Discovering Transformations for Generalization in Semantic Vision

Deep neural networks have emerged as a widely popular architectural choice for modeling tasks in multiple domains such as computer vision [22] and natural language processing [18]. Although highly capable of learning from training data, recent studies show that neural networks underperform on new test sets or under distribution shift [17], natural corruptions [9], adversarial attacks [8], spurious correlations [2], and many other types of "unseen" changes in test samples. For instance, digit recognition models trained on the black-and-white MNIST training images are almost perfect ($> 99.5\%$ accuracy) on the corresponding *i.i.d.* test set, yet their performance on colored digits and real-world digits from street number plates is only around 70%. Similarly, state-of-the-art NLP models have been shown to fail when negation is introduced in the input [11]. These findings pose a significant challenge to the practical adoption and reliability of computer vision models in the real-world, especially with sensitive data such as biomedical and satellite imagery.

My work addresses these shortcomings in the domain of semantic vision (i.e. computer vision tasks designed for assigning meaning to the image); this includes tasks such as image classification, visual question answering, vision-language inference, image captioning, etc. The focus of my thesis is to identify the various situations under which machine learning models may fail while making predictions for semantic vision tasks, and to develop machine learning techniques to mitigate risks posed therein. The approach taken by my thesis can be summarized as:

1. Identifying failure modes – situations under which computer vision systems may fail for semantic vision tasks,
2. Developing machine learning algorithms and data transformations to mitigate the risks posed by such situations,
3. Creating evaluation and analysis tools such as creation of evaluation datasets, protocols, and metrics.

Robustness under Attribute Shift (Gokhale et al. [5]): Previous work on robust image classification has considered pixel-level perturbations – Adversarial Training [14, 20] is one class of methods effective on pixel-level adversarial attacks. However, in real world scenarios, complex and larger domain shifts can be encountered – for instance, lighting, camera angle, background, textures, weather, geometric transformations and other natural perturbations. Unfortunately, these cannot be covered by methods that utilize norm-bounded and additive pixel-level perturbations.

We consider a relaxed setting of the generalization problem – in this setting, information about the target domain is available only in terms of a set of attributes that are known to differ at test time (target samples are not available). Figure 1 shows an illustrative example from CLEVR-Singles [5] where the task is color classification, but attributes that are known to change are: *size*, *shape*, *position*, and *material* of the object. Note that the classifier is expected to be invariant to such attribute shifts, without observing the target images, and without knowing the magnitudes or combinations of attribute shifts that may exist. Only the set of attributes are given – we do not know which attributes will change at test time, by what magnitude, or in what combinations.

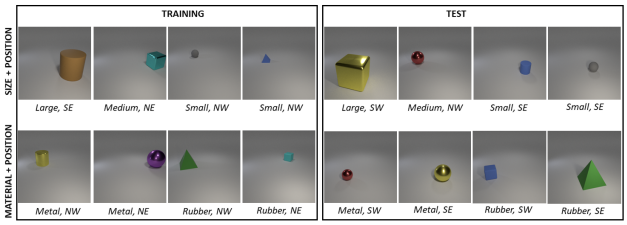

Figure 1. CLEVR-Singles evaluates robustness to attribute shift.

We introduced a solution : *Attribute-Guided Adversarial Training (AGAT)* that leverages adversarial training to generate new samples so as to maximize exposure of the classifier to variations in the attribute space. Furthermore, our proposed approach is flexible to support a wide-range of attribute specifications, which we demonstrate with three different use-cases:

1. Object-level shifts from a conditional GAN for adversarial training on a new variant of the CLEVR dataset;
2. Geometric transformations implemented using a spatial transformer for MNIST data; and
3. Synthetic image corruptions [9] on CIFAR-10 data.

Adversarially Learned Transformations for Domain Generalization (In Review): Next, we focus on the problem of *single source domain generalization (SSDG)* – where the model has access only to a single training domain, and is expected to generalize to multiple different testing domains. This is especially hard because of the limited information available to train the model with just a single source. To be successful in SSDG, maximizing diversity of synthesized domains has emerged as one of the most effective strategies. However, naïve pre-specified augmentations such as AugMix [10] or RandConv [21] do not work effectively for domain generalization either because they cannot model large semantic shifts, or the span of transforms that are pre-specified, do not cover shifts commonly occurring in domain generalization. We also observe that the effectiveness of augmentation techniques is often dataset dependent – for instance, common data augmentation such as rotation-translation-scaling-cropping might improve in-domain accuracy and robustness to adversarial attacks, but does not help when domain discrepancies are large, as is

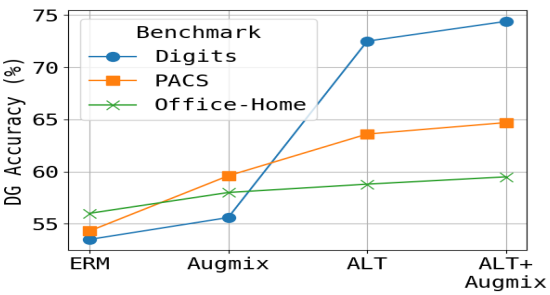

Figure 2. The plot summarizes our results – while diversity alone improves performance over the naive ERM baseline, adapting this diversity using adversarially learned transformations (ALT) provides a significant boost for single source domain generalization.

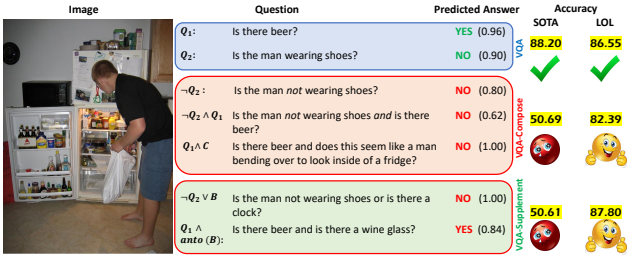

Figure 3. Illustration of logical composition of questions. State-of-the-art models answer questions from the VQA dataset (Q_1, Q_2) correctly, but fail when asked a logical composition including negation, conjunction, disjunction. Our models improve on this metric substantially and retain performance on VQA data.

the case for SSDG benchmarks such as PACS [12] and OfficeHome [19].

To address this issue, we propose the design of a novel framework that uses adversarially learned transformations (ALT) using a neural network to model plausible, yet hard image transformations that maximize classification error. ALT potentially offers an interplay between diversity and adversity and over time, a synergistic partnership is expected to emerge, exposing the model to increasingly unique, challenging, and semantically diverse examples – ideally suited for single source domain generalization. Results on three benchmarks in SSDG are summarized in Figure 2.

VQA Robustness to Logical Transformations (Gokhale et al. [4]): Multi-modal tasks involving both vision and language (V&L) inputs, such as visual question answering (VQA), open up many more types of domain discrepancies that can affect model performance of test time. For the VQA task, given an image and a question about it, models are trained to predict the answers to those questions. In VQA-LOL [4], we discovered that existing VQA models fail when logical transformations such as negation, conjunction, and disjunction are introduced in the questions as shown in Figure 3. This surprising finding led us to develop a data augmentation tool that allows us to produce logical combinations of multiple questions in the source dataset, and a training objective that is based on Frechet inequalities to guide the predicted probabilities of answers to questions with negation, conjunction, and disjunction. Thus, given a known transformation between source and target domains, we developed a method that can leverage data augmentation for improving robustness of VQA models.

In Mutant (Gokhale et al. [3]), we make use of simple image transformations which remove objects or change their colors, in addition to the logical transformations developed above. Empirical results on the VQA-CP challenge [1] show that this method achieves robustness under changing question-answer priors (i.e. when the probability of answers given a question type varies between train and test domains).

V&L Robustness to Linguistic and Semantic Perturbations (Gokhale et al. [6]): We follow-up on VQA-LOL and investigate robustness to a broader set of linguistic transformations for the task of vision-and-language inference (VLI), i.e., predicting whether a sentence is true of false about a given image or video. We define a set of text transformations called "SISP transforms" which allow us a controlled method to semantically manipulate text to generate augmented data that is semantics-inverting (SI) or semantics-preserving (SP). Our key idea is a model-agnostic min-max adversarial training optimization called *Semantically Distributed Robust Optimization (SDRO)*, which utilizes SISP transforms in a distributed robust optimization setting. SDRO is motivated by group-DRO style approaches [15], but in our case we can leverage linguistics-informed SISP transformations as groups for worst-group error minimization. Experimental results show that SDRO benefits both in-domain accuracy, robustness to linguistic transformations and text attacks, as well as improved calibration, on three datasets: NLVR^2 [16] (image-based reasoning), VIOLIN [13] (video-based reasoning), and binary VQA.

To summarize, my work addresses tasks in semantic vision and seeks to design and discover data transformations that can improve the robustness and generalization of models in semantic vision. Ongoing work investigates the connections between OOD generalization and adversarial robustness (starting with (Gokhale et al. [7])). Planned work for the remainder of my PhD will include an effort towards analyzing and improving robustness of pre-trained V&L models and designing new measures for detecting and quantifying OOD shift.

Website: https://tejas-gokhale.github.io/
Google Scholar: https://scholar.google.com/citations?user=_ILTlEwAAAAJ

References

[1] Agrawal, A., Batra, D., Parikh, D., and Kembhavi, A. Don't just assume; look and answer: Overcoming priors for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 4971–4980, 2018.

[2] Beery, S., Van Horn, G., and Perona, P. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 456–473, 2018.

[3] Gokhale, T., Banerjee, P., Baral, C., and Yang, Y. MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 878–892, 2020. URL https://www.aclweb.org/anthology/2020.emnlp-main.63.

[4] Gokhale, T., Banerjee, P., Baral, C., and Yang, Y. Vqa-lol: Visual question answering under the lens of logic. In *European conference on computer vision*. Springer, 2020.

[5] Gokhale, T., Anirudh, R., Kailkhura, B., Thiagarajan, J. J., Baral, C., and Yang, Y. Attribute-guided adversarial training for robustness to natural perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7574–7582, 2021.

[6] Gokhale, T., Chaudhary, A., Banerjee, P., Baral, C., and Yang, Y. Semantically distributed robust optimization for vision-and-language inference. *Findings of the ACL*, 2022.

[7] Gokhale, T., Mishra, S., Luo, M., Sachdeva, B. S., and Baral, C. Generalized but not robust? comparing the effects of data modification methods on out-of-domain generalization and adversarial robustness. *Findings of the ACL*, 2022.

[8] Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1412.6572.

[9] Hendrycks, D. and Dietterich, T. G. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

[10] Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1gmrxHFvB.

[11] Kassner, N. and Schütze, H. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7811–7818, 2020. URL https://www.aclweb.org/anthology/2020.acl-main.698.

[12] Li, D., Yang, Y., Song, Y., and Hospedales, T. M. Deeper, broader and artier domain generalization. In *IEEE International Conference on Computer Vision, ICCV*, 2017. URL https://doi.org/10.1109/ICCV.2017.591.

[13] Liu, J., Chen, W., Cheng, Y., Gan, Z., Yu, L., Yang, Y., and Liu, J. Violin: A large-scale dataset for video-and-language inference. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 10897–10907, 2020. URL https://doi.org/10.1109/CVPR42600.2020.01091.

[14] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

[15] Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

[16] Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., and Artzi, Y. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6418–6428, 2019. URL https://www.aclweb.org/anthology/P19-1644.

[17] Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18583–18599, 2020.

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

[19] Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5385–5394, 2017. URL https://doi.org/10.1109/CVPR.2017.572.

[20] Volpi, R., Namkoong, H., Sener, O., Duchi, J. C., Murino, V., and Savarese, S. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, pp. 5339–5349, 2018.

[21] Xu, Z., Liu, D., Yang, J., Raffel, C., and Niethammer, M. Robust and generalizable visual representation learning via random convolutions. In *International Conference on Learning Representations*, 2020.

[22] Yuille, A. L. and Liu, C. Deep nets: What have they ever done for vision? *International Journal of Computer Vision*, 129(3):781–802, 2021.