INVESTIGATING THE REPRODUCIBILITY OF EMERGENCE OF SHAPE BIAS IN CONVOLUTIONAL NEURAL NETWORKS THROUGH ACTIVATION SPARSITY

Natthapath Rungseesiripak, Rashmi Rawat, Daniel Turato, Abhash Shrestha, & Ritam Behwal
Department of Electronics & Computer Science
University of Southampton
{nr2n23, rr4u23, dt3n23, as14n23, rb3n23}@soton.ac.uk

1 INTRODUCTION

Recent research has found that deep learning models for object recognition are heavily biased towards texture, while human visual perception relies more on shape and structure cues. The paper "Emergence of Shape Bias in Convolutional Neural Networks through Activation Sparsity" by Li et al. (2023) proposes a sparse coding constraint in convolutional neural networks (CNNs) can lead to the emergence of shape bias, similar to the human visual system. This report aims to assess the reproducibility of the key findings presented in this paper by replicating the experiments and analysing the results. Our code is hosted at: https://github.com/COMP6258-Reproducibility-Challenge/emergence-of-shape-bias.

2 REPRODUCTION APPROACH

2.1 TOP-K TRAINING INDUCES SHAPE BIAS IN RECOGNITION NETWORKS

2.1.1 MODEL SPECIFICATIONS

A review of the authors' code revealed three unmentioned details within the paper: (1) The stride of the first convolutional layer is set to 1, not the conventional 2 for ResNet-18 (He et al., 2016). (2) The input dimension of the classification layer is adjusted to 25,088, instead of the standard 512. (3) The SGD optimiser includes a momentum factor of 0.9 and a weight decay of $5e^{-4}$, which deviates from the mentioned standard SGD optimiser.

Two models were trained and evaluated to verify if the results were reproducible with only the details mentioned in the paper. The first set, Code-Spec (CS), matched the architecture and training details from the authors' code, which is assumed to produce the reported results. The second set, Paper-Spec (PS), strictly adhered to the architectural and training details outlined in the paper. Due to time constraints, both sets used a stride of 2 for the first convolutional layer instead of 1 since a stride of 1 preserves more input spatial details but significantly increases computational load and memory usage—extending training time from 1.5 minutes to 19 minutes per epoch on an NVIDIA RTX 3080.

2.1.2 IMPLEMENTATION DETAILS

Two additional non-overlapping subsets, IN-S3 and IN-S4, were created to validate the paper's results. The code used in this section is the refactored version of the code provided by the author. The code for the ResNet-18 model remains mostly the same, with changes to variable names and additional comments added to improve readability, albeit with the removal of parts deemed unnecessary for this experiment. However, the code regarding Top-K Mean Replacement was implemented from scratch using the equation described in the paper (Li et al., 2023) since it was not present in the author's code. The code for image transformation and splitting the dataset into training and test sets remains the same. The part of the code for the training loop was rewritten using `Torchbearer`.

2.2 TOWARDS SHAPE BIASED FEW SHOT IMAGE SYNTHESIS

2.2.1 DATASET

To validate the paper's results, we created two additional datasets of 100 images per class sampled from IN-S3 (hairy dog) and IN-S1 (French horn). The classes were selected to be texture-dependent

and intricate, respectively. This was done to evaluate the performance on texture and detail-heavy images, which require the model to understand texture instead of shape.

2.2.2 Model Specifications

After investigating the authors' code, it was noted that inter-layer noise injection was not allowed to be added even if you set the hyperparameter to True. FastGAN uses noise injection between generator layers for better generalization. Normal FastGANs trained in this investigation did not use noise-injection for comparison to Li et al. (2023) results. The code was modified, and additional tests with noise injections were conducted.

2.3 Shape Bias Benchmark

To reproduce the inference for the Top-K operation, we will re-use the authors' code, which integrates with the ModelvsHuman benchmark, proposed by Geirhos et al. (2021). We choose this over re-developing the inference as the benchmark is well-established and tested.

3 Results

3.1 Top-K Neurons Encode Structural Information

Shape information is mainly encoded in Top-K responses, while non-Top-K responses encode textures. This was confirmed using the original Jeep dataset. Top-K & non-Top-K images retained structural information, but using only non-Top-K responses resulted in a complete loss of structure.

To further test the validity of the claim and robustness of the Top-K neurons in terms of retaining the shape and structural information of the target object, cat images from The Oxford-IIIT Pet Dataset were used. This dataset is valuable for its variety: images with sharp subject-background distinctions and images with similar colours and textures. In an image of a cat on a similar-textured carpet, Top-K and non-Top-K layers retained some structure, but non-Top-K layers alone lost all structure. Likewise, non-Top-K layers still failed to retain structure in an image with a clear background-subject distinction and no textural similarity. See the README file within the texture-synthesis folder for the referenced cat images [1].

Further continuing with the Top-K neuron visualisation, another reconstruction approach was to update our optimisation objective. We were to replicate the original image in three different outputs: Top-K-mask only, non-Top-K-mask, and a combination of Top-K and non-Top-K-mask reconstruction. Of all three, only the non-Top-K-mask reconstruction was successfully implemented. Possible reasons behind the failed optimisation convergence and unfit visualisation may be the improper implementation of the masking technique or instability in the gradient descent process.

3.2 Top-K Responses already have Shape Bias without Training

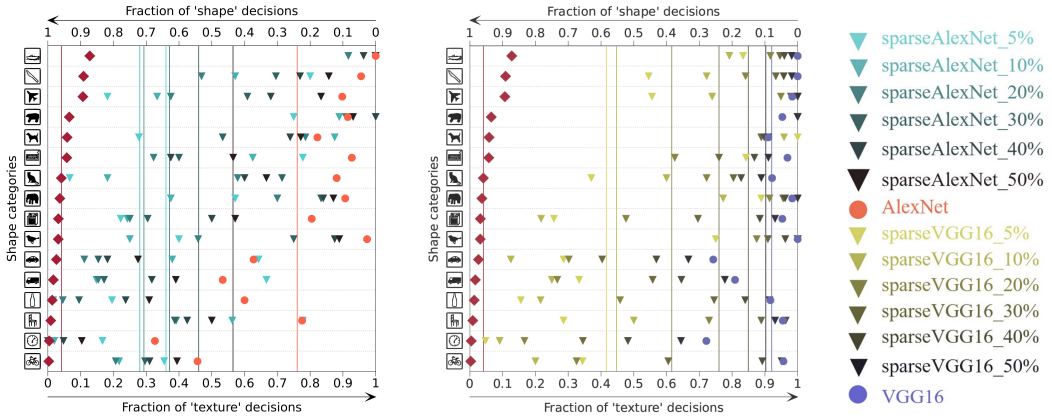
The ModelvsHuman benchmark was completed using the Iridis supercomputer, taking less than 10 minutes to run on all available models. As displayed in Fig.1, the benchmark produced surprisingly similar results as shown in the original paper. Unfortunately, the paper failed to disclose specific shape bias scores, so we could only compare them using their provided images. However, it is clear from the reproduced results that as the activations become more sparse, the fraction of shape decisions increases. See the results folder in our codebase for all benchmark results.[2]

3.3 Top-K training induces Shape Bias in Recognition Networks

Each model setting was trained and evaluated three times to obtain an error bar. All models were trained and evaluated on an NVIDIA RTX 3080, each training session taking around 1 hour and 14 minutes. Table 1 details the top-1 accuracies for the original and stylised subsets of ImageNet-1k (Deng et al., 2009) under the Code-Spec and Paper-Spec settings.

[1] https://github.com/emergence-of-shape-bias/texture-synthesis

[2] https://github.com/emergence-of-shape-bias/cnns-inference-top-k



(a) AlexNet (b) VGG

Figure 1: Shape Bias Benchmark using ModelvsHuman for a) AlexNet and b) VGG models

Table 1: The top-1 accuracy (%) of Code-Spec and Paper-Spec models. The performance on the stylised subset is reported in parentheses.

Model	IN-S1	IN-S2	IN-S3	IN-S4
CS ResNet-18	88.4 ± 0.3 (67.2 ± 1.8)	78.1 ± 1.8 (53.6 ± 3.1)	81.2 ± 0.8 (68.3 ± 0.9)	86.2 ± 0.4 (71.5 ± 0.7)
CS ResNet-18 w. Top-K	**89.3 ± 0.9 (70.1 ± 1.3)**	**80.7 ± 1.6 (61.3 ± 2.4)**	**85.4 ± 1.3 (75.3 ± 2.3)**	**87.4 ± 1.0 (74.1 ± 1.5)**
CS ResNet-18 w. Top-K Mean Rpl.	83.6 ± 0.7 (65.6 ± 1.0)	74.3 ± 0.5 (51.2 ± 2.6)	73.5 ± 6.6 (54.0 ± 3.9)	73.7 ± 0.5 (55.3 ± 1.0)
PS ResNet-18	92.1 ± 0.3 (**75.2 ± 0.7**)	**88.7 ± 0.3 (73.4 ± 0.4)**	88.6 ± 0.1 (73.6 ± 0.8)	**92.4 ± 0.4** (76.6 ± 1.0)
PS ResNet-18 w. Top-K	**93.0 ± 0.1 (75.2 ± 1.0)**	87.9 ± 0.3 (71.5 ± 0.7)	**89.7 ± 0.3 (76.8 ± 1.9)**	92.2 ± 0.2 (**78.7 ± 0.9**)
PS ResNet-18 w. Top-K Mean Rpl.	90.5 ± 0.5 (68.6 ± 1.5)	82.3 ± 0.7 (68.0 ± 1.0)	86.7 ± 0.4 (68.8 ± 1.3)	88.3 ± 0.2 (68.6 ± 0.5)

3.4 Towards Shape Biased Few Shot Image Synthesis

This experiment used the Iridis supercomputer. The training time was 12 hours per dataset per model to ensure FID would no longer decrease. Due to this high training time, the models were only run once. Additionally, each dataset was restricted to 100 samples to ensure a comparison between FID scores.

Model (FastGAN)	Jeep FID	Fish FID	Train FID	Table FID
Baseline	50.65	46.76	44.34	70.04
Top K	46.38	40.84	**40.23**	**59.31**
Top K + Noise	**40.92**	**37.06**	42.56	61.44

(a) GAN training results on the same dataset

Model (FastGAN)	IN-S1 (C-5) FID	IN-S3 (C-1) FID
Baseline	154.35	88.81
Top K=5	175.88	94.66
Top K=15	**151.80**	81.71
Top K=15 + Noise	162.50	**79.15**

(b) GAN training results on new datasets

Table 2: GAN training results for a) existing dataset, b) new datasets

4 Discussion

4.1 Top-K training induces Shape Bias in Recognition Networks

Table 1 indicates that, using the Code-Spec setting, introducing the Top-K operation during ResNet-18 training (Li et al., 2023) significantly enhanced performance on texture-transferred subsets and also improved performance on the original subsets (Deng et al., 2009). These findings are consistent with those reported in the paper. However, the enhanced performance claimed for the Top-K Mean Replacement (Li et al., 2023), which is reported to outperform traditional ResNet-18 (He et al., 2016) on texture-transferred subsets, could not be reproduced.

The results also show that the improvement in ResNet-18's performance with the Top-K operation on texture-transferred subsets was less pronounced in the Paper-Spec setting, with conventional ResNet-18 outperforming ResNet-18 with Top-K in some stylised subsets. Additionally, the accuracy on the original subsets neither dropped nor improved, and the improvement claimed from implementing the Top-K Mean Replacement remained unreproducible.

4.2 Towards Shape Biased Few Shot Image Synthesis

It can be seen from Table 2 that inducing Top-K operations in the FastGAN generator improves image generation quality. On the 4 datasets, Top-K operations resulted in an 8-15% reduction in FID scores. Surprisingly, noise injection in generator layers decreased FID in Jeep and Fish datasets. This trend was not consistent with every dataset.

Top-K operations also resulted in lower FID scores than vanilla FastGAN on the texture and detailed heavy additional dataset. Both datasets performed better with K=15 (15% of neurons active) at the 32x32 layer of the generator. Using noise injection increased FID in IN-S1-C5 but minimally affected in IN-S3-C1. Noise injection with Top-K does not seem to produce consistently better results, which theoretically it should Feng et al. (2020). Further experimentation is required to understand this discrepancy. This could be why the authors, Li et al. (2023), did not allow noise injection in their code.

5 Conclusion

This report largely reproduced the key findings from the original paper, confirming that the Top-K sparsity operation can induce shape bias in object recognition and image synthesis networks. The Top-K operation improved ResNet-18's performance on texture-transformed images, and inference results showed increased shape decisions with sparser activations. However, the enhanced performance of Top-K mean replacement could not be reproduced, and combining Top-K with noise injections in GANs did not produce consistent results. These findings support the conclusion that sparse coding can induce shape in deep learning networks, suggesting sparsity's potential role in human visual systems.

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, 2009. doi: 10.1109/CVPR.2009.5206848.

Ruili Feng, Deli Zhao, and Zhengjun Zha. On noise injection in generative adversarial networks. *CoRR*, abs/2006.05891, 2020. URL `https://arxiv.org/abs/2006.05891`.

Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. In *Advances in Neural Information Processing Systems 34*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Tianqin Li, Ziqi Wen, Yangfan Li, and Tai Sing Lee. Emergence of shape bias in convolutional neural networks through activation sparsity. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=QzcZb3fWmW`.