COS 454: Project Proposal

Inductive Biases in Computer Vision

Shikhar Tuli
stuli@princeton.edu

November 7, 2020

1 Introduction

Convolutional Neural Networks (CNNs) are currently the de-facto standard for many computer vision tasks including object detection [1], image classification [2], segmentation [3], facial recognition [4], and captioning [5]. Recently, work has also been done in Transformers for vision-based tasks, and some have even outperformed state-of-the-art CNNs [6, 7]. However, studies have shown that CNNs trained on popular datasets like ImageNet, inherently gain texture-bias, i.e., such networks tend to classify images by texture rather than by shape [8]. On the other hand, humans preferentially use shape information for classification [9, 10].

Further, there has been an increasing trend of model accuracy in computer vision tasks with a proportionate increase in the training dataset size. However, as dataset sizes increase, the computation required for training escalates [11]. For context, state-of-the-art CNNs need thousands of days of GPU/TPU training [7]. Forcing such vision systems to learn based on cognitively-inspired inductive biases should substantially reduce training time and thus, computational cost.

2 Related Work

Many popular datasets have been used to train CNNs including CIFAR-100 [12], VTAB [13], and ImageNet [14]. However, none of these datasets probe into the inductive biases of such networks. To improve the robustness of trained networks, adversarial perturbations have been used [15]. Even though adversarial training has proved to improve accuracy on out-of-distribution data, such models are still not necessarily robust enough to other forms of imperceptible perturbations [16, 17]. Nevertheless, recently proposed natural adversarial examples could be used to further test this [18]. Data augmentation has also been used to improve both generalization and out-of-distribution robustness of CNNs [19, 20]. However, such methods often tend to decrease model accuracy [21].

A recent work by Hermann et al. shows that naturalistic data augmentation involving color distortion, noise, and blur substantially decreases texture bias, whereas random-crop augmentation increases texture bias in ImageNet-trained CNNs [22]. With the use of specially designed datasets (namely Geirhos Style-Transfer dataset [23], Navon dataset [24] and ImageNet-C dataset [25]), Hermann et al. also show that CNNs can learn shape bias as easily as texture bias. However, these datasets have limitations that could be countered only with new datasets [22]. Moreover, this work does not consider other datasets for pre-training and their effect on bias. Drawing out a trade-off between the amount of data needed to 'blindly' train a network and the use of appropriate inductive biases in the pre-train process is also worth exploring. It has also been shown that the use of attention in place of convolution

appears to have little effect upon texture bias [22]. However, this needs to be tested on state-of-the-art Transformer models as well.

3 Goal of this Project

In this project, we aim to develop a novel dataset (or a combination of existing datasets) that could probe inductive biases in vision-based systems. These cognitively-inspired inductive biases could include - shape, texture, color, rotation-invariance, and shift-invariance.

Further, by training on multiple-sized samples of the same dataset and using these inductive biases for pre-training, we hope to find trade-offs between the number of training samples and the use of such biases. These tests would not only be done on CNNs but also on emerging Transformer networks for vision-based applications. Since, doing these tests on the state-of-the-art models like BiT [26], iGPT [6], ViT [7], etc. might be computationally expensive, we could use smaller models. This could make comparisons much easier with less computational requirements and could help us get a perspective if at all such a trade-off exists. But then again, literature shows that these biases are highly dependent on the data (e.g. scale of the ImageNet dataset) and training methodology [22]. So it might be hard to show if these tests on such toy models represent those on the state-of-the-art.

Recently proposed error-consistency results between humans and CNNs could also be exploited for better evaluation [27]. Consequently, guidelines could be proposed for the use of specific biases for training in certain tasks (or otherwise for generalization) to reduce the dataset and model size, which would in turn, decrease the training time and computational cost.

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[5] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5659–5667.

[6] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, P. Dhariwal, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[7] Anonymous, "An image is worth 16x16 words: Transformers for image recognition at scale," https://openreview.net/pdf?id=YicbFdNTTy, (Accessed on 10/11/2020).

[8] N. Baker, H. Lu, G. Erlikhman, and P. J. Kellman, "Deep convolutional networks do not classify based on global object shape," *PLoS computational biology*, vol. 14, no. 12, p. e1006613, 2018.

[9] S. C. Kucker, L. K. Samuelson, L. K. Perry, H. Yoshida, E. Colunga, M. G. Lorenz, and L. B. Smith, "Reproducibility and a unifying explanation: Lessons from the shape bias," *Infant Behavior and Development*, vol. 54, pp. 156–165, 2019.

[10] B. Landau, L. B. Smith, and S. S. Jones, "The importance of shape in early lexical learning," *Cognitive Development*, vol. 3, no. 3, pp. 299 – 321, 1988. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0885201488900147

[11] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 843–852.

[12] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.

[13] X. Zhai, J. Puigcerver, A. Kolesnikov, P. Ruyssen, C. Riquelme, M. Lucic, J. Djolonga, A. S. Pinto, M. Neumann, A. Dosovitskiy *et al.*, "A large-scale study of representation learning with the visual task adaptation benchmark," *arXiv preprint arXiv:1910.04867*, 2019.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[15] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 427–436.

[16] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song, "Spatially transformed adversarial examples," *arXiv preprint arXiv:1801.02612*, 2018.

[17] F. Tramèr and D. Boneh, "Adversarial training and robustness for multiple perturbations," in *Advances in Neural Information Processing Systems*, 2019, pp. 5866–5876.

[18] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," *arXiv preprint arXiv:1907.07174*, 2019.

[19] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6023–6032.

[20] E. Rusak, L. Schott, R. Zimmermann, J. Bitterwolf, O. Bringmann, M. Bethge, and W. Brendel, "Increasing the robustness of dnns against image corruptions by playing the game of noise," *arXiv preprint arXiv:2001.06057*, 2020.

[21] R. G. Lopes, D. Yin, B. Poole, J. Gilmer, and E. D. Cubuk, "Improving robustness without sacrificing accuracy with patch gaussian augmentation," *arXiv preprint arXiv:1906.02611*, 2019.

[22] K. L. Hermann and S. Kornblith, "Exploring the origins and prevalence of texture bias in convolutional neural networks," *arXiv preprint arXiv:1911.09071*, 2019.

[23] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," *arXiv preprint arXiv:1811.12231*, 2018.

[24] D. Navon, "Forest before trees: The precedence of global features in visual perception," *Cognitive psychology*, vol. 9, no. 3, pp. 353–383, 1977.

[25] D. Hendrycks and T. G. Dietterich, "Benchmarking neural network robustness to common corruptions and surface variations," *arXiv preprint arXiv:1807.01697*, 2018.

[26] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," *arXiv preprint arXiv:1912.11370*, 2019.

[27] R. Geirhos, K. Meding, and F. A. Wichmann, "Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency," *arXiv preprint arXiv:2006.16736*, 2020.