CONVOLUTION VS ATTENTION FOR IMAGE CLASSIFICATION

Abhay Puri
abhay.puri@umontreal.ca

Jizhou Wang
jizhou.wang@umontreal.ca

Axel Bogos
axel.bogos@umontreal.ca

Pulkit Madan
pulkit.madan@umontreal.ca

ABSTRACT

Vision transformers are biased towards shape and convolutions are biased towards texture. We want to do an exploration on such inductive biases for both the architectures and compare how the latest developments such as ViT (Dosovitskiy et al., 2021) and ConvneXt (Liu et al., 2022) methods compare on a task in a suitable domain and how they compare to human vision. If time permits, we would try to come up with an architecture that combines the best of both worlds where we can combine the right inductive biases of shape and texture for the target task. As an initial proxy measure of such biases, performance on a stylized dataset such as the stylized ImageNet dataset (Geirhos et al., 2019) will be used. However, further datasets where shape and texture present particular challenges may be used either as complimentary evaluation of existing biases in the models we are interested in or as benchmarks of a proposed architecture.

1 ROADMAP

1. Comparative study of biases of shape and texture with existing convolution-based and attention-based architectures.
2. Researching ways how the biases can be controlled in each type of architecture.
3. Researching an architecture that is able to combine the inductive biases of shape and bias, for a task where such biases are required.*
4. Testing if controlling the biases acts as a proxy task for out-of-distribution generalization by evaluating the performance on other datasets. *

2 DATASET

The stylized ImageNet dataset (Geirhos et al., 2019) aims to provide means to quantify and benchmark shape vs texture biases of various models. In short, it transfers textures of certain image classes to other ones, for example conserving the shape of a cat while applying the texture of another class, e.g. an Indian elephant as demonstrated in Fig. 1 of Tuli et al. (2021) below.

3 EVALUATION METRICS

While the initial aim is of stylizing the dataset is to nudge traditional CNNs such as the AlexNet introduced by Krizhevsky et al. (2012) and more recently ResNet by He et al. (2015) towards having a shape bias, we aim to use the performance on the Stylized ImageNet dataset as a proxy measure of the existing biases of different architectures such as CNNs, ViT and ConvneXt. A simple measure of these biases is introduced by Geirhos et al. where they consider a correct classification as either the original image label from ImageNet (henceforth the "shape label") and the applied texture transform (henceforth the "texture label"). The shape-vs-texture bias may then be simply expressed as a ratio

*Time permitting

between the correct-by-texture-label and correct-by-shape-label over the total correct classifications of a particular model.

A process for “stylizing” an arbitrary dataset has been made public here , hence our initial exploration would be conducted on a stylized version of Tiny-ImageNet (Le & Yang, 2015) obtained through stylizing the original Tiny-ImageNet dataset.

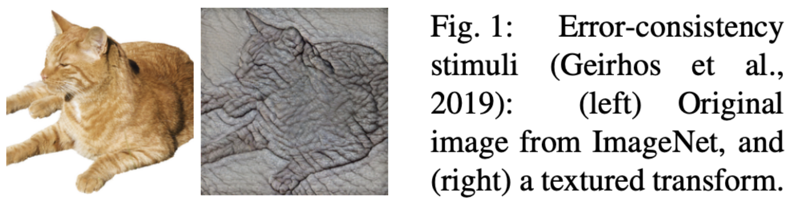

Figure 1: Figure 1 of Tuli et al. (2021)

4 Models

Below are some of the models and architectures we will consider to evaluate during this project:

- ResNet (He et al., 2015)
- ViT (Dosovitskiy et al., 2021)
- ConvNeXt (Liu et al., 2022)
- Swin Transformer * (Liu et al., 2021)
- Florence * (Yuan et al., 2021)
- CLIP * (Radford et al., 2021)
- CoAtNet (Dai et al., 2021)

5 Work Distribution

As of now, a rigid distribution of the tasks has not yet been done. We plan on collectively participating in the planning and literature review phases. A more precise distribution of the coding work will be established further into the project.

References

Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

*Time permitting

Md Amirul Islam, Matthew Kowal, Patrick Esser, Sen Jia, Björn Ommer, Konstantinos G. Derpanis, and Neil Bruce. Shape or texture: Understanding discriminative features in {cnn}s. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=NcFEZOi-rLa.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L. Griffiths. Are convolutional neural networks or transformers more like human vision?, 2021.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision, 2021.