Cognitive Models in Deep Learning

Idan Schwartz

Cognitive Models in Deep Learning

Research Thesis

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Idan Schwartz

Submitted to the Senate
of the Technion — Israel Institute of Technology
Tamuz 5781 Haifa July 2021

This research was carried out under the supervision of Prof. Tamir Hazan and Prof. Alexander Schwing in the Faculty of Computer Science.

Some results in this thesis have been published as articles by the author and research collaborators in conferences and journals during the course of the author's doctoral research period, the most up-to-date versions of which being:

Itai Gat, Idan Schwartz, and Alexander Schwing. Perceptual score: Measuring perceptiveness of multi-modal classifiers. *Advances in Neural Information Processing Systems*, 34, 2021.

Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *Advances in Neural Information Processing Systems*, 33, 2020.

Idan Schwartz. Ensemble of mrr and ndcg models for visual dialog. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

Idan Schwartz, Alexander G Schwing, and Tamir Hazan. High-order attention models for visual question answering. *Advances in Neural Information Processing Systems*, 30, 2017.

Idan Schwartz, Alexander G Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. Factor graph attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Acknowledgements

I wish to thank my advisors, Alexander Schwing, and Tamir Hazan. It was a pleasure working with two advisors: Tamir's insights always added spice to our work, and Alex was willing to take on challenging projects with me. Our research meetings quickly devolved into philosophical discussions, enabling us to question elementary aspects of machine learning. Tamir and Alex have taught me everything I need to know to become a good researcher.

Further, I would like to thank Benny Kimelfield for his guidance during my master's degree.

My studies at the Technion have allowed me to meet many friends and have fruitful discussions that have inspired my research and sparked many exciting ideas. In particular, I would like to thank Yftah Ziser and Jalil Moroney, my good friends from the legendary 216 office. Thanks for adding laughter and unforgettable memories to the office. Itai Gat, one of my closest collaborators, worked with me late at night on deadlines. Thanks also to Nimrod Raifer, who always has excellent advice. Last but not least, Maayan Kislev for supporting and encouraging frequent breaks from routine for travel and relaxation. There are many others I've been fortunate to know, collaborate with, and become friends with: Edward Vitking, Guy Uziel, Michal Badian,

Gali Sheffi, Shoval Lagziel, Yonatan Geifman, Michal Friedman, Eytan Singher, Idan Hasson, Matan Peled.

I had the opportunity to work at eBay's research team and collaborate with Kira Radinsky and Ido Guy. Thanks for mentoring me and providing me with a wonderful experience. At eBay, it was a pleasure to work with Yotam Eshel, Nir Levin, Shai Haim, and Nir Ofek.

I have also enjoyed working at Microsoft and collaborating with Tom Braude, Shlomi Maliah, Lera Shtotland, Tzoof Avny Brosh, Ran Bernstein, Sagi Hilleli, and Ido Priness.

Finally, I would like to thank my parents, Nely and Zeev, for supporting my passion for computers from a young age. Also, I would like to express gratitude to my brother Dor, who has always been my best friend. My grandparents, Gila and Morris, for their constant love and care. Thank you from the bottom of my heart for doing everything possible to ensure I could fulfill my dreams without worry.

The generous financial help of the Technion is gratefully acknowledged.

Contents

List of Figures

Abstract **1**

1 Introduction **3**
1.1 Overview . . . 3
1.2 Related Work . . . 5
1.3 Contributions and Outline . . . 6

2 Deep Learning Background **11**
2.1 Deep Neural Networks . . . 11
2.2 Convolutional Neural Networks . . . 12
2.3 Recurrent Neural Network . . . 14
2.3.1 Long Short Term Memory . . . 14

3 Visual Question Answering **17**
3.1 Related Work . . . 18
3.2 Higher order attention models . . . 20
3.2.1 Data Embedding . . . 20
3.2.2 Attention . . . 21
3.2.3 Decision Making . . . 23
3.3 High-order Attention for Visual Question Answering . . . 24
3.3.1 Data Embedding . . . 24
3.3.2 Decision Making . . . 26
3.3.3 Results . . . 26
3.4 Conclusion . . . 28

4 Visual Dialog **31**
4.1 Related Work . . . 33
4.2 Factor Graph Attention . . . 35
4.2.1 Local Factors . . . 37
4.2.2 Joint Factors . . . 37
4.2.3 Attention, Messages and Beliefs . . . 38

4.3 Factor Graph Attention for Visual Dialog . . . 39
4.3.1 Utilities and Embeddings . . . 39
4.3.2 Attention Module . . . 39
4.3.3 Fusion Step . . . 40
4.4 Results . . . 41
4.5 Ensemble of MRR and NDCG models for Visual Dialog . . . 49
4.5.1 Related Work . . . 49
4.5.2 Two-step Rank Ensemble . . . 51
4.5.3 MRR Step . . . 52
4.5.4 NDCG Step . . . 53
4.6 Results . . . 54
4.7 Conclusions . . . 56

5 Visual Storytelling **59**
5.1 Related Work . . . 61
5.2 Method . . . 62
5.2.1 Ordered Image Attention (OIA) . . . 63
5.2.2 Image-Sentence Attention (ISA) . . . 66
5.2.3 Story Decoding . . . 67
5.3 Results . . . 68
5.3.1 Training Setup . . . 68
5.3.2 Quantitative Analysis . . . 70
5.3.3 Human Evaluation . . . 71
5.3.4 Qualitative Evaluation . . . 73
5.4 Conclusion . . . 73

6 Audio-Visual Scene-Aware Dialog **75**
6.1 Related Work . . . 77
6.2 Audio Visual Scene-Aware Dialog Baselines . . . 79
6.2.1 Answer Generation . . . 79
6.2.2 Attention . . . 81
6.2.3 Data Representation . . . 82
6.3 Results . . . 83
6.3.1 AVSD v0.1 Dataset . . . 84
6.3.2 Implementation Details . . . 84
6.3.3 Training . . . 84
6.3.4 Performance Evaluation: . . . 85
6.3.5 Quantitative Results and Insights for a Good Baseline . . . 86
6.3.6 Qualitative Results . . . 88
6.4 Conclusion . . . 89

7 Perceptual Score: Measuring Perceptiveness of Multi-Modal Classifiers 91
7.1 Related Work . . . 93
7.2 The Perceptual Score . . . 94
7.2.1 Setup . . . 95
7.2.2 Perceptual Score of a Data Modality . . . 95
7.3 Evaluation of Perceptual Scores . . . 97
7.3.1 Visual Question Answering . . . 98
7.3.2 Video Social Reasoning . . . 101
7.3.3 Visual Dialog . . . 102
7.4 Conclusion . . . 103

8 Removing Bias in Multi-modal Classifiers: Regularization by Maximizing Functional Entropies 105
8.1 Related Work . . . 106
8.2 Background . . . 107
8.2.1 Functional entropy . . . 108
8.2.2 Functional Fisher information . . . 108
8.2.3 Tensorization and multi-modal data . . . 109
8.3 Regularization by Maximizing Functional Entropies . . . 109
8.3.1 Tensorization . . . 110
8.4 Connection Between Functional Entropy and Variance . . . 111
8.4.1 Regularization using Variance . . . 112
8.5 Experiments . . . 112
8.5.1 Colored MNIST . . . 113
8.5.2 VQA-CPv2 . . . 115
8.5.3 SocialIQ . . . 115
8.5.4 Dogs and Cats . . . 116
8.6 Conclusion . . . 116

9 Conclusions 117

Hebrew Abstract i

List of Figures

1.1 As cognitive tasks evolved, datasets included more modalities. There are two challenges: the ability to attend the important things and perceive all the input data. Note, the third-row figure illustrates a sound-enabled video and taken from [ACD+18]. 4

1.2 Multi-modal datasets often have an undesired bias: a classifier exploits shortcuts and predicts the correct answer based on parts of the data. For instance, in SocialIQ the task requires to understand social situations shown in video data. However, a classifier can achieve high accuracy by only perceiving textual cues because the answer correctness with subtle question cues. 5

2.1 Illustration of one layer linear classification 12

2.2 Translational invariant functions $f : \mathbb{R}^2 \to \mathbb{R}$ it is $f(A) = f(A+t)$. Image patterns have translational invariance characteristics. Source: Stephan Kulla. 13

2.3 A filter (kernel) convolved over an image of $32 \times 32 \times 3$ dimension. The output depth is dependent on numbers of filters. In this case the output depth is 5, which means the convolutional layer have 5 filters. Source: cs231n. 13

2.4 Convolutional networks were motivated by biological processes in that the connectivity pattern between neurons resembles the animal visual cortex's organization. Source: Kubilius, Jonas (`https://doi.org/10.6084/m9.figshare.106794.v3`) 13

2.5 Illustration of an RNN layer. RNN is build feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video). Source: Nature. 14

2.6 Memory cell introduces gates to avoid vanishing/exploding gradient problems by learning how to update the hidden state. Diagram is taken from [Che18]. 15

3.1 Results of our multi-modal attention for one image and two different questions (1st column). The unary image attention is identical by construction. The pairwise potentials differ for both questions and images since both modalities are taken into account (3rd column). The final attention is illustrated in the 4th column. 18

3.2 Our state-of-the-art VQA system 20

3.3 Illustration of our $k-$order attention. (a) unary attention module (e.g., visual). (b) pairwise attention module (e.g., visual and question) marginalized over its two data modalities. (c) ternary attention module (e.g., visual, question and answer) marginalized over its three data modalities. 22

3.4 Illustration of correlation units used for decision making. (a) MCB unit approximately sample from outer product space of two attention vectors, (b) MCT unit approximately sample from outer product space of three attention vectors. . . 24

3.5 For each image (1st column) we show the attention generated for two different questions in columns 2-4 and columns 5-7 respectively. The attentions are ordered as unary attention, pairwise attention and combined attention for both the image and the question. We observe the combined attention to significantly depend on the question. 26

3.6 The attention generated for two different questions over three modalities. We find the attention over multiple choice answers to emphasis the unusual answers. 26

3.7 Comparison of our attention results (2nd column) with attention provided by [LYBP16] (3rd column) and [FPY+16] (4th column). The fourth column provides the question and the answer of the different techniques. 27

3.8 Failure cases: Unary, pairwise and combined attention of our approach. Our system focuses on the colorful umbrella as opposed to the table in the first row. 29

4.1 Illustration of our factor graph attention. We show two consecutive questions in a dialog. The image attention correlates well with the question. Attention over history interactions allows our model to attend to subtle nuances. The caption focuses on the last word due to given potential priors. Attention over the answers focuses on specific options. The attended options usually correlate with the correct answer. Note: for readability, we chose to display only the top-10 answers out of 100 possible ones. 32

4.2 Our state-of-the-art architecture for the Visual Dialog task. Implementations details can be found in Sec. 4.3. 35

4.4 An illustration of question and image attention over a series of interactions for the same dialog. In addition we provide the ground truth answer, *i.e.*, GT, and our predicted answer, *i.e.*, A. 43

4.5 Illustration of history attention for 2 interactions. We observe small nuances of history to be useful to answer questions, and improve co-reference resolution. 45

4.6 Two images each with two questions. We illustrate scores obtained from different types of factors. Local-info denotes 'Image-Local-Information,' Question refers to 'Image-Question,' *etc.* We observe 'Image-Question' to have the highest variance between different questions, since its heat map differs the most. 'Image-Question' also correlates the most with the final attention. 47

4.7 Illustration of 2 step interaction using visual question generation and illustration of the involved modalities. The classifier receives the previous question and answer, to predict a new one. 48

4.8 A visual dialog interaction. The question asks, "what is the nightstand made of ?". We show our final ranking, created by the ensemble of an MRR/NDCG models' rankings. The MRR/NDCG models are trained to optimize the MRR/NDCG metric. The MRR metric measures the number of retrievals to retrieve the human-derived answer. Hence, the MRR model favors human-like and detailed answers. On the other hand, the NDCG metric measures the rank of all the correct candidates based on dense annotation, which are often general and uncertain. Our ensemble approach seeks a minimal candidate set that is likely to contain the human-derived answer. The remaining candidates are ranked according to the NDCG model. 50

4.9 Performance of a naïve score ensemble of the MRR model and the NDCG model on the VisDialv1.0 val set. We calibrate the importance of each model with a scalar α. 54

4.10 MRR and NDCG scores for different hyperparameter values. 54

4.11 An illustration of two visual dialog samples. Each sample includes the MRR candidate set and four answers from the remaining NDCG candidates. We find that the MRR candidate set has more certain answers. We colorize the *high-certainty* candidates ($\mathcal{H}$) with orange, the *NDCG-agreement* candidates ($\mathcal{N}$) with purple, and the *top-answers* subset ($\mathcal{T}$) with red. Note, if a candidate belongs to more than one set, we sketch the colors in the following order: orange→red→purple. 56

5.1 We propose Ordered Image Attention (OIA) to form the structure of a sentence and to encourage coherency. Each row shows the spatial attention of the five images created when generating a specific sentence. We find important objects by collecting directional interactions. The relative order to the sentence-corresponding image determines the connection type, illustrated as the blue and orange edges for preceding and proceeding connections. The attended images' border indicates the image attention importance formed by the Image-Sentence Attention (ISA). *E.g.*, red indicates a high attention score, meaning the image is essential for generating that sentence. Our model performs this step for all five images in parallel, creating a total of 25 spatial attention maps, that are fed into the decoder to create the sentences in order. 60

5.2 Our architecture for Visual Storytelling synthesis. 61

5.3 Illustration of Ordered Image Attention. Each node represents an image attention belief. For each sentence, we connect all the images with the sentence-corresponding image. The relative position to this image determines whether the connection is modeled with the Ψ_{bwd} factor (for preceding images) or the Ψ_{fwd} factor (for subsequent images). We infer the attention belief by collecting interactions and local object information within the image. We use scalars to calibrate the importance of each factor. In total, we generate 25 attention maps, one per image for every sentence. 65

5.4 Illustration of ISA. The attention selects the attended image representation per sentence. We model interactions between attended images of the same sentence to compute each image's importance. Note, each node represents a sentence attention belief over the attended images. 66

5.5 Human evaluation to compare human-like and coherency properties. . . 69

5.6 An illustration of an image sequence along with three different stories generated by: (1) AREL baseline [WCfWW18], (2) No History: a model without intra-repetition regularization and BOW prior (see Sec. 5.2.3); and (3) With History: the final model. Repeated sentences are highlighted with a yellow colored marker. Repeated words in a sentence are emphasized in red color. 72

5.7 Illustration of OIA and ISA attention maps, the ground-truth story and the final generated story. Each row corresponds to a story sentence and shows objects OIA highlights. The attended images' border specifies the relevancy to sentence generation, from red (important) to blue (not important). 74

6.1 We present 4 different questions and the generated answer. Our attention unit is illustrated as well. Our model samples 4 frames, and attends to each frame separately, along with the question and the audio. We observe attention for each frame to differ, where first and fourth frames are widespread, while the second and third are more specific. Also, the question attention attends to relevant words. We also include the audio modality as input to the attention computation. 76

6.2 Overview of our approach for the AVSD task. More details can be found in Sec. 6.2. 77

6.3 Our decoder for audio-visual scene-aware dialog. We start with encoding of attended audio and video vectors using the Aud-Vis LSTM (orange colored), followed by the Ans-Generation LSTM that receives the textual data concatenated with the previous answer word (green colored). 80

6.4 Multimodal Attention model for audio-visual scene-aware dialog. We treat each frame as a modality, along with audio and question modality, to total of 6 modalities. Each element attention score is affected not only from local evidence, but also via cross-data interactions of all other elements. 81

6.5 Perplexity values for our model *vs.* baseline [HAW$^+$18] 83

6.6 An illustration of out 4-framed samples from a video along with the relevant attention variables. Our attention treats any frame as different component. This allows the attention module to learn different attention behaviors for different temporal locations. We observe the first and fourth samples are noisier, while the second and third attend to specific interesting locations. Our multimodal attention also generates attention for questions, illustrated over the question via a word heat map. We provide generated answers for different baseline models: q+h+att, is a model with only history and question input; i3d-rgb-temporal is a model with temporal features instead of spatial; q+h+vgg-spatial+audio is a model without attention. We also compare to the generated answer by [HAW$^+$18]. the ground-truth is denoted by GT, and our final model denoted by Ours. 85

7.1 Multi-modal datasets often have undesired biases: (a) To identify those biases we suggest the perceptual score as a new metric. It assesses the change in prediction when a model's input for some modalities is permuted during testing. If the classifier output remains identical despite permutation, a model doesn't perceive the modality. (b) Using the perceptual score we identify that recent progress of VQA models may not be entirely due to better reasoning. 92

7.2 SocialIQ data samples. On the left, we show a sample with a high perceptual score towards video data. Neither a positive nor a negative sentiment is evident in this sample. Hence, the video is required for prediction. We illustrate two samples (marked with a red border) that received a low perceptual score. There is a sentiment-based correlation between the label and the answer in these samples. For simplicity, we highlight with red color words that exhibit sentiment. 101

8.1 We illustrate our approach. In the visual question answering task, we are given a question about an image. Thus, we can partition our input into two modalities: a textual modality, and a visual modality. We measure the modalities' functional Fisher information by evaluating the sensitivity of the prediction by perturbing each modality. We maximize the functional Fisher information by incorporating it into our loss as a regularization term. Our results show that our regularization permits higher utilization of the visual modality. 106

8.2 Proportions of the Fisher information values during training for SocialIQ, Colored MNIST, VQA-CPv2 and Dogs&Cats. Using our proposed regularization brings the modalities Fisher information value closer than training without our regularization, a desired property in multi-modal learning. In ColoredMNIST, we observe that training a model with our regularization, the prediction is based on both the shape and the color. Unlike, a model trained without our regularization which makes predictions based on the color only. 114

8.3 Training process with and without regularization. We note that generalization significantly improves when using our proposed regularization. . 114

Abstract

The quest for algorithms that enable cognitive abilities is an integral part of machine learning and has many facets, such as visual question answering and dialog generation. A common trait of these cognitive-like tasks is that they consider different data modalities, for example, visual and lingual data.

Attention mechanisms have emerged as a prominent common theme to address these tasks. They provide not only some form of interpretability but also often improve performance. The latter effect is attributed to more concise forms of the various data modalities. However, present-day attention mechanisms are often geared towards a specific form of input and therefore hand-crafted for a particular task in an ad-hoc and entangled manner. As datasets continue to grow in size, the ad-hoc paradigm is no longer tractable and can lead to biased models.

To address these issues, we propose a novel and generally applicable form of attention mechanism, namely 'Factor Graph Attention,' that learns high-order correlations between various parts of the data input. For example, the second-order correlation factor can model interactions between two data modalities, *e.g.*, an image and a question, and more generally, $k-$th order correlation can model interactions between k modalities. Learning these correlations directs the appropriate attention to the relevant elements in the different data modalities required to solve the joint task.

We demonstrate our novel attention mechanism's effectiveness in various cognitive tasks, such as Visual Question Answering, Visual Dialog, Visual Storytelling, and Audio-Visual Scene-Aware Dialog.

Despite the substantial improvements that the attention mechanism achieves, large datasets are hard to annotate and contain biases that we are often unaware of. Attention-based classifiers, in turn, are prone to exploit those biases and to find shortcuts. As a consequence, current methods may solve the dataset but not directly the task. To address this concern, we introduce perceptual scores that assess the degree to which a model relies on the input features' different subsets (*i.e.*, modalities). For instance, a high image perceptual score indicates that the model relied on the image for its decision. We also study regularization to increase perceptiveness, by maximizing the functional entropy of modalities during training. We validate the efficacy of the proposed method on the synthetic 'Colored MNIST,' and other datasets, such as 'VQA,' 'SocialIQ,' and 'SNLI.'

Chapter 1

Introduction

1.1 Overview

Several large-scale datasets have emerged in recent years. They aim to enable development of methods that mimic the human ability of reasoning about the world given a plethora of cognitive inputs (see Fig. 1.1). A fundamental example is the Visual Question Answering task, in which the cognitive input is an accompanying visual modality that is needed to accomplish the task. A more recent example, such as Audio-Visual Scene-Aware dialog, uses as cognitive input a video, a dialog history, and audio. One of the traits of these cognitive-like tasks is that they take into account many data modalities. In this dissertation, attention and perception are examined in depth

> *Attention is taking possession of the mind, in clear and vivid form, of one out of what seems several simultaneously possible objects or trains of thought. Focalization, concentration of consciousness are of its essence. It implies a withdrawal from some things in order to deal effectively with others.*
>
> — William James (1890)

Throughout history, philosophers have held the suspicion that we perceive much more than we are able to notice. Despite the vast amount of different information that is permanently occupying our nervous system, we are often easily able to quickly discern important cues from data that is irrelevant. Telling apart useful information from distracting aspects is also an important ability for virtual assistants, car navigation systems, or smart speakers. However, present-day technology uses a chain of components from speech recognition and dialog management to sentence generation and speech synthesis, making it hard to design a holistic and entirely data-driven approach.

To address these tasks, recently, attention mechanisms have emerged as a powerful common theme, which provides not only some form of interpretability if applied to deep net models, but also often improves performance [HKG+15]. The latter effect is attributed to more expressive yet concise representations of the various data modalities. Nonetheless, multi-modal attention is typically tailored to specific tasks individually.

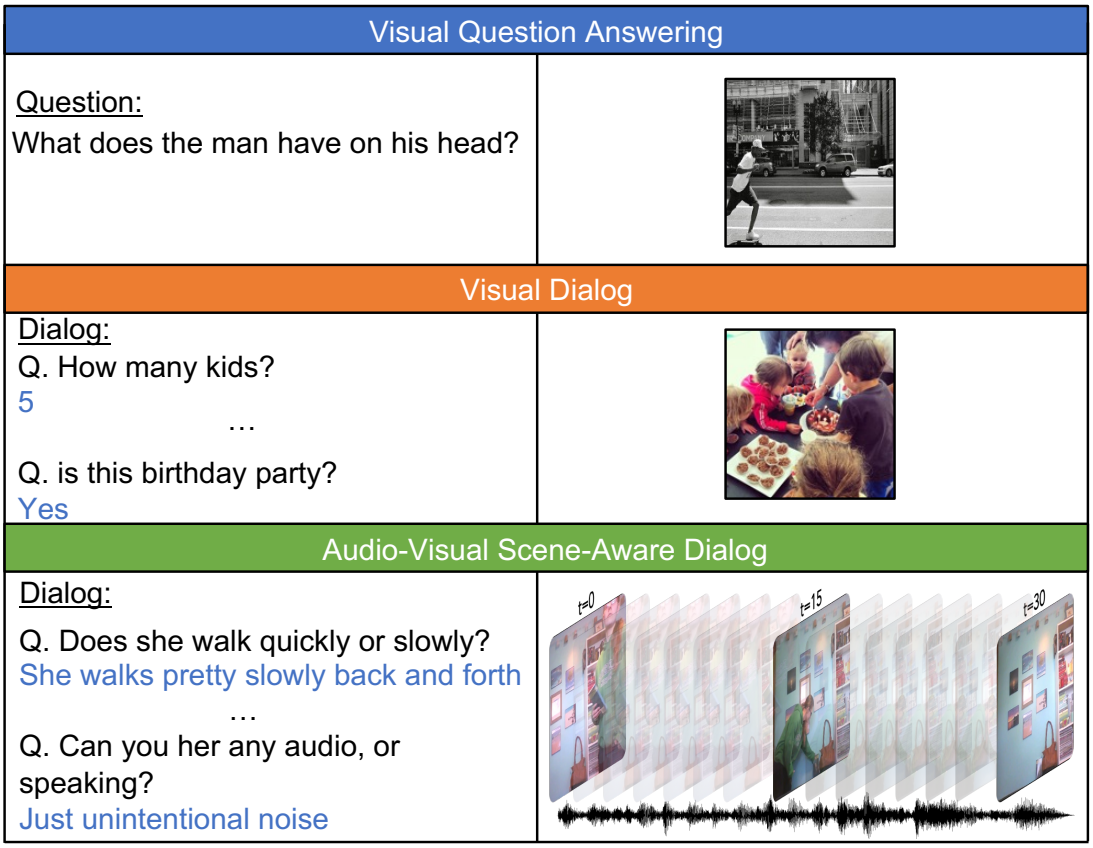


Figure 1.1: As cognitive tasks evolved, datasets included more modalities. There are two challenges: the ability to attend the important things and perceive all the input data. Note, the third-row figure illustrates a sound-enabled video and taken from [ACD+18].

To this end, we propose a generally applicable form of attention mechanism that learns high-order correlations between various data modalities. For example, second-order correlations can model interactions between two data modalities, *e.g.*, an image and a question, and more generally, $k-$th order correlations can model interactions between k modalities. Learning these correlations effectively directs the appropriate attention to the relevant elements in the different data modalities that are required to solve the joint task.

Perception, or the ability to perceive all of the modalities, is another aspect of cognition. This ability can be viewed as complementary to attention. Attention aims to compress the representation by selecting only relevant information, whereas perceptive models rely on all available data (*i.e.*, all modalities). When training deep net classifiers on those multi-modal datasets, the modalities get exploited at different scales, *i.e.*, some modalities can more easily contribute to the classification results than others (see Fig. 1.2). This is suboptimal because the classifier is inherently biased towards a subset of the modalities. To address this, we introduce the perceptual score to identify these biases by indicating whether a model relies on specific parts of the data. Our perceptual

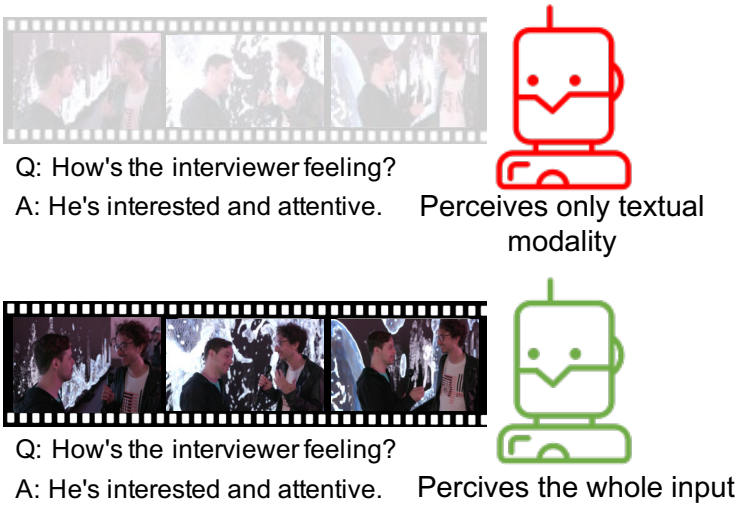


Figure 1.2: Multi-modal datasets often have an undesired bias: a classifier exploits shortcuts and predicts the correct answer based on parts of the data. For instance, in SocialIQ the task requires to understand social situations shown in video data. However, a classifier can achieve high accuracy by only perceiving textual cues because the answer correctness with subtle question cues.

score is further used to produce weights, which effectively reduce the chance of a deep network exploiting these biases. Further, we propose a novel regularization term based on the functional entropy. Intuitively, this term encourages to balance the contribution of each modality to the classification result.

1.2 Related Work

Multimodal Problems: In recent years various machine learning techniques were developed to tackle cognitive-like multimodal tasks, which involve both vision and language processing. Image captioning [MXY+15, KFF15, WSL17, CS18, ADS18] was an instrumental language+vision task, followed by visual question answering [LYBP16, KJZ18, NLS18, FPY+16]. Instrumental to cognitive tasks are attention models, that enable interpretation of the machine's cognition and often improve performance. Attention has been a prominent tool as it models interactions to select the important elements. In early work, Xu *et al.* [XBK+15] used interaction-based attention with the image at each caption generation step. This idea was later extended to visual question answering [XS16]. To imitate multi-step reasoning, Yang *et al.* [YHG+16] stacked attention modules sequentially. Later, many works concentrated on better vector-fusion modeling [FPY+16, KOL+17, BYCCT17, YYX+18]. Importantly, Lu *et al.* [LYBP16] suggested attending to the visual and textual modalities separately. Afterward, Kim *et al.* [KJZ18] proposed a bilinear module that efficiently generates attention for every pair. Following Lu *et al.* [LYBP16]. We improve upon those ideas by suggested a general framework that extends attention to any number of utilities via local and

interaction-based factors.

Attention in General: Atention models have been applied to graphical data structures. For example, Graph Attention Networks use an MRF approach to embed graph-structured data, *e.g.*, protein-protein interactions [VCC+18]. Also, attention for non-structured tasks (*e.g.*, chain, tree) were discussed in the past [KDHR17]. These works differ from ours in important aspects: they are used to embed a structure based model, *e.g.*, a graph, and provide a probability distribution across nodes of the graph. Instead, our model provides attention for entities within each node of the graph, *e.g.*, the words of a question or the pixels in an image.

Multimodal Perception: Recently, datasets were proposed to study whether a model can generalize and address the task or whether it uses a single modalities' features. Usually, this evaluation is performed by partitioning data into train and test sets using different distributions. For example, VQA-CP [ABPK18] is a reshuffle of the VQA [GKS+17] dataset ensuring that question-type distributions differ between train and test splits. Another well-known dataset is Colored MNIST [KKK+18]. In this dataset, each digit class is colored differently in the train set, while samples in the test set remain gray-scale. Different approaches were proposed to deal with such problems: Arjovsky *et al.* [ABGLP19] propose to improve generalization by ensuring that the optimal classifier equals all training distributions. Methods like REPAIR [LV19] prevent a model from exploiting dataset biases by re-sampling the training data. Kim *et al.* [KKK+18] use an adversarial approach to learn unbiased feature representations. Clark *et al.* [CYZ19] and Cadene *et al.* [CDC+19] suggest methods to overcome language priors using a bias-only model in VQA tasks.

1.3 Contributions and Outline

In this dissertation, we begin by introducing our novel attention unit, the Factor Graph Attention (FGA), on the Visual Dialog (VD) task. We subsequently describe the different cognitive tasks we address by employing special variants of the FGA module. We then conclude by discussing perceptiveness in multimodal datasets, and describe different techniques to measure and increase perceptiveness.

In **Chapter 2**, we provide the necessary deep learning background for text and image encoding methods.

In **Chapter 3**, we propose a novel and generally applicable form of attention mechanism that learns high-order correlations between various data modalities. We show that high-order correlations effectively direct the appropriate attention to the relevant elements in the different data modalities that are required to solve the joint task. We demonstrate the effectiveness of our high-order attention mechanism on the task of visual question answering (VQA), where we achieve state-of-the-art performance on the standard VQA dataset. Source code is available at `https://github.com/idansc/HighOrderAtten`.

In **Chapter 4**, we specify our solution for the VD task. Dialog is an effective way to exchange information, but subtle details and nuances are extremely important. While significant progress has paved a path to address visual dialog with algorithms, details and nuances remain a challenge. Attention mechanisms have demonstrated compelling results to extract details in visual question answering and also provide a convincing framework for visual dialog due to their interpretability and effectiveness. However, the many data utilities that accompany visual dialog challenge existing attention techniques. We address this issue and develop a general attention mechanism for visual dialog which operates on any number of data utilities. To this end, we design a factor graph based attention mechanism which combines any number of utility representations. We illustrate the applicability of the proposed approach on the challenging and recently introduced VisDial datasets, outperforming recent state-of-the-art methods by 1.1% for VisDial0.9 and by 2% for VisDial1.0 on MRR. Our ensemble model improved the MRR score on VisDial1.0 by more than 6%. Source code is available at `https://github.com/idansc/fga`.

Despite the significant boost in performance, we note that Assessing an AI agent that can converse in human language and understand visual content is challenging. Generation metrics, such as BLEU scores favor correct syntax over semantics. Hence a discriminative approach is often used, where an agent ranks a set of candidate options. The mean reciprocal rank (MRR) metric evaluates the model performance by taking into account the rank of a single human-derived answer. This approach, however, raises a new challenge: the ambiguity and synonymy of answers, for instance, semantic equivalence (*e.g.*, 'yeah' and 'yes'). To address this, the normalized discounted cumulative gain (NDCG) metric has been used to capture the relevance of all the correct answers via dense annotations. However, the NDCG metric favors the usually applicable uncertain answers such as 'I don't know.' Crafting a model that excels on both MRR and NDCG metrics is challenging [MBPD20]. Ideally, an AI agent should answer a human-like reply and validate the correctness of any answer. To address this issue, we describe a two-step non-parametric ranking approach that can merge strong MRR and NDCG models. Using our approach, we manage to keep most MRR state-of-the-art performance (70.41% *vs.* 71.24%) and the NDCG state-of-the-art performance (72.16% *vs.* 75.35%). Moreover, our approach won the recent Visual Dialog 2020 challenge. Source code is available at `https://github.com/idansc/mrr-ndcg`.

In **Chapter 5**, we address the problem of visual storytelling, *i.e.*, generating a story for a given sequence of images. Such a multi-modal problem requires to combine both visual and linguistic components. Different from image captioning, a coherent story needs to be consistent and relate to both future and past images. For this, we develop ordered image attention (OIA), a visual attention which combines information from an ordered set of images and highlights important regions based on order-aware interactions between objects across the sequence. The contextualized attention vectors of all images are used in the language generation module to generate the sentences of

the story. To alleviate common linguistic mistakes like repetitiveness and in-coherence, the decoder generates a novel sentence while considering the story up until the current sentence. We present results on the VIST dataset that improve upon the state-of-the-art both quantitatively and qualitatively. The work has been collaboratively created with Tom Braude and Arik Shamir.

In **Chapter 6**, we address the recently proposed audio-visual scene-aware dialog task that paves the way to a more data-driven way of learning virtual assistants, smart speakers and car navigation systems. Very little is known to date about how to effectively we describe how to effectively extract meaningful information from a plethora of sensors that pound the computational engine of those devices. The recently proposed audio-visual scene-aware dialog task paves the way to a more data-driven way of learning virtual assistants, smart speakers and car navigation systems. Therefore, in this chapter, we provide and carefully analyze a simple baseline for audio-visual scene-aware dialog which is trained end-to-end. Our method differentiates in a data-driven manner useful signals from distracting ones using an attention mechanism. We evaluate the proposed approach on the recently introduced and challenging audio-visual scene-aware dataset, and demonstrate the key features that permit to outperform the current state-of-the-art by more than 20% on CIDEr. Source code is available at `https://github.com/idansc/simple-avsd`.

In **Chapter 7**, we discuses method to measure the perceptiveness of different modalities. Machine learning advances in the last decade have relied significantly on large-scale datasets that continue to grow in size. Increasingly, those datasets also contain different data modalities. However, large multi-modal datasets are hard to annotate, and annotations may contain biases that we are often unaware of. Deep-net-based classifiers, in turn, are prone to exploit those biases and to find shortcuts. To study and quantify this concern, we introduce the perceptual score, a metric that assesses the degree to which a model relies on the different subsets of the input features, *i.e.*, modalities. Using the perceptual score, we find a surprisingly consistent trend across four popular datasets: recent, more accurate state-of-the-art multi-modal models for visual question-answering or visual dialog tend to perceive the visual data less than their predecessors. This trend is concerning as answers are hence increasingly inferred from textual cues only. Using the perceptual score also helps to analyze model biases by decomposing the score into data subset contributions. We hope to spur a discussion on the perceptiveness of multi-modal models and also hope to encourage the community working on multi-modal classifiers to start quantifying perceptiveness via the proposed perceptual score.

In **Chapter 8**, we reduce multimodal bias with a novel regularization term based on the functional entropy. Intuitively, this term encourages to balance the contribution of each modality to the classification result. However, regularization with the functional entropy is challenging. To address this, we develop a method based on the log-Sobolev inequality, which bounds the functional entropy with the functional-Fisher-

information. Intuitively, this maximizes the amount of information that the modalities contribute. On the two challenging multi-modal datasets VQA-CPv2 and SocialIQ, we obtain state-of-the-art results while more uniformly exploiting the modalities. In addition, we demonstrate the efficacy of our method on Colored MNIST. Source code is available at `https://github.com/itaigat/removing-bias-in-multi-modal-classifiers`. The work has been collaboratively created with Itai Gat.

Chapter 2

Deep Learning Background

2.1 Deep Neural Networks

Deep Neural Networks(DNNs) are used to estimate or approximate functions. For example, consider an image classification. Given an image input $x \in \mathbb{R}^{2\times2\times1}$ we want to approximate a score function for each class (*e.g.*, cat, dog, ship). A single neuron can be seen as a linear operation in the following form:

$$s = Wx + b \tag{2.1}$$

Where $W \in \mathbb{R}^{4\times1}, b \in \mathbb{R}^1$ (Fig. 2.1) . It resembles a real neuron because it receives input from other units and computes its output. DNNs are called networks because they are organized in layers that are made up of many interconnected neurons. Each neuron gets data from the previous layer, which in his turn forwards his outputs to the next layer. A 2 layer neural network, in which the outputs of the first layer are transferred to another layer can be formulated as

$$s = W_2(W_1x + b_1) + b_2 \tag{2.2}$$

Where $W_1 \in \mathbb{R}^{3\times k_1}, W_2 \in \mathbb{R}^{3\times k_2}, b_1 \in \mathbb{R}^{k_1}, b_2 \in \mathbb{R}^{k_2}$. The k_1, k_2 terms indicate the number of neurons in each layer. The model is associated with a directed acyclic graph describing how the functions are composed together. The number of layers defines the depth of the model. In the above form, the two matrices could be collapsed into a single matrix. Therefore the predicted class scores would again be a linear function of the input, therefore to allow non-linearity, we add a non-linear function σ called activation function. for instance $\sigma = \tanh$. The two-layer network can be formulated as

$$s = W_2 \tanh(W_1x + b_1) + b_2 \tag{2.3}$$

The term 'learning' refers to minimize the scoring error by changing the weights W and the bias term b. The error is typically achieved with respect to a loss function.

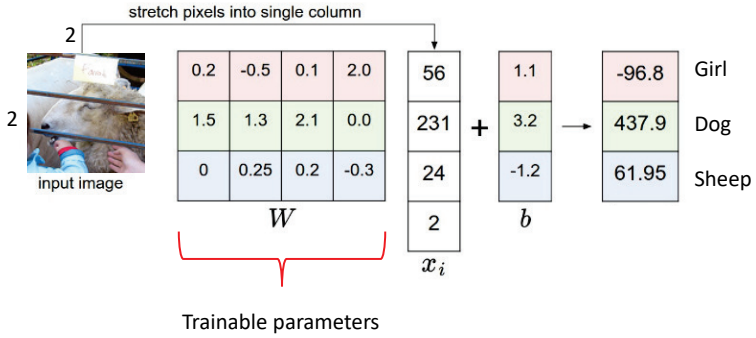


Figure 2.1: Linear architecture for classification. The model has 1-layer with three neurons. Each neuron outputs a score. The correct answer is sheep; therefore, the score for the ship class should be the highest. The error is reduced by training the weights using the backpropagation algorithm.

For instance, the cross-entropy loss function interprets the score as Gibbs probability over the classes using softmax$(\cdot)$ operator, and indicates the difference between the interpolated probability to ground truth probability. Thus, we can obtain the output we want for specific inputs. The process of adjusting the weights in order to obtain desired output is called learning or training. This process is done by gradient descent and is known as the backpropagation algorithm.

2.2 Convolutional Neural Networks

The network described so far are often called fully-connected networks. These networks do not scale to high-dimensional input since each pixel is connected to all neurons. If we consider an image with size of $200 \times 200 \times 3$ it would lead to $120,000$ parameters for a single neuron. Many parameters usually lead to overfitting problems, where the model excels on the training data but fails to generalize.

However, images are 2D spatial input, built with hierarchical patterns. Further, the patterns have translation invariance characteristics, *i.e.*, the pattern's spatial location does not change the pattern meaning (See Fig. 2.2). A convolutional layer arranges its neurons in three dimensions (width, height, depth). The convolution neurons act as filters (also known as kernels), convolve across the current volume, and compute dot products between the entries of the filter and the input at any position. The convolutional layer transforms the 3D input volume to a 3D output volume of neuron activations. Convolutional layers are locally connected to small subsets of neurons in the previous layer. Therefore, on the scale of connectedness and complexity, CNNs are on the lower extremity.

Convolutional networks were motivated by biological processes in that the connectivity pattern between neurons resembles the animal visual cortex's organization (See Fig. 2.4) [SWB$^+$07]. Individual cortical neurons respond to stimuli only in a restricted

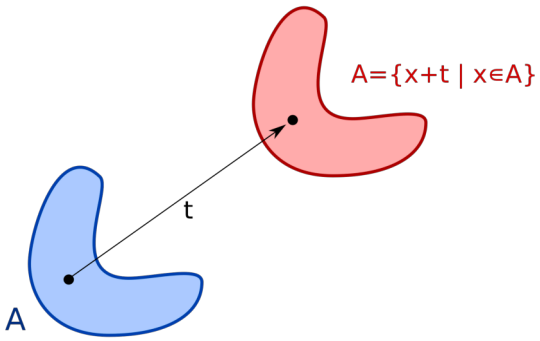


Figure 2.2: Translational invariant functions $f : \mathbb{R}^2 \to \mathbb{R}$ it is $f(A) = f(A + t)$. Image patterns have translational invariance characteristics. Source: Stephan Kulla.

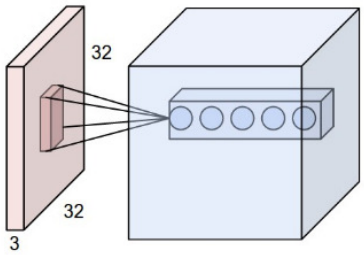


Figure 2.3: A filter (kernel) convolved over an image of $32 \times 32 \times 3$ dimension. The output depth is dependent on numbers of filters. In this case the output depth is 5, which means the convolutional layer have 5 filters. Source: cs231n.

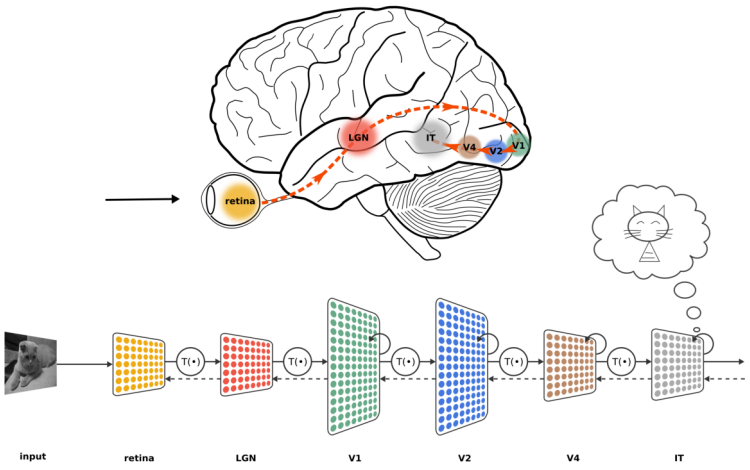


Figure 2.4: Convolutional networks were motivated by biological processes in that the connectivity pattern between neurons resembles the animal visual cortex's organization. Source: Kubilius, Jonas (`https://doi.org/10.6084/m9.figshare.106794.v3`)

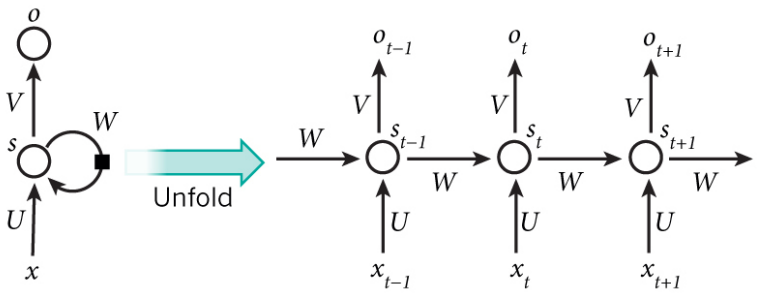


Figure 2.5: Illustration of an RNN layer. RNN is build feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video). Source: Nature.

region of the visual field known as the receptive field. The receptive fields of different neurons partially overlap such that they cover the entire visual field.

2.3 Recurrent Neural Network

The idea behind Recurrent neural networks (RNNs) is to make use of sequential information. If one wants to predict the next word in a sentence, it is better to know which words came before. Similar to CNNs, RNNs are a group of networks that process input with a spatial dimension. Specifically, a sentence is a sequence of words that can be seen as input with 1D spatial dimension (*i.e.*, the length of the sentence). RNNs are called recurrent because they perform the same task for every element of a sequence, with the output being dependent on the previous computations. Another way to think about RNNs is that they have a "memory" which captures information about what has been calculated so far. For instance, a two-layer linear classifier unfolded to work on sentencing length (Fig. 2.5)

2.3.1 Long Short Term Memory

During the backpropagation phase of RNN, the gradient signal is being multiplied a large number of times (proportional to the number of timestamps) by the linear weight matrix. If the weights in this matrix are small (leading eigenvalue of the weight matrix is smaller than 1.0), it can lead to vanishing gradients where the gradient signal gets so small that learning either becomes very slow or stops working altogether [HS97b]. This makes the task of learning long-term dependencies in the data difficult. Conversely, if the weights in this matrix are large, it can lead to a situation where the gradient signal is so large that it can cause learning to diverge. This is often referred to as exploding gradients.

LSTM model introduces a new structure called a memory cell (see Fig. 2.6). A memory cell is composed of four main elements: an input gate, a neuron with a self-recurrent connection (a connection to itself), a forget gate, and an output gate. The

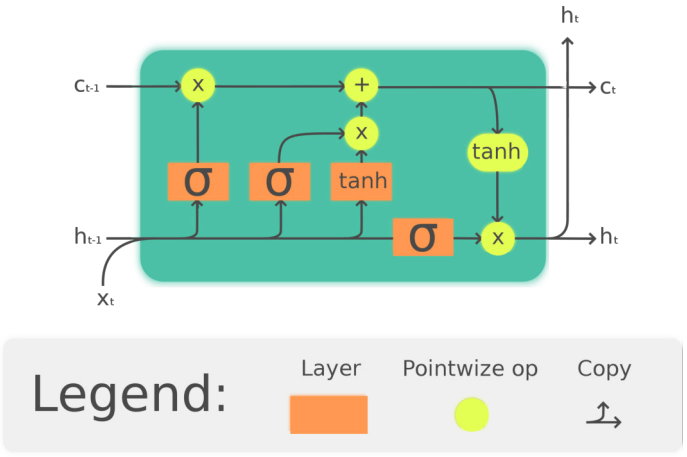


Figure 2.6: Memory cell introduces gates to avoid vanishing/exploding gradient problems by learning how to update the hidden state. Diagram is taken from [Che18].

self-recurrent connection has a weight of 1.0 and ensures that, barring any outside interference, the state of a memory cell can remain constant from one time-step to another. The gates serve to modulate the interactions between the memory cell itself and its environment. The input gate can allow the incoming signal to alter the memory cell's state or block it. On the other hand, the output gate can allow the state of the memory cell to affect other neurons or prevent it. Finally, the forget gate can modulate the memory cell's self-recurrent connection, allowing the cell to remember or forget its previous state, as needed.

Chapter 3

Visual Question Answering

The quest for algorithms which enable cognitive abilities is an important part of machine learning and appears in many facets, *e.g.*, in visual question answering tasks [DAZ+16], image captioning [XBK+15], visual question generation [MMD+16, JZS17a] and machine comprehension [HKG+15]. A common trait in these recent cognitive-like tasks is that they take into account different data modalities, for example, visual and textual data.

To address these tasks, recently, attention mechanisms have emerged as a powerful common theme, which provides not only some form of interpretability if applied to deep net models, but also often improves performance [HKG+15]. The latter effect is attributed to more expressive yet concise forms of the various data modalities. Present day attention mechanisms, like for example [LYBP16, XBK+15], are however often lacking in two main aspects. First, the systems generally extract abstract representations of data in an ad-hoc and entangled manner. Second, present day attention mechanisms are often geared towards a specific form of input and therefore hand-crafted for a particular task.

To address both issues, we propose a novel and generally applicable form of attention mechanism that learns high-order correlations between various data modalities. For example, second order correlations can model interactions between two data modalities, *e.g.*, an image and a question, and more generally, $k-$th order correlations can model interactions between k modalities. Learning these correlations effectively directs the appropriate attention to the relevant elements in the different data modalities that are required to solve the joint task.

We demonstrate the effectiveness of our novel attention mechanism on the task of visual question answering (VQA), where we achieve state-of-the-art performance on the VQA dataset [AAL+15]. Some of our results are visualized in Fig. 3.1, where we show how the visual attention correlates with the textual attention.

We begin by reviewing the related work. We subsequently provide details of our proposed technique, focusing on the high-order nature of our attention models. We then conclude by presenting the application of our high-order attention mechanism to

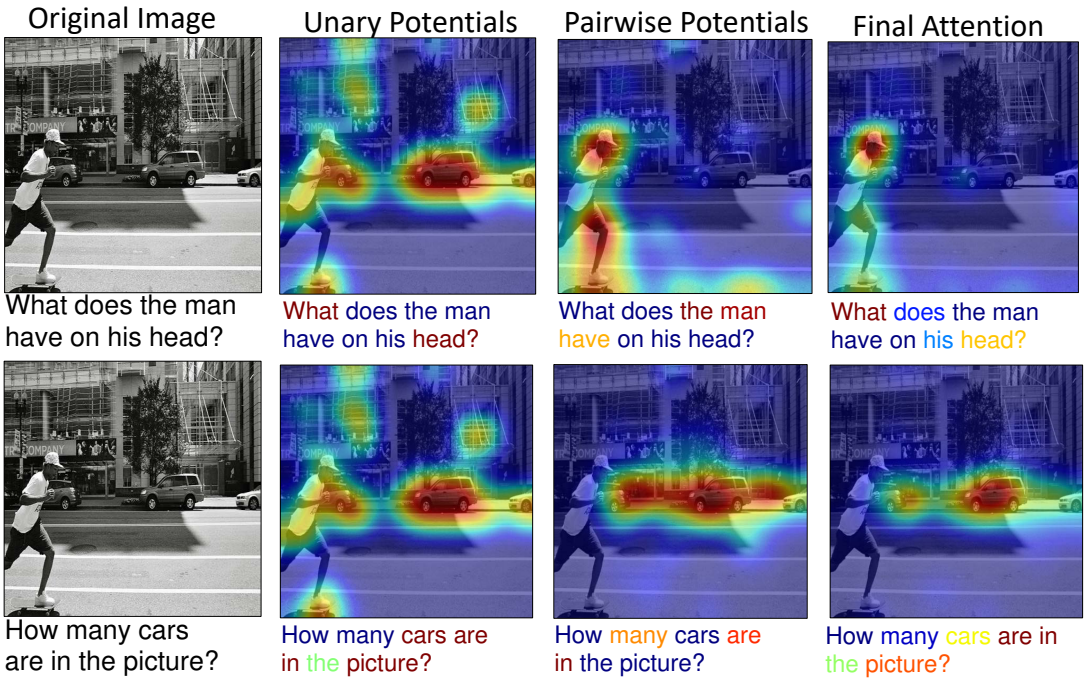


Figure 3.1: Results of our multi-modal attention for one image and two different questions (1st column). The unary image attention is identical by construction. The pairwise potentials differ for both questions and images since both modalities are taken into account (3rd column). The final attention is illustrated in the 4th column.

VQA and compare it to the state-of-the-art.

3.1 Related Work

Attention mechanisms have been investigated for both image and textual data. In the following we review mechanisms for both.

Image attention mechanisms: Over the past few years, single image embeddings extracted from a deep net (*e.g.*, [MRF15, MLL16]) have been extended to a variety of image attention modules, when considering VQA. For example, a textual long short term memory net (LSTM) may be augmented with a spatial attention [ZGBFF16]. Similarly, Andreas *et al.* [ARDK16] employ a language parser together with a series of neural net modules, one of which attends to regions in an image. The language parser suggests which neural net module to use. Stacking of attention units was also investigated by Yang *et al.* [YHG+16]. Their stacked attention network predicts the answer successively. Dynamic memory network modules which capture contextual information from neighboring image regions has been considered by Xiong *et al.* [XMS16]. Shih *et al.* [SSH16] use object proposals and and rank regions according to relevance. The multi-hop attention scheme of Xu *et al.* [XS16] was proposed to extract fine-grained details. A joint attention mechanism was discussed by Lu *et al.* [LYBP16] and Fukui *et al.* [FPY+16] suggest an efficient outer product mechanism to combine visual representation and text representation before applying attention over the combined representation. Additionally, they suggested the use of glimpses. Very recently, Kazemi

et al. [KE17] showed a similar approach using concatenation instead of outer product. Importantly, all of these approaches model attention as a single network. The fact that multiple modalities are involved is often not considered explicitly which contrasts the aforementioned approaches from the technique we present.

Very recently Kim *et al.* [KDHR17] presented a technique that also interprets attention as a multi-variate probabilistic model, to incorporate structural dependencies into the deep net. Other recent techniques are work by Nam *et al.* [NHK17] on dual attention mechanisms and work by Kim *et al.* [KOL$^+$17] on bilinear models. In contrast to the latter two models our approach is easy to extend to any number of data modalities.

Textual attention mechanisms: We also want to provide a brief review of textual attention. To address some of the challenges, *e.g.*, long sentences, faced by translation models, Hermann *et al.* [HKG$^+$15] proposed RNNSearch. To address the challenges which arise by fixing the latent dimension of neural nets processing text data, Bahdanau *et al.* [BCB14] first encode a document and a query via a bidirectional LSTM which are then used to compute attentions. This mechanism was later refined in [RGH$^+$16] where a word based technique reasons about sentence representations. Joint attention between two CNN hierarchies is discussed by Yin *et al.* [YSXZ16].

Among all those attention mechanisms, relevant to our approach is work by Lu *et al.* [LYBP16] and the approach presented by Xu *et al.* [XS16]. Both discuss attention mechanisms which operate jointly over two modalities. Xu *et al.* [XS16] use pairwise interactions in the form of a similarity matrix, but ignore the attentions on individual data modalities. Lu *et al.* [LYBP16] suggest an alternating model, that directly combines the features of the modalities before attending. Additionally, they suggested a parallel model which uses a similarity matrix to map features for one modality to the other. It is hard to extend this approach to more than two modalities. In contrast, our model develops a probabilistic model, based on high order potentials and performs mean-field inference to obtain marginal probabilities. This permits trivial extension of the model to any number of modalities.

Additionally, Jabri *et al.* [JJvdM16] propose a model where answers are also used as inputs. Their approach questions the need of attention mechanisms and develops an alternative solution based on binary classification. In contrast, our approach captures high-order attention correlations, which we found to improve performance significantly.

Overall, while there is early work that propose a combination of language and image attention for VQA, *e.g.*, [LYBP16, XS16], attention mechanism with several potentials haven't been discussed in detail yet. In the following we present our approach for joint attention over any number of modalities.

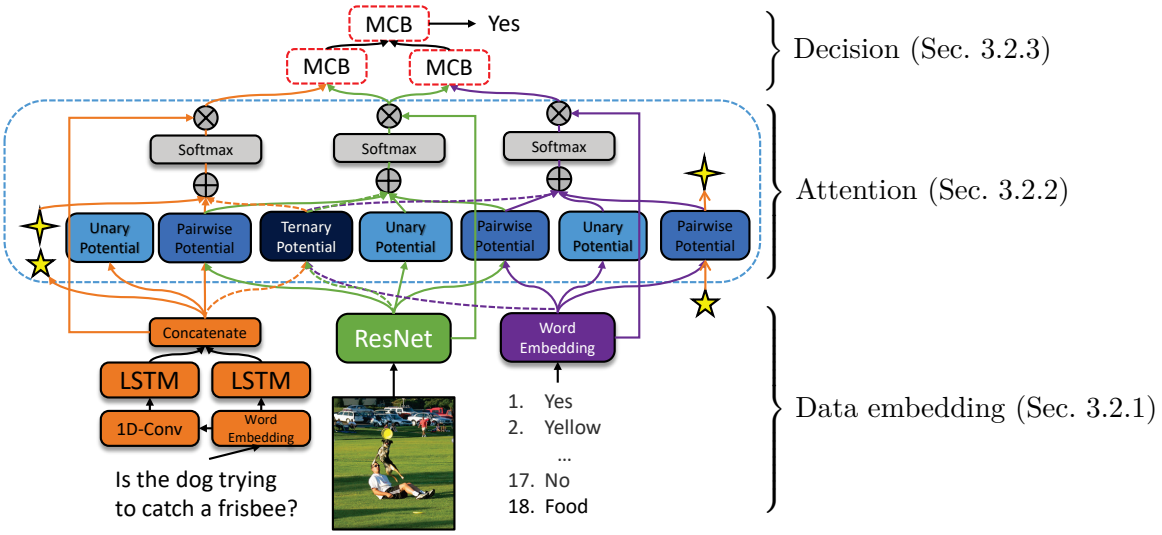


Figure 3.2: Our state-of-the-art VQA system

3.2 Higher order attention models

Attention modules are a crucial component for present day decision making systems. Particularly when taking into account more and more data of different modalities, attention mechanisms are able to provide insights into the inner workings of the oftentimes abstract and automatically extracted representations of our systems.

An example of such a system that captured a lot of research efforts in recent years is Visual Question Answering (VQA). Considering VQA as an example, we immediately note its dependence on two or even three different data modalities, the visual input V, the question Q and the answer A, which get processed simultaneously. More formally, we let

$$V \in \mathbb{R}^{n_v \times d}, \quad Q \in \mathbb{R}^{n_q \times d}, \quad A \in \mathbb{R}^{n_a \times d}$$

denote a representation for the visual input, the question and the answer respectively. Hereby, n_v, n_q and n_a are the number of pixels, the number of words in the question, and the number of possible answers. We use d to denote the dimensionality of the data. For simplicity of the exposition we assume d to be identical across all data modalities.

Due to this dependence on multiple data modalities, present day decision making systems can be decomposed into three major parts: (i) the data embedding; (ii) attention mechanisms; and (iii) the decision making. For a state-of-the-art VQA system such as the one we developed here, those three parts are immediately apparent when considering the high-level system architecture outlined in Fig. 3.2.

3.2.1 Data Embedding

Attention modules deliver to the decision making component a succinct representation of the relevant data modalities. As such, their performance depends on how we rep-

resent the data modalities themselves. Oftentimes, an attention module tends to use expressive yet concise data embedding algorithms to better capture their correlations and consequently to improve the decision making performance. For example, data embeddings based on convolutional deep nets which constitute the state-of-the-art in many visual recognition and scene understanding tasks. Language embeddings heavily rely on LSTM which are able to capture context in sequential data, such as words, phrases and sentences. We give a detailed account to our data embedding architectures for VQA in Sec. 3.3.1.

3.2.2 Attention

As apparent from the aforementioned description, attention is the crucial component connecting data embeddings with decision making modules.

Subsequently we denote attention over the n_q words in the question via $P_Q(i_q)$, where $i_q \in \{1, \ldots, n_q\}$ is the word index. Similarly, attention over the image is referred to via $P_V(i_v)$, where $i_v \in \{1, \ldots, n_v\}$, and attention over the possible answers are denoted $P_A(i_a)$, where $i_a \in \{1, \ldots, n_a\}$.

We consider the attention mechanism as a probability model, with each attention mechanism computing "potentials." First, unary potentials θ_V, θ_Q, θ_A denote the importance of each feature (*e.g.*, question word representations, multiple choice answers representations, and image patch features) for the VQA task. Second, pairwise potentials, $\theta_{V,Q}, \theta_{V,A}, \theta_{Q,A}$ express correlations between two modalities. Last, third-order potential, $\theta_{V,Q,A}$ captures dependencies between the three modalities.

To obtain marginal probabilities P_Q, P_V and P_A from potentials, our model performs mean-field inference. We combine the unary potential, the marginalized pairwise potential and the marginalized third order potential linearly including a bias term:

$$\begin{aligned} P_V(i_v) &= \text{smax}(\alpha_1\theta_V(i_v)+\alpha_2\theta_{V,Q}(i_v)+\alpha_3\theta_{A,V}(i_v)+\alpha_4\theta_{V,Q,A}(i_v)+\alpha_5), \\ P_Q(i_q) &= \text{smax}(\beta_1\theta_Q(i_q)+\beta_2\theta_{V,Q}(i_q)+\beta_3\theta_{A,Q}(i_q)+\beta_4\theta_{V,Q,A}(i_q)+\beta_5), \\ P_A(i_a) &= \text{smax}(\gamma_1\theta_A(i_a)+\gamma_2\theta_{A,V}(i_a)+\gamma_3\theta_{A,Q}(i_a)+\gamma_4\theta_{V,Q.A}(i_a)+\gamma_5). \end{aligned} \tag{3.1}$$

Hereby α_i, β_i, and γ_i are learnable parameters and $\text{smax}(\cdot)$ refers to the soft-max operation over $i_v \in \{1, \ldots, n_v\}$, $i_q \in \{1, \ldots, n_q\}$ and $i_a \in \{1, \ldots, n_a\}$ respectively. The soft-max converts the combined potentials to probability distributions, which corresponds to a single mean-field iteration. Such a linear combination of potentials provides extra flexibility for the model, since it can learn the reliability of the potential from the data. For instance, we observe that question attention relies more on the unary question potential and on pairwise question and answer potentials. In contrast, the image attention relies more on the pairwise question and image potential.

Given the aforementioned probabilities P_V, P_Q, and P_A, the attended image, question and answer vectors are denoted by $a_V \in \mathbb{R}^d$, $a_Q \in \mathbb{R}^d$ and $a_A \in \mathbb{R}^d$. The attended

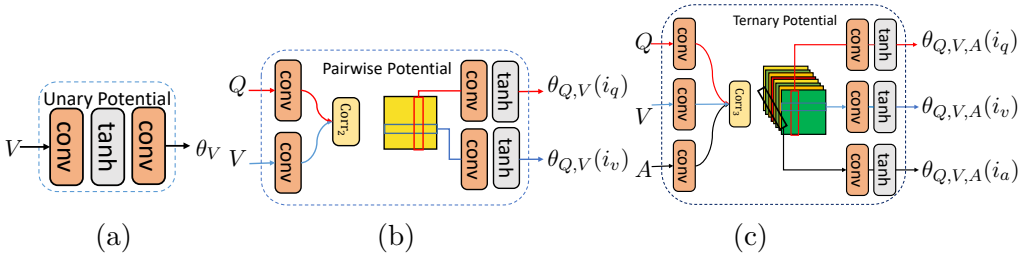


Figure 3.3: Illustration of our $k-$order attention. (a) unary attention module (e.g., visual). (b) pairwise attention module (e.g., visual and question) marginalized over its two data modalities. (c) ternary attention module (e.g., visual, question and answer) marginalized over its three data modalities.

modalities are calculated as the weighted sum of the image features $V = [v_1, \ldots, v_{n_v}]^T \in \mathbb{R}^{n_v \times d}$, the question features $Q = [q_1, \ldots, q_{n_q}]^T \in \mathbb{R}^{n_q \times d}$, and the answer features $A = [a_1, \ldots, a_{n_a}]^T \in \mathbb{R}^{n_a \times d}$, *i.e.*,

$$a_V = \sum_{i_v=1}^{n_v} P_V(i_v) v_{i_v}, \quad a_Q = \sum_{i_q=1}^{n_q} P_Q(i_q) q_{i_q}, \quad \text{and} \quad a_V = \sum_{i_a=1}^{n_a} P_A(i_a) a_{i_a}.$$

The attended modalities, which effectively focus on the data relevant for the task, are passed to a classifier for decision making, *e.g.*, the ones discussed in Sec. 3.2.3. In the following we now describe the attention mechanisms for unary, pairwise and ternary potentials in more detail.

Unary potentials

We illustrate the unary attention schematically in Fig. 3.3 (a). The input to the unary attention module is a data representation, *i.e.*, either the visual representation V, the question representation Q, or the answer representation A. Using those representations, we obtain the 'unary potentials' θ_V, θ_Q and θ_A using a convolution operation with kernel size 1×1 over the data representation as an additional embedding step, followed by a non-linearity (tanh in our case), followed by another convolution operation with kernel size 1×1 to reduce embedding dimensionality. Since convolutions with kernel size 1×1 are identical to matrix multiplies we formally obtain the unary potentials via

$$\theta_V(i_v) = \tanh(VW_{v_2})W_{v_1}, \quad \theta_Q(i_q) = \tanh(QW_{q_2})W_{q_1}, \quad \theta_A(i_a) = \tanh(AW_{a_2})W_{a_1}.$$

where $W_{v_1}, W_{q_1}, W_{a_1} \in \mathbb{R}^{d \times 1}$, and $W_{v_2}, W_{q_2}, W_{a_2} \in \mathbb{R}^{d \times d}$ are trainable parameters.

Pairwise potentials

Besides the mentioned mechanisms to generate unary potentials, we specifically aim at taking advantage of pairwise attention modules, which are able to capture the correlation between the representation of different modalities. Our approach is illustrated in Fig. 3.3 (b). We use a similarity matrix between image and question modalities $C_2 = QW_q(VW_v)^\top$. Alternatively, the (i, j)-th entry is the correlation (inner-product)

of the i-th column of QW_q and the j-th column of VW_v:

$$(C_2)_{i,j} = \text{corr}_2((QW_q)_{:,i}, (VW_v)_{:,j}), \qquad \text{corr}_2(q, v) = \sum_{l=1}^{d} q_l v_l.$$

where $W_q, W_v \in \mathbb{R}^{d \times d}$ are trainable parameters. We consider $(C_2)_{i,j}$ as a pairwise potential that represents the correlation of the i-th word in a question and the j-th patch in an image. Therefore, to retrieve the attention for a specific word, we convolve the matrix along the visual dimension using a 1×1 dimensional kernel. Specifically,

$$\theta_{V,Q}(i_q) = \tanh\left(\sum_{i_v=1}^{n_v} w_{i_v}(C_2)_{i_v,i_q}\right), \quad \text{and} \quad \theta_{V,Q}(i_v) = \tanh\left(\sum_{i_q=1}^{n_q} w_{i_q}(C_2)_{i_v,i_q}\right).$$

Similarly, we obtain $\theta_{A,V}$ and $\theta_{A,Q}$, which we omit due to space limitations. These potentials are used to compute the attention probabilities as defined in Eq. (3.1).

Ternary Potentials

To capture the dependencies between all three modalities, we consider their high-order correlations.

$$(C_3)_{i,j,k} = \text{corr}_3((QW_q)_{:,i}, (VW_v)_{:,j}, (AW_a)_{:,k}), \qquad \text{corr}_3(q, v, a) = \sum_{l=1}^{d} q_l v_l a_l.$$

Where $W_q, W_v, W_a \in \mathbb{R}^{d \times d}$ are trainable parameters. Similarly to the pairwise potentials, we use the C_3 tensor to obtain correlated attention for each modality:

$$\theta_{V,Q,A}(i_q) = \tanh\left(\sum_{i_v=1}^{n_v}\sum_{i_a=1}^{n_a} w_{i_v,i_a}(C_3)_{i_q,i_v,i_a}\right), \; \theta_{V,Q,A}(i_v) = \tanh\left(\sum_{i_q=1}^{n_q}\sum_{i_a=1}^{n_a} w_{i_q,i_a}(C_3)_{i_q,i_v,i_a}\right),$$

$$\text{and} \quad \theta_{V,Q,A}(i_a) = \tanh\left(\sum_{i_v=1}^{n_v}\sum_{i_q=1}^{n_q} w_{i_q,i_a}(C_3)_{i_q,i_v,i_a}\right).$$

These potentials are used to compute the attention probabilities as defined in Eq. (3.1).

3.2.3 Decision Making

The decision making component receives as input the attended modalities and predicts the desired output. Each attended modality is a vector that consists of the relevant data for making the decision. While the decision making component can consider the modalities independently, the nature of the task usually requires to take into account correlations between the attended modalities. The correlation of a set of attended modalities are represented by the outer product of their respective vectors, *e.g.*, the

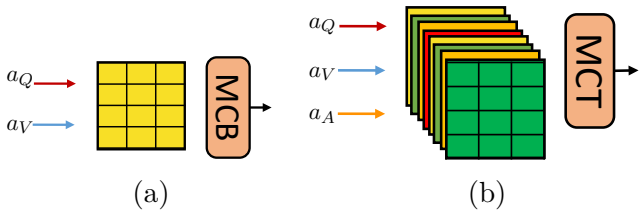


Figure 3.4: Illustration of correlation units used for decision making. (a) MCB unit approximately sample from outer product space of two attention vectors, (b) MCT unit approximately sample from outer product space of three attention vectors.

correlation of two attended modalities is represented by a matrix and the correlation of k-attended modalities is represented by a k-dimensional tensor.

Ideally, the attended modalities and their high-order correlation tensors are fed into a deep net which produces the final decision. The number of parameters in such a network grows exponentially in the number of modalities, as seen in Fig. 3.4. To overcome this computational bottleneck, we follow the tensor sketch algorithm of Pham and Pagh [PP13], which was recently applied to attention models by Fukui *et al.* [FPY+16] via Multimodal Compact Bilinear Pooling (MCB) in the pairwise setting or Multimodal Compact Trilinear Pooling (MCT), an extension of MCB that pools data from three modalities. The tensor sketch algorithm enables us to reduce the dimension of any rank-one tensor while referring to it implicitly. It relies on the count sketch technique [CCFC02] that randomly embeds an attended vector $a \in \mathbb{R}^{d_1}$ into another Euclidean space $\Psi(a) \in \mathbb{R}^{d_2}$. The tensor sketch algorithm then projects the rank-one tensor $\otimes_{i=1}^{k} a_i$ which consists of attention correlations of order k using the convolution $\Psi(\otimes_{i=1}^{k} a_i) = *_{i=1}^{k} \Psi(a_i)$. For example, for two attention modalities, the correlation matrix $a_1 a_2^\top = a_1 \otimes a_2$ is randomly projected to $\mathbb{R}^{d_2}$ by the convolution $\Psi(a_1 \otimes a_2) = \Psi(a_1) * \Psi(a_2)$. The attended modalities $\Psi(a_i)$ and their high-order correlations $\Psi(\otimes_{i=1}^{k} a_i)$ are fed into a fully connected neural net to complete decision making.

3.3 High-order Attention for Visual Question Answering

In the following we evaluate our approach qualitatively and quantitatively. Before doing so we describe the data embeddings.

3.3.1 Data Embedding

The attention module requires the question representation $Q \in \mathbb{R}^{n_q \times d}$, the image representation $V \in \mathbb{R}^{n_v \times d}$, and the answer representation $A \in \mathbb{R}^{n_a \times d}$, which are computed as follows.

Image Embedding: To embed the image, we use pre-trained convolutional deep nets (*i.e.*, VGG-19, ResNet). We extract the last layer before the fully connected units. Its

Table 3.1: Comparison of results on the Multiple-Choice VQA dataset for a variety of methods. We observe the combination of all three unary, pairwise and ternary potentials to yield the best result.

	test-dev				test-std
Method	Y/N	Num	Other	All	All
Naive Bayes [LYBP16]	79.7	40.1	57.9	64.9	-
HieCoAtt (ResNet) [LYBP16]	79.7	40.0	59.8	65.8	66.1
RAU (ResNet) [NH16]	81.9	41.1	61.5	67.7	67.3
MCB (ResNet) [FPY$^+$16]	-	-	-	68.6	-
DAN (VGG) [NHK17]	-	-	-	67.0	-
DAN (ResNet) [NHK17]	-	-	-	69.1	69.0
MLB (ResNet) [KOL$^+$17]	-	-	-	-	68.9
2-Modalities: Unary+Pairwis (ResNet)	80.9	36.0	61.6	66.7	-
3-Modalities: Unary+Pairwise (ResNet)	**82.0**	42.7	63.3	68.7	68.7
3-Modalities: Unary + Pairwise + Ternary (VGG)	81.2	42.7	62.3	67.9	-
3-Modalities: Unary + Pairwise + Ternary (ResNet)	81.6	**43.3**	**64.8**	**69.4**	**69.3**

dimension in the VGG net case is $512 \times 14 \times 14$ and the dimension in the ResNet case is $2048 \times 14 \times 14$. Hence we obtain $n_v = 196$ and we embed both the 196 VGG-19 or ResNet features into a $d = 512$ dimensional space to obtain the image representation V.

Question Embedding: To obtain a question representation, $Q \in \mathcal{R}^{n_q \times d}$, we first map a 1-hot encoding of each word in the question into a d-dimensional embedding space using a linear transformation plus corresponding bias terms. To obtain a richer representation that accounts for neighboring words, we use a 1-dimensional temporal convolution with filter of size 3. While a combination of multiple sized filters is suggested in the literature [LYBP16], we didn't find any benefit from using such an approach. Subsequently, to capture long-term dependencies, we used a Long Short Term Memory (LSTM) layer. To reduce overfitting caused by the LSTM units, we used two LSTM layers with $d/2$ hidden dimension, one uses as input the word embedding representation, and the other one operates on the 1D conv layer output. Their output is then concatenated to obtain Q. We also note that n_q is a constant hyperparameter, *i.e.*, questions with more than n_q words are cut, while questions with less words are zero-padded.

Answer Embedding: To embed the possible answers we use a regular word embedding. The vocabulary is specified by taking only the most frequent answers in the training set. Answers that are not included in the top answers are embedded to the same vector. Answers containing multiple words are embedded as n-grams to a single vector. We assume there is no real dependency between the answers, therefore there is no need of using additional 1D conv, or LSTM layers.

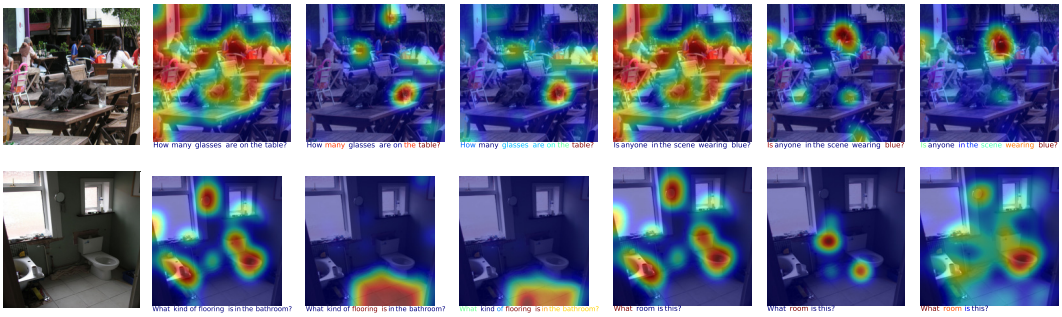


Figure 3.5: For each image (1st column) we show the attention generated for two different questions in columns 2-4 and columns 5-7 respectively. The attentions are ordered as unary attention, pairwise attention and combined attention for both the image and the question. We observe the combined attention to significantly depend on the question.

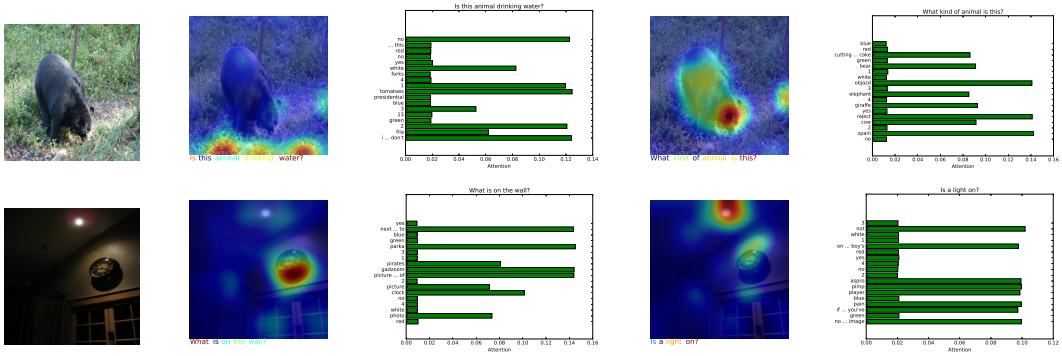


Figure 3.6: The attention generated for two different questions over three modalities. We find the attention over multiple choice answers to emphasis the unusual answers.

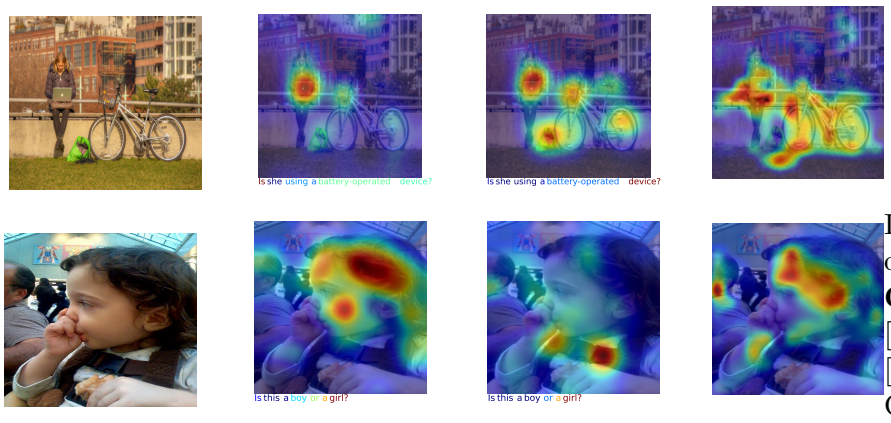
3.3.2 Decision Making

For our VQA example we investigate two techniques to combine vectors from three modalities. First, the attended feature representation for each modality, *i.e.*, a_V, a_A and a_Q, are combined using an MCT unit. Each feature element is of the form $((a_V)_i \cdot (a_Q)_j \cdot (a_A)_k)$. While this first solution is most general, in some cases like VQA, our experiments show that it is better to use our second approach, a 2-layer MCB unit combination. This permits greater expressiveness as we employ features of the form $((a_V)_i \cdot (a_Q)_j \cdot (a_Q)_k \cdot (a_A)_t)$ therefore also allowing image features to interact with themselves. Note that in terms of parameters both approaches are identical as neither MCB nor MCT are parametric modules.

Beyond MCB, we tested several other techniques that were suggested in the literature, including element-wise multiplication, element-wise addition and concatenation [KOL+17, LYBP16, KE17], optionally followed by another hidden fully connected layer. The tensor sketching units consistently performed best.

3.3.3 Results

Experimental Setup: We use the RMSProp optimizer with a base learning rate of $4e^{-4}$ and $\alpha = 0.99$ as well as $\epsilon = 1e^{-8}$. The batch size is set to 300. The dimension d of



Is she using a battery device?
Ours: yes
[LYBP16]: no
[FPY$^+$16]: no
GT: yes

Is this boy or a girl?
Ours: girl
[LYBP16]: boy
[FPY$^+$16]: girl
GT: girl

Figure 3.7: Comparison of our attention results (2$^{\text{nd}}$ column) with attention provided by [LYBP16] (3$^{\text{rd}}$ column) and [FPY$^+$16] (4$^{\text{th}}$ column). The fourth column provides the question and the answer of the different techniques.

all hidden layers is set to 512. The MCB unit feature dimension was set to $d = 8192$. We apply dropout with a rate of 0.5 after the word embeddings, the LSTM layer, and the first conv layer in the unary potential units. Additionally, for the last fully connected layer we use a dropout rate of 0.3. We use the top 3000 most frequent answers as possible outputs, which covers 91% of all answers in the train set. We implemented our models using the Torch framework[1] [CKF11].

As a comparison for our attention mechanism we use the approach of Lu *et al.* [LYBP16] and the technique of Fukui *et al.* [FPY$^+$16]. Their methods are based on a hierarchical attention mechanism and multi-modal compact bilinear (MCB) pooling. In contrast to their approach we demonstrate a relatively simple technique based on a probabilistic intuition grounded on potentials. For comparative reasons only, the visualized attention is based on two modalities: image and question.

We evaluate our attention modules on the VQA real-image test-dev and test-std datasets [AAL$^+$15]. The dataset consists of 123, 287 training images and 81, 434 test set images. Each image comes with 3 questions along with 18 multiple choice answers.

Quantitative Evaluation: We first evaluate the overall performance of our model and compare it to a variety of baselines. Tab. 3.1 shows the performance of our model and the baselines on the test-dev and the test-standard datasets for multiple choice (MC) questions. To obtain multiple choice results we follow common practice and use the highest scoring answer among the provided ones. Our approach (Fig. 3.2) for the multiple choice answering task achieved the reported result after 180,000 iterations, which requires about 40 hours of training on the 'train+val' dataset using a TitanX GPU. Despite the fact that our model has only 40 million parameters, while techniques like [FPY$^+$16] use over 70 million parameters, we observe state-of-the-art behavior. Additionally, we employ a 2-modality model having a similar experimental setup. We

[1] https://github.com/idansc/HighOrderAtten

observe a significant improvement for our 3-modality model, which shows the importance of high-order attention models. Due to the fact that we use a lower embedding dimension of 512 (similar to [LYBP16]) compared to 2048 of existing 2-modality models [KOL+17, FPY+16], the 2-modality model achieves inferior performance. We believe that higher embedding dimension and proper tuning can improve our 2-modality starting point.

Additionally, we compared our proposed decision units. MCT, which is a generic extension of MCB for 3-modalities, and 2-layers MCB which has greater expressiveness (Sec. 3.3.2). Evaluating on the 'val' dataset while training on the 'train' part using the VGG features, the MCT setup yields 63.82% where 2-layer MCB yields 64.57%. We also tested a different ordering of the input to the 2-modality MCB and found them to yield inferior results.

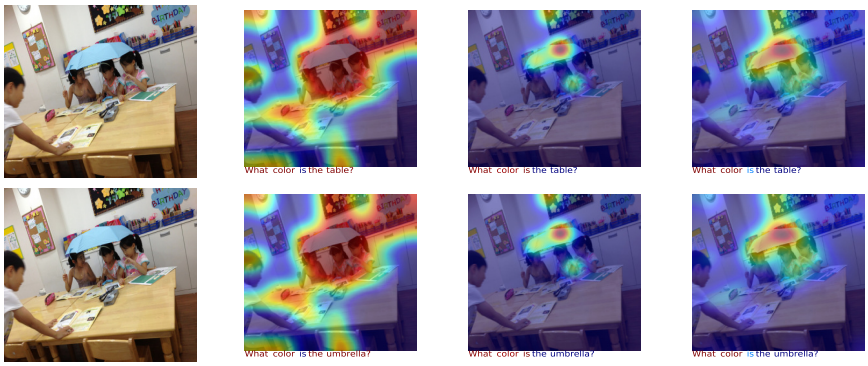
Qualitative Evaluation: Next, we evaluate our technique qualitatively. In Fig. 3.5 we illustrate the unary, pairwise and combined attention of our approach based on the two modality architecture, without the multiple choice as input. For each image we show multiple questions. We observe the unary attention usually attends to strong features of the image, while pairwise potentials emphasize areas that correlate with question words. Importantly, the combined result is dependent on the provided question. For instance, in the first row we observe for the question "How many glasses are on the table?," that the pairwise potential reacts to the image area depicting the glass. In contrast, for the question "Is anyone in the scene wearing blue?" the pairwise potentials reacts to the guy with the blue shirt. In Fig. 3.6, we illustrate the attention for our 3-modality model. We find the attention over multiple choice answers to favor the more unusual results.

In Fig. 3.7, we compare the final attention obtained from our approach to the results obtained with techniques discussed in [LYBP16] and [FPY+16]. We observe that our approach attends to reasonable pixel and question locations. For example, considering the first row in Fig. 3.7, the question refers to the battery operated device. Compared to existing approaches, our technique attends to the laptop, which seems to help in choosing the correct answer. In the second row, the question wonders "Is this a boy or a girl?". Both of the correct answers were produced when the attention focuses on the hair.

In Fig. 3.8, we illustrate a failure case, where the attention of our approach is identical, despite two different input questions. Our system focuses on the colorful umbrella as opposed to the object queried for in the question.

3.4 Conclusion

In this thesis, we investigated a series of techniques to design attention for multimodal input data. Beyond demonstrating state-of-the-art performance using relatively simple models, we hope that this work inspires researchers to work in this direction.



What color is the table?
GT: brown
Ours: blue

What color is the umbrella?
GT: blue
Ours: blue

Figure 3.8: Failure cases: Unary, pairwise and combined attention of our approach. Our system focuses on the colorful umbrella as opposed to the table in the first row.

Chapter 4

Visual Dialog

Dialog is an effective way for humans to exchange information. Due to this effectiveness it is an important research goal to develop artificial intelligence based agents for human-computer conversation. However, when humans talk to each other, subtle details and nuances are often very important. This importance of subtle details and nuances makes development of agents for visual dialog a challenging endeavor.

Recent efforts to facilitate human-computer conversation about images focus on image captioning, visual question answering, visual question generation and very recently also visual dialog. To this end, Das *et al.* [DKG+18] collected, curated and provided to the general public an impressive dataset, which allows to design virtual assistants that can converse. Different from image captioning datasets, such as MSCOCO [LMB+14], or visual question answering datasets, such as VQA [GKS+17], the visual dialog dataset contains short dialogs about a scene between two people. To direct the dialog, the dataset was collected by showing a caption to the first person ('questioner') which attempts to inquire more about the hidden image. The second person ('answerer') could see both the image and its caption to provide answers to these questions. Beyond releasing the Visual Dialog dataset, to ensure a fair comparison, Das *et al.* [DKG+18] propose a particular task that can be evaluated precisely. It asks the AI system to predict the next answer given the image, the question, and a history of question-answer pairs. A variety of discriminative and generative techniques have been discussed, ranging from deep nets with Long-Short-Term-Memory (LSTM) units [HS97a] to more involved ones with memory nets [WCB14] and hierarchical LSTM architectures [SSL+17].

One of the successful techniques to improve visual question answering is the attention mechanism [LYBP16]. Due to the similarity of visual question answering and visual dialog, we envision similar improvements to be realizable. In fact, some approaches point in this direction and use a subset of the available data utilities to direct question answering [LKY+17]. However, in visual dialog many more "data parts," *i.e.*, the image, the question, the history and the caption are involved and have been referred to as 'modalities.' To avoid confusion with the original convention/sense of the word modality, we coin the term "utilities" to refer to different parts of the available data.

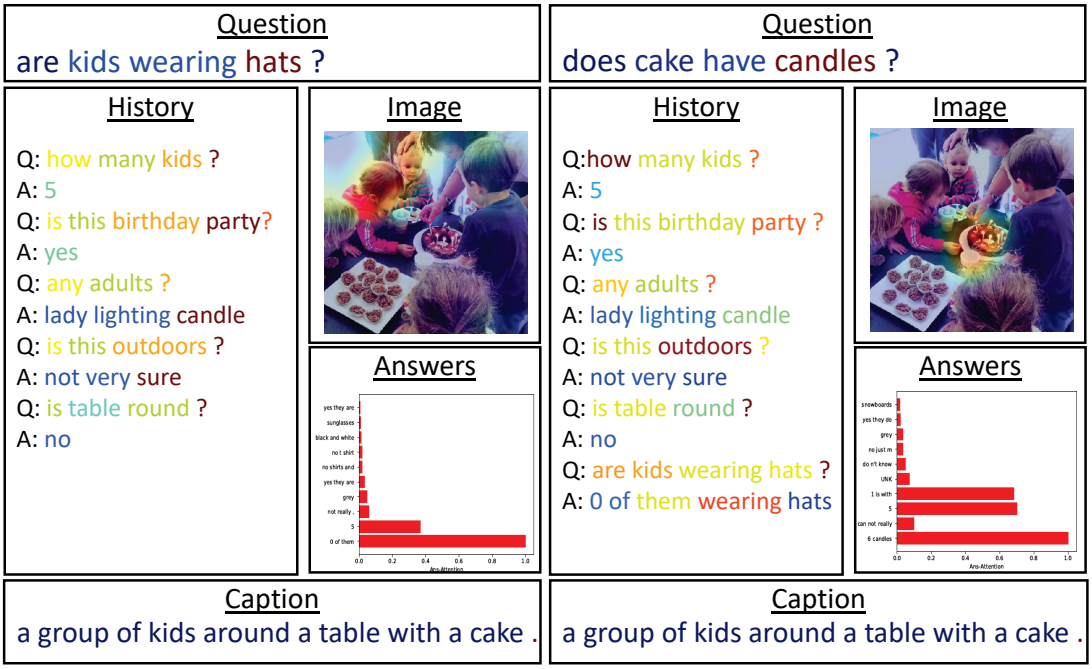


Figure 4.1: Illustration of our factor graph attention. We show two consecutive questions in a dialog. The image attention correlates well with the question. Attention over history interactions allows our model to attend to subtle nuances. The caption focuses on the last word due to given potential priors. Attention over the answers focuses on specific options. The attended options usually correlate with the correct answer. Note: for readability, we chose to display only the top-10 answers out of 100 possible ones.

Taking all utilities into account makes it computationally and conceptually much more challenging to develop an effective attention mechanism. While ignoring utilities when computing attention is always an option, we argue that subtle details and nuances can only be captured adequately if we focus on all available signals.

To address this issue we develop a general factor graph based attention mechanism which combines representations of any number of utilities. Inspired by graphical models, we use a graph based formulation to represent the attention framework, where nodes correspond to utilities and factors model their interactions. A message passing like procedure aggregates information from modalities which are connected by edges in the graph.

We demonstrate the efficacy of the proposed multi-utility attention mechanism on the challenging and recently introduced Visual Dialog dataset, realizing improvements up to 1% on MRR. Moreover, we examine our model behavior using question generation proposed by [JLS18]. Examples of the computed attention for visual question answering are illustrated in Fig. 4.1.

4.1 Related Work

Cognitive Tasks and Attention: Instrumental to cognitive tasks are attention models, that enable interpretation of the machine's cognition and often improve performance. While attention mechanisms have been applied to visual question answering [FPY+16, LYBP16, KDHR17, XS16], few works have addressed visual dialog because of the many different data utilities. Here, we develop an attention mechanism for visual dialog, a cognitive task that was created to imitate human-like decisions [DKG+18]. We build a general attention mechanism that is capable of capturing details. In the following we briefly review visual question answering and visual dialog, focusing on the use of attention.

Visual Question Answering (VQA): Visual question answering is considered a simplified version of visual dialog since it consists of a single interaction with a given image. Some discriminative approaches include a pre-trained convolutional neural network with question embedding to predict the correct answer [SZ15, MRF15]. Quickly, attention mechanisms have emerged as a tool to augment the spatial attention of the image. Yang *et al.* [YHG+16] created a multi-step reasoning system via an attention model. Fukui *et al.* [FPY+16] and Kim *et al.* [KOL+17] suggested an efficient multimodal pooling method before applying attention using a compact outer product which was later improved using the Hadamard product. Zhu *et al.* [ZZH+17] treated image attention as a structured prediction task over regions, by first generating attention beliefs via unary and pairwise potentials, for which a probability distribution is inferred via loopy belief propagation.

Alternatively, Lu *et al.* [LYBP16] suggested to produce Co-Attention for the image and question separately, using a hierarchical formulation. We extended this approach

for the multiple-choice VQA variant, applying attention over image, question and answer via unary, pairwise and ternary potentials (See Sec. 3.2.2).

Visual Dialog: D. Geman *et al.* [GGHY15] were among the first to generate dialogs over images. These early attempts used only street scene images, and also restricted the conversation to templated, binary questions. A discriminative and generative approach was later introduced by Das *et al.* [DKG+18], along with the largest visual dialog dataset, VisDial. Concurrently, GuessWhat, another visual dialog dataset was published [DVSC+17]. GuessWhat is a goal driven dialog dataset for object identification, while VisDial focuses on human-like interactions. For instance, in Fig. 4.8, the answer for the question "are kids wearing hats?" is "0 of them wearing hats," while a goal-driven interaction will answer with a simple "no." While both types of dialogs are challenging, VisDial interactions typically consider more subtle nuances.

The VisDial dataset is accompanied with three baselines. A vanilla approach which encodes the image, dialog and history separately and combines them subsequently (*i.e.*, late fusion). A more complex approach based on a memory network [WCB14], which maintains previous question and answer as facts in a memory bank, and learns to retrieve the appropriate fact. Lastly, a hierarchical encoding approach to capture the history [SSL+17]. Seo *et al.* [SLHS17] propose a memory network based on attention, which also addressed co-referential issues. Later, Lu *et al.* [LKY+17, LYBP16] combined a generative and discriminative model to choose generated answers, and also proposed history attention conditioned on the image using hierarchical co-attention developed for visual question answering. Kottur *et al.* [KMP+18] focused on visual co-reference resolution for visual dialog. While co-reference resolution is not the focus of our work, we found our attention model to exhibit some co-reference resolution abilities. Jain *et al.* [JLS18] developed a discriminative model that produces a binary score for each possible answer by concatenating representations of all utilities. While Jain *et al.* [JLS18] also consider all utilities for interaction prediction, our work differs in important aspects: (1) we develop an attention mechanism that weights different representations; (2) when predicting an answer, we take information from other possible answers into account.

Among all attention-based techniques for Visual Dialog, the most relevant to our approach is work by Lu *et al.* [LKY+17] that use Co-Attention over the image, the question and the history representation in a hierarchical fashion. Their hierarchical approach is based on a sequential process, computing attention for one utility first and using the obtained result to generate attention for another utility subsequently. As the ordering is important, their framework is not straightforward to extend to a general multi-utility setting.

In contrast, we develop a general attention model for any number of utilities. In the visual dialog setting, those utilities are the question in the history (10 utilities), each answer in the history (10 utilities), the caption (1 utility), the image (1 utility) and the answer representation (1 utility). To work with a total of 23 utilities, we constructed

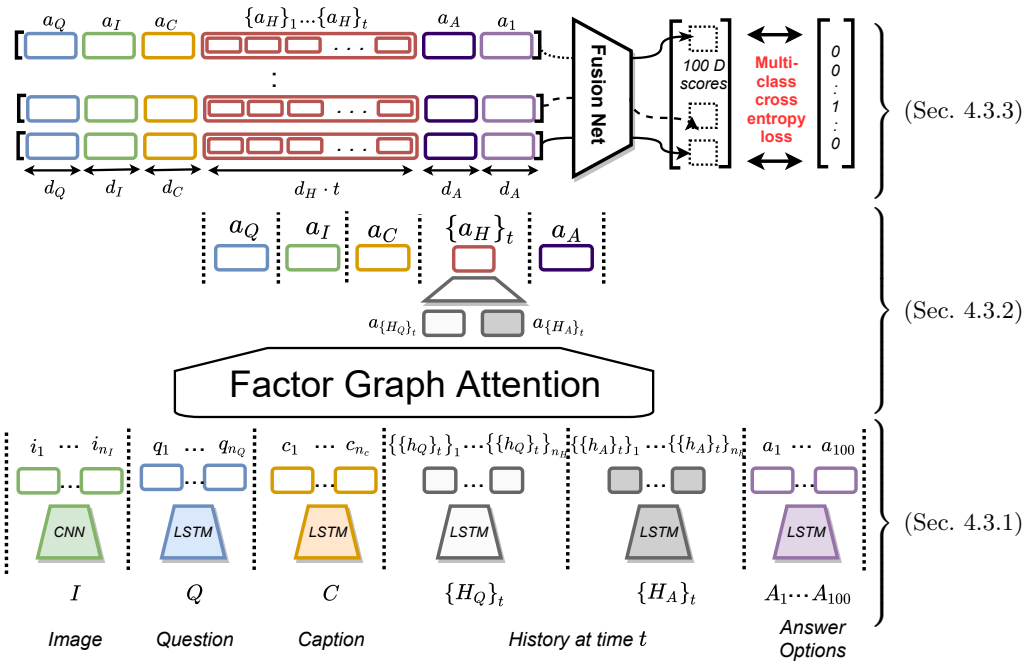


Figure 4.2: Our state-of-the-art architecture for the Visual Dialog task. Implementations details can be found in Sec. 4.3.

a general attention framework that may be applied to any high-order utility setting. With our general purpose attention model we improve results and achieve state-of-the-art performance.

To demonstrate the generality of the approach, we also follow Jain *et al.* [JLS18] and evaluate the proposed approach on choosing an appropriate question given the previous question and answer. There too we obtain state-of-the-art results.

4.2 Factor Graph Attention

In the following, we describe a general framework to construct a multi-utility attention model using factor graphs. We use the visual dialog task as guidance since it requires the encoding of many modalities, which is our primary goal.

The factor graph is defined over *utilities*, which, in the visual dialog setting, consists of an image I, an answer A, a caption C, and a history of past interactions $(H_{Q_t}, H_{A_t})_{t\in\{1,\dots,T\}}$. We subsume all utilities within the set:

$$\mathcal{U} = \{I, A, C, (H_{Q_t}, H_{A_t})_{t\in\{1,\dots,T\}}\}.$$

In our work we have 23 utilities (10 history questions, 10 history answers, the image, answer and caption). For notational convenience and to demonstrate the generality of the formulation we also refer to the set of utilities via $\mathcal{U} = \{U_1, \dots, U_{|\mathcal{U}|}\}$. Each utility $U_i \in \mathcal{U}$, for $i \in \{1, \dots, |\mathcal{U}|\}$ consists of basic entities, *e.g.*, a question is composed of a

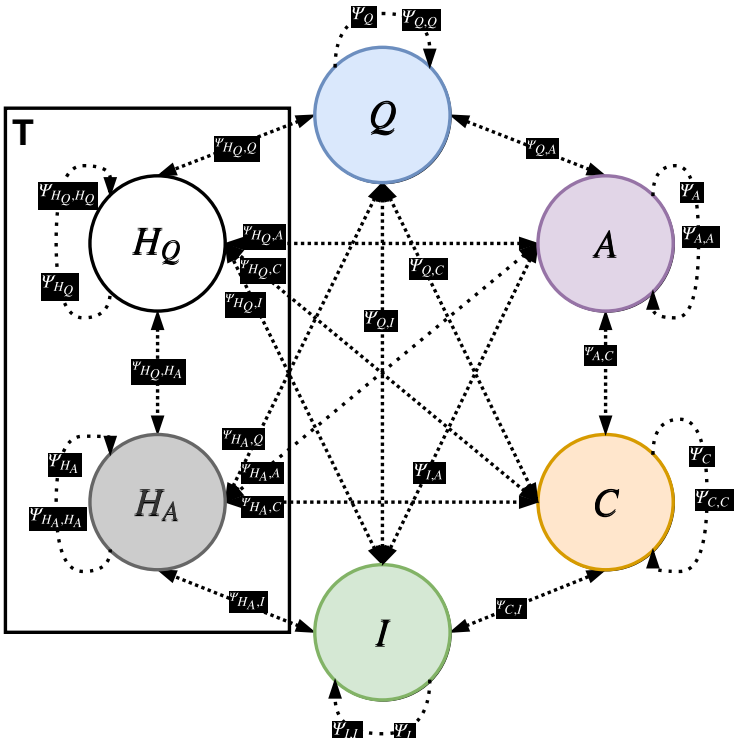


Figure 4.3: A graphical representation of our attention unit. Each node represents an attention probability over the utilities' entities. To infer the probability we aggregate two types of messages: 1) A joint factor message, constructed from interactions of entities from different utilities, *e.g.*, $\Psi_{Q,I}$. 2) A local factor: learned from the entity representation, *e.g.*, Ψ_Q, and the self entity interactions, *e.g.*, $\Psi_{Q,Q}$. T is the number of history dialog interactions.

sequence of words and an image is composed of spatially ordered regions.

Formally, the i-th utility U_i is a $d_i \times n_i$ matrix which consists of n_i entities $\hat{u}_i \in U_i$, which are the d_i-dimensional columns of the matrix. Each vector $\hat{u}_i \in U_i$ is embedded in its respective Euclidean space, *i.e.*, $\hat{u}_i \in \mathbb{R}^{d_i}$, where d_i is the embedding dimension of the i-th utility. We use the index $u_i \in \{1, \ldots, n_i\}$ to refer to a specific column inside the matrix U_i, *i.e.*, we extract the u_i-th column via $\hat{u}_i = U_{i,u_i}$.

The $|\mathcal{U}|$ nodes in the factor graph each represent attention distributions over their n_i utility elements, which we call beliefs. To infer the probability we take into account two types of factors: 1) Local factors which capture information within a utility, such as their entity representation and their local interactions. 2) Joint factors which capture interactions of any subset of utilities. Due to the high number of utilities, in our attention model, we limit ourselves to pairwise factors. See Fig. 4.3 for graphical illustration of the modeled dependencies.

Next we will explain our construction of local factors and joint factors. Note, bias terms are omitted for readability.

4.2.1 Local Factors

The local factors capture the local information in an employed utility U_i. Each utility contains entities, *i.e.*, words in a sentence or regions in an image. There are two types of information within a utility U_i: *Entity information*, which is extracted from an entity's vector representation $\hat{u}_i \in U_i$ and *Entity interactions*, which capture dependencies between two entities, such as two words in the same question or two regions in the same image.

Entity Information: This representation is obtained as the result of an embedding model, such as a Long-Short-Term-Memory (LSTM) net for sentences or a convolutional layer for image regions. Each vector representation $\hat{u}_i \in U_i$ has the potential to focus the model's attention to the entity the vector is representing. The potential function $\psi_i(u_i)$ is parametrized by the i-th utility's parameters V_i and v_i, and is obtained via

$$\psi_i(u_i) = v_i^\top \operatorname{relu}(V_i \hat{u}_i).$$

Hereby, $v_i \in \mathbb{R}^{d_i}, V_i \in \mathbb{R}^{d_i \times d_i}$ are trainable parameters. Recall that the index $u_i \in \{1, \ldots, n_i\}$ refers to a specific entity. During training we also apply a dropout operation after the first linear embedding (*i.e.*, $V_i \hat{u}_i$).

Entity Interactions: The factor dependency between two elements is extracted from their vector representation. Given two indices $u_i^1, u_i^2 \in \{1, \ldots, n_i\}$, we embed the two corresponding entity representation vectors $\hat{u}_i^1, \hat{u}_i^2$ in the same Euclidean space, and compute the factor dependency on both entities using the dot product operation, *i.e.*,

$$\psi_{ii}(u_i^1, u_i^2) = \left(\frac{L_i \hat{u}_i^1}{\|L_i \hat{u}_i^1\|} \right)^\top \left(\frac{R_i \hat{u}_i^2}{\|R_i \hat{u}_i^2\|} \right),$$

where $L_i \in \mathbb{R}^{d_i \times d_i}$, $R_i \in \mathbb{R}^{d_i \times d_i}$ are trainable parameters, governing the left and right arguments respectively.

4.2.2 Joint Factors

Joint factors capture interactions between two elements of different utilities, *e.g.*, between a word in the question and a region in the image. Similarly to entity interaction factors within a utility, we use

$$\psi_{ij}(u_i, u_j) = \left(\frac{L_{ij} \hat{u}_i}{\|L_{ij} \hat{u}_i\|} \right)^\top \left(\frac{R_{ji} \hat{u}_j}{\|R_{ji} \hat{u}_j\|} \right),$$

where $L_{ij} \in \mathbb{R}^{d_i \times d}$, $R_{ji} \in \mathbb{R}^{d_j \times d}$ are trainable parameters. For simplicity we let $d = \max\{d_i, d_j\}$ be the maximum dimension between the two utilities.

To avoid a situation where pairwise scores (*e.g.*, image and question) negatively bias another one (*e.g.*, image and caption), proper normalization is necessary. Since

the pairwise interaction scores are generated during training, we chose a batch normalization [IS15] operation which fixes the bias during training. Additionally, we applied an L_2 normalization on u_i and u_j to be of unit norm before the multiplication, *i.e.*, we use the cosine similarity.

4.2.3 Attention, Messages and Beliefs

For each utility U_i we infer the amount of attention that should be given to each of its elements $\hat{u}_i \in U_i$. Motivated by classical message-passing algorithms, we first collect all dependencies of a given utility element via

$$\mu_{j \to i}(u_i) = \sum_{u_j \in \{1,\dots,n_j\}} W_{ij}(u_i, u_j)\psi_{ij}(u_i, u_j),$$

where $W_{ij}(u_i, u_j) \in \mathbb{R}$ is a trainable parameter. We aggregate these messages from all pairwise factor dependencies and send them to a utility, in order to infer its attention belief. The inferred attention belief

$$b_i(u_i) \propto \exp\left(\hat{w}_i p_i(u_i) + w_i \psi_i(u_i) + \sum_{j=1}^{|\mathcal{U}|} w_{ij}\mu_{j \to i}(u_i)\right),$$

also uses local entity information.

Hereby w_{ij}, w_i are scalar weights learned per utility. These scalars reflect the importance of one utility with respect to the others. For instance, for the image belief, we find by examining these weights that the question utility is more important than the caption utility. This makes sense since we want to look at relevant places for the question. Moreover, p_i is a prior potential for the i-th utility, and $\hat{w}_i$ is a trainable parameter to calibrate the prior potential's importance. For instance, the question utility prior encourages focus of its attention onto the last word in the question, a common practice in LSTM networks. Using priors, we are able to steer the desired belief for a utility, while still allowing guidance of other utilities via pairwise interactions. We also experimented with priors that are updated after we infer the attention through steps, but we didn't find it to improve the results in our setup.

Once the attention belief $b_i(u_i)$ is computed for each entity representation $\hat{u}_i \in U_i$, we obtain the attended vector of this utility as the average representation. This reduces the utility representation to a single vector, which is dependent on the other utilities via the belief $b_i(u_i)$:

$$a_i = \sum_{u_i \in \{1,\dots,n_i\}} b_i(u_i) \cdot \hat{u}_i.$$

Note that a_i is the attended representation of utility U_i.

4.3 Factor Graph Attention for Visual Dialog

We use visual dialog to demonstrate the generality of the discussed attention mechanism because many utilities are available. A general overview of the approach is illustrated in Fig. 4.2. We detail next how the general factor graph attention model is applied to visual dialog by describing (1) the utility embeddings, (2) the attention module, and (3) the fusion of attended representations for prediction.

4.3.1 Utilities and Embeddings

In the following, we describe the embeddings of the image and textual utilities.

Image utility: To represent the image regions, we use a conv net, pre-trained on ImageNet [DDS+09]. Taking the output of the last convolutional layer we obtain a representation of $7 \times 7 \times 512$. Specifically, 7×7 is the spatial dimension of the convolutional layer and 512 is the number of channels/features of the representation. Following our notation in Sec. 4.2, the visual utility U_i has dimensions $n_i = 49$ and $d_i = 512$. To fine-tune this representation to our task, we feed it into another convolutional layer, with a 1×1 kernel, followed by a ReLU activation and a dropout.

Textual utilities: Our textual utilities are the caption, the question, the possible answers and the history interactions. For each textual utility U_i we embed up to n_i words. Sentences with a shorter length are zero padded, while sentences of longer length are truncated. The embedding starts with a one-hot encoding representation of the word index, followed by a linear transformation. The linear transformation embeds the word index into the Euclidean space. This embedding is identical for all textual utilities. Intuitively, usage of the same embedding ensures a better consistency between the textual utilities and we also found it to improve the results.

Each embedded representation for each textual utility is fed into an LSTM layer, which yields a representation with the appropriate embedding dimension. The caption utility C and the question utility Q are generated by applying a dedicated LSTM on the respective embedded representation. In contrast, we embed all history questions $(H_{Q_t})_{t\in\{1,\dots,T\}}$ using the same LSTM model. We also embed all history answers $(H_{A_t})_{t\in\{1,\dots,T\}}$ using another LSTM model.

The answer utility subsumes n_A possible answers and it consists of the final decision of the model in our visual dialog system. Our answer utility uses the same LSTM to embed each of the $n_A = 100$ answers separately, the embedding of each possible answer is the LSTM hidden state of the last word in the answer.

4.3.2 Attention Module

The attention step infers the importance of each entity in each utility, using our Factor Graph Attention (see Sec. 4.2), and creates an attended representation. In the visual dialog setting, for each answer generation step we use an image I, a question Q, an

answer A, a caption C, and a history of past interactions $(H_{Q_t}, H_{A_t})_{t\in\{1,\ldots,T\}}$ (see Fig. 4.3 for an illustration). In the following we describe the special treatment of the different entities as well as their respective priors.

Group utilities and dependency-relaxation: Our factor graph attention model may have a large number of trainable parameters, as it grows quadratically with the number of utilities. To address this concern, we observe that we can group some utilities, *e.g.*, the history answers $(H_{A_t})_{t\in\{1,\ldots,T\}}$, and the history questions $(H_{Q_t})_{t\in\{1,\ldots,T\}}$. To take advantage of the dependency between the group of utilities, we share the factor weights across all the group utilities. For example, for two utilities $U_{i_1}, U_{i_2} \in H_{A_t}$ we enforce the parameter sharing $v_{i_1} = v_{i_2}$, $V_{i_1} = V_{i_2}$, $L_{i_1} = L_{i_2}$, $R_{i_1} = R_{i_2}$, $L_{i_1,j} = L_{i_2,j}$ and $R_{j,i_1} = R_{j,i_2}$. Not only did it contribute to a reduced memory consumption, but we also observed this grouping to improve the results. We attribute the improvement to better generalization of the factors.

The answer utility U_i encodes each of the possible n_i answers in a d_i-dimensional vector, using the LSTM hidden state at the last word. Fig. 4.8 shows that the attention beliefs correlate with the correct answer. Note that we didn't attend separately to each possible answer. Doing so would have resulted in increased computational demand and we didn't find improved model performance. We conjecture that due the fact that the number of words within an answer is usually small, a complete attention model on each and every word of the answer does not seem to be necessary.

Priors: The prior potentials for the question and caption utilities are important in practice. For both utilities we set the prior to emphasize the last word by focusing the energy onto the last hidden state index. We use a one hot vector with the high bit set for the last hidden state index.

4.3.3 Fusion Step

The fusion step, outlined in Fig. 4.2 combines the attended representations a_i from all utilities $\{I, A, C, (H_{Q_t}, H_{A_t})_{t\in\{1,\ldots,T\}}\}$ to find the best answer. This is performed by creating a probability distribution $p(u_A|I, Q, C, A, H)$ for each answer index $u_A \in \{1, \ldots, n_A\}$, where $n_A = 100$ is the number of possible answers.

We denote by $a_I \in \mathbb{R}^{d_I}$ the attended image vector, $a_A \in \mathbb{R}^{d_A}$ the attended answer vector, and $a_C \in \mathbb{R}^{d_C}$ the attended caption vector. We construct the attended history vector $a_H \in \mathbb{R}^{d_I}$ from the attended history utilities $(H_{Q_t}, H_{A_t})_{t\in\{1,\ldots,T\}}$. For this purpose, we start by concatenating the attended vector of each history question a_{Q_t} with the concurrent history answer a_{A_t}, and fuse them using a linear transformation with a bias term to obtain a_t, which is a d_t-dimensional vector. We then concatenate the attended history vectors a_t for the entire dialog history $t \in \{1, \ldots, T\}$, which results in an attended history representation $a_H \in \mathbb{R}^{d_H}$. Note that $d_H = \sum_{t=1}^{T} d_t$. We concatenate the image, question, caption and history attended representations, which yields an attention representation $a \in \mathbb{R}^L$ of length $L = d_I + d_Q + d_C + d_A + d_H$.

Table 4.1: Performance of discriminative models on VisDial v0.9. Higher is better for MRR and recall@k, while lower is better for mean rank. (*) denotes use of external knowledge.

Model	MRR	R@1	R@5	R@10	Mean
LF [DKG+18]	0.5807	43.82	74.68	84.07	5.78
HRE [DKG+18]	0.5846	44.67	74.50	84.22	5.72
HREA [DKG+18]	0.5868	44.82	74.81	84.36	5.66
MN [DKG+18]	0.5965	45.55	76.22	85.37	5.46
HieCoAtt-QI [LYBP16]	0.5788	43.51	74.49	83.96	5.84
HCIAE-NP-ATT [LKY+17]	0.6222	48.48	78.75	87.59	4.81
SF-QIH-se-2 [JLS18]	0.6242	48.55	78.96	87.75	4.70
CorefNMN [KMP+18]*	0.636	50.24	79.81	88.51	4.53
CorefNMN (ResNet-152) [KMP+18]*	0.641	50.92	80.18	88.81	4.45
FGA (VGG)	0.6525	51.43	82.08	89.56	4.35
FGA (F-RCNNx101)	0.6712	54.02	83.21	90.47	4.08
9×FGA (VGG)	**0.6892**	**55.16**	**86.26**	**92.95**	**3.39**

Table 4.2: Performance on the question generation task. Higher is better for MRR and recall@k, while lower is better for mean rank.

Model	MRR	R@1	R@5	R@10	Mean
SF-QIH-se-2 [JLS18]	0.4060	26.76	55.17	70.39	9.32
FGA	**0.4138**	**27.42**	**56.33**	**71.32**	**9.1**

Next, we combine the image, question, caption and history attended representation $a \in \mathbb{R}^L$ with the $n_A = 100$ possible answers to compute a probability for each answer. Let $U_A \in \mathbb{R}^{n_A \times d_A}$ be the answer utility, with $N = n_A = 100$ answers, while each answer is embedded in a d_A-dimensional space. For each answer, we denote by $\hat{u}_A \in \mathbb{R}^{d_A}$ its embedded vector. We concatenate each answer embedding with the system attention $(a, \hat{u}_A)$ to obtain a $(L+d_A)$-dimensional vector and feed it into a multi-layer perception with two layers of size $(L + d_A)/2$ and $(L + d_A)/4$ respectively. Between each layer we perform batch normalization followed by a ReLU activation. We used a dropout layer before the last fully connected layer. The obtained scores are turned into probabilities, for each answer, using a softmax $(\cdot)$ operation, which yields the posterior probability for each answer $p(u_A|I, Q, C, A, H)$. The approach is trained using maximum likelihood.

4.4 Results

In the following we evaluate the proposed factor graph attention (FGA) approach on the Visual dialog dataset, which we briefly describe first. Our code is publicly available[1].

Visual Dialog Dataset: We used VisDial v0.9 to train the model. The dataset consists of approx. 120k images from COCO [LMB+14]. Each image is annotated with a dialog of 10 questions and corresponding answers, for a total of approx. 1.2M dialog

[1] https://github.com/idansc/fga

Table 4.3: Performance of discriminative models on VisDial v1.0 test-std. Higher is better for MRR and recall@k, while lower is better for mean rank and NDCG. (*) denotes use of external knowledge.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
LF [DKG+18]	0.554	40.95	72.45	82.83	5.95	0.453
HRE [DKG+18]	0.542	39.93	70.45	81.50	6.41	0.455
MN [DKG+18]	0.555	40.98	72.30	83.30	5.92	0.475
CorefNMN (ResNet-152) [KMP+18]*	0.615	47.55	78.10	88.80	4.40	0.547
NMN (ResNet-152) [HAR+17]*	0.588	44.15	76.88	86.88	4.81	**0.581**
FGA (VGG)	0.637	49.58	80.97	88.55	4.51	0.521
FGA (F-RCNNx101)	0.662	52.75	82.92	91.07	3.8	0.569
5×FGA (VGG)	0.673	53.40	85.28	92.70	3.54	0.545
5×FGA (F-RCNNx101)	**0.693**	**55.65**	**86.73**	**94.05**	**3.14**	0.572

Table 4.4: Attention-related ablation analysis.

Model	MRR	R@1	R@5	R@10	Mean
No Attention	0.6249	48.67	78.95	87.73	4.69
No BatchNorm	0.6301	49.23	79.65	88.32	4.55
No Local-Interactions	0.6369	50.17	79.92	88.33	4.55
No Local-Information	0.6425	50.12	81.49	89.34	4.37
No Priors	0.6451	50.57	81.37	89.00	4.47
FGA	**0.6525**	**51.43**	**82.08**	**89.56**	**4.35**

question-answer pairs. In the discriminative setup, each question-answer pair is given 100 plausible possible answers, the model needs to choose from. We follow [DKG+18] and split the data into 80k images for train, 40k for test and 3k for validation.

Experimental setup: We used a batch size of 64. We set the word embedding dimension to $d_E = 128$, and the utility embeddings to $d_Q = 512$ and $d_C = 128$. For each question or answer in the history we use $d_{H_{Q_i}} = d_{H_{A_i}} = 128$. For each possible answer we use $d_a = 512$. The lengths are set equally for all textual utilities $n_Q = n_C = n_a = n_{H_Q} = n_{H_A} = 20$. The VisDial history consists of $T = 10$ questions with their answers. For our image representation we use the last conv layer of VGG having dimensions of $7 \times 7 \times 512$. After flattening the 2D spatial dimension, $n_I = 49$. The dropout parameter after the image embedding is set to 0.5, the dropout parameter before the last fc layer is set to 0.3.

Training: The total amount of trainable parameters in our model is $17,848,416$. We initialized all the weights in the model using Kaiming normal initialization [HZRS15b]. To train the model we used a multi-class cross entropy loss, where each possible answer represents a class. We used Adam optimizer with a learning rate of 10^{-3}. We evaluate our performance on the validation set after each epoch to determine when to stop our training.

Model	MRR	R@1	R@5	R@10	Mean
No Answer Utility	0.6294	49.35	79.31	88.10	4.63
No History Attention	0.6449	50.74	81.07	88.86	4.48
Answers Fine-attention	0.6478	50.80	81.86	89.25	4.46
History No Fine-attention	0.6494	51.17	81.56	89.13	4.43
FGA	**0.6525**	**51.43**	**82.08**	**89.56**	**4.35**

Table 4.5: Utility-related ablation analysis.

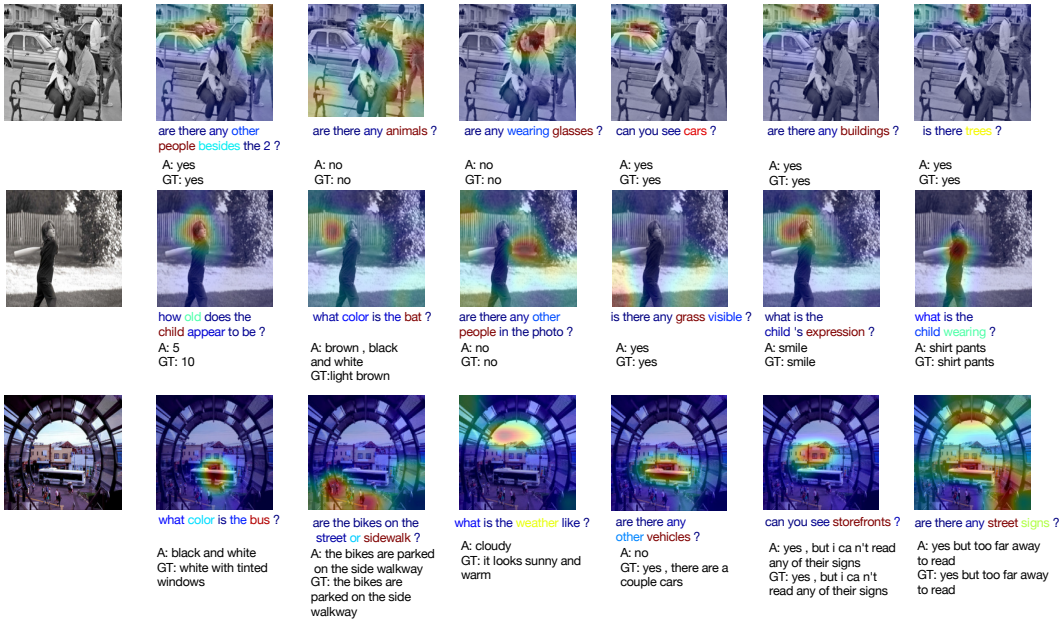


Figure 4.4: An illustration of question and image attention over a series of interactions for the same dialog. In addition we provide the ground truth answer, *i.e.*, GT, and our predicted answer, *i.e.*, A.

Quantitative Evaluation

Evaluation metrics: Evaluating dialog systems, or any other generative tasks is challenging [LLS+16]. We follow [DKG+18] and evaluate each individual response at each of the $T = 10$ rounds in a multiple-choice setup. The model is hence evaluated on retrieval metrics: Recall@k is the percentage of questions where the human response was part of the top k predicted answers. Mean rank is the average rank allotted by a model to the human response, hence a lower score is desired. Mean Reciprocal Rank (MRR) is defined as $\frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}$, where rank_i is the rank of the human response, and Q is the set of all questions. The perfect score, *i.e.*, MRR $= 1$ is achieved when the human response is consistently ranked first.

Visual question answering comparison: We first compare against a variety of baselines (see Tab. 4.1). Note that almost all of the baselines (except LF, HRE and MN and SF-QIH-se-2) use attention, *i.e.*, attention is an important element in any model. Note that our model uses the entire set of answers to predict each answer's score, *i.e.*, we use $p(u_i|A, I, Q, C, H)$ This is in contrast to SF-QIH-se-2, which doesn't

use attention and models $p(u_i|\hat{u}_i, I, Q, C, H)$. Notable as well, the current state-of-the-art model, CoAtt-GAN [WWS+17], used the largest amount of utilities to attend to, *i.e.*, image, question and history. Because CoAtt-GAN uses a hierarchical approach, the ability to further improve the reasoning system is challenging and manual work. In contrast, our general attention mechanism allows to attend to the entire set of cues in the dataset, letting the model automatically choose the more relevant cues. We refer the readers to the appendix for analysis of utility-importance via importance score. As can be seen from Tab. 4.1, this results in a significant improvement of performance, even when compared to the very recently published baselines [JLS18, WWS+17, KMP+18]. We also report an ensemble of 9 models which differ only by the initial seed. We emphasize that our approach only uses VGG16. Lastly, some baselines report to use GloVe to initialize the word embeddings, while we didn't use any pre-trained embedding weights.

Our attention model is very efficient to train. Our state-of-the-art score is achieved after only 4 epochs. Each epoch takes approximately 2 hours on a standard machine with an Nvidia Tesla M40 GPU. In contrast, CorefNMN [KMP+18], has 100M parameters and takes 33 hours to train on a Titan X. Both [LKY+17, WWS+17] report that more than 25 epochs 101M parameters and 50 hours were required for training.

Visual question generation comparison: To assess question generation, [JLS18] proposed to predict the next question given the previous question and answer. Their introduced question prediction dataset is based on VisDial v0.9, along with a collected set of 100 question candidates.

We adapted to this task, by changing the input utilities to the previous interaction $(Q + A)_{t-1}$ instead of the current question Q_t. Our model also improves previous state-of-the-art results (see Tab. 4.2).

Visual Dialog Challenge: Recently, VisDial v1.0 was released as part of the Visual Dialog challenge, where 123,287 images are used for training, 2,000 images for validation, and 8,000 images for testing. For the test split each image consists of only 1 interaction, at some point of time in the dialog. Furthermore, an additional metric, normalized discounted cumulative gain (NDCG), was introduced. NDCG uses dense annotations, *i.e.*, the entire set of candidate answers is annotated as true or wrong. The metric penalizes low ranking correct answers, addressing issues when the set of answers contains more than one plausible result.

Our submission to the challenge significantly improved all metrics except for NDCG. We report our results in Tab. 4.3 on test-std, a 4,000 image split, the other 4,000 image split was preserved for the challenge. While the challenge did allow use of any external resources to improve the model, we only changed our approach to use an ensemble of 5 trained Factor Graph Attention models which were initialized randomly. All other top teams used external data in form of detection features on top of ResNet-152, inspired by Top-Bottom attention [AHB+18]. These features are expensive to extract, and use external detector information.

Figure 4.5: Illustration of history attention for 2 interactions. We observe small nuances of history to be useful to answer questions, and improve co-reference resolution.

Our model used only the single ground truth answer to train. Therefore it is expected that our model isn't optimized w.r.t. the NDCG metric. However, given the small subset of densely annotated samples (2,000 out of the 123,287 train images), it is hard to carefully analyze this result.

Ablation Study: We asses (1) design choices of our factor graph attention; and (2) utility ablation focusing on history and answer cues as they are a unique aspect of our work. (1) In Tab. 4.4 we see that FGA improves the MRR of a model without attention by 3% (0.6249 *vs.* 0.6653). This ablation study shows that attention is crucial for VisDial. Removing local-information drops MRR to 0.6425. When omitting local-interactions, *i.e.*, a score based on interactions of embedding representations of a utility, the MRR drops to 0.6369. BatchNorm over pairwise interactions is crucial. Without BatchNorm MRR drops to 0.6301. Removing prior information, *e.g.*, a high prior potential for the last word in the question is less crucial, dropping MRR to 0.6451. (2) Our history attention attends separately to questions and answers in the history. In contrast, classical methods [SLHS17] attend over history locations only. Based on Tab. 4.5, we note that our fine-grained history attention improves MRR from 0.6494 to 0.6525. Without the answers utility, performance on MRR drops significantly from 0.6525 to 0.6294. If we attend to each word in the answers separately, *i.e.*, 'Answers Fine-Attention,' performance drops to 0.6478.

Other Datasets: When we replace the attention unit of other methods with our FGA unit we observe improvements in visual question answering (VQA) and audio-visual scene aware dialog (AVSD) [ACD$^+$18]. For VQA v1.0 we increase validation set accuracy from 57.0 to 57.3 (no tuning) by replacing the alternating and parallel attention [LYBP16]. For AVSD, we improve Hori *et al.* [HAW$^+$18] which report a CIDEr score of 0.733 to 0.806. We used FGA to attend to all video cues as well as the question. This differs from Hori *et al.* who mix the question representation with video-related cues (*e.g.*, I3D features, optical flow and audio features), and aggregate them to generate attention. Other components remain the same. Our flexible framework is instrumental for this improvement.

Bottom-up Features: We follow Anderson *et al.* [AHB+18] and use bottom-up features of 36 proposals from images. Equipped with bottom-up features as image representation our ensemble network increase MRR score on VisDial v1.0 by 2% (0.673 vs 0.693). For a single model we observe a similar boost in performance (0.6525 vs 0.6712) on VisDial v0.9.

Ensemble model for VisDial v0.9**:** For VisDial v1.0, a simple ensemble technique has significantly improved the results. We observe a similar effect for VisDial v0.9, pushing the current state of the art for MRR from 0.6525 to 0.6892 as summarized in Tab. 4.6. We achieve this result with an ensemble of 9 models which differ only by the initial seed. For VisDial v1.0 we report a 5 model ensemble score. Due to restriction of the number of submissions to the evaluation server we could not evaluate a larger ensemble model. The results in Tab. 4.6 suggest that the VisDial v1.0 score can be further improve with a larger ensemble model.

Analysis of Factor Graph Attention weights: To infer the attention belief for a utility, *i.e.*, $b_i(u_i)$, we aggregate marginalized joint and local interactions and also local-information and prior terms. To calibrate each cue, we use scalar weights, *i.e.*, $\hat{w}_i, w_i, w_{i,j}$. To obtain a better understanding of the reasoning process and analyze attention, we suggest an importance score:

$$S(\gamma) = \frac{|m_\gamma \cdot \gamma|}{\sum_{\delta \in \{\hat{w}_i, w_i, (w_{i,j})_{j \in \mathcal{U}}\}} |m_\delta \cdot \delta|}, \tag{4.1}$$

where $\gamma \in \{\hat{w}_i, w_i, (w_{i,j})_{j \in \mathcal{U}}\}$ is the weight of a cue and m_γ is the mean term of the corresponding cue γ, which was calculated over the entire validation set. Note that γ are the scalar weights. $S(w_{i,j})\ \forall j \in \mathcal{U}$ captures the importance of the j-th cue for utility i. A high score means the i-th utility attention belief heavily relies on cue j. Similarly, $S(w_{i,i}), S(w_i),\ S(\hat{w}_i)$ capture the importance of local-interactions, local-information and prior cues for the i-th utility. We report the scores in Tab. 4.7. We observe that the answer utility relies mostly on local-interactions. The question heavily relies on the prior, but also makes use of history answers and question cues. The caption ignores all utilities other than the prior. The image question utility is the most important cue. Interestingly, we observe importance of priors. Image attention relies on the captions, while the caption ignores all the cues and preserves the prior behavior. The history question and answers rely on the question and the local factors.

Computation and insignificant interactions: Upon training interactions may be found to be unnecessary. Our model can be optimized easily: 1) The score in Eq. (4.1), can be used to omit less significant interactions. Previous multimodal attention doesn't model pairwise interaction scores, making it hard to eliminate computations. 2) For the same image but different question, we can re-use calculated joint interactions, such as local-interaction, image-caption, *etc.* This is impossible for approaches that pool cues since the question changes. 3) It's possible to share weights between similar utilities, *e.g.*, different history questions/answers.

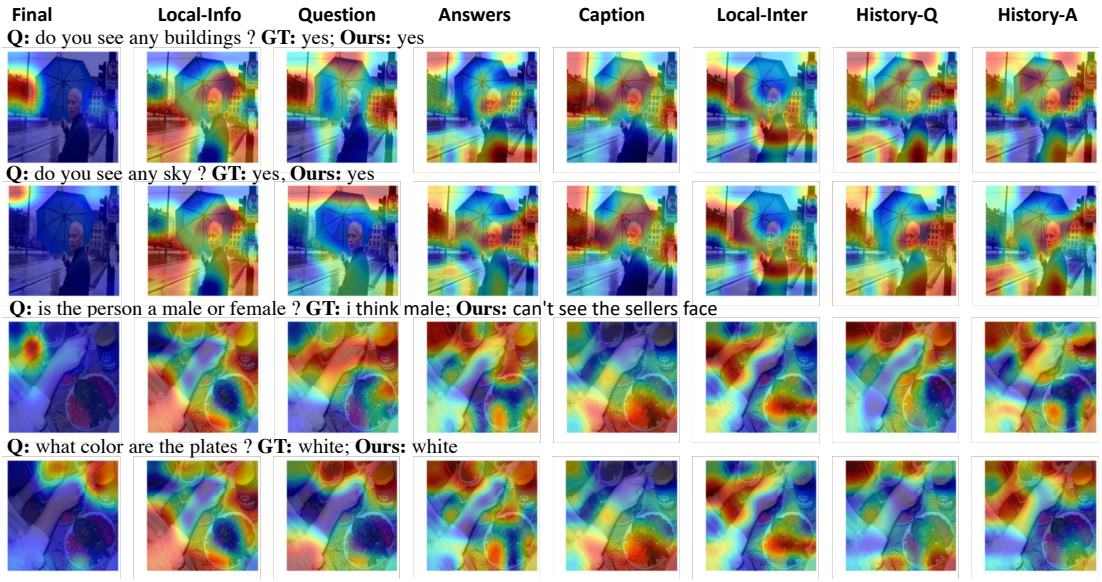


Figure 4.6: Two images each with two questions. We illustrate scores obtained from different types of factors. Local-info denotes 'Image-Local-Information,' Question refers to 'Image-Question,' *etc.* We observe 'Image-Question' to have the highest variance between different questions, since its heat map differs the most. 'Image-Question' also correlates the most with the final attention.

Currently, we don't consider most of those options, as the model trains quickly (8 hours vs. 33 hours of previous state-of-the-art) and fits into a single 12GB GPU.

Qualitative Evaluation

Attention is an important tool not only because it boosts performance, but also because it yields a weak form of interpretability. By illustrating the attention beliefs, we can observe the reasoning process of the model. In Fig. 4.4 we provide co-attention of image and question. The first row shows dialogs with yes/no questions. We observe the question attention to focus on the indicative word, *e.g.*, people, animals, buildings, cars, *etc.*, while the image attention performs detection and attends to the relevant area of the image. For the second row, again we observe plausible attention behavior.

Table 4.6: Analysis of ensemble models for VisDial v0.9. With an ensemble of 9 models we observe an improvement of more than 3% over the single model.

Model	MRR	R@1	R@5	R@10	Mean
FGA	0.6525	51.43	82.08	89.56	4.35
Ensemble of 2 FGA	0.6711	53.56	83.83	90.97	3.92
Ensemble of 3 FGA	0.6786	54.28	84.71	91.69	3.73
Ensemble of 4 FGA	0.6819	54.56	85.19	92.10	3.62
Ensemble of 5 FGA	0.6848	54.82	85.57	92.38	3.55
Ensemble of 6 FGA	0.6860	54.95	85.71	92.52	3.50
Ensemble of 7 FGA	0.6869	54.97	85.91	92.67	3.47
Ensemble of 8 FGA	0.6881	55.10	86.04	92.77	3.44
Ensemble of 9 FGA	**0.6892**	**55.16**	**86.26**	**92.95**	**3.39**

Table 4.7: For each utility (column) we show the three most related cues based on the S score given in Eq. (4.1). We provide S in parenthesis. P, L_1 and L_2 indicate prior, local-information and local-interactions of the utility in the column.

A	Q	C	I	H_Q	H_A
L_1 (0.125)	P(0.851)	P (0.988)	Q (0.593)	Q (0.205)	L_2 (0.607)
H_Q (0.122)	H_A(0.052)	I (0.004)	L_1 (0.186)	L_1 (0.121)	Q (0.304)
Q (0.075)	H_Q(0.031)	A (0.001)	C (0.123)	L_2 (0.085)	L_1 (0.017)

Figure 4.7: Illustration of 2 step interaction using visual question generation and illustration of the involved modalities. The classifier receives the previous question and answer, to predict a new one.

An interesting failure-case: when asked about the color of the bat, the ground-truth answer was "light brown," while our model answered "brown, black and white" instead. A possible explanation is related to the fact that the image is in black and white. The last line shows that question-answering type of task is always debatable. For the question "what is the weather like?" the model answered "cloudy," while the ground truth is "it looks sunny and warm." While it does look sunny, the model attends to clouds and the model answer likely isn't entirely wrong.

Next, in Fig. 4.5, we show how attention is useful when applied over each question in the history. In the first row, for the question "is this at a skateboarder park?", the skateboard related terms in the history are given more weight. Another use case of attention is co-reference resolution. We highlight those results in the second row: the word "they" in the second question refers to people in the background, which remain the focus of the attention model.

Lastly, in Fig. 4.7, we evaluate question generation and let the model interact with the answer predictor. We show how complete dialogs can be generated in a discriminative manner. We first observe that attention for question generation is noisier. This seems intuitive because asking a question requires a broader focus than answering. Nonetheless, visual input is important. For the second row second image, "at what angle is the man sitting?" the model attends mostly to the man, and for the question "is it a laptop or desktop?" image attention focuses on the laptop. Also, in both cases the caption attention is useful. For instance, in the first row, the word "tie" is picked to generate two relevant questions. This nicely illustrates how the proposed model adapts to tasks, when the importance of different data cues changes.

Factors visualization: We provide additional visualization in Fig. 4.6. We visualize scores for each image region obtained from different types of factors. 'Image-Local-Information,' 'Image-Caption' and 'Image-Local-Interaction' are constant for different questions, while 'Image-Question,' 'Image-Answer,' 'Image-History-Q' and 'Image-History-A' change for every question. We calculated the variance of interactions and observe that 'Image-Question' has the highest variance (≈ 3), while 'Image-Answer,' 'Image-History-Q' and 'Image-History-A' have a variance of ≈ 1. Beyond the importance score, the high-variance also suggests that the 'Image-Question' cue is most important.

4.5 Ensemble of MRR and NDCG models for Visual Dialog

[DKG+18] introduced the task of Visual Dialog, which requires an agent to converse about visual input. Evaluating visually aware conversation should examine both linguistic properties and visual reasoning. Analysis of generative metrics for dialog often shows no correlation with human judgments [LLS+16]. Hence, to evaluate the correctness of the candidate answers, a retrieval approach is preferred. Two metrics are standard, MRR and NDCG. The MRR metric focuses on a single human-derived ground-truth answer. Despite preferring the more human-like answer, the metric ignores many correct candidate answers. Differently, the NDCG considers the rank of all the correct answers. The metric relies on dense annotation, where three annotators were asked to mark all the correct candidate answers. However, the candidate answers are generated plausible answers. The analysis shows that the NDCG metric favors uncertain, generally correct answers, such as "not sure" [MBPD20, QNHZ20].

Prior work in visual dialog focused on a single metric. Ideally, an AI agent should answer human-like and detailed reply (the MRR metric) and be able to validate the correctness of any answer (the NDCG metric). However, crafting a model that excels in both metrics is challenging [MBPD20]. To this end, we propose principals to ensemble the rankings of strong MRR and NDCG models. Our approach is to find a minimal set that is likely to hold the human-derived answer. This permits ranking the rest of the candidates according to the NDCG model. Our approach won the recent Visual Dialog 2020 challenge and achieved strong performance on both the MRR and the NDCG metrics simultaneously.

4.5.1 Related Work

Visual conversation evaluation: Early attempts to marry conversation with vision used street scene images, and binary questions [GGHY15]. While binary answers are easy to verify, such an approach is limiting for an AI agent. On the other hand,

Figure 4.8: A visual dialog interaction. The question asks, "what is the nightstand made of ?". We show our final ranking, created by the ensemble of an MRR/NDCG models' rankings. The MRR/NDCG models are trained to optimize the MRR/NDCG metric. The MRR metric measures the number of retrievals to retrieve the human-derived answer. Hence, the MRR model favors human-like and detailed answers. On the other hand, the NDCG metric measures the rank of all the correct candidates based on dense annotation, which are often general and uncertain. Our ensemble approach seeks a minimal candidate set that is likely to contain the human-derived answer. The remaining candidates are ranked according to the NDCG model.

analysis of generative metrics for dialog often show no correlation with human judgements [LLS+16]. Intuitively, metrics like BLEU-scores rely on corresponding words with the ground-truth answer and often miss synonyms or the subjective nature. More importantly, generative metrics are geared toward textual assessment rather than visual reasoning, which results in models mainly relying on textual cues [SSH19]. Malinowski *et al.* [MF14] suggest Wu-Palmer similarity metric that calculates similarity based on the depth of two words based on the WordNet taxonomy [Mil95]. A different approach suggested in the VQA dataset focus only on brief, mostly 1-word answers [GKS+17]. In this setup, the task turns into popular answers classification, alleviating many text-generation challenges. Notably, VQA requires 3 out of 10 annotators to agree on the answer, which is robust to inter-person variation. Still, accuracy ignores the reasoning process. Hudson *et al.* [HM19] propose GQA, which extends the accuracy metric and uses a scene graph for both question generation and evaluation. Following, Das *et al.* [DKG+18] propose the VisDial dataset for the visual dialog task, which formulates multiple image-language interactions via a dialog. Concurrently, H de Veris *et al.* [DVSC+17] propose GuessWhat, a goal-driven dialog dataset for object identification. Different from VQA and goal-driven dialogs, the VisDial answers

are detailed and more human-like. For instance, in Fig. 4.8, the answer is "Can't tell...cloth", while a VQA answer would be "cloth". Therefore, metrics that require exact matching are no longer suitable. Instead, each question is accompanied with 100 candidate answers. Consequently, the metric has been shifted from accuracy to retrieval-based metrics, *e.g.*, MRR and NDCG. Prior works focus on optimizing a single metric [GXT19, JYQ+20, HAR+17, GCK+19]. Differently, Murahari *et al.* [MBPD20] attempt to optimize both metrics with a joint loss. Still, a dedicated single metric model is superior. Instead, we propose principals to ensemble two dedicated models, one for NDCG and one for MRR. Our approach allows most of the MRR and NDCG to be preserved simultaneously.

Visual dialog models: Various approaches were proposed to solve the Visual Dialog task. Most of them focus on dialog history reasoning per interaction. Serban *et al.* [SSL+17] propose history hierarchical encoding.Seo *et al.* [SLHS17] introduce a memory network based on attention, which also addressed co-referential issues. Kottur *et al.* [KMP+18] focus on visual co-reference. Jain *et al.* [JLS18] concatenate representations of all the cues (*e.g.*, image, question, history, and caption) per candidate answer. Zheng *et al.* [ZWQZ19] employ a graph structure learning. Schwartz *et al.* [SYHS19] propose a model, namely Factor Graph Attention (FGA), that lets all entities (*e.g.*, question-words, image-regions, answer-candidate, and caption-words) interact to infer an attention map for each modality. An ensemble of five FGA models achieves the state-of-the-art MRR performance. However, FGA optimizes using the sparse annotations, *i.e.*, the human-derived answer. Murahari *et al.* [MBPD20] recently propose Large-Scale(LS) model, which pre-trains on related vision-language datasets, *e.g.*, Conceptual Captions and Visual Question Answering[SDGS18, AAL+15]. Concurrently, Wang *et al.* [WJL+20] leverage the pretrained BERT language models. Both methods mentioned above finetune using the dense annotation (*i.e.*, human assessment of all the candidates), resulting in a substantial improvement on the NDCG metric. Importantly, Murahari *et al.* find that finetuning a model for NDCG hurts MRR performance. This work demonstrates that re-ranking MRR model (*e.g.*, FGA) and NDCG model (*e.g.*, LS) with simple principles keeps most MRR and NDCG performance.

4.5.2 Two-step Rank Ensemble

The MRR metric depends on a single human-derived answer. Hence, given that this answer is ranked highly, the remaining candidates can be ranked according to the NDCG model. In the following, we describe two steps: (i) the MRR step responsible for preserving the human-derived rank high, and (ii) the NDCG step responsible for ranking the remaining candidates based on the NDCG model.

Setup

We are given a set of dialog questions $\{(q, \mathcal{C}_q)_i\}_{i=1}^{d}$, where d is the dataset size, q is a dialog question, and $\mathcal{C}_q = \{c_{q,j}\}_{j=1}^{100}$ are the corresponding candidates. The MRR metric, *i.e.*, the inverse harmonic mean of rank, is defined as:

$$\text{MRR} = \frac{1}{d}\sum_{i=1}^{d}\frac{1}{r_{\text{mrr}}}, \tag{4.2}$$

where r_{mrr} is the rank of the human response. The DCG, *i.e.*, discounted cumulative gain over the K correct answers, is defined as:

$$\text{DCG}_\text{K} = \sum_{i=1}^{K}\frac{s_i}{\log_2(i+1)}, \tag{4.3}$$

where s_i is a binary score, representing the fraction of annotators that marked the candidate at position as correct. We normalize by the ideal DCG_K score (IDCG_K), *i.e.*, $\text{NDCG}_\text{K} = \frac{\text{DCG}_\text{K}}{\text{IDCG}_\text{K}}$. We denote the set of MRR models as $\mathcal{M} = \{M_1, \ldots, M_{n_m}\}$ where n_m is the number of MRR models. Each MRR model is built by altering the initial conditions. We denote the NDCG model as N. We define an operator $\text{T}(M, n, q)$ that returns the model M's top n responses given a question q. Next, we describe the MRR step that aims to keep the MRR score.

4.5.3 MRR Step

The purpose of the MRR step is to find a minimal candidate set $\mathcal{C}_{\text{MRR},q}$ that is likely to contain the human-derived answer given a question q. We build this set as a union of three sets, as follows:

$$\mathcal{C}_{\text{MRR},q} = \mathcal{T}_q \cup \mathcal{N}_q \cup \mathcal{H}_q, \tag{4.4}$$

where $\mathcal{T}_q$ is a set of first ranked candidates according to MRR models, $\mathcal{N}_q$ is a set of high ranked candidates by both MRR and NDCG models, $\mathcal{H}_q$ is a set of high-certainty candidates agreed by all the MRR models. All sets are conditioned by the question q. In the following, we formally define those sets.

High-certainty answers

One of the most significant signals to be the human-derived answer is being a top MRR-model's answer. However, in many subjective questions, the MRR model is not certain. We found that in those cases, the top answers often varies between different MRR models. Thus, to verify the top candidate's certainty, we require an agreement

of MRR-models. Let q be a dialog question, we define the *high-certainty* set as follows:

$$\mathcal{H}_q = \{c | \, (\forall M \in \mathcal{M}; c \in \mathrm{T}(M, \rho_h, q))\}, \tag{4.5}$$

where $\rho_h \in \mathbb{R}$ is an hyperparameter. Intuitively, a low ρ_h results in higher certainty. We Next, we add the MRR-models' answer at first retrieval.

Top answers

The MRR metric prioritizes the first-ranked answer (see Eq.(4.2)). This property suits the nature of dialog models that reply with a single response. Consequently, we keep the first responses of the MRR models. Let q be a dialog question, the *top-answers* set is defined as:

$$\mathcal{T}_q = \{c | \, (\exists M \in \mathcal{M}; c \in \mathrm{T}(M, \rho_t, q))\}, \tag{4.6}$$

where $\rho_t \in \mathbb{R}$ is an hyperparameter. We note that ρ_t should be low to maintain candidates' certainty. In the next step, we consider top NDCG candidates.

NDCG-agreement answers

When the NDCG model and the MRR model agree that a candidate is likely to be correct, it implies that both the NDCG and MRR metrics gain by ranking this candidate high. Thus, we want to rank it high. We note that the MRR set is ranked first, so we include these candidates in the MRR set. Let q be a dialog question, the *ndcg-agreement* set is defined as:

$$\mathcal{N}_q = \{c | \exists M \in \mathcal{M}; c \in \mathrm{T}(N, \rho_n^n, q) \cap \mathrm{T}(M, \rho_m^n, q)\}, \tag{4.7}$$

where $\rho_n^n, \rho_m^n \in \mathbb{R}$ are hyperparameters that indicate relevancy to NDCG and MRR, respectively. *I.e.*, as ρ_{nn} increases, we may include more relevant candidates according to the NDCG model.

Up until this stage we have built a minimal set $C_{\mathrm{MRR},q}$ that is likely to hold the human-derived answer. In the following we describe how we rank this set.

MRR ranking

Let $r_{M_i,c,q}$ denote the rank according to $M_i \in \mathcal{M}$ of candidate c for a question q. We compute the MRR rank of candidate $c \in \mathcal{C}_{\mathrm{MRR},q}$ via geometric mean: $r_{\mathrm{MRR},c,q} = \prod_{i=1}^{n_m} r_{M_i,c,q}$.

4.5.4 NDCG Step

In this step, we rank the remaining candidates $\mathcal{C}_{\mathrm{NDCG},q} = \mathcal{C}_q \setminus \mathcal{C}_{\mathrm{MRR},q}$. We assume the correct MRR answer is in $\mathcal{C}_{\mathrm{MRR}}$. Thus, we rank the remaining candidates, according

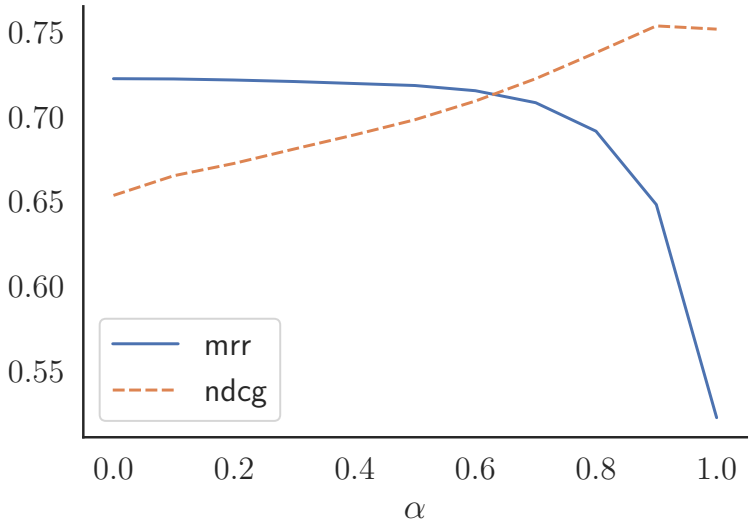


Figure 4.9: Performance of a naïve score ensemble of the MRR model and the NDCG model on the VisDialv1.0 val set. We calibrate the importance of each model with a scalar α.

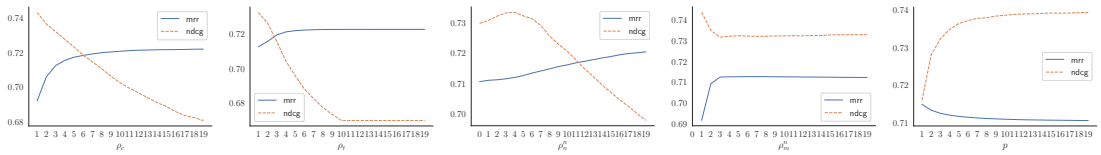


Figure 4.10: MRR and NDCG scores for different hyperparameter values.

to the NDCG model via geometric mean: $r_{\text{NDCG},c,q} = (r_{N,c,q})^p \cdot r_{M,c,q}$, where $M \in \mathcal{M}$ is the most accurate MRR model, and $p \in \mathbb{R}$ is a calibration hyperparameter which controls the trade-off between MRR and NDCG.

To conclude, let q be a dialog question and $\mathcal{C}_q$ the corresponding candidates. We first find $\mathcal{C}_{\text{MRR},q}$, and rank the set according to $r_{\text{MRR},c,q}$. We then rank the remaining candidates, according to $r_{\text{NDCG},c,q}$.

4.6 Results

We show our results on the VisDial v1.0 dataset, where 123,287 images are used for training, 2,000 images for validation, and 8,000 images for testing [DKG+18]. Each image is associated with ten questions, and each question has 100 corresponding answer candidates. We use two MRR models (*i.e.*, $n_m = 2$), FGA [SYHS19] and an ensemble of LS [MBPD20] with FGA. We use LS(CE) as the NDCG model. We set $\rho_h = 3$, $\rho_t = 1$, $\rho_n^n = 5$, $\rho_m^n = 10$, and $p = 3$. We tune these parameters using the validation set.

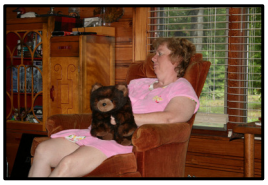
Comparison to state-of-the-art: In Tab. 4.8 we compare our method to naïve ensembles and previous baselines. We first ensemble the LS's output with the FGA's

output. By combining them, we achieve the new MRR state-of-the-art (71.24% *vs.* 69.37%). Next, we build a naïve ensemble of the MRR model and the NDCG model. We do so by adding the MRR ensemble scores (denoted by $\mathcal{S}_M$) and LS(CE) scores (denoted by $\mathcal{S}_N$), as follows: $\alpha \cdot \mathcal{S}_M + (1-\alpha)\mathcal{S}_N$, where $\alpha \in \mathbb{R}$ calibrates the trade-off between MRR and NDCG performance. We show in Fig. 4.9 an analysis of different α values on the validation set. In Tab. 4.8, we report results for $\alpha = 0.8$. Our two-step method outperforms the MRR (70.41% *vs.* 68.78%) and NDCG (72.16% *vs.* 69.22%) metrics, despite lacking the output scores and only requiring rankings.

We also compare our approach to previous baselines. Most methods use the sparse annotations, *i.e.*, the human-derived answer, while MReal-BDAI, VD-BERT, and LS(CE) finetune using the dense annotations. Finetuning with the dense annotations tremendously boosts the NDCG performance but loses MRR performance. The MRR performance decline can be attributed to NDCG being biased toward uncertain answers. We also note that LS leverages large-scale image-text corpora. LS(CE+NSP) optimizes both the dense and sparse annotations but still suffers from a performance drop compared to metric-dedicated LS models, *i.e.*, MRR (63.92% *vs.* 67.50%) and NDCG (68.08% *vs.* 74.47%). Unlike the method mentioned above, our method re-rank the candidates based on two distinct models, with two distinct steps, to keep the human-derived answer high. In doing so, we achieve a good MRR performance (70.41% *vs.* 71.24%), yet notably with limited NDCG drop (72.25% *vs.* 75.35%). This property comes in handy in the recent Visual Dialog challenge, where the winners were picked based on both the NDCG and MRR evaluation metrics. Our method performs well on both metrics simultaneously and won the challenge.

Ablation analysis: The MRR candidate set consists of different subsets. In Tab. 4.9 we show the influence of each of subset independently on the retrieval metrics. Further, omitting a subset harms the performance, *i.e.* each component is essential to preserve both the MRR and NDCG metrics. We also report the average size of the MRR-candidate set, and the validation performance of the MRR model (*i.e.*, 5xFGA) and the NDCG model (*i.e.*, LS(CE)). In addition we provide the results of the MRR ensemble, and the naïve NDCG and MRR ensemble for $\alpha = 0.8$.

In Fig. 4.10, we examine how the NDCG and MRR metrics are affected by modifying one hyperparameter while maintaining the others. On the first figure from the left, we alter ρ_c. The higher ρ_c, we require higher agreement between the MRR models, resulting in higher certainty for elements in the MRR set. Because the MRR models are responsible for the MRR set ranking, an MRR set that is too large hurts the NDCG metric. For the same reason, in the second image from the left, increasing ρ_t, significantly harms the NDCG performance. In the third figure from the left, we show that considering more candidates that both NDCG and MRR models agree upon (*i.e.*, increasing ρ_n^n) helps both metrics' performance. However, adding too many candidates harms the NDCG metric. In the fourth image from the left, we show that the performance remains stable when ρ_m^n is larger than three. Last, on the fifth image from the



how old is the woman?

MRR candidate set

1. she 's probably 60
2. middle aged
3. late UNK ‘s
4. i can not tell
5. unable to tell

Top 4 from the remaining NDCG candidates:

1. ca n't tell
2. looks middle aged
3. mid sixties
4. i ca n't tell

is the cage made of metal ?

MRR candidate set

1. yes
2. metal and wood - looks like a zoo enclosure
3. yes it is
4. it appears to be
5. yes , it is

Top 4 from the remaining NDCG candidates:

1. yes it is ,
2. i think so
3. yep
4. wood

Figure 4.11: An illustration of two visual dialog samples. Each sample includes the MRR candidate set and four answers from the remaining NDCG candidates. We find that the MRR candidate set has more certain answers. We colorize the *high-certainty* candidates ($\mathcal{H}$) with orange, the *NDCG-agreement* candidates ($\mathcal{N}$) with purple, and the *top-answers* subset ($\mathcal{T}$) with red. Note, if a candidate belongs to more than one set, we sketch the colors in the following order: orange→red→purple.

left, we show the effect of changing p, which calibrates the trade-off between MRR and NDCG during the NDCG ranking step.

Qualitative analysis: In Fig. 4.11, we show two sample visual dialogs from test-std. For each sample, we provide the ranked MRR candidate set and the next 4 NDCG candidates. The analysis reveals the answers' ambiguity and that the MRR candidate set mostly consists of certain responses. In addition, we highlight the candidates within each MRR candidate subset with different colors. Additional samples can be found in the appendix.

4.7 Conclusions

We developed a general factor graph based attention mechanism which can operate on any number of utilities. We showed applicability of the proposed attention mechanism on the recently introduced visual dialog dataset and outperformed existing baselines by 1.1% on MRR. Next, we describe a non-parametric method to ensemble the candidate ranks of two strong MRR and NDCG models into a single ranking that excels on both NDCG and MRR. Intuitively, we use the MRR-model for non-ambiguous questions with certain answers. The dense-annotations cue is more applicable in ambiguous questions than the sparse annotations. Thus, in the case of low certainty, our method relies almost entirely on the NDCG model. We hope the proposed principles can guide the

community towards a parametric model that can employ answers' semantics to measure certainty.

Model	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓	NDCG↑
NMN [KMP+18]	58.80	44.15	76.88	86.88	4.81	58.10
NN [ZWQZ19]	61.37	47.33	77.98	87.83	4.57	52.82
CorefNMN [KMP+18]	61.50	47.55	78.10	88.80	4.40	54.70
RvA [NZZ+19]	63.03	49.03	80.40	89.83	4.18	55.59
HACAN [YZZ19]	64.22	50.88	80.63	89.45	4.20	57.17
MReal - BDAI‡ [QNHZ20]	52.62	40.03	68.85	79.15	6.76	74.02
ReDAN [GCK+19]	53.13	41.38	66.07	74.50	8.91	61.86
ReDAN+† [GCK+19]	53.74	42.45	64.68	75.68	6.64	64.47
DualVD [JYQ+20]	63.23	49.25	80.23	89.70	4.11	56.32
DL-61 [GXT19]	62.20	47.90	80.43	89.95	4.17	57.32
DL-61† [GXT19]	63.42	49.30	80.77	90.68	3.97	57.88
DAN [KLZ19]	63.20	49.63	79.75	89.35	4.30	57.59
DAN† [KLZ19]	64.92	51.28	81.60	90.88	3.92	59.36
FGA [SYHS19]	63.75	49.58	80.975	88.55	4.51	52.12
5×FGA† [SYHS19]	69.37	55.65	86.73	94.05	3.14	57.29
LS(CE)*‡ [MBPD20]	50.74	37.95	64.13	80.00	6.28	74.47
LS(CE+NSP)*‡ [MBPD20]	63.92	50.78	79.53	89.60	4.28	68.08
LS* [MBPD20]	67.50	53.85	84.68	93.25	3.32	63.87
VD-BERT*†‡ [WJL+20]	51.17	38.90	62.82	77.98	6.69	**75.35**
5xFGA + LS*†	**71.24**	**58.28**	**87.55**	**94.45**	**2.96**	64.04
5xFGA + LS + LS(CE)*†‡	68.78	55.72	85.02	93.55	3.26	69.22
Ours*†‡	70.41	58.18	83.85	90.83	3.66	72.16
Visual Dialog Challenge 2020 Leaderboard						
LS	68.79	55.20	86.15	93.88	3.12	63.34
VD-BERT	51,84	39.91	63.45	78.56	6.57	75.92
MReaL Lab (3rd Place)	64.12	50.81	80.03	90.92	3.83	75.70
SES-100M (2nd Place)	63.84	55.62	72.20	83.70	5.84	75.86
Ours (1st Place)	70.42	58.59	82.85	88.84	3.96	73.35

Table 4.8: Performance on VisDial v1.0 test-std. (*) denotes the use of external knowledge. (†) indicates ensemble model, and (‡) signifies fine-tuning on the dense annotations. Shown are the MRR, NDCG, the mean rank of the human-derived answer, and the recall at a certain number of retrievals.

$\mathcal{H}$	$\mathcal{T}$	$\mathcal{N}$	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓	NDCG↑	$\lvert\mathcal{C}_{\text{MRR}}\rvert$
✓	✗	✗	70.83	58.87	84.32	90.67	3.67	74.32	2.34
✗	✓	✗	68.63	59.12	79.08	88.53	4.15	74.31	1.12
✗	✗	✓	61.75	51.74	70.35	85.88	4.94	**74.69**	3.97
✗	✓	✓	69.21	59.17	78.68	88.53	4.11	74.33	4.27
✓	✗	✓	71.15	59.11	84.38	90.67	3.65	73.29	4.87
✓	✓	✗	71.06	59.15	84.49	90.78	3.64	72.98	2.39
✓	✓	✓	**71.26**	**59.18**	**84.62**	**90.78**	**3.62**	73.23	4.91
LS(CE)			52.21	39.92	65.04	80.62	6.16	75.24	-
LS			69.00	55.80	85.36	93.13	3.35	64.89	-
5xFGA			69.38	56.17	86.15	92.95	3.32	58.68	-
5xFGA + LS			72.25	59.20	88.55	94.52	2.84	65.34	-
5xFGA + LS + LS(CE)			69.14	56.79	84.24	92.37	3.43	73.78	-

Table 4.9: MRR candidates set ablation analysis. Performance reported on the VisDialv1.0 val set.

Chapter 5

Visual Storytelling

Visual Storytelling (VST) [PK15, HFM$^+$16] – the task of generating a story based on a sequence of images – goes beyond a basic understanding of visual scenes and can be applied in many real-world scenarios, *e.g.*, to support the visually impaired. Moreover, VST reflects on the creative ability of intelligent systems. Although similar in concept to other cognitive tasks such as image captioning and visual question answering, VST differs as it requires to reason over a *sequence* of images while simultaneously ensuring coherence across multiple generated sentences. To achieve this, VST methods need to address two major challenges: the first is visual and relates to grounding the story's text to the images. The second is linguistic and relates to the quality of the story. Both challenges can be described in terms of coherency: the story should be coherent by itself, and coherent with the images.

Prior research on VST started to address the aforementioned challenges. Early works expand captioning [VTBE14, XBK$^+$15, CZ15], focusing sentence generation mainly on the current image [GRP18, WMZ$^+$19]. This limits the ability to incorporate complex semantic information, which is necessary for visual reasoning. Prior work also makes limited use of temporal dependence and history, *e.g.*, sentences that have already been generated are not used. Consequently, the output lacks narrative consistency and is prone to linguistic errors such as *repetitiveness* and *incoherence* [MP19]. To mitigate these issues, later works strive to generate more meaningful stories via adversarial and reinforcement learning [WCfWW18, HG $^+$18], which remain delicate to train.

Importantly, images are not independent. For example, if the first image in a sequence shows a protest, the model may want to focus on signs in later images. Conversely, if the last image shows a ring on a finger, then the model should pay attention to wedding-related objects and activities in the preceding images. This is important for VST because sentences are created per image but are part of a story. Hence, objects that the model is focusing on in one image should be conditioned on the selection in other images.

To do this we develop a novel model which (1) implicitly reasons over objects, activities, and their temporal dependencies in each image; and which (2) improves the

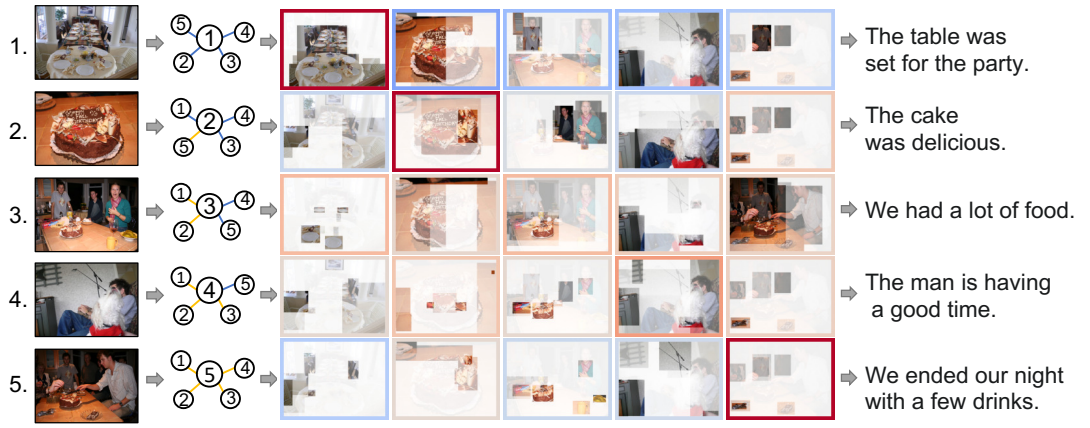


Figure 5.1: We propose Ordered Image Attention (OIA) to form the structure of a sentence and to encourage coherency. Each row shows the spatial attention of the five images created when generating a specific sentence. We find important objects by collecting directional interactions. The relative order to the sentence-corresponding image determines the connection type, illustrated as the blue and orange edges for preceding and proceeding connections. The attended images' border indicates the image attention importance formed by the Image-Sentence Attention (ISA). *E.g.*, red indicates a high attention score, meaning the image is essential for generating that sentence. Our model performs this step for all five images in parallel, creating a total of 25 spatial attention maps, that are fed into the decoder to create the sentences in order.

coherency of the narrative. To reason over objects and activities in each image, *i.e.*, to understand their dependencies and their temporal ordering, we introduce *ordered image attention* (OIA). As illustrated in Fig. 5.1, for each image, OIA accumulates representation information from objects detected within the corresponding image into an attended image representation. Importantly, accumulation factors depend on whether the image precedes or succeeds the image for which we are currently generating the sentence, which permits to establish an order. The attended image representations are subsequently summarized into a context embedding via an Image-Sentence Attention (ISA) unit, before being used for sentence decoding.

In addition, to alleviate common linguistic mistakes like repetitiveness and to promote coherence in the story, we incorporate information from the story generated up to the current sentence into the sentence generation decoder. Specifically, the decoding strategy decays the probability of a word if it has already been used in the story. The decoder also maintains a separate prior over the output probability distribution, independent from the language generation unit. This prior is based on counts of the words that were already predicted in the story. Both the prior, and the Recurrent Neural Net (RNN) decoder output are combined to predict the next word in the sentence.

Empirical results on the challenging VIST dataset [HFM+16] demonstrate that the proposed method generates stories with an improved narrative quality. The method outperforms prior state-of-the-art by 1% on the METEOR score. Examples of stories generated by the approach are shown in Fig. 5.1. We also present a user study demonstrating the advantage of the model in terms of coherency.

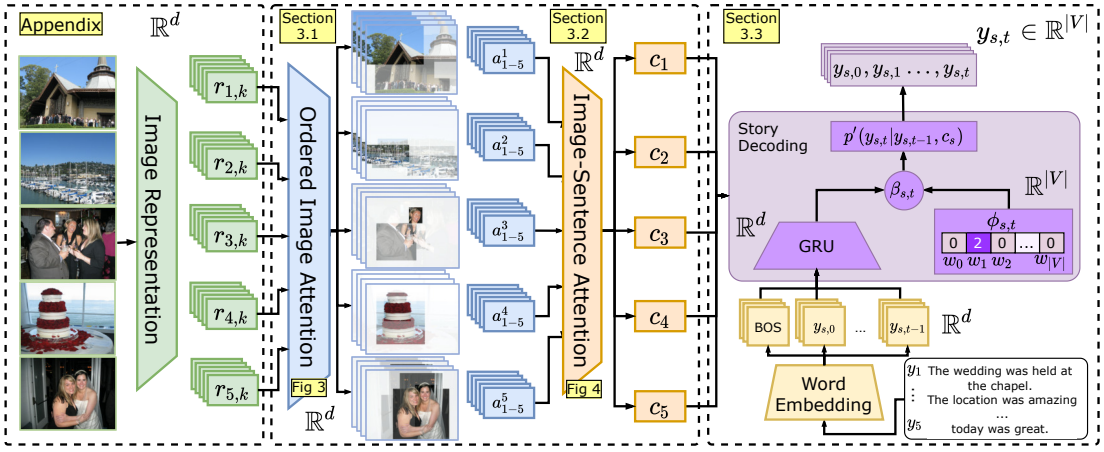


Figure 5.2: Our architecture for Visual Storytelling synthesis.

5.1 Related Work

Vision+Language has been an active area of research for many years, addressing tasks such as image/video captioning, paragraph generation, and visual question answering. We briefly review those related areas in the following. **Image Captioning** Bernard *et al.* [BSF+03] first explored annotating images with text. Since then, image/video captioning has seen a surge of research activity. Initial work utilized pre-trained image embeddings from a CNN network. The success of attention mechanisms for language translation quickly transferred to image captioning as well [XBK+15]. Later work leveraged advances in object detection and proposed a bottom-up/top-down attention approach to attend to specific objects in the image instead of fixed spatial regions [AHB+18]. Different from image captioning, for visual storytelling, both story coherency and visual grounding are important. **Multimodal Attention** Multimodal problems are characterized by input data that comes from different domains, *e.g.*, visual and linguistic. This raises two challenges: 1) how to model interactions between different domains, and 2) how to manage the large input data. Considering those challenges, attention has been a prominent tool as it models interactions to select the important elements. In early work, Xu *et al.* [XBK+15] used interaction-based attention with the image at each caption generation step. This idea was later extended to visual question answering [XS16]. To imitate multi-step reasoning, Yang *et al.* [YHG+16] stacked attention modules sequentially. Later, many works concentrated on better vector-fusion modeling [FPY+16, KOL+17, BYCCT17, YYX+18]. Importantly, Lu *et al.* [LYBP16] suggested attending to the visual and textual modalities separately. Afterward, Kim *et al.* [KJZ18] proposed a bilinear module that efficiently generates attention for every pair. Following Lu *et al.* [LYBP16], Schwartz *et al.* [SSH17, SYHS19] suggested a general framework that extends attention to any number of utilities via local and interaction-based factors. We improve upon those ideas by suggesting an ordered attention. This ensures that interaction modeling is affected by the image position in a

sequence.

Visual Storytelling Huang *et al.* [HFM+16] introduced the Visual Storytelling task. Initially, Gonzalez *et al.* [GRP18] adapted work by Vinyals *et al.* [VTBE14] used for captioning. Kim *et al.* [KHS+18] presented a Seq2Seq [SVL14] approach with a decoding sampling strategy aimed to reduce the amount of repetition based on a word list. We improve their strategy by using a data-driven approach, penalizing each word differently based on its average counts. Wang *et al.* [WCfWW18] employ adversarial learning to improve output stories. Huang *et al.* [HG +18] utilize a reinforcement learning (RL) approach based on inter-image relations. Later works by Li *et al.* [LST+19] and Zhang *et al.* [ZHS20] rely on preprocessing the data to better ground visual elements to the text while Yang *et al.* [YLC+19] and Hsu *et al.* [HCH+20] enrich the data with an external word common-sense knowledge graph. Our approach captures inter-image relations via ordered attention and is trained in an end-to-end manner alleviating the computational drawbacks of preprocessing or RL. Recently, state-of-the-art results were obtained by generating scene graphs for each image in the sequence [WWL+19]. Conversely, our image representations are dependant on all the images in the sequence.

5.2 Method

The goal of visual storytelling is to generate a story, composed of N *ordered* sentences $\{y_s|1 \leq s \leq N\}$, given an *ordered* sequence of images $I = \{I_s|1 \leq s \leq N\}$. Each sentence $y_s = (y_{s,0}, \ldots, y_{s,t}, \ldots)$ is composed of words $y_{s,t} \in \mathcal{Y}$ from vocabulary $\mathcal{Y}$.

The order in which the images are given is essential as it defines the plot line of the story. The story should be focused, *i.e.*, each sentence should be related to the remainder of the story. Importantly, the sentences should form a coherent body of text describing the set of images, and not only a set of related information. For instance, the story *"The church was beautiful. The bride and groom walk down the aisle. The cake was amazing."* is less coherent than: *"We went to the church for the wedding today. The bride and groom were excited for the day. Both cut the cake together."*

Overview: To address this challenge, we develop the model illustrated in Fig. 5.2. It infers conditional probabilities $p'(y_{s,t}|y_{s,t-1}, c_s)$ for the t-th word $y_{s,t} \in \mathcal{Y}$ in sentence y_s given the previous word $y_{s,t-1}$ and the context embedding c_s for sentence s. The context embedding c_s summarizes region representations $r_{i,k}$ of all K object regions across all N images I_i ($i \in [1, N]$, $k \in [1, K]$) via Ordered Image Attention (OIA) (Sec. 5.2.1) and Image-Sentence Attention (ISA) (Sec. 5.2.2). Specifically, when generating sentence s, OIA computes an attended image representation a_i^s for every image I_i by attending to the K region representations $r_{i,k}$ (Sec. 5.2.1). These attended image representations a_i^s are subsequently summarized into the context embedding c_s via an image-sentence attention (Sec. 5.2.2).

Below we first discuss computation of the attended image representation a_i^s (Sec. 5.2.1), before detailing computation of the context embedding c_s (Sec. 5.2.2) and computation

of the conditional probabilities $p'(y_{s,t}|y_{s,t-1}, c_s)$ (Sec. 5.2.3).

5.2.1 Ordered Image Attention (OIA)

Ordered Image Attention (OIA) is designed to 1) form a structure across ordered images and to 2) select the relevant objects per image. For this we model preceding and proceeding interactions separately using different attention factors. We calibrate each factor's importance with trainable scalars, which forms a graph of dependencies between the images. For each sequence of N images, the model infers a total of N^2 attention maps, one per image for each sentence. We detail this module next.

Attention Belief

For each image $I_i = \{r_{i,1}, \dots r_{i,K}\}$ we consider a set of K regions, represented by their feature vectors $r_{i,k} \in \mathbb{R}^d$, where d is the objects' embedding dimension. Suppose we are currently generating sentence y_s ($1 \leq s \leq N$). To do this we first compute an attended image representation a_i^s as follows

$$a_i^s = \sum_{k=1}^{K} b_{i,k}^s r_{i,k}, \tag{5.1}$$

where $b_{i,k}^s \geq 0$ is the attention belief highlighting the importance of the k-th object in the i-th image when generating the s-th sentence. Importantly, for every image I_i we require $b_{i,k}^s$ to be a valid probability distribution, *i.e.*, we also enforce $\sum_{k=1}^{K} b_{i,k}^s = 1$ $\forall s, i$.

The object attention belief $b_{i,k}^s$ is dependent on all the input data, *i.e.*, other objects and images. To avoid complex computation, we factorize the belief $b_{i,k}^s$ into two pairwise dependencies that preserve the order, and a local term. For the pairwise terms we use $\mu_{j\to i}^{\text{bwd}}$, which is a message from a preceding image I_j, or $\mu_{j\to i}^{\text{fwd}}$, which is a message from a subsequent image I_j. We also use $\mu_{i\to i}$ for self-messages. Additionally, we include a local factor $\Psi_i(r_{i,k})$ that considers the object representation. Unlike the messages mentioned before, the local factor does not rely on interactions with other objects. We aggregate all the messages along with the local factor as illustrated in Fig. 5.3. For normalization we employ a softmax.

Formally we compute the attention belief $b_{i,k}^s$ by distinguishing three cases. If $i = s$ we have

$$\begin{aligned} b_{i,k}^s \quad \propto \quad & \exp(\alpha_i^s \Psi_i(r_{i,k}) + \alpha_{i,i}^s \mu_{i\to i}(r_{i,k}) + \\ & \sum_{j<i} \alpha_{i,j}^s \mu_{j\to i}^{\text{bwd}}(r_{i,k}) + \sum_{j>i} \alpha_{i,j}^s \mu_{j\to i}^{\text{fwd}}(r_{i,k})). \end{aligned} \tag{5.2}$$

If $i < s$ we use

$$\begin{aligned} b^s_{i,k} \quad &\propto \quad \exp(\alpha^s_i \Psi_i(r_{i,k}) + \\ &\qquad \alpha^s_{i,i} \mu_{i \to i}(r_{i,k}) + \alpha^s_{i,s} \mu^{\text{bwd}}_{s \to i}(r_{i,k})). \end{aligned} \tag{5.3}$$

If $i > s$ we obtain

$$\begin{aligned} b^s_{i,k} \quad &\propto \quad \exp(\alpha^s_i \Psi_i(r_{i,k}) + \\ &\qquad \alpha^s_{i,i} \mu_{i \to i}(r_{i,k}) + \alpha^s_{i,s} \mu^{\text{fwd}}_{s \to i}(r_{i,k})). \end{aligned} \tag{5.4}$$

In all three cases $\alpha^s_i, \alpha^s_{i,i}, \alpha^s_{i,j} \in \mathbb{R}$ are scalars used to calibrate the importance of different messages for a given sentence. These scalars form a dependency structure between images for each of the generated sentence indices. Intuitively, when we generate the first sentence, the attention belief might depend more on subsequent images, to correctly identify the story event, *e.g.*, a wedding, a parade, *etc.* Thus, the scalars will promote interaction with later images. An analysis of these scalars is provided in the appendix. Next, we define the different types of messages.

Pairwise Messages and Factors

A message aggregates interaction scores from an image to an object. The three messages $\mu^{\text{bwd}}_{j \to i}, \mu^{\text{fwd}}_{j \to i}$ and $\mu_{i \to i}(r_{i,k})$ are computed as follows:

$$\mu^{\text{bwd}}_{j \to i}(r_{i,k}) = \sum_{k'=1}^{K} \Psi_{\text{bwd}}(r_{i,k}, r_{j,k'}), \tag{5.5}$$

$$\mu^{\text{fwd}}_{j \to i}(r_{i,k}) = \sum_{k'=1}^{K} \Psi_{\text{fwd}}(r_{i,k}, r_{j,k'}), \text{ and} \tag{5.6}$$

$$\mu_{i \to i}(r_{i,k}) = \sum_{k'=1}^{K} \Psi_{i,i}(r_{i,k}, r_{i,k'}). \tag{5.7}$$

Importantly, these messages collect three different types of order-dependent interaction factors: (1) A backward image interaction, namely $\Psi_{\text{bwd}}(r_{i,k}, r_{j,k'})$. This interaction models relations to the preceding j-th image in the sequence. (2) A forward image interaction, namely $\Psi_{\text{fwd}}(r_{i,k}, r_{j,k'})$. This interaction models relations to the subsequent j-th image in the sequence. (3) The self interaction factor, namely $\Psi_{i,i}(r_{i,k}, r_{i,k'})$, which takes into account interactions between objects within the image. We formally define the different factors next.

Interaction factors: A commonly used practice to capture interactions across attention mechanisms is to first embed the elements into a joint Euclidean space followed by a dot-product [VSP+17, SSH17, GJY+19, SYHS19]. While we follow the same practice, we define three types of interaction factors to preserve the order. Consider two objects,

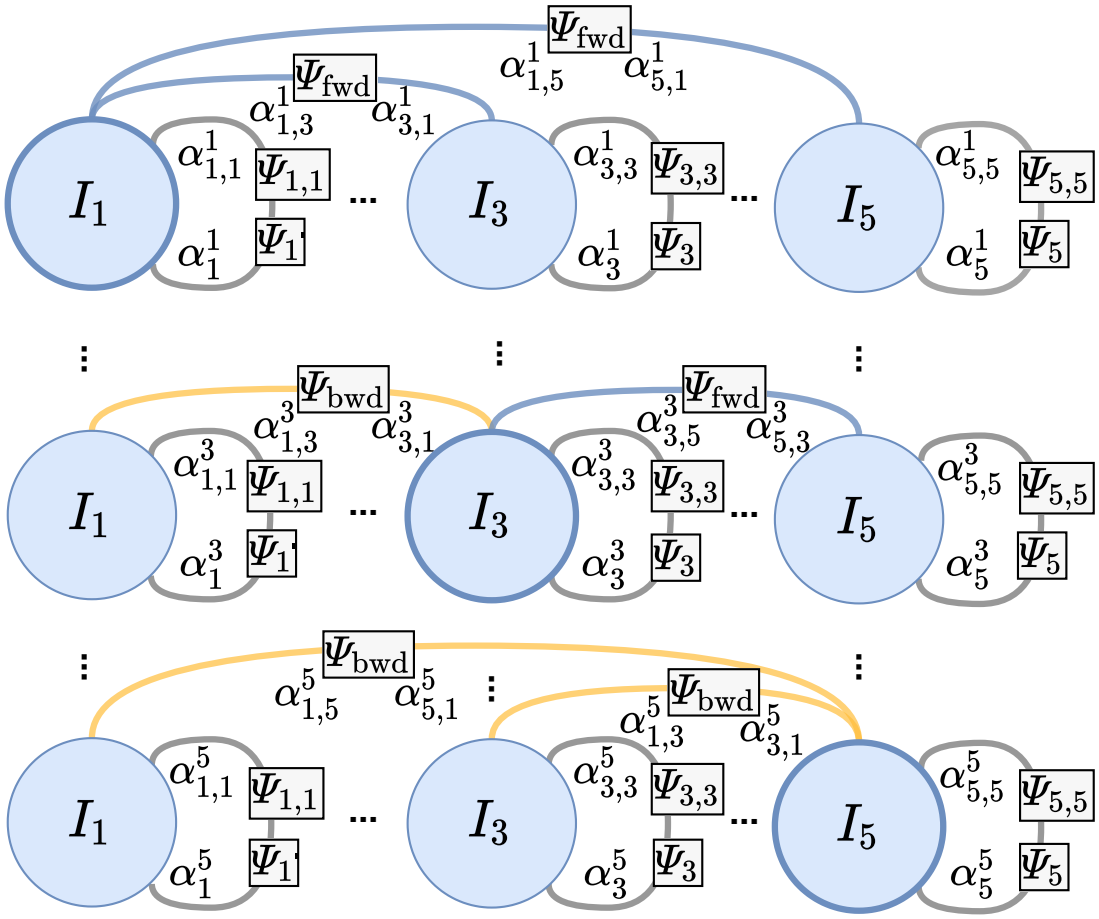


Figure 5.3: Illustration of Ordered Image Attention. Each node represents an image attention belief. For each sentence, we connect all the images with the sentence-corresponding image. The relative position to this image determines whether the connection is modeled with the Ψ_{bwd} factor (for preceding images) or the Ψ_{fwd} factor (for subsequent images). We infer the attention belief by collecting interactions and local object information within the image. We use scalars to calibrate the importance of each factor. In total, we generate 25 attention maps, one per image for every sentence.

$r_{i,k} \in I_i$ from the sentence-corresponding image and $r_{j,k'} \in I_j$ from the interacting image. We describe three types of interactions: for interactions with subsequent images (*i.e.*, $j > i$) we use

$$\Psi_{\text{fwd}}(r_{i,k}, r_{j,k'}) = \left(\frac{L_{\text{fwd}} r_{i,k}}{\|L_{\text{fwd}} r_{i,k}\|_2}\right)^\top \left(\frac{R_{\text{fwd}} r_{j,k'}}{\|R_{\text{fwd}} r_{j,k'}\|_2}\right). \tag{5.8}$$

For interactions with preceding images (*i.e.*, $j < i$) we use

$$\Psi_{\text{bwd}}(r_{i,k}, r_{j,k'}) = \left(\frac{L_{\text{bwd}} r_{i,k}}{\|L_{\text{bwd}} r_{i,k}\|_2}\right)^\top \left(\frac{R_{\text{bwd}} r_{j,k'}}{\|R_{\text{bwd}} r_{j,k'}\|_2}\right). \tag{5.9}$$

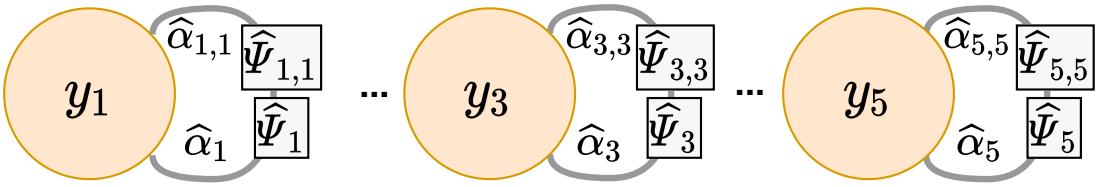


Figure 5.4: Illustration of ISA. The attention selects the attended image representation per sentence. We model interactions between attended images of the same sentence to compute each image's importance. Note, each node represents a sentence attention belief over the attended images.

For interactions within the image (*i.e.*, $j = i$) we have

$$\Psi_{i,i}(r_{i,k}, r_{i,k'}) = \left(\frac{L_{i,i} r_{i,k}}{\|L_{i,i} r_{i,k}\|_2}\right)^\top \left(\frac{R_{i,i} r_{i,k'}}{\|R_{i,i} r_{i,k'}\|_2}\right). \tag{5.10}$$

Note, $L_{\text{fwd}}, R_{\text{fwd}}, L_{\text{bwd}}, R_{\text{bwd}}, L_{i,i}, R_{i,i} \in \mathbb{R}^{d \times d}$ are trainable shared weights across the entire image sequence. Also, the object from the sentence-corresponding image will always be on the left side of the factor equation. Thus, the factor embeddings preserve the order.

Local factor: Differently from the previous interactions the following factor captures how important an object is based solely on the object representation. Given an object $r_{i,k} \in I_i$, we define the local factor as,

$$\Psi_i(r_{i,k}) = v^\top \operatorname{ReLU}(V r_{i,k}), \tag{5.11}$$

where $v \in \mathbb{R}^d, V \in \mathbb{R}^{d \times d}$ are trainable weights.

5.2.2 Image-Sentence Attention (ISA)

In a next step we summarize the attended image representations a_i^s produced by OIA to compute the context embedding c_s for the sentence s that we wish to generate. For this we use the Image-Sentence Attention (ISA) unit. It picks the relevant image context for generating the specific sentence. Formally we obtain the context embedding via

$$c_s = \sum_{i=1}^{N} \hat{b}_{s,i} a_i^s, \tag{5.12}$$

where attention factors

$$\hat{b}_{s,i} \propto \exp\left(\hat{\alpha}_s \hat{\Psi}_i(a_i^s) + \hat{\alpha}_{s,s} \hat{\mu}_{s \to s}(a_i^s)\right), \tag{5.13}$$

and where $\hat{\alpha}_s, \hat{\alpha}_{s,s} \in \mathbb{R}$ are scalars. To avoid spurious correlations between sentences, we consider only self interactions and a local factor. This is illustrated in Fig. 5.4. The

self-message of the attended image representation a_i^s is

$$\hat{\mu}_{s\to s}(a_i^s) = \sum_{j=1}^{N} \hat{\Psi}(a_i^s, a_j^s). \tag{5.14}$$

Finally, the self and local factors are defined with a different set of weights following Eq. (5.10) and Eq. (5.11) respectively.

5.2.3 Story Decoding

The goal at each timestep of decoding is to compute the conditional probability $p(y_{s,t}|y_{s,t-1}, c_s)$ where $y_{s,t} \in \mathcal{Y}$ is the t-th word in sentence y_s, $\mathcal{Y}$ is the vocabulary and c_s is the context embedding detailed in Sec. 5.2.2. For this we use a GRU recurrent unit, tasked with generating probabilities over the vocabulary conditioned on the context embedding c_s and the previously generated token $y_{s,t-1}$:

$$\begin{aligned} p(y_{s,t} = w|y_{s,t-1}, c_s) \propto \exp(&\beta_{s,t} \cdot g_w(y_{s,t-1}, h_{s,t-1}, c_s) \\ &+(1-\beta_{s,t}) \cdot f_w(\phi_{s,t})), \end{aligned} \tag{5.15}$$

where g_w is the output of a GRU unit for the word w. We set the GRU hidden dimension to d. $h_{s,t-1} \in \mathbb{R}^d$ is the hidden state at timestep $t-1$ for sentence s. $f : \mathbb{R}^{|\mathcal{Y}|} \to \mathbb{R}^{|\mathcal{Y}|}$ is a learned prior over the vocabulary based on a bag-of-words prior histogram $\phi_{s,t}$, which we describe in the next paragraph. The purpose of f is to reduce text repetitions. f_w denotes the value of f for a word w. We also incorporate a calibration gate $\beta_{s,t} : \mathbb{R}^d \to [0, 1]$ for functions f and g using

$$\beta_{s,t} = \sigma\left(v_\beta^\top \tanh(G_g h_{s,t} + G_f W_1(\phi_{s,t}))\right). \tag{5.16}$$

Here, $G_g \in \mathbb{R}^{d\times d}$ and $G_f \in \mathbb{R}^{\gamma\times d}$ are trained projections of the GRU hidden state and the bottleneck layer respectively, $v_\beta \in \mathbb{R}^d$ are learned weights and σ is the sigmoid function. W_1 is obtained from the prior as discussed next.

Bag-of-words (BOW) prior: Remembering history during storytelling permits to stay on topic and advance the story in the desired direction. Although quite intuitive, mimicking this ability is not trivial. *E.g.*, most approaches for VST generate all the sentences in parallel. Converting the parallel sentence generation into a sequential one implies a major computational overhead during training.

To address this, we propose a simple yet effective learnable framework that does not require sequential training while still exploiting information found in prior sentences. The history is represented via a bag-of-words histogram $\phi_{s,t}$, which includes all words that have been used until timestep t for the s-th sentence. During training, we initialize $\phi_{s,t=0}$ with the ground truth history counts found in the previous $s-1$ sentences. We update the statistics at each timestep with the predicted word $y_{s',t}$ for $s' < s$,

Method	M	B-1	B-2	B-3	B-4	R	C	Img Feat
seq2seq [HFM+16]	31.4	-	-	-	3.5	-	6.84	FC
h-attn-rank [YBB17]	33.9	-	29.8	-	-	29.8	7.4	FC
Contextualize, Show & Tell [GRP18]	34.4	60.1	36.5	21.1	12.7	29.2	7.1	FC
AREL [WCfWW18]	35.0	63.8	39.1	23.2	14.1	29.5	9.4	FC
KnowledgeableStoryteller [YLC+19]	35.2	66.4	39.2	23.1	12.8	29.9	**12.1**	FC
HSRL [HG +18]	35.2	-	-	-	12.3	29.5	8.4	Spatial
StoryAnchor [ZHS20]	35.5	65.1	40.0	23.4	14.0	30.0	9.9	FC
SGVST [WWL+19]	35.8	65.1	40.1	23.8	14.7	29.9	9.8	F-RCNN
Ours - ResNet	36.3	66.3	41.5	23.7	14.5	30.0	9.8	Spatial
Ours - Full	**36.8**±0.1	**68.4**±0.7	**42.7**±0.3	**25.2**±0.2	**15.3**±0.2	**30.2**±0.1	10.1±0.2	F-RCNN

Table 5.1: Quantitative results on the VIST dataset for METEOR, BLEU-1...4, ROUGE-L and CIDEr. The primary metric is METEOR. The 'Img Feat' column describes the pretrained image features. All models utilize a ResNet [HZRS15a] backbone except CS&T which employs an Inception v3 model [SVI+15]. FC and Spatial refer to features extracted from the penultimate layer and the preceding one accordingly. F-RCNN are bottom up features [AHB+18].

and produce the next state of the counter $\phi_{s,t+1}$. At inference we generate sentences sequentially and update $\phi_{s,t}$ with the predicted words. $\phi_{s,t}$ is fed through a shallow bottleneck network to obtain the prior f, composed of two layers $W_1 \in \mathbb{R}^{|\mathcal{Y}|\times\gamma}$ and $W_2 \in \mathbb{R}^{\gamma\times|\mathcal{Y}|}$ without activation, where γ is the bottleneck dimension:

$$f(\phi_{s,t}) = W_2(W_1(\phi_{s,t})). \tag{5.17}$$

Also note the use of $W_1(\phi_{s,t})$ in the gate (Eq. (5.16)).

Intra-repetition regularization: To regularize intra-repetitions, we decay the probability of previously used words during sentence generation. A critical aspect of this approach is to exclude words that appear frequently in the language (*e.g.*, was, were, am). For this we pre-process the training set to calculate the average story frequency $\rho(w)$ of a word w via $\rho(w) = \frac{\#\text{ appearances of word } w}{\#\text{ stories } w \text{ was used}}$. The final count for word w at timestep t is calculated as $\phi'_{s,t}(w) = \max[0, (\phi_{s,t}(w) - \rho(w) + 1)]$. Intuitively, a word will not be penalized before it is used more than the prior belief average $\rho(w)$. The final probability for word w being used is given by

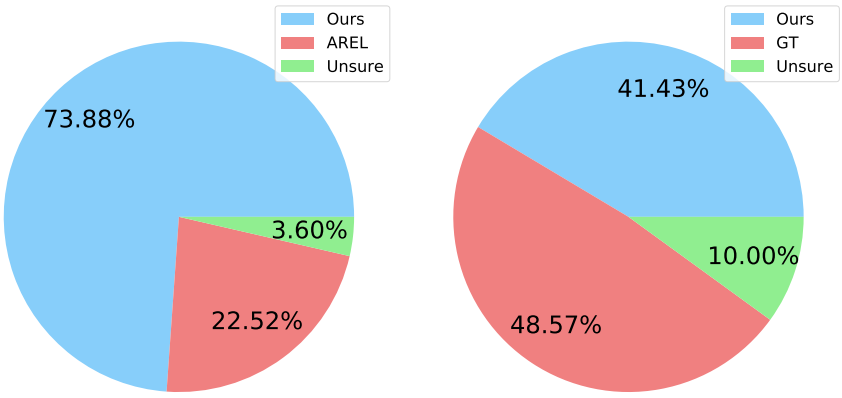
$$p'(y_{s,t} = w|y_{s,t-1}, c_s) = \frac{p(y_{s,t} = w|y_{s,t-1}, c_s)}{\pi \cdot \phi'_{s,t}(w) + 1}, \tag{5.18}$$

where $\pi \geq 0$ is a constant hyper-parameter. A penalty of 2 proved to work best on the validation set.

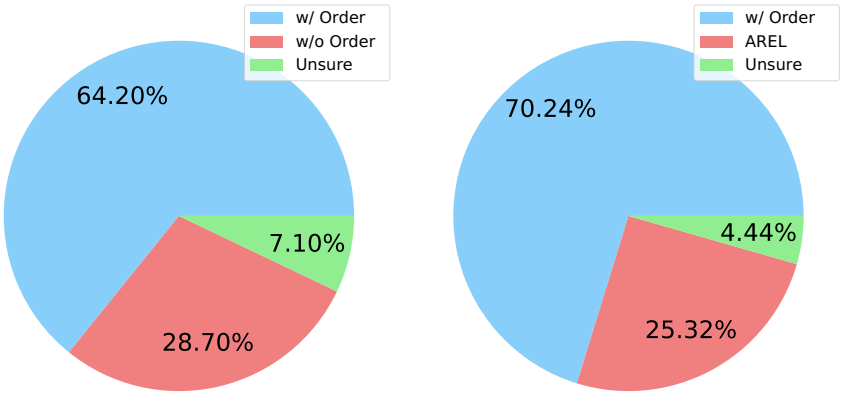
5.3 Results

5.3.1 Training Setup

Dataset: To train and test the model we use the VIST dataset [HFM+16]. This dataset is composed of stories. Each story has 5 images and $N = 5$ corresponding



(a) Human-like property comparison.



(b) Coherence property comparison.

Figure 5.5: Human evaluation to compare human-like and coherency properties.

sentences. All images were collected from Flickr albums. Sequences of images belong to the same album. Each image sequence is annotated with 5 ground-truth reference stories. On average, around 2.5 stories are based on the images, and the rest are rewrites. The overall numbers are 40,098 training stories, 4,988 validation stories, and 5,050 test stories.

Training: We extracted the image features using a pre-trained F-RCNN model with a ResNet152 backbone [HZRS15a, RHGS15, AHB+18]. We set the number of extracted objects $K = 36$. Bounding box coordinates were normalized between 0 and 1. Words that appear less than 3 times in the training set are represented by an $<UNK>$ token. The vocabulary size is 12,210 words. Word representations were initialized using GloVe embeddings [PSM14]. We set the decay parameter $\pi = 2$ and the image representation dimension $d = 512$. We set the dropout parameter to 0.3. We use cross-entropy loss to maximize likelihood of ground-truth stories. At decoding time we employ a beam search algorithm, with beam width set to 3. We use Adam [KB14] optimizer with a learning-rate of 0.0004, which is decayed by a factor of 0.8 if the validation score

Model	METEOR	B-4	#Params
w/o OIA	36.0	14.1	11M
w/o ISA	35.9	14.2	11M
w/o attention	35.8	13.6	11M
no-direction	36.1	14.5	12M
w/o rep. regularization	36.0	14.2	13M
w/o count norm	36.1	14.6	13M
w/o BOW prior	36.2	14.5	13M
Full model	**36.8**	**15.3**	13M

Table 5.2: Components ablation analysis.

Model		Text Rep.	Sent. Rep.
AREL [WCfWW18]		0.16	0.4
BOG prior	**Intra-repetition reg.**		
No	No	0.14	0.33
Yes	No	0.10	0.18
No	Yes	0.04	0.04
Yes	Yes	**0.008**	**0.0**

Table 5.3: Story generation ablation analysis.

Model			Metric						
Local	**Self**	**Directional**	**R**	**C**	**B-1**	**B-2**	**B-3**	**B-4**	**M**
×	✓	✓	30.0	9.3	67.4	42.4	24.2	14.5	36.2
✓	×	✓	29.8	9.2	67.8	42.3	24.2	14.4	36.0
✓	✓	×	29.9	8.5	67.6	42.2	24.0	14.2	35.9
✓	✓	✓	**30.2**	**10.1**	**68.4**	**42.7**	**25.2**	**15.3**	**36.8**

Table 5.4: Factor ablation analysis.

(METEOR) does not improve after 4 epochs. The total amount of trainable parameters is 13,092,194. Training converges after ∼20 epochs. Each epoch needs 20 minutes on an Nvidia V100 GPU.

5.3.2 Quantitative Analysis

Evaluation metrics: As suggested by the creators of VIST [HFM+16], METEOR [BL05] correlates best with human judgement. Following their example, we use METEOR as the primary metric. We also compute BLEU [PRWZ01], ROUGE [Lin04], and CIDEr [VZP14] and compare to prior work where available. For evaluation we use the evaluation script of Yu *et al.* [YBB17][1].

Comparison to state-of-the-art: In Tab. 5.1 we compare the method to recent baselines. Early methods did not take into account visual-spatial information, which harms the performance (*e.g.*, 35.5% *vs.* 36.8% on METEOR) [HFM+16, WCfWW18, GRP18]. Wang *et al.* [WWL+19] utilize image representations similar to our approach but do not consider relations between different images, resulting in a 1% drop on METEOR,

[1] `http://github.com/lichengunc/vist_eval` - Codebase for commonly used evaluation scripts.

showing that ordered structure encoding with OIA is beneficial. SGVST and StoryAnchor [YBB17, ZHS20] use different methods for mapping the image sequence to distinct topics. Differently, our approach is trained end-to-end. Finally, Yang *et al.* [YLC+19] utilize an external commonsense dataset to enrich the input. Their CIDEr score is significantly higher, yet this improvement does not translate to all other metrics. The approach improves upon the current state-of-the-art by a margin (36.8% *vs.* 35.8% on METEOR). Note, the ROUGE-L metric is based on finding the longest subsequence matched to human generated stories. However, this score is almost identical for all prior works, indicating that this metric doesn't capture story generation improvements. We also report the performance with spatial ResNet152 features [HZRS15a], which outperforms the state-of-the-art as well. This shows that the method is stable irrespective of image features.

Ablation study: In Tab. 5.2 we show the importance of different components via an ablation study. In 'w/o OIA,' we replace the OIA module (Sec. 5.2.1) with simple averaging of the K object representations of image I_i, resulting in a 0.8% drop on METEOR. Similarly, in 'w/o ISA,' we replace the ISA unit (Sec. 5.2.2) with averaging, leading to a 0.9% drop on METEOR. In 'w/o attention,' we removed both OIA and ISA, which dropped the METEOR score to 35.8%. For the method referred to as 'no-direction,' we use the same factor for preceding and proceeding interaction (*i.e.*, $L_{\text{bwd}} = L_{\text{fwd}}$ and $R_{\text{bwd}} = R_{\text{fwd}}$). Here, METEOR results drop by 0.7%. Hence, ordered interactions are beneficial. Next, we assess the decoding components (Sec. 5.2.3). We first remove the intra-repetition regularization (*i.e.*, $\rho(w)$), which causes METEOR score to drop by 0.8%. Removing the popular words count ($\phi'_{s,t}$), results in a 0.7% drop on METEOR. The METEOR score drops by 0.6% when we remove the BOW prior. Next, we evaluate the effect of the decoding strategy for reducing repetitions directly.

In Tab. 5.3, we show the ability to reduce repetitions. As proposed by Bertoldi *et al.* [BCF13], text repetitiveness is measured by the repetition rate of non-singleton n-grams within each story. In our experiment, we use up to 4-grams. The use of intra-repetition regularization reduces text repetition (0.14 to 0.04). Combined with the trainable bag-of-words prior module, we further improve this measure (0.008 *vs.* 0.14). We also report sentence repetitiveness, *i.e.*, the average number of repeated sentences in a story.

In Tab. 5.4 we show an ablation of the different factors. We found that each factor contributes to the model's performance, and the directional factors (*i.e.*, Ψ_{fwd} and Ψ_{bwd}) have the biggest impact.

5.3.3 Human Evaluation

The subjective nature of the VST task calls for a human evaluation. We use a sample of 150 image sequences and test different story qualities by asking 3 MTurk annotators to rank or compare them to other methods. We compare our results to the AREL baseline

AREL	The kids had a great time at the pool. The little boy was excited to see the kids. We had a great time at the park. We had a great time at the pool. We had a great time at the park.
No History	The kids had a great time at the beach. The baby was happy to see the baby. We had a great time at the park. The had a great time at the pool. We had a great time at the park.
With History	**The family went to the pool. The baby was very happy. The kids had a great time. The kids played in the pool. The little girl is having a good time.**

Figure 5.6: An illustration of an image sequence along with three different stories generated by: (1) AREL baseline [WCfWW18], (2) No History: a model without intra-repetition regularization and BOW prior (see Sec. 5.2.3); and (3) With History: the final model. Repeated sentences are highlighted with a yellow colored marker. Repeated words in a sentence are emphasized in red color.

Method	Focused	Coherent	Share	Human-like	Grounded	Detailed
AREL	3.49	3.18	3.18	3.26	3.32	3.15
Ours	3.67	3.52	3.20	3.56	3.54	3.32
GT	**3.72**	**3.57**	**3.34**	**3.64**	**3.56**	**3.53**

Table 5.5: Human evaluation results for rating survey (scores are between 1-5).

since none of the more recent baselines are publicly available. Note that we also compare coherency against a model without ordered-factors, which already improves upon the prior state-of-the-art.

In Fig. 5.5a we provide the results when asking annotators to pick the most human-like story. We use the majority vote to decide the best model per story. The generated stories outperform the AREL baseline (73.87% *vs.* 22.53%). Surprisingly, in many cases, the annotators found the generated stories to be more human-like than the ground truth stories (41% *vs.* 48.57%). In Fig. 5.5b, we assess coherency. An important aspect of our work are the directional factors for coherency. To validate their effectiveness, we compared to a model that does not incorporate direction into the attention representation (*i.e.*, we use the same factor for preceding and proceeding interactions). The comparison shows a significant coherency improvement (64.2% *vs.* 28.7%). Also, a comparison against the AREL baseline demonstrates a more significant improvement (70.24% *vs.* 25.32%).

To further evaluate the quality of the stories, we follow the criteria set by the Visual Storytelling Challenge[2] and conduct a survey where judges are asked to rate six categories between 1-5: *1. Focused*: the story contains information that is "naturally" relevant to the rest of the story; *2. Coherence*: the sentences in the story are related and consistent; *3. Share*: the inclination to share the story; *4. Human-like*: the story was likely written by a human; *5. Grounded*: the story directly reflects concrete entities in the image; and *6. Detailed*: the story provides an appropriate level of detail.

[2] `http://visionandlanguage.net/workshop2018`

To obtain the final score, we average the annotators' scores per sample, followed by averaging across the entire sample set. From Tab. 5.5 we observe: the model improved on all the criteria compared to the AREL model. Importantly, the generated stories are comparable to the ground-truth stories, indicating success in reducing the shortcomings found in prior methods. Nonetheless, the level of detail is still lacking, supporting the observation of Holtzman *et al.* [HBD+20] that current decoding strategies tend to generate well-formed yet somewhat generic text.

5.3.4 Qualitative Evaluation

In Fig. 5.6, we show the ability of the method in reducing repetitions. We observe the AREL baseline to repeat the same sentences, for example, "...had a great time at...". We also observe this repetitiveness when we remove the bag-of-words prior and the intra-sentence regularization (*i.e.*, No History column). Nevertheless, the method remains on topic, *i.e.*, family in the pool.

In Fig. 5.7 we sketch the attention maps along with the generated story. The first sentence, "We went to the mountains," sets the theme for the story, which requires the processing of subsequent images. Notably, the ISA module picked the proceeding images. In contrast, for the second sentence, the attention focuses mostly on the second image resulting in a description of the lake observed exclusively in this image. The third sentence relates to the scenery. Hence the attention focuses on preceding and proceeding images.

5.4 Conclusion

We present a novel approach for VST, which encourages coherency of generated story. We incorporate structure between images with a new attention method that selects the important objects in an ordered image sequence. Human evaluation and quantitative analysis demonstrate that the approach outperforms existing methods. Further, we perform ablation and qualitative analysis to show effectiveness.

Ground Truth	The clouds compliment the mountain peak. They find a lovely forested mountain with a lake. The misty clouds roll in and obscure the scene. The height of the mountains can be seen by the snow covering them. On the road again moving towards another place.
Ours	**We went to the mountains for a hike. The view of the lake was amazing. The scenery was breathtaking. We saw some old buildings. The view of the mountain was spectacular.**

Figure 5.7: Illustration of OIA and ISA attention maps, the ground-truth story and the final generated story. Each row corresponds to a story sentence and shows objects OIA highlights. The attended images' border specifies the relevancy to sentence generation, from red (important) to blue (not important).

Chapter 6

Audio-Visual Scene-Aware Dialog

We are interacting with a dynamic environment which constantly stimulates our brain via visual and auditory signals. Despite the huge amount of different information that is permanently occupying our nervous system, we are often easily able to quickly discern important cues from data that is irrelevant. Telling apart useful information from distracting aspects is also an important ability for virtual assistants, car navigation systems, or smart speakers. However present day technology uses a chain of components from speech recognition and dialog management to sentence generation and speech synthesis, making it hard to design a holistic and entirely data-driven approach.

For instance, in computer vision, a tremendous amount of recent work has focused on image captioning [VTBE14, DHG+15, MXY+15, WSL17, ADS18, CS18], visual question answering [GKS+17, SSH16, RHGS15, MRF15, XMS16, XBK+15], and visual dialog [DKG+18, JLS18]. While those meticulously engineered algorithms have shown promising results in their specific domain, little is known about the end-to-end performance of an entire system. This is partly due to the fact that little data is publicly available to design such an end-to-end algorithm.

Recent work on audio-visual scene aware dialog [ACD+18, HAW+18] partly addresses this shortcoming and proposes a novel dataset. Different from classical datasets like MSCOCO [LMB+14], VQA [GKS+17] or Visual Dialog [DKG+18], this new dataset contains short video clips, the corresponding audio stream and a sequence of question-answer pairs. While development of an end-to-end data driven system isn't feasible just yet due to the missing speech signal, the new audio-visual scene aware dialog dataset at least permits to develop a holistic dialog management and sentence generation approach taking audio and video signals into account.

In recent work [ACD+18, HAW+18], a baseline for a system based on audio, video and language data was proposed. Compelling results were achieved, demonstrating accurate question answering. The authors demonstrate that multimodal features based on I3D-Kinetics (RGB+Flow) [CZ17] refined via a carefully designed attention-based mechanism improve the quality of the generated dialog.

However, since much effort was dedicated to collecting the dataset, little analysis of

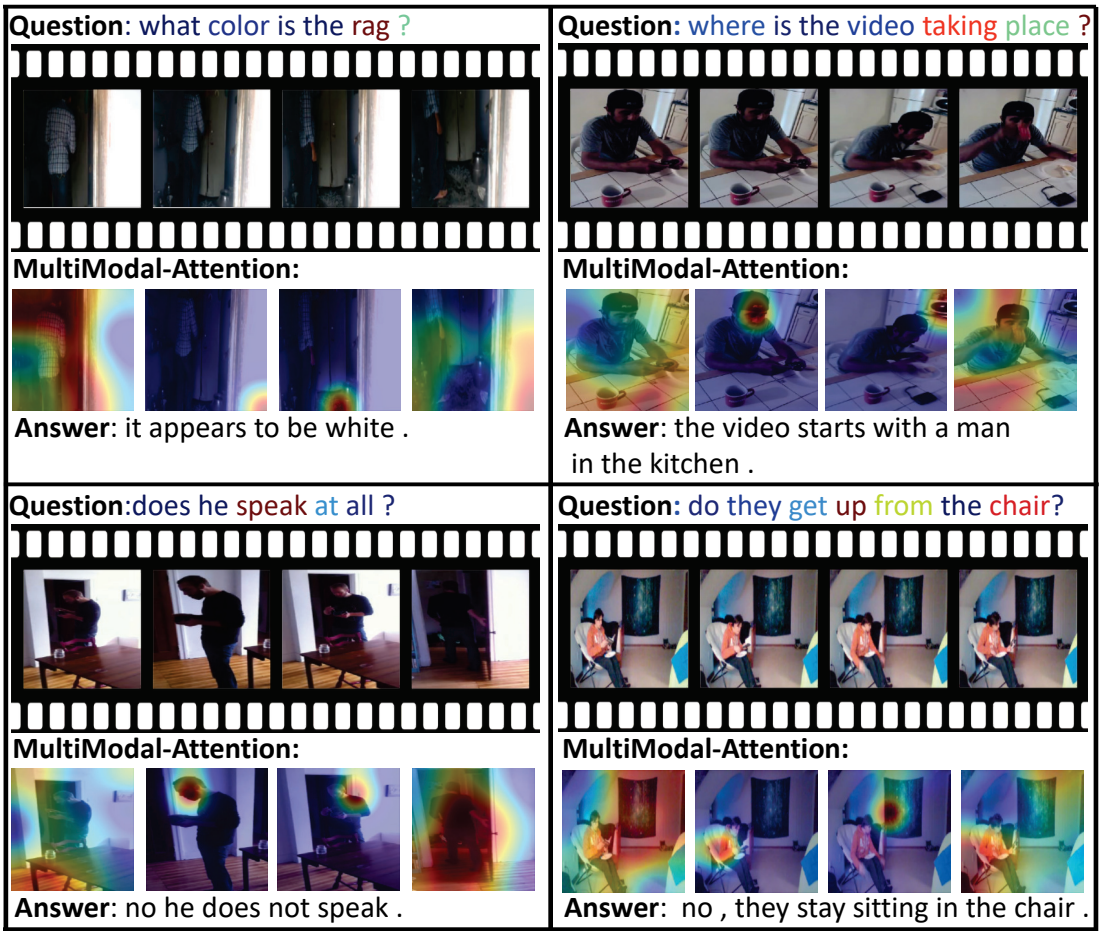


Figure 6.1: We present 4 different questions and the generated answer. Our attention unit is illustrated as well. Our model samples 4 frames, and attends to each frame separately, along with the question and the audio. We observe attention for each frame to differ, where first and fourth frames are widespread, while the second and third are more specific. Also, the question attention attends to relevant words. We also include the audio modality as input to the attention computation.

such a holistic system was provided. Moreover, due to tremendous amounts of available data (certainly a ten-fold increase compared to classical visual dialog data) this is by no means trivial. To provide this missing information and to share some insights with the community about how and where to improve, in this thesis, we follow the spirit of [JJvdM16] and demonstrate (1) that simply using the question as a signal already permits to outperform the current state-of-the-art; (2) that it is crucial to maintain spatial features for the video signal (either VGG19 [SZ15] or I3D-Kinetics [CZ17]). Reducing every video frame into a single representation drops performance significantly; (3) that temporally subsampling the video frames improves the accuracy; (4) that using attention over all available data (including different frames) is beneficial. To this end we analyze how to fuse the attended vectors for different data modalities.

Our simple baseline, which consists of three jointly trained components (data representation extraction, attention and answer generation) outperforms state-of-the-art by

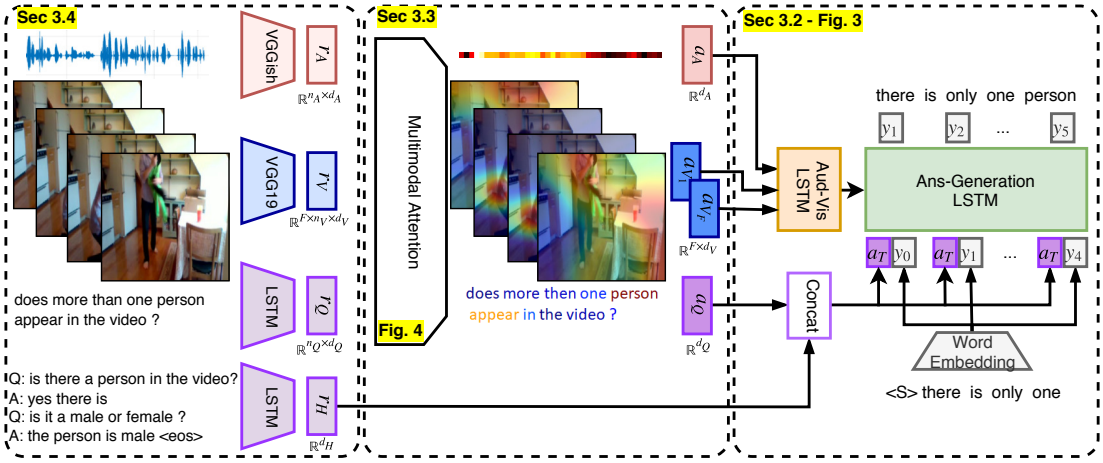


Figure 6.2: Overview of our approach for the AVSD task. More details can be found in Sec. 6.2.

a large margin of 20% on CIDEr. Improvements of the proposed approach are largely due to the aforementioned four points. Results of generated answers are contrasted to the current state-of-the-art in Fig. 6.1. We observe plausible answers to many questions and attention that focuses on important parts in both video and text.

6.1 Related Work

A significant amount of research has been conducted regarding image captioning, visual question generation, visual question answering, visual dialog, video data, audio data and multimodal attention models. We briefly review those related areas in the following.

Image Captioning: Originally image captioning was formulated as a retrieval problem. The best fitting caption from a set of considered options was found by matching features obtained from the available textual descriptions and the given image. Importantly, the matching function is typically learned using a dataset of image-caption pairs. While such a formulation permits end-to-end training, assessing the fit of image descriptors to a large pool of captions is computationally expensive. Moreover, it's likely prohibitive to construct a database of captions that is sufficient for describing even a modestly large fraction of plausible images.

To address this challenge, recurrent neural nets (RNNs) decompose captions into a product space of individual words. This technique has recently found widespread use for image captioning because remarkable results have been demonstrated which are, despite being constructed word by word, syntactically correct most of the time. For instance, a CNN to extract image features and a language RNN that shares a joint embedding layer was trained [MXY+15]. Joint training of a CNN with a language RNN to generate sentences one word at a time was demonstrated in [XBK+15]. A bi-directional RNN was employed along with a structured loss function in a shared vision-language space [KFF15]. Diversity was considered, *e.g.*, by Wang *et al.* [WSL17].

Visual Question Answering: Beyond generating a caption for an image, a large

amount of work has focused on answering a question about a given image. On a plethora of datasets [MRF15, RKZ15, AAL+15, GMZ+15, ZGBFF16, JHvdM+17], models with multi-modal attention [LYBP16, YHG+16, ARDK16, FPY+16, SSH16], deep net architecture developments [BYCCT17, MRF15, MLL16] and memory nets [XMS16] have been investigated.

Visual Question Generation: In spirit similar to question answering is the task of visual question generation, which is still very much an open-ended topic. For example, Ren *et al.* [RKZ15] discuss a rule-based method, converting a given sentence into a corresponding question which has a single word answer. Mostafazadeh *et al.* [MMD+16] learned a question generation model with human-authored questions rather than machine-generated descriptions. Vijayakumar *et al.* [VCS+18] have shown results for this task as well. Different from the two aforementioned techniques, Jain *et al.* [JZS17b] argued for more diverse predictions and use a variational auto-encoder approach. Li *et al.* [LDZ+17] discuss VQA and VQG as dual tasks and suggest a joint training. They take advantage of the state-of-the art VQA model by Ben-younes *et al.* [BYCCT17] and report improvements for both VQA and VQG.

Visual Dialog: Visual dialog [DKG+18] combines the three aforementioned tasks. Strictly speaking it requires both generation of questions and corresponding answers. Originally, visual dialog required to only predict the answer for a given question, a given image and a provided history of question-answer pairs. While this resembles the VQA task, different approaches, *e.g.*, also based on reinforcement learning, have been proposed recently [KMP+18, JLS18, WWS+17].

Video Data: A variety of tasks like video paragraph captioning [YWH+16], video object segmentation [PWGSH15], video classification [KFF15], and action recognition [SZ15] have used video data for a long time. Probably most related to our approach are video classification and action recognition since both techniques also extract a representation from a video. While the extracted representation is subsequently used for either classification or action recognition, we employ the representation to more accurately answer a question. Commonly used feature representations for either video classification or action recognition are I3D-based features by Carreira *et al.* [CZ17], extracted from an action recognition dataset. With proper fine-tuning the I3D-based features proved to be better than the classical approaches, such as C3D [TBF+15] that capture spatiotemporal information via a 3D CNN. In this work, we assess a naïve feature extractor based on VGG [SZ15], and demonstrate that for video-reasoning, careful reduction of the spatial dimension is more crucial than the type of extracted features used to embed the video frames. Wang *et al.* [WXW+16] showed that working with video frame samples, achieves not only efficiency, but also improves performance compared to a conservative dense temporal representation. Recently, Zhou *et al.* [ZAT17] further extended those ideas, and suggested to capture relational temporal relationships between the sampled frames, relying on the relational-networks concept [SRB+17]. We follow those ideas by also sub-sampling a small set of frames uniformly. Our model

further advances those concepts, by exploiting spatial relationships between sampled temporal frames via a high-order multimodal attention module, where each video frame is treated as a separate modality. Li *et al.* [LGG+18] propose the Video-LSTM model, which uses attention to emphasis relevant locations, during LSTM video encoding. Our approach differs in that attention on one frame can influence attention on other frames which isn't the case in their model.

Audio Data: Audio data gained popularity in the vision community recently. For instance, prediction of pose given audio input [SDSKS18], learning of audio-visual object models from unlabeled video for audio source separation in novel videos [OE18], use of video and audio data for acoustic scene/object classification [AVT16], source separation was also considered in [EML+18] and learning to see using audio [OWM+18].

Multimodal Attention: Multimodal attention has been a prominent component in tasks which operate on different input data. Xu *et al.* [XBK+15] showed an encoder decoder attention model for image captioning, which was extended to visual question answering [XS16]. Yang *et al.* [YHG+16] propose a multi-step reasoning system using an attention model. Multimodal pooling methods were also explored [FPY+16, KOL+17]. Lu *et al.* [LYBP16] suggest to produce co-attention for the image and question separately, using a hierarchical and parallel formulation. Schwartz *et al.* [SSH17, SYHS19] later extend this approach to high-order attention applied over image, question and answer modalities via potentials. Similarly, in the visual dialog task, co-attention models have held the state-of-the-art [WWS+17, LKY+17] attending over image, question and history in hierarchical manner. For audio-visual scene-aware dialog, [HAW+18] also use a sum-pooling type of attention, using the question feature along with audio and video modalities separately. In contrast, here we compute attention over each modality via local and cross data evidence, letting all the modalities interact with each other.

6.2 Audio Visual Scene-Aware Dialog Baselines

Our method has three building blocks: answer generation, attention and data representation as shown in Fig. 6.2.

6.2.1 Answer Generation

We are interested in predicting an answer $y = (y_1, \ldots, y_n)$ consisting of n words $y_i \in \mathcal{Y}_i = \{1, \ldots, |\mathcal{Y}_i|\}$ each arising from a vocabulary of possible words $\mathcal{Y}_i$. Given data $x = (Q, V, A, H)$ which subsumes, a question Q, a subsampled video $V = (V_1, \ldots, V_F)$ composed of F frames, the corresponding audio signal A, and a history of past question-answer pairs H, we construct a probability model over the set of possible words for the answer generation task. To this end, we formulate prediction of the answer as inference in a recurrent model where the joint probability is given by the product of conditionals,

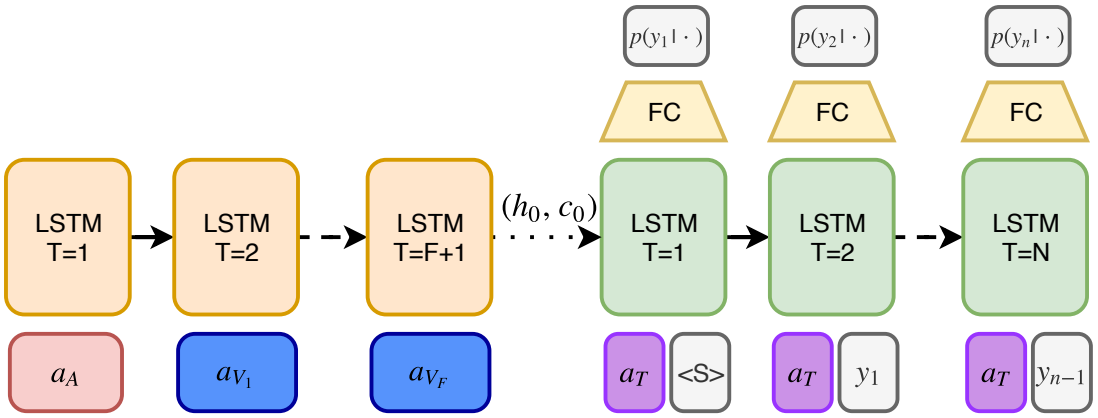


Figure 6.3: Our decoder for audio-visual scene-aware dialog. We start with encoding of attended audio and video vectors using the Aud-Vis LSTM (orange colored), followed by the Ans-Generation LSTM that receives the textual data concatenated with the previous answer word (green colored).

i.e.,

$$p(y|x) = \prod_{i=1}^{n} p(y_i|y_{<i}, x).$$

Note that, for now, we condition on all the data x for readability and provide details later. Instead of conditioning the probability of the current word $p(y_i|y_{<i}, x)$ on its entire past $y_{<i}$, we combine two recurrent nets: an audio-visual recurrent net that generates the temporal information which is fed as an initialization to the answer generating recurrent net. See Fig. 6.3 for a schematic.

Audio-visual LSTM-net: It operates on an attended audio embedding a_A and attended video embeddings $a_{V_1}, ..., a_{V_F}$ for each of the F frames $f \in \{1, \ldots, F\}$. This LSTM-net has $F + 1$ units, the first unit's input is the attended audio vector, and the input to the F subsequent units are the attended video representations $a_{V_1}, \ldots, a_{V_F}$. The context vector that is generated from this LSTM, *i.e.*, (h_0, c_0) summarizes the audio-visual attention and is provided as input to the answer generation LSTM-net.

Answer generation LSTM-net: It computes conditional probabilities for the possible words $y_i \in \mathcal{Y}_i$ of the answer $y = (y_1, \ldots, y_n)$. This probability considers the last word and captures context via a representation h_{i-1} obtained from the previous time-step.

$$p(y_i|y_{i-1}, h_{i-1}, x) = g_w(y_i, y_{i-1}, h_{i-1}, x).$$

We illustrate the LSTM-net g_w in Fig. 6.3. Using the initial state (h_0, c_0), the LSTM-net g_w predicts in its i-th step a probability distribution $p(y_i|y_{i-1}, h_{i-1}, x)$ over words $y_i \in \mathcal{Y}_i$ using as input y_{i-1} and the textual attention vector $a_T = (a_Q, r_H)$: the attended textual vector is a concatenation of the attended question vector a_Q and the history vector r_H, which represents information about question and history data. The output of the LSTM-net is transformed via a FC-layer with a dropout and a softmax to obtain the probability distribution $p(y_i|y_{i-1}, h_{i-1}, x)$.

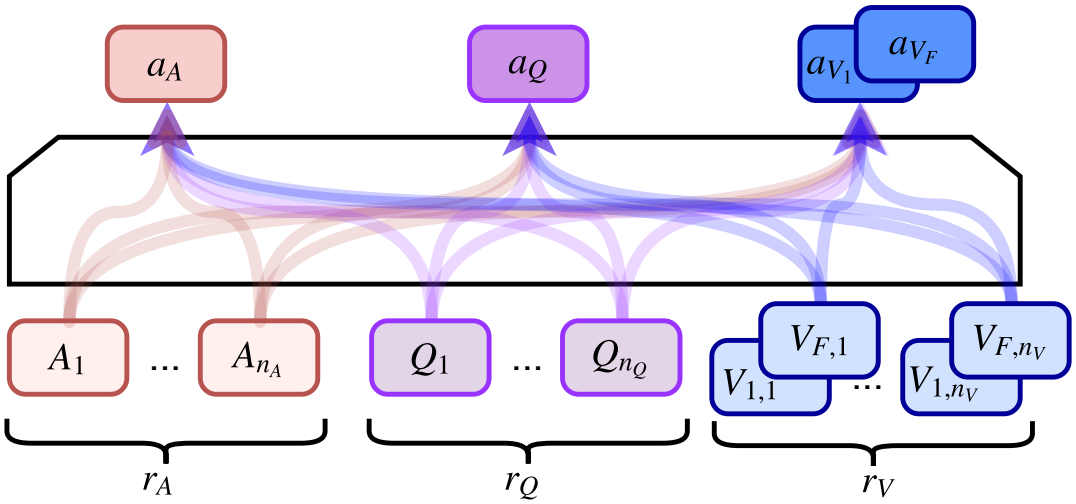


Figure 6.4: Multimodal Attention model for audio-visual scene-aware dialog. We treat each frame as a modality, along with audio and question modality, to total of 6 modalities. Each element attention score is affected not only from local evidence, but also via cross-data interactions of all other elements.

6.2.2 Attention

The attention step provides an attended representation for the data components, *i.e.*, $a_{V_f} \in \mathbb{R}^{d_V}$ for frame $f \in \{1, \ldots, F\}$ of the video data, $a_A \in \mathbb{R}^{d_A}$ for the audio data, and $a_T \in \mathbb{R}^{d_T}$ for the textual data. These attended representations are obtained by transforming the representations extracted from the raw data, *i.e.*, $r_{V_f} \in \mathbb{R}^{n_V \times d_V}$ for the video data, $r_A \in \mathbb{R}^{n_A \times d_A}$ for the audio data, and for the textual data, $r_Q \in \mathbb{R}^{n_Q \times d_Q}$ as well as $r_H \in \mathbb{R}^{d_H}$ which capture signals from the question and history respectively. We outline the general procedure in Fig. 6.4.

Formally, we obtain the attended representation

$$a_\alpha = \sum_{k=1}^{n_\alpha} \alpha_k p_\alpha(k),$$

where $\alpha \in \{A, Q, V_1, \ldots, V_F\}$ is used to index the available data components (audio, question, visual frames), n_α is the number of entities in a data component (*e.g.*, the number of words in a question), and $p_\alpha(k) \geq 0 \ \forall \alpha$ is a probability distribution ($\sum_{k=1}^{n_\alpha} p_\alpha(k) = 1 \ \forall \alpha$) over the n_α entity representations of data α. For instance, if we let $\alpha = A$ we obtain the attended audio representation $a_A = \sum_{k=1}^{n_A} A_k p_A(k)$.

We compute the attention via a factor graph attention approach [SSH17, SYHS19]. The attention probability distribution over a data source α consists of a log-prior distribution π_α, a local evidence l_α that relies solely on its data representation r_α and a cross data evidence c_α that accounts for correlations between the different data representations r_α, r_β, for $\beta \in \{A, Q, V_1, \ldots, V_F\}$. This probability distribution takes the

form:

$$p_\alpha(k) \propto \exp\left(\hat{w}_\alpha \pi_\alpha(k) + l_\alpha(k) + c_\alpha(k)\right).$$

The local evidence is $l_\alpha(k) = w_\alpha \left(v_\alpha^\top \operatorname{relu}(V_\alpha \alpha_k)\right)$, the log-prior is $\pi_\alpha(k)$ and the cross data evidence is

$$c_\alpha(k) = \sum_{\beta \in \mathcal{D}} \frac{w_{\alpha,\beta}}{n_\beta} \sum_{j=1}^{n_\beta} \left(\left(\frac{L_\alpha \alpha_k}{\|L_\alpha \alpha_k\|} \right)^\top \left(\frac{R_\beta \beta_j}{\|R_\beta \beta_j\|} \right) \right).$$

The set $\mathcal{D} = \{A, Q, V_1, \ldots, V_F\}$ consists of the possible data types. The trainable parameters of the model are: (1) $V_\alpha, L_\alpha, R_\alpha$ which re-embed the data representation to tune the attention; (2) v_α which scores the local modality; and (3) $\hat{w}_\alpha, w_\alpha, w_{\alpha,\beta}$ which weight the three components with respect to each other.

We found the use of attention for history to not yield improvements. Therefore, we obtain the attended textual representation $a_T \in \mathbb{R}^{d_T}$ by concatenating the attended question representation $a_Q \in \mathbb{R}^{d_Q}$ with the history representation $r_H \in \mathbb{R}^{d_H}$. Consequently, $d_T = d_Q + d_H$.

6.2.3 Data Representation

The proposed approach relies on representations r_α obtained for a variety of data components which we briefly discuss subsequently.

Video: Containing both temporal and spatial information, video data is among the most memory consuming. Common practice is to reduce the spatial information while maintaining attention over the temporal dimension. Instead, we first reduce the temporal dimension, maintaining the ability for spatial attention to reason about the video content. To ensure fast training, we reduce the temporal dimension by sampling F frames uniformly. For each sampled frame we extract a representation from a deep net trained on ImageNet (in our case VGG19). We then fine tune the representation of each frame using a 1D conv layer with a bias term. This conv layer is identical for all the F frames. Consequently, we obtain the video representation $r_V \in \mathbb{R}^{F \times n_V \times d_V}$, where F is the number of sampled frames, n_V is the spatial dimension and d_V is the embedding dimension.

Audio: For audio, we extracted features from a strong audio classification model (*i.e.*, VGGish [HCE+17]) by taking the last representation before the final FC-layer. This representation has adaptive temporal length. For each batch we find the maximal temporal length of the audio signal, and zero-padded the shorter audio representations. We then fine-tune each audio file using a 1D conv layer with a bias. We obtain the audio representation $r_A \in \mathbb{R}^{n_A \times d_A}$, where n_A is the maximal temporal length of a given batch and d_A is the embedding dimension.

Question: We start with an adaptive-length list of 1-hot word-representations. For

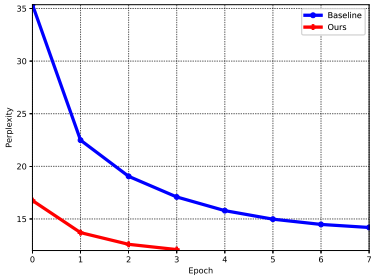


Figure 6.5: Perplexity values for our model *vs.* baseline [HAW$^+$18]

each batch we find the longest sentence, and zero-pad shorter ones. We embed each word using a linear-embedding layer, followed by a single layer LSTM-net with dropout. The last hidden state of the LSTM is the question representation $r_Q \in \mathbb{R}^{n_Q \times d_Q}$, where n_Q is the length of the maximal sentence for the given batch and d_Q is the embedding dimension.

History: The history data source consists of the past T question-answer pairs, which we denote by $H = (Q, A)_{t \in \{1,\dots,T\}}$. The history embedding consists of two components: we first embed each question-answer pair $(Q, A)_t$ using a LSTM-net to get T representations of the history. We then feed these representations into another LSTM-net to obtain the vector representation $r_H \in \mathbb{R}^{d_H}$, where d_H is the history embedding dimension.

We embed each question-answer pair $(Q, A)_t$ following the question embedding above. A question-answer pair starts with a list of 1-hot word-representations of the words in the question followed by 1-hot word-representations of the words in the answer. For each batch we find the longest question-answer sequence, and zero-pad the shorter ones. We embed each 1-hot vector using a linear-embedding layer, followed by a two layer LSTM-net with a dropout. The last hidden state of this LSTM-net is the vector representation of $(Q, A)_t$, which we denote by r_t.

We embed the history by feeding $r_1, \dots, r_T$ to a one layer LSTM-net with dropout, in order to capture the temporal aspect of the question-answer history. To deal with the adaptive length of history interactions, for each batch we find the interaction with the longest history, and zero-pad question-answer pairs with shorter history. The final LSTM-net hidden state is the history representation $r_H \in \mathbb{R}^{d_H}$, where d_H is the history embedding dimension.

6.3 Results

In the following we evaluate the discussed baseline on the Audio Visual Scene-Aware Dialog (AVSD) dataset. We follow the proposed protocol and assess the generated answers to a user question given a dialog context [ACD$^+$18, HAW$^+$18]. This context consists of a dialog history (previous questions and answers) in addition to video and

audio information about the scene. Our code is publicly available[1].

6.3.1 AVSD v0.1 Dataset

The AVSD dataset consists of annotated conversations about short videos. The dataset contains 9,848 videos taken from CHARADES, a multi-action dataset with 157 action categories [SVW+16]. Each dialog is obtained from two Amazon Mechanical Turk (AMT) workers, who discuss about events in a video. One of the workers takes the role of an answerer who had already watched the video. The answerer replies to questions asked by another AMT worker, the questioner.

The questioner was not shown the whole video but only the first, middle and last frames of the video. The dialog revolves around the events in and other aspects of the video. The AVSD v0.1 dataset is split into 7,659 train dialogs, 1,787 validation and 1,710 test dialogs. Because the test set doesn't currently include ground truth, we follow [HAW+18] and evaluate on the 'prototype test-set' with 733 dialogs. Because the 'prototype test-set' is part of the 'v0.1 validation-set,' we use the 'prototype validation-set' with 732 dialogs, which doesn't overlap with the 'prototype test-set.'

6.3.2 Implementation Details

Our system relies on textual, visual and audio data representations, *i.e.*, r_α for $\alpha \in \{A, Q, V_1, \ldots, V_F\}$. For the **video** representation we randomly sample $F = 4$ equally spaced frames, and use the last conv layer of a VGG19 having a dimensions of $7\times7\times512$. Therefore the visual embedding dimension is $d_V = 512$. After flattening the 2D spatial dimension, we obtain the spatial dimension $n_V = 49$. For **audio** features we use VGGish that operates on 0.96s log-Mel spectrogram patches extracted from 16kHz audio, and outputs a $d_A = 128$ dimensional vector. VGGish inputs overlap by 50%, therefore an output is provided every 0.48s. Dropout parameters before the last FC layer, and the LSTM layers are set to 0.5. For the **question** representation we set the word embedding dimension to 128. The questions are embedded to $d_Q = 256$ dimensional vectors, extracted from the last hidden state of their LSTM-net. The **history** consists of $T = 10$ question-answer pairs, which we denote by $H = (Q, A)_{t\in\{1,\ldots,T\}}$. We use an LSTM-net with a hidden state of $d_H = 128$ to encode the history.

6.3.3 Training

We use a cross-entropy loss on the probabilities, $p(y_i|y_{<i}, x)$ to train the answer generator, the attention and the embedding layers jointly end-to-end. The total amount of trainable parameters are 8,359,107. We use the Adam optimizer [KB14] with a learning rate of 0.001 and a batch size of 64. During training after each epoch we evaluate our

[1] https://github.com/idansc/simple-avsd

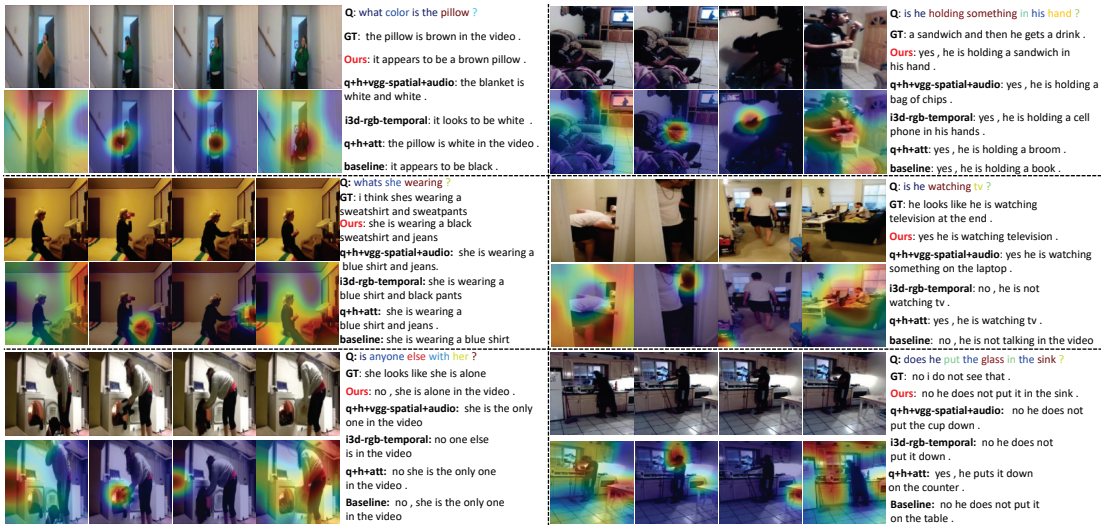


Figure 6.6: An illustration of out 4-framed samples from a video along with the relevant attention variables. Our attention treats any frame as different component. This allows the attention module to learn different attention behaviors for different temporal locations. We observe the first and fourth samples are noisier, while the second and third attend to specific interesting locations. Our multimodal attention also generates attention for questions, illustrated over the question via a word heat map. We provide generated answers for different baseline models: q+h+att, is a model with only history and question input; i3d-rgb-temporal is a model with temporal features instead of spatial; q+h+vgg-spatial+audio is a model without attention. We also compare to the generated answer by [HAW+18]. the ground-truth is denoted by GT, and our final model denoted by Ours.

performance on the validation set using a perplexity metric. We stop our training after two consecutive epochs with no improvement.

We use a standard machine with an Nvidia Tesla M40 GPU for all our experiments. Training our system takes 4 epochs to converge *vs.* 9 epochs for the baseline (see Fig. 6.5). Each epoch takes 8 minutes *vs.* 13 minutes for the baseline. In total, training our model takes approximately 30 minutes.

6.3.4 Performance Evaluation:

We evaluate the performance of our system using several metrics. Our prime metric is CIDEr, the Consensus-based Image Description Evaluation, which measures the similarity of a sentence to the consensus [VZP14]. We also evaluate our performance on the ROUGE-L metric (Recall Oriented Understudy of Gisting Evaluation). This is a recall-based metric that measures the longest common subsequence of tokens [Lin04]. The METEOR metric is a unigram precision and recall that allows for matchings between candidates and references [BL05]. We also evaluate our performance using the traditional BLEU score, which measures the effective overlap between a reference sentence and a candidate sentence. We measure the geometric mean of the effective n-gram precision scores, for $n = 1, \ldots, 4$ and refer to these as BLEU1,..., BLEU4.

6.3.5 Quantitative Results and Insights for a Good Baseline

We compare to the baseline discussed in [HAW+18]. In the following we explore the various components of audio-visual dialog systems and present our insights for constructing a simple and effective baseline. These insights cover all aspects of our system: feature embedding, attention, fusion and training techniques. We particularly emphasize the importance of spatial features for AVSD, which we contrast with the action recognition based I3D features.

Question Bias and Basic Baselines: We revisit the scores published by [HAW+18] and assess a basic seq2seq-type baseline, with no attention [SVL14]. In this variant, which we call **q** in Tab. 6.1, we encode the question using a word embedding (with embedding dimension of 128) and a 1-layer LSTM-net (with hidden state dimension of 256 compared to a dimension of 128 in the baseline), without any video or history related features. For decoding, another 1-layer LSTM-net (with hidden state dimension of 256 compared to a dimension of 128 in the baseline) is used. Surprisingly, this model alone was able to surpass the current baseline of [HAW+18]. Similar results are also reported in [SPM19]. This indicates that there might be bias-problem within the AVSD dataset, no visual information is needed. For instance a common question is "How many people are in the video?", but videos in many cases feature only one person. Another example are questions of the form "is it indoor?" which are meaningless since the CHARADES dataset focuses on indoor activities. Another possible explanation for this good result is the encoding of the answer in the question. For instance, a question "this person is standing in a kitchen correct?" is answered with "yes he is in the kitchen." Moreover, generative evaluation is also more prone to biases, as the evaluation emphasizes correct sentence structure rather than correctness of the answer. Very recently, a discriminative approach was proposed [ACD+18]. The bias problem is not unique to AVSD, and was also discussed for Visual Question Answering [GKS+17].

To further improve the most basic baseline **q**, we add more modalities. We use the fusion and embedding techniques of the proposed model but omit attention. Instead of attention, we use a mean over the representation for visual and auditory data sources, and the last hidden state of the LSTM-net is used to represent the question data source. We found that our model can utilize any modality supplement, even without attention. In the 'basic baselines+attention' section of Tab. 6.1 we assess versions with attention, which brings us closer to our full model.

Spatial vs. Temporal Information: Current methods focus on temporal models and often naïvely reduce the spatial dimension [HAW+18, WXW+16, ZAT17]. In contrast, for closely related visual reasoning tasks, such as visual dialog and visual question answering, it is broadly accepted that spatial attention is necessary. Therefore, it is unlikely that video reasoning is effective when simply reducing the spatial dimension. Indeed, we find better results when reducing the temporal dimension with sampling techniques and employing attention to reduce the spatial dimension. In Fig. 6.6 we

observe that a small subset of frames (*e.g.*, 4) is usually enough for an almost complete understanding of the video. In the 'i3d-features-&-spatial-temporal' section of Tab. 6.1, we compare spatial-based features to temporal-based ones. The temporal features are computed on a stack of 16 video frames, and are treated as an input modality to our attention mechanism. Attention choses the relevant temporal locations. The temporal attended representation was fed to the Aud-Vis LSTM-net along with the audio attended-features. For the i3d-rgb-flow version we also use the I3D model based on optical flow features as an additional data component. This resulted in a drop in performance compared to the spatial-based i3d-features reported in the i3d-rgb-spatial-10 line of Tab. 6.1. We also test different number of sampled frames. Interestingly, only one frame is already very useful for AVSD, and too many VGG-frames harm performance. Note that each frame is coupled to an attention-score and treated as a modality, which explains why too many frames can add noise to the inferred multimodal probability.

I3D Features *vs.* VGG: I3D features are widely used as video-based feature extractor (*cf.* [CZ17]), discarding the classical image-based features, *e.g.*, VGG. They are extracted from a model trained on the Kinetics Dataset, a dataset for action recognition, and have been shown to improve many video tasks. We find that while I3D features have repeatedly been shown to improve on action-recognition tasks, they are not as useful in the answer generation task of AVSD. Equipped with VGG features we were able to achieve comparable results to the i3d-rgb-spatial-20 version. The i3d-rgb-spatial features are 4 times bigger (7x7x512 *vs.* 2x7x7x1024), as well as more complicated to extract. Seeking simplicity, we report scores with the VGG-based features subsequently. This may also indicate a weakness in the dataset, as this solution seems to be suboptimal for action-related questions (*e.g.*, classifying sequences of actions). Not only do we naïvely sample temporal frames, but also do we not use I3D features that were extracted from a network trained for action-recognition, yet we achieve good results.

Attention Model: We assess different components of the attention model. See Sec. 6.2.2 for details about local evidence and cross data evidence. We found that every component contributes to the model, especially the cross-data component. The cross-data component determines the attention score of an element by considering interactions with other modalities. For instance, a region in the second frame can affect a region in the third frame, or perhaps a word in the question.

To find the simplest attention module, we also explored the option of grouping together the parameters for all video frames, *i.e.*, $V_{V_1} = \ldots = V_{V_F}$, $L_{V_1} = \ldots = L_{V_F}$, and $R_{V_1} = \ldots = R_{V_F}$, which yields good results despite 2 million fewer parameters. This version allows to increase the number of processed frames, with no additional memory cost. Those results are reported in the 'sharing-weights' line of Tab. 6.1.

Multimodal Decoding Fusion: We experimented with several variants that reduce $a_A, a_{V_1}, \ldots, a_{V_F}$. In Tab. 6.1, section 'decoder-input,' we show a version that uses an additional multimodal attention step over the video-related attended vector, called temporal-attention. Another attempt is summation polling of the vectors, and weighted

summation with scalers. Instead, we note the sequential information of $a_{V_1}, \ldots, a_{V_F}$ that naturally calls for the use of an additional LSTM unit, which we call Aud-Vis (see Fig. 5.7). We think audio is a more general cue while frames have more specific information. Ordering is guided by the intuition that LSTM-based encoding commonly starts with more general information. To verify this intuition, in video-audio-lstm, we performed additional experiments with ordering of $a_{V_1}, \ldots, a_{V_F}, a_A$.

Next we find a good way to input elements into the answer generation LSTM-net. We first analyze the basic **q** model. A classic decoder, where encoded **q** are fed as first hidden state to the LSTM-net is reported in the 'q-first-state' row in Tab. 6.1 (decoder-input section). This suggest that textual data should be concatenated to the decoder inputs. Concatenating all modalities to the input, which is reported in the 'all-concat-input' line in Tab. 6.1 drops the performance, suggesting that a dichotomy of video-related and textual-related features is useful. To incorporate the audio signal, we find it's best to use it as a first state in the Aud-Vis LSTM-net. A version where we concatenated the audio attended vector to a_T is referred to as 'q+h+a-concat-input+s-first-state.' The model behaves the best when the fused video related features were used as the initial state h_0 of the Ans-Generation LSTM-net. Our state-of-the-art model further improves the fusion technique by using the Aud-Vis LSTM-net to generate h_0 which captures the temporal information of audio attention a_A and the visual attention $a_{V_1}, \ldots, a_{V_F}$.

Weight Initialization: An important aspect is the initialization of the deep net parameters. We observed a significant improvement using Kaiming normal initialization or Xavier initialization for all LSTM models [HZRS15b, GB10].

Beam Search Width: In an attempt to improve the overall evaluation time, we experimented with different beam width. We found that although beam search is useful for generation, a width of 2 achieves almost as good results. Our version use 3-width beam search.

6.3.6 Qualitative Results

In Fig. 6.6, we show several examples of generated answers of five models, our final model, a version without any attention (q+h+vgg-spatial+audio), a version with temporal I3D features (i3d-rgb-temporal), a version with only textual modalities (q+h+att), and the baseline [HAW$^+$18]. The ground-truth is referred to via GT.

Additionally, we take advantage of the interpretability of attention modules to also illustrate the attention probabilities of our final modal on 5 different modalities, *i.e.*, our 4-frames, and the question. First, we observe an interesting behavior of our attention model: each sampled frame is attended a differently, which captures different features from different frames. The first and fourth frames are noisier and extract general concepts, while the second and third capture unique aspects of the video, *e.g.*, a person, a couch. This behavior can be associated with the temporal aspect of the frames.

Meaning it is more important to capture general aspects at the end and at the beginning, but in the middle we reveal the important specific concepts. Additionally, the question attention attends to the informative words. Our generated answers are usually more aware of the scene, and less prone to bias. For instance, in the first row, the question is "what color is the pillow?." We observe our model to be able to answer the correct color, while all other model variants answer with white, the most-common color of a pillow. In another question "whats she is wearing," our model was the only one to relate to her black sweatshirt.

6.4 Conclusion

We propose a simple baseline for Audio-Visual Scene-Aware Dialog that surpasses current techniques by 20% on the CIDEr metric. Pioneering on this task, we carefully evaluated our approach. We hope our analysis can bridge the gap between video-reasoning and image-reasoning.

Table 6.1: Results for the AVSD dataset for CIDEr, BLEU1, ..., BLEU4, ROUGE-L, METEOR. We provide a comparison to the baseline and a detailed ablation study separated into categories and discussed in Sec. 6.3.5. We also report the number of parameters for each baseline.

Model	C	B4	B3	B2	B1	R	M	P
baseline[HAW+18]	0.766	0.084	0.117	0.173	0.273	0.291	0.117	6.15M
basic baselines								
q	0.815	0.088	0.122	0.178	0.279	0.297	0.121	3.1M
q+h	0.843	0.089	0.123	0.178	0.277	0.296	0.122	4.51M
q+h+vgg-spatial	0.869	0.089	0.124	0.180	0.279	0.302	0.123	5.12M
q+h+vgg-spatial+audio	0.874	0.091	0.125	0.182	0.282	0.305	0.124	5.23M
basic baselines+attention								
q+att	0.849	0.090	0.124	0.179	0.278	0.298	0.121	3.35M
q+h+att	0.861	0.090	0.124	0.177	0.271	0.298	0.122	4.57M
q+h+vgg-spatial+att	0.908	0.093	0.129	0.185	0.283	0.307	0.125	7.4M
attention-model								
w/o-cross-data-evidence	0.896	0.095	0.131	0.190	0.292	0.309	**0.128**	7.5M
w/o-local-evidence	0.917	0.096	0.132	0.191	0.293	0.309	**0.128**	8.35M
w/o-question-prior	0.906	0.096	0.132	0.190	0.292	0.309	0.127	8.35M
sharing–weights	0.923	**0.097**	**0.133**	**0.191**	**0.293**	0.309	0.127	6.18M
video-fusion								
temporal-attention	0.877	0.091	0.126	0.182	0.281	0.302	0.124	8.4M
summation	0.890	0.093	0.128	0.183	0.283	0.303	0.124	7.35M
weighted-summation	0.876	0.094	0.130	0.187	0.289	0.304	0.126	7.85M
video-audio-lstm	0.865	0.076	0.101	0.141	0.210	0.286	0.108	8.35M
decoder-input								
q-first-state	0.704	0.078	0.110	0.163	0.257	0.279	0.112	8.35M
all-first-state	0.714	0.079	0.114	0.171	0.271	0.276	0.113	10.1M
all-concat-decoder-input	0.797	0.089	0.125	0.183	0.285	0.297	0.121	9.53M
q+h+a-concat-input	0.857	0.090	0.123	0.177	0.274	0.298	0.121	7.72M
i3d-features-&-spatial-temporal								
i3d-rgb-temporal	0.886	0.094	0.130	0.188	0.289	0.306	0.126	7.23M
i3d-rgb-flow-temporal	0.851	0.091	0.127	0.185	0.286	0.303	0.125	7.82M
i3d-rgb-spatial-10	0.928	**0.097**	**0.133**	0.190	0.290	0.310	0.127	6.58M
vgg-spatial-1	0.919	0.095	0.130	0.187	0.287	0.309	0.126	6.18M
vgg-spatial-16	0.903	0.093	0.128	0.186	0.287	0.307	0.127	28.88M
initialization								
default	0.877	0.090	0.123	0.178	0.274	0.300	0.121	8.35M
xavier	0.848	0.087	0.119	0.171	0.262	0.297	0.119	8.35M
he	0.913	0.095	0.131	0.189	0.290	0.308	0.127	8.35M
beam-search hyper-parameters								
w/o beam	0.924	0.082	0.109	0.152	0.226	0.298	0.114	8.35M
2-width	0.934	0.094	0.128	0.183	0.279	0.311	0.126	8.35M
4-width	0.931	0.096	0.131	0.188	0.287	0.310	0.127	8.35M
5-width	0.926	0.096	0.132	0.188	0.289	0.309	0.127	8.35M
Ours	**0.941**	0.096	0.131	0.187	0.285	**0.311**	**0.128**	8.35M

Chapter 7

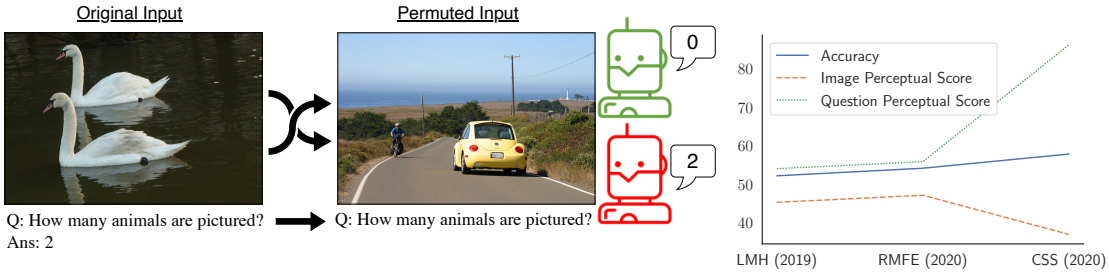
Perceptual Score: Measuring Perceptiveness of Multi-Modal Classifiers

Machine learning advances over the last decade are remarkable. Challenges that seemed daunting merely ten years ago are now a breeze, and new applications that we barely dared to dream about seem achievable within the next few years. Indeed, accuracy metrics on tasks like visual question answering and reasoning suggest significant improvements.

Reported improvements are to a large extent due to the availability of large datasets [AAL+15, DDS+09, LMB+14], computational performance advances, *e.g.*, for GPUs, and a better understanding about how to encode inductive biases into deep-nets, *e.g.*, by using rectified linear units [NH10], normalization [IS15], skip connections [HZRS15a], encoder-decoder structures [SVL14], *etc.* However, importantly, developed deep-net architectures are not guaranteed to solve a given task. There is a chance that they may instead exploit dataset biases as illustrated in Fig. 7.1a.

This concern is surely in part due to non-robust training techniques, and a plethora of methods improve classifier robustness [LZ14, SHK+14, SC18]. However, datasets play an important role in controlling the extracted bias as well. For instance, if correct answers in a question-answering task are significantly shorter than incorrect ones, classifier training should not use answer length as a cue. Although this seems reasonable, for audio-visual scene aware dialog,Schwartz *et al.* [SSH19] find for example that in many cases the question alone is sufficient to generate a scene-aware dialog response, avoiding the need to look at the video. Hence, in order to assess the suitability of a classifier, we need to understand how much it relies on different data modalities.

To quantify how much a classifier relies on its different input modalities, we introduce the perceptual score. The perceptual score assesses the degree to which a model relies on a modality. To do so the perceptual score permutes the features of a modality



(a) Two models: one perceives the image (green), and the other only perceive the question (red).

(b) Progress of VQA-CP models.

Figure 7.1: Multi-modal datasets often have undesired biases: (a) To identify those biases we suggest the perceptual score as a new metric. It assesses the change in prediction when a model's input for some modalities is permuted during testing. If the classifier output remains identical despite permutation, a model doesn't perceive the modality. (b) Using the perceptual score we identify that recent progress of VQA models may not be entirely due to better reasoning.

across samples in the test set after the classifier was trained. If the classifier's performance drops to or below chance level, the perceptual score is high. This intuitively applies to single-modality models too: randomly permuting test data and labels after training results in chance-level classification accuracy.

Using the perceptual score, we find a surprisingly consistent trend across four popular datasets (VQA, VQA-CP, Visual Dialog, SocialIQ): recent, more accurate state-of-the-art multi-modal models for visual question-answering or visual dialog tend to perceive the visual data less than their predecessors (see Fig. 7.1b). This trend is concerning as answers are hence increasingly inferred from textual cues only. Using the perceptual score also helps to analyze model biases by decomposing the score into data subset contributions. For example, the perception of an image and question varies depending on the question type. None of the recent VQA-CP models showed high image perception scores for 'number'-type questions. A surprisingly low image perception score is obtained for the state-of-the-art model when confronted with 'yes/no' questions.

We hope the perceptual score spurs a discussion regarding the perceptiveness of multi-modal models and we also hope to encourage the community working on multi-modal classifiers to start quantifying perceptiveness of models.

Our contributions:

- We propose the Perceptual score, a simple yet effective method for assessing the perceptiveness of multi-modal models towards a modality.
- Our experiments span multiple datasets and models. We find that multi-modal models tend to ignore some modalities while taking shortcuts.
- Consequently, we investigate the sources of bias on popular multi-modal datasets

such as VQA-CP and SocialIQ: We find that SocialIQ is biased by sentiment, and bias in the VQA-CP model results from shifting the training priors to more closely resemble those in the test set.

7.1 Related Work

The perceptual score assesses the degree to which a classifier relies on a particular input modality. This is related to studying datasets and their biases, methods which aim to reduce the biases captured by classifiers and work which studies the importance of features. We review all three areas next.

Datasets and bias: Data has been a central element for machine learning progress [LBBH98, NNM96, TKSDM03] in the last two to three decades. The ImageNet challenge [DDS+09] and the development of AlexNet [KSH12] sparked the deep learning era. But as datasets grow, biases emerge which may go undetected for a long time. For instance, the background in ImageNet can reveal information about the object class [SWY+15]. Also, with the increasing popularity of crowdsourcing systems like Amazon Mechanical Turk, many datasets are annotated in uncontrolled environments. Different annotators are hence injecting unknown socio-economic properties into dataset annotations [GGB19]. Those dataset biases can be detrimental to the considered task. For instance, meticulously collected visual question answering (VQA) data [AAL+15] aims to provide a platform for exciting research to advance image-language understanding. However, it is non-trivial to remove biases from this type of data. Indeed, it was reported that the question solely is sufficient to detect the correct answer [JJvdM16], *i.e.*, no image information is required. In an attempt to fix this bias, the dataset was re-annotated [GKS+17], or the train and test split were re-organized [ABPK18]. Similarly, Schwartz *et al.* [SSH19] reported that in Audio-visual-scene-aware dialog (AVSD) [ACD+18], the question cue is often stronger, making the desired video and sound reasoning implausible or unnecessary. Likewise, SNLI [BAPM15], aims to determine the correctness of a hypothesis given a premise. However, Gururangan *et al.* [GSL+18] point out that an internal bias exists: linguistic features not related to the premise correlate with the label. Further, the Story Cloze [MCH+16] dataset permits to develop models which estimate the correct ending of a story.Schwartz *et al.* [SSK+17] show that, for this dataset, length alone is a powerful feature to determine the correct ending.

Recently, Zadeh *et al.* [ZCL+19] proposed SocialIQ, a dataset intended to reason about social situations in videos, specifically, emotion detection in a social situation. In this dataset, given a video and a question, the correct social situation should be recognized, *e.g.*, "The man is upset because he is being insulted." We show that it is easy to pick the correct statement using only the text data. This is possible because the statement's correctness often correlates with the statement's sentiment. Again, much like for VQA and AVSD, image information doesn't seem to be necessary for reasonably accurate performance. Hence, biases exist which permit to 'address the

dataset' without addressing the task. For SocialIQ, we attribute these biases to the fact that social reasoning is considered difficult, even for humans. Hence an annotator's expertise is particularly important.

Importantly, going forward, we think it is elusive that we will be able to create unbiased datasets. We hence need techniques to automatically measure biases. In this work, we provide a mechanism that permits to do this for any dataset.

Methods to reduce bias: Several methods have been suggested to reduce bias from a dataset. Some techniques require prior knowledge of the biased variables, for instance, gender bias in vision is addressed by masking related features (*e.g.*, faces) [AZN18, BB19, DBW19, HBS+18, JOM+19, KJ19, KKK+18, WZY+19, ZWY+18, ZMWC19]. Also, some techniques require access to the test set to re-balance it [ZWY+17]. Additionally, various approaches were proposed for visual question answering [CDC+19, CYZ19, RAL18]. These methods use a classifier trained only on the question modality to regularize bias directly.Cadene *et al.* [CDC+19] suggest to mask a model's softmax prediction with a softmax prediction of a subset classifier trained on the question modality.Clark *et al.* [CYZ19] use an ensemble method of a full classifier paired with a question subset classifier using products of experts [Hin02]. Common to all those approaches is the use of a subset classifier to indicate reliance on subset of the data. In contrast, we assess the perceptiveness of the original model based on permutations of subset of data.

Feature importance methods: Also relevant to our work are techniques that measure feature-importance. These works generally focus on understanding why a model made a particular prediction. To this end, Tulio *et al.* [RSG16] introduce LIME, a local explanation method that approximates a linear explanation model around a given example. Later, Shrikumar *et al.* [SGK17] proposed DeepLift specifically for deep net models. Lundberg *et al.* [LL17] introduced SHapley Additive exPlanations (SHAP), a method to estimate feature importance using an approximation of the features Shapley values. LIME assumes that the local model is linear, while SHAP does not have any such assumptions. Lundberg *et al.* [LL17] show a connection between DeepLIFT and Shapely values introduced as 'Deep SHAP.' Each of those techniques concentrates on the importance of features, whereas we focus on the importance of a modality as a whole.

7.2 The Perceptual Score

The *perceptual score* quantifies the degree to which a model relies on the input data or a subset thereof. Said differently, the perceptual score assesses the importance of data or a subset of the data for a model's result. For instance, if a model answers questions about an image without using cues extracted from the image modality, the model does not perceive the image modality. In this case we want the perceptual score for the image modality to be lower than that of a model which relies heavily on the

image modality.

We believe that reporting the perceptual score of a model in addition to its accuracy is particularly important in a multi-modal setting. As a community we are developing increasingly complex models. However, to date we know very little about what parts of the data these models rely on. We think this lack in our understanding is due to a missing easy way to quantify how much a model relies on available data modalities. To rectify this we hope to illustrate that a simple yet intuitive score like the proposed perceptual score is very useful and easy to report too.

The following sections describe the perceptual score of a modality for a given model and dataset.

7.2.1 Setup

Let $\mathcal{D} = \{(x_1, y_1), \ldots, (x_{|\mathcal{D}|}, y_{|\mathcal{D}|})\}$ denote the test set data, where $|\mathcal{D}|$ refers to the number of samples in the dataset. Each sample is a pair consisting of the input data $x_i \in \mathcal{X}$ and its corresponding label $y_i \in \mathcal{Y}$. Here, $\mathcal{Y}$ is a finite set of possible classes. In multi-modal datasets which we consider here, the data x_i can be separated naturally into different parts. For instance, in the SocialIQ task, we can partition the data into video-related, question-related and answer-related parts. Formally, let $\mathcal{M} = \{M_1, \ldots, M_{|\mathcal{M}|}\}$ be a set of modalities of size $|\mathcal{M}|$, *e.g.*, the video-, the question- and the answer-modality. We partition the data x_i into its modalities using a set notation, *i.e.*, $x_i = \{x_i^{M_1}, \ldots, x_i^{M_{|\mathcal{M}|}}\}$. We use $x_i^{\{M_1, M_2\}} = \{x_i^{M_1}, x_i^{M_2}\}$ to refer to the first two modalities, *i.e.*, the superscript can be a set.

7.2.2 Perceptual Score of a Data Modality

The perceptual score $P_{f,\mathcal{D}}(M_m)$ of a model $f : \mathcal{X} \to \mathcal{Y}$ towards modality M_m on data $\mathcal{D}$ is defined as

$$P_{f,\mathcal{D}}(M_m) = \frac{1}{Z}\left(\mathbb{E}_{(x,y)\sim\mathcal{D}}[P_{f,x,y}(M_m)]\right), \tag{7.1}$$

i.e., as the normalized expectation of sample perceptual scores $P_{f,x,y}(M_m)$. Here, Z indicates the normalization factor, which can either be determined by the dataset alone (*i.e.*, $Z = Z_{\mathcal{D}}$) or based on both the dataset and the model (*i.e.*, $Z = Z_{f,\mathcal{D}}$). We discuss the normalization in Sec. 7.2.2.

The sample perceptual score $P_{f,x,y}(M_m)$ aims to measure the degree to which a model $f : \mathcal{X} \to \mathcal{Y}$ relies on a modality M_m for prediction of sample $x \in \mathcal{X}$. To do so we define the sample perceptual score for modality $M_m \in \mathcal{M}$ as the normalized difference between the accuracy of a model which uses all data modalities, and the accuracy of a model which doesn't use modality M_m for prediction, *i.e.*, as

$$P_{f,x,y}(M_m) = \mathrm{Acc}_{f,x,y}(\mathcal{M}) - \mathrm{Acc}_{f,x,y}(\mathcal{M} \setminus \{M_m\}). \tag{7.2}$$

Here, $\mathcal{M} \setminus \{M_m\}$ is an operator that removes the influence of modality M_m from

the set of all modalities $\mathcal{M}$. We define this operation formally in Sec. 7.2.2. Note, $\text{Acc}_{f,x,y}(\mathcal{M})$ refers to the classical prediction accuracy of a dataset sample (x, y) for a given trained model f with $\mathcal{M}$ the set of all modalities used for prediction, *i.e.*, $\text{Acc}_{f,x,y}(\mathcal{M}) = \mathbb{1}_{f(x^{\mathcal{M}})=y}$.

Intuitively, the sample perceptual score $P_{f,x,y}(M_m)$ is high if the accuracy of a model that does not consider modality M_m for prediction is significantly smaller than the accuracy of a model which uses all data modalities $\mathcal{M}$. Conversely, if the accuracy doesn't change, irrespective of whether modality M_m is available or not, the model f doesn't perceive the modality M_m. Note that in cases where the modality M_m irritates the model, the perceptual score can be negative.

In the following we discuss how to 'remove' a modality M_m from a model f (Sec. 7.2.2) and how to compute the normalization constant Z (Sec. 7.2.2).

Removing Modality Influence

Removing a modality from a trained model is difficult since typical models entangle modalities and compute high-order correlations. Ideally, we need a tool that minimizes the impact of one modality while maintaining the other components' functionality. To achieve this, we study a permutation-based approach, *i.e.*, we randomly permute the modality-related features among the test set data $\mathcal{D}$. We think permutation is particularly useful for the perceptual score because it is hyper-parameter free. This ensures that the perceptual score defined in Eq. (7.1) is unambiguous.

Formally, to compute $\text{Acc}_{f,x,y}(\mathcal{M} \setminus \{M_m\})$ we don't use all modalities from data x_i. Instead, we use all modalities but M_m, *i.e.*, $x_i^{\mathcal{M}\setminus M_m}$ and append the data $x_j^{M_m}$ of modality M_m from another data point x_j. Hereby j is drawn uniformly from $\{1, \ldots, |\mathcal{D}|\}$, *i.e.*, from $\mathcal{U}(1, |\mathcal{D}|)$. Taken together we compute the accuracy via

$$\text{Acc}_{f,x,y}(\mathcal{M} \setminus \{M_m\}) = \mathbb{E}_{j\sim\mathcal{U}(1,|\mathcal{D}|)}[\mathbb{1}_{f(\{x^{\mathcal{M}\setminus M_m}, x_j^{M_m}\})=y}]. \tag{7.3}$$

In the following we discuss how to compute the normalization Z employed in Eq. (7.1).

Normalization

Normalization enables comparability. For this, normalization aims for consistency of the perceptual score between models designed for the same task and between models designed for different tasks. Two types of normalization are useful to consider: 1) a *task-normalization*, which enables a more meaningful comparison of the perceptual score across different tasks; and 2) a *model-normalization* that enables a meaningful comparison of the perceptual score within the same task. We think models should be analyzed with both kinds of normalization in mind.

Task-normalization: The degree of difficulty of the task matters when comparing

a model's perceptual score for a modality. Without normalization, the comparison of models designed for different tasks is inconsistent. For instance, if a task is relatively easy, removing a modality won't significantly affect the accuracy. This would result in a perceptual score close to zero, which isn't compelling because, in this case, even a marginal reduction in performance might be significant.

To incorporate the difficulty of a task we compare the perfect accuracy to the accuracy of a majority vote classifier. We note that the majority vote classifier always predicts the majority class of the employed training set. We use $\widehat{\text{Acc}}_{\mathcal{D}}$ to refer to the accuracy of the majority vote classifier evaluated on the test set $\mathcal{D}$ and compute the normalization factor via

$$Z_{\mathcal{D}} = 1 - \widehat{\text{Acc}}_{\mathcal{D}}. \tag{7.4}$$

Intuitively, the normalization factor is the gap between the perfect accuracy (*i.e.*, 1) and the accuracy of the majority vote classifier, *i.e.*, $\widehat{\text{Acc}}_{\mathcal{D}}$. We note that this normalization may result in perceptual score higher than one since the majority vote may be superior to the permuted accuracy. However, this case is unlikely in practice and did not occur in any of our experiments. Notably, this normalization is limited by the fact that the model accuracy is not considered, as we will discuss next.

Model-normalization: The performance of a model is an important factor. Suppose for two different classifiers permutation of a modality results in the same accuracy as that of the majority vote classifier. As a result, the initially stronger model will attain a higher perceptual score. This property can be desirable since one may want the perceptual score to reflect both perception and the ability to address a task. However, in this case the perceptual score does not solely reflect the degree to which a model considers a particular modality for decision making. Instead the perceptual score for a modality would be conflated with the model's accuracy.

To obtain a score which solely reflects the degree to which a model considers a particular modality, we normalize by the model's accuracy. Formally, we normalize via

$$Z_{\mathcal{D},f} = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\text{Acc}_{f,x,y}(\mathcal{M})]. \tag{7.5}$$

7.3 Evaluation of Perceptual Scores

In the following, we assess the perceptual score using popular multi-modal datasets. Specifically, in Sec. 7.3.1, we study visual question answering (VQA, VQA-CP). We examine video social reasoning (*i.e.*, SocialIQ) in Sec. 7.3.2. In Sec. 7.3.3, we assess visual dialog models. Our analysis shows that state-of-the-art models exploit biases that haven't been documented. We study the bias by investigating samples with low perceptual scores and discover its cause.

Experimental setup: We train each model with five different seeds and report the mean accuracy. We compute the perceptual score based on five permutations per

Table 7.1: Accuracy and perceptual scores on VQAv2 and VQA-CP for different baselines and question types: number (Num), yes/no (Y/N), and other. We report the accuracy ($\text{Acc}_{\mathcal{M}}$), the accuracy after removing a modality's influence ($\text{Acc}_{\mathcal{M}\setminus\{V\}}$, $\text{Acc}_{\mathcal{M}\setminus\{Q\}}$), the perceptual score without normalization (P_V, P_Q), the perceptual score with task normalization ($P_V/Z_{\mathcal{D}}$, $P_Q/Z_{\mathcal{D}}$), the perceptual score with model normalization ($P_Q/Z_{\mathcal{D},f}$, $P_V/Z_{\mathcal{D},f}$), and majority-vote accuracy ($\widehat{\text{Acc}}_{\mathcal{D}}$). Means and standard deviations are provided.

			Image				Question				
Model	Q. Type	$\text{Acc}_{\mathcal{M}}$	$\text{Acc}_{\mathcal{M}\setminus\{V\}}$	P_V	$P_V/Z_{\mathcal{D}}$	$P_V/Z_{\mathcal{D},f}$	$\text{Acc}_{\mathcal{M}\setminus\{Q\}}$	P_Q	$P_Q/Z_{\mathcal{D}}$	$P_Q/Z_{\mathcal{D},f}$	$\widehat{\text{Acc}}_{\mathcal{D}}$
VQAv2											
LXMERT	All	68.97	36.46	32.51	47.40 ± 0.40	47.16 ± 0.37	27.41	41.56	60.60 ± 0.38	60.26 ± 0.38	31.42
LMH	All	54.33	27.90	26.43	38.54 ± 0.32	48.64 ± 0.29	22.82	31.51	45.94 ± 0.30	57.99 ± 0.31	31.42
BAN	All	65.67	35.09	30.58	44.59 ± 0.38	46.56 ± 0.39	28.25	37.43	54.57 ± 0.36	56.99 ± 0.32	31.42
BUTD	All	63.09	34.38	28.71	41.86 ± 0.36	45.52 ± 0.38	28.78	34.31	50.03 ± 0.35	54.33 ± 0.34	31.42
LXMERT	Num	52.73	14.55	38.18	47.14 ± 0.28	72.40 ± 0.00	13.04	39.69	49.01 ± 0.25	49.00 ± 0.25	19.01
LMH	Num	37.58	15.64	21.94	27.09 ± 0.26	58.38 ± 0.26	13.75	23.82	29.41 ± 0.22	63.40 ± 0.27	19.01
BAN	Num	48.62	18.55	30.07	37.13 ± 0.28	61.84 ± 0.29	16.07	32.55	40.19 ± 0.24	66.95 ± 0.27	19.01
BUTD	Num	42.46	21.04	21.41	26.44 ± 0.25	50.47 ± 0.24	18.39	24.06	29.71 ± 0.22	56.29 ± 0.26	19.01
LXMERT	Other	60.86	16.81	44.05	45.63 ± 0.27	72.37 ± 0.27	3.47	57.39	59.45 ± 0.08	94.30 ± 0.09	3.47
LMH	Other	54.29	13.72	40.58	42.04 ± 0.21	74.73 ± 0.21	4.30	49.99	51.79 ± 0.09	92.08 ± 0.08	3.47
BAN	Other	56.99	16.37	40.62	42.08 ± 0.25	71.27 ± 0.25	4.17	52.82	54.72 ± 0.10	92.68 ± 0.06	3.47
BUTD	Other	54.99	14.29	40.70	42.16 ± 0.22	73.93 ± 0.22	4.53	50.46	52.27 ± 0.09	91.74 ± 0.07	3.47
LXMERT	Y/N	85.30	64.58	20.72	41.02 ± 0.40	24.29 ± 0.32	63.84	21.46	42.49 ± 0.32	25.16 ± 0.38	49.49
LMH	Y/N	60.24	50.80	9.43	18.68 ± 0.33	15.66 ± 0.42	50.29	9.94	19.68 ± 0.29	16.50 ± 0.39	49.49
BAN	Y/N	83.03	65.44	17.59	34.82 ± 0.36	21.18 ± 0.29	64.09	18.93	37.48 ± 0.33	22.80 ± 0.32	49.49
BUTD	Y/N	80.94	65.42	15.52	30.73 ± 0.32	19.27 ± 0.33	64.24	16.69	33.05 ± 0.30	20.63 ± 0.25	49.49
VQA-CP											
CSS	All	57.89	36.46	21.43	26.43 ± 0.41	37.01 ± 0.27	7.93	49.96	61.61 ± 0.21	86.30 ± 0.31	18.91
RMFE	All	54.20	29.91	24.29	27.11 ± 0.34	47.17 ± 0.31	7.65	46.55	51.95 ± 0.17	55.91 ± 0.24	10.40
LMH	All	52.23	27.14	25.09	28.00 ± 0.32	45.35 ± 0.24	7.13	45.10	50.33 ± 0.14	54.08 ± 0.27	10.40
CSS	Num	51.34	44.50	6.84	7.20 ± 0.38	13.32 ± 0.08	13.04	38.30	40.34 ± 0.29	74.60 ± 0.31	5.06
RMFE	Num	44.03	37.25	6.77	7.13 ± 0.33	18.84 ± 0.33	34.05	9.97	10.51 ± 0.30	26.51 ± 0.33	5.06
LMH	Num	37.40	30.35	7.05	7.43 ± 0.28	15.39 ± 0.28	27.48	9.92	10.45 ± 0.25	22.65 ± 0.28	5.06
CSS	Other	46.48	10.12	36.36	37.57 ± 0.15	78.22 ± 0.15	4.12	42.36	43.77 ± 0.16	91.13 ± 0.14	3.23
RMFE	Other	45.97	10.08	35.89	37.09 ± 0.17	77.94 ± 0.18	4.15	41.82	43.22 ± 0.11	91.14 ± 0.21	3.23
LMH	Other	46.10	10.17	35.93	37.13 ± 0.18	78.07 ± 0.19	4.08	42.02	43.42 ± 0.11	90.96 ± 0.23	3.23
CSS	Yes/No	83.11	82.52	0.59	0.71 ± 0.04	3.16 ± 0.04	43.84	39.27	100.0 ± 0.00	47.25 ± 0.33	64.46
RMFE	Yes/No	74.47	62.49	11.98	18.58 ± 0.31	17.99 ± 0.31	59.31	15.16	23.51 ± 0.28	21.68 ± 0.31	35.52
LMH	Yes/No	73.75	60.49	13.27	20.58 ± 0.32	16.08 ± 0.32	57.76	16.00	24.81 ± 0.29	20.35 ± 0.32	35.52

sample and report mean along with the standard deviations. For all the models, we used the official implementations.

7.3.1 Visual Question Answering

The visual question answering task reasons about an image given a question. We use the VQAv2 dataset [GKS+17], which contains 443,757 image-question pairs in the train set and 214,354 in the validation set. We also assess the perceptiveness of models trained on Visual Question Answering: Changing Priors (VQA-CP) data [ABPK18], which was released after several studies suggested that VQAv2 models heavily rely on answer priors. For instance, 'how many' questions are typically answered with '2.' To overcome this shortcoming, VQA-CP suggested a new train-test-split. As a result, the train and test sets have different prior distributions for each question type. The new split consists of 438,183 training samples, and 219,928 samples for validation.

Table 7.2: Proportion of yes/no ratios for different kinds of questions. The initial token categorizes questions. We report proportion in the train set, the test set, and the model prediction. '# Train' indicates the number of samples in the train set, '# Test' is the number of samples in the test set.

Token	Model	Predicted Yes	Predicted No	Test Yes	Test No	Train Yes	Train No	# Test	# Train
'has'	CSS	1.0	0.0	0.88	0.12	0.5	0.5	1330	2784
	LMH	0.47	0.53	0.88	0.12	0.0	1.0	1330	1392
'can'	CSS	0.94	0.06	0.65	0.35	0.5	0.5	2664	2670
	LMH	0.38	0.62	0.65	0.35	0.0	1.0	2664	1235
'is a'	CSS	0.01	0.99	0.0	1.0	0.5	0.5	305	610
	LMH	0.37	0.63	0.0	1.0	0.89	0.11	305	305
'do'	CSS	0.99	0.01	0.9	0.1	0.5	0.5	1328	10844
	LMH	0.42	0.58	0.9	0.1	0.39	0.61	1328	5422

Baselines: We use four baselines for VQAv2: 1) BUTD [AHB+18], an early competitive approach that used detector-based features pre-trained on VisualGenome [KZG+17]; 2) BAN [KJZ18], a baseline that uses an effective multi-modal bilinear attention; 3) LMH [CYZ19], originally crafted for VQA-CP, this approach removes superficial question patterns; and 4) LXMERT [TB19] a large-scale Transformer-based model that is pre-trained with large amounts of image-text pairs.

Quantitative Analysis

In Tab. 7.1, we provide the perceptual score and the accuracy for different baselines and questions. We start by analyzing the perceptual scores of the vision and language modalities. In most cases, models perceive the question better (*i.e.*, $P_Q > P_V$). Studying the accuracy, LXMERT has the highest accuracy (68.97%). In contrast, LMH, which reduces the reliance on question priors, achieves the lowest accuracy (54.33%). However, the model-normalized perceptual score for the visual modality ($P_V/Z_{\mathcal{D},f}$) suggests: LMH perceives image data similarly to other models. Studying the task normalized perceptual score for the visual modality ($P_V/Z_{\mathcal{D}}$) suggests: despite the high accuracy on 'Y/N' questions, their image and question perceptual scores are very low, *i.e.*, models mostly ignore the visual data and rely on priors.

Next, we show metrics for VQA-CP, a variant designed to reduce bias caused by answer priors. We compare different models on all the data by analyzing the model-normalized perceptual score for both visual and question modalities ($P_V/Z_{\mathcal{D},f}$ and $P_Q/Z_{\mathcal{D},f}$). Interestingly, the state-of-the-art model, CSS, has the lowest image perceptual score (37.01%) and the highest question perceptual score (86.30%), suggesting that the question modality may serve as a shortcut to answer without perceiving the image. Further, by analyzing the different question types via the task-normalized score ($P_V/Z_{\mathcal{D}}$), we note that CSS has a perceptual score for the image modality of only 3.6% for 'Y/N' questions. One possible explanation: CSS generates new samples, which in turn alter priors. We further investigate this bias next.

Table 7.3: Accuracy and perceptual scores on SocialIQ. For different modalities (1st column), we report the accuracy ($\text{Acc}_{\mathcal{M}}$), the accuracy after removing the modality's influence ($\text{Acc}_{\mathcal{M}\setminus\{M\}}$), the perceptual score without normalization (P_M), the perceptual score with task normalization ($P_M/Z_{\mathcal{D}}$), the perceptual score with model normalization ($P_M/Z_{\mathcal{D},f}$), and the majority-vote accuracy ($\widehat{\text{Acc}}_{\mathcal{D}}$).

Modality M	Model	$\text{Acc}_{\mathcal{M}}$	$\text{Acc}_{\mathcal{M}\setminus\{M\}}$	P_M	$P_M/Z_{\mathcal{D}}$	$P_M/Z_{\mathcal{D},f}$	$\widehat{\text{Acc}}_{\mathcal{D}}$
Answer	Baseline	64.84	56.73	8.11	18.92 ± 0.13	12.51 ± 0.03	57.14
	FGA	67.38	57.11	10.27	23.96 ± 0.21	15.24 ± 0.10	57.14
Question	Baseline	64.84	64.32	0.52	1.21 ± 0.02	0.80 ± 0.03	57.14
	FGA	67.38	65.19	2.19	5.11 ± 0.11	3.25 ± 0.08	57.14
Video	Baseline	64.84	63.79	1.05	2.45 ± 0.05	1.62 ± 0.03	57.14
	FGA	67.38	64.42	2.96	6.91 ± 0.16	4.39 ± 0.15	57.14
-	NLTK-Sentimant	66.70	56.14	11.18	26.08 ± 0.19	16.36 ± 0.09	57.14
-	Answer-Only	68.65	57.29	11.39	26.57 ± 0.21	16.59 ± 0.06	57.14

Bias Analysis for CSS

The Counterfactual Samples Synthesizing (CSS) model produces counterfactual training samples by masking either critical objects in images or words in questions, and by assigning different ground-truth answers. Our perceptiveness study above shows that the CSS model has a significantly lower perceptual score for the visual modality, despite being state-of-the-art on VQA-CP with a substantial accuracy gap of 6.5% over LMH. Why?

The first thing to note: CSS generates new samples which may shift prior distributions. Recalling that VQA-CP was introduced to prevent benefiting from answer priors in VQAv2 reveals a potential reason for the improvements of CSS.

Use of the sample perceptual score permits a more in-depth analysis. In Tab. 7.2, we identify popular question start tokens with low sample perceptual scores for the visual modality ('do,' 'has,' 'can,' 'is a'). We further examine their prediction accuracy using both CSS and LMH. We find that the proportion of 'yes' and 'no' answers between the train and the test set differ: the yes answer is correct for 88% of the questions in the test set starting with 'has,' while the yes answer is *never* correct for the corresponding training set questions. For CSS, counterfactual samples, however, produce equal proportions. As a result, CSS seemingly alleviates the prior inconsistency and adjusts the majority of its predictions to 'yes,' which more closely resembles the test set.

Use of the perceptual scores hence permits to hypothesize: improvements in CSS can be attributed to a shifted prior distribution instead of a complex counterfactual data-manipulation. To test this hypothesis we train the LMH model using samples obtained from CSS training. Importantly, we did not modify the image or the question. *Without even any changes to the input*, we obtain an accuracy of 57.54%, only 0.3% lower than CSS and within standard deviation.

Figure 7.2: SocialIQ data samples. On the left, we show a sample with a high perceptual score towards video data. Neither a positive nor a negative sentiment is evident in this sample. Hence, the video is required for prediction. We illustrate two samples (marked with a red border) that received a low perceptual score. There is a sentiment-based correlation between the label and the answer in these samples. For simplicity, we highlight with red color words that exhibit sentiment.

Table 7.4: Perceptual scores on VisDial v1.0. We show that the perceptual score can be computed using metrics other than accuracy (*i.e.*, MRR and NDCG).

		MRR						NDCG					
Modality M	Model	$\text{MRR}_{\mathcal{M}}$	$\text{MRR}_{\mathcal{M}\setminus\{M\}}$	P_M	$P_M/Z_{\mathcal{D}}$	$P_M/Z_{\mathcal{D},f}$	$\widehat{\text{MRR}}_{\mathcal{D}}$	$\text{NDCG}_{\mathcal{M}}$	$\text{NDCG}_{\mathcal{M}\setminus\{M\}}$	P_M	$P_M/Z_{\mathcal{D}}$	$P_M/Z_{\mathcal{D},f}$	$\widehat{\text{NDCG}}_{\mathcal{D}}$
Question	LS (CE)	52.21	32.48	19.73	29.11 ± 0.02	37.79 ± 0.03	32.22	75.24	20.35	54.89	73.86 ± 0.04	72.95 ± 0.03	25.68
	LS	69.00	47.96	21.04	31.04 ± 0.03	30.49 ± 0.03	32.22	64.89	23.92	40.97	55.13 ± 0.02	63.14 ± 0.01	25.68
	FGA	66.14	32.23	33.91	50.03 ± 0.02	51.26 ± 0.03	32.22	56.00	30.74	25.26	33.99 ± 0.02	45.11 ± 0.02	25.68
Image	LS (CE)	52.21	34.94	17.27	25.48 ± 0.02	33.08 ± 0.02	32.22	75.24	57.45	17.79	23.94 ± 0.02	23.64 ± 0.03	25.68
	LS	69.00	52.06	16.94	24.99 ± 0.01	24.55 ± 0.03	32.22	64.89	51.50	13.39	18.02 ± 0.01	20.63 ± 0.03	25.68
	FGA	66.14	52.15	14.00	20.65 ± 0.01	21.16 ± 0.02	32.22	56.00	46.73	9.27	9.27 ± 0.03	16.55 ± 0.02	25.68
Caption	LS (CE)	52.21	52.18	0.03	0.00 ± 0.00	0.06 ± 0.00	32.22	75.24	75.19	0.05	0.07 ± 0.00	0.07 ± 0.01	25.68
	LS	69.00	69.00	0.00	0.00 ± 0.00	0.00 ± 0.00	32.22	64.89	64.89	0.00	0.00 ± 0.00	0.00 ± 0.00	25.68
	FGA	66.14	64.93	1.21	1.79 ± 0.02	1.83 ± 0.02	32.22	56.00	55.32	0.68	0.92 ± 0.01	1.22 ± 0.01	25.68

7.3.2 Video Social Reasoning

SocialIQ [ZCL+19] proposes an unconstrained benchmark, specifically designed to understand social situations. More concretely, given an input tuple of a video, a question, and an answer, the task is to predict whether the answer is correct or not. The videos were collected from YouTube and annotated by students. The dataset is split into 37,191 training samples, and 5,320 validation set samples.

Baseline: We were able to reproduce the SocialIQ baseline [ZCL+19] and achieve an accuracy of 64.84% using all feature modalities during training. In fact, we achieved an accuracy of 67.38% by changing the baseline's model to FGA [SYHS19] and by using GloVe [PSM14] instead of BERT features [DCLT19].

Quantitative Analysis

In Tab. 7.3 we show scores for different metrics. The task-normalized perceptual scores ($P_M/Z_{\mathcal{D}}$) for different modalities M (1st column) of FGA and the baseline reveals heavy reliance on the answer modality, and low dependence on the video modality (*i.e.*, 15.24 *vs.* 3.25). To verify that the answer is indeed the only important modality, we train a classifier using *only* the answer modality and observe: we can surpass the original paper's baseline by 4% when *only* using answer features, achieving 68.65% accuracy. Again, the perceptual score helped us understand shortcomings of a model. Upon analyzing the bias source, we observe a sentiment bias, which we discuss next.

Bias Analysis for SocialIQ

Assessing samples with low sample perceptual scores can reveal biases. Our study suggests that the labels correlate strongly with the sentiment. In Fig. 7.2, we marked with a red border two videos with a low sample perceptual score for the video modality. Studying those and similar samples we find: 1) when the answer is True, the answer has positive sentiment; 2) in contrast, when the answer is False, the answer contains words with negative connotation (*e.g.*, 'hostile,' 'unfriendly').

We hypothesize: successful prediction of the answer by just looking at the answer modality for SocialIQ data is due to sentiment-biased annotations. To validate this hypothesis, we use an off-the-shelf sentiment classifier from the NLTK package [LB04]. When applied to SocialIQ *without any training*, we obtain a remarkable answer prediction accuracy of 66.7%. This matches our reported result of 68.65% quite reasonably and outperforms the SocialIQ baseline [ZCL+19].

7.3.3 Visual Dialog

The visual dialog task encourages models to ask *and* answer questions about visual input. Notably, each dialog-interaction employs many modalities (*e.g.*, image, question, caption, dialog-history). We show our results on the VisDial v1.0 dataset, where 123,287 images are used for training, 2,000 images for validation, and 8,000 images for testing [DKG+18]. Each image is associated with ten questions, and each question has 100 corresponding answer candidates.

Instead of accuracy $\text{Acc}_{f,x,y}(\mathcal{M})$, here, we use ranking-based metrics as the Visual Dialog dataset primarily uses two metrics: MRR and NDCG. In short, the MRR metric examines the rank of a single ground-truth response (*i.e.*, sparse annotations), while the NDCG metric measures the cumulative gain in the case of multiple correct answers (*i.e.*, dense annotations). We computed the majority ranking based on answer frequency in the train set. See appendix for more.

Baselines: We use two baselines for VisDial v1.0: 1) FGA [SYHS19], an attention unit inspired from graphical models that infers an attention map for each modality; and 2) LS [MBPD20], which pre-trains on related vision-language datasets, *e.g.*, Conceptual Captions and Visual Question Answering [SDGS18, GKS+17]. We also report LS (CE), which finetunes on the dense annotations, at the expense of MRR performance.

Quantitative Analysis

In Tab. 7.4, we show the perceptual scores for different modalities (*i.e.*, M), and for two metrics: MRR and NDCG. We find: 1) using the MRR metric, the model-normalized perceptual score for the question modality of FGA is relatively high compared to LS (*i.e.*, 51.26% *vs.* 30.49%). However, when we employ the NDCG metric, the LS model perceives the question better (*i.e.*, 63.14% *vs.* 45.11%). 2) analyzing the only model optimized for NDCG, *i.e.*, LS (CE), reveals: When analyzed using the NDCG metric,

the model-normalized perceptual score for both image and question is significantly higher than for MRR. Interestingly, the LS (CE) model-normalized perceptual score for the image modality is on par with the other models when we use the MRR metric. However, the question perceptual score is low, which suggests the reason for the low MRR performance may be related to the low utilization of the question. Finally, we note that all models, on both metrics, seem to ignore the caption, which suggests caption information is redundant.

7.4 Conclusion

We introduced the perceptual score of a multi-modal classifier towards a data modality. The perceptual score assesses a classifier's perceptiveness of a modality and reveals exciting insights if analyzed carefully. We hope that this is demonstrated by our studies which reveal that 1) shifted prior distributions seam to help CSS achieve state-of-the-art results, 2) SocialIQ data exhibits a sentiment bias, and 3) Visual Dialog caption information appears less informative. We hope that researchers working on multi-modal models see the use of the perceptual score and will start to report the perceptual score in addition to classical accuracy metrics.

Limitations: We propose perceptual scores, a novel metric that conveys information about multi-modal classifiers. The two limitations we can see: 1) a small computational overhead; 2) a false sense of security. Perceptual scores don't alleviate the need to carefully study results. However we think the societal and scientific benefits of reporting this novel metric outweigh any concerns.

Chapter 8

Removing Bias in Multi-modal Classifiers: Regularization by Maximizing Functional Entropies

Multi-modal data is ubiquitous and commonly used in many real-world applications. For instance, discriminative visual question answering systems take into account the question, the image and a variety of answers. In general, we treat data as multi-modal if it can be partitioned into semantic features, *e.g.*, color and shape can be treated as multi-modal data.

Training of discriminative classifiers on multi-modal datasets like discriminative visual question answering almost always follows the classical machine learning paradigm: use a common loss function like cross-entropy and employ a standard ℓ_2-norm regularizer (a.k.a. weight decay). The regularizer favors 'simple' classifiers over more complex ones. These classical regularizers are suitable in traditional machine learning settings that predominantly use a single data modality. Unfortunately, because they favor 'simple' models, their use is detrimental when learning from multi-modal data. Simplicity encourages use of information from a single modality, which often ends up biasing the learner. For instance, visual question answering models end up being driven by a language prior rather than visual understanding [ABPK18, JHvdM$^+$17, GKS$^+$17, ABP16]. *E.g.*, answering 'how many...?' questions with '2' regardless of the question. Another popular example consists of colored images whose label is correlated with their color modality and their shape modality. In these cases, standard learners often focus on the 'simple' color modality and largely ignore the shape modality [LV19, KKK$^+$18].

To address this issue, we develop a novel regularization term based on the functional entropy. Intuitively, this term encourages to balance the contribution of each modality to classification. To address the computational challenges of computing the functional entropy we develop a method based on the log-Sobolev inequality which bounds the functional entropy with the functional Fisher information.

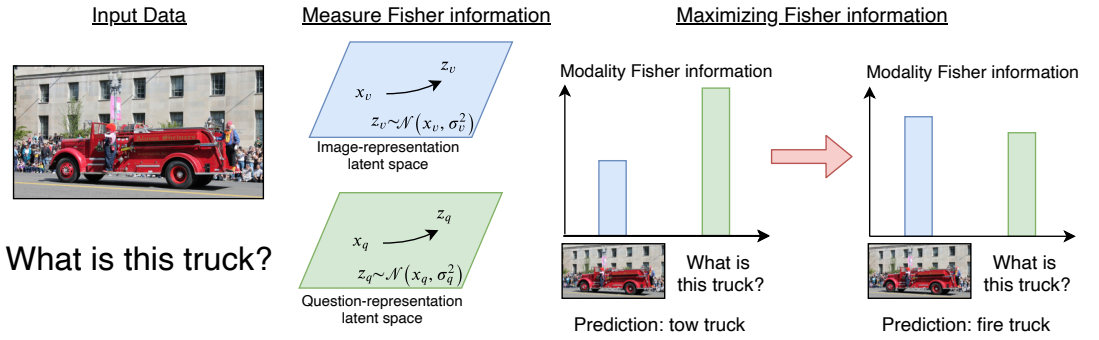


Figure 8.1: We illustrate our approach. In the visual question answering task, we are given a question about an image. Thus, we can partition our input into two modalities: a textual modality, and a visual modality. We measure the modalities' functional Fisher information by evaluating the sensitivity of the prediction by perturbing each modality. We maximize the functional Fisher information by incorporating it into our loss as a regularization term. Our results show that our regularization permits higher utilization of the visual modality.

We illustrate the efficacy of the proposed approach on the three challenging multi-modal datasets Colored MNIST, VQA-CPv2, and SocialIQ. We find that our regularization maximizes the utilization of essential information. We verify this empirically on the synthetic dataset Colored MNIST. We also evaluate on popular benchmarks, finding that our method permits a state-of-the-art performance on two datasets: SocialIQ (68.53% *vs.* 64.82%) and VQA-CPv2 (54.55% *vs.* 52.05%).

8.1 Related Work

Multi-modal datasets. Over the years, the amount and variety of data that has been used across tasks has grown significantly. Unsurprisingly, present-day tasks are increasingly sophisticated and combine multiple data modalities like vision, text, and audio. In particular, in the past few years, many large-scale multi-modal datasets have been proposed [JHvdM+17, GKS+17, ZCL+19, ZBFC19, HM19, DKG+18]. Subsequently, multiple works developed strong models to address these datasets [SSH17, KJZ18, SYHS19, LBPL19, AHB+18, TB19, LYBP16, FPY+16, HM18, ADS18, JZS17a, JLS18]. However, recent work also suggests that many of these advanced models predict by leveraging one of the modalities more than the others, *e.g.*, utilizing question type to determine the answer in VQA problems [ABPK18, SSH19, GSL+18, SSK+17]. This property is undesirable since multi-modal tasks consider all data essential to solve the challenge without overfitting to the dataset.

Bias in datasets. Recently, datasets were proposed to study whether a model can generalize and solve the task or whether it uses a single modalities' features. Usually, this evaluation is performed by partitioning data into train and test sets using different distributions. For example, VQA-CP [ABPK18] is a reshuffle of the VQA [GKS+17]

dataset ensuring that question-type distributions differ between train and test splits. Another well-known dataset is Colored MNIST [LV19, KKK$^+$18, ABGLP19]. In this dataset, each digit class is colored differently in the train set, while samples in the test set remain gray-scale. Different approaches were proposed to deal with such problems: Arjovsky *et al.* [ABGLP19] propose to improve generalization by ensuring that the optimal classifier equals all training distributions. Wang *et al.* [WTF20] suggest to regularize the overfitting behavior to different modalities. Methods like REPAIR [LV19] prevent a model from exploiting dataset biases by re-sampling the training data. Kim *et al.* [KKK$^+$18] use an adversarial approach to learn unbiased feature representations. Clark *et al.* [CYZ19] and Cadene *et al.* [CDC$^+$19] suggest methods to overcome language priors using a bias-only model in VQA tasks.

Entropy and information in deep nets. Entropy plays a pivotal role in machine learning and has been extensively used in losses and for regularization [JGJS99]. However, its use is confined to probability distributions while we use functional entropy, which has a different form and is defined for any non-negative function. More broadly, other components of information theory have been studied in deep nets, for example, the information bottleneck criteria [TZ15, SZT17]. Other works use information theory to overcome generalization [KKK$^+$18, KBFF19, RAL18]. For instance, Krishna *et al.* [KBFF19] propose to maximize the mutual information of the modalities by regularizing differences between modality representations. Fisher information is also used in various machine learning and deep learning settings, *e.g.*, monitoring of the learning process [LDRC20]. In contrast, our work considers the *functional* Fisher information of a non-negative function that represents a multi-modal learner, while Fisher information is defined over probability density functions. Also, we use the log-Sobolev inequality between the functional entropy and the functional Fisher information, which does not hold for entropy and Fisher information.

8.2 Background

Discriminative learning constructs mapping between data-instance $x \in \mathcal{X}$ and labels $y \in \mathcal{Y}$ given training data S. We are particularly interested in multi-modal data, where each data-instance x is composed of multiple modalities. For example, a monochrome image x is composed of two data modalities $x = (x_c, x_s)$ where $x_c \in \mathbb{R}^3$ is the monochromatic color tone and $x_s \in \mathbb{R}^{d \times d}$ is the $d \times d$ intensity map of the image capturing the shape. Similarly, in discriminative visual question answering, x is composed of a visual modality, a question modality and an answer modality, *i.e.*, $x = (x_v, x_q, x_a)$ with $x_v \in \mathbb{R}^{d_v}$, $x_q \in \mathbb{R}^{d_q}$ and $x_a \in \mathbb{R}^{d_a}$ respectively. Generally, x may have n modalities, *i.e.*, $x = (x_1, \ldots, x_n)$, each residing in Euclidean space, *i.e.*, $x_i \in \mathbb{R}^{d_i}$.

Discriminative learning searches for the parameters w of a function which assign a score to each label y given data x. In this work we focus on the softmax function $p_w(\hat{y}|x)$. Its goodness of fit is measured by a loss function, often the cross-entropy

loss $\mathrm{CE}(\mathbb{1}[\cdot = y], p_w(\cdot|x)) = -\sum_{\hat{y}} \mathbb{1}[\hat{y} = y] \log p_w(\hat{y}|x)$, where $\mathbb{1}$ refers to the indicator function. More generally, the cross-entropy loss between two distributions $p_w(\hat{y}|x), q(\hat{y})$ is

$$\mathrm{CE}(q, p_w) = -\sum_{\hat{y}} q(\hat{y}) \log p_w(\hat{y}|x). \tag{8.1}$$

Beyond the loss, a typical learning process employs a regularization term which encourages use of the 'simplest' function. Various regularization terms that favor 'simple' functions pose a considerable difficulty for multi-modal problems: deep learners easily find simple functions that ignore one of the modalities. For example, a simple discriminator for Colored MNIST, which consists of monochromatic images whose colors correlate with their labels, focuses almost exclusively on the color vector to predict the label rather than also assessing the shape of the image. Formally, if the monochromatic images are represented by their color and shape modalities $x = (x_c, x_s)$ then the simplest discriminator will only consider the 3-dimensional color x_c. In this setting, the learned function $p_w(\hat{y}|x)$ avoids all important information within the shape modality x_s.

In the following we describe the notion of functional entropy in Section 8.2.1. In Section 8.2.2 we present the log-Sobolev inequality, which bounds the functional entropy of a non-negative function by the functional Fisher information. We conclude with the notion of tensorization, which decomposes these components according to their multi-modal spaces.

8.2.1 Functional entropy

In this work we consider the functional entropy that is encapsulated in multi-modal problems. Functional entropies are defined over a continuous random variable, *i.e.*, a function $f(z)$ over the Euclidean space $z \in \mathbb{R}^d$ with a probability measure μ. Here and throughout we use z to refer to a stochastic variable, which we integrate over. The functional entropy of a non-negative function $f(z) \geq 0$ is

$$\mathrm{Ent}_\mu(f) \triangleq \int_{\mathbb{R}^d} f(z) \log f(z) d\mu(z) - \left(\int_{\mathbb{R}^d} f(z) d\mu(z)\right) \log \left(\int_{\mathbb{R}^d} f(z) d\mu(z)\right) \tag{8.2}$$

The functional entropy is non-negative, namely $\mathrm{Ent}_\mu(f) \geq 0$ and equals zero only if $f(z)$ is a constant. This is in contrast to differential entropy of a continuous random variable with probability density function $q(z)$: $h(q) = -\int_{\mathbb{R}^d} q(z) \log q(z) dz$, which is defined for $q(z) \geq 0$ with $\int_{\mathbb{R}^d} q(z) dz = 1$ and may be negative.

8.2.2 Functional Fisher information

Unfortunately, the functional entropy is hard to estimate empirically, since it involves the term $\log(\int_{\mathbb{R}^d} f(z) d\mu(z))$. Since the integral can only be estimated by sampling, the

logarithm of its estimate is hard to compute in practice. Instead of estimating the functional entropy directly, we use the log-Sobolev inequality for Gaussian measures (cf. [BGL13], Section 5.1.1). This permits to bound the functional entropy with the functional Fisher information. Specifically, for any non-negative function $f(z) \geq 0$ we obtain

$$\mathrm{Ent}_\mu(f) \leq \frac{1}{2} \int_{\mathbb{R}^d} \frac{\|\nabla f(z)\|^2}{f(z)} d\mu(z). \tag{8.3}$$

Hereby, $\|\nabla f(z)\|$ is the ℓ_2 norm of the gradient of f. The functional Fisher information is non-negative, since it is defined for non-negative functions. It is a natural extension of the Fisher information, which is defined for probability density functions.

8.2.3 Tensorization and multi-modal data

Functional entropy naturally fits into multi-modal settings that correspond to product probability spaces. For example, when considering discriminative visual-question answering, a data point $x = (x_v, x_q, x_a)$ resides in the Euclidean product space of the visual modality x_v, the question modality x_q and the answer modality x_a. This product space property is called tensorization and informally relates the functional entropy of each modality to the overall functional entropy of the system. Generally, consider the product space $\hat{z} = (\hat{z}_1, \ldots, \hat{z}_n)$, where each modality resides in the d_i-dimensional Euclidean space $\hat{z}_i \in \mathbb{R}^{d_i}$. Consider the product measure $\mu = \mu_1 \otimes \cdots \otimes \mu_n$ and let

$$f_i(z_i) = f(\hat{z}_1, \ldots, \hat{z}_{i-1}, z_i, \hat{z}_{i+1}, \ldots, \hat{z}_n). \tag{8.4}$$

The tensorization of the functional entropy amounts to

$$\mathrm{Ent}_\mu(f) \quad \leq \quad \sum_{i=1}^{n} \int_{\mathbb{R}^d} \mathrm{Ent}_{\mu_i}(f_i) d\mu(\hat{z}), \tag{8.5}$$

Here the dimension d is the dimension of $\hat{z} = (\hat{z}_1, \ldots, \hat{z}_n)$, namely $\hat{z} \in \mathbb{R}^d$ and $d = \sum_{i=1}^n d_i$. Tensorization is well-suited for multi-modal settings, as it provides the means to bound the overall functional entropy of the system using the functional entropies of its modalities.

8.3 Regularization by Maximizing Functional Entropies

Functional entropy requires a probability measure. In the following we differentiate between multi-modal training points $x = (x_1, \ldots, x_n)$ and general multi-modal points in the probability measure space, which we denote by $z = (z_1, \ldots, z_n)$. We use the training points $x = (x_1, \ldots, x_n)$ to determine the measure and we denote by $z = (z_1, \ldots, z_n)$ the variable of the integrands. In our work we consider a Gaussian product. Given a

training point $x \in S$ that resides in the multi-modal space $x = (x_1, \ldots, x_n)$ we define the measure μ_i^x for the i-th modality to be the Gaussian distribution with mean x_i and variance $\sigma_{x_i}^2$, where x_i is the i-th modality of the training point x and $\sigma_{x_i}^2$ is the variance of the coordinate of x_i:

$$\mu_i^x \triangleq \mathcal{N}(x_i, \sigma_{x_i}^2). \tag{8.6}$$

The measure μ^x is the product measure over the different modalities $\mu^x \triangleq \mu_1^x \otimes \cdots \otimes \mu_n^x$. For example, given a monochromatic image $x = (x_c, x_s)$ in the training data, the distribution employed by the functional entropy in Eq. (8.2) is $\mu^x = \mathcal{N}(x_c, \sigma_{x_c}^2) \otimes \mathcal{N}(x_s, \sigma_{x_s}^2)$.

For each training data point $x \in S$, we define the functional entropy over the deep net softmax function $p_w(\cdot|x)$ as

$$f^x(z_1, \ldots, z_n) \triangleq \mathrm{CE}(p_w(\cdot|z), p_w(\cdot|x)). \tag{8.7}$$

This function measures the sensitivity of the softmax prediction to Gaussian perturbations z of the input, since the random perturbation z is sampled from a Gaussian with an expected value x, as described in Eq. (8.6).

The cross-entropy function is a non-negative function, therefore, it is natural to apply the log-Sobolev inequality for Gaussian measures to bound the functional entropy using the functional Fisher information, in Eq. (8.3):

$$\mathrm{Ent}_{\mu^x}\left(\mathrm{CE}(p_w(\cdot|z), p_w(\cdot|x))\right) \leq \int_{\mathbb{R}^d} \frac{\|\nabla_z \mathrm{CE}(p_w(\cdot|z), p_w(\cdot|x))\|^2}{\mathrm{CE}(p_w(\cdot|z), p_w(\cdot|x))} d\mu^x(z). \tag{8.8}$$

We use the functional Fisher information bound in Eq. (8.3) to regularize the training process, in order to implicitly encourage to maximize the information of each modality, while minimizing the training loss. In order to account for both the loss minimization and the information maximization, we take the inverse information. Given multi-modal training data S, our learning objective is

$$\sum_{(x,y)\in S} \mathrm{CE}(\mathbb{1}[\cdot = y], p_w(\cdot|x)) + \lambda \sum_{(x,y)\in S} \left(\int_{R^d} \frac{\|\nabla_{z^x} \mathrm{CE}(p_w(\cdot|z^x), p_w(\cdot|x))\|^2}{\mathrm{CE}(p_w(\cdot|z^x), p_w(\cdot|x))} d\mu_x(z) \right)^{-1} \tag{8.9}$$

The hyperparameter λ balances between the training loss and the inverse information.

8.3.1 Tensorization

The tensorization argument in Section 8.2.3 determines a bound on the functional entropy by its functional entropy over each modality. The tensorization argument is favorable since it permits to consider the functional entropy of each modality separately in the integral of Eq. (8.5), given a point $\hat{z} = (\hat{z}_1, \ldots, \hat{z}_n)$. The tensorization also

permits to efficiently approximate the functional entropy, given a training point x: Let $z_i^x \triangleq (x_1, \ldots, x_{i-1}, z_i, x_{i+1}, \ldots, x_n)$ and set $\tilde{f}_i^x(z_i) \triangleq f^x(z_i^x)$. Given this definition, the tensorization in Eq. (8.5) reduces to

$$\mathrm{Ent}_{\mu^x}(f^x) \quad \leq \quad \sum_{i=1}^{n} \int_{\mathbb{R}^d} \mathrm{Ent}_{\mu_i^x}(f_i^x) d\mu(\hat{z}) \approx \sum_{i=1}^{n} \mathrm{Ent}_{\mu_i^x}(\tilde{f}_i^x). \tag{8.10}$$

We combine this approximation with the log-Sobolev inequality to measure the amount of the functional Fisher information added by each modality, for a given multi-modal training point $x = (x_1, \ldots, x_n)$:

$$\sum_{i=1}^{n} \mathrm{Ent}_{\mu_i^x}\left(\mathrm{CE}(p_w(\cdot|z_i^x), p_w(\cdot|x))\right) \leq \sum_{i=1}^{n} \int_{\mathbb{R}^{d_i}} \frac{\left\|\nabla_{z_i^x} \mathrm{CE}(p_w(\cdot|z_i^x), p_w(\cdot|x))\right\|^2}{\mathrm{CE}(p_w(\cdot|z_i^x), p_w(\cdot|x))} d\mu_i^x(z_i) \tag{8.11}$$

We recall that $z_i^x \triangleq (x_1, \ldots, x_{i-1}, z_i, x_{i+1}, \ldots, x_n)$ and $z_i \in \mathbb{R}^{d_i}$ is the variable that is being integrated while all other modalities remain fixed to the training point input modality.

Similarly to Eq. (8.9), we may use the tensorized functional Fisher information bound in Eq. (8.11) to regularize the training process. Given multi-modal training data S, our tensorized learning objective is

$$\sum_{(x,y)\in S} \mathrm{CE}(\mathbb{1}[\cdot = y], p_w(\cdot|x)) + \lambda \sum_{(x,y)\in S} \sum_{i=1}^{n} \left(\int_{\mathbb{R}^{d_i}} \frac{\left\|\nabla_{z_i^x} \mathrm{CE}(p_w(\cdot|z_i^x), p_w(\cdot|x))\right\|^2}{\mathrm{CE}(p_w(\cdot|z_i^x), p_w(\cdot|x))} d\mu_i^x(z_i) \right)^{-1} \tag{8.12}$$

8.4 Connection Between Functional Entropy and Variance

Rothaus [Rot85] has shown a connection between the functional entropy of a non-negative function and its variance.

$$\mathrm{Var}_\mu(f) \quad \triangleq \quad \int_{\mathbb{R}^d} f^2(z) d\mu(z) - \left(\int_{\mathbb{R}^d} f(z) d\mu(z) \right)^2. \tag{8.13}$$

Particularly, when the values of the non-negative function $f(z)$ are small, one can expand the Taylor series of $1 + f(z)$ to show that

$$\mathrm{Ent}_\mu(1 + f) = \mathrm{Var}_\mu(f) + o(\|f\|_\infty^2), \tag{8.14}$$

where the residual function $o(t)$ is non-negative and approaches zero faster than t approaches zero, *i.e.*, $\lim_{t\to 0} \frac{o(t)}{t} = 0$. Interestingly, a similar bound to the log-Sobolev inequality (Eq. (8.3)) exists for the variance of continuous random variables $f(z)$, which is widely known as the Poincaré inequality:

$$\mathrm{Var}_\mu(f) \leq \int_{\mathbb{R}^d} \|\nabla f(z)\|^2 \, d\mu(z). \tag{8.15}$$

The relation between the functional entropy and the variance, expressed in Eq. (8.14), suggests that these bounds should behave similarly in practice. To fit the variance into multi-modal settings we need to show tensorization (as in Sec. 8.2.3). In the variance case, this property is called the Efron-Stein theorem (cf. [Led01], Proposition 2.2),

$$\mathrm{Var}_{\mu}(f) \quad \leq \quad \sum_{i=1}^{n} \int_{\mathbb{R}^d} \mathrm{Var}_{\mu_i}(f_i) d\mu(\hat{z}). \tag{8.16}$$

Next, we present a similar regularization term to the one described in Sec. 8.3. This time we use variance and Poincaré inequality.

8.4.1 Regularization using Variance

In Sec. 8.3, we were interested in bounding the functional entropy (Eq. (8.8)), for each training point $x \in S$, of $\mathrm{CE}(p_w(\cdot|z), p_w(\cdot|x))$. Similarly, we want to bound the variance of $\mathrm{CE}(p_w(\cdot|z), p_w(\cdot|x))$. For this purpose, we can use the Poincaré inequality, described in Eq. (8.15),

$$\mathrm{Var}_{\mu^x}\left(\mathrm{CE}(p_w(\cdot|z), p_w(\cdot|x))\right) \leq \int_{\mathbb{R}^d} \left\|\nabla_z \mathrm{CE}(p_w(\cdot|z), p_w(\cdot|x))\right\|^2 d\mu^x(z). \tag{8.17}$$

We use the above inequality to regularize the training process. To consider both the loss minimization and the regularization term we formulate the learning objective,

$$\sum_{(x,y)\in S} \mathrm{CE}(\mathbb{1}[\cdot = y], p_w(\cdot|x)) + \lambda \sum_{(x,y)\in S} \left(\int_{\mathbb{R}^d} \left\|\nabla_{z^x} \mathrm{CE}(p_w(\cdot|z^x), p_w(\cdot|x))\right\|^2 d\mu_x(z) \right)^{-1} \tag{8.18}$$

To fit our multi-modal settings, we need to follow the tensorization process as illustrated in Sec. 8.3.1. The same tensorization process can be applied to the variance using the Poincaré bound given in Eq. (8.15). For tensorized Poincaré bound leads to the learning objective

$$\sum_{(x,y)\in S} \mathrm{CE}(\mathbb{1}[\cdot = y], p_w(\cdot|x)) + \lambda \sum_{(x,y)\in S} \sum_{i=1}^{n} \left(\int_{\mathbb{R}^{d_i}} \left\|\nabla_{z_i^x} \mathrm{CE}(p_w(\cdot|z_i^x), p_w(\cdot|x))\right\|^2 d\mu_i^x(z_i) \right)^{-1} \tag{8.19}$$

8.5 Experiments

In the following, we evaluate our proposed regularization on four different datasets. One of the datasets is a synthetic dataset (Colored MNIST), which permits to study whether a classifier leverages the wrong features. We show that adding the discussed regularization improves the generalization of a given classifier. We briefly describe each dataset and discuss the results of the proposed method.

Table 8.1: Comparison between our proposed regularization terms on the Colored MNIST (multi-modal settings, gray-scale test set), SocialIQ [ZCL+19] and Dogs & Cats [KKK+18] datasets. We report maximum accuracy observed and accuracy after convergence of the model (Convg). We compare the 4 regularizers specified by the equation numbers. We underline the highest maximum accuracy and bold the highest results after convergence. Using functional Fisher information regularization (Eq. (8.12)) leads to a smaller difference between the maximum accuracy and accuracy after convergence. * refers to results we achieve without using our proposed regularization. ** denotes training with weight-decay (ℓ_2 regularization).

Model	Colored MNIST	
	Convg.	Max
Baseline*	41.11±2.13	98.31
Baseline**	47.32±1.12	98.23
Eq. (8.2)	93.68±0.75	94.44
Eq. (8.13)	94.87±1.03	96.37
Eq. (8.12)	**96.17**±0.63	98.38
Eq. (8.19)	**96.24**±0.74	<u>98.52</u>

(a) Comparison on Colored MNIST.

Model	SocialIQ	
	Convg.	Max
Baseline*	63.91±0.26	66.16
Baseline**	65.28±0.23	66.95
Eq. (8.2)	63.87±0.34	64.22
Eq. (8.13)	64.36±0.31	64.93
Eq. (8.12)	**67.93**±0.18	<u>68.53</u>
Eq. (8.19)	67.41±0.21	68.19

(b) Comparison on SocialIQ.

Model	Dogs & Cats (TB1)		Dogs & Cats (TB2)	
	Convg.	Max	Convg.	Max
Baseline*	79.22±0.45	80.12	65.51±1.54	67.38
Baseline**	81.24±0.23	84.31	68.47±0.29	71.36
Eq. (8.2)	92.92±0.46	93.48	85.32±0.41	85.79
Eq. (8.13)	93.38±0.27	94.15	85.14±0.29	85.41
Eq. (8.12)	**94.71**±0.37	<u>95.99</u>	**88.11**±0.17	<u>88.48</u>
Eq. (8.19)	**94.43**±0.24	95.35	87.81±0.31	88.12

(c) Comparison on Dogs & Cats.

8.5.1 Colored MNIST

Dataset: Colored MNIST [LV19, KKK+18] is a colored variant of MNIST [LBBH98]. The train and validation set consist of 60,000 and 10,000 samples, respectively. Each sample is biased with a color that correlates with its digit. The biasing process assigns to each digit an RGB vector which represents a mean color. Then, each sample receives its color, sampled from a normal distribution with a fixed variance around the digit's mean color. This process results in a monochromatic image and high correlation between the digit's color and its label. To introduce a bias in a multi-modal approach, we split each sample x into a color modality x_c and a shape (gray-scale representation of the image) modality x_s. For humans it is evident that a digit should be classified based on its shape and not its color. For a learner this fact is not as clear. To minimize the loss, it is much easier for a classifier to leverage the color modality, which correlates very well with the label. In its nature, Colored MNIST evaluates the generalization of a model since it has a test set that assesses whether a classifier relies solely on color or both the color and the shape.

Baseline: A simple deep net achieves high accuracy on both colored train and

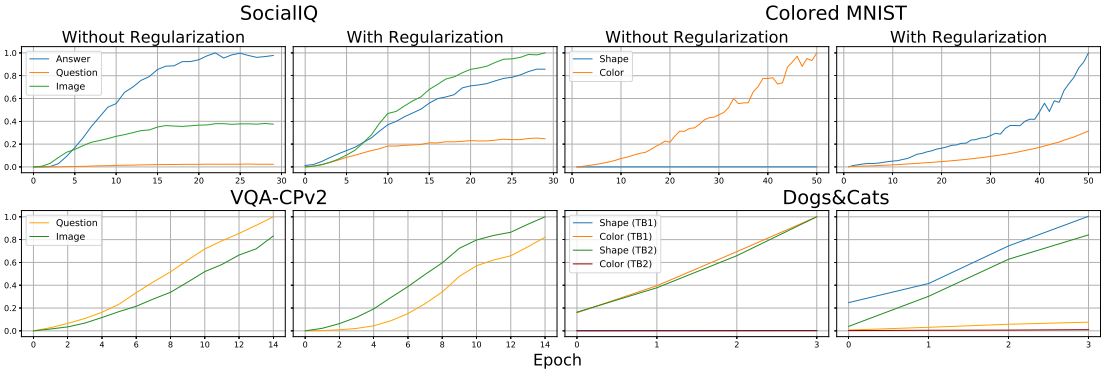


Figure 8.2: Proportions of the Fisher information values during training for SocialIQ, Colored MNIST, VQA-CPv2 and Dogs&Cats. Using our proposed regularization brings the modalities Fisher information value closer than training without our regularization, a desired property in multi-modal learning. In ColoredMNIST, we observe that training a model with our regularization, the prediction is based on both the shape and the color. Unlike, a model trained without our regularization which makes predictions based on the color only.

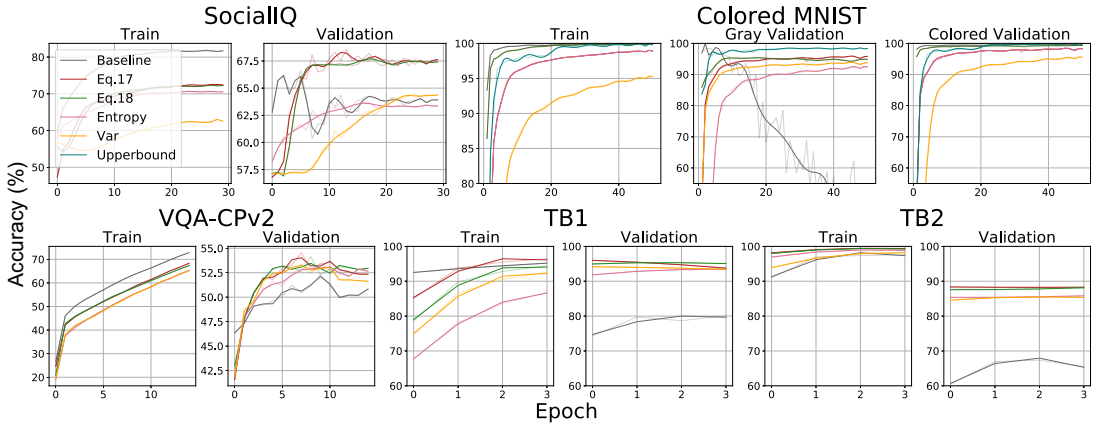


Figure 8.3: Training process with and without regularization. We note that generalization significantly improves when using our proposed regularization.

colored validation set. However, on the gray-scale validation set, the network fails drastically, achieving only a 41.11% accuracy when using the model from the last training epoch. We note that the more we train the more the baseline relies on color rather than shape. We also compute an upper-bound by training the deep net on a gray-scale version. The upper-bound accuracy on the gray-scale validation set is 98.47%.

Results: Adding our proposed regularization encourages to exploit information from both shape and color modalities. We provide results in Tab. 8.1. Fig. 8.2 shows that without entropy regularization, the Fisher information value of the shape is almost zero while adding the regularization results in a higher shape information value than the color. This fact complements the classifier's performance on the gray-scale validation set shown in Fig. 8.3. Using functional Fisher information based regularization outperforms the same classifier trained without regularization by almost 55%.

Table 8.2: Comparison between the state-of-the-art on the VQA-CPv2 test set. The best results for each category are in bold. * denotes models that make use of external data.

Model	Overall	Answer type		
		Yes/No	Number	Other
BUTD [AHB+18]	39.34	42.13	12.29	45.29
AdvReg [RAL18]	41.17	65.49	15.48	35.48
HINT [SLS+19]*	46.73	67.27	10.61	45.88
RUBi [CDC+19]	47.11	68.65	20.28	43.18
SCR [WM19]*	49.45	72.36	10.93	**48.02**
LMH [CYZ19]	52.05	72.58	31.12	46.97
LMH +Ours Eq. (8.19)	54.01 ± 0.27	73.02 ± 1.21	43.15 ± 1.01	47.02 ± 0.28
LMH +Ours Eq. (8.12)	**54.55** ± 0.29	**74.03** ± 1.13	**49.16** ± 1.22	45.82 ± 0.37

8.5.2 VQA-CPv2

Dataset: VQA-CPv2 [ABPK18] is a re-shuffle of the VQAv2 [GKS+17] dataset. Visual question answering (VQA) requires to answer a given question-image pair. [ABPK18] observed that the original split of the VQAv2 dataset permits to leverage language priors. To challenge models to not use these priors, the question type distributions of the train and validation set were changed to differ from one another. VQA-CPv2 consist of 438,183 samples in the train set and 219,928 samples in the test set.

Results: We evaluated our method by adding functional Fisher information regularization to the current state-of-the-art [CYZ19]. In doing so, the result improves by 2.5%, achieving 54.55% accuracy. We provide a comparison with recent state-of-the-art methods in Tab. 8.2.

The authors of [RSK20, GB19] raise the concern that new regularization methods mainly boost the performance of yes/no questions. Investigating the improvements due to our result shows that this is not the case. The accuracy difference to the previous state-of-the-art on the different answer types is: yes/no +1.5%, number +18%, and other -1%.

8.5.3 SocialIQ

Dataset: The SocialIQ dataset is designed to develop models for understanding of social situations in videos. Each sample consists of a video clip, a question, and an answer. The task is to predict whether the answer is correct or not given this tuple. The dataset is split into 37,191 training samples, and 5,320 validation set samples. Note that an inherent bias exists in this dataset: specifically the sentiment of the answer provides a good cue.

Baseline: A simple classifier based on only the answer modality performs significantly better than chance level accuracy (using our settings ~6% more). Such biases in the train set lead to a classic case of overfitting.

Results: As seen in Fig. 8.3, training without functional Fisher information regularization leads to ~80% accuracy on the train set and ~64% accuracy on the validation set. Although, functional Fisher information regularization results in 70% accuracy on the train set, it improves validation set accuracy to 67.93% accuracy.

We further investigate the information values during the training phase with and without functional Fisher information regularization. In Fig. 8.2 we observe that without our regularization, the answer modality has the highest information value while the question modality is almost entirely ignored. Adding the proposed regularization balances the information between modalities, the desired behavior in multi-modal learning.

8.5.4 Dogs and Cats

Dataset: Following the settings of Kim *et al.* [KKK+18], we evaluate our models on the biased "Dogs and Cats" dataset. This dataset comes in two splits: The TB1 set consists of bright dogs and dark cats and contains 10,047 samples. The TB2 set consist of dark dogs and bright cats and contains 6,738 samples. We use the image as a single-modality.

Baseline: The authors show that training of ResNet-18 [HZRS15a] on TB1 and testing on TB2 results in a poor performance of 74.98%. The authors also show that using TB2 as the train set and TB1 as the test set results in even worse accuracy of 66.45%.

Functional Fisher information regularization training on TB1 and testing on TB2 with λ (see Eq. (8.12)) set to equal 3e-10 results in 94.71% accuracy, exceeding [KKK+18] by 3.5%. Training on TB2 while testing on TB1 achieves an accuracy of 88.11%, 1% higher than [KKK+18].

8.6 Conclusion

Classical regularizers applied on multi-modal datasets lead to models which may ignore one or more of the modalities. This is sub-optimal as we expect all modalities to contribute to classification. To alleviate this concern we study regularization via the functional entropy. It encourages the model to more uniformly exploit the available modalities.

Chapter 9

Conclusions

Since computer science emerged a century ago, it has grown exponentially and developed algorithms that can classify images better than humans [HZRS15a].

Nevertheless, the intricacies of how a machine is different from the human mind remain a mystery. Although a machine has a massive amount of computational power that can easily surpass human capabilities at the game of Go [SHM+16], it is still unable to achieve an adequate level of fidelity for human-like conversation task (*i.e.*, the Turing test[1]). To help reduce the fog around the human-like cognitive goal of machines, we find it helpful to break down the task into the following specific objectives: 1) Perception, *i.e.*, the organization and identification of sensory information 2) Comprehension, *i.e.*, the ability to process abstract or physical objects, understand their meaning and integrate with the already known parts. 3) Attention, *i.e.*, the process of selectively concentrating on a discrete aspect of information, whether considered subjective or objective, while ignoring other perceivable information. 4) Problem solving, *i.e.*, the process of using a set of strategies such as simulation, computer modeling, and experiment to find solutions to problems in an orderly manner. 5) Decision-making, *i.e.*, selecting a belief or a course of action among several possible alternative options, could be rational or irrational.

In this dissertation, we've thoroughly studied two of these concepts using deep learning models: (i) attention; and (ii) perception. Despite their opposite natures, both are essential to cognitive development.

We study attention with a proposal of a new general model, namely Factor Graph Attention, that can attend any input type with any number of modalities. Using our model, we address many recent and challenging tasks and reached new state-of-the-art performance.

We first introduced an attention unit with three types of potentials: unary, pairwise, and trenary. While previous works concentrated on co-attention for two modalities, our new attention unit allowed us to address a task with three modalities, the multiple-choice visual question answering, question modality, image modality, and an-

[1] `https://en.wikipedia.org/wiki/Turing_test`

swer modality.

We extended our attention unit to any input modalities, introducing Factor Graph Attention, a module inspired from graphical models and built with pairwise interaction factors. With this module, we achieved state-of-the-art performance for the task of Visual Dialog. We also study ways to train ensemble models with dense and sparse annotations. Notably, our study results in the winning entry to the Visual Dialog 2020 challenge [2].

We employ factor graph attention to the task of visual storytelling. This task requires generating a story for a given sequence of images. To achieve this, we develop ordered image attention (OIA).

Finally, we propose a simple baseline for Audio-Visual Scene-Aware Dialog that includes different modalities, *e.g.*, audio, video, and dialog.

Next, we studied perception and introduced the perceptual score of a multi-modal classifier towards a data modality. The perceptual score assesses a classifier's perceptiveness of a modality and reveals exciting insights if analyzed carefully.

We study regularization via functional entropy that encourages the model to exploit the available modalities more uniformly.

Finally, we observe that despite recent rapid progress, deep learning models still have much room for improvement in solving problems. A neural network's decision is instinctive and lacks essential elements, such as variables (*i.e.*, memory storage), logic, and planning. For instance, in the visual question answering task, counting questions are challenging [ZHPB18], requiring a counting variable and iterative logic. A promising line of works employs symbolic program execution for reasoning [YWG$^+$18, YGL$^+$19, VDL$^+$19]. The facilitation of AI that is able to plan and reason with logic is an essential and fascinating area of future research.

[2] `https://visualdialog.org/challenge/2020`

Bibliography

[AAL+15] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015.

[ABGLP19] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[ABP16] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. In *EMNLP*, 2016.

[ABPK18] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, 2018.

[ACD+18] H. Alamri, V Cartillier, A. Das, J. Wang, J. Essa, D. Batra, D. Parikh, A. Cherian, T. K. Marks, and C. Hori. Audio visual scene-aware dialog (avsd) challenge at dstc7. *arXiv preprint arXiv:1806.00525*, 2018.

[ADS18] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. Convolutional image captioning. In *CVPR*, 2018.

[AHB+18] P. Anderson, 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 *CVPR*, 2018.

[ARDK16] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Learning to compose neural networks for question answering. *NAACL*, 2016.

[AVT16] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *NIPS*, 2016.

[AZN18] Mohsan S. Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *ECCV (Workshop)*, 2018.

[BAPM15] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015.

[BB19] Shikha Bordia and Samuel R. Bowman. Identifying and reducing gender bias in word-level language models. In *NAACL (Workshop)*, 2019.

[BCB14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[BCF13] Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. Cache-based online adaptation for machine translation enhanced computer assisted translation. In *MT Summit*, 2013.

[BGL13] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*. SBM, 2013.

[BL05] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL*, 2005.

[BSF+03] Kobus Barnard, Pinar Duygulu Sahin, David A. Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. Matching words and pictures. *JMLR*, 2003.

[BYCCT17] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *ICCV*, 2017.

[CCFC02] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *ICALP*, 2002.

[CDC+19] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. In *NeurIPS*, 2019.

[Che18] Guillaume Chevalier. Larnn: linear attention recurrent neural network. *arXiv preprint arXiv:1808.05578*, 2018.

[CKF11] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, 2011.

[CS18] Moitreya Chatterjee and Alexander G Schwing. Diverse and coherent paragraph generation from images. In *ECCV*, 2018.

[CYZ19] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *EMNLP*, 2019.

[CZ15] Xinlei Chen and C. Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*, 2015.

[CZ17] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[DAZ+16] Abhishek Das, Harsh Agrawal, C Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *EMNLP*, 2016.

[DBW19] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In *ACL*, 2019.

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019.

[DDS+09] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[DHG+15] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015.

[DKG+18] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Stefan Lee, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *PAMI*, 2018.

[DVSC+17] H. De Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, 2017.

[EML+18] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation. In *SIGGRAPH*, 2018.

[FPY+16] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *EMNLP*, 2016.

[GB10] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010.

[GB19] Gabriel Grand and Yonatan Belinkov. Adversarial regularization for visual question answering: Strengths, shortcomings, and side effects. In *NAACL*, 2019.

[GCK+19] Zhe Gan, Yu Cheng, Ahmed EI Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. Multi-step reasoning via recurrent dual attention for visual dialog. *ACL*, 2019.

[GGB19] M. Geva, Y. Goldberg, and J. Berant. Are we modeling the task or the annotator? an investigation ofannotator bias in natural language understanding datasets. In *EMNLP*, 2019.

[GGHY15] D. Geman, S. Geman, N. Hallonquist, and L. Younes. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, 2015.

[GJY+19] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *CVPR*, 2019.

[GKS+17] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*, 2017.

[GMZ+15] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? Dataset and Methods for Multilingual Image Question Answering. In *NIPS*, 2015.

[GRP18] Diana Gonzalez-Rico and Gibran Fuentes Pineda. Contextualize, show and tell: A neural visual storyteller. In *Storytelling Workshop, NAACL*, 2018.

[GSL+18] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *NAACL (Short Papers)*, 2018.

[GXT19] Dalu Guo, Chang Xu, and Dacheng Tao. Image-question-answer synergistic network for visual dialog. In *CVPR*, 2019.

[HAR+17] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko. Learning to reason: End-to-end module networks for visual question answering. *ICCV*, 2017.

[HAW+18] C. Hori, H. Alamri, J. Wang, G. Winchern, T. Hori, A. Cherian, T.K. Marks, V. Cartillier, R.G. Lopes, A. Das, I. Essa, D. Batra, and D. Parikh. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. *arXiv*, 2018.

[HBD+20] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLR*, 2020.

[HBS+18] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, 2018.

[HCE+17] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson. Cnn architectures for large-scale audio classification. In *ICASSP*, 2017.

[HCH+20] Chao-Chun Hsu, Zi-Yuan Chen, Chi-Yang Hsu, Chih-Chia Li, Tzu-Yuan Lin, Ting-Hao Huang, and Lun-Wei Ku. Knowledge-enriched visual storytelling. In *AAAI*, 2020.

[HFM+16] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. In *NAACL*, 2016.

[HG +18] Qiuyuan Huang, Zhe Gan, Asli Çelikyilmaz, Dapeng Wu, Jianfeng Wang, and Xiaodong He. Hierarchically structured reinforcement learning for topically coherent visual story generation. In *AAAI*, 2018.

[Hin02] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 2002.

[HKG+15] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *NIPS*, 2015.

[HM18] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In *ICLR*, 2018.

[HM19] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

[HS97a] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997.

[HS97b] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. In *Neural Computation*, 1997.

[HZRS15a] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2015.

[HZRS15b] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *CVPR*, 2015.

[IS15] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[JGJS99] Michael Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence Saul. An introduction to variational methods for graphical models. *Learning in Graphical Models*, 1999.

[JHvdM+17] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

[JJvdM16] Allan Jabri, Armand Joulin, and Laurens van der Maaten. Revisiting visual question answering baselines. In *ECCV*, 2016.

[JLS18] U. Jain, S. Lazebnik, and A. G. Schwing. Two can play this game: Visual dialog with discriminative question generation and answering. *CVPR*, 2018.

[JOM+19] Julio Cezar Silveira Jacques, Cagri Ozcinar, Marina Marjanovic, Xavier Baró, Gholamreza Anbarjafari, and Sergio Escalera. On the effect of age perception biases for real age regression. *FG 2019*, 2019.

[JYQ+20] Xiaoze Jiang, Jing Yu, Zengchang Qin, Yingying Zhuang, Xingxing Zhang, Yue Hu, and Qi Wu. Dualvd: An adaptive dual encoding model for deep visual understanding in visual dialogue. *AAAI*, 2020.

[JZS17a] U. Jain*, Z. Zhang*, and A. G. Schwing. Creativity: Generating Diverse Questions using Variational Autoencoders. In *CVPR*, 2017. * equal contribution.

[JZS17b] Unnat Jain*, Ziyu Zhang*, and A. G. Schwing. Creativity: Generating Diverse Questions using Variational Autoencoders. In *CVPR*, 2017. * equal contribution.

[KB14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *CoRR*, 2014.

[KBFF19] Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Information maximizing visual question generation. In *CVPR*, 2019.

[KDHR17] Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. Structured attention networks. *ICLR*, 2017.

[KE17] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*, 2017.

[KFF15] A Karpathy and L Fei-Fei. Deep visual-semantic alignments for generating image descriptions. arxiv e-prints. *CVPR*, 2015.

[KHS+18] Taehyeong Kim, Min-Oh Heo, Seonil Son, Kyoung-Wha Park, and Byoung-Tak Zhang. Glac net: Glocal attention cascading networks for multi-image cued story generation. In *CoRR*, 2018.

[KJ19] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age, 2019.

[KJZ18] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *NeurIPS*, 2018.

[KKK+18] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *CVPR*, 2018.

[KLZ19] Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. Dual attention networks for visual reference resolution in visual dialog. *ACL*, 2019.

[KMP+18] S. Kottur, J. Moura, D. Parikh, D. Batra, and M. Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In *ECCV*, 2018.

[KOL+17] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *ICLR*, 2017.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[KZG+17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[LB04] Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *ACL*, 2004.

[LBBH98] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *IEEE*, 1998.

[LBPL19] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

[LDRC20] Zhibin Liao, Tom Drummond, Ian Reid, and Gustavo Carneiro. Approximate fisher information matrix to characterize the training of deep neural networks. *TPAMI*, 2020.

[LDZ+17] Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, and Xiaogang Wang. Visual Question Generation as Dual Task of Visual Question Answering. In *CVPR*, 2017.

[Led01] Michel Ledoux. *The Concentration of Measure Phenomenon*. AMS, 2001.

[LGG+18] Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees GM Snoek. Videolstm convolves, attends and flows for action recognition. *Computer Vision and Image Understanding*, 2018.

[Lin04] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL*, 2004.

[LKY+17] J. Lu, A. Kannan, J. Yang, D. Parikh, and D. Batra. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *NIPS*, 2017.

[LL17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017.

[LLS+16] C.W. Liu, R. Lowe, I.V. Serban, M. Noseworthy, L. Charlin, and J. Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*, 2016.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[LST+19] Jiacheng Li, Haizhou Shi, Siliang Tang, Fei Wu, and Yueting Zhuang. Informative visual storytelling with cross-modal rules. In *MM*, 2019.

[LV19] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *CVPR*, 2019.

[LYBP16] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *NIPS*, 2016.

[LZ14] A. Liu and Brian Ziebart. Robust classification under sample selection bias. In *NIPS*, 2014.

[MBPD20] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. *ECCV*, 2020.

[MCH+16] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *NAACL*, 2016.

[MF14] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, 2014.

[Mil95] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995.

[MLL16] Lin Ma, Zhengdong Lu, and Hang Li. Learning to answer questions from image using convolutional neural network. *AAAI*, 2016.

[MMD+16] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. *ACL*, 2016.

[MP19] Yatri Modi and Natalie Parde. The steep road to happily ever after: an analysis of current visual storytelling models. In *Workshop on Shortcomings in Vision and Language, NAACL*, 2019.

[MRF15] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015.

[MXY+15] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep Captioning with Multimodal Recurrent Neural Networks (m-rnn). In *ICLR*, 2015.

[NH10] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

[NH16] Hyeonwoo Noh and Bohyung Han. Training recurrent answering units with joint loss minimization for vqa. *arXiv:1606.03647*, 2016.

[NHK17] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. *CVPR*, 2017.

[NLS18] Medhini Narasimhan, Svetlana Lazebnik, and Alexander G Schwing. Out of the box: Reasoning with graph convolution nets for factual visual question answering. *NeurIPS*, 2018.

[NNM96] Sameer A. Nene, Shree K. Nayar, and Hiroshi Murase. Columbia object image library (coil-20). Technical report, Columbia, 1996.

[NZZ+19] Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. Recursive visual attention in visual dialog. In *CVPR*, 2019.

[OE18] A. Owens and A. A. Efros. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features. In *Proc. ECCV*, 2018.

[OWM+18] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Learning Sight from Sound: Ambient Sound Provides Supervision for Visual Learning. *IJCV*, 2018.

[PK15] Cesc C. Park and Gunhee Kim. Expressing an image stream with a sequence of natural sentences. In *NeurIPS*, 2015.

[PP13] Ninh Pham and Rasmus Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *SIGKDD*, 2013.

[PRWZ01] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2001.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[PWGSH15] Federico Perazzi, Oliver Wang, Markus Gross, and Alexander Sorkine-Hornung. Fully connected object proposals for video segmentation. In *ICCV*, 2015.

[QNHZ20] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. *CVPR*, 2020.

[RAL18] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *NeurIPS*, 2018.

[RGH+16] Tim Rocktäschel, Edward Grefenstette, Moritz Hermann, Karl, Tomáš Kočiskỳ, and Phil Blunsom. Reasoning about entailment with neural attention. In *ICLR*, 2016.

[RHGS15] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *IEEE PAMI*, 2015.

[RKZ15] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *NIPS*, 2015.

[Rot85] OS Rothaus. Analytic inequalities, isoperimetric inequalities and logarithmic sobolev inequalities. *JFA*, 1985.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *SIGKDD*, 2016.

[RSK20] Kushal Kafle Robik Shrestha and Christopher Kanan. A negative case analysis of visual grounding methods for vqa. *https://arxiv.org/abs/2004.05704*, 2020.

[SC18] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *ECCV*, 2018.

[SDGS18] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

[SDSKS18] E. Shlizerman, L. Dery, H. Schoen, and I. Kemelmacher-Shlizerman. Audio to Body Dynamics. In *CVPR*, 2018.

[SGK17] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *ICML*, 2017.

[SHK+14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 2014.

[SHM+16] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[SLHS17] P. H. Seo, A. Lehrmann, B. Han, and L. Sigal. Visual reference resolution using attention memory for visual dialog. In *NIPS*, 2017.

[SLS+19] Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *ICCV*, 2019.

[SPM19] Ramon Sanabria, Shruti Palaskar, and Florian Metze. Cmu sinbad's submission for the dstc7 avsd challenge. In *DSTC7 at AAAI2019 workshop*, 2019.

[SRB+17] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lillicrap. A simple neural network module for relational reasoning. In *NIPS*, 2017.

[SSH16] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, 2016.

[SSH17] Idan Schwartz, Alexander G. Schwing, and Tamir Hazan. High-order attention models for visual question answering. In *NIPS*, 2017.

[SSH19] Idan Schwartz, Alexander G Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. In *CVPR*, 2019.

[SSK+17] Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *CoNLL*, 2017.

[SSL+17] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, 2017.

[SVI+15] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2015.

[SVL14] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[SVW+16] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016.

[SWB+07] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. *PAMI*, 2007.

[SWY+15] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[SYHS19] Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. Factor graph attention. In *CVPR*, 2019.

[SZ15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[SZT17] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *https://arxiv.org/abs/1703.00810*, 2017.

[TB19] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *EMNLP*, 2019.

[TBF+15] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

[TKSDM03] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *NAACL*, 2003.

[TZ15] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *ITW*, 2015.

[VCC+18] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *ICLR*, 2018.

[VCS+18] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra. Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. In *AAAI*, 2018.

[VDL+19] Ramakrishna Vedantam, Karan Desai, Stefan Lee, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Probabilistic neural symbolic models for interpretable visual question answering. In *International Conference on Machine Learning*. PMLR, 2019.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[VTBE14] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2014.

[VZP14] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2014.

[WCB14] J. Weston, S. Chopra, and A. Bordes. Memory networks. In *NIPS*, 2014.

[WCfWW18] Xin Wang, Wenhu Chen, Yuan fang Wang, and William Yang Wang. No metrics are perfect: Adversarial reward learning for visual storytelling. In *ACL*, 2018.

[WJL+20] Yue Wang, Shafiq Joty, Michael R Lyu, Irwin King, Caiming Xiong, and Steven CH Hoi. Vd-bert: A unified vision and dialog transformer with bert. *EMNLP*, 2020.

[WM19] Jialin Wu and Raymond Mooney. Self-critical reasoning for robust visual question answering. In *NeurIPS*, 2019.

[WMZ+19] Bairui Wang, Lin Ma, Wei Zhang, Wenhao Jiang, and Feng Zhang. Hierarchical photo-scene encoder for album storytelling. In *AAAI*, 2019.

[WSL17] Liwei Wang, Alexander G Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. *NIPS*, 2017.

[WTF20] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multimodal classification networks hard? In *CVPR*, 2020.

[WWL+19] Ruize Wang, Zhongyu Wei, Piji Li, Qi Zhang, and Xuanjing Huang. Storytelling from an image stream using scene graphs. In *AAAI*, 2019.

[WWS+17] Q. Wu, P. Wang, C. Shen, I. Reid, and A. van den Hengel. Are you talking to me? reasoned visual dialog generation through adversarial learning. *arXiv preprint arXiv:1711.07613*, 2017.

[WXW+16] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.

[WZY+19] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *ICCV*, 2019.

[XBK+15] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[XMS16] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. *ICML*, 2016.

[XS16] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, 2016.

[YBB17] Licheng Yu, Mohit Bansal, and Tamara L. Berg. Hierarchically-attentive rnn for album summarization and storytelling. In *EMNLP*, 2017.

[YGL+19] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

[YHG+16] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.

[YLC+19] Pengcheng Yang, Fuli Luo, Peng Chen, Lei Li, Zhiyi Yin, Xiaodong He, and Xu Sun. Knowledgeable storyteller: A commonsense-driven generative model for visual storytelling. In *IJCAI*, 2019.

[YSXZ16] Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *TACL*, 2016.

[YWG+18] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Neurips*, 2018.

[YWH+16] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *CVPR*, 2016.

[YYX+18] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. In *NeurIPS*, 2018.

[YZZ19] Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. Making history matter: History-advantage sequence training for visual dialog. In *ICCV*, 2019.

[ZAT17] Bolei Zhou, Alex Andonian, and Antonio Torralba. Temporal relational reasoning in videos. *arXiv preprint arXiv:1711.08496*, 2017.

[ZBFC19] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019.

[ZCL+19] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *CVPR*, 2019.

[ZGBFF16] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, 2016.

[ZHPB18] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. *ICLR*, 2018.

[ZHS20] Bowen Zhang, Hexiang Hu, and Fei Sha. Visual storytelling via predicting anchor word embeddings in the stories. In *ICCV*. Workshop on Closing the Loop Between Vision and Language, 2020.

[ZMWC19] Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *ACL*, 2019.

[ZWQZ19] Zilong Zheng, Wenguan Wang, Siyuan Qi, and Song-Chun Zhu. Reasoning visual dialogs with structural and partial observations. In *CVPR*, 2019.

[ZWY+17] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017.

[ZWY+18] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *NAACL*, 2018.

[ZZH+17] C. Zhu, Y. Zhao, S. Huang, K. Tu, and Y. Ma. Structured attentions for visual question answering. In *ICCV*, 2017.

באמצעות למידה עמוקה. פרקים 3-6 עוסקים בבעיות שפתרנו באמצעות מודל הקשב החדשני שלנו: פרק 3 עוסק בפתרון מענה על שאלות חזותיות; פרק 4 עוסק ביצירת מודל בינה אינטראקטיבי היכול לקיים דיאלוג על תמונה; פרק 5 עוסק ביצירת סיפורים על תמונה. משימה זו מאתגרת במיוחד שכן היא דורשת קוהרנטיות וניתוח של רצף תמונות במקביל; בפרק 6 אנו מציגים מודל בינה שמתאר כדיאלוג ווידיאו. בתהליך זה עליו גם להבין קלט חזותי (ווידיאו), גם לנתח קלט קולי (אודיו), וגם להבין את השאלות הנשאלות בדיאלוג. החלק השני של התזה עוסק בתפיסה של המודל. בפרק 7, אנו מציגים את הציון החדשני שלנו לכימות יכולת התפיסה של מודל עבור חלקים שונים של הקלט. בפרק 8 אנחנו מציגים שיטה להגברת יכולת התפיסה של מודל באמצעות רגולריזציה בזמן האימון.

תקציר

המסע לעבר אלגוריתם בעל יכולות קוגניטיביות הוא חלק בתחום מערכות לומדות ומופיע במגוון היבטים, כגון: מענה על שאלות חזותיות, יצירה אוטומטית של תיאור תמונה, מענה על שאלות חזותיות, וקריאת מכונה. מכנה משותף לכל הבעיות הללו הוא השימוש בריבוי מודאלים. למשל, לצורך מענה על שאלות חזותיות, על האלגוריתם להבין את השאלה, ולהסיק מהתמונה את התשובה. דוגמא נוספת היא יצירת כיתוב המתאר תמונה, בה האלגוריתם צריך להבין את הסצנה בתמונה, ובנוסף לבנות מודל שפה היוצר שפה תקנית. בשני המקרים המודאלים הם תמונה וטקסט.

מנגנון עיקרי שמאפשר לחקות מערכת קוגניציה הוא "מנגנוני קשב" אשר צמחו לאחרונה ומהווים גורם נפוץ בפתרון בעיות אלו. למשל במענה על שאלות חזותיות מנגנוני קשב מאפשרים לרשתות עמוקות להסיק היכן להסתכל בתמונה כדי לענות על השאלה. במקרה של יצירת כיתוב המתאר תמונה, מערכת הקשב מאפשרת למודל להסתכל על אזור מסוים בכל שלב יצירת הכיתוב לתמונה. אחד היתרונות של מנגנוני קשב הוא האפשרות לניתוח האזורים בהם מנגנון הקשב למד להתמקד, דבר המספק פרשנות על אופן פעולת ההסקה של הרשת. אך החשוב מכל הוא שלעיתים קרובות מנגנונים אלה משפרים את הביצועים. הסיבה לתופעה זו מיוחסת ליכולת של הרשת להפיק מהמודלים יותר מידע ובאופן יותר תמציתי. כיום למנגנוני קשב חסרים שני אספקטים: הראשון, מנגנוני קשב מותאמים לפתרון של קלט ספציפי, ולכן נבנים לפתרון משימה ספציפית; השני, מנגנוני קשב לרוב מתבססים על פתרון אד־הוק מסובך, ולא מנומק. כדיי להתמודד עם בעיות אלו, אנחנו מציעים מנגנון קשב חדשני וגנרי שלומד קורלציות מסדר גבוהה של מגוון רחב של מודאליים. לדוגמא, קורלציות מסדר שני ממדלות קשרים בין שני מודאליים, טקסט ותמונה, ובהכללה לסדר גבוהה, קורלציות מסדר k מייצגות קשרים בין k מודאליים שונים. למידה של הקורלציות הללו מאפשרות לכוון כמות קשב הולמת לכל מודאל לצורך פתרון המשימה המשותפת.

למרות ההתקדמות המרשימה שמנגנון הקשב מאפשר, מאגרי ידע גדולים קשים לתיוג ומכילים הטיות שלעתים קרובות איננו מודעים להן. מסווגים מבוססי קשב, בתורם, נוטים לנצל את ההטיות הללו ולמצוא קיצורי דרך לקבלת ההחלטה. כתוצאה מכך, השיטות הנוכחיות עשויות לא לפתור את המשימה, אלא להצליח להתאים את ההחלטה להטיות שקיימות במאגרי הידע. כדיי לטפל בדאגה זו, אנו מגדירים ציון חדש: ציון תפיסה. ציון זה מעריך את מידת הסתמכותו של המודל על כל חלקי הקלט. למשל, ציון תפיסה גבוהה עבור התמונה יעיד שהמודל מסתמך על התמונה בהחלטתו. בנוסף, אנחנו לומדים שיטת רגולריזציה להגברת התפיסה, על ידי מקסום האנטרופיה הפונקציונאלית של חלקי הקלט השונים במהלך האימון. אנו מאמתים את היעילות של שיטה זו על מגוון רחב של בעיות, הכוללות ניתוח ווידיאו, תמונה וטקסט.

תזה זו מחולקת לשמונה פרקים. בפרק 2, אנחנו מספקים את הרקע הדרוש לקידוד טקסט ותמונה

והאכפתיות המתמידים. תודה מקרב לב שעשיתם הכל כדי להבטיח שאוכל להגשים את חלומותיי ללא דאגה.

אני מודה לטכניון על התמיכה הכספית הנדיבה בהשתלמותי.

המחקר בוצע בהנחייתו של פרופסור תמיר חזן ופרופסור אלכסנדר שווינג, בפקולטה למדעי המחשב.

חלק מן התוצאות בחיבור זה פורסמו כמאמרים מאת המחבר ושותפיו למחקר בכנסים ובכתבי-עת במהלך תקופת מחקר הדוקטורט של המחבר, אשר גרסאותיהם העדכניות ביותר הינן:

Itai Gat, Idan Schwartz, and Alexander Schwing. Perceptual score: Measuring perceptiveness of multi-modal classifiers. *Advances in Neural Information Processing Systems*, 34, 2021.

Itai Gat, Idan Schwartz, Alexander Schwing, and Tamir Hazan. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *Advances in Neural Information Processing Systems*, 33, 2020.

Idan Schwartz. Ensemble of mrr and ndcg models for visual dialog. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

Idan Schwartz, Alexander G Schwing, and Tamir Hazan. High-order attention models for visual question answering. *Advances in Neural Information Processing Systems*, 30, 2017.

Idan Schwartz, Alexander G Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. Factor graph attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

תודות

אני רוצה להודות למנחים שלי, אלכסנדר שווינג ותמיר חזן. היה תענוג לעבוד עם שני מנחים: התובנות של תמיר תמיד הוסיפו פלפל לעבודה שלנו, ואלכס תמיד היה מוכן לעבוד על פרויקטים מאתגרים. פגישות המחקר שלנו התפתחו במהרה לדיונים פילוסופיים, ואפשרו לנו להטיל ספק בהיבטים בסיסיים של למידת מכונה. תמיר ואלכס לימדו אותי כל מה שאני צריך לדעת כדי להיות חוקר טוב.

בנוסף, ברצוני להודות לבני קימלפלד על ההנחיה שלו במהלך התואר השני שלי.

הלימודים שלי בטכניון אפשרו לי להכיר חברים רבים ולקיים דיונים פוריים שהיוו השראה למחקר שלי ועוררו הרבה רעיונות מרגשים. במיוחד ברצוני להודות ליפתח זיסר ולג'ליל מורוני, חבריי הטובים ממשרד 216 האגדי. תודה על הצחוקים והזיכרונות הבלתי נשכחים. איתי גת, אחד ממשתפי הפעולה הקרובים שלי, עבד איתי רבות בשעות לילה מאוחרות. תודה גם לנמרוד רייפר שתמיד יש לו עצות מצוינות. אחרונה חביבה, מעין כסלו על תמיכה ועידוד הפסקות תכופות מהשגרה לטיולים ורגיעה. יש עוד רבים שזכיתי להכיר, לשתף איתם פעולה ולהתיידד איתם: אדוארד ויטקינג, גיא עוזיאל, מיכל בדיאן, גלי שפי, יונתן גייפמן, מיכל פרידמן, איתן זינגר, שובל לגזיאל, עידן חסון, מתן פלד.

הייתה לי הזדמנות לעבוד בצוות המחקר של איביי ולשתף פעולה עם קירה רדינסקי ועידו גיא. תודה על ההדרכה. באיביי היה תענוג לעבוד עם יותם אשל, ניר לוין, שי חיים וניר אופק.

נהניתי גם לעבוד במיקרוסופט ולשתף פעולה עם תום בראודה, שלומי מליח, לרה שטוטלנד, צוף אבני ברוש, רן ברנשטיין, שגיא הלי ועידו פרינס.

לסיום, ברצוני להודות להורי, נלי וזאב, על תמיכה באהבה שלי למחשבים מגיל צעיר. כמו כן, ברצוני להביע תודה לאחי דור, שתמיד היה החבר הכי טוב שלי. סבי וסבתי, גילה ומוריס, על האהבה

מודלים קוגנטיבים בלמידה עמוקה

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר
דוקטור לפילוסופיה

עידן שורץ

הוגש לסנט הטכניון — מכון טכנולוגי לישראל
תמוז התשפ"א חיפה יולי 2021

מודלים קוגנטיבים בלמידה עמוקה

עידן שורץ