REPRODUCABILITY CHALLENGE - WHITE NOISE ANALYSIS OF NEURAL NETWORKS

Zeyu Yan, Zixu Guo, Peilin Zhan, Tianqi Huang
Electronics and Computer Science
University of Southampton
Southampton, UK
{zy10u21, zg4g21, pz1n21, th3u21}@soton.ac.uk

ABSTRACT

The International Conference on Representational Learning (ICLR) has received an increasing number of papers in recent years, and the organization is encouraging researchers to try to replicate the work in papers submitted to the conference. This paper aims to reproduce the work of the paper "White noise analysis of neural networks" The performance of the original paper will be analyzed and reviewed to verify reproduction and summarize the advantages and disadvantages of the project.

1 INTRODUCTION

Classification images and pulse-triggered analysis have been widely used in psychophysics and neurophysiology. This paper will verify the original paper "White Noise Analysis of Neural Networks" by Borji & Lin (2020). We will conduct experiments based on the CNN and MNIST datasets, using classified images and spike-triggered analysis to reveal hidden biases in deep networks. Moreover, our reproduction mainly focus on classifier bias, adversarial attack, filter visualization and psychometric curves in the original paper.

2 CONCEPTS

In the original article Borji & Lin (2020), classification images technique and spike triggered analysis Marmarelis (2012) are recruited to study the biases in neural networks. In this method, a signal and a noise image are summed to produce the stimulus n. Classification image is calculated as:

$$c = (\bar{n}^{12} + \bar{n}^{22}) - (\bar{n}^{11} + \bar{n}^{12}). \tag{1}$$

The principle of this method is that the noise pattern in some experiments has similar features to one of the signals, so it will bias the observer to choose the signal. The stimulus is a linear combination of noise plus signal as:

$$t = \gamma * s + (1 - \gamma) * n; \gamma \in [0, 1], \tag{2}$$

which combines the images with the random generated white noise. Different value of γ in an image of digit 9 is shown in Fig1.

The spike triggered analysis(STA) is also known as "white-noise analysis". STA is the average stimulus preceding a spike. When the stimulus distribution is spherically symmetric, an unbiased estimate of the neuron's receptive field is analysed, which is defined by Sandler & Marmarelis (2015) as

$$STA : \mu = \frac{1}{n_{sp}} \sum_{i=1}^{N} y_i x_i, \tag{3}$$

In our work, we will base on Eq. 1 to reproduce the classification in neural networks mainly using the MNIST dataset. And make trials in different digits in MNIST and different value of γ in Eq. 2. In this way, we will verify whether this method is efficient to classify features in different models and in different proportion of noise in the images. Also, we will use STA to analyse classifier bias and the adversarial attack based on Eq. 3.

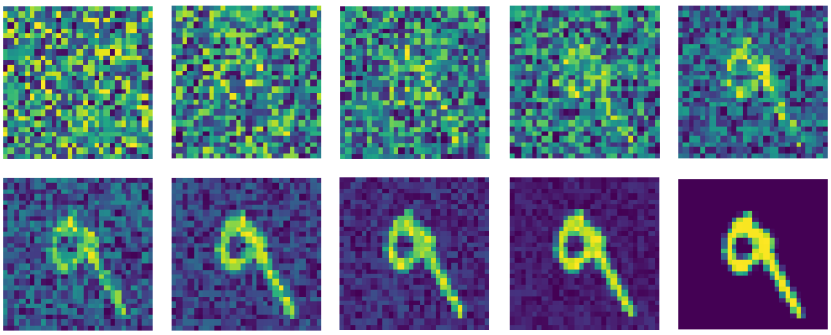

Figure 1: Images combined with different proportion of noise

3 IMPLEMENTATION

We reproduced three use cases of classification images and STA to examine neural networks including studying classifier analysis, adversarial attack and filter visualization. The whole reproduction work was done on the Google Colab Platform using the PyTorch framework and the majority of the code we used and modified is based on the description in the paper. The repository of our code is https://github.com/COMP6248-Reproducability-Challenge/White-Noise-Analysis.git

For the first task, we used a CNN with 2 convolutional layers, 2 pooling layers and 1 fully-connected layer and trained it on the MNIST dataset. And then we generated 1 million white noise images to feed into the model, tested the classification results and draw the confusion matrix. In addition, we created the average noise maps on the basis of the noise map and fed them back into the CNN model.

Deep neural networks have been very successful in a variety of visual recognition tasks, but they can sometimes be influenced by some perturbation that are imperceptible to a human. Therefore, we combined different bias maps with the input digits and computed the classification results on CNN. The combination of input is calculated by Eq. 3. By changing different values of γ, we can see how the decision of CNN is influenced. Also, we added the bias map to noise to see if the model is influenced by that combination.

In filter visualization, we randomly generated patterns to feed into the model and recorded the average response of the neurons at each layer. The visualization of the filter in the model is shown in the following section. And we also drew Psychometric curves to indicate the change of accuracy over different magnitude of the signal added to the noise.

4 RESULT AND ANALYSIS

4.1 CLASSIFIER BIASES

In this section, we reproduce the visualization of the classifier bias. We generate 10M white noise images and feed them to the CNN. Humans can easily identify which number the bias map represents for various digits, as shown in Fig 2 A, and we reproduce the generated image results that are mainly similar to those of the original paper. However, We notice the frequency of 8 is huge which means the majority of the noise patterns are classified as 8, possibly because this number has many characteristics with other numbers.

Additionally, the 8 is classified as 2. As seen in Fig 2 B, the largest dot product is not 8 which determined the output digit. This supports the original paper's conclusion that CNN extracts characteristics in the same way that the human visual system does, and so has similar mechanisms and biases.

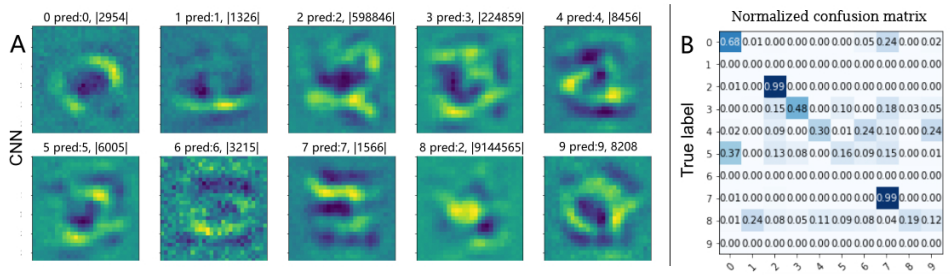

Figure 2: A) Classification images of a CNN trained on MNIST. Image titles show ground truth, predicted class for the bias map, and the frequency of the noise patterns classified as that digit. B) Confusion matrices, the classification was done via template matching using dot product.

4.2 Adversarial Attack

In this section, adversarial attacks and defenses of deep neural networks in visual recognition will be reproduced. In order to verify the conclusions in the original paper, we selected the number 7 as the target class for experiments. After adding the bias of the numbers 0 to 9 and the average image to the numbers and noise respectively, the classifier is used to make predictions, from Figure 3A we can see that in many cases adding a bias to a number will recognize it as target category. And in almost all cases, adding a bias to the noise turns the noise into the target number. This coincides with the conclusion in the original text. Additionally, by comparing Figure 3A and B, we found that adding the mean image was more effective than adding the bias.

This shows that by adding bias to digits and noise, the CNN's decision for a specific digit class (for MNIST) is indeed influenced for the purpose of adversarial attacks.

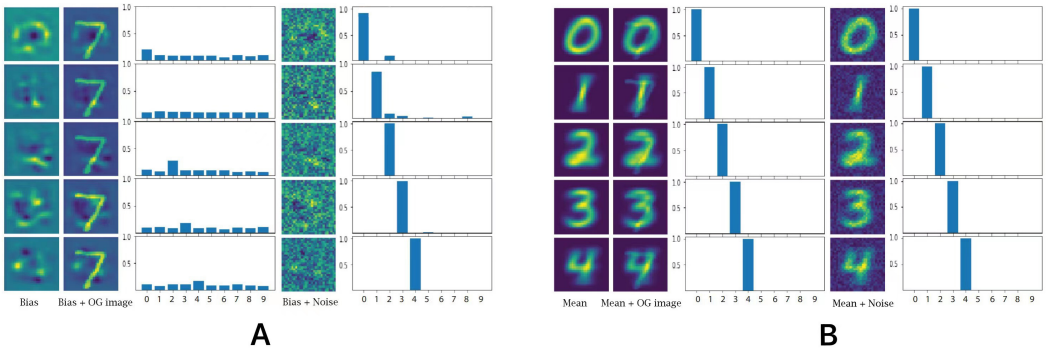

Figure 3: A) In many cases, adding bias to a number will change it to the target category. In almost all cases, adding a bias to the noise will turn the noise into the target number. B) Same as A, but using the mean.

4.3 Filter Visualization and Psychometric Curves

In this section, we reproduce the visualization of the filter in the neural networks. As $ReLU$ resulted in dead filters, the activation functions in the convolution layers of CNN is changed to $tanh$. We also plot the Psychometric curves of different convolution layers of CNN according to the original paper.

It is shown in Fig 4 that the filters using STA for the first two convolution layers of CNN trained on MNIST is visualized just as the same in the original paper. But the Psychometric curves in our reproduction has lower accuracy comparing to the paper, where the redder curve means more bias. The result of the Psychometric curves in original paper shows that increase the bias enhances the recognition towards the target digits. However, the red curve in our reproduction does not have sufficient accuracy shown in Fig 4. Despite, We can still get the same result that the larger γ (higher proportion of noise in images) enhances the accuracy of the classification.

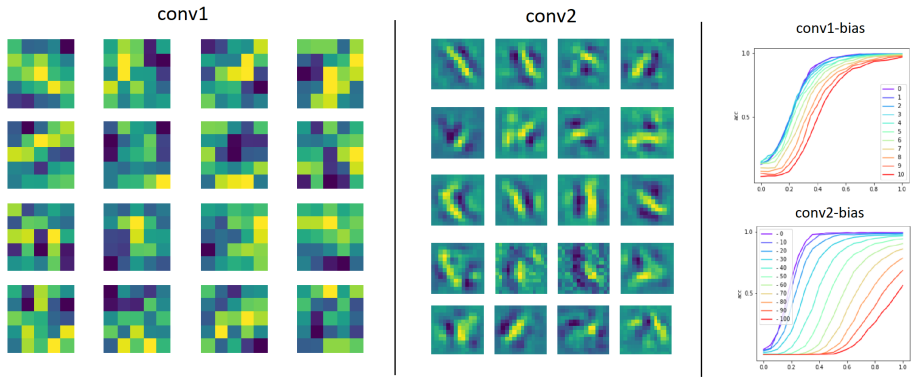

Figure 4: Filters using STA for the first two conv layers and Psychometric curves of CNN trained on MNIST

5 Conclusion

In our reproduction, we reproduce classification images technique based on spike triggered analysis successfully. This report demonstrates the core ideas of this paper. Two popular methods in computational neuroscience, classification image and spike triggered averaging, are used to understand and explain the behavior of artificial neural networks.

Classification results are evaluated in the MNIST dataset by adding CNN bias maps and mean images, and these methods achieve good performance and better results with mean images. In addition, the conclusion of the original paper is verified by reproducing the Psychometric curves, which shows that increasing the bias enhances the recognition of the target digits as well as larger γ (higher proportion of noise in images) enhances the accuracy.

We also find that the article has some shortages during our reading and reproduction. The theoretical dimension of this article is weak, and the existing model is not improved, but used to analyze the results on a variety of data sets. And some results are not as accurate as describes in the article, which is shown in Section 4.

Due to the limitation of computational resource and time, we only reproduce with MNIST data set and modify some parameters. We believe better results can be produced with further modifications (e.g., using better distance measures between an image and the average noise map for each class). Also, it is likely that increasing the number of samples will lead to better performance.

References

Ali Borji and Sikun Lin. Unveiling hidden biases in deep networks with classification images and spike triggered analysis. In *International Conference on Learning Representations*, 2020.

Vasilis Marmarelis. *Analysis of physiological systems: The white-noise approach*. Springer Science & Business Media, 2012.

Roman A Sandler and Vasilis Z Marmarelis. Understanding spike-triggered covariance using wiener theory for receptive field identification. *Journal of vision*, 15(9):16–16, 2015.