New Initialization Mechanisms for Convolutional Neural Networks

Keyu Long kelong@ucsd.edu
Grace Lam pklam@ucsd.edu
Tiffany Yu z5yu@ucsd.edu
Daniel Shi Dshi@ucsd.edu
Licheng Hu l2hu@ucsd.edu
Yian Ma (Mentor) yim013@ucsd.edu
Misha Belkin (Mentor) mbelkin@ucsd.edu

Abstract

The exploration of feature evolution during the training of neural networks is crucial for advancing our understanding of modern machine learning methodologies. Equally important is the role of initialization, which significantly impacts model performance and convergence. This research aims to investigate the integration of Neural Feature Matrix (NFM) and Average Gradient Outer Product (AGOP) into initialization methods to enhance neural network training. Employing the VGG11 model across diverse datasets (SVHN, CIFAR-10, CIFAR-100, Tiny-ImageNet), we compare traditional and NFM/AGOP-based initialization methods to assess their effects on model accuracy and learning efficiency. Our study seeks to demonstrate the potential advantages of NFM and AGOP in initialization, potentially offering new insights into the optimization of neural networks for improved feature learning and convergence. The findings could have significant implications for the development of more efficient and effective deep learning models, marking a step forward in the field.

Website: https://hulicheng117.github.io/DSC180-website/
Code: https://github.com/KULcoder/DSC180

1 Introduction

Neural networks, mirroring the human brain's pattern recognition abilities, have revolutionized machine learning. Their success spans various fields, from healthcare to finance, showcasing exceptional problem-solving and data interpretation capabilities. As neural networks increasingly outperform traditional algorithms, it becomes crucial to unravel their complex learning mechanisms.

Previous studies Beaglehole et al. (2023) and Radhakrishnan et al. (2023) formulated Convolutional Neural Feature Ansatz, demonstrating that features selected by convolutional networks can be recovered by computing the average gradient outer product (AGOP) of the trained network with respect to image patches given by empirical covariance matrices of filters at any given layer. Concurrently, these investigations identified Average Gradient Outer Product (AGOP) and Neural Feature Matrix (NFM) as key elements characterizing feature learning in neural networks.

Yet, the initialization process of neural networks, a fundamental aspect influencing their performance and ability to converge, remains a challenging domain. Improper initialization, exacerbated by backpropagation's nuances, can trigger the vanishing or exploding gradient problem, hindering the training process.

Our study aims to combine the concepts of NFM and AGOP with initialization methods. This exploration seeks to address the question: **How does the application of the Neural Feature Matrix and Average Gradient Outer Product as initialization affect the performance of neural networks?**

- **NFM**, denoted by $W^T W$, is the neural feature matrix resulted by multiplying model's weight matrices.
- **AGOP** is the average gradient outer product over patches, it is the gradient with respect to that patch average over data.
- Previous studies posit the **Convolutional Neural Feature Ansatz** Beaglehole et al. (2023), which states that there is a positive correlation between AGOP and NFM.

$$w^T w \propto \frac{1}{n} \sum_{p=1}^{n} \nabla f(x_p) \nabla f(x_p)^T$$

- The significance of AGOP and NFM is highlighted by findings suggesting that these measures, of early layers, perform operations similar to edge detection.
- As shown in the Figure 1, Patch-AGOPs and NFMs from pre-trained VGG11 identified edges in images and progressively highlighted regions of images used for prediction.

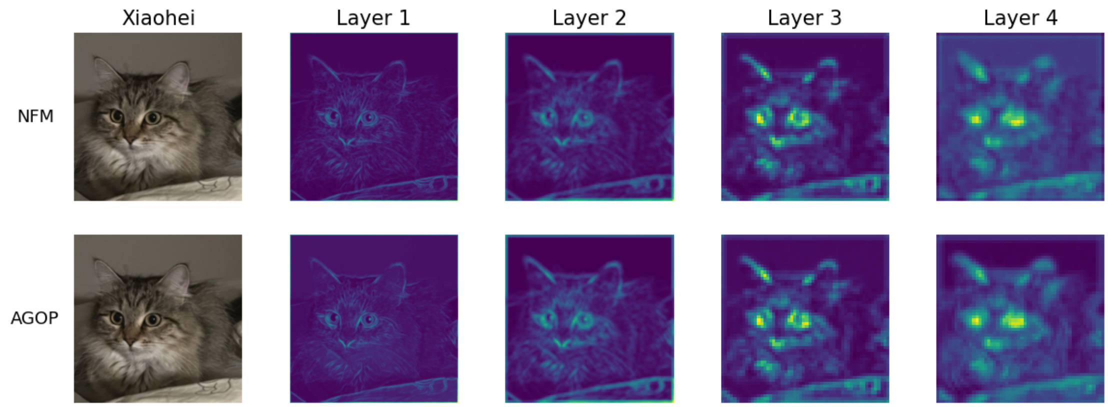

Figure 1: AGOP and NFMs of pre-trained networks identified relevant features for prediction.

2 Methods

2.1 Dataset

To investigate the application of the Neural Feature Matrix (NFM) and Average Gradient Outer Product (AGOP) as initialization methods, we will be examining their performance across four different datasets: SVHN, CIFAR-10, CIFAR-100, and Tiny ImageNet.

SVHN (Street View House Numbers) is a dataset that contains a total of 600,000 digit images from Google Street View. Each image is of the size 32x32 pixels in size. This dataset is instrumental in assessing how well the initialization methods facilitate the model's ability to recognize numerical digits in varied conditions.

CIFAR-10 (Canadian Institute for Advanced Research, 10 classes) is comprised of 60,000 32x32 color images across 10 classes, providing a balanced platform to evaluate the impact of initialization on general object recognition across a moderate number of classes.

CIFAR-100 (Canadian Institute for Advanced Research, 100 classes) an extension of CIFAR-10 with 100 classes, introduces a higher level of complexity and granularity in classification. This dataset tests the initialization methods' efficacy in a more challenging and diversified setting.

Tiny-Imagenet offers a scaled-down version of the ImageNet challenge, featuring images that are significantly reduced in resolution and variety compared to the full ImageNet dataset. It serves as an intermediate step towards complex image recognition tasks, further pushing the boundaries for the initialization methods under investigation.

2.2 Methods Overview

The core of our investigation revolves around the application of AGOP and NFM in the initialization phase of neural network training, specifically using the VGG11 architecture.

The VGG11 model, known for its simplicity and effectiveness in image classification tasks, provides a solid foundation for our experiments.

- **Model**: We used a VGG11 model which has 8 convolutional layers followed by 3 fully connected layers.
- **Method:** Investigate the performance of AGOP and NFM initialization compared to other common initializations by examining accuracy and loss.
- **Dataset:** SVHN, CIFAR-10, CIFAR-100, Tiny-ImageNet.
- **Initialization Methods:**
  - **Normal**: extract weights from a normal distribution.
  - **Uniform**: extract weights from a uniform distribution.
  - **AGOP and NFM:** We use the pre-trained model from PyTorch to extract AGOP and NFM, and use these as the covariant matrix to generate a multivariate Gaussian for sampling weights for initialization.
  - **Kaiming**: Kaiming initialization is one of the most effective methods for initializing convolutional layer weights, it is also the default initialization method for Conv2d in PyTorch. The method changes the parameters (e.g. std for normal, range for uniform, covariance matrix NFM) of the distribution based on layer width and the type of activation function.
- **Experiment and Hyper-parameters:** Our experimental setup is designed to isolate the effect of initialization on model performance. By employing Stochastic Gradient Descent (SGD) with a learning rate of 0.001 and cross-entropy loss, we maintain consistency across experiments, ensuring that any observed differences in performance can be attributed to the initialization method. The overarching goal is to ascertain the extent to which AGOP and NFM can enhance neural network training, reflected in improvements in accuracy and loss metrics.

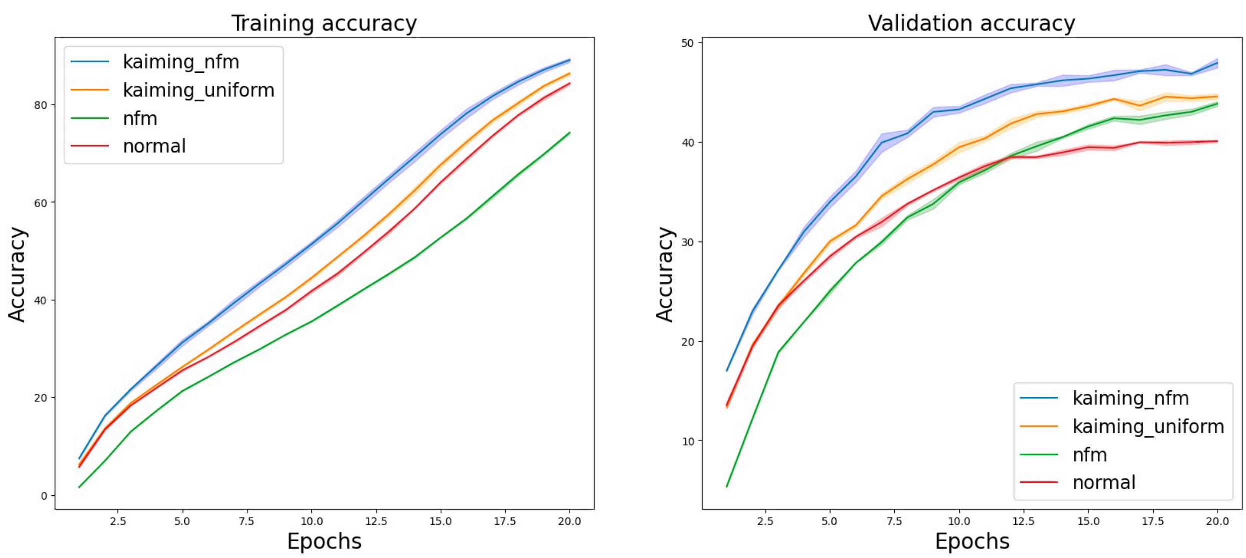

Figure 2: Training graph on CIFAR-100 with different initialization functions

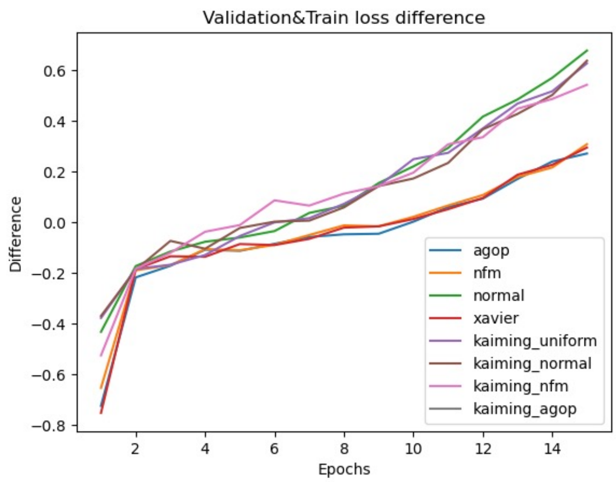

Figure 3: Difference between Validation and Training Loss on CIFAR-100

In our training graph we can see both training and validation accuarcy of Kaiming_NFM outperform the Kaiming_uniform which is the default initialization method used in pytorch on CIFAR-100. We also ploted the difference of validation loss and training loss for each initialization method acrossed training process. When the difference is negative, the validation loss is lower than the training loss. This can be a sign of underfitting, meaning the model is not capturing the underlying trends in the training data enough. When the difference is positive (above 0 on the y-axis), the validation loss is higher than the training loss and indicate overfitting. From the graph, our initialization method of agop and nfm showed robust to overfit than other initialization method and Kaiming_nfm also outperformed among all initialization methods with Kaiming initialization scaling.

3 Results

Table 1: Best Validation Accuracy Across Different Initialization Methods and Datasets

Initialization Method	SVHN	CIFAR10	CIFAR100	Tiny-ImageNet
Normal	92.35	74.26	40.18	27.73
Xavier Normal	92.66	76.56	43.09	28.49
Kaiming Uniform	93.02	77.32	44.76	33.53
Kaiming Normal	93.13	77.10	45.31	32.38
AGOP	91.71	78.61	42.84	29.56
NFM	92.79	**79.32**	44.09	31.34
Kaiming NFM	**93.23**	78.69	**48.33**	**33.78**

The standout performer among the initialization methods tested was the Kaiming NFM, which demonstrated superior performance in three out of the four datasets, achieving the highest validation accuracies of 93.23% on SVHN, 48.33% on CIFAR-100, and 33.78% on

Tiny ImageNet. Notably, NFM initialization led the performance in CIFAR-10 with an impressive 79.32% accuracy, underscoring its effectiveness in diverse learning environments.

These results underscore the robustness of NFM and Kaiming NFM initialization methods, particularly their ability to improve model performance across a range of complex datasets. Furthermore, our findings suggest that these methods offer a degree of protection against overfitting, contributing to their superior performance metrics.

4 Discussion

Our initialization methods can be viewed as a soft way of transferring learning. We used the pre-trained model on ImageNet and train the model on simpler datasets. It is worth investigating what if we initialize with model pre-trained on simpler datasets and train on more complex datasets. Some other potential areas of improvement include:

- Perform hyperparameter tuning to obtain the best performance of each method.
- Try to repeated training the model and extracting AGOP/NFM to see if it can improve the performance.
- Try to expand the method to transformers.

5 Conclusion

In conclusion, our study observed the impact of advanced initialization methods on the performance of neural networks, with a specific focus on the Neural Feature Matrix (NFM) and Average Gradient Outer Product (AGOP). The investigation centered on the application of these methods to the VGG11 model, trained on a range of datasets from SVHN to Tiny ImageNet.

Our findings suggest that the integration of NFM and AGOP with traditional initialization methods can lead to substantial improvements in validation accuracy. Notably, the Kaiming NFM initialization outperformed standard practices in several datasets, marking a significant step forward in neural network training strategies. The implications of this advancement are profound, offering potential enhancements in various applications where deep learning models are pivotal.

6 Acknowledgments

We would like thank our mentors, Misha Belkin, Yian Ma, for their invaluable guidance and support throughout this project.

Special thanks to Daniel Beaglehole for his guidance.His insights have been instrumental in shaping the success of this project.

References

Beaglehole, Daniel, Adityanarayanan Radhakrishnan, Parthe Pandit, and Mikhail Belkin. 2023. "Mechanism of feature learning in convolutional neural networks."

Radhakrishnan, Adityanarayanan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. 2023. "Mechanism of feature learning in deep fully connected networks and kernel machines that recursively learn features."