al.* (2020c). Teney *et al.* (2020a) have shown that many V&L, image classification, and sentiment analysis models are sensitive to image editing. There has been a recent effort of model-in-the-loop dataset collection to guide humans to create harder VQA samples (Li *et al.*, 2021b; Sheng *et al.*, 2021).

Robustness in NLP: Generation of SP adversarial examples (Jia and Liang, 2017a; Ribeiro *et al.*, 2018a; Iyyer *et al.*, 2018a; Alzantot *et al.*, 2018a), and approaches to defend against word substitution (Jia *et al.*, 2019a) have been explored. Evaluation datasets have also been proposed for textual entailment that are manually crafted (Gardner *et al.*, 2020a) or template-based (McCoy *et al.*, 2019a; Glockner *et al.*, 2018a; Naik *et al.*, 2018a). Our method uses automated linguistically-informed SI and SP transforms for both training and inference.

7.6 Discussion

On Ensembling Coefficients. While designing our ensembling approach, we used $\alpha = 0.5$, i.e., equal contribution from the original output and the average of all outputs for transformed samples. This choice is generic and does not rely on dataset- or model-specific characteristics of SISP accuracy. While treating α as a hyperparameter and tuning it on validation datasets could lead to further gains, our intuitive choice of $\alpha=0.5$ is effective by itself.

On SI Samples. Tables 7.9, 7.10, 7.11 show that existing models perform well on SP transforms, implying that equivalent semantics are captured in transformer-based models. However, these models fail on SI samples, and thus the average SISP accuracy is close to random (50%). When images are perturbed with noise, blur, weather, or digital artifacts ([Hendrycks and Dietterich, 2019](#)), they retain semantics – an image of a “cat” remains a cat after perturbation. However, for text inputs, minimal changes to the sample, such as a single word changing from “sitting” to “standing” or “not sitting”, inflict large changes in meaning. We hope that future work on design of V&L evaluation criterion along the SI axis, could benefit from our findings. While we generated SI and SP text for VLI tasks, the idea could be extended to design SISP transformations for images, by operating at object-level instead of pixel-level

On combination of AT and SDRO. We show that combining AT with SDRO can improve VLI performance and incorporate domain knowledge into the training process, such as semantic knowledge that often exists in natural language or linguistic rules. This is explicitly observed with VILLA, which is pre-trained and fine-tuned using standard adversarial training ([Gan et al., 2020](#)). When fine-tuned with SDRO, VILLA+SDRO further improves compared to UNITER+SDRO. The combination

of standard adversarial training, (which accounts for local adversaries inside a ϵ norm-ball) and SDRO, (which accounts for linguistic adversaries and contrastive examples, typically outside the norm-ball as shown in Figure 7.2) could lead to improved generalization in many other V&L tasks.

On differentiability. Linguistic transforms are not differentiable and prohibit gradient-based solutions to the inner maximization in SDRO. Nevertheless, we show that explicitly choosing the *argmax* over a pre-defined set of transformations does indeed lead to model-agnostic improvements for multiple V&L tasks. This is a first step towards incorporating domain knowledge into adversarial training. More sophisticated methods may emerge in the future to address non-differentiability by leveraging proximal point or trust-region methods (Eckstein, 1993; Conn *et al.*, 2000) or Interval Bound Propagation ([Dvijotham *et al.*, 2018](#); [Jia *et al.*, 2019a](#)).

Chapter 8

BEYOND VISION-LANGUAGE ALIGNMENT: ENHANCING VIDEO CAPTIONING VIA COMMONSENSE DESCRIPTIONS

Captioning is a crucial and challenging task for video understanding. In videos that involve active agents such as humans, the agent’s actions can bring about myriad changes in the scene. Observable changes such as movements, manipulations, and transformations of the objects in the scene, are reflected in conventional video captioning. Unlike images, actions in videos are also inherently linked to social aspects such as intentions (why the action is taking place), effects (what changes due to the action), and attributes that describe the agent. Thus for video understanding, such as when captioning videos or when answering questions about videos, one must have an understanding of these commonsense aspects. We present the first work on generating *commonsense* captions directly from videos, to describe latent aspects such as intentions, effects, and attributes. We present a new dataset “Video-to-Commonsense (V2C)” that contains $\sim 9k$ videos of human agents performing various actions, annotated with 3 types of commonsense descriptions. Additionally we explore the use of open-ended video-based commonsense question answering (V2C-QA) as a way to enrich our captions. Both the generation task and the QA task can be used to enrich video captions.

8.1 Introduction

When humans watch videos they can typically understand and reason about various aspects of the scene beyond the visible objects and actions. This involves understanding that some objects are *active agents* that not only perform actions and



Conventional Caption	Group of runners get prepared to run a race.
Commonsense-Enriched Caption	In order to win a medal , a group of runners get prepared to run a race. As a result they are congratulated at the finish line . They are athletic .
Commonsense Question Answering	What happens next to the runners? { Are congratulated at the finish line become tired

Figure 8.1: Comparison of conventional video captioning with our commonsense-enriched captioning. Our captions describe intention behind the action (**red**), attribute of the agent (**blue**), and effect of the action on the agent (**green**).

manipulate objects, but are motivated by intentions, have pre-conditions, and that their actions have an effect on the world and their own mental states. For instance, in analyzing the video clip in Figure 8.1, humans employ various capabilities such as perception, reasoning, inference, and speculation, to come up with a description for the observable sequence of events, but also reason about latent aspects such as the intention of the group of runners “*to win the medal*”, the effect of being “*congratulated at the finish line*”, and the attribute “*athletic*”.

The above example also illustrates that recognition of objects, actions, and events is often not enough; understanding causal relationships, social interactions, and commonsense aspects behind them provides context and a more semantic interpretation of the video (Gupta *et al.*, 2009). A model that can provide such detailed interpretations facilitates answering inferential questions, such as “*Will the player get angry later?*”. However, existing visual understanding systems are unable to perform such tasks that require speculative reasoning. A critical missing element in complex video understanding is the capability of performing commonsense inference, especially a

generative model. Existing efforts seek to find textual explanations or intentions of human activities as a classification task ([Vondrick *et al.*, 2016](#)) or a vision-to-text alignment problem ([Zhu *et al.*, 2015](#)).

In this paper we propose the **V**ideo to **C**ommonsense (V2C) framework to generate visually grounded commonsense descriptions about the underlying event in the video, enriching the factual description provided by a caption. Under this framework a system is expected to generate captions as well as three types of commonsense descriptions (intention, effect, attribute) directly from an input video. The V2C model can also be used as a building block for downstream tasks such as video question answering for questions requiring commonsense. Inspired by ([Bosselut *et al.*, 2019](#)), our model – the “V2C-Transformer” utilizes: (1) a video encoder to extract global representations of the video, (2) a transformer decoder that generates captions and commonsense descriptions, and (3) a cross-modal self-attention module that exploits joint visual-textual embeddings.

We curate the V2C dataset for training and benchmarking models on this task. We adopt the MSR-VTT video description dataset ([Xu *et al.*, 2016](#)) as a source of videos and captions. We first utilize the ATOMIC machine commonsense dataset ([Sap *et al.*, 2019a](#)) to get a list of candidate commonsense texts (intentions, effects, and attributes), and rank these using a BERT-based ([Devlin *et al.*, 2019](#)) model. Since these candidates are retrieved without using the video and may not be accurate, we instruct humans to watch the videos and select, remove, or rewrite the texts retrieved from ATOMIC. The text retrieved by ATOMIC helps our human annotators to understand the format of desired annotations, and also gives them a list of suggestions. The human component in our annotation procedure makes our data visually grounded and relevant, linguistically diverse, and natural.

We additionally explore the use of our V2C-Transformer architecture for a open-

ended video question answering task, where the questions are about commonsense aspects from the video. For this, we create a QA addendum of the V2C dataset called V2C-QA. By asking questions about the latent aspects in the video, our models are able to enrich caption generation with three specific types of commonsense knowledge.

Our contributions are summarized below:

1. We formulate the “V2C” task for enriching video captioning by generating descriptions of commonsense aspects.
2. We curate a video dataset annotated with captions and commonsense descriptions.
3. We present our V2C-Transformer architecture that generates relevant commonsense descriptions, and serves as a strong baseline.
4. We pose V2C as a video question answering task and show that it can assist commonsense caption generation.

8.2 Video to Commonsense (V2C)

Problem Formulation: Consider a video V consisting of N_v frames described by sentence S . Our Video-to-Commonsense (V2C) framework can be used for generating commonsense descriptions C under two settings. In the first setting (**V2C-Completion**), we use ground-truth captions to guide commonsense-enriched caption generation. This task can be viewed as providing supplementary explanations to the caption. In the second setting (**V2C-Generation**), we first learn to generate captions from videos, $\mathbf{g}(V)$, and then use them to generate commonsense descriptions.

$$\begin{aligned} \textbf{V2C-Completion} \quad C &= \mathbf{f}(V, S). \\ \textbf{V2C-Generation} \quad C &= \mathbf{f}(V, \mathbf{g}(V)). \end{aligned} \tag{8.1}$$

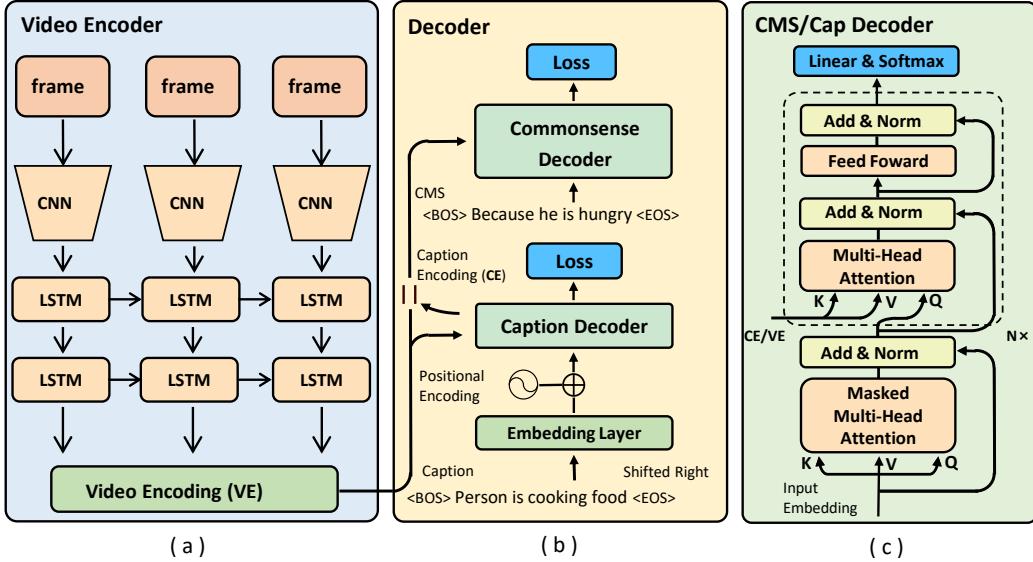


Figure 8.2: The V2C-Transformer model architecture contains: (a) Video Encoder designed to take video frames as input and encode them into frame-wise representations, (b) Decoder module consisting of a Caption Decoder and a Commonsense Decoder, and (c) Transformer Decoder module containing a stack of N consecutive transformer blocks (shown inside the dashed area).

8.2.1 V2C-Transformer

The proposed Video2Commonsense Transformer is a cross-modal model that generates captions and commonsense-enriched descriptions from videos. Our approach (Figure 8.2) adopts the “encoder-decoder” design: a video encoder that extracts global representations of the input video, and a transformer decoder that produces relevant commonsense knowledge along with captions.

Video Encoder: We obtain per-frame ResNet-152 (He *et al.*, 2016) features for video V and process them using an LSTM model (Sundermeyer *et al.*, 2012), a standard architecture for modeling long temporal sequences, and use the last hidden states of the LSTM as the video representations. We concatenate all previous hidden

states from each LSTM module as a final global video encoding \mathbf{v} , to provide the model with explicit context using the temporal attention mechanism.

Decoder: The video encoding is used as input to two decoder networks that use a transformer language model (Radford *et al.*, 2018) to generate a caption and commonsense description, using an inference mechanism similar to Bosselut *et al.* (2019). Our model is a two-stage process that first predicts the current events directly from videos, and then produces the corresponding commonsense captions. During training, the caption decoder \mathbf{D}_{CAP} takes the video encoding (\mathbf{v}) and ground truth caption (\mathbf{s}) as input to generate caption encoding ($\hat{\mathbf{s}}$), while the commonsense decoder \mathbf{D}_{CMS} uses the concatenation of video and caption encoding to obtain the commonsense description (\mathbf{c}), as shown in Figure 8.1 (b). This arrangement enables the attention module in commonsense decoder to attend to both the video and caption context.

$$\hat{\mathbf{s}} = \mathbf{D}_{\text{CAP}}(\mathbf{v}, \mathbf{s}), \quad \mathbf{c} = \mathbf{D}_{\text{CMS}}(\mathbf{v}, \hat{\mathbf{s}}). \quad (8.2)$$

Transformer Decoder is composed of a stack of transformer blocks (dashed area in (c) Figure 8.2), whose main component is a self-attention architecture. It takes as input the summation of word embedding and the positional encoding offset by 1 position through masked multi-head attention, which prevents the future words been seen. In our model, we deploy two stacked decoder architectures for both caption decoding and commonsense knowledge decoding. The Transformer Block consists of consecutive linear transformation: a multi-head attention module (denoted as $\mathcal{H}_{\text{M-ATT}}$), a two-layer feed forward network (\mathcal{H}_{FFN}), a layer normalization operation, and a residual connection.

Multi-head Attention module To enable our transformer decoder to generate commonsense descriptions by using both the visual and textual content, we modify

the multi-head attention module (which acts as the basic unit in recent transformer based language generation models (Radford *et al.*, 2018, 2019)) as a cross-modal module. $\mathcal{H}_{\text{M-ATT}}$ takes the input of the embedding of key (K), value (V) and query (Q). The key and value in transformer block are the video encoding (caption decoder) or concatenation of video/caption encoding (commonsense decoder), while the query is the output from the previous transformer block. In the masked multi-head attention module, K, V and Q are the identical vectors of input embedding. For a self-attention block with h heads,

$$\mathcal{H}_{\text{M-ATT}}(\mathbf{K}, \mathbf{V}, \mathbf{Q}) = \mathcal{H}_{\text{FFN}}([x_1, \dots, x_h]), \quad (8.3)$$

where x_i is computed by scaled dot-product attention operation, for head-index i , key-dimension $d_k n$, and transformation parameters \mathbf{w}_i .

$$\begin{aligned} \text{for } \mathbf{D}_{\text{CAP}}, \quad x_i &= \text{SOFTMAX}\left(\frac{\mathbf{w}_i^Q \mathbf{Q} \cdot \mathbf{w}_i^K \mathbf{K}'}{\sqrt{d_k}}\right) \mathbf{w}_i^V \mathbf{V}, \\ \text{for } \mathbf{D}_{\text{CMS}}, \quad x_i &= \text{SOFTMAX}\left(\frac{\mathbf{w}_i^Q [\mathbf{v}, \mathbf{s}] \cdot \mathbf{w}_i^K [\mathbf{v}, \mathbf{s}]'}{\sqrt{d_k}}\right) \mathbf{w}_i^V \mathbf{V}. \end{aligned}$$

8.3 The V2C Dataset

For the V2C task we need video clips annotated with commonsense descriptions about the agents in the video, as shown in Figure 8.1. While there are video captioning datasets such as MSR-VTT (Xu *et al.*, 2016), the captions in these datasets describe only the observable objects in the image, but do not describe latent and commonsense aspects. We are the first to curate such a dataset with annotations describing the intention of agent to perform an action, the effect of the action and the attribute of the agent given the action.

MSR-VTT contains around 10k videos each 10 to 30 seconds long, belonging to 20 categories covering a variety of topics such as sports, music, news, and home

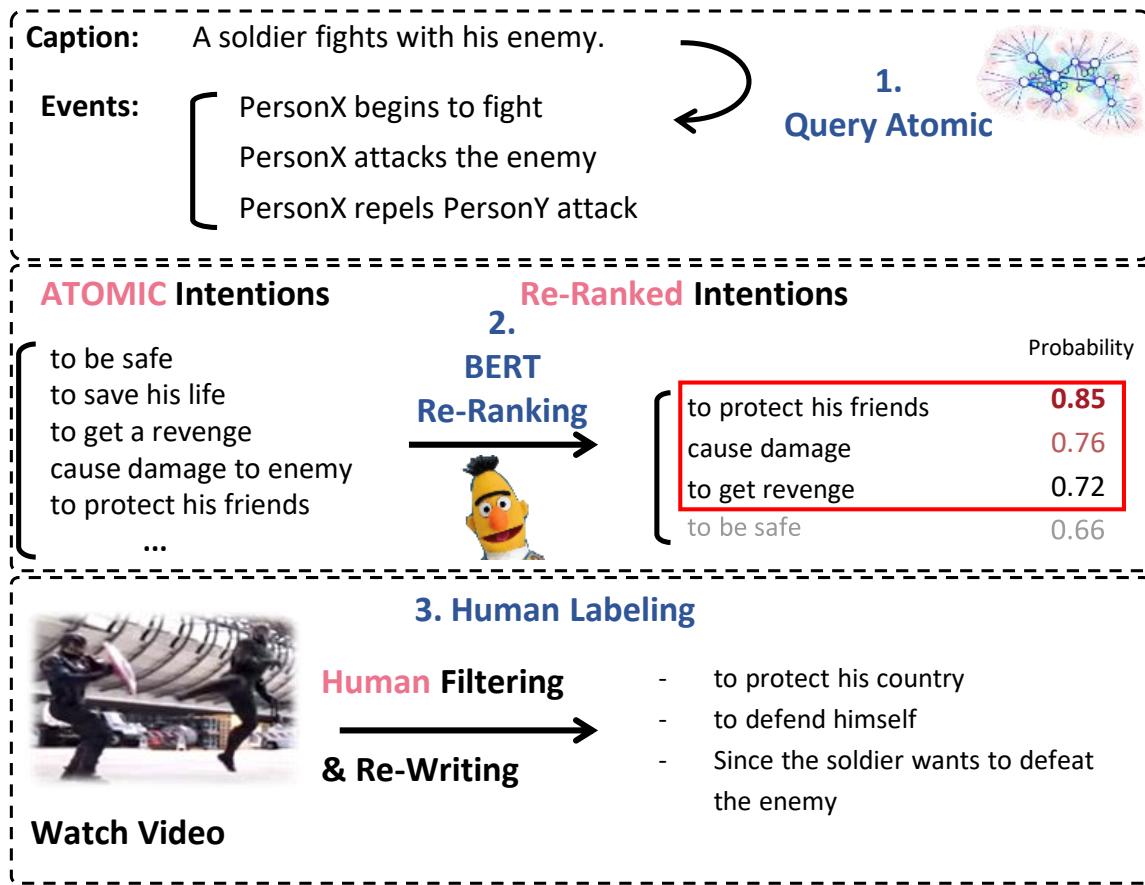


Figure 8.3: The overall three-step pipeline (retrieval from ATOMIC, BERT re-ranking, and human labeling) to construct our V2C dataset.

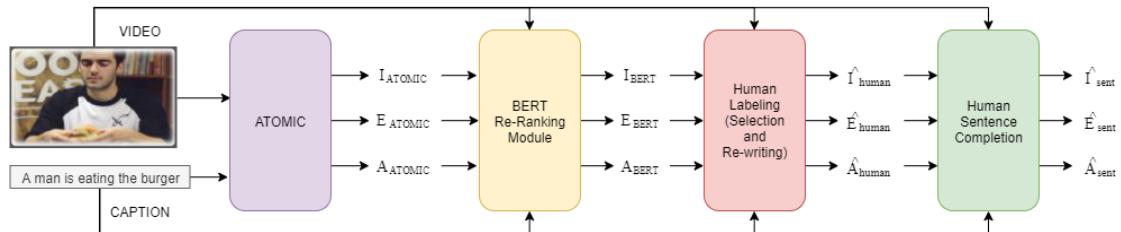


Figure 8.4: The data creation flow for V2C. We use the retrieved videos and captions from MSR-VTT and use the BERT re-ranking module to obtain a list of top-3 intentions (I), effects (E), and attributes (A). These are then further improved by human labeling. A subset of annotations is also converted to full sentences by human annotators.

videos. Each video is accompanied by 20 human-annotated textual descriptions on average. For training and benchmarking the novel V2C task, we further complement MSR-VTT with event-level commonsense annotations, i.e. event descriptions with intentions, effects and attributes. We remove captions and videos that do not have clear human activities. This is because having such videos leads to an imbalance in the number of captions for each video, thus making it inappropriate to just evaluate caption generation using BLEU scores.

ATOMIC ([Sap *et al.*, 2019a](#)) is an atlas of everyday commonsense knowledge and contains 880k triplets about causes and effects of human activities, organized as *if-then* relations, annotated by crowd-sourced workers. This data can be categorized based on causal relations, thereby giving us the categories “cause”, “effect” and “attribute”, e.g., “*if* X wants to relax, *then* he will play video game.”

8.3.1 Querying from ATOMIC and Re-ranking

Since inferential knowledge in ATOMIC only covers human activities, we first retain only those captions in Msr-vtt that describe human activities. We then select three queries from ATOMIC most similar to the caption, and extract the commonsense descriptions corresponding to these queries. In order to select a more reasonable subset of commonsense descriptions, we first train a ranking model. We use the BERT ([Devlin *et al.*, 2019](#)) architecture for the ranking model, trained on the ATOMIC dataset for a binary classification task, to predict the relevance of a candidate commonsense description with respect to the event. We select the top three relevant intentions, effects, and attributes for each caption. This allows us to obtain a preliminary set of 9 commonsense annotations per video directly from the ATOMIC dataset, relevant to the caption, albeit with noise and annotations that are not relevant to the video.

Type	Video Caption	Commonsense
Intention	Two guys are wrestling	to beat the opponent
	A man and woman are singing	to express themselves musically
Attribute	A guy is singing in a crowd	outgoing
	Group of riders race on motorcycles.	adventurous
Effect	A person is making a paper airplane	gets excited to fly it
	A man and a woman are talking to each other	share ideas and opinions

Table 8.1: Examples of commonsense annotations (intentions, attributes and effects) retrieved from ATOMIC for captions in MSR-VTT.

8.3.2 Detailed Human Annotation

Since we do not use the video to retrieve commonsense descriptions from ATOMIC, we employ human workers to annotate our dataset. We recruit two sets of human workers to watch the video, read the caption and select/annotate the relevant commonsense descriptions for each video. The first set is Amazon Mechanical Turkers (AMT) who select relevant descriptions. The second set is skilled human annotators, screened from a set of university students proficient in English, who are asked to provide annotations in their own words, and remove or edit irrelevant annotations that were provided by ATOMIC and AMT workers. This makes our annotations not only grounded in the video, but also more descriptive, linguistically diverse, and of higher quality (see Figure 8.3). The descriptions from ATOMIC, although not relevant to the video in some cases, give our workers an idea about the format of annotations desired. The skilled humans reported that 95% of the captions were relevant, and

Relatio	Model	#	Param (M)	CIDE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Attribute	RNN Encoder	+	66.0	-	21.7	-	-	-	-	-
	Decoder									
	Attention	+	52.5	-	36.5	-	-	-	-	-
	Video2Text									
	Transformer	En-	69.7	-	40.7	-	-	-	-	-
	coder + Decoder									
	Video CMS Trans-		69.2	-	47.3	-	-	-	-	-
	former									
	RNN Encoder	+	66.0	12.2	21.2	11.5	8.6	7.5	10.2	16.7
	Decoder									
Effect	Attention	+	52.5	18.5	27.7	16.9	13.3	11.5	16.0	23.9
	Video2Text									
	Transformer	En-	69.7	37.7	35.3	26.6	23.2	21.0	21.4	31.1
	coder + Decoder									
	Video CMS Trans-		69.2	40.8	36.5	28.1	24.6	22.4	22.2	32.3
Intention	RNN Encoder	+	66.0	19.3	41.7	26.4	16.6	11.7	16.5	36.4
	Decoder									
	Attention	+	52.5	23.2	54.3	40.0	27.4	24.7	19.4	45.6
	Video2Text									
	Transformer	En-	69.7	57.4	58.3	45.7	36.3	31.1	27.4	53.2
	coder + Decode									
	Video CMS Trans-		69.2	62.0	60.8	48.4	39.1	34.1	28.5	54.6
	former									

Table 8.2: Evaluation of V2C completion task using CIDE, BLEU, Rouge, and Meteor metrics. We use only BLEU-1 to evaluate the attribute generation since the average length of the ground truth is just less than 2.

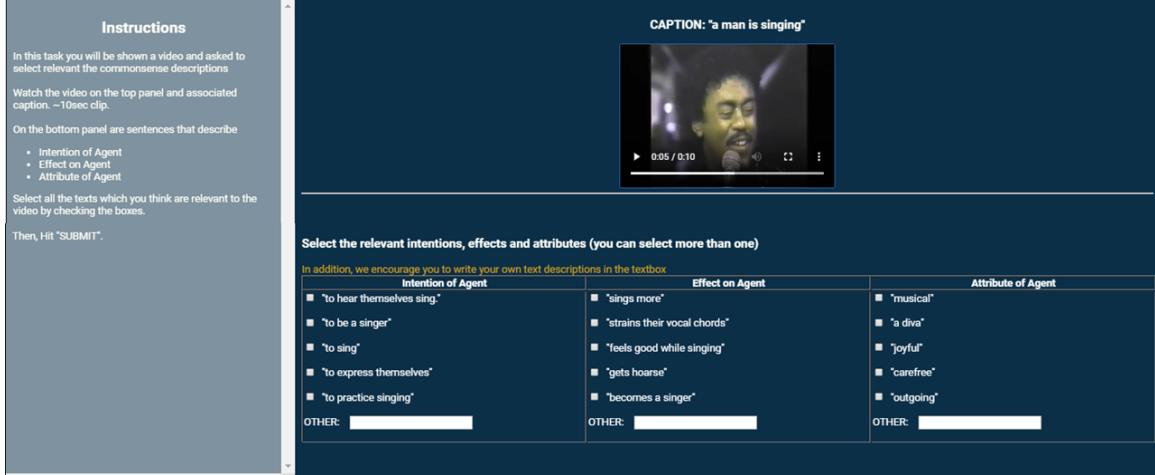


Figure 8.5: Our human labeling interface. We ask human workers to select relevant commonsense descriptions as well provide additional texts in their own words.

65% of the ATOMIC descriptions were useful in understanding the annotation task. Through this procedure, we obtain 6819 videos for training and 2906 videos for testing, a total of 121,651 captions (~ 12 captions/video), each caption accompanied with 5 commonsense knowledge annotations (V2C-Raw set). In experiment, we use video captioning technique to conduct the V2C completion task on V2C-Raw set. In addition, we instruct human annotators to select and rewrite one raw phrase into complete sentences that complement the captions. In total we have 3 complete sentences per video for intention/effect/attribute respectively, and this yields a subset that allows our model to generate complete story-like sentences (V2C-Clean Set). Table 8.1 shows examples from the newly compiled dataset. We conduct rigorous human evaluation to evaluate the quality of our V2C dataset (“Gold Annotations” in Table 8.3) Our annotation interface is shown in Figure 8.5.

We show additional samples from our V2C dataset in Figure 8.8, word cloud in Figure 8.6 and word frequency in Figure 8.7.

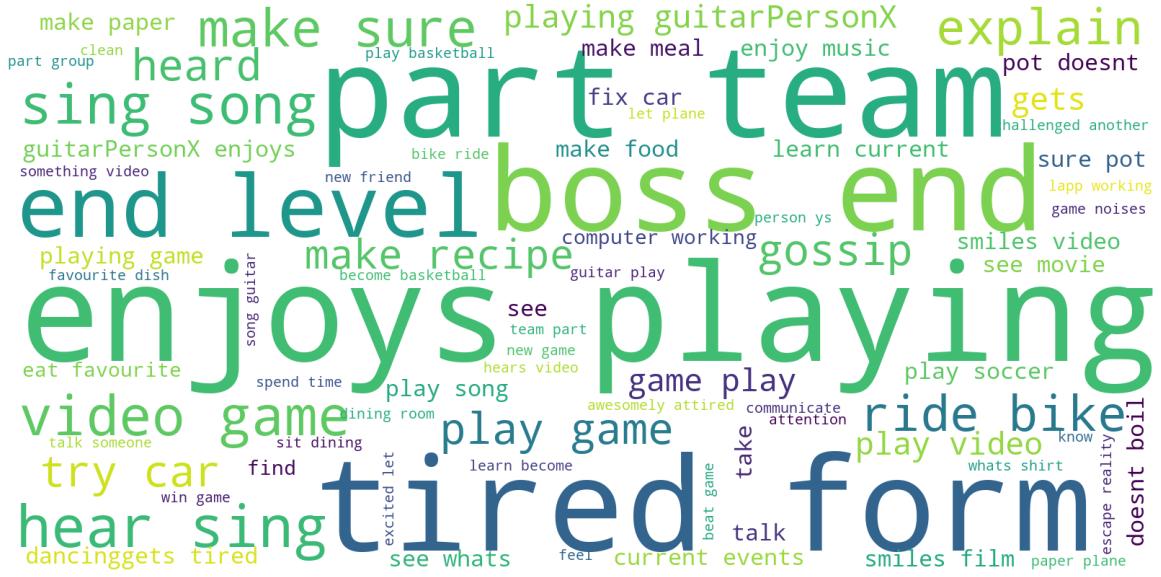


Figure 8.6: Word cloud figure of the intention commonsense annotations from our V2C dataset.

Task	Model	Effect	Attribute	Intention	Average	Caption
E2C-Completion (Text-Only)	9ENC9DEC (Sap et al., 2019a)	44.23	52.01	49.72	49.47	-
	COMET (Bosselut et al., 2019)	54.98	56.28	66.32	59.22	-
V2C-Completion	Att-Enc-Dec(Gao et al., 2017)	66.09	52.40	56.26	58.25	-
	VCT-Completion	66.83	63.45	67.37	65.88	-
V2C-Generation	Att-Enc-Dec(Gao et al., 2017)	55.93	<u>74.87</u>	65.54	64.78	<u>74.67</u>
	VCT-Generation	62.99	73.54	66.74	67.76	73.17
Gold Annotations	V2C Dataset	75.19	83.03	80.11	79.44	95.01

Table 8.3: Human evaluation scores for V2C. Captions are an input for the **V2C-Completion** task, and generated for the **V2C-Generation** task. The best model is given in bold, while the overall best is underlined.

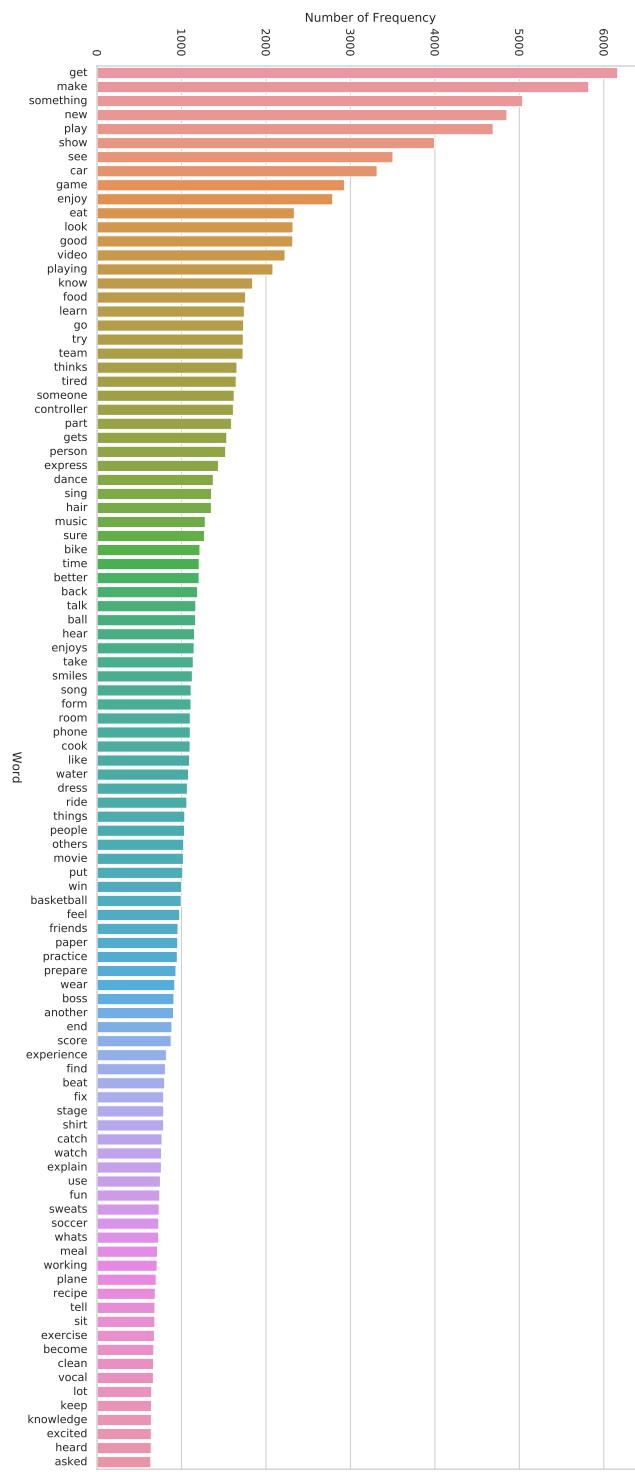


Figure 8.7: Top-100 most frequent words in our V2C dataset (stop words are ignored).

The person wants:	As an effect, the person:			The person is:
to express themselves to sing a song to make life more pleasant		put it on YouTube learns a new dance gets into their rhythm		outgoing enthusiastic energetic
	Caption A guy sings a song in a music video			
	Rewritten Story: Because he wants to express himself, a guy sings a song in a music video, and he will upload it to YouTube soon. He is quite an enthusiastic guy.			
to show off at the gym to acquire new footwear to wear appealing clothes		becomes obsessed gets their picture taken grabs attention from people		stylish trendy fashionable
	Caption Girls trying on new sports bra.			
	Rewritten Story: In order to purchase new sportswear, girls trying on new sports bra, and they may grab attention from people later. They are all stylish person.			
to win the race to earn a medal to win the competition		runs a race is congratulated at the finish line focuses on the race		athletic competitive determined
	Caption Groups of runner get prepared to run a race.			
	Rewritten Story: Since the athletes are trying to win the race, groups of runner get prepared to run a race, and they will run and get congratulated at the finish line soon. They are athletic.			
to show appreciation to be accommodating to talk to them		sweats from nervousness shares information communicates		empathetic talkative conversational
	Caption President Obama calls a team to congratulate them.			
	Rewritten Story: To show his appreciation to the winners, President Obama calls a team to congratulate them, the girls will got sweats because of that. The Obama is so talkative.			
to have a conversation convey information to give speech		gives a rebuttal gets to meet the host loses their voice due to loud talking/yelling		extroverted polite speaker
	Caption A group of males speaking to each other at a meeting.			
	Rewritten Story: In order to convey with each other the information, a groups of males speaking to each other at a meeting, they will get into a rebuttal soon. The people have the attribute to be extroverted.			
to get to her destination to get somewhere to drive fast		travels to a different city arrives at their destination enjoy driving		traveling a good driver speedy
	Caption A man drives a vehicle through the countryside.			
	Rewritten Story: To get to his destination as soon as possible, a man drives a vehicle through the countryside, he may soon arrives at his destination. The man is a good driver.			
to get the computer working set up system to clean the viruses from his computer		turns off the computer boots up the computer spends money on a new computer		busy smart informative
	Caption A woman in a business suit looking at a computer monitor.			
	Rewritten Story: Because the computer is not working and the woman is trying to fix it, a woman in a business suit looking at a computer monitor, she will boots the computer first soon. She is a very informative person.			

Figure 8.8: Qualitative examples of our V2C dataset.

8.4 Experiments

In this section we describe the loss function used for training our model, additional details about video pre-processing, hyper-parameters, and baseline models, and the metrics used for evaluation.

Loss Function: The decoder parameters Θ are trained to maximize the log-likelihood over the training set given by $\mathcal{L} = \mathcal{L}_{cap} + \mathcal{L}_{cms}$, where

$$\begin{aligned}\mathcal{L}_{cap} &= \sum_{t=1}^{N_S} \log \Pr(\mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{v}; \Theta), \text{ and} \\ \mathcal{L}_{cms} &= \sum_{t=1}^{N_C} \log \Pr(\mathbf{y}_t | \mathbf{y}_{t-1}, [\mathbf{v}, \tilde{\mathbf{s}}]; \Theta).\end{aligned}\tag{8.4}$$

\mathbf{y}_t denotes the one-hot vector probability of each word at time t , and N_S, N_C denote the length of the caption and commonsense respectively.

Setting: In order to obtain video representations, we uniformly sample 40 frames from each video and extract features using feed ResNet (He *et al.*, 2016) pre-trained on Imagenet ILSVRC12 dataset (Deng *et al.*, 2009) and get a 2048-d output from the last layer. We use one-hot input (1-of- N encoding) of the text input and pass it through an embedding layer to produce a 1028-d hidden vector. We use independent vocabularies for captioning and commonsense generation with sizes 27,603 and 24,010 respectively. Note that, as the generated

Hyperparameters: Our decoder is a lightweight transformer decoder consisting of 6 transformer blocks with 8 attention heads each. We use Adam optimizer with 5000 warm-up steps, and learning rate initialized at $1e-4$, and a dropout probability of 0.1 after the residual layer. Our model is trained on a machine with single NVIDIA 1080-Ti GPU.

Baseline Model: We compare our method with strong video captioning baseline models like, S2VT (Venugopalan *et al.*, 2015), “Attention-Enc-Dec” (Gao *et al.*, 2017) – LSTM based models which reach competitive performing on MSR-VTT dataset. and “Dense Captioning” (Zhou *et al.*, 2018), which is a transformer based video captioning model. As “Dense Captioning” is proposed to generate multiple continuous captions for a long untrimmed videos, we modify this by removing the temporal bounding boxes prediction module, and produce two continuous captions (caption + commonsense sentence) together without corresponded starting and ending time. All baselines are trained to predict commonsense descriptions from video on the V2C dataset. We do not compare with VideoBERT (Sun *et al.*, 2019) which is trained on a limited set of cooking videos and hence non-transferable, and requires individual captions for multiple segments of the video.

Metrics: We report both the performances evaluated by automatic scores and human evaluations following the protocols from (Bosselut *et al.*, 2019; Sap *et al.*, 2019a). We evaluate our method using BLEU (n=1-4) (Papineni *et al.*, 2002), Meteor (Banerjee and Lavie, 2005), Rouge (Lin, 2004), score of the generation on its corpus. We further conduct human evaluations using AMT workers, who are asked to identity whether the generated commonsense justifiably completes the events (V2C-completion). We follow the setup in (Sap *et al.*, 2019a) and randomly sample 100 videos from test set and collect 10 generations for each. To guarantee the objectiveness of the human evaluations, we hire 5 workers for each sample, yielding **30k** ratings in total for each model.

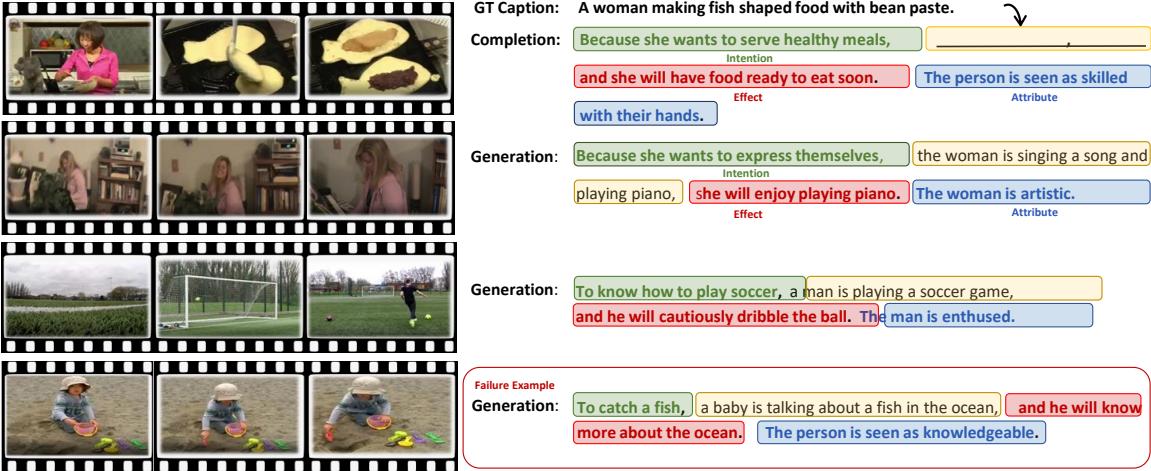


Figure 8.9: Examples of outputs of our model for the V2C Completion and Generation tasks along with the ground-truth (GT) caption. A failure example shown in the bottom red box.

8.4.1 Results

Natural Language Generation Metrics: We show evaluation of the common-sense completion task in Table 8.2. Compared to the baseline model, our method exhibits a consistent and overall improvement on almost all metrics. Our V2C-Transformer significantly outperforms the attention mechanism based LSTM model ([Venugopalan *et al.*, 2015](#)) by 6.5% at BLEU-4 for the intention prediction. Because the V2C-Transformer and the LSTM model share a similar video decoder, our performance improvement could be attributed to the use of self-attention mechanisms in the transformer block in decoding phase. This observation is consistent with the conclusion from ([Bosselut *et al.*, 2019](#)), and yields further support to the transformer architecture being suited for commonsense inference tasks. Moreover, when compared with a transformer encoder + decoder architecture which have similar learnable parameters with our model, our model exhibits better evaluation scores, verifying it as a strong baseline model for the V2C task. For a fair comparison, all baseline models are pre-trained for 600 epochs with on par learnable parameters.

Human Evaluation In Table 8.3, E2C (Event to Commonsense) is the task of commonsense completion given only textual events (Sap *et al.*, 2019a; Bosselut *et al.*, 2019) as opposed to V2C which uses both text and video. 9ENC9DEC (Sap *et al.*, 2019a) is composed of nine GRU based encoder-decoders as a baseline model for commonsense completion on text, and COMET (Bosselut *et al.*, 2019) is a large-scale generative pre-trained transformer (GPT) model (Radford *et al.*, 2018). We would like to highlight that our transformer model is light-weight with only half of the parameters in GPT without any pre-training.

We evaluate our model on the tasks of caption generation with human evaluations, and also compare it with the gold annotations. Our gold annotation for ground-truth captions (sourced from the MSR-VTT dataset) points to the fact that a small percentage of captions from MSR-VTT are not relevant to the video, and this is amended by our human workers.

For the V2C-Completion task, our V2C-Transformer model is substantially better (by 7.73%) than the LSTM-based model from (Gao *et al.*, 2017), and shows consistent lead on each dimension. Thus, when the ground-truth caption is given, our model is able to generate much more relevant commonsense descriptions, thereby consolidating its ability of commonsense generation.

For the task of V2C-Generation, the difference between human scores for LSTM vs V2C-Transformer is reduced, but our VTC outperforms on average by 2.98%. This may be attributed to the fact that the LSTM-based model is slightly better at generating captions.

Generating Textual Stories with Commonsense In order to generate story-like textual descriptions that complement the factual captions, we additionally train our model to exploit our diverse complete-sentence annotations. Specifically, instead of

Instructions

In this task you will be shown a video and asked to rate the commonsense descriptions generated by our algorithm on a scale of 1-5.

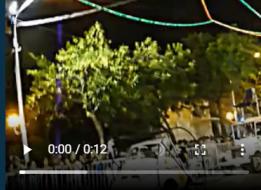
Watch the video on the top panel and associated caption. ~10sec clip.

On the bottom panel are sentences that describe

- Effect of Agent's Action

Judge the effects with regards to their relevance to the video by providing a rating for each.

CAPTION: "a group of cars lineup for a race"



0:00 / 0:12

Effect	Rating for Effect
"gets a speeding ticket"	1 ● 2 ● 3 ● 4 ● 5 ●
"focus on driving"	1 ● 2 ● 3 ● 4 ● 5 ●

Instructions

In this task you will be shown a video and asked to rate the commonsense descriptions generated by our algorithm on a scale of 1-5.

Watch the video on the top panel and associated caption. ~10sec clip.

On the bottom panel are sentences that describe

- Agent's Attribute

Read the attributes, understand them, and judge their relevance to the video by providing a rating for each.

CAPTION: "a boy and girl dance together"



0:00 / 0:12

Attribute	Rating for Attribute
"graceful"	1 ● 2 ● 3 ● 4 ● 5 ●
"musical"	1 ● 2 ● 3 ● 4 ● 5 ●

Instructions

In this task you will be shown a video and asked to rate the commonsense descriptions generated by our algorithm on a scale of 1-5.

Watch the video on the top panel and associated caption. ~10sec clip.

On the bottom panel are sentences that describe

- Effect of Agent's Action

Judge the effects with regards to their relevance to the video by providing a rating for each.

CAPTION: "a man is giving a speech"



0:15 / 0:18

Effect	Rating for Effect
"makes voice and opinions known"	1 ● 2 ● 3 ● 4 ● 5 ●
"has to go back to own desk"	1 ● 2 ● 3 ● 4 ● 5 ●

Figure 8.10: Snapshot of our AMT human evaluation interface for V2C-completion task.

Instructions

In this task you will be shown a video and asked to rate the commonsense descriptions generated by our algorithm on a scale of 1-5.

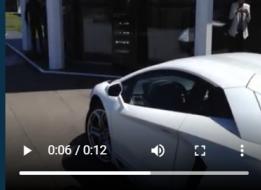
Watch the video on the top panel (typically with an agent performing some actions). ~10sec clip.

On the bottom panel are two types of textual descriptions:

- Generated Caption
- Intention of Agent's Action

Judge them independently on the basis of relevance to the video.

Provide a rating on the scale of 1-5.



Type	Text	Rating
CAPTION:	"a person is talking about a car"	1 ● 2 ● 3 ● 4 ● 5 ●
INTENTION:	"to tweak their car to their liking"	1 ● 2 ● 3 ● 4 ● 5 ●
CAPTION:	"person is recording the car"	1 ● 2 ● 3 ● 4 ● 5 ●
INTENTION:	"transportation to be able to go somewhere"	1 ● 2 ● 3 ● 4 ● 5 ●

Instructions

In this task you will be shown a video and asked to rate the commonsense descriptions generated by our algorithm on a scale of 1-5.

Watch the video on the top panel (typically with an agent performing some actions). ~10sec clip.

On the bottom panel are textual descriptions of two types:

- Generated Caption
- Effect of Agent's Action

Judge these independently on the basis of relevance to the video.

Rate them on the scale of 1-5.



Type	Text	Rating
CAPTION:	"a man is showing how to use a toy"	1 ● 2 ● 3 ● 4 ● 5 ●
EFFECT:	"personx gains new skills"	1 ● 2 ● 3 ● 4 ● 5 ●
CAPTION:	"a person is showing how to make a good item"	1 ● 2 ● 3 ● 4 ● 5 ●
EFFECT:	"gets a lot of effort"	1 ● 2 ● 3 ● 4 ● 5 ●

Instructions

In this task you will be shown a video and asked to rate the commonsense descriptions generated by our algorithm on a scale of 1-5.

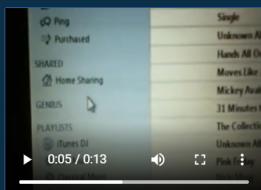
Watch the video on the top panel (typically with an agent performing some actions). ~10sec clip.

On the bottom panel are textual descriptions of two types:

- Generated Caption
- Agent's Attribute

Judge them independently on the basis of relevance to the video.

Provide a rating on the scale of 1-5.



Type	Text	Rating
CAPTION:	"a person is showing how to use a computer program"	1 ● 2 ● 3 ● 4 ● 5 ●
ATTRIBUTE:	"smart"	1 ● 2 ● 3 ● 4 ● 5 ●
CAPTION:	"computer screen is shown in the video"	1 ● 2 ● 3 ● 4 ● 5 ●
ATTRIBUTE:	"curious"	1 ● 2 ● 3 ● 4 ● 5 ●

Figure 8.11: Snapshot of our AMT human evaluation interface for V2C-generation task.

producing the commonsense knowledge given the videos and captions, we finetune our pre-trained V2C-Transformer model on predicting the human rewritten texts, and generate complete story-like captions. Since we do not have enough annotations per sample to compute a fair BLEU score for comparisons, we showcase some sample generated descriptions for qualitative analysis (see Figure 8.9). With that, we observe V2C-Transformer is able to produce complete stories that contain simple, while logically consistent storylines that complement both the visual content and the factual descriptions. We believe that collecting a set of story-like sentences will further enrich our models, and allow us to generate much more contextual, creative, and natural commonsense descriptions from a video.

8.4.2 Qualitative Generation Results

We show additional **V2C-Completion** samples by our V2C-Transformer model in Table. 8.4.

8.5 V2C-QA

Another way of generating commonsense descriptions about the video is by asking pointed questions. Consider the example in 8.1 where we ask the question “*What happens next to the runners*”, about the *effect* of the action “*prepare*” performed by the agents “*group of runners*” observed in the video. We propose a V2C-QA – an open-ended commonsense video question-answering task, where we ask questions about the intents, effects and attributes of the agents in the video.

Dataset: We use the caption and commonsense annotations in the V2C dataset to create question-answer pairs for each video. We first extract the action and subject from the caption using SpaCy linguistic features ([Honnibal and Johnson, 2015](#)). For

Intention	Caption	Effect	Attribute
to entertain people	a band is performing for a crowd	gets applause	acting
to try out PersonY's new car	a man checks out detail on a car	gets a speeding ticket	helpful
to learn about current events	a complex news host gives an update on rappers.	gets informed about current political events	talkative
to be in a good mood	a group of people trying to perform an exorcism on a girl	gets applause	fun
to show his knowledgeable	there is an old man is answering to somebody questions	gets another question	sporty
to score a point	a man is shooting a basketball ground	gets exercise	helpful
to share their message	a man giving a speech to important people	gets applause	orator
to be safe from anything that lurks in the dark	a group of people are being chased by crocodiles	gets tired from taking pictures	scared
to be informed about the world	a girl is describing about hot news	learns about whats happening worldwide	gossipy
to watch something interesting	a children s television show clip	smiles at the screen	entertained
to enjoy the evening with the concert band	a band composed of older gentlemen are playing blue grass music on a small stage and people are dancing along to the music swing-style	gets tired form dancing	fun
to be part of the team	there is a woman playing badminton in a court	gets tired after exercise	athletic
to try out person ys new car	a boy explaining the features of a car	they check car websites online to look at deals	helpful

Table 8.4: Illustrative samples generated by our V2C-Transformer model on **V2C-completion** task.



Conventional Video QA

Who is fighting?

the soldier

V2C - QA

What is the intention of the person on the left?

to protect the country

What could happen to the person after this?

gets injured

What is the characteristic of the person?

brave, powerful

Does the person want to protect his country?

Yes

Figure 8.12: Example questions from V2C-QA compared with conventional video question answering.

each intention, attribute and effect for a video, we use template-based generation to get 7 types of questions – yielding 21 questions per sample, including negative questions as in Gokhale *et al.* (2020c). In total, we have 1,250 training videos and 250 test videos, and a total of 37k questions. We have a set of 5,555 unique answers for our questions. Each question can have multiple possible true answers as shown in the example in Figure 8.12. The V2C-QA task asks questions that require commonsense reasoning about internal mental states, motivations, and latent aspects of agents in the video as opposed to the conventional video-QA questions about visible objects and actions.

Models: We utilize our V2C-Encoder followed by an open-ended answering module. We jointly predict the type of the question and combine it with the V2C encoding using



Question Type	Question	Answer
Intention	What might be the goal of the person?	to record a music video
Intention (Negative)	What could the person not want to achieve?	to bake a cake
Intention (Action)	What prompts the person to do the action?	to express themselves
Intention (Action, Negative)	What did not lead the person to act like that?	to feed the dog
Intention (Why)	Why might the person be doing the action?	to entertain viewers
Intention (Yes-No)	Does the person wish to express himself?	Yes
Intention (Yes-No, Negative)	Does the person want to not get recognition?	No
Effect	What will the person do after this?	puts the video on YouTube
Effect (Negative)	What does not happen as a result?	the person gets sad
Effect (Action)	What does the dancing end up in?	becomes tired
Effect (Action, Negative)	What will not happen due to the action?	feels tense
Effect (How)	How does the person feel after performing?	feels accomplished
Effect (Yes-No)	Could the person put it on YouTube as a result?	Yes
Effect (Yes-No, Negative)	Will the person not learn a new dance?	No
Attribute	What trait does the man possess?	musical
Attribute (Negative)	What attribute does not match with the person?	angry
Attribute (How)	How can the person be described?	entertaining
Attribute (Action, How)	How can the dancing person be characterized?	rhythmic
Attribute (Yes-No, Action)	Is the person who is singing smiling?	Yes
Attribute (Yes-No)	Is the person entertaining?	Yes
Attribute (Yes-No, Negative)	Is the person not tense?	Yes

Table 8.5: Examples of open-ended V2C-QA samples

a feed-forward network. For textual features, we use embeddings from BERT-base (Devlin *et al.*, 2019). Our models are trained on the open-ended QA task and set-up as a multi-label classification task similar to VQA (Antol *et al.*, 2015), with an answering module design inspired by LXMERT (Tan and Bansal, 2019). Our loss function includes the classification loss for answering, the attention loss for question-type, and a label-ranking loss.

Results: MSR-VTT QA (Xu *et al.*, 2017) is as a good baseline since it is trained on a conventional videoQA task on the MSR-VTT videos, and only takes video and query as input, unlike recent video understanding models (Lei *et al.*, 2018) that take additional supervision, such as subtitles. However this model is trained for a multiple-choice QA scheme, so we modify it with our open-ended answering module. We compare our models when we use our encoder pretrained on the V2C caption generation task, and then finetune it on the V2C-QA task. We also train models with ground-truth factual captions as input. Our results are shown in Table 8.6, where we evaluate on prediction of top-k (1,3,5) answers, and report precision and recall. Our encoder pre-trained on the V2C task outperforms all other models. Attribute-related questions are easier to answer, while the models struggle the most for questions about intention. Captions help in questions about effects. The overall text-only baseline shows an insignificant bias between the question and answer-options.

8.6 Related Work

Video Captioning: Captioning is crucial for understanding visuals; however it is typically limited to describing observable objects and events (Yang *et al.*, 2011; Thomason *et al.*, 2014; Gan *et al.*, 2017)), or for generating paragraphs or multi-sentence captions about the image or video (Krause *et al.*, 2017; Krishna *et al.*, 2017).

Model		top-1		top-3		top-5	
		p	r	p	r	p	r
Intention	MSR-VTT QA	9.68	2.13	7.15	4.68	6.07	6.60
	V2C-T	10.34	2.31	7.69	5.03	6.37	6.87
	V2C-T + Captions	10.72	2.54	8.08	5.47	6.39	7.20
	Pretrained V2C-T	10.77	2.69	8.01	5.58	6.71	7.88
Effect	Pretrained V2C-T + Cap.	11.04	2.68	7.96	5.70	6.63	7.79
	MSR-VTT QA	19.89	5.02	8.04	5.91	5.30	6.49
	V2C-T	20.95	5.43	8.65	6.57	5.65	7.06
	V2C-T + Captions	20.95	5.32	8.50	6.48	5.76	7.26
	Pretrained V2C-T	20.95	5.32	8.63	6.55	5.82	7.49
Attribute	Pretrained V2C-T + Cap.	21.12	5.60	8.70	6.89	5.83	7.68
	MSR-VTT QA	46.10	37.22	16.02	49.45	7.49	41.03
	V2C-T	59.52	48.30	22.39	51.40	13.97	52.57
	V2C-T + Captions	59.74	48.22	23.12	52.44	14.64	54.35
	Pretrained V2C-T	60.72	49.00	23.18	52.73	14.98	55.40
	Pretrained V2C-T +Cap.	59.57	48.24	23.10	52.54	14.94	54.91
	Text-Only Baseline	12.36	11.70	13.84	12.35	14.77	14.10

Table 8.6: Precision (p) and Recall (r) for V2C-QA for each type of question.

However, for detailed video understanding, one needs to obtain descriptions that go beyond observable visual entities and use background knowledge and commonsense to reason about objects and actions. Work for inferring motivations of human actions in static images by incorporating commonsense knowledge are reflected in [Pirsavash et al. \(2014\)](#); [Vondrick et al. \(2016\)](#). Commonsense caption generation has been approached on abstract scenes and clip-art images in [Vedantam et al. \(2015a\)](#). We present the first generative model for commonsense video captioning.

Video Question Answering: Since caption generation can only describe observable events, recent work seeks to move closer to comprehension, by learning to answer complex questions about videos. However, the datasets used for Video QA ([Yang et al., 2003](#); [Xu et al., 2016](#); [Zhu et al., 2017](#)) focus only on directly evident visual concepts and construct the questions mostly about “where” and “what” aspects. Question answering on movie videos has been explored by [Tapaswi et al. \(2016\)](#) who collect questions about “why” and “how” aspects. Recently [Lei et al. \(2018\)](#); [Zadeh et al. \(2019\)](#) have propose video-based QA tasks with open-ended high-order questions that need multi-modal understanding, social intelligence modeling, and spatio-temporal reasoning. We introduce a novel open-ended video question answering task in this paper, where the questions are about three aspects of commonsense human behavior.

Visual Reasoning: Aspects of visual reasoning have been explored by [Yatskar et al. \(2016\)](#) as a situation recognition task on single images, and in Visual Madlibs ([Yu et al., 2015](#)) as a “fill-in-the-blanks” task for single-image captioning that contains some categories which require reasoning about internal mental states and future events. [Kim et al. \(2018\)](#) provide textual explanations for actions in a self-driving scene. [Zellers et al. \(2019\)](#) propose a visual question answering task that requires commonsense

reasoning to answer a question and to provide a rationale behind the answer. Spatial and compositional reasoning is required to answer questions about synthetic images in CLEVR (Johnson *et al.*, 2017). Critical aspects of visual reasoning also include the model’s ability to conduct object grounding by natural language descriptions (Rohrbach *et al.*, 2016; Fang *et al.*, 2018, 2019). Another aspect of visual reasoning is the ability predict a sequence of actions (procedure planning), or to reason about intermediate video frames (walkthrough planning) between two frames, explored in Gokhale *et al.* (2019); Chang *et al.* (2020).

Textual Commonsense: Commonsense-based question answering is an area of active research with several datasets and challenges requiring reasoning about conceptual commonsense (Talmor *et al.*, 2019), physical commonsense (Bisk *et al.*, 2020), social commonsense (Sap *et al.*, 2019b), and abductive commonsense (Bhagavatula *et al.*, 2020). On the other hand, challenges such as ProPara (Dalvi *et al.*, 2018) and bAbI (Weston *et al.*, 2016) require tracking elements, actions, and effects of actions. Commonsense-based text generation has recently been explored via the ATOMIC dataset (Sap *et al.*, 2019a), a corpus of 877k textual descriptions of inferential knowledge organized as *if-then* relations. Bosselut *et al.* (2019) adopt the ATOMIC dataset to learn a generative model of commonsense knowledge. To the best of our knowledge, ours is the first work on *generating* commonsense descriptions from visual inputs.

8.7 Outlook

A video typically contains one or many objects (sometimes performing actions) in different backgrounds, scenes, or situations. Some objects may be “passive” such as trees or buildings, while some objects may be “active” such as people performing

actions like walking, singing, and driving. This paper is focused on describing such active agents in terms of their intentions, effects of their actions, and attributes that characterize these agents.

We distinguish V2C from the traditional video captioning task. Video captions describe observable objects, background, and actions, while commonsense descriptions in our task seek to describe the unobservable intentions of the agent (pre-conditions or mental conditions), effects of the action (that happen in the future), and attributes which characterize the agent. Thus commonsense generation goes *beyond the visible*. Ours is the first attempt at developing a generative video-based commonsense model. We anticipate that our framework can be utilized for many applications in video understanding, comprehension, human-robot interaction, and learning commonsense in a multi-modal setting.

8.8 Conclusion

In this paper, we explore a novel and challenging task to generate video descriptions with rich commonsense descriptions that complement the factual captions. We expand an existing video captioning dataset for the V2C task through automated retrieval from a textual commonsense corpus followed by human labeling, and present a novel V2C-Transformer model to serve as a strong baseline method for the V2C task. Our evaluation verifies the effectiveness of our method, while also indicating a scope for further study, enhancement, and extensions in the future. Our experiments on using the V2C-Transformer as a component for the V2C-QA task show that the model has transfer learning capabilities that can be applied to other vision-and-language tasks such as question-answering, that require commonsense reasoning.

Chapter 9

BENCHMARKING SPATIAL RELATIONSHIPS IN TEXT-TO-IMAGE GENERATION

Spatial understanding is a fundamental aspect of computer vision and integral for human-level reasoning about images, making it an important component for grounded language understanding. While recent text-to-image synthesis (T2I) models have shown unprecedented improvements in photorealism, it is unclear whether they have reliable spatial understanding capabilities. We investigate the ability of T2I models to generate correct spatial relationships among objects and present VISOR, an evaluation metric that captures how accurately the spatial relationship described in text is generated in the image. To benchmark existing models, we introduce a dataset, SR_{2D} , that contains sentences describing two objects and the spatial relationship between them. We construct an automated evaluation pipeline to recognize objects and their spatial relationships, and employ it in a large-scale evaluation of T2I models. Our experiments reveal a surprising finding that, although state-of-the-art T2I models exhibit high image quality, they are severely limited in their ability to generate multiple objects or the specified spatial relations between them. Our analyses demonstrate several biases and artifacts of T2I models such as the difficulty with generating multiple objects, a bias towards generating the first object mentioned, spatially inconsistent outputs for equivalent relationships, and a correlation between object co-occurrence and spatial understanding capabilities. We conduct a human study that shows the alignment between VISOR and human judgement about spatial understanding. We offer the SR_{2D} dataset and the VISOR metric to the community in support of T2I

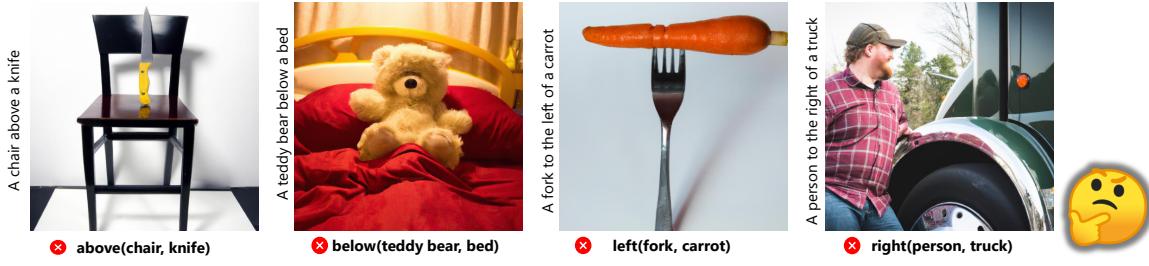


Figure 9.1: We benchmark T2I models on their competency with generating appropriate spatial relationships in their visual renderings. Although text inputs may explicitly mention these spatial relationships, T2I models lack such spatial understanding.

reasoning research.¹

9.1 Introduction

Text to image synthesis (T2I), has advanced rapidly with capabilities for generating high-definition images in response to text prompts. Models are being used as tools for art, graphic design, and image editing. The power of T2I models for generating photorealistic objects and scenes is well-known. We less understand the ability of the models to faithfully render spatial relationships in its compositions.

We pursue the question: Do T2I models have the ability to render the spatial relationships among objects that are specified in text prompts? Fig. 9.1 illustrates images generated by a state-of-the-art model (DALLE-v2 ([Ramesh et al., 2022](#))) for sentences that contain a spatial relationship between two objects. In these examples, although both objects mentioned in the text are generated, the specified spatial relationship is not rendered.

Spatial relations and larger scene geometries are integral aspects of computer vision. Rendering and reasoning about these relationships is crucial for many applications such as language-guided navigation and object manipulation ([Anderson et al., 2018b](#);

¹<https://visort2i.github.io/>

Mees *et al.*, 2020; Nair *et al.*, 2022). A lack of spatial understanding by T2I models can be frustrating to creators seeking to render specific configurations of objects. The assertion of spatial relationship is common in natural communication among humans and poor capabilities in this realm will rapidly come to the fore in navigational and instructional applications.

Prior work on evaluation metrics for T2I models have focused on photorealism (Salimans *et al.*, 2016; Heusel *et al.*, 2017), object accuracy (Hinz *et al.*, 2020), and image-text vector similarity (via CLIP (Hessel *et al.*, 2021), retrieval (Xu *et al.*, 2018), and captioning (Hong *et al.*, 2018)). We find that these metrics are insensitive to errors with generating spatial relationships (Sec. 9.5.1). This finding highlights the need for a metric to quantify competencies and progress in spatial understanding. We develop an automated evaluation pipeline that employs computer vision to recognize objects and their spatial relationships, and harness this pipeline to conduct a large-scale evaluation of the spatial understanding capabilities of T2I models. We create the “SR_{2D}” dataset, containing 25,280 sentences describing two-dimensional spatial relationships (*left/right/above/below*) between pairs of commonly occurring objects from MS-COCO (Lin *et al.*, 2014), as shown in Table 9.1. We study several state-of-the-art models: GLIDE (Nichol *et al.*, 2022), DALLE-mini (Dayma *et al.*, 2021), CogView2 (Ding *et al.*, 2022), DALLE-v2 (Ramesh *et al.*, 2022), Stable Diffusion (Rombach *et al.*, 2022), and Composable Diffusion (Liu *et al.*, 2022b). For each model we generate and evaluate four images per SR_{2D} example, i.e., a large-scale study of 101,120 images per model. Our study makes significant advances to evaluation of T2I reasoning capabilities since we evaluate photorealistic images rather than synthetic objects on solid background.

We introduce a new evaluation metric we refer to as VISOR (for verifying spatial object relationships), to compare the spatial understanding abilities of T2I models.

We define three variants of the metric: (1) VISOR: verifies spatial correctness for each image w.r.t. its text input, (2) VISOR_n : consider whether at least n of the multiple generated images for each text input are spatially correct, (3) $\text{VISOR}_{\text{cond}}$: verifies the spatial correctness in images, conditioned on both objects being generated by the model. While VISOR provides a macro-perspective on the performance gap in the spatial capabilities of T2I models, VISOR_n reflects the practical value of the model to users who can select one of many images generated by the model. The conditional formulation $\text{VISOR}_{\text{cond}}$ disentangles two capabilities: (i) the generation of multiple objects and (ii) generation of correct spatial relationships between the rendered objects. We conduct a human study on Amazon Mechanical Turk and find that the VISOR metric is correlated with human judgment.

Our experiments reveal several interesting findings. First, we find that all existing models are significantly worse at generating two objects as compared to their capability to render single objects. While previous work shows exceptional zero-shot compositionality of colors, styles, and attributes ([Ramesh et al., 2022](#); [Saharia et al., 2022](#); [Yu et al., 2022](#)), we found challenges with compositionality for multiple objects. Second, we find poor spatial understanding: even in cases where both objects are generated, models tend to ignore spatial relationships specified in language. VISOR scores for all models show that even the best model in our benchmark generates correct spatial relationships on less than 40% of test cases. When we consider a strict metric (VISOR_4) that requires that all generated images for text prompts to have correct spatial relationships, the best model (DALLE-v2) achieves the goal in 7.49% cases. Third, we discover several biases in T2I models: to generate only the first object mentioned in the text and ignoring the second, to show better performance on commonly occurring object pairs, to have a tendency to merge two objects into one, and to have inconsistent outputs for equivalent text inputs.

A	B	R	Text
microwave	sink	left	A microwave to the left of a sink
elephant	cat	right	An elephant to the right of a cat
donut	airplane	above	A donut above an airplane
suitcase	chair	below	A suitcase below a chair

Table 9.1: Examples text inputs from the SR2D dataset for a pair of objects (A, B) and relationship R between them.

To summarize, our contributions are as follows:

- We introduce a metric called VISOR to quantify spatial reasoning performance. VISOR can be used off-the-shelf with any text-to-image model, disentangles correctness of object generation with the ability of spatial understanding.
- We construct and make available a large-scale dataset: SR_{2D}, which contains sentences that describe spatial relationships between a pair of 80 commonly occurring objects along with linguistic variations.
- With SR_{2D}, we conduct a large-scale benchmarking of state-of-the-art T2I models with automated and human evaluation of spatial reasoning abilities of state-of-the-art T2I models using the VISOR metric. We find that although existing T2I models have improved photorealism, they lack spatial and relational understanding with multiple objects, and indicate several biases.

9.2 Related Work

Text-to-Image Synthesis. Earlier work (Reed *et al.*, 2016; Zhang *et al.*, 2017) trained and evaluated models on human-labeled datasets (Welinder *et al.*, 2010;

Nilsback and Zisserman, 2008; Lin *et al.*, 2014). Recent work on T2I has focused on zero-shot capabilities by taking advantage of implicit knowledge from pretrained language models and V+L models like CLIP, and the diffusion technique to train on large-scale web data.

Biases in Vision+Language models have been studied from a linguistic perspective, such as question-answer priors in VQA (Agrawal *et al.*, 2018; Kervadec *et al.*, 2021), gender bias in captioning (Hendricks *et al.*, 2018; Zhao *et al.*, 2017), shortcut effects in commonsense reasoning (Ye and Kovashka, 2021), and failure modes in logic-based VQA (Ray *et al.*, 2019; Gokhale *et al.*, 2020c; Goel *et al.*, 2021). The difficulty of spatial understanding has been studied for visual grounding (Liu *et al.*, 2019b), image-text matching (Liu *et al.*, 2022a), VQA (Johnson *et al.*, 2017; Hudson and Manning, 2018), and navigation (Chen *et al.*, 2019).

Human Study about Relational Understanding. Conwell *et al.*(Conwell and Ullman, 2022) conducted a human study (1350 images) of DALLE-v2 on a set of eight physical relations and seven action-based relations between 12 object categories. Our human study is significantly larger in scale, considers diverse text inputs, several state of the art models, and establishes an alignment with the automated VISOR metric.

Empirical Evaluation of Visual Reasoning Skills. DALL-Eval (Cho *et al.*, 2022) evaluates reasoning skills of T2I models trained and tested on a synthetically generated dataset PAINTSKILLS with black backgrounds and 21 rendered object categories. In our work, we instead focus on the evaluation of photorealistic and open-domain images with commonly occurring real-world objects and backgrounds on a large scale. Most importantly, we devise a new human-aligned metric (VISOR) that disentangles object accuracy from spatial understanding to accordingly measure

progress in spatial reasoning despite the model’s capabilities in object generation.

Other Failure Modes of T2I Models. Preliminary stress-testing of DALLE-v2 (Marcus *et al.*, 2022) (14 prompts), (Leivada *et al.*, 2022) (40 prompts) and (Saharia *et al.*, 2022) (200 prompts) illustrated anecdotal failures of the model in terms of compositionality, grammar, binding, and negation. However, since these studies rely on human judgment, there is a need for automated evaluation techniques for comparing the reasoning abilities of T2I models. Our paper fills this gap with the automated VISOR metric for spatial relationships and the large-scale SR_{2D} dataset.

9.3 Spatial Relationships Challenge Dataset

Predicate Generation. Our goal is to collect a set of sentences that describe spatial relationships between two objects. Let \mathcal{C} be the set of object categories. Let \mathcal{R} be the set of spatial relationships between objects. In this paper, we focus on two-dimensional relationships, i.e. $\mathcal{R} = \{\text{left}, \text{right}, \text{above}, \text{below}\}$, and 80 object categories derived from the MS-COCO dataset (Lin *et al.*, 2014). Then, for every $A \in \mathcal{C}$, $B \in \mathcal{C}$, and $R \in \mathcal{R}$, let the predicate $R(A, B)$ indicate that the spatial relationship R exists between object A and object B . For example `left(cat, dog)` describes a scene where a cat is to the left of a dog. For each pair, we construct 8 types of spatial relationships as shown below:

`left(A, B), right(A, B), above(A, B), below(A, B)`

`left(B, A), right(B, A), above(B, A), below(B, A)`

Sentence Generation. For each predicate $R(A, B)$, we convert it into a template `<A> <R> `, and paraphrase it into natural language. Appropriate articles “*a*”/“*an*” are prepended to object names A and B , to obtain four templates:

A/an <A> to the left of a/an
A/an <A> to the right of a/an
A/an <A> above a/an
A/an <A> below a/an

The template-based procedure has several advantages. First, it avoids linguistic ambiguity, subjectivity, and grammatical errors. Second, it is extensible to new object categories and additional spatial relationships. While we focus on two-dimensional relationships in this paper, our templates can be extended for generating test inputs for studying more complex spatial relationships and geometric features of objects, as we discuss in Sec. 9.7.

Dataset Statistics. We use $|\mathcal{C}| = 80$ object categories from MS-COCO and therefore obtain $\binom{80}{2} = 3160$ unique combinations of object pairs (A, B). For each pair, we construct 8 types of spatial relationships listed above, which leads to a total of $3,160 \times 8 = 25,280$ predicates. The SR_{2D} dataset contains 25,280 text examples, uniformly distributed across 80 COCO object categories, with each object being found in 632 images. Tab. 9.1 lists a few illustrative examples.

9.4 VISOR Metric

We propose VISOR as an automated metric for quantifying spatial understanding abilities of text-to-image models.

Definition 1 (Object Accuracy) *Let h be an oracle function that returns a set of detected objects in image x from set \mathcal{C} . Then, object accuracy for an image x , generated by a sentence containing objects A and B is:*

$$\text{OA}(x, A, B) = \mathbb{1}_{h(x)}(\exists A \cap \exists B). \quad (9.1)$$

t	x^1	x^2	x^3	x^4	$VISOR$	$VISOR_{1/2/3/4}$
An orange above a giraffe					50	100/100/0/0
An airplane to the right of a clock					0	0/0/0/0
A sports ball to the left of a bird					0	0/0/0/0
A surfboard above an oven					75	100/100/100/0
OVERALL	$VISOR_{cond} = \frac{5}{5+6} = 45.45\%$		$VISOR = \frac{5}{16} = 31.25\%$		$VISOR_{1/2/3/4} = 50 / 50 / 25 / 0$	
	$OA_x = 0$	$OA_x = 1; R_{gen} \neq R$			$OA_x = 1; R_{gen} = R$	

Figure 9.2: Examples illustrating the intuition behind OA, VISOR, $VISOR_{cond}$, and $VISOR_{1/2/3/4}$. **Purple box**: cases where one or both objects are not generated; **Red box**: both objects are generated but with a wrong spatial relationship; **Green box**: successful cases.

Note that, the oracle function h here could be either a pluggable learned model or a human detecting the presence of objects mentioned in the sentence. In our experiments, we show results for both cases and a correlational analyses between the two. Object accuracy is agnostic to the relationship R , whose presence is instead captured in the VISOR metric.

Definition 2 (VISOR) Let R_{gen} be the generated spatial relationship, while R is the ground-truth relationship mentioned in text. Then, for each image x ,

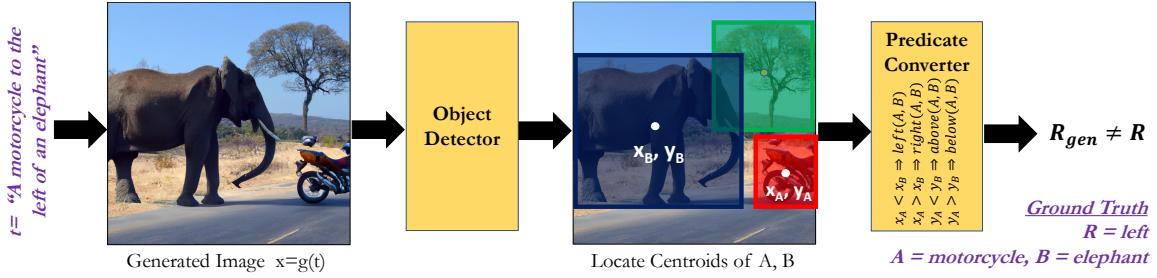


Figure 9.3: For text t and corresponding generated image $x = g(t)$, object centroids are located and converted into predicates indicating the spatial relationship between them. These predicates are compared with the ground truth relationship R to obtain the VISOR score.

$$\text{VISOR}(x, A, B, R) = \begin{cases} 1, & \text{if } (R_{gen} = R) \cap \exists A \cap \exists B \\ 0, & \text{otherwise.} \end{cases} \quad (9.2)$$

A useful feature of T2I models for artists and designers is the ability to generate multiple images for each input text prompt. This allows the creators to pick an appropriate image from N generated images. We define VISOR_n to reflect how good T2I models are at generating at least n spatially correct images given a text input that mentions a spatial relationship. From a usability perspective, the relaxed version of VISOR is useful for measuring if it is possible to find at least n images that would satisfy the input sentence where the task is one where the creator has the option to pick from the output image set.

Definition 3 (VISOR_n) VISOR_n is the probability of generating images such that for every text prompt t , at least n out of N images have $\text{VISOR}=1$:

$$\text{VISOR}_n(x, A, B, R) = \begin{cases} 1, & \text{if } \sum_{i=1}^N \text{VISOR}(x_i, A, B, R) \geq n \\ 0, & \text{otherwise.} \end{cases} \quad (9.3)$$

The relationship between VISOR and VISOR_n is given below. The proof is presented

in the supplementary materials.

$$\text{VISOR} = \frac{1}{N} \sum_{n=0}^N n(\text{VISOR}_n - \text{VISOR}_{n+1}). \quad (9.4)$$

In our study we use $N = 4$ images per text prompt and, therefore, report VISOR_1 , VISOR_2 , VISOR_3 , and VISOR_4 . Fig. 9.2 shows an example computation of all VISOR metrics.

Note that $\text{VISOR} = 1$ only if both objects are generated in the image, i.e. $\text{OA} = 1$. However, as we will see in Sec. 9.5, T2I models fail to generate multiple objects in a large subset of images. As such, it is important to disentangle the two abilities of the models to (1) generate multiple objects and (2) to generate them according to the spatial relationships described in the text of the prompt. For this purpose, we define conditional VISOR:

Definition 4 (Conditional VISOR) *is defined as the conditional probability of correct spatial relationships being generated, given that both objects were generated correctly.*

$$\text{VISOR}_{\text{cond}} = P(R_{\text{gen}}=R | \exists A \cap \exists B) \quad (9.5)$$

Implementation. The VISOR computation process is summarized in Fig. 9.3. Given any text prompt t and a T2I model g , we first generate images $x = g(t)$, and use an object detector to localize objects in x . Object accuracy OA is computed using Eq. (9.1). We obtain centroid coordinates of objects A and B from the the bounding boxes of the detected objects. Based on the centroids, we deduce the spatial relationship R_{gen} between them using the rules shown in the “Predicate Converter” box in Fig. 9.3. Finally, the generated relationship is compared with the ground-truth relationship R , and VISOR scores are computed using Eqs. (9.2), (9.3) and (9.5).

We use OWL-ViT ([Minderer et al., 2022](#)), a state of the art open-vocabulary object detector, with a CLIP backbone and ViT-B/32 transformer architecture and

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDER	SPICE	CLIPScore	VISOR
GLIDE	0.29 / 0	0.14 / 1.9e-6	0.05 / 2.9e-6	0.02 / 7.4e-7	0.13 / -9.9e-6	0.35 / -2.6e-6	0.18 / 1.1e-6	0.11 / -8.1e-6	0.70 / 1.3e-3	0.02 / 0.01
GLIDE + CDM	0.31 / 0	0.15 / 2.9e-5	0.06 / 7.6e-6	0.02 / 1.7e-6	0.15 / 6.4e-5	0.36 / 1.7e-6	0.22 / 3.1e-5	0.14 / -7.7e-5	0.75 / -8.8e-6	0.06 / 0.05
DALLE-mini	0.34 / 0	0.19 / -4.4e-5	0.09 / -1.6e-5	0.04 / -4.5e-6	0.19 / 2.1e-6	0.41 / -7.3e-6	0.34 / -6.0e-5	0.20 / 3.3e-5	0.80 / 1.5e-3	0.16 / 0.12
CogView-2	0.30 / 0	0.16 / 4.4e-6	0.07 / 1.3e-6	0.03 / 5.6e-6	0.16 / -6.6e-6	0.36 / -2.9e-6	0.25 / 8.7e-6	0.15 / 7.0e-5	0.72 / 1.5e-5	0.12 / 0.10
DALLE-v2	0.36 / 0	0.21 / -1.9e-5	0.11 / -4.8e-6	0.04 / -1.5e-6	0.21 / 1.7e-4	0.44 / 8.8e-6	0.40 / -2.8e-5	0.22 / -4.1e-5	0.84 / 1.8e-3	0.38 / 0.28
SD	0.33 / 0	0.18 / 3.3e-6	0.08 / 9.7e-7	0.03 / 2.8e-7	0.19 / 1.0e-5	0.40 / -2.6e-6	0.31 / 4.3e-6	0.19 / 7.4e-5	0.79 / 1.5e-3	0.19 / 0.15
SD + CDM	0.32 / 0	0.17 / 1.1e-5	0.07 / 5.1e-6	0.03 / 1.3e-6	0.17 / 1.6e-4	0.38 / 4.4e-6	0.28 / 1.2e-5	0.18 / -4.5e-5	0.77 / 3.6e-4	0.15 / 0.12
SD 2.1	0.35 / 0	0.20 / -1.3e-5	0.09 / 4.2e-6	0.038 / -1.3e-6	0.20 / 7.1e-5	0.42 / 5.4e-6	0.35 / -1.8e-5	0.20 / 3.5e-5	0.82 / 1.0e-3	0.30 / 0.24

Table 9.2: s/Δ_s scores for T2I metrics shown in the 0 to 1 range. All previous metrics have low Δ_s (**magenta**) whereas VISOR has high Δ_s (**green**), showing they are ineffective in quantifying and benchmarking spatial understanding. s/Δ_s for all VISOR variants are in Supp.Mat.

confidence threshold 0.1. The supplementary material also contains results using DETR-ResNet-50 (Carion *et al.*, 2020) trained on MS-COCO. The results using both object detectors are similar and lead to an identical ranking of models in our benchmark. However, the open-vocabulary functionality of OWL-ViT ensures that VISOR is widely applicable to other datasets, categories, and vocabularies. This removes dependence on specific datasets, making VISOR widely applicable for any freeform text input.

9.5 Experiments

Baselines. We study state-of-the-art T2I models as baselines: GLIDE (Nichol *et al.*, 2022), DALLE-mini (Dayma *et al.*, 2021), CogView2 (Ding *et al.*, 2022), DALLE-v2 (Ramesh *et al.*, 2022), and Stable-Diffusion (SD and SD 2.1.) (Rombach *et al.*, 2022), and two versions of Composable Diffusion Models (Liu *et al.*, 2022b) (GLIDE + CDM and SD + CDM). We generate $N=4$ images for each text prompt from our SR_{2D} dataset, to obtain 126,720 images per model and compare performance in terms of

Model	OA (%)	VISOR (%)					
		uncond	cond	1	2	3	4
GLIDE (Nichol <i>et al.</i> , 2022)	3.36	1.98	59.06	6.72	1.02	0.17	0.03
GLIDE + CDM (Liu <i>et al.</i> , 2022b)	10.17	6.43	63.21	20.07	4.69	0.83	0.11
DALLE-mini (Dayma <i>et al.</i> , 2021)	27.10	16.17	59.67	38.31	17.50	6.89	1.96
CogView2 (Ding <i>et al.</i> , 2022)	18.47	12.17	65.89	33.47	11.43	3.22	0.57
DALLE-v2 (Ramesh <i>et al.</i> , 2022)	63.93	37.89	59.27	73.59	47.23	23.26	7.49
SD (Rombach <i>et al.</i> , 2022)	29.86	18.81	62.98	46.60	20.11	6.89	1.63
SD + CDM (Liu <i>et al.</i> , 2022b)	23.27	14.99	64.41	39.44	14.56	4.84	1.12
SD 2.1	47.83	30.25	63.24	64.42	35.74	16.13	4.70

Table 9.3: Comparison of the performance of all models in terms of object accuracy (OA) and each version of VISOR.

OA, VISOR, $\text{VISOR}_{\text{cond}}$, and $\text{VISOR}_{1/2/3/4}$.

9.5.1 Ineffectiveness of Existing Metrics

T2I models have been primarily compared in terms of photorealism (purely visual) and human judgment about image quality (subjective). We quantify whether existing automated multimodal metrics are useful for evaluating spatial relationships generated by T2I models. We consider CLIPScore (Hessel *et al.*, 2021) (cosine similarity between image and text embeddings) and image captioning-based evaluation (BLEU (Papineni *et al.*, 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), CIDEr (Vedantam *et al.*, 2015b), SPICE (Anderson *et al.*, 2016)) which are used by generating a caption c for the synthesized image $x = g(t)$ and computing the captioning score with respect to the reference input text t . Note that purely visual metrics (FID and Inception Score (Heusel *et al.*, 2017; Salimans *et al.*, 2016)) ignore the text, while

semantic object accuracy ([Hinz et al., 2020](#)) ignores all words except nouns, making them incapable of scoring spatial relationships.

Let s^t be the score for (x, t) where x is the generated image and t is the input text. Let t_{flip} be the transformed version of t obtained by inverting/flipping the spatial relationship in t (for example, left→right). Let s_{flip}^t be the score for (x, t_{flip}) . For each metric, we define Δ_s as the average difference between s^t and s_{flip}^t over the entire SR_{2D} dataset:

$$\Delta_s = \mathbb{E}_t[s^t - s_{flip}^t], \quad (9.6)$$

Thus, Δ_s captures the ability of metric s to understand spatial relationships. Table 9.2 shows s and Δ_s values for each previous metric and VISOR for each model. It can be seen that, for all previous metrics, Δ_s is negligible and close to zero, which implies that they return similar scores even if the text is flipped. For some cases, the difference is negative, implying that the score for the image and the flipped caption is higher. On the other hand, the Δ values for VISOR are high implying that VISOR assigns significantly lower scores for the flipped samples. These results establish the need for a new evaluation metric since none of the existing metrics are able to quantify spatial relationships reliably, and show the efficacy of VISOR for this purpose.

9.5.2 Benchmarking Results

Table 9.3 shows the results of benchmarking on our SR_{2D} dataset. We first note that the object accuracy of all models except DALLE-v2 is lower than 30%. While DALLE-v2 (63.93%) significantly outperforms other models, it still shows a large number of failures in generating both objects that are mentioned in the prompt. For the unconditional metrics VISOR and VISOR_{1/2/3/4}, DALLE-v2 is the best performing model. However, in terms of VISOR_{cond}, CogView2 has the highest performance. This implies that, although CogView2 is better than other models on those examples

	<p>(1) Rate the quality of the image. <small>("1" being artificial (e.g. a sketch or cartoon) and "5" being natural (a real photograph))</small></p> <p><input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input checked="" type="radio"/> 4 <input type="radio"/> 5</p> <p>(2) How likely is the scene to occur in real life? <small>(Rate from "1" (least likely) to "5" (most likely))</small></p> <p><input type="radio"/> 1 <input type="radio"/> 2 <input type="radio"/> 3 <input type="radio"/> 4 <input checked="" type="radio"/> 5</p> <p>(3) How many objects are in this image?</p> <p><input type="text" value="3"/> <input type="button" value="▼"/></p> <p>(4) Object A: The image contains a wine glass</p> <p><input checked="" type="radio"/> True <input type="radio"/> False</p> <p>(5) Object B: The image contains a sandwich</p> <p><input checked="" type="radio"/> True <input type="radio"/> False</p> <p>(6) Choose the spatial relationship between the wine glass and the sandwich. <small>Multiple Options may be possible. If there are more instances of the same type (example: two dogs and one cat) then select all possible relationships between each dog and the cat. [IMPORTANT] Choose "N/A" if you answered "False" for either question (2) or (3)</small></p> <p><input checked="" type="checkbox"/> wine glass to the left of sandwich <input type="checkbox"/> wine glass to the right of sandwich <input type="checkbox"/> wine glass above sandwich <input type="checkbox"/> wine glass below sandwich <input type="checkbox"/> N/A</p> <p>(7) If you answered True for both (4) and (5), are the two objects merged or distinct</p> <p><input type="radio"/> Merged <input checked="" type="radio"/> Distinct <input type="radio"/> N/A</p> <p><small>[IMPORTANT] Choose "N/A" if you answered "False" for either question (2) or (3)</small></p>						
---	---	--	--	--	--	--	--

Figure 9.4: The human study interface with an image on the left and seven multiple choice questions about it.

where both objects are generated, the large failures of CogView2 in OA result in a lower unconditional VISOR score. VISOR₄ is extremely low for all models including DALLE-v2 (8.54%), revealing a large gap in performance.

9.5.3 Human Study

Methodology. We conducted a human evaluation study to understand the alignment of our metrics with human judgment, and to quantify the gap between object detector performance and human assessments of object presence. For the human study, we used four models: CogView2, DALLE-v2, Stable Diffusion (SD), and SD + CDM. Annotators were shown (via Amazon Mechanical Turk) an image generated by one of the four models, and were asked seven questions about it, as shown in Fig. 9.4. The questions assessed human evaluation of image quality and scene realism (*scene likelihood*) on a Likert scale (1 through 5), the number of objects, answering True or False for presence of objects, selecting valid spatial relationships, and responding if two objects were merged in the image. We used a sample size of 1000 images per model and 3 workers per sample.

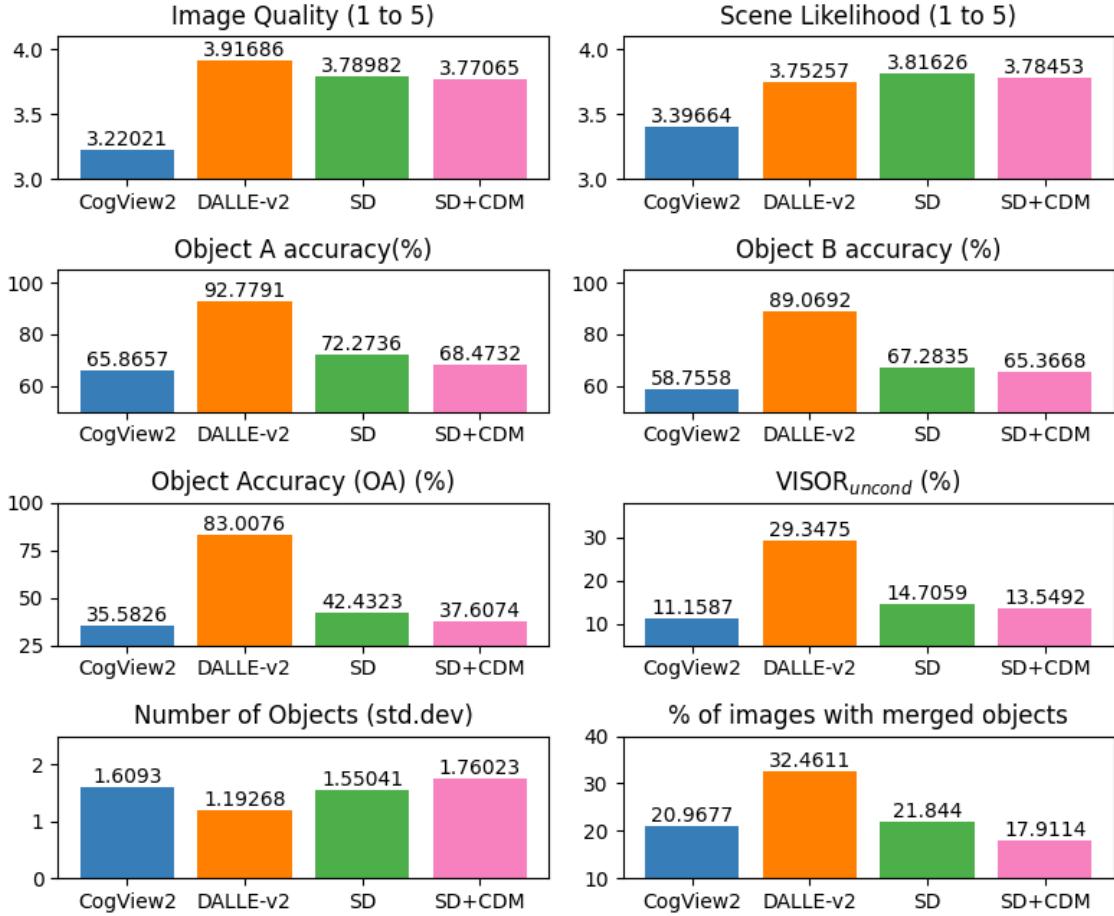


Figure 9.5: Summary of responses to each question in the human study, compared across all four models.

Results. Fig. 9.5 shows a summary of responses for each question in the human study. While DALLE-v2 received the highest image quality rating, SD and SD+CDM received higher scene likelihood rating. Interestingly, DALLE-v2 also had the largest number of images with merged objects (32.46%); several cases of this phenomenon are shown in Fig. 9.15. Inter-annotator agreement was high for all questions in terms of majority (agreement between at least 2 out of 3 workers) and unanimous agreement (agreement between all 3 out of 3 workers) as reported in Table 9.4.

Response	CogView2	DALLE-v2	SD	SD + CDM
Image Quality	65.47 / 52.93	75.02 / 62.33	69.86 / 55.31	72.59 / 57.99
Scene Likelihood	64.40 / 50.78	72.13 / 59.62	69.47 / 52.35	67.19 / 53.99
Num. Objects	79.63 / 50.03	87.09 / 46.39	81.41 / 46.06	80.28 / 45.74
Object A	100.0 / 33.00	99.64 / 8.02	100.0 / 18.56	100.0 / 20.04
Object B	100.0 / 32.75	100.0 / 13.39	100.0 / 22.44	100.0 / 25.51
Spatial Relation	100.0 / 23.33	100.0 / 47.90	100.0 / 30.79	100.0 / 25.00
Merged/Distinct	100.0 / 43.02	99.64 / 58.85	100.0 / 39.95	100.0 / 38.60

Table 9.4: Majority / Unanimous inter-worker agreement (%) for each question in our human study.

Metric	CogView2	DALLE-v2	SD	SD-CDM
OA	73.07	73.87	79.25	80.21
VISOR	88.48	77.41	88.43	88.80
VISOR _{cond}	75.02	75.62	76.95	74.69

Table 9.5: Agreement(%) of human responses with automated metrics

Alignment of VISOR with Human Responses. We observe that the ranking of models in terms of both object accuracy (OA) and VISOR is identical for the human study and for the automated VISOR scores in Table 9.3, i.e. *DALLE-v2* \dot{e} *SD* \dot{e} *SD-CDM* \dot{e} *CogView2*. Table 9.5 shows the percentage of samples for which responses from humans matched our automated evaluation using object detectors.

9.6 Analysis

Qualitative Results. Fig. 9.6 shows examples of images generated by all baselines for each prompt, with more visualizations in the appendix. Although the photorealism



Figure 9.6: Illustrative examples of text prompts from our SR2D dataset and corresponding images generated by each T2I model.

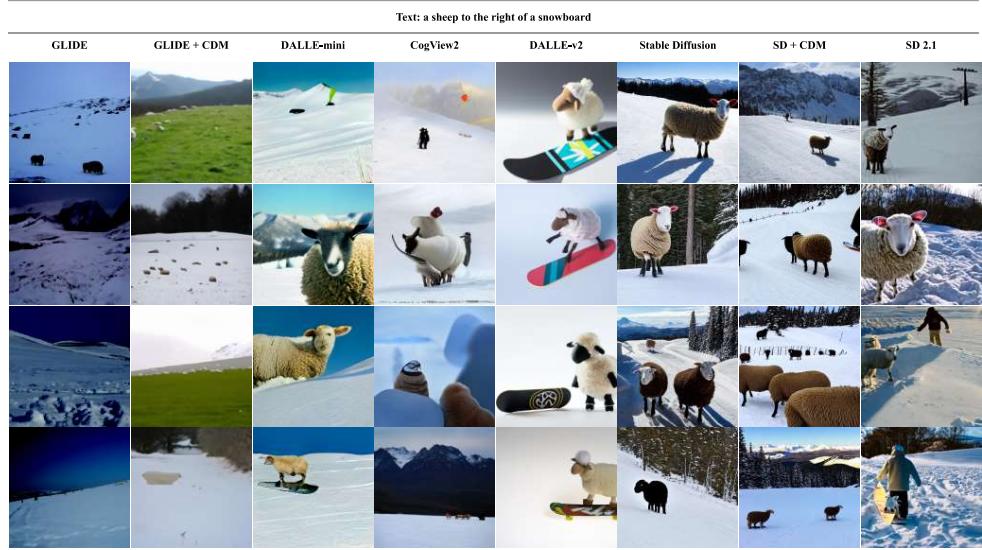


Figure 9.7: Illustrative examples of images generated by each of the 7 benchmark models using text prompts (top row) from the SR_{2D} dataset.

of recent models, such as DALLE-v2, SD, and SD+CDM, is much higher, all models are equally poor at generating accurate spatial relationships. More examples are shown in Figures 9.7, 9.8, 9.9, 9.10, 9.11, 9.12, 9.13, 9.14.

Merged Objects. Fig. 9.15 shows examples of a few common types of merging between objects that we observed, especially with DALLE-v2. Common patterns observed include animals being rendered as patterns on inanimate objects (a, b) and both objects retaining their typical shape but getting merged (c, d). As our human study in Fig. 9.5 shows, a large proportion (more than 20%) of images have merged objects – this poses a significant challenge for generating distinct objects and their relationships using T2I models.

Performance per relationship is shown in Table 9.6. Interestingly, five of the seven models have the best VISOR_{cond} scores for horizontal relationships (left or

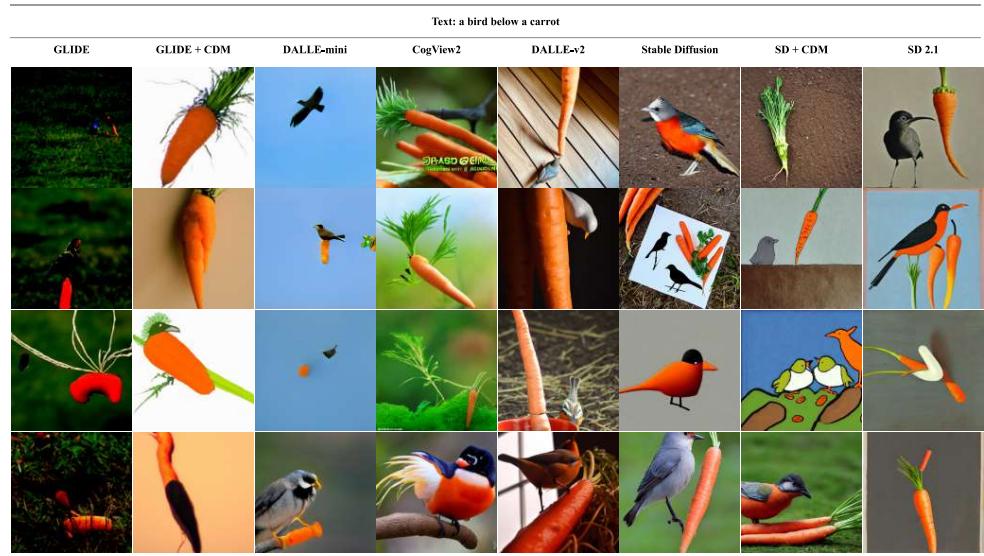


Figure 9.8: Illustrative examples of images generated by each of the 7 benchmark models using text prompts (top row) from the SR_{2D} dataset.

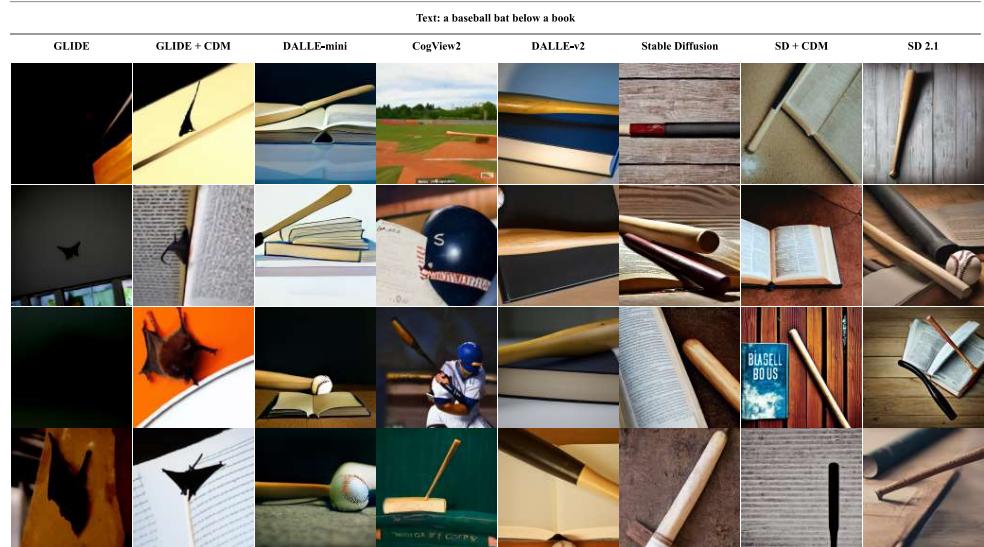


Figure 9.9: Illustrative examples of images generated by each of the 7 benchmark models using text prompts (top row) from the SR_{2D} dataset.

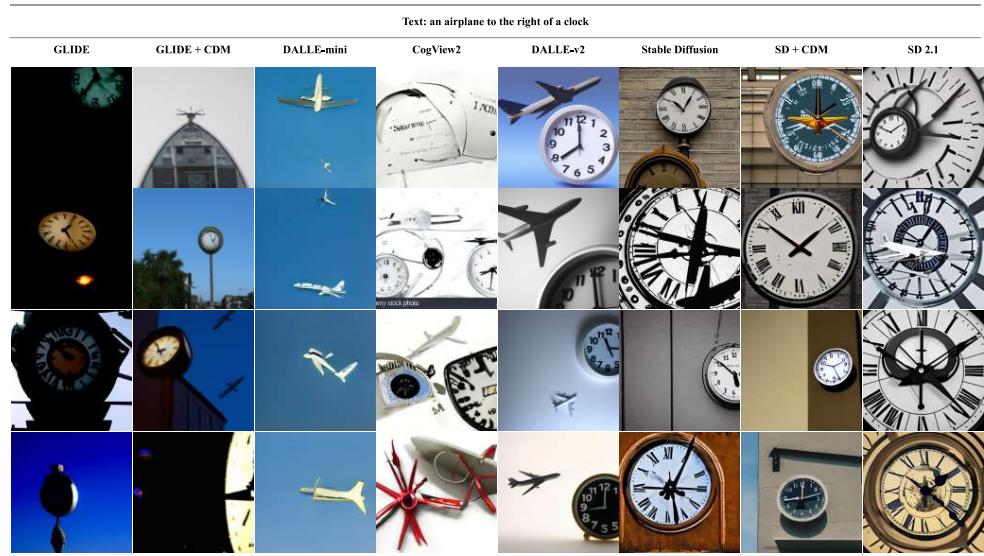


Figure 9.10: Illustrative examples of images generated by each of the 7 benchmark models using text prompts (top row) from the SR_{2D} dataset.

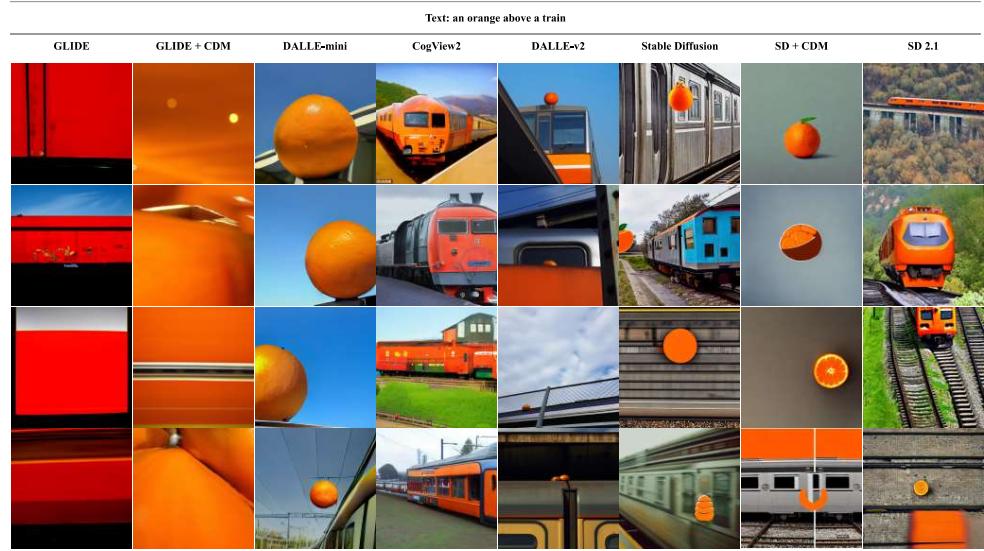


Figure 9.11: Illustrative examples of images generated by each of the 7 benchmark models using text prompts (top row) from the SR_{2D} dataset.

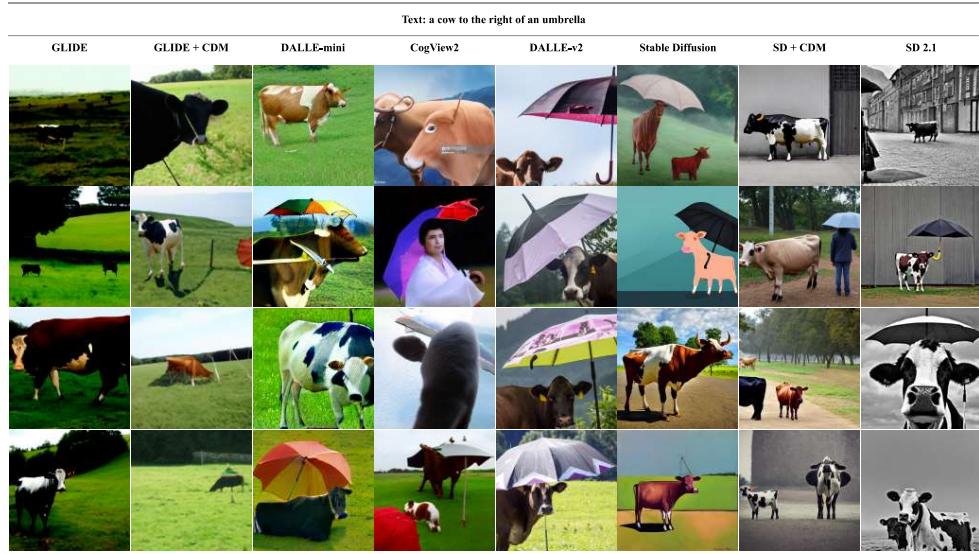


Figure 9.12: Illustrative examples of images generated by each of the 7 benchmark models using text prompts (top row) from the SR_{2D} dataset.

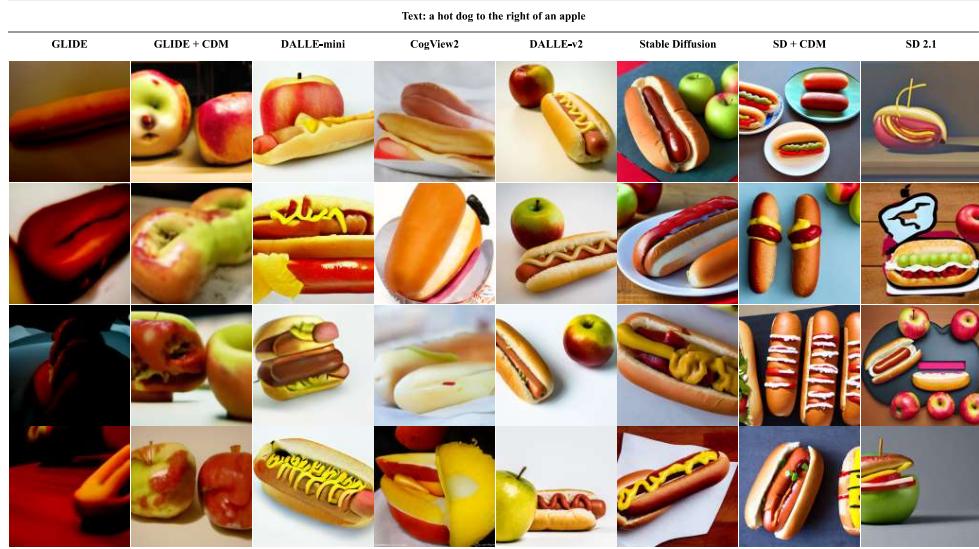


Figure 9.13: Illustrative examples of images generated by each of the 7 benchmark models using text prompts (top row) from the SR_{2D} dataset.

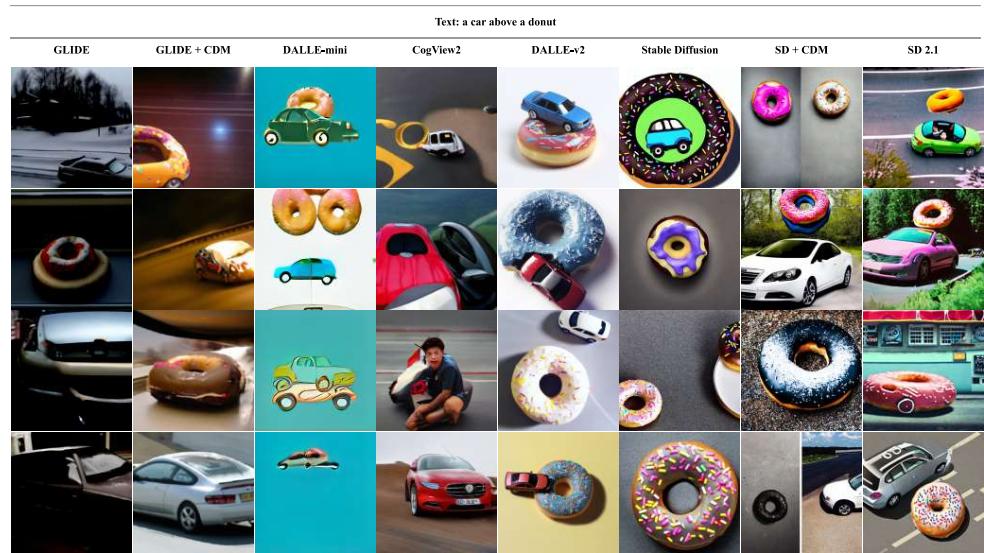


Figure 9.14: Illustrative examples of images generated by each of the 7 benchmark models using text prompts (top row) from the SR_{2D} dataset.

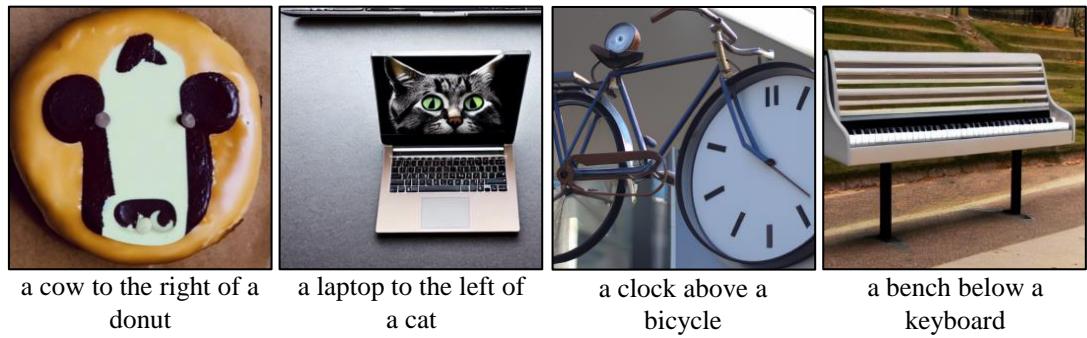


Figure 9.15: Illustrative examples where the two objects from the text input appear to be merged. From left to right: a, b, c, d.

Model	VISOR _{cond} (%)				Object Accuracy (%)			
	left	right	above	below	left	right	above	below
GLIDE	57.78	61.71	60.32	56.24	3.10	3.46	3.49	3.39
GLIDE + CDM	65.37	65.46	59.40	59.84	12.78	12.46	7.75	7.68
DALLE-mini	57.89	60.16	63.75	56.14	22.29	21.74	33.62	30.74
CogView2	68.50	68.03	63.72	62.51	20.34	19.30	17.71	16.54
DALLE-v2	56.47	56.51	60.99	63.24	64.30	64.32	65.66	61.45
SD	64.44	62.73	61.96	62.94	29.00	29.89	32.77	27.8
SD + CDM	69.05	66.52	62.51	59.94	23.66	21.17	23.66	24.61

Table 9.6: Comparison of Visor and OA split by relationship type

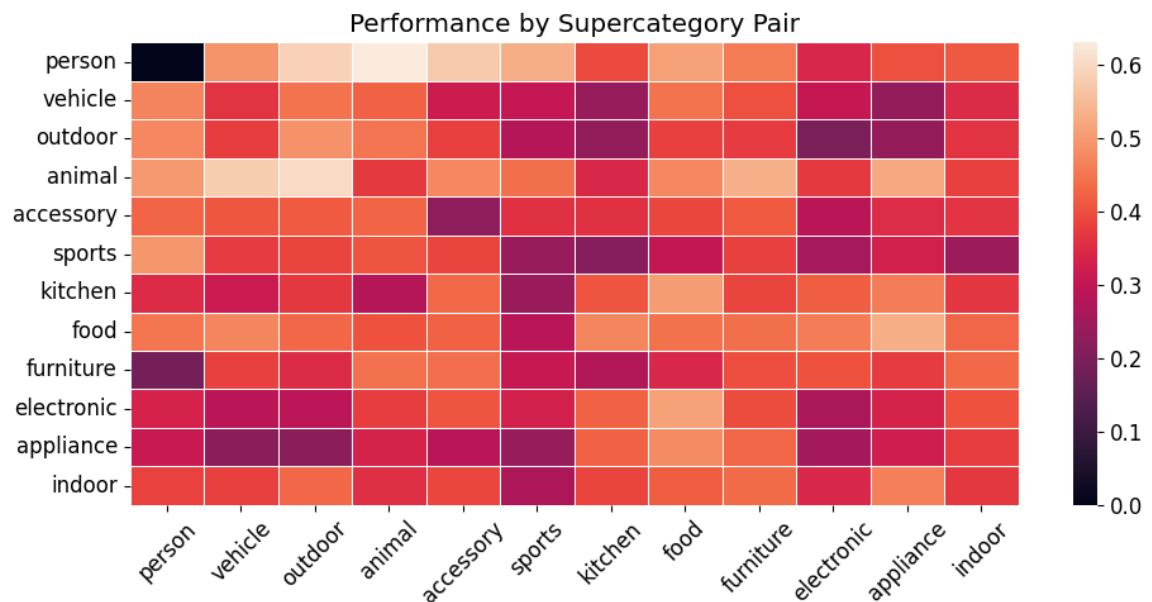


Figure 9.16: VISOR scores for each supercategory pair.

right). However, five of the seven models have the best object accuracy for vertical relationships (above or below).

Performance per Supercategory. The 80 object categories in SR_{2D} belong to 11 MS-COCO “supercategories”. We investigate VISOR scores for each supercategory pair and report the results for the best model (DALLE-v2) in Fig. 9.16 (results for other models are in the appendix). VISOR scores for commonly co-occurring supercategories such as “*animal, outdoor*” are highest whereas unlikely combinations of indoor-outdoor objects such as “*vehicle, appliance*” and “*electronic, outdoor*” have low performance.

Correlation between VISOR and Object Co-occurrence. The object categories in our dataset span a wide range of commonly occurring objects from MS-COCO such as wild animals, vehicles, appliances, and humans, found in varying contexts, including combinations that do not appear together often in real life. For instance, an elephant is unlikely to be found indoors near a microwave oven. To understand how object co-occurrence affects VISOR, we first obtain $P_{\text{COCO}}(A, B)$, the probability of co-occurrence for each object pair (A, B) as a proxy for real-world object co-occurrence. Then, we plot the correlation of VISOR and object accuracy for pair (A, B) with its $P_{\text{COCO}}(A, B)$. As Fig. 9.17 shows, the correlation is positive for all models, for both OA and $\text{VISOR}_{\text{cond}}$, implying that the quality of outputs is likely to be better for commonly co-occurring objects, clearly establishing a bias towards real-world likelihood. This correlation shows the difficulty in generating unlikely relationships such as “*an elephant to the left of a microwave*” even though such unlikely combinations may be desired by creators, pursuing artistic compositions.

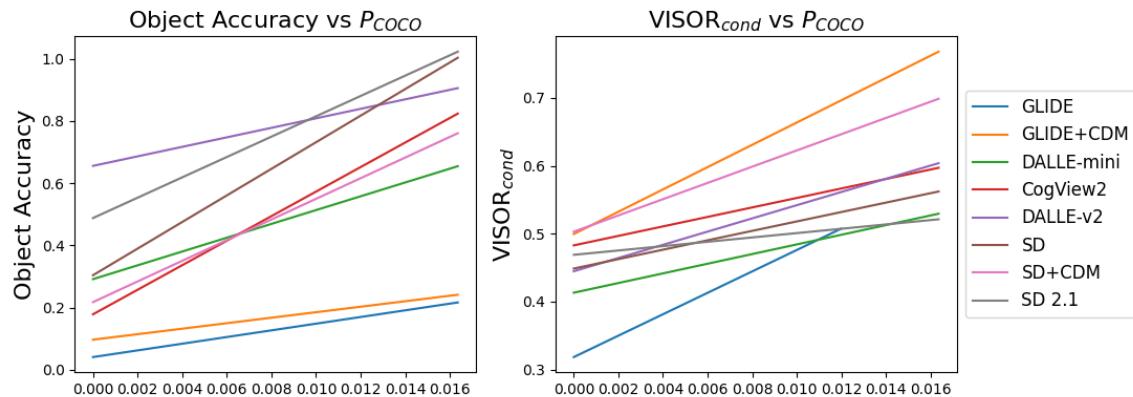


Figure 9.17: Correlation of our metrics with P_{COCO} , the object co-occurrence probability in MS-COCO.

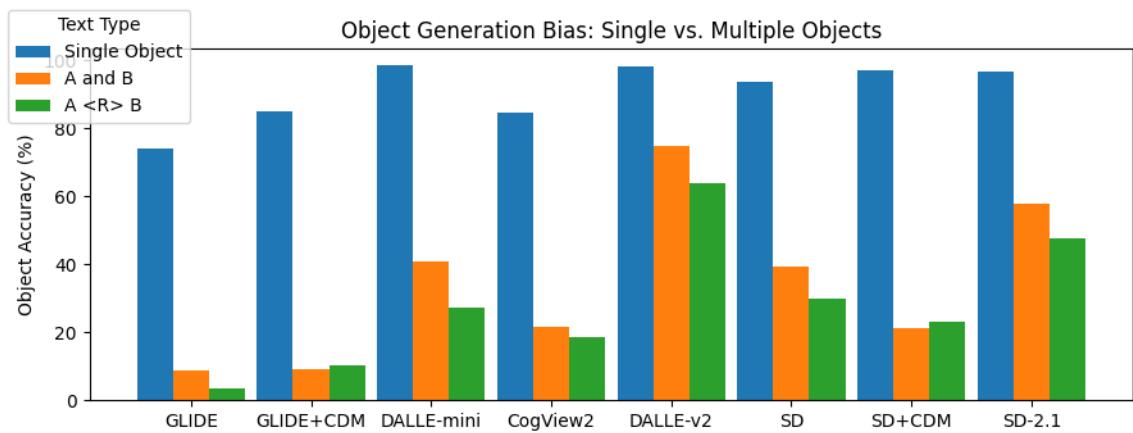


Figure 9.18: Comparison of object accuracy for text with single and multiple objects reveals a bias towards single objects.

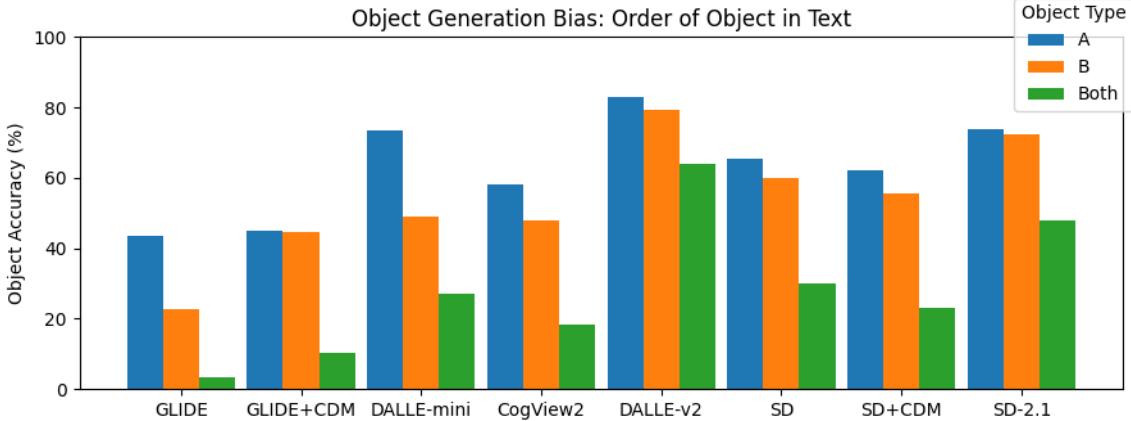


Figure 9.19: Comparison of object accuracy for object A and B reveals a bias towards A, the first object appearing in the prompt).

Object Generation Bias. We compare object accuracy with three types of inputs to generate images: (1) single objects text such as “*an elephant*”, (2) multiple object conjunction such as “*an elephant and a cat*”, and (3) relational texts such as “*an elephant to the right of a cat*”. Fig. 9.18 shows that, for all models, OA is significantly higher for single objects; while composition using conjunction is challenging, systems perform better with this generation than spatial composition.

Text-Order Bias. In Fig. 9.19, we show that for all models, OA for the first mentioned object (A) in the text is significantly higher than OA for the second object (B); generating both objects together is most challenging.

Consistency between equivalent phrases. Ideally, given two equivalent inputs such as “*a cat above a dog*” and “*a dog below a cat*”, the model should generate images with the same spatial relationship. To evaluate this consistency, we consider cases in which both objects are detected and report the consistency for each relationship type in Table 9.7. Surprisingly, the best model *DALLE-v2* is the least consistent, while

Model	left	right	above	below	Average
GLIDE	<u>45.90</u>	58.93	63.16	52.63	55.16
GLIDE + CDM	61.99	59.15	54.79	56.15	58.02
DALLE-mini	54.75	52.28	54.64	55.77	54.36
CogView2	67.32	65.38	65.67	66.95	66.33
DALLE-v2	48.81	<u>48.10</u>	<u>48.72</u>	<u>48.15</u>	<u>48.45</u>
SD	58.71	61.36	55.36	55.39	57.71
SD + CDM	64.69	65.71	61.35	57.71	62.37
SD 2.1	53.96	55.50	54.73	54.38	54.64

Table 9.7: Consistency (%) of generated spatial relationships for equivalent inputs. Bold: highest, Underline: lowest consistency.

CogView2 is the most consistent model. This result shows that merely rephrasing the input can have a large influence on the spatial correctness of the output.

Effect of Attributes on Spatial Understanding. We conduct a case study with Stable Diffusion (SD) to seek an understanding of the impact of sentence complexity on a model’s VISOR performance. We increase the complexity of text prompts by randomly assigning two attributes (size Z and color C) to the object category, via templates of the form $[Z_A] [C_A] <A> <R> [Z_B] [C_B] $. We focus on 11 object categories representative of each supercategory in COCO, 8 colors, and 4 sizes. As shown in Fig. 9.20, compared to generation without attributes, there is a drop in performance in 13 out of 15 types of attribute combinations. Addition of the color attribute (C) leads to a large drop in performance. Adding size descriptors (Z) may improve performance. While concurrent work ([Feng et al., 2022](#)) has reported difficulty

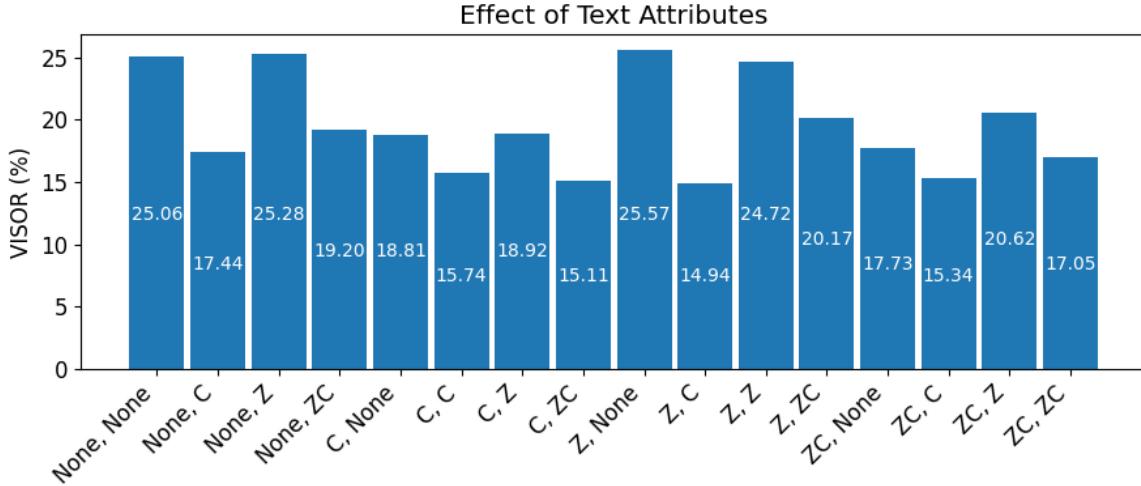


Figure 9.20: Comparing VISOR performance with different combinations of attributes. “Z, ZC” indicates a prompt describing object A with a size attribute and object B with both size and color.

Prompt Type	OA (%)	VISOR (%)					
		uncond	cond	1	2	3	4
Phrases	29.86	18.81	62.98	46.60	20.11	6.89	1.63
Sentences	32.48	20.67	63.64	48.54	22.94	8.92	2.25
Split Sentences	24.98	16.44	65.82	41.91	16.29	5.66	1.91

Table 9.8: Effect of prompt variations on OA and VISOR scores. All three versions use the same Stable Diffusion (SD) model .

in attribute binding, our analysis suggests that attributes may negatively influence spatial compositionality.

Effect of Rephrased Text Prompts. We compare variations of the prompt: (1) *phrases*: the default version of SR_{2D} used in Tab. 9.3 (for e.g. “a cat to the left of a dog”); (2) *sentence* (for e.g. “There is a cat to the left of a dog”), and (3) *split*

sentences (for e.g. “There is a cat to the left. There is a dog to the right”). Compared to phrases, Tab. 9.8 shows higher OA and VISOR_{cond} for *sentences*; lower OA and higher VISOR_{cond} for *split sentences*. Prompt engineering for grounded generation such as spatial aspects is a promising future direction.

9.7 Discussion and Conclusion

We studied the spatial capabilities of text-to-image generators by introducing spatial relationship metrics (VISOR measures), building a dataset (SR_{2D}), and developing an automated evaluation pipeline. Our experiments reveal that existing T2I models have poor spatial interpretation and rendering abilities, as characterized by their low VISOR scores, making them unreliable for uses that depend on the correctness in generated images of spatial relations specified in prompts. Our analysis also reveals several biases and artifacts of T2I models, such as proclivity for generating single objects (especially the first mentioned object), correlation of spatial correctness with likelihood of object co-occurrence, sensitivity to equivalent phrasings of spatial relations, and negative influences of the inclusion in prompts of several commonly used modifiers. We hope that the metrics, methods, and dataset will help to stimulate a stream of research on the spatial rendering capabilities of generative models, leading to enhancements of these capabilities over time. For uses of today’s technologies, we hope our findings can provide creators with guidance for prompt engineering. We note that the SR_{2D} data generation pipeline can be extended to study spatial relationships of more than two objects, including three-dimensional and complex relations such as *inside*, *outside*, *contains*, *behind*, *in front*, *covers*, *touching*, as well as semantic and action-based relationships. We hope the VISOR metric will serve as a complement to prior metrics for evaluating photorealism and image-text similarity.

Chapter 10

CONCLUSIONS

This chapter concludes the dissertation with an outline of key findings, contributions, and broader impact of the research presented in the previous chapters.

10.1 Summary of Contributions

The primary goal of this work was to study the reliability of computer vision models from a holistic perspective – the perspective which states that we must view reliability from two lenses:

1. a machine learning lens which seeks to better understand and measure the failures of computer vision systems in terms of different measurable quantities such as accuracy, distributional robustness, adversarial robustness, uncertainty quantification, etc.
2. a human-centered lens which seeks to improve the utility of exciting recent advances in semantic vision by offering benchmarking tools, evaluation protocols and metrics, and by developing new functionalities that allow non-expert users to interact and collaborate with these models.

In this dissertation, we have made contributions via both of these lenses – this is only the first step towards a larger goal. This is an exciting research frontier with many impending research problems that need to be addressed in order to take another step towards reliable semantic vision.

In Chapters 2 and 3 we saw how the idea of discovering data transformations during training was a powerful concept to improve the diversity of training data; we

saw how an adversarial training pipeline guided by such transformations and domain knowledge can help improve the robustness of image classifiers. In Chapters 4 we complemented these improvements in domain generalization with a mechanism to flag out-of-domain inputs – this method turned out to be effective for both image classification as well as text classification tasks.

Multi-modal tasks involving both vision and language (V&L) inputs, such as visual question answering (VQA), open up intriguing domain discrepancies that can affect model performance of test time. For the VQA task, models are trained to predict the answers to questions about images. In Chapter 6 we discovered that existing VQA models fail when logical transformations such as negation, conjunction, and disjunction are introduced in the questions. We built on this surprising finding to develop a data augmentation tool that produces logical combinations of multiple questions. We then designed a logic-inspired training objective based on Frechet inequalities to guide the predicted probabilities of answers to questions with logical connectives. VQA-LOL was instrumental as a reality check for VQA performance and was included (by other researchers) as part of a compendium of datasets for testing VQA robustness ([Li *et al.*, 2020b](#)). VQA-LOL led to a series of papers ([Gokhale *et al.*, 2020b](#)), ([Gokhale *et al.*, 2022a](#)), ([Varshney *et al.*, 2022](#)) that adopted linguistic and semantic transformations for image–text alignment, video–text reasoning, and natural language inference.

In Chapter 7, we identified that knowledge of linguistic transformations can inform the algorithm design for V+L tasks. This led to the development of the “SISP transformation” suite – a controlled method to semantically manipulate text to generate augmented data that is semantics-inverting (SI) or semantics-preserving (SP). I showed that these SISP transformations can be leveraged to train robust models by developing a new knowledge-guided adversarial training algorithm called The combination of SISP (data engineering) and SDRO (robust optimization) led to

improvements on image-based reasoning, video-based reasoning, and visual question answering, along several dimensions of robustness – in-domain and out-of-domain accuracy, adversarial robustness, and calibration, and also on my previous VQA-LOL benchmark ([Gokhale *et al.*, 2020c](#)).

We further expanded our approach for synthetic data generation that enabled design of weakly-supervise VQA models for limited-data settings ([Banerjee *et al.*, 2021a](#)), ([Banerjee *et al.*, 2021c](#)), and for creating video QA benchmarks for reasoning about physical properties of objects ([Patel *et al.*, 2022](#)), in Chapter 8 for commonsense reasoning about people’s actions ([Fang *et al.*, 2020](#)), and in Chapter 9 for benchmarking spatial reasoning abilities of text-to-image generative models.

10.2 Impact

AI has undergone a paradigm shift in the past decade – the connection between vision and language (V+L) is now an integral part of AI, with deep impact beyond vision and NLP – robotics, graphics, cybersecurity, and HCI are utilizing V+L tools and there are direct industrial implications for software, arts, and media. As V+L models are being widely adopted, new types of challenges and failure modes are emerging due to the multimodal and non-trivial relationships between images and text (as I have shown through my research). This means that we will need to simultaneously (1) discover failure by rigorous testing and benchmarking and (2) develop exciting new functionalities and capabilities with improved accuracy. The biggest challenge in robust multimodal learning is the scarcity of task-specific and functionality-specific data. While recent pre-trained models use millions of image-text pairs from the web to learn representations – they often fail when reasoning capabilities and fine-grained understanding is required. My research identifies these performance gaps and offers the unique combination of **semantic data engineering** and **knowledge-guided**

adversarial training as a solution.

The goal of this dissertation : to research and develop robust and reliable AI systems by leveraging the complex interactions between vision and language, has allowed an interdisciplinary work at the wonderful intersection of machine learning, computer vision, and natural language processing. The findings from my research together show that active design and discovery of data transformations and adversarial training algorithms is the key for improving robustness, in multimodal (V+L) tasks as well as robust image classification. This work has been published in premier AI, vision, and NLP conferences, and has served as the foundation of grant proposals that I helped write. My research contributions mentioned above have had a transformative effect on discourse about reliability of multimodal vision-language systems. The methodologies, frameworks, and techniques that I have developed find applications beyond vision and language and I am excited to build on them to solve important societal problems and civilizational challenges.

I have led collaborative projects with ASU, Lawrence Livermore National Laboratory, Microsoft Research, Carnegie Mellon, and Adobe Research. This work has coincidentally been directly aligned with the recent clarion calls for safe and robust AI from government agencies (DARPA¹, White House OSTP²), and academia (ACL³, AAAI⁴).

My work has been appeared in several top-tier publications including computer vision venues such as ICCV, ECCV, WACV; NLP venues such as ACL, NAACL, EMNLP; as well as AAAI, along with workshop articles at CVPR, ICLR, and NeurIPS. In recognition of my work, I have been invited to be part of program committees of

¹<https://www.darpa.mil/work-with-us/ai-next-campaign>

²<https://www.whitehouse.gov/ostp/ai-bill-of-rights/>

³https://2023.aclweb.org/calls/main_conference/#theme-track-reality-check

⁴<https://aaai.org/Conferences/AAAI-23/safeandrobustai/>

prestigious conferences such as CVPR, ICCV, ECCV, ICML, ICLR, NeurIPS, AAAI, ACL, NAACL, EMNLP, IROS, ICRA, RA-L, as well as several workshops and demo tracks at these conferences. I have also been invited to host two instances (2022 and 2023) of the CVPR workshop on Open-Domain Reasoning Under Multimodal Settings (ODRUM) and the WACV tutorial on Semantic Data Engineering under Multimodal Settings (SERUM).

I have also collaborated with other researchers in ASU leading to multiple publications such as debiasing vision–language datasets ([Luo *et al.*, 2022b](#)), improving biomedical information retrieval ([Luo *et al.*, 2022a](#)), and learning to generate images from sparse semantic label maps ([Kulkarni *et al.*, 2021](#)). Finally my research activities have provided research opportunities for masters and doctoral students who I have had the privilege to mentor ([Gokhale *et al.*, 2022a](#); [Patel *et al.*, 2022](#); [Wisdom *et al.*, 2023](#)).

10.3 Future Research Agenda

Over the last decade, the nature of AI research has changed considerably. Research communities that were largely isolated are now actively leveraging the connecting elements between the visual world and human-assigned meaning (language). However the link between V and L goes beyond image–text similarity. Language is ideally suited for developing reasoning capabilities beyond the visible – these reasoning capabilities are key for allowing V+L models to interact with humans. My goal is to integrate vision, language, and human collaboration together for active decision making, complex reasoning, and for learning novel concepts, without sacrificing robustness.

In the short term, my focus will be on developing reasoning capabilities that are geared towards *correctness* of outputs, for instance reasoning about spatial relationships and scene geometry and reasoning about everyday actions.

Spatial reasoning is a fundamental aspect of computer vision. In WeaQA (Banerjee *et al.*, 2021b) we showed that VQA models lacked this understanding, but their performance can be improved via weak geometric supervision. My ongoing work involves investigating the spatial understanding of text-to-image synthesis (T2I) – I am developing an evaluation framework called “VISOR” for quantifying the fidelity of T2I models in generating spatial relationships between objects. VISOR reveals the surprising finding that although recent SOTA models like DALLE exhibit high photorealism, they are ineffective in composing images with two or more distinct objects, especially when a spatial relationship such as left/right/above/below is specified. I plan to explore this direction further and develop prompting and finetuning techniques to improve spatial reasoning of image generators. T2I models are ideally poised to serve a crucial purpose in computer vision research – my research will investigate how the ability of generating images corresponding to text prompts can be leveraged for data generation, augmentation, and transformation for low-resource settings and to enhance the robustness in V+L.

Reasoning about Actions. In our previous work (Fang *et al.*, 2020), we developed commonsense video captioning to speculate about effect of actions. However, can V+L models reason about unlikely and atypical actions (eg. people often kick footballs, but rarely kick walls)? I plan to investigate how V+L models like CLIP (Radford *et al.*, 2021) can be used for reasoning about everyday actions and commonsense aspects, for both typical and atypical situations, by using counterfactual text-based image manipulation to reflect atypical situations. This study is expected to reveal that V+L models are biased towards spurious correlations between actions and objects. My research will develop debiasing techniques and constraint-based learning for reasoning about actions and their consequences.

Human-Computer Collaborative Reasoning. In the last five years, the nature of work in V+L has evolved from research prototypes to bringing about a paradigm shift in AI. I am convinced that the use of language has immense potential in changing the way we interact with AI and for democratizing and simplifying access to graphics and robotics. While we have begun to develop visual grounding and reasoning frameworks, how can we improve these abilities and embed dialog and cooperation with humans into the reasoning process? I am excited about starting this new research program of “Human-Computer Collaborative Reasoning”, an under-explored direction, which will bring together reasoning and human-aware AI research, to improve visual reasoning capabilities of computer vision models. Allowing humans to directly interact with V+L models has immense potential, as the use of human language will lead to ease-of-use when using AI. This will engender exciting new technologies, but they will be accompanied by risks and threats ([Horvitz, 2022](#)). My previous experience in discovering and mitigating failure modes will continue to be a core element of this agenda, by studying how human collaboration and feedback can help avoid such failures. I plan to expand my work into the domains of Human-Computer Interaction to develop techniques that make the fullest use of language, while mitigating the security threats and failure modes.

Connections between Adversarial and Distributional Robustness. While standard notions of distribution shift in ML are limited to single modalities and static train–test splits, the theoretical investigation of effect of interactions between different modalities remains unexplored. I am interested in understanding fundamental connections between adversarial and distributional robustness, especially when multiple modalities and data formats are involve. This will be particularly challenging effort for models that will interact and continually learn and reason with human collaboration.

My recent empirical investigation ([Gokhale *et al.*, 2022c](#)) found that data filtering methods with good intentions of removing spurious correlations, can hurt adversarial robustness. This finding has been recently corroborated by [Moayeri *et al.* \(2022\)](#); [Teney *et al.* \(2022\)](#). I plan to pursue this direction and expect it to lead to actionable design considerations for building robust vision and language models.

In sum, this dissertation takes one step towards the pursuit of knowledge-guided, human-aware, and robust learning and reasoning about the things that cameras see.

“The end of a melody is not its goal: but nonetheless, had the melody not reached its end it would not have reached its goal either. A parable.”

- Friedrich Nietzsche

REFERENCES

- Aditya, S., Y. Yang and C. Baral, “Integrating knowledge and reasoning in image understanding”, in “Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10–16, 2019”, edited by S. Kraus, pp. 6252–6259 (ijcai.org, 2019), URL <https://doi.org/10.24963/ijcai.2019/873>.
- Agarwal, V., R. Shetty and M. Fritz, “Towards causal VQA: revealing and reducing spurious correlations by invariant and covariant semantic editing”, in “2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13–19, 2020”, pp. 9687–9695 (IEEE, 2020), URL <https://doi.org/10.1109/CVPR42600.2020.00971>.
- Agrawal, A., D. Batra, D. Parikh and A. Kembhavi, “Don’t just assume; look and answer: Overcoming priors for visual question answering”, in “2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018”, pp. 4971–4980 (IEEE Computer Society, 2018), URL http://openaccess.thecvf.com/content_cvpr_2018/html/Agrawal_Dont_Just_Assume_CVPR_2018_paper.html.
- Agrawal, A., A. Kembhavi, D. Batra and D. Parikh, “C-vqa: A compositional split of the visual question answering (vqa) v1.0 dataset”, arXiv preprint arXiv:1704.08243 (2017).
- Alzantot, M., Y. Sharma, A. Elgohary, B.-J. Ho, M. Srivastava and K.-W. Chang, “Generating natural language adversarial examples”, in “Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing”, pp. 2890–2896 (Association for Computational Linguistics, Brussels, Belgium, 2018a), URL <https://www.aclweb.org/anthology/D18-1316>.
- Alzantot, M., Y. Sharma, A. Elgohary, B.-J. Ho, M. Srivastava and K.-W. Chang, “Generating natural language adversarial examples”, in “Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing”, pp. 2890–2896 (Association for Computational Linguistics, Brussels, Belgium, 2018b), URL <https://www.aclweb.org/anthology/D18-1316>.
- Anderson, P., B. Fernando, M. Johnson and S. Gould, “Spice: Semantic propositional image caption evaluation”, in “European conference on computer vision”, pp. 382–398 (Springer, 2016).
- Anderson, P., X. He, C. Buehler, D. Teney, M. Johnson, S. Gould and L. Zhang, “Bottom-up and top-down attention for image captioning and visual question answering”, in “2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018”, pp. 6077–6086 (IEEE Computer Society, 2018a), URL http://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.html.

- Anderson, P., Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. D. Reid, S. Gould and A. van den Hengel, “Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments”, in “2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018”, pp. 3674–3683 (IEEE Computer Society, 2018b), URL http://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Vision-and-Language_Navigation_Interpreting_CVPR_2018_paper.html.
- Antol, S., A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick and D. Parikh, “VQA: visual question answering”, in “2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015”, pp. 2425–2433 (IEEE Computer Society, 2015), URL <https://doi.org/10.1109/ICCV.2015.279>.
- Arjovsky, M., L. Bottou, I. Gulrajani and D. Lopez-Paz, “Invariant risk minimization”, arXiv preprint arXiv:1907.02893 (2019).
- Arunkumar, A., S. Mishra, B. Sachdeva, C. Baral and C. Bryan, “Real-time visual feedback for educative benchmark creation: A human-and-metric-in-the-loop workflow”, (2020).
- Asai, A. and H. Hajishirzi, “Logic-guided data augmentation and regularization for consistent question answering”, in “Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics”, pp. 5642–5650 (Association for Computational Linguistics, Online, 2020), URL <https://www.aclweb.org/anthology/2020.acl-main.499>.
- Banerjee, P., *Implicitly Supervised Neural Question Answering*, Ph.D. thesis, Arizona State University (2022).
- Banerjee, P., T. Gokhale and C. Baral, “Self-supervised test-time learning for reading comprehension”, in “Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies”, pp. 1200–1211 (Association for Computational Linguistics, Online, 2021a), URL <https://www.aclweb.org/anthology/2021.naacl-main.95>.
- Banerjee, P., T. Gokhale, Y. Yang and C. Baral, “Weakly supervised relative spatial reasoning for visual question answering”, Proceedings of the IEEE/CVF International Conference on Computer Vision <https://arxiv.org/abs/2109.01934> (2021b).
- Banerjee, P., T. Gokhale, Y. Yang and C. Baral, “Weqa: Weak supervision via captions for visual question answering”, in “Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021”, pp. 3420–3435 (2021c).
- Banerjee, S. and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments”, in “Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization”, pp. 65–72 (Association for Computational Linguistics, Ann Arbor, Michigan, 2005), URL <https://www.aclweb.org/anthology/W05-0909>.

Beery, S., G. Van Horn and P. Perona, “Recognition in terra incognita”, in “Proceedings of the European conference on computer vision (ECCV)”, pp. 456–473 (2018).

Belinkov, Y. and Y. Bisk, “Synthetic and natural noise both break neural machine translation”, in “6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings”, (OpenReview.net, 2018), URL <https://openreview.net/forum?id=BJ8vJebC->.

Bender, E. M., T. Gebru, A. McMillan-Major and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?”, in “Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency”, pp. 610–623 (2021).

Benton, G. W., M. Finzi, P. Izmailov and A. G. Wilson, “Learning invariances in neural networks from training data”, in “Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual”, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin (2020), URL <https://proceedings.neurips.cc/paper/2020/hash/cc8090c4d2791cdd9cd2cb3c24296190-Abstract.html>.

Bhagavatula, C., R. L. Bras, C. Malaviya, K. Sakaguchi, A. Holtzman, H. Rashkin, D. Downey, W. Yih and Y. Choi, “Abductive commonsense reasoning”, in “8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020”, (OpenReview.net, 2020), URL <https://openreview.net/forum?id=Byg1v1HKDB>.

Bhargava, P., A. Drozd and A. Rogers, “Generalization in NLI: Ways (not) to go beyond simple heuristics”, in “Proceedings of the Second Workshop on Insights from Negative Results in NLP”, pp. 125–135 (Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021), URL <https://aclanthology.org/2021.insights-1.18>.

Bhattacharjee, S., D. Mandal and S. Biswas, “Multi-class novelty detection using mix-up technique”, in “Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)”, (2020).

Bhattacharya, N., Q. Li and D. Gurari, “Why does a visual question have different answers?”, in “2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019”, pp. 4270–4279 (IEEE, 2019), URL <https://doi.org/10.1109/ICCV.2019.00437>.

Biggio, B., I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto and F. Roli, “Evasion attacks against machine learning at test time”, in “Joint European conference on machine learning and knowledge discovery in databases”, pp. 387–402 (Springer, 2013).

Bisk, Y., R. Zellers, R. LeBras, J. Gao and Y. Choi, “PIQA: reasoning about physical commonsense in natural language”, in “The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications

of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020”, pp. 7432–7439 (AAAI Press, 2020), URL <https://aaai.org/ojs/index.php/AAAI/article/view/6239>.

Blender Online Community, A., *Blender - a 3D modelling and rendering package*, Blender Foundation, Stichting Blender Foundation, Amsterdam, URL <http://www.blender.org> (2018).

Bobrow, D. G., “Natural language input for a computer problem solving system”, (1964).

Bolukbasi, T., K. Chang, J. Y. Zou, V. Saligrama and A. T. Kalai, “Man is to computer programmer as woman is to homemaker? debiasing word embeddings”, in “Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain”, edited by D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon and R. Garnett, pp. 4349–4357 (2016), URL <https://proceedings.neurips.cc/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html>.

Boole, G., *An investigation of the laws of thought: on which are founded the mathematical theories of logic and probabilities* (Dover Publications, 1854).

Bordes, A., N. Usunier, A. García-Durán, J. Weston and O. Yakhnenko, “Translating embeddings for modeling multi-relational data”, in “Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States”, edited by C. J. C. Burges, L. Bottou, Z. Ghahramani and K. Q. Weinberger, pp. 2787–2795 (2013), URL <https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html>.

Bosselut, A., H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz and Y. Choi, “COMET: Commonsense transformers for automatic knowledge graph construction”, in “Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics”, pp. 4762–4779 (Association for Computational Linguistics, Florence, Italy, 2019), URL <https://www.aclweb.org/anthology/P19-1470>.

Bowman, S. R., G. Angeli, C. Potts and C. D. Manning, “A large annotated corpus for learning natural language inference”, in “Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing”, pp. 632–642 (Association for Computational Linguistics, Lisbon, Portugal, 2015a), URL <https://www.aclweb.org/anthology/D15-1075>.

Bowman, S. R., C. Potts and C. D. Manning, “Recursive neural networks can learn logical semantics”, in “Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality”, pp. 12–21 (Association for Computational Linguistics, Beijing, China, 2015b), URL <https://www.aclweb.org/anthology/W15-4002>.

- Bras, R. L., S. Swayamdipta, C. Bhagavatula, R. Zellers, M. E. Peters, A. Sabharwal and Y. Choi, “Adversarial filters of dataset biases”, in “Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event”, vol. 119 of *Proceedings of Machine Learning Research*, pp. 1078–1088 (PMLR, 2020), URL <http://proceedings.mlr.press/v119/bras20a.html>.
- Bulusu, S., B. Kailkhura, B. Li, P. K. Varshney and D. Song, “Anomalous example detection in deep learning: A survey”, IEEE Access **8**, 132330–132347 (2020).
- Carey, S., *Conceptual change in childhood* (MIT press, 1985).
- Carion, N., F. Massa, G. Synnaeve, N. Usunier, A. Kirillov and S. Zagoruyko, “End-to-end object detection with transformers”, in “European conference on computer vision”, pp. 213–229 (Springer, 2020).
- Carlini, N. and D. Wagner, “Towards evaluating the robustness of neural networks”, in “2017 ieee symposium on security and privacy (sp)”, pp. 39–57 (IEEE, 2017).
- Carlucci, F. M., A. D’Innocente, S. Bucci, B. Caputo and T. Tommasi, “Domain generalization by solving jigsaw puzzles”, in “IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019”, pp. 2229–2238 (Computer Vision Foundation / IEEE, 2019), URL http://openaccess.thecvf.com/content_CVPR_2019/html/Carlucci_Domain_Generalization_by_Solving_Jigsaw_Puzzles_CVPR_2019_paper.html.
- Cesana-Arlotti, N., A. Martín, E. Téglás, L. Vorobyova, R. Cetnarski and L. L. Bonatti, “Precursors of logical reasoning in preverbal human infants”, Science **359**, 6381, 1263–1266, URL <https://science.sciencemag.org/content/359/6381/1263> (2018).
- Chai, L., J.-Y. Zhu, E. Shechtman, P. Isola and R. Zhang, “Ensembling with deep generative views”, in “Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition”, pp. 14997–15007 (2021).
- Chan, Y.-H. and Y.-C. Fan, “A recurrent BERT-based model for question generation”, in “Proceedings of the 2nd Workshop on Machine Reading for Question Answering”, pp. 154–162 (Association for Computational Linguistics, Hong Kong, China, 2019), URL <https://www.aclweb.org/anthology/D19-5821>.
- Chang, C.-Y., D.-A. Huang, D. Xu, E. Adeli, L. Fei-Fei and J. C. Niebles, “Procedure planning in instructional videos”, European Conference on Computer Vision (2020).
- Chao, W., H. Hu and F. Sha, “Cross-dataset adaptation for visual question answering”, in “2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018”, pp. 5716–5725 (IEEE Computer Society, 2018), URL http://openaccess.thecvf.com/content_cvpr_2018/html/Chao_Cross-Dataset_Adaptation_for_CVPR_2018_paper.html.
- Chen, H., A. Suhr, D. Misra, N. Snavely and Y. Artzi, “TOUCHDOWN: natural language navigation and spatial reasoning in visual street environments”, in “IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach,

CA, USA, June 16-20, 2019”, pp. 12538–12547 (Computer Vision Foundation / IEEE, 2019), URL http://openaccess.thecvf.com/content_CVPR_2019/html/Chen_TOUCHDOWN_Natural_Language_Navigation_and_Spatial_Reasoning_in_Visual_Street_CVPR_2019_paper.html.

Chen, L., X. Yan, J. Xiao, H. Zhang, S. Pu and Y. Zhuang, “Counterfactual samples synthesizing for robust visual question answering”, in “2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020”, pp. 10797–10806 (IEEE, 2020a), URL <https://doi.org/10.1109/CVPR42600.2020.01081>.

Chen, Y.-C., L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng and J. Liu, “Uniter: Universal image-text representation learning”, in “Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX”, pp. 104–120 (Springer, 2020b).

Cheng, S., T. Gokhale and Y. Yang, “Adversarial bayesian augmentation for single-source domain generalization”, in “preprint”, (2023).

Cho, J., A. Zala and M. Bansal, “Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers”, arXiv preprint arXiv:2202.04053 (2022).

Clark, C., M. Yatskar and L. Zettlemoyer, “Don’t take the easy way out: Ensemble based methods for avoiding known dataset biases”, in “Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)”, pp. 4069–4082 (Association for Computational Linguistics, Hong Kong, China, 2019), URL <https://www.aclweb.org/anthology/D19-1418>.

Cohen, T. S. and M. Welling, “Transformation properties of learned visual representations”, in “3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings”, edited by Y. Bengio and Y. LeCun (2015), URL <http://arxiv.org/abs/1412.7659>.

Conn, A. R., N. I. Gould and P. L. Toint, *Trust region methods* (SIAM, 2000).

Conwell, C. and T. Ullman, “Testing relational understanding in text-guided image generation”, arXiv preprint arXiv:2208.00005 (2022).

Corcoran, J., “Completeness of an ancient logic”, The journal of symbolic logic **37**, 4, 696–702 (1972).

Cordts, M., M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth and B. Schiele, “The cityscapes dataset for semantic urban scene understanding”, in “2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016”, pp. 3213–3223 (IEEE Computer Society, 2016), URL <https://doi.org/10.1109/CVPR.2016.350>.

Csurka, G., “Domain adaptation for visual applications: A comprehensive survey”, arXiv preprint arXiv:1702.05374 (2017).

Cubuk, E. D., B. Zoph, D. Mané, V. Vasudevan and Q. V. Le, “Autoaugment: Learning augmentation strategies from data”, in “IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019”, pp. 113–123 (Computer Vision Foundation / IEEE, 2019), URL http://openaccess.thecvf.com/content_CVPR_2019/html/Cubuk_AutoAugment_Learning_Augmentation_Strategies_From_Data_CVPR_2019_paper.html.

Cubuk, E. D., B. Zoph, J. Shlens and Q. Le, “RandAugment: Practical automated data augmentation with a reduced search space”, in “Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual”, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin (2020), URL <https://proceedings.neurips.cc/paper/2020/hash/d85b63ef0ccb114d0a3bb7b7d808028f-Abstract.html>.

Dalvi, B., L. Huang, N. Tandon, W.-t. Yih and P. Clark, “Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension”, in “Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)”, pp. 1595–1604 (Association for Computational Linguistics, New Orleans, Louisiana, 2018), URL <https://www.aclweb.org/anthology/N18-1144>.

Dalvi, N., P. Domingos, S. Sanghai and D. Verma, “Adversarial classification”, in “Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining”, pp. 99–108 (2004).

Dancette, C., R. Cadene, D. Teney and M. Cord, “Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering”, in “Proceedings of the IEEE/CVF International Conference on Computer Vision”, pp. 1574–1583 (2021).

Davis, J. and M. Goadrich, “The relationship between precision-recall and ROC curves”, in “Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006”, edited by W. W. Cohen and A. W. Moore, vol. 148 of *ACM International Conference Proceeding Series*, pp. 233–240 (ACM, 2006), URL <https://doi.org/10.1145/1143844.1143874>.

Dayma, B., S. Patil, P. Cuenca, K. Saifullah, T. Abraham, P. Le Khac, L. Melas and R. Ghosh, “Dall·e mini”, URL <https://github.com/borisdayma/dalle-mini> (2021).

Demszky, D., K. Guu and P. Liang, “Transforming question answering datasets into natural language inference datasets”, arXiv preprint arXiv:1809.02922 (2018).

Deng, J., W. Dong, R. Socher, L. Li, K. Li and F. Li, “Imagenet: A large-scale hierarchical image database”, in “2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA”, pp. 248–255 (IEEE Computer Society, 2009), URL <https://doi.org/10.1109/CVPR.2009.5206848>.

- Denker, J., W. Gardner, H. Graf, D. Henderson, R. Howard, W. Hubbard, L. Jackel, H. Baird and I. Guyon, “Neural network recognizer for hand-written zip code digits”, in “Proceedings of the 1st International Conference on Neural Information Processing Systems”, pp. 323–331 (1988).
- Devlin, J., M.-W. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding”, in “Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)”, pp. 4171–4186 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019), URL <https://www.aclweb.org/anthology/N19-1423>.
- DeVries, T. and G. W. Taylor, “Improved regularization of convolutional neural networks with cutout”, arXiv preprint arXiv:1708.04552 (2017).
- Dhillon, G. S., K. Azizzadenesheli, Z. C. Lipton, J. D. Bernstein, J. Kossaifi, A. Khanna and A. Anandkumar, “Stochastic activation pruning for robust adversarial defense”, in “International Conference on Learning Representations”, (2018).
- Ding, M., W. Zheng, W. Hong and J. Tang, “Cogview2: Faster and better text-to-image generation via hierarchical transformers”, arXiv preprint arXiv:2204.14217 (2022).
- Dong, Y., F. Liao, T. Pang, H. Su, J. Zhu, X. Hu and J. Li, “Boosting adversarial attacks with momentum”, in “Proceedings of the IEEE conference on computer vision and pattern recognition”, pp. 9185–9193 (2018).
- Dvijotham, K., R. Stanforth, S. Gowal, T. A. Mann and P. Kohli, “A dual approach to scalable verification of deep networks”, in “Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018”, edited by A. Globerson and R. Silva, pp. 550–559 (AUAI Press, 2018), URL <http://auai.org/uai2018/proceedings/papers/204.pdf>.
- Ebrahimi, J., A. Rao, D. Lowd and D. Dou, “HotFlip: White-box adversarial examples for text classification”, in “Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)”, pp. 31–36 (Association for Computational Linguistics, Melbourne, Australia, 2018), URL <https://www.aclweb.org/anthology/P18-2006>.
- Eckstein, J., “Nonlinear proximal point algorithms using bregman functions, with applications to convex programming”, Mathematics of Operations Research **18**, 1, 202–226 (1993).
- Ethayarajh, K., “How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings”, in “Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)”, pp. 55–65 (Association for Computational Linguistics, Hong Kong, China, 2019), URL <https://www.aclweb.org/anthology/D19-1006>.

Ettinger, A., “What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models”, *Transactions of the Association for Computational Linguistics* **8**, 34–48, URL <https://www.aclweb.org/anthology/2020.tacl-1.3> (2020).

Everingham, M., L. Van Gool, C. K. Williams, J. Winn and A. Zisserman, “The pascal visual object classes (voc) challenge”, *International journal of computer vision* **88**, 2, 303–338 (2010).

Fang, Z., T. Gokhale, P. Banerjee, C. Baral and Y. Yang, “Video2Commonsense: Generating commonsense descriptions to enrich video captioning”, in “Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)”, pp. 840–860 (Association for Computational Linguistics, Online, 2020), URL <https://www.aclweb.org/anthology/2020.emnlp-main.61>.

Fang, Z., S. Kong, C. C. Fowlkes and Y. Yang, “Modularized textual grounding for counterfactual resilience”, in “IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019”, pp. 6378–6388 (Computer Vision Foundation / IEEE, 2019), URL http://openaccess.thecvf.com/content_CVPR_2019/html/Fang_Modularized_Textual_Grounding_for.Counterfactual_Resilience_CVPR_2019_paper.html.

Fang, Z., S. Kong, T. Yu and Y. Yang, “Weakly supervised attention learning for textual phrases grounding”, arXiv preprint arXiv:1805.00545 (2018).

Feinglass, J. and Y. Yang, “SMURF: SeMantic and linguistic UndeRstanding fusion for caption evaluation via typicality analysis”, in “Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)”, pp. 2250–2260 (Association for Computational Linguistics, Online, 2021), URL <https://aclanthology.org/2021.acl-long.175>.

Feng, W., X. He, T.-J. Fu, V. Jampani, A. Akula, P. Narayana, S. Basu, X. E. Wang and W. Y. Wang, “Training-free structured diffusion guidance for compositional text-to-image synthesis”, arXiv preprint arXiv:2212.05032 (2022).

Fidler, S., R. Mottaghi, A. L. Yuille and R. Urtasun, “Bottom-up segmentation for top-down detection”, in “2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013”, pp. 3294–3301 (IEEE Computer Society, 2013), URL <https://doi.org/10.1109/CVPR.2013.423>.

Fisch, A., A. Talmor, R. Jia, M. Seo, E. Choi and D. Chen, “MRQA 2019 shared task: Evaluating generalization in reading comprehension”, in “Proceedings of the 2nd Workshop on Machine Reading for Question Answering”, pp. 1–13 (Association for Computational Linguistics, Hong Kong, China, 2019), URL <https://www.aclweb.org/anthology/D19-5801>.

Fréchet, M., “Généralisation du théoreme des probabilités totales”, *Fundamenta mathematicae* **1**, 25, 379–387 (1935).

Fukui, A., D. H. Park, D. Yang, A. Rohrbach, T. Darrell and M. Rohrbach, “Multi-modal compact bilinear pooling for visual question answering and visual grounding”, in “Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing”, pp. 457–468 (Association for Computational Linguistics, Austin, Texas, 2016), URL <https://www.aclweb.org/anthology/D16-1044>.

Gan, Z., Y. Chen, L. Li, C. Zhu, Y. Cheng and J. Liu, “Large-scale adversarial training for vision-and-language representation learning”, in “Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual”, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin (2020), URL <https://proceedings.neurips.cc/paper/2020/hash/49562478de4c54fafd4ec46fdb297de5-Abstract.html>.

Gan, Z., C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin and L. Deng, “Semantic compositional networks for visual captioning”, in “2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017”, pp. 1141–1150 (IEEE Computer Society, 2017), URL <https://doi.org/10.1109/CVPR.2017.127>.

Ganin, Y. and V. S. Lempitsky, “Unsupervised domain adaptation by backpropagation”, in “Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015”, edited by F. R. Bach and D. M. Blei, vol. 37 of *JMLR Workshop and Conference Proceedings*, pp. 1180–1189 (JMLR.org, 2015), URL <http://proceedings.mlr.press/v37/ganin15.html>.

Ganin, Y., E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand and V. Lempitsky, “Domain-adversarial training of neural networks”, *The journal of machine learning research* **17**, 1, 2096–2030 (2016).

Gao, L., Z. Guo, H. Zhang, X. Xu and H. T. Shen, “Video captioning with attention-based lstm and semantic consistency”, *IEEE Transactions on Multimedia* **19**, 9, 2045–2055 (2017).

Gardner, M., Y. Artzi, V. Basmov, J. Berant, B. Bogin, S. Chen, P. Dasigi, D. Dua, Y. Elazar, A. Gottumukkala, N. Gupta, H. Hajishirzi, G. Ilharco, D. Khashabi, K. Lin, J. Liu, N. F. Liu, P. Mulcaire, Q. Ning, S. Singh, N. A. Smith, S. Subramanian, R. Tsarfaty, E. Wallace, A. Zhang and B. Zhou, “Evaluating models’ local decision boundaries via contrast sets”, in “Findings of the Association for Computational Linguistics: EMNLP 2020”, pp. 1307–1323 (Association for Computational Linguistics, Online, 2020a), URL <https://www.aclweb.org/anthology/2020.findings-emnlp.117>.

Gardner, M., Y. Artzi, V. Basmov, J. Berant, B. Bogin, S. Chen, P. Dasigi, D. Dua, Y. Elazar, A. Gottumukkala, N. Gupta, H. Hajishirzi, G. Ilharco, D. Khashabi, K. Lin, J. Liu, N. F. Liu, P. Mulcaire, Q. Ning, S. Singh, N. A. Smith, S. Subramanian, R. Tsarfaty, E. Wallace, A. Zhang and B. Zhou, “Evaluating models’ local decision boundaries via contrast sets”, in “Findings of the Association for

Computational Linguistics: EMNLP 2020”, pp. 1307–1323 (Association for Computational Linguistics, Online, 2020b), URL <https://www.aclweb.org/anthology/2020.findings-emnlp.117>.

Geirhos, R., P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann and W. Brendel, “Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness”, in “7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019”, (OpenReview.net, 2019), URL <https://openreview.net/forum?id=Bygh9j09KX>.

Geirhos, R., C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge and F. A. Wichmann, “Generalisation in humans and deep neural networks”, in “Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada”, edited by S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, pp. 7549–7561 (2018), URL <https://proceedings.neurips.cc/paper/2018/hash/0937fb5864ed06ffb59ae5f9b5ed67a9-Abstract.html>.

Gidaris, S., P. Singh and N. Komodakis, “Unsupervised representation learning by predicting image rotations”, in “6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings”, (OpenReview.net, 2018), URL <https://openreview.net/forum?id=S1v4N210->.

Glockner, M., V. Shwartz and Y. Goldberg, “Breaking NLI systems with sentences that require simple lexical inferences”, in “Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)”, pp. 650–655 (Association for Computational Linguistics, Melbourne, Australia, 2018a), URL <https://www.aclweb.org/anthology/P18-2103>.

Glockner, M., V. Shwartz and Y. Goldberg, “Breaking NLI systems with sentences that require simple lexical inferences”, in “Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)”, pp. 650–655 (Association for Computational Linguistics, Melbourne, Australia, 2018b), URL <https://www.aclweb.org/anthology/P18-2103>.

Goel, V., M. Chandak, A. Anand and P. Guha, “Iq-vqa: intelligent visual question answering”, in “Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II”, pp. 357–370 (Springer, 2021).

Gokhale, T., R. Anirudh, B. Kailkhura, J. J. Thiagarajan, C. Baral and Y. Yang, “Attribute-guided adversarial training for robustness to natural perturbations”, arXiv preprint arXiv:2012.01806 (2020a).

Gokhale, T., R. Anirudh, B. Kailkhura, J. J. Thiagarajan, C. Baral and Y. Yang, “Attribute-guided adversarial training for robustness to natural perturbations”, in “Proceedings of the AAAI Conference on Artificial Intelligence”, vol. 35, pp. 7574–7582 (2021).

Gokhale, T., R. Anirudh, J. J. Thiagarajan, B. Kailkhura, C. Baral and Y. Yang, “Improving diversity with adversarially learned transformations for domain generalization”, in “Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)”, pp. 434–443 (2023).

Gokhale, T., P. Banerjee, C. Baral and Y. Yang, “MUTANT: A training paradigm for out-of-distribution generalization in visual question answering”, in “Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)”, pp. 878–892 (Association for Computational Linguistics, Online, 2020b), URL <https://www.aclweb.org/anthology/2020.emnlp-main.63>.

Gokhale, T., P. Banerjee, C. Baral and Y. Yang, “Vqa-lol: Visual question answering under the lens of logic”, in “European conference on computer vision”, (Springer, 2020c).

Gokhale, T., A. Chaudhary, P. Banerjee, C. Baral and Y. Yang, “Semantically distributed robust optimization for vision-and-language inference”, in “Findings of the Association for Computational Linguistics: ACL 2022”, pp. 1493–1513 (2022a).

Gokhale, T., J. Feinglass and Y. Yang, “Covariate shift detection via domain interpolation sensitivity”, in “First Workshop on Interpolation Regularizers and Beyond at NeurIPS 2022”, (2022b), URL <https://openreview.net/forum?id=YkPjTHZDdm>.

Gokhale, T., S. Mishra, M. Luo, B. Sachdeva and C. Baral, “Generalized but not Robust? comparing the effects of data modification methods on out-of-domain generalization and adversarial robustness”, in “Findings of the Association for Computational Linguistics: ACL 2022”, pp. 2705–2718 (Association for Computational Linguistics, 2022c), URL <https://aclanthology.org/2022.findings-acl.213>.

Gokhale, T., H. Palangi, B. Nushi, V. Vineet, E. Horvitz, E. Kamar, C. Baral and Y. Yang, “Benchmarking spatial relationships in text-to-image generation”, arXiv preprint arXiv:2212.10015 (2022d).

Gokhale, T., S. Sampat, Z. Fang, Y. Yang and C. Baral, “Cooking with blocks: A recipe for visual reasoning on image-pairs”, in “Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops”, pp. 5–8 (2019).

Goodfellow, I. J., J. Shlens and C. Szegedy, “Explaining and harnessing adversarial examples”, in “3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings”, edited by Y. Bengio and Y. LeCun (2015), URL <http://arxiv.org/abs/1412.6572>.

Gopnik, A., A. N. Meltzoff and P. K. Kuhl, *The scientist in the crib: Minds, brains, and how children learn.* (William Morrow & Co, 1999).

Goyal, Y., T. Khot, D. Summers-Stay, D. Batra and D. Parikh, “Making the V in VQA matter: Elevating the role of image understanding in visual question answering”, in “2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017”, pp. 6325–6334 (IEEE Computer Society, 2017), URL <https://doi.org/10.1109/CVPR.2017.670>.

- Gulrajani, I. and D. Lopez-Paz, “In search of lost domain generalization”, in “International Conference on Learning Representations”, (2021), URL <https://openreview.net/forum?id=1QdXeXDoWtI>.
- Guo, C., G. Pleiss, Y. Sun and K. Q. Weinberger, “On calibration of modern neural networks”, in “Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017”, edited by D. Precup and Y. W. Teh, vol. 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330 (PMLR, 2017), URL <http://proceedings.mlr.press/v70/guo17a.html>.
- Gupta, A., P. Srinivasan, J. Shi and L. S. Davis, “Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos”, in “2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA”, pp. 2012–2019 (IEEE Computer Society, 2009), URL <https://doi.org/10.1109/CVPR.2009.5206492>.
- Hannun, A., C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, “Deep speech: Scaling up end-to-end speech recognition”, arXiv preprint arXiv:1412.5567 (2014).
- He, K., X. Zhang, S. Ren and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification”, in “Proceedings of the IEEE international conference on computer vision”, pp. 1026–1034 (2015).
- He, K., X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition”, in “2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016”, pp. 770–778 (IEEE Computer Society, 2016), URL <https://doi.org/10.1109/CVPR.2016.90>.
- He, Z., W. Zuo, M. Kan, S. Shan and X. Chen, “Attgan: Facial attribute editing by only changing what you want”, *IEEE Transactions on Image Processing* **28**, 11, 5464–5478 (2019).
- Hegel, G. W. F., “Hegel’s science of logic”, (1929).
- Hendricks, L. A., K. Burns, K. Saenko, T. Darrell and A. Rohrbach, “Women also snowboard: Overcoming bias in captioning models”, in “European Conference on Computer Vision”, pp. 793–811 (Springer, 2018).
- Hendrycks, D., S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo *et al.*, “The many faces of robustness: A critical analysis of out-of-distribution generalization”, in “Proceedings of the IEEE/CVF International Conference on Computer Vision”, pp. 8340–8349 (2021).
- Hendrycks, D. and T. G. Dietterich, “Benchmarking neural network robustness to common corruptions and perturbations”, in “7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019”, (OpenReview.net, 2019), URL <https://openreview.net/forum?id=HJz6tiCqYm>.

Hendrycks, D. and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks”, in “5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings”, (OpenReview.net, 2017), URL <https://openreview.net/forum?id=Hkg4TI9xl>.

Hendrycks, D., X. Liu, E. Wallace, A. Dziedzic, R. Krishnan and D. Song, “Pretrained transformers improve out-of-distribution robustness”, in “Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics”, pp. 2744–2751 (Association for Computational Linguistics, Online, 2020a), URL <https://www.aclweb.org/anthology/2020.acl-main.244>.

Hendrycks, D., X. Liu, E. Wallace, A. Dziedzic, R. Krishnan and D. Song, “Pretrained transformers improve out-of-distribution robustness”, in “Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics”, pp. 2744–2751 (Association for Computational Linguistics, Online, 2020b), URL <https://www.aclweb.org/anthology/2020.acl-main.244>.

Hendrycks, D., N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer and B. Lakshminarayanan, “Augmix: A simple data processing method to improve robustness and uncertainty”, in “8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020”, (OpenReview.net, 2020c), URL <https://openreview.net/forum?id=S1gmrxFvB>.

Hessel, J., A. Holtzman, M. Forbes, R. Le Bras and Y. Choi, “Clipscore: A reference-free evaluation metric for image captioning”, in “Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing”, pp. 7514–7528 (2021).

Heusel, M., H. Ramsauer, T. Unterthiner, B. Nessler and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium”, in “Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA”, edited by I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan and R. Garnett, pp. 6626–6637 (2017), URL <https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html>.

Hinz, T., S. Heinrich and S. Wermter, “Semantic object accuracy for generative text-to-image synthesis”, IEEE transactions on pattern analysis and machine intelligence (2020).

Hochreiter, S. and J. Schmidhuber, “Long short-term memory”, Neural computation **9**, 8, 1735–1780 (1997).

Hoeffding, W., “Probability inequalities for sums of bounded random variables”, in “The collected works of Wassily Hoeffding”, pp. 409–426 (Springer, 1994).

Hong, S., X. Yan, T. E. Huang and H. Lee, “Learning hierarchical semantic image manipulation through structured representations”, in “Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information

Processing Systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada”, edited by S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, pp. 2713–2723 (2018), URL <https://proceedings.neurips.cc/paper/2018/hash/602d1305678a8d5fdb372271e980da6a-Abstract.html>.

Honnibal, M. and M. Johnson, “An improved non-monotonic transition system for dependency parsing”, in “Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing”, pp. 1373–1378 (Association for Computational Linguistics, Lisbon, Portugal, 2015), URL <https://www.aclweb.org/anthology/D15-1162>.

Honnibal, M. and I. Montani, “spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing”, To appear **7**, 1 (2017).

Horn, L. R. and Y. Kato, *Negation and polarity: Syntactic and semantic perspectives* (OUP Oxford, 2000).

Hornik, K., M. Stinchcombe and H. White, “Multilayer feedforward networks are universal approximators”, Neural networks **2**, 5, 359–366 (1989).

Horvitz, E., “On the horizon: Interactive and compositional deepfakes”, in “Proceedings of the 2022 International Conference on Multimodal Interaction”, pp. 653–661 (2022).

Hu, W., G. Niu, I. Sato and M. Sugiyama, “Does distributionally robust supervised learning give robust classifiers?”, in “Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018”, edited by J. G. Dy and A. Krause, vol. 80 of *Proceedings of Machine Learning Research*, pp. 2034–2042 (PMLR, 2018a), URL <http://proceedings.mlr.press/v80/hu18a.html>.

Hu, W., G. Niu, I. Sato and M. Sugiyama, “Does distributionally robust supervised learning give robust classifiers?”, in “Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018”, edited by J. G. Dy and A. Krause, vol. 80 of *Proceedings of Machine Learning Research*, pp. 2034–2042 (PMLR, 2018b), URL <http://proceedings.mlr.press/v80/hu18a.html>.

Huang, R., A. Geng and Y. Li, “On the importance of gradients for detecting distributional shifts in the wild”, in “Advances in Neural Information Processing Systems”, (2021).

Hudson, D. A. and C. D. Manning, “Compositional attention networks for machine reasoning”, in “6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings”, (OpenReview.net, 2018), URL <https://openreview.net/forum?id=S1Euwz-Rb>.

Hudson, D. A. and C. D. Manning, “Gqa: A new dataset for real-world visual reasoning and compositional question answering”, (2019).

Isola, P., J. Zhu, T. Zhou and A. A. Efros, “Image-to-image translation with conditional adversarial networks”, in “2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017”, pp. 5967–5976 (IEEE Computer Society, 2017), URL <https://doi.org/10.1109/CVPR.2017.632>.

Iyyer, M., J. Wieting, K. Gimpel and L. Zettlemoyer, “Adversarial example generation with syntactically controlled paraphrase networks”, in “Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)”, pp. 1875–1885 (Association for Computational Linguistics, New Orleans, Louisiana, 2018a), URL <https://www.aclweb.org/anthology/N18-1170>.

Iyyer, M., J. Wieting, K. Gimpel and L. Zettlemoyer, “Adversarial example generation with syntactically controlled paraphrase networks”, in “Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)”, pp. 1875–1885 (Association for Computational Linguistics, New Orleans, Louisiana, 2018b), URL <https://www.aclweb.org/anthology/N18-1170>.

Jaderberg, M., K. Simonyan, A. Zisserman and K. Kavukcuoglu, “Spatial transformer networks”, in “Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada”, edited by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett, pp. 2017–2025 (2015), URL <https://proceedings.neurips.cc/paper/2015/hash/33ceb07bf4eeb3da587e268d663aba1a-Abstract.html>.

Jang, Y., T. Zhao, S. Hong and H. Lee, “Adversarial defense via learning to generate diverse attacks”, in “Proceedings of the IEEE/CVF International Conference on Computer Vision”, pp. 2740–2749 (2019).

Jascob, B., “Lemminflect. a python module for english word lemmatization and inflection.”, <https://github.com/bjascob/LemmInflect> (v0.2.1 (February 22, 2020)).

Jia, R. and P. Liang, “Adversarial examples for evaluating reading comprehension systems”, in “Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing”, pp. 2021–2031 (Association for Computational Linguistics, Copenhagen, Denmark, 2017a), URL <https://www.aclweb.org/anthology/D17-1215>.

Jia, R. and P. Liang, “Adversarial examples for evaluating reading comprehension systems”, in “Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing”, pp. 2021–2031 (Association for Computational Linguistics, Copenhagen, Denmark, 2017b), URL <https://www.aclweb.org/anthology/D17-1215>.

Jia, R., A. Raghunathan, K. Göksel and P. Liang, “Certified robustness to adversarial word substitutions”, in “Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on

Natural Language Processing (EMNLP-IJCNLP)”, pp. 4129–4142 (Association for Computational Linguistics, Hong Kong, China, 2019a), URL <https://www.aclweb.org/anthology/D19-1423>.

Jia, R., A. Raghunathan, K. Göksel and P. Liang, “Certified robustness to adversarial word substitutions”, in “Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)”, pp. 4129–4142 (Association for Computational Linguistics, Hong Kong, China, 2019b), URL <https://www.aclweb.org/anthology/D19-1423>.

Jiang, H., P. He, W. Chen, X. Liu, J. Gao and T. Zhao, “SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization”, in “Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics”, pp. 2177–2190 (Association for Computational Linguistics, Online, 2020), URL <https://www.aclweb.org/anthology/2020.acl-main.197>.

Johnson, J., B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick and R. B. Girshick, “CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning”, in “2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017”, pp. 1988–1997 (IEEE Computer Society, 2017), URL <https://doi.org/10.1109/CVPR.2017.215>.

Jones, E., R. Jia, A. Raghunathan and P. Liang, “Robust encodings: A framework for combating adversarial typos”, in “Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics”, pp. 2752–2765 (Association for Computational Linguistics, Online, 2020), URL <https://www.aclweb.org/anthology/2020.acl-main.245>.

Joshi, A., A. Mukherjee, S. Sarkar and C. Hegde, “Semantic adversarial attacks: Parametric transformations that fool deep classifiers”, in “2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019”, pp. 4772–4782 (IEEE, 2019), URL <https://doi.org/10.1109/ICCV.2019.00487>.

Kafle, K. and C. Kanan, “An analysis of visual question answering algorithms”, in “IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017”, pp. 1983–1991 (IEEE Computer Society, 2017), URL <https://doi.org/10.1109/ICCV.2017.217>.

Kannan, H., A. Kurakin and I. Goodfellow, “Adversarial logit pairing”, arXiv preprint arXiv:1803.06373 (2018).

Karpathy, A. and F. Li, “Deep visual-semantic alignments for generating image descriptions”, in “IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015”, pp. 3128–3137 (IEEE Computer Society, 2015), URL <https://doi.org/10.1109/CVPR.2015.7298932>.

Karras, T., S. Laine and T. Aila, “A style-based generator architecture for generative adversarial networks”, in “IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019”, pp. 4401–4410 (Computer Vision Foundation / IEEE, 2019), URL http://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html.

Kassner, N. and H. Schütze, “Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly”, in “Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics”, pp. 7811–7818 (Association for Computational Linguistics, Online, 2020a), URL <https://www.aclweb.org/anthology/2020.acl-main.698>.

Kassner, N. and H. Schütze, “Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly”, in “Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics”, pp. 7811–7818 (2020b).

Kervadec, C., G. Antipov, M. Baccouche and C. Wolf, “Roses are red, violets are blue... but should vqa expect them to?”, in “Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition”, pp. 2776–2785 (2021).

Khot, T., A. Sabharwal and P. Clark, “Scitail: A textual entailment dataset from science question answering”, in “Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018”, edited by S. A. McIlraith and K. Q. Weinberger, pp. 5189–5197 (AAAI Press, 2018), URL <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17368>.

Kiela, D., M. Bartolo, Y. Nie, D. Kaushik, A. Geiger, Z. Wu, B. Vidgen, G. Prasad, A. Singh, P. Ringshia, Z. Ma, T. Thrush, S. Riedel, Z. Waseem, P. Stenetorp, R. Jia, M. Bansal, C. Potts and A. Williams, “Dynabench: Rethinking benchmarking in NLP”, in “Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies”, pp. 4110–4124 (Association for Computational Linguistics, Online, 2021), URL <https://aclanthology.org/2021.naacl-main.324>.

Kim, J., A. Rohrbach, T. Darrell, J. Canny and Z. Akata, “Textual explanations for self-driving vehicles”, in “Proceedings of the European conference on computer vision (ECCV)”, pp. 563–578 (2018).

Kingma, D. P. and J. Ba, “Adam: A method for stochastic optimization”, in “3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings”, edited by Y. Bengio and Y. LeCun (2015), URL <http://arxiv.org/abs/1412.6980>.

Koh, P. W., S. Sagawa, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee *et al.*, “Wilds: A benchmark of in-the-wild distribution shifts”, in “International Conference on Machine Learning”, pp. 5637–5664 (PMLR, 2021).

Krause, J., J. Johnson, R. Krishna and L. Fei-Fei, “A hierarchical approach for generating descriptive image paragraphs”, in “2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017”, pp. 3337–3345 (IEEE Computer Society, 2017), URL <https://doi.org/10.1109/CVPR.2017.356>.

Krishna, R., K. Hata, F. Ren, L. Fei-Fei and J. C. Niebles, “Dense-captioning events in videos”, in “IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017”, pp. 706–715 (IEEE Computer Society, 2017), URL <https://doi.org/10.1109/ICCV.2017.83>.

Krizhevsky, A., “Learning multiple layers of features from tiny images”, Master’s thesis, University of Toronto (2009).

Krizhevsky, A., I. Sutskever and G. E. Hinton, “Imagenet classification with deep convolutional neural networks”, in “Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States”, edited by P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, pp. 1106–1114 (2012), URL <https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html>.

Krueger, D., E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol and A. Courville, “Out-of-distribution generalization via risk extrapolation (rex)”, in “International Conference on Machine Learning”, pp. 5815–5826 (PMLR, 2021).

Kulkarni, K., T. Gokhale, R. Singh, P. Turaga and A. Sankaranarayanan, “Halluci-net: Scene completion by exploiting object co-occurrence relationships”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops <https://arxiv.org/abs/2004.08614> (2021).

Kullback, S., R. Leibler *et al.*, “On information and sufficiency”, Annals of Mathematical Statistics **22**, 1, 79–86 (1951).

Kwiatkowski, T., J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le and S. Petrov, “Natural questions: A benchmark for question answering research”, Transactions of the Association for Computational Linguistics **7**, 452–466, URL <https://www.aclweb.org/anthology/Q19-1026> (2019).

LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition”, Neural computation **1**, 4, 541–551 (1989).

LeCun, Y., L. Bottou, Y. Bengio and P. Haffner, “Gradient-based learning applied to document recognition”, Proceedings of the IEEE **86**, 11, 2278–2324 (1998).

- Lei, J., L. Yu, M. Bansal and T. Berg, “TVQA: Localized, compositional video question answering”, in “Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing”, pp. 1369–1379 (Association for Computational Linguistics, Brussels, Belgium, 2018), URL <https://www.aclweb.org/anthology/D18-1167>.
- Leivada, E., E. Murphy and G. Marcus, “Dall-e 2 fails to reliably capture common syntactic processes”, arXiv preprint arXiv:2210.12889 (2022).
- Lewis, M. and M. Steedman, “Combined distributional and logical semantics”, Transactions of the Association for Computational Linguistics 1, 179–192, URL <https://www.aclweb.org/anthology/Q13-1015> (2013).
- Li, D., Y. Yang, Y. Song and T. M. Hospedales, “Deeper, broader and artier domain generalization”, in “IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017”, pp. 5543–5551 (IEEE Computer Society, 2017), URL <https://doi.org/10.1109/ICCV.2017.591>.
- Li, D., Y. Yang, Y. Song and T. M. Hospedales, “Learning to generalize: Meta-learning for domain generalization”, in “Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018”, edited by S. A. McIlraith and K. Q. Weinberger, pp. 3490–3497 (AAAI Press, 2018a), URL <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16067>.
- Li, D., Y. Zhang, H. Peng, L. Chen, C. Brockett, M.-T. Sun and W. B. Dolan, “Contextualized perturbation for textual adversarial attack”, in “Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies”, pp. 5053–5069 (2021a).
- Li, L., Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu and J. Liu, “HERO: Hierarchical encoder for Video+Language omni-representation pre-training”, in “Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)”, pp. 2046–2065 (Association for Computational Linguistics, Online, 2020a), URL <https://www.aclweb.org/anthology/2020.emnlp-main.161>.
- Li, L., Z. Gan and J. Liu, “A closer look at the robustness of vision-and-language pre-trained models”, arXiv preprint arXiv:2012.08673 (2020b).
- Li, L., J. Lei, Z. Gan and J. Liu, “Adversarial vqa: A new benchmark for evaluating the robustness of vqa models”, in “International Conference on Computer Vision (ICCV)”, (2021b).
- Li, X., X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, “Oscar: Object-semantics aligned pre-training for vision-language tasks”, in “European Conference on Computer Vision”, pp. 121–137 (Springer, 2020c).
- Li, Y., Y. Li and N. Vasconcelos, “Resound: Towards action recognition without representation bias”, in “Proceedings of the European Conference on Computer Vision (ECCV)”, pp. 513–528 (2018b).

- Li, Y. and N. Vasconcelos, “REPAIR: removing representation bias by dataset resampling”, in “IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019”, pp. 9572–9581 (Computer Vision Foundation / IEEE, 2019), URL http://openaccess.thecvf.com/content_CVPR_2019/html/Li_REPAIR_Removing_Representation_Bias_by_Dataset_Resampling_CVPR_2019_paper.html.
- Liang, S., Y. Li and R. Srikant, “Enhancing the reliability of out-of-distribution image detection in neural networks”, in “6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings”, (OpenReview.net, 2018), URL <https://openreview.net/forum?id=H1VGkIxRZ>.
- Lin, C.-Y., “ROUGE: A package for automatic evaluation of summaries”, in “Text Summarization Branches Out”, pp. 74–81 (Association for Computational Linguistics, Barcelona, Spain, 2004), URL <https://www.aclweb.org/anthology/W04-1013>.
- Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. L. Zitnick, “Microsoft coco: Common objects in context”, in “European conference on computer vision”, pp. 740–755 (Springer, 2014).
- Liu, E. Z., B. Haghgoo, A. S. Chen, A. Raghunathan, P. W. Koh, S. Sagawa, P. Liang and C. Finn, “Just train twice: Improving group robustness without training group information”, in “International Conference on Machine Learning”, pp. 6781–6792 (PMLR, 2021).
- Liu, F., G. Emerson and N. Collier, “Visual spatial reasoning”, arXiv preprint arXiv:2205.00363 (2022a).
- Liu, H. D., M. Tao, C. Li, D. Nowrouzezahrai and A. Jacobson, “Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer”, in “7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019”, (OpenReview.net, 2019a), URL <https://openreview.net/forum?id=SJ12niR9KQ>.
- Liu, J., W. Chen, Y. Cheng, Z. Gan, L. Yu, Y. Yang and J. Liu, “Violin: A large-scale dataset for video-and-language inference”, in “2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020”, pp. 10897–10907 (IEEE, 2020a), URL <https://doi.org/10.1109/CVPR42600.2020.01091>.
- Liu, N., S. Li, Y. Du, A. Torralba and J. B. Tenenbaum, “Compositional visual generation with composable diffusion models”, European Conference on Computer Vision (2022b).
- Liu, R., C. Liu, Y. Bai and A. L. Yuille, “Clevr-ref+: Diagnosing visual reasoning with referring expressions”, in “IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019”, pp. 4185–4194 (Computer Vision Foundation / IEEE, 2019b), URL <http://openaccess>.

[thecvf.com/content_CVPR_2019/html/Liu_CLEVR-Ref_Diagnosing_Visual_Reasoning_With_Referring_Expressions_CVPR_2019_paper.html](https://thevcf.com/content_CVPR_2019/html/Liu_CLEVR-Ref_Diagnosing_Visual_Reasoning_With_Referring_Expressions_CVPR_2019_paper.html).

Liu, W., X. Wang, J. Owens and Y. Li, “Energy-based out-of-distribution detection”, Advances in Neural Information Processing Systems (NeurIPS) (2020b).

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach”, arXiv preprint arXiv:1907.11692 (2019c).

Longpre, S., Y. Lu, Z. Tu and C. DuBois, “An exploration of data augmentation and sampling techniques for domain-agnostic question answering”, in “Proceedings of the 2nd Workshop on Machine Reading for Question Answering”, pp. 220–227 (Association for Computational Linguistics, Hong Kong, China, 2019), URL <https://www.aclweb.org/anthology/D19-5829>.

Lu, J., D. Batra, D. Parikh and S. Lee, “Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks”, in “Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada”, edited by H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. B. Fox and R. Garnett, pp. 13–23 (2019), URL <https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html>.

Luo, M., A. Mitra, T. Gokhale and C. Baral, “Improving biomedical information retrieval with neural retrievers”, in “Proceedings of the AAAI Conference on Artificial Intelligence”, vol. 36, pp. 11038–11046 (2022a).

Luo, Y., P. Banerjee, T. Gokhale, Y. Yang and C. Baral, “To find waldo you need contextual cues: Debiasing who’s waldo”, in “Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)”, pp. 355–361 (2022b).

Madry, A., A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, “Towards deep learning models resistant to adversarial attacks”, in “6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings”, (OpenReview.net, 2018a), URL <https://openreview.net/forum?id=rJzIBfZAb>.

Madry, A., A. Makelov, L. Schmidt, D. Tsipras and A. Vladu, “Towards deep learning models resistant to adversarial attacks”, in “6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings”, (OpenReview.net, 2018b), URL <https://openreview.net/forum?id=rJzIBfZAb>.

Malinowski, M. and M. Fritz, “A multi-world approach to question answering about real-world scenes based on uncertain input”, in “Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada”, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger, pp.

1682–1690 (2014), URL <https://proceedings.neurips.cc/paper/2014/hash/d516b13671a4179d9b7b458a6ebdeb92-Abstract.html>.

Mao, J., C. Gan, P. Kohli, J. B. Tenenbaum and J. Wu, “The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision”, in “7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019”, (OpenReview.net, 2019), URL <https://openreview.net/forum?id=rJgM1hRctm>.

Marcus, G., E. Davis and S. Aaronson, “A very preliminary analysis of dall-e 2”, arXiv preprint arXiv:2204.13807 (2022).

McCoy, T., E. Pavlick and T. Linzen, “Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference”, in “Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics”, pp. 3428–3448 (Association for Computational Linguistics, Florence, Italy, 2019a), URL <https://www.aclweb.org/anthology/P19-1334>.

McCoy, T., E. Pavlick and T. Linzen, “Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference”, in “Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics”, pp. 3428–3448 (Association for Computational Linguistics, Florence, Italy, 2019b), URL <https://www.aclweb.org/anthology/P19-1334>.

Mees, O., A. Emek, J. Vertens and W. Burgard, “Learning object placements for relational instructions by hallucinating scene representations”, in “2020 IEEE International Conference on Robotics and Automation (ICRA)”, pp. 94–100 (IEEE, 2020).

Miller, J., K. Krauth, B. Recht and L. Schmidt, “The effect of natural distribution shift on question answering models”, in “Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event”, vol. 119 of *Proceedings of Machine Learning Research*, pp. 6905–6916 (PMLR, 2020), URL <http://proceedings.mlr.press/v119/miller20a.html>.

Miller, J. P., R. Taori, A. Raghunathan, S. Sagawa, P. W. Koh, V. Shankar, P. Liang, Y. Carmon and L. Schmidt, “Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization”, in “International Conference on Machine Learning”, pp. 7721–7735 (PMLR, 2021).

Minderer, M., A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf and N. Houlsby, “Simple open-vocabulary object detection with vision transformers”, arXiv preprint arXiv:2205.06230 (2022).

Mintz, M., S. Bills, R. Snow and D. Jurafsky, “Distant supervision for relation extraction without labeled data”, in “Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP”, pp. 1003–1011 (Association for

Computational Linguistics, Suntec, Singapore, 2009), URL <https://www.aclweb.org/anthology/P09-1113>.

Mirza, M. and S. Osindero, “Conditional generative adversarial nets”, arXiv preprint arXiv:1411.1784 (2014).

Mishra, S. and A. Arunkumar, “How robust are model rankings: A leaderboard customization approach for equitable evaluation”, in “Proceedings of the AAAI Conference on Artificial Intelligence”, vol. 35, pp. 13561–13569 (2021).

Mishra, S., A. Arunkumar, C. Bryan and C. Baral, “Our evaluation metric needs an update to encourage generalization”, arXiv preprint arXiv:2007.06898 (2020a).

Mishra, S., A. Arunkumar, B. Sachdeva, C. Bryan and C. Baral, “Dqi: A guide to benchmark evaluation”, arXiv preprint arXiv:2008.03964 (2020b).

Mishra, S. and B. S. Sachdeva, “Do we need to create big datasets to learn a task?”, in “Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing”, pp. 169–173 (Association for Computational Linguistics, Online, 2020), URL <https://www.aclweb.org/anthology/2020.sustainlp-1.23>.

Miyato, T., A. M. Dai and I. J. Goodfellow, “Adversarial training methods for semi-supervised text classification”, in “5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings”, (OpenReview.net, 2017), URL https://openreview.net/forum?id=r1X3g2_xl.

Moayeri, M., K. Banihashem and S. Feizi, “Explicit tradeoffs between adversarial and natural distributional robustness”, in “Advances in Neural Information Processing Systems”, (2022).

Moosavi-Dezfooli, S.-M., A. Fawzi and P. Frossard, “Deepfool: a simple and accurate method to fool deep neural networks”, in “Proceedings of the IEEE conference on computer vision and pattern recognition”, pp. 2574–2582 (2016).

Morante, R. and C. Sporleder, “Modality and negation: An introduction to the special issue”, Computational Linguistics **38**, 2, 223–260, URL <https://www.aclweb.org/anthology/J12-2001> (2012).

Morris, J., E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin and Y. Qi, “TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP”, in “Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations”, pp. 119–126 (Association for Computational Linguistics, Online, 2020), URL <https://www.aclweb.org/anthology/2020.emnlp-demos.16>.

Mu, N. and J. Gilmer, “Mnist-c: A robustness benchmark for computer vision”, arXiv preprint arXiv:1906.02337 (2019).

- Naeini, M. P., G. F. Cooper and M. Hauskrecht, “Obtaining well calibrated probabilities using bayesian binning”, in “Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA”, edited by B. Bonet and S. Koenig, pp. 2901–2907 (AAAI Press, 2015), URL <http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9667>.
- Naik, A., A. Ravichander, N. Sadeh, C. Rose and G. Neubig, “Stress test evaluation for natural language inference”, in “Proceedings of the 27th International Conference on Computational Linguistics”, pp. 2340–2353 (Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018a), URL <https://www.aclweb.org/anthology/C18-1198>.
- Naik, A., A. Ravichander, N. Sadeh, C. Rose and G. Neubig, “Stress test evaluation for natural language inference”, in “Proceedings of the 27th International Conference on Computational Linguistics”, pp. 2340–2353 (Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018b), URL <https://www.aclweb.org/anthology/C18-1198>.
- Nair, S., E. Mitchell, K. Chen, S. Savarese, C. Finn *et al.*, “Learning language-conditioned robot behavior from offline data and crowd-sourced annotation”, in “Conference on Robot Learning”, pp. 1303–1315 (PMLR, 2022).
- Nam, H., H. Lee, J. Park, W. Yoon and D. Yoo, “Reducing domain gap by reducing style bias”, in “Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition”, pp. 8690–8699 (2021).
- Neelakantan, A., B. Roth and A. McCallum, “Compositional vector space models for knowledge base completion”, in “Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)”, pp. 156–166 (Association for Computational Linguistics, Beijing, China, 2015), URL <https://www.aclweb.org/anthology/P15-1016>.
- Netzer, Y., T. Wang, A. Coates, A. Bissacco, B. Wu and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning”, (2011).
- Nichol, A. Q., P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever and M. Chen, “Glide: Towards photorealistic image generation and editing with text-guided diffusion models”, in “International Conference on Machine Learning”, pp. 16784–16804 (PMLR, 2022).
- Niculescu-Mizil, A. and R. Caruana, “Predicting good probabilities with supervised learning”, in “Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005”, edited by L. D. Raedt and S. Wrobel, vol. 119 of *ACM International Conference Proceeding Series*, pp. 625–632 (ACM, 2005), URL <https://doi.org/10.1145/1102351.1102430>.
- Nie, Y., A. Williams, E. Dinan, M. Bansal, J. Weston and D. Kiela, “Adversarial NLI: A new benchmark for natural language understanding”, in “Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics”,

pp. 4885–4901 (Association for Computational Linguistics, Online, 2020), URL <https://www.aclweb.org/anthology/2020.acl-main.441>.

Nilsback, M.-E. and A. Zisserman, “Automated flower classification over a large number of classes”, in “2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing”, pp. 722–729 (IEEE, 2008).

Oren, Y., S. Sagawa, T. Hashimoto and P. Liang, “Distributionally robust language modeling”, in “Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)”, pp. 4227–4237 (Association for Computational Linguistics, Hong Kong, China, 2019), URL <https://www.aclweb.org/anthology/D19-1432>.

Ott, M., S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier and M. Auli, “fairseq: A fast, extensible toolkit for sequence modeling”, in “Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)”, pp. 48–53 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019), URL <https://www.aclweb.org/anthology/N19-4009>.

Papert, S. A., “The summer vision project”, (1966).

Papineni, K., S. Roukos, T. Ward and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation”, in “Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics”, pp. 311–318 (Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002), URL <https://www.aclweb.org/anthology/P02-1040>.

Park, J. S., C. Bhagavatula, R. Mottaghi, A. Farhadi and Y. Choi, “Visualcomet: Reasoning about the dynamic context of a still image”, in “In Proceedings of the European Conference on Computer Vision (ECCV)”, (2020).

Patel, M., T. Gokhale, C. Baral and Y. Yang, “CRIPP-VQA: Counterfactual reasoning about implicit physical properties via video question answering”, in “Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing”, pp. 9856–9870 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022), URL <https://aclanthology.org/2022.emnlp-main.670>.

Patel, M. J., “Implicit hypothetical reasoning about intrinsic physical properties”, Arizona State University, Masters Thesis (2022).

Pennington, J., R. Socher and C. Manning, “GloVe: Global vectors for word representation”, in “Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)”, pp. 1532–1543 (Association for Computational Linguistics, Doha, Qatar, 2014), URL <https://www.aclweb.org/anthology/D14-1162>.

Piattelli-Palmarini, M., “Language and learning: the debate between jean piaget and noam chomsky”, (1980).

Pirsiavash, H., C. Vondrick and A. Torralba, “Inferring the why in images”, Massachusetts Inst of Tech Cambridge (2014).

Pruthi, D., B. Dhingra and Z. C. Lipton, “Combating adversarial misspellings with robust word recognition”, in “Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics”, pp. 5582–5591 (Association for Computational Linguistics, Florence, Italy, 2019), URL <https://www.aclweb.org/anthology/P19-1561>.

Qiao, F., L. Zhao and X. Peng, “Learning to learn single domain generalization”, in “2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020”, pp. 12553–12562 (IEEE, 2020), URL <https://doi.org/10.1109/CVPR42600.2020.01257>.

Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision”, in “International Conference on Machine Learning”, pp. 8748–8763 (PMLR, 2021).

Radford, A., K. Narasimhan, T. Salimans and I. Sutskever, “Improving language understanding by generative pre-training”, (2018).

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei and I. Sutskever, “Language models are unsupervised multitask learners”, OpenAI Blog **1**, 8 (2019).

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer”, Journal of Machine Learning Research **21**, 1–67 (2020).

Raghunathan, A., J. Steinhardt and P. Liang, “Certified defenses against adversarial examples”, in “6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings”, (OpenReview.net, 2018), URL <https://openreview.net/forum?id=Bys4ob-Rb>.

Raghunathan, A., S. M. Xie, F. Yang, J. C. Duchi and P. Liang, “Understanding and mitigating the tradeoff between robustness and accuracy”, in “Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event”, vol. 119 of *Proceedings of Machine Learning Research*, pp. 7909–7919 (PMLR, 2020), URL <http://proceedings.mlr.press/v119/raghunathan20a.html>.

Raju, P., “The principle of four-cornered negation in indian philosophy”, The Review of Metaphysics pp. 694–713 (1954).

Ramakrishnan, S. K., A. Pal, G. Sharma and A. Mittal, “An empirical evaluation of visual question answering for novel objects”, in “2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017”, pp. 7312–7321 (IEEE Computer Society, 2017), URL <https://doi.org/10.1109/CVPR.2017.773>.

Ramesh, A., P. Dhariwal, A. Nichol, C. Chu and M. Chen, “Hierarchical text-conditional image generation with clip latents”, arXiv preprint arXiv:2204.06125 (2022).

Ramshaw, L. and M. Marcus, “Text chunking using transformation-based learning”, in “Third Workshop on Very Large Corpora”, (1995), URL <https://www.aclweb.org/anthology/W95-0107>.

Ratner, A. J., H. R. Ehrenberg, Z. Hussain, J. Dunnmon and C. Ré, “Learning to compose domain-specific transformations for data augmentation”, in “Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA”, edited by I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan and R. Garnett, pp. 3236–3246 (2017), URL <https://proceedings.neurips.cc/paper/2017/hash/f26dab9bf6a137c3b6782e562794c2f2-Abstract.html>.

Rauber, J., W. Brendel and M. Bethge, “Foolbox: A python toolbox to benchmark the robustness of machine learning models”, in “Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning”, (2017), URL <http://arxiv.org/abs/1707.04131>.

Ray, A., K. Sikka, A. Divakaran, S. Lee and G. Burachas, “Sunny and dark outside?! improving answer consistency in VQA through entailed question generation”, in “Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)”, pp. 5860–5865 (Association for Computational Linguistics, Hong Kong, China, 2019), URL <https://www.aclweb.org/anthology/D19-1596>.

Recht, B., R. Roelofs, L. Schmidt and V. Shankar, “Do cifar-10 classifiers generalize to cifar-10?”, arXiv preprint arXiv:1806.00451 (2018).

Reed, S. E., Z. Akata, X. Yan, L. Logeswaran, B. Schiele and H. Lee, “Generative adversarial text to image synthesis”, in “Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016”, edited by M. Balcan and K. Q. Weinberger, vol. 48 of *JMLR Workshop and Conference Proceedings*, pp. 1060–1069 (JMLR.org, 2016), URL <http://proceedings.mlr.press/v48/reed16.html>.

Ren, M., R. Kiros and R. S. Zemel, “Exploring models and data for image question answering”, in “Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada”, edited by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett, pp. 2953–2961 (2015a), URL <https://proceedings.neurips.cc/paper/2015/hash/831c2f88a604a07ca94314b56a4921b8-Abstract.html>.

Ren, S., Y. Deng, K. He and W. Che, “Generating natural language adversarial examples through probability weighted word saliency”, in “Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics”, pp. 1085–1097

(Association for Computational Linguistics, Florence, Italy, 2019), URL <https://www.aclweb.org/anthology/P19-1103>.

Ren, S., K. He, R. B. Girshick and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks”, in “Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada”, edited by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett, pp. 91–99 (2015b), URL <https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html>.

Ribeiro, M. T., C. Guestrin and S. Singh, “Are red roses red? evaluating consistency of question-answering models”, in “Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics”, pp. 6174–6184 (Association for Computational Linguistics, Florence, Italy, 2019a), URL <https://www.aclweb.org/anthology/P19-1621>.

Ribeiro, M. T., C. Guestrin and S. Singh, “Are red roses red? evaluating consistency of question-answering models”, in “Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics”, pp. 6174–6184 (Association for Computational Linguistics, Florence, Italy, 2019b), URL <https://www.aclweb.org/anthology/P19-1621>.

Ribeiro, M. T., S. Singh and C. Guestrin, “Semantically equivalent adversarial rules for debugging NLP models”, in “Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)”, pp. 856–865 (Association for Computational Linguistics, Melbourne, Australia, 2018a), URL <https://www.aclweb.org/anthology/P18-1079>.

Ribeiro, M. T., S. Singh and C. Guestrin, “Semantically equivalent adversarial rules for debugging NLP models”, in “Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)”, pp. 856–865 (Association for Computational Linguistics, Melbourne, Australia, 2018b), URL <https://www.aclweb.org/anthology/P18-1079>.

Ribeiro, M. T., T. Wu, C. Guestrin and S. Singh, “Beyond accuracy: Behavioral testing of NLP models with CheckList”, in “Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics”, pp. 4902–4912 (Association for Computational Linguistics, Online, 2020), URL <https://www.aclweb.org/anthology/2020.acl-main.442>.

Riedel, S., L. Yao, A. McCallum and B. M. Marlin, “Relation extraction with matrix factorization and universal schemas”, in “Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies”, pp. 74–84 (Association for Computational Linguistics, Atlanta, Georgia, 2013), URL <https://www.aclweb.org/anthology/N13-1008>.

Robey, A., G. J. Pappas and H. Hassani, “Model-based domain generalization”, Advances in Neural Information Processing Systems **34**, 20210–20229 (2021).

- Rocktäschel, T., M. Bosnjak, S. Singh and S. Riedel, “Low-dimensional embeddings of logic”, in “Proceedings of the ACL 2014 Workshop on Semantic Parsing”, pp. 45–49 (Association for Computational Linguistics, Baltimore, MD, 2014), URL <https://www.aclweb.org/anthology/W14-2409>.
- Rohrbach, A., M. Rohrbach, R. Hu, T. Darrell and B. Schiele, “Grounding of textual phrases in images by reconstruction”, in “European Conference on Computer Vision”, pp. 817–834 (Springer, 2016).
- Rombach, R., A. Blattmann, D. Lorenz, P. Esser and B. Ommer, “High-resolution image synthesis with latent diffusion models”, in “Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition”, pp. 10684–10695 (2022).
- Rosenfeld, A., *Digital picture processing* (Academic press, 1976).
- Rudin, L. I., S. Osher and E. Fatemi, “Nonlinear total variation based noise removal algorithms”, *Physica D: nonlinear phenomena* **60**, 1-4, 259–268 (1992).
- Rudinger, R., J. Naradowsky, B. Leonard and B. Van Durme, “Gender bias in coreference resolution”, in “Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)”, pp. 8–14 (Association for Computational Linguistics, New Orleans, Louisiana, 2018), URL <https://www.aclweb.org/anthology/N18-2002>.
- Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, “Imagenet large scale visual recognition challenge”, *International journal of computer vision* **115**, 3 (2015).
- Russell, B., “On denoting”, *Mind* **14**, 56, 479–493 (1905).
- Sagawa, S., P. W. Koh, T. B. Hashimoto and P. Liang, “Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization”, in “International Conference on Learning Representations”, (2020).
- Saharia, C., W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, “Photorealistic text-to-image diffusion models with deep language understanding”, *Advances in Neural Information Processing Systems* **35**, 36479–36494 (2022).
- Salimans, T., I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford and X. Chen, “Improved techniques for training gans”, in “Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain”, edited by D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon and R. Garnett, pp. 2226–2234 (2016), URL <https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html>.

Sap, M., R. L. Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith and Y. Choi, “ATOMIC: an atlas of machine commonsense for if-then reasoning”, in “The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019”, pp. 3027–3035 (AAAI Press, 2019a), URL <https://doi.org/10.1609/aaai.v33i01.33013027>.

Sap, M., H. Rashkin, D. Chen, R. Le Bras and Y. Choi, “Social IQa: Commonsense reasoning about social interactions”, in “Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)”, pp. 4463–4473 (Association for Computational Linguistics, Hong Kong, China, 2019b), URL <https://www.aclweb.org/anthology/D19-1454>.

Selvaraju, R. R., P. Tendulkar, D. Parikh, E. Horvitz, M. Ribeiro, B. Nushi and E. Kamar, “Squinting at vqa models: Interrogating vqa models with sub-questions”, in “IEEE Conference on Computer Vision and Pattern Recognition (CVPR)”, (2020).

Shen, W. B., D. Xu, Y. Zhu, F. Li, L. J. Guibas and S. Savarese, “Situational fusion of visual representation for visual navigation”, in “2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019”, pp. 2881–2890 (IEEE, 2019), URL <https://doi.org/10.1109/ICCV.2019.00297>.

Sheng, S., A. Singh, V. Goswami, J. Magana, T. Thrush, W. Galuba, D. Parikh and D. Kiela, “Human-adversarial visual question answering”, Advances in Neural Information Processing Systems **34**, 20346–20359 (2021).

Shi, H., J. Mao, T. Xiao, Y. Jiang and J. Sun, “Learning visually-grounded semantics from contrastive adversarial samples”, in “Proceedings of the 27th International Conference on Computational Linguistics”, pp. 3715–3727 (Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018), URL <https://www.aclweb.org/anthology/C18-1315>.

Shrestha, R., K. Kafle and C. Kanan, “Answer them all! toward universal visual question answering models”, in “IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019”, pp. 10472–10481 (Computer Vision Foundation / IEEE, 2019), URL http://openaccess.thecvf.com/content_CVPR_2019/html/Shrestha_Answer_Them_All_Toward_Universal_Visual_Question_Answering_Models_CVPR_2019_paper.html.

Shrestha, R., K. Kafle and C. Kanan, “An investigation of critical issues in bias mitigation techniques”, in “Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision”, pp. 1943–1954 (2022).

Sinha, A., H. Namkoong and J. C. Duchi, “Certifying some distributional robustness with principled adversarial training”, in “6th International Conference

on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings”, (OpenReview.net, 2018a), URL <https://openreview.net/forum?id=Hk6kPgZA->.

Sinha, A., H. Namkoong and J. C. Duchi, “Certifying some distributional robustness with principled adversarial training”, in “6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings”, (OpenReview.net, 2018b), URL <https://openreview.net/forum?id=Hk6kPgZA->.

Socher, R., D. Chen, C. D. Manning and A. Y. Ng, “Reasoning with neural tensor networks for knowledge base completion”, in “Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States”, edited by C. J. C. Burges, L. Bottou, Z. Ghahramani and K. Q. Weinberger, pp. 926–934 (2013), URL <https://proceedings.neurips.cc/paper/2013/hash/b337e84de8752b27eda3a12363109e80-Abstract.html>.

Spinoza, B. D., “Ethics, translated by andrew boyle, introduction by ts gregory”, (1934).

Stanovsky, G., N. A. Smith and L. Zettlemoyer, “Evaluating gender bias in machine translation”, in “Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics”, pp. 1679–1684 (Association for Computational Linguistics, Florence, Italy, 2019), URL <https://www.aclweb.org/anthology/P19-1164>.

Storkey, A., “When training and test sets are different: characterizing learning transfer”, Dataset shift in machine learning pp. 3–28 (2009).

Su, N. M. and D. J. Crandall, “The affective growth of computer vision”, in “Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition”, pp. 9291–9300 (2021).

Suhr, A., S. Zhou, A. Zhang, I. Zhang, H. Bai and Y. Artzi, “A corpus for reasoning about natural language grounded in photographs”, in “Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics”, pp. 6418–6428 (Association for Computational Linguistics, Florence, Italy, 2019), URL <https://www.aclweb.org/anthology/P19-1644>.

Sun, C., A. Myers, C. Vondrick, K. Murphy and C. Schmid, “Videobert: A joint model for video and language representation learning”, in “2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019”, pp. 7463–7472 (IEEE, 2019), URL <https://doi.org/10.1109/ICCV.2019.00756>.

Sun, Y., X. Wang, Z. Liu, J. Miller, A. A. Efros and M. Hardt, “Test-time training with self-supervision for generalization under distribution shifts”, in “Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event”, vol. 119 of *Proceedings of Machine Learning Research*, pp. 9229–9248 (PMLR, 2020), URL <http://proceedings.mlr.press/v119/sun20b.html>.

Sundermeyer, M., R. Schlüter and H. Ney, “Lstm neural networks for language modeling”, in “Thirteenth annual conference of the international speech communication association”, (2012).

Szegedy, C., W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow and R. Fergus, “Intriguing properties of neural networks”, in “2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings”, edited by Y. Bengio and Y. LeCun (2014), URL <http://arxiv.org/abs/1312.6199>.

Talmor, A., J. Herzig, N. Lourie and J. Berant, “CommonsenseQA: A question answering challenge targeting commonsense knowledge”, in “Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)”, pp. 4149–4158 (Association for Computational Linguistics, Minneapolis, Minnesota, 2019), URL <https://www.aclweb.org/anthology/N19-1421>.

Tan, H. and M. Bansal, “LXMERT: Learning cross-modality encoder representations from transformers”, in “Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)”, pp. 5100–5111 (Association for Computational Linguistics, Hong Kong, China, 2019), URL <https://www.aclweb.org/anthology/D19-1514>.

Taori, R., A. Dave, V. Shankar, N. Carlini, B. Recht and L. Schmidt, “Measuring robustness to natural distribution shifts in image classification”, in “Advances in Neural Information Processing Systems”, vol. 33, pp. 18583–18599 (2020).

Tapaswi, M., Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun and S. Fidler, “Movieqa: Understanding stories in movies through question-answering”, in “2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016”, pp. 4631–4640 (IEEE Computer Society, 2016), URL <https://doi.org/10.1109/CVPR.2016.501>.

Teney, D., E. Abbasnejad and A. v. d. Hengel, “Learning what makes a difference from counterfactual examples and gradient supervision”, in “European conference on computer vision”, (Springer, 2020a).

Teney, D. and A. v. d. Hengel, “Zero-shot visual question answering”, arXiv preprint arXiv:1611.05546 (2016).

Teney, D., K. Kafle, R. Shrestha, E. Abbasnejad, C. Kanan and A. v. d. Hengel, “On the value of out-of-distribution testing: An example of goodhart’s law”, arXiv preprint arXiv:2005.09241 (2020b).

Teney, D., Y. Lin, S. J. Oh and E. Abbasnejad, “Id and ood performance are sometimes inversely correlated on real-world datasets”, arXiv preprint arXiv:2209.00613 (2022).

Thiagarajan, J. J., R. Anirudh, V. Narayanaswamy and P.-T. Bremer, “Single model uncertainty estimation via stochastic data centering”, NeurIPS (2022).

Thomason, J., S. Venugopalan, S. Guadarrama, K. Saenko and R. Mooney, “Integrating language and vision to generate natural language descriptions of videos in the wild”, in “Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers”, pp. 1218–1227 (Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 2014), URL <https://www.aclweb.org/anthology/C14-1115>.

Tsipras, D., S. Santurkar, L. Engstrom, A. Turner and A. Madry, “Robustness may be at odds with accuracy”, in “7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019”, (OpenReview.net, 2019), URL <https://openreview.net/forum?id=SyxAb30cY7>.

Valiant, L. G., “A theory of the learnable”, Communications of the ACM **27**, 11, 1134–1142 (1984).

Van der Maaten, L. and G. Hinton, “Visualizing data using t-sne.”, Journal of machine learning research **9**, 11 (2008).

Vapnik, V., “Principles of risk minimization for learning theory”, in “Proceedings of the 4th International Conference on Neural Information Processing Systems”, pp. 831–838 (1991).

Vapnik, V. N. and A. Chervonenkis, “The necessary and sufficient conditions for consistency of the method of empirical risk minimization”, Pattern Recognition and Image Analysis **1**, 3, 284–305 (1991).

Varshney, N., P. Banerjee, T. Gokhale and C. Baral, “Unsupervised natural language inference using phl triplet generation”, in “Findings of the Association for Computational Linguistics: ACL 2022”, pp. 2003–2016 (2022).

Vasiljevic, I., A. Chakrabarti and G. Shakhnarovich, “Examining the impact of blur on recognition by convolutional networks”, arXiv preprint arXiv:1611.05760 (2016).

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, “Attention is all you need”, in “Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA”, edited by I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan and R. Garnett, pp. 5998–6008 (2017), URL <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fdb053c1c4a845aa-Abstract.html>.

Vedantam, R., X. Lin, T. Batra, C. L. Zitnick and D. Parikh, “Learning common sense through visual abstraction”, in “2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015”, pp. 2542–2550 (IEEE Computer Society, 2015a), URL <https://doi.org/10.1109/ICCV.2015.292>.

Vedantam, R., C. L. Zitnick and D. Parikh, “Cider: Consensus-based image description evaluation”, in “IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015”, pp. 4566–4575 (IEEE Computer Society, 2015b), URL <https://doi.org/10.1109/CVPR.2015.7299087>.

- Venkateswara, H., J. Eusebio, S. Chakraborty and S. Panchanathan, “Deep hashing network for unsupervised domain adaptation”, in “2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017”, pp. 5385–5394 (IEEE Computer Society, 2017), URL <https://doi.org/10.1109/CVPR.2017.572>.
- Venugopalan, S., M. Rohrbach, J. Donahue, R. J. Mooney, T. Darrell and K. Saenko, “Sequence to sequence - video to text”, in “2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015”, pp. 4534–4542 (IEEE Computer Society, 2015), URL <https://doi.org/10.1109/ICCV.2015.515>.
- Volpi, R. and V. Murino, “Addressing model vulnerability to distributional shifts over image transformation sets”, in “Proceedings of the IEEE/CVF International Conference on Computer Vision”, pp. 7980–7989 (2019).
- Volpi, R., H. Namkoong, O. Sener, J. C. Duchi, V. Murino and S. Savarese, “Generalizing to unseen domains via adversarial data augmentation”, in “Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada”, edited by S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, pp. 5339–5349 (2018), URL <https://proceedings.neurips.cc/paper/2018/hash/1d94108e907bb8311d8802b48fd54b4a-Abstract.html>.
- Vondrick, C., D. Oktay, H. Pirsiavash and A. Torralba, “Predicting motivations of actions by leveraging text”, in “2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016”, pp. 2997–3005 (IEEE Computer Society, 2016), URL <https://doi.org/10.1109/CVPR.2016.327>.
- Wei, J. and K. Zou, “EDA: Easy data augmentation techniques for boosting performance on text classification tasks”, in “Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)”, pp. 6382–6388 (Association for Computational Linguistics, Hong Kong, China, 2019), URL <https://www.aclweb.org/anthology/D19-1670>.
- Welinder, P., S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie and P. Perona, “Caltech-ucsd birds 200”, Tech. Rep. CNS-TR-201, Caltech, URL [/se3/wp-content/uploads/2014/09/WelinderEtal10_CUB-200.pdf](https://se3/wp-content/uploads/2014/09/WelinderEtal10_CUB-200.pdf), <http://www.vision.caltech.edu/visipedia/CUB-200.html> (2010).
- Weston, J., A. Bordes, S. Chopra and T. Mikolov, “Towards ai-complete question answering: A set of prerequisite toy tasks”, in “4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings”, edited by Y. Bengio and Y. LeCun (2016), URL <http://arxiv.org/abs/1502.05698>.
- Williams, A., N. Nangia and S. Bowman, “A broad-coverage challenge corpus for sentence understanding through inference”, in “Proceedings of the 2018 Conference

- of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)”, pp. 1112–1122 (Association for Computational Linguistics, New Orleans, Louisiana, 2018), URL <https://www.aclweb.org/anthology/N18-1101>.
- Wisdom, E., T. Gokhale, C. Xiao and Y. Yang, “Mole recruitment: Poisoning of image classifiers via selective batch sampling”, arXiv preprint arXiv:2303.17080 (2023).
- Wittgenstein, L., *Tractatus logico-philosophicus* (Routledge, 1921).
- Wong, E. and J. Z. Kolter, “Learning perturbation sets for robust machine learning”, in “International Conference on Learning Representations”, (2020).
- Wren, P. and H. Martin, “English grammar and composition”, New Delhi: S Chand & Company Ltd (2000).
- Wu, M., N. S. Moosavi, A. Rücklé and I. Gurevych, “Improving QA generalization by concurrent modeling of multiple biases”, in “Findings of the Association for Computational Linguistics: EMNLP 2020”, pp. 839–853 (Association for Computational Linguistics, Online, 2020), URL <https://www.aclweb.org/anthology/2020.findings-emnlp.74>.
- Wu, Y. and K. He, “Group normalization”, in “Proceedings of the European conference on computer vision (ECCV)”, pp. 3–19 (2018).
- Xiao, K., L. Engstrom, A. Ilyas and A. Madry, “Noise or signal: The role of image backgrounds in object recognition”, in “International Conference on Learning Representations”, (2021).
- Xie, Q., Z. Dai, E. Hovy, T. Luong and Q. Le, “Unsupervised data augmentation for consistency training”, Advances in Neural Information Processing Systems **33** (2020).
- Xu, D., Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He and Y. Zhuang, “Video question answering via gradually refined attention over appearance and motion”, in “Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017”, pp. 1645–1653 (2017), URL <https://doi.org/10.1145/3123266.3123427>.
- Xu, J., T. Mei, T. Yao and Y. Rui, “MSR-VTT: A large video description dataset for bridging video and language”, in “2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016”, pp. 5288–5296 (IEEE Computer Society, 2016), URL <https://doi.org/10.1109/CVPR.2016.571>.
- Xu, T., P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang and X. He, “AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks”, in “2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018”, pp. 1316–1324 (IEEE Computer Society, 2018), URL http://openaccess.thecvf.com/content_cvpr_2018/html/Xu_AttnGAN_Fine-Grained_Text_CVPR_2018_paper.html.

- Xu, Y., L. Chen, Z. Cheng, L. Duan and J. Luo, “Open-ended visual question answering by multi-modal domain adaptation”, in “Findings of the Association for Computational Linguistics: EMNLP 2020”, pp. 367–376 (Association for Computational Linguistics, Online, 2020a), URL <https://www.aclweb.org/anthology/2020.findings-emnlp.34>.
- Xu, Z., D. Liu, J. Yang, C. Raffel and M. Niethammer, “Robust and generalizable visual representation learning via random convolutions”, in “International Conference on Learning Representations”, (2020b).
- Yang, H., L. Chaisorn, Y. Zhao, S.-Y. Neo and T.-S. Chua, “Videoqa: question answering on news video”, in “Proceedings of the eleventh ACM international conference on Multimedia”, pp. 632–641 (ACM, 2003).
- Yang, J., K. Zhou, Y. Li and Z. Liu, “Generalized out-of-distribution detection: A survey”, arXiv preprint arXiv:2110.11334 (2021).
- Yang, Y., C. Teo, H. Daumé III and Y. Aloimonos, “Corpus-guided sentence generation of natural images”, in “Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing”, pp. 444–454 (Association for Computational Linguistics, Edinburgh, Scotland, UK., 2011), URL <https://www.aclweb.org/anthology/D11-1041>.
- Yang, Y.-Y., C. Rashtchian, H. Zhang, R. R. Salakhutdinov and K. Chaudhuri, “A closer look at accuracy vs. robustness”, Advances in neural information processing systems **33**, 8588–8601 (2020).
- Yatskar, M., L. S. Zettlemoyer and A. Farhadi, “Situation recognition: Visual semantic role labeling for image understanding”, in “2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016”, pp. 5534–5542 (IEEE Computer Society, 2016), URL <https://doi.org/10.1109/CVPR.2016.597>.
- Ye, K. and A. Kovashka, “A case study of the shortcut effects in visual commonsense reasoning”, in “Proceedings of the AAAI conference on artificial intelligence”, vol. 35, pp. 3181–3189 (2021).
- Yi, M., L. Hou, J. Sun, L. Shang, X. Jiang, Q. Liu and Z. Ma, “Improved ood generalization via adversarial training and pretraing”, in “International Conference on Machine Learning”, pp. 11987–11997 (PMLR, 2021).
- Young, P., A. Lai, M. Hodosh and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions”, Transactions of the Association for Computational Linguistics **2**, 67–78, URL <https://www.aclweb.org/anthology/Q14-1006> (2014).
- Yu, J., Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan *et al.*, “Scaling autoregressive models for content-rich text-to-image generation”, Transactions on Machine Learning Research (2022).

- Yu, L., E. Park, A. C. Berg and T. L. Berg, “Visual madlibs: Fill in the blank description generation and question answering”, in “2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015”, pp. 2461–2469 (IEEE Computer Society, 2015), URL <https://doi.org/10.1109/ICCV.2015.283>.
- Yu, Z., J. Yu, Y. Cui, D. Tao and Q. Tian, “Deep modular co-attention networks for visual question answering”, in “IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019”, pp. 6281–6290 (Computer Vision Foundation / IEEE, 2019), URL http://openaccess.thecvf.com/content_CVPR_2019/html/Yu_Deep_Modular_Co-Attention_Networks_for_Visual_Question_Answering_CVPR_2019_paper.html.
- Yuan, X., P. He, Q. Zhu and X. Li, “Adversarial examples: Attacks and defenses for deep learning”, *IEEE transactions on neural networks and learning systems* **30**, 9, 2805–2824 (2019).
- Yuille, A. L. and C. Liu, “Deep nets: What have they ever done for vision?”, *International Journal of Computer Vision* **129**, 3, 781–802 (2021).
- Yun, S., D. Han, S. Chun, S. J. Oh, Y. Yoo and J. Choe, “Cutmix: Regularization strategy to train strong classifiers with localizable features”, in “2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019”, pp. 6022–6031 (IEEE, 2019), URL <https://doi.org/10.1109/ICCV.2019.00612>.
- Zadeh, A., M. Chan, P. P. Liang, E. Tong and L. Morency, “Social-iq: A question answering benchmark for artificial social intelligence”, in “IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019”, pp. 8807–8817 (Computer Vision Foundation / IEEE, 2019), URL http://openaccess.thecvf.com/content_CVPR_2019/html/Zadeh_Social-IQ_A_Question_Answering_Benchmark_for_Artificial_Social_Intelligence_CVPR_2019_paper.html.
- Zellers, R., Y. Bisk, A. Farhadi and Y. Choi, “From recognition to cognition: Visual commonsense reasoning”, in “IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019”, pp. 6720–6731 (Computer Vision Foundation / IEEE, 2019), URL http://openaccess.thecvf.com/content_CVPR_2019/html/Zellers_From_Recognition_to_Cognition_Visual_Commonsense_Reasoning_CVPR_2019_paper.html.
- Zellers, R., Y. Bisk, R. Schwartz and Y. Choi, “SWAG: A large-scale adversarial dataset for grounded commonsense inference”, in “Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing”, pp. 93–104 (Association for Computational Linguistics, Brussels, Belgium, 2018), URL <https://www.aclweb.org/anthology/D18-1009>.
- Zettlemoyer, L. S. and M. Collins, “Learning to map sentences to logical form: structured classification with probabilistic categorial grammars”, in “Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence”, pp. 658–666 (2005).

- Zhang, H., M. Cissé, Y. N. Dauphin and D. Lopez-Paz, “mixup: Beyond empirical risk minimization”, in “6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings”, (OpenReview.net, 2018), URL <https://openreview.net/forum?id=r1Ddp1-Rb>.
- Zhang, H., T. Xu and H. Li, “Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks”, in “IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017”, pp. 5908–5916 (IEEE Computer Society, 2017), URL <https://doi.org/10.1109/ICCV.2017.629>.
- Zhang, P., X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi and J. Gao, “Vinvl: Revisiting visual representations in vision-language models”, in “Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition”, pp. 5579–5588 (2021).
- Zhang, X., Q. Wang, J. Zhang and Z. Zhong, “Adversarial autoaugment”, in “8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020”, (OpenReview.net, 2020), URL <https://openreview.net/forum?id=ByxdUySKvS>.
- Zhao, J., T. Wang, M. Yatskar, V. Ordonez and K.-W. Chang, “Men also like shopping: Reducing gender bias amplification using corpus-level constraints”, in “Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing”, pp. 2979–2989 (Association for Computational Linguistics, Copenhagen, Denmark, 2017), URL <https://www.aclweb.org/anthology/D17-1323>.
- Zhong, Z., L. Zheng, G. Kang, S. Li and Y. Yang, “Random erasing data augmentation”, in “Proceedings of the AAAI Conference on Artificial Intelligence”, vol. 34, pp. 13001–13008 (2020).
- Zhou, B., H. Zhao, X. Puig, S. Fidler, A. Barriuso and A. Torralba, “Scene parsing through ADE20K dataset”, in “2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017”, pp. 5122–5130 (IEEE Computer Society, 2017), URL <https://doi.org/10.1109/CVPR.2017.544>.
- Zhou, K., Y. Yang, T. Hospedales and T. Xiang, “Learning to generate novel domains for domain generalization”, in “European Conference on Computer Vision”, pp. 561–578 (Springer, 2020a).
- Zhou, L., H. Palangi, L. Zhang, H. Hu, J. J. Corso and J. Gao, “Unified vision-language pre-training for image captioning and vqa.”, in “AAAI”, pp. 13041–13049 (2020b).
- Zhou, L., Y. Zhou, J. J. Corso, R. Socher and C. Xiong, “End-to-end dense video captioning with masked transformer”, in “2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018”, pp. 8739–8748 (IEEE Computer Society, 2018), URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhou_End-to-End_Dense_Video_CVPR_2018_paper.html.

Zhu, C., Y. Cheng, Z. Gan, S. Sun, T. Goldstein and J. Liu, “Freelb: Enhanced adversarial training for natural language understanding”, in “8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020”, (OpenReview.net, 2020), URL <https://openreview.net/forum?id=BygzbbyHFvB>.

Zhu, L., Z. Xu, Y. Yang and A. G. Hauptmann, “Uncovering the temporal context for video question answering”, International Journal of Computer Vision **124**, 3, 409–421 (2017).

Zhu, Y., R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba and S. Fidler, “Aligning books and movies: Towards story-like visual explanations by watching movies and reading books”, in “2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015”, pp. 19–27 (IEEE Computer Society, 2015), URL <https://doi.org/10.1109/ICCV.2015.11>.

Zoph, B., E. D. Cubuk, G. Ghiasi, T.-Y. Lin, J. Shlens and Q. V. Le, “Learning data augmentation strategies for object detection”, in “European Conference on Computer Vision”, pp. 566–583 (Springer, 2020).

APPENDIX A
STATEMENT ON PREVIOUSLY PUBLISHED ARTICLES

Tejas Gokhale confirms that for all previously published work mentioned and included in this dissertation, co-authors of such work have granted their permission to use these articles.

ProQuest Number: 30426752

INFORMATION TO ALL USERS

The quality and completeness of this reproduction is dependent on the quality
and completeness of the copy made available to ProQuest.



Distributed by ProQuest LLC (2023).

Copyright of the Dissertation is held by the Author unless otherwise noted.

This work may be used in accordance with the terms of the Creative Commons license
or other rights statement, as indicated in the copyright statement or in the metadata
associated with this work. Unless otherwise specified in the copyright statement
or the metadata, all rights are reserved by the copyright holder.

This work is protected against unauthorized copying under Title 17,
United States Code and other applicable copyright laws.

Microform Edition where available © ProQuest LLC. No reproduction or digitization
of the Microform Edition is authorized without permission of ProQuest LLC.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346 USA