A REPRODUCE REPORT OF PAPER: IMAGENET-TRAINED CNNS ARE BIASED TOWARDS TEXTURE; INCREASING SHAPE BIAS IMPROVES ACCURACY AND ROBUSTNESS

Shaobo Xu* & Peihao Ren*
School of Electronics and Computer Science
University of Southampton
Southampton, SO17 1BJ, United Kingdom
{sx2n18, pr1y18}@soton.ac.uk

ABSTRACT

We reproduced an oral ICLR paper, named "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness". We implemented the most important part in the original paper and briefly discussed experiments on a subset of the ImageNet dataset (only 16 classes). We found that ResNet-50 trained on Stylized ImageNet is more accurate and robust than the same network trained only on ImageNet. We also verified that shape-based representations are more robust than the texture representations. All codes except the style transform part are written by the two authors.

1 INTRODUCTION

Ever since AlexNet Krizhevsky et al. (2012) won the competition, deep neural networks (DNN) becomes a more and more hot topics in the image-relevant field. However, the author of the paper "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness" has found that the popular DNN structures are actually biased towards texture which stands opposite to human behaviour. The author of the original paper has proposed one way to eliminate or partially eliminate the bias, that is by training the NN model on the stylized dataset.

2 DATA

2.1 16-CLASSES IMAGENET DATASET

We downloaded the famous ImageNet dataset Deng et al. (2009). It contains 1000 classes of images and each of the classes has about 1000 pictures. This data set is too big for us, thus we choose 16 classes of them as listed in the paper. The 16 classes are knife, keyboard, elephant, bicycle, airplane, clock, oven, chair, bear, boat, cat, bottle, truck, car, bird, dog. Each of these classes has 1300 pictures which is still too big since we would introduce stylized images in the following section which multiply the number of images by 8.

Hence our final data structure contains a training set and a validation set. Each set has 16 classes of different types of images. The training set has 400 images in each class while the validation set has 100 images in each class.

2.2 STYLIZED IMAGE

As is done in the original paper, the input images are stylized. We use the algorithm from GitHub (https://github.com/rgeirhos/Stylized-ImageNet) to transfer our original image input to several different styles. This stylizing method is from Huang & Belongie (2017). It uses the pictures from Kaggle's painter-by-numbers dataset (https://www.kaggle.com/c/painter-by-numbers/data) as style

pictures, then use an encoder to get the representation of the style picture and the original picture. The data is then fed to an Adaptive Instance Normalization (AdaIN) layer and a decoder. The final stylized picture is given by the original picture and the data from the decoder.

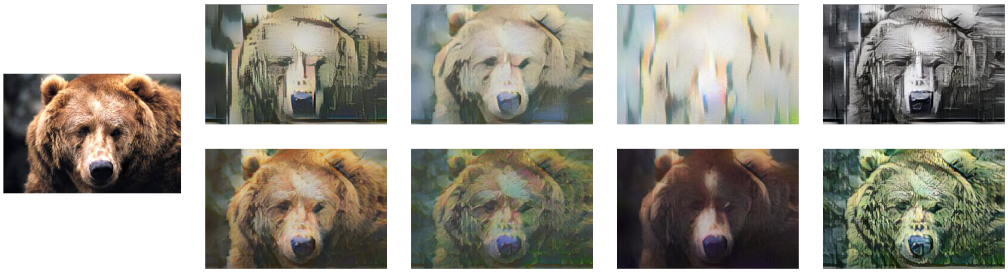

Figure 1: Example output of the stylized image. For each input image, we put it into the stylizing algorithm and then got 8 output images with random styles.

The stylized images are shown in figure 1. Human can easily find that these 8 stylized images are from a same original image. That is because we focus on the shape of the bear in stead of its texture.

In this report paper, we use IN to represent 16 classes ImageNet, and use SIN to represent Stylized 16 classes ImageNet.

3 EXPERIMENT

3.1 TRAINING CONVERGENCE

Since our dataset is different from the original paper, we plot the accuracy curve during each epoch in order to show the convergence performance on our dataset.

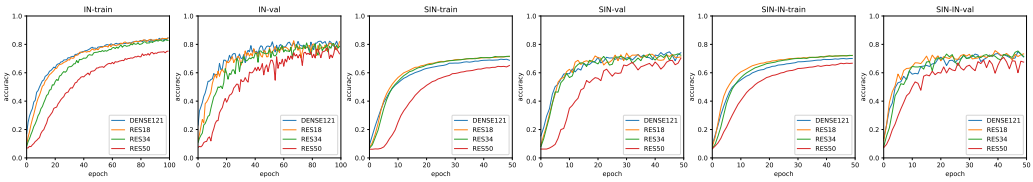

Figure 2: Top 1 accuracy on different datasets. In the figure above, IN train represents the top 1 training accuracy on IN dateset and SIN represents the top 1 accuracy on SIN dataset.

The results of our loss and accuracy curve on different datasets are shown in Figure 2. Our dataset only contains 8000 images (6400 for training and 1600 for validation). This means a complex DNN may cause overfitting thus leads to the situation that a simple DNN may get better performance than a complex DNN.

On the ImageNet dataset, we trained the model for 100 epochs while on the SIN dataset and the SIN+IN dataset, we trained the model for 50 epochs. This is because we need to get the converged accuracy of each model. IN is too small that needs more epochs to converge. SIN, on the other hand, is 8 times as large as IN, thus needs much less epochs to converge. Other parameters are the same as that in the original paper.

3.2 TEXTURE SHAPE BIAS

The key point of the reproduced paper is, the DNNs which were trained on ImageNet, are biased towards texture. Hence the author of the original paper did the an experiment to show this bias.

Table 1 shows our results on the experience. In this experiment, we trained 4 different DNNs on different datasets, and evaluated them on different datasets as the author of the original paper did. We

Table 1: Top 1 and top 5 accuracy of different neural networks training on different datasets. We trained different neural networks on different datasets and evaluated them on these datasets. The number of training epochs we choose are, IN:100, SIN:50.

Model	**IN → IN**	**IN → SIN**	**SIN → SIN**	**SIN → IN**
ResNet-18	**82.6** / 96.9	27.3 / 62.4	73.8 / 94.1	78.8 / 95.6
ResNet-34	81.1 / 96.8	28.3 / 62.2	73.3 / 93.6	79.3 / 94.8
ResNet-50	78.4 / 95.8	20.9 / 54.6	70.3 / 92.3	77.5 / 94.7
DenseNet-121	82.2 / **97.0**	25.0 / 58.9	74.8 / 94.2	**79.6 / 95.8**

used the DenseNet Huang et al. (2017) instead of the BagNet as the author of the original paper did because the BagNet has many large convolution kernels which needs long time and large memory to train.

The results in the table show that DenseNet-121 outperforms than every ResNet in all datasets. The models trained on IN have poor performance on SIN while the models trained on SIN have much better performance on IN. This may be a prove to the author's opinion that ImageNet trained models are biased towards texture since the texture of SIN and IN are almost totally different.

3.3 IMPROVE DNN TOWARDS SHAPE

The next experiment we did is to show that the SIN dataset can make the original DNN pay more attention to the shape of the input image.

Table 2: Results of RestNet 50 training on different training data. The parameters are set the same as that in the former subsection. The fine-tune part is trained on the IN dataset with 50 epochs.

training set	**fine-tune**	**top 1 acc**	**top 5 acc**
IN	-	78.4	95.8
SIN	-	77.5	94.8
SIN+IN	-	78.8	95.6
SIN+IN	IN	**89.0**	**97.9**

Table 2 clearly shows that using SIN+IN and fine-tune will improve the original model. As is shown in the table, if we train the ResNet-50 model on SIN+IN dataset, and then use IN to fine tune it, we got a better accuracy result than the results of all original networks on IN in the former subsection.

3.4 ROBUSTNESS

We also tested the robustness of the models. As it is proposed in the original paper, we tried to add several distortions to the validation set, including adding noises, changing the contrast and blurring. We added uniform noise and Gaussian noise with different amplitudes.

We blurred images using low-pass Gaussian filter with a fixed kernel size ($N = 15$) and different variance σ.

We can conclude from figure 3 that the ResNet-50 model trained on the original IN dataset is the least robust model. It performs the worst when getting any noise. However, the model trained on SIN and SIN+IN has similar performance. This means that training on the SIN dataset do have a positive effect on robustness.

4 DISCUSSION

The original paper contains many psychophysical experiments which are hard to reproduce. Also, training on such a large dataset (9 times ImageNet size for SIN+IN dataset) is almost impossible

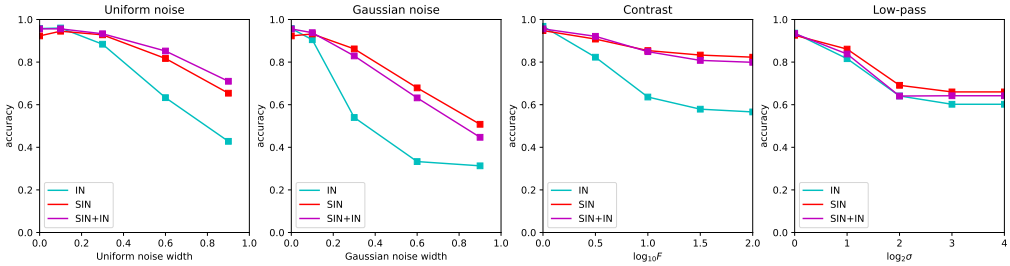

Figure 3: Top 5 accuracy on distorted images in validation set. ResNet-50 trained on SIN is more robust towards distortions than the same network trained on IN.

for our computing resources. Hence we need to weaken some of the experiments and focus on the important experiments which give the main idea of the author. But even we use the 16 classes ImageNet, it still needs about 2.5 minutes to train a single epoch on the SIN dataset. This means the computing time is another handicap in our experiments.

The method proposed by the original paper's author is essentially a data augmentation method. It can more or less eliminate the bias towards texture and improve the accuracy a little bit. However, the most important idea of the original paper is that we can use a stylized dataset as a complementary dataset of the original one. This is a new perspective of improve the networks since we used to change the structure of the model in order to get better performance.

5 Conclusion

In this report, we have successfully reproduced most important parts of the original paper and has proved the author's idea that the models trained on ImageNet is biased towards texture rather than shape. We've also proved that training on SIN or SIN+IN and fine-tuning on IN will eliminate this bias to some extent. This paper warns us that we should build dataset that has less bias towards texture since this is the way we humans behave. Or the model we trained may have poor generalization performance.

Acknowledgments

We would like to thank Dr.Jonathon Hare and Dr.Kate Farrahi for delivering the Differentiable Programming (and Deep Learning) module. The deep learning module and this reproducibility challenge is challenging but very meaningful. During reproducing the original paper, we've read a lot and learnt a lot about the state-of-the-art technologies in the deep learning field.

References

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.