The report of FishNet: A Versatile Backbone for Image, Region, and Pixel Level Prediction

Chengbo Zang (cz2678)
cz2678@columbia.edu

Yuqing Cao (yc3998)
yc3998@columbia.edu

Fengyang Shang (fs2752)
fs2752@columbia.edu

ABSTRACT

With increasingly demanding tasks in computer vision problems, neural network models go deeper and are divided into image level, region level and pixel level tasks to deal with more complex situations. In this project, we reviewed the paper *FishNet: A Versatile Backbone for Image, Region, and Pixel Level Predictions*, which proposed a new model suitable for all level predictions. We reproduced the model and conducted several experiments under necessary modifications with Mnist, Cifar10 and Cifar100 datasets. It can be deduced that FishNets could achieve better classification accuracy with less computational cost. All codes of our project can be found here.

I. INTRODUCTION

It has been a long time since human beings showed great interest in brains. After the computer was invented, some experiments which mimicked the work style of brains appeared. In 1949, Hebb firstly used some rules or principles and this is the first step of machine learning [2]. As time passed, people began to try to make their computers think and learn by themselves. Then, a person named Frank Rosenblatt who can be seen as the father of deep learning came up with a model [3]. This model was not as complicated as today's neural networks, but it is the first model which can do some judgements as a human. This was fantastic and encouraging news to make every one working in this field feel excited. However, due to the limitations of the technology at that time, deep learning halted for many years.

But nowadays in the 21st century, due to the dramatically improved computers and a famous person, Hinton, deep learning has become one of the most popular fields. In [4], he illustrates that multilayer neural networks have a very effective way to predict or learn features from any datasets theoretically. However, the model is still simple, and has lots of limitations or challenges that need to be resolved.

Then nowadays deep learning can use the models to classify, recognize and make some predictions for complex tasks with considerably high accuracy. Challenges lie that all models are built for image-level predictions. But using these models in different levels such as region-level, or pixel-level without changing the models requires a very high resolution, and at the same time, the pixel level and region level tasks still want information with high semantic meaning. Most of the models forget to consider such situations.

In this paper we studied, the authors came up with a new kind of model, which tries to get all the features from all resolutions with high semantic information and send all of them to predict the final results, and the features can directly propagate from the deep layers to the shallow layers.

II. LITERATURE REVIEW

In [1], the authors introduced the FishNet including its design purposes, improvements and why it has such advantages compared with traditional neural network models.

Firstly, the paper shows the related works, such as the traditional solutions to solve the problems of direct propagation to tackle vanishing or exploding gradients. Conventionally, to make the prediction easier and improve generalization, identity mapping and concatenation are the most common structures. The principle of weighted identity mapping can be illustrated as follows:

$$x_{l+1} = h(x_l) + F(x_l, w_l) \quad (1)$$

$$h(x_l) = \lambda * x_l \quad (2)$$

Recursively, we apply (3) to calculate the x_{l+1},

$$x_{l+1} = \prod_{i=1}^{L} \lambda_i x_0 + \sum_{i=1}^{L} F(x_i, w_i) \quad (3)$$

And the back propagation equals to:

$$\frac{\partial \epsilon}{\partial x_l} = \frac{\partial \epsilon}{\partial x_L} \left(\left(\prod_{i=1}^{L} \lambda_i \right) + \frac{\partial}{\partial x_l} \sum_{i=1}^{L} F(x_i, w_i) \right) \quad (4)$$

Observe that in Isolation Convolution as is stated above, the additive term is $\prod_{i=1}^{L} \lambda_i x_0$. For deep networks, this factor will finally become exponentially large or small, and will cause the parameter to explode or varnish and hinder the back propagation. These two situations will make the optimization more difficult. Unluckily, identity mapping [5] and concatenation will not solve the problem when dealt with different resolutions and different input channels.

In order to address this problem, the author elaborates the FishNet's structure. The FishNet can be divided into three parts, fish tail, fish body and fish head. In the tail part, the authors implement existing networks such as CNN and ResNet. There is a concept of "stage", which is a stack of all the layers with the same resolutions. A bunch of data going through residual blocks in the tail will be directly passed to the body part where they will be refined and preserved by concatenation. In the head part, the structure will preserve and refine all the data from the previous stage of the head and the body part. Finally, the data will be sent to the head part and also preserved and refined with the previous data in the head in the DR-block to keep the whole information to make a prediction. All of these will be carefully constructed to avoid the isolated convolution in the fish body and fish head. In this way, the gradients can be easily propagated directly.

III. METHODS

A. Network Construction

The model used in this project is reconstructed using Tensorflow Keras API. The code maximally preserves the configurability and versatility of the original code by decomposing the network into different models. All application structures are assembled through factories with highly customized configurations.

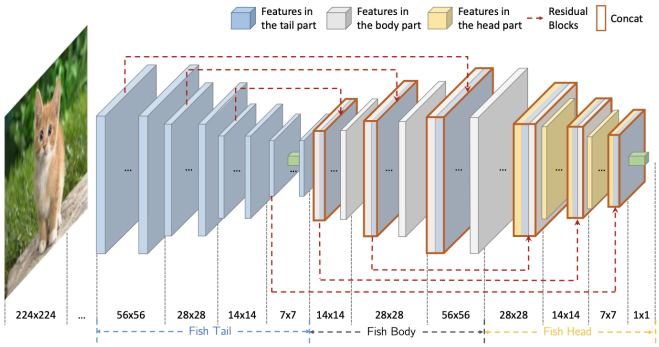

FIG. 1. Overall network structure [1]

Considering the original dataset (ImageNet) that Fishnet was used on has as much as 1000 classes with more than 13 million images, it is necessary for us to reduce the model capacity and complexity before using it on simpler datasets for illustration purposes. Based on the original Fish-Net99 model (Table 1), we constructed two other simplified networks, FishNet77 (Table 2), which reduces the number of channels of every stage to ¼, and FishNet55 (Table 3), which cuts the number of

stages in every body part to 2 and further reduces the number of channels to ⅛.

TABLE 1. Network structure of FishNet99

	Tail			Body				Head		
Stage	1	2	3	(se)	3	2	1	1	2	3
Shape	28	14	7	7	14	28	56	28	14	7
Channels	128	256	512	512	512	384	256	320	832	1600

TABLE 2. Network structure of FishNet77

	Tail			Body				Head		
Stage	1	2	3	(se)	3	2	1	1	2	3
Shape	28	14	7	7	14	28	56	28	14	7
Channels	32	64	128	128	128	96	64	80	208	400

TABLE 3. Network structure of FishNet55

	Tail		Body			Head	
Stage	1	2	(se)	2	1	1	2
Shape	28	14	7	14	28	14	7
Channels	16	32	32	32	24	32	48

Apart from modifying the whole network structure, details with individual block construction are also to be examined. Specifically, the structure of "Score Blocks" (the output layer) in the original network is shown on the left side of Fig. 2. Although this kind of structure is able to significantly reduce the total number of parameters, the network performances on simple datasets are drastically poor (as is shown in Fig. 4). Therefore, we further substituted the last two layers (GlobalAverage- Pooling and Convolution2D) with a flatten layer and a dense layer.

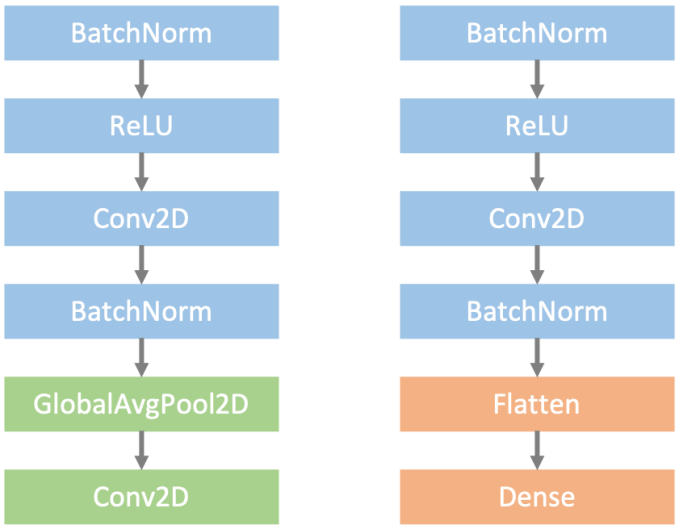

FIG. 2. Score Block structure based on convolution layer and dense layer

The results of our experiment on the Cifar10 dataset show that the modified structure is learning much more effectively compared to the original structure with almost identical training time per step.

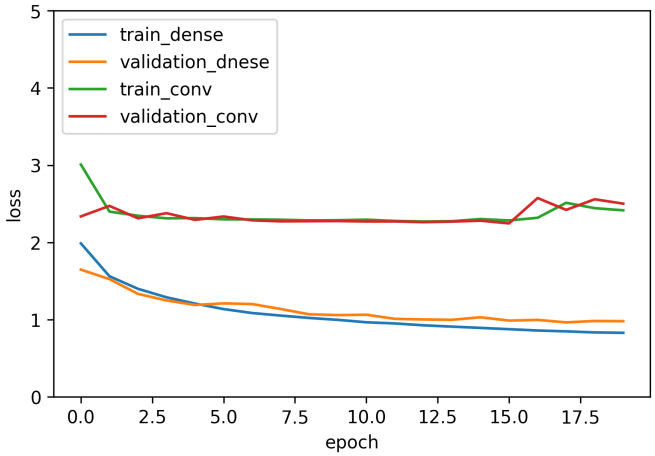

FIG. 3. Loss with dense and convolutional outputs on Cifar10 dataset

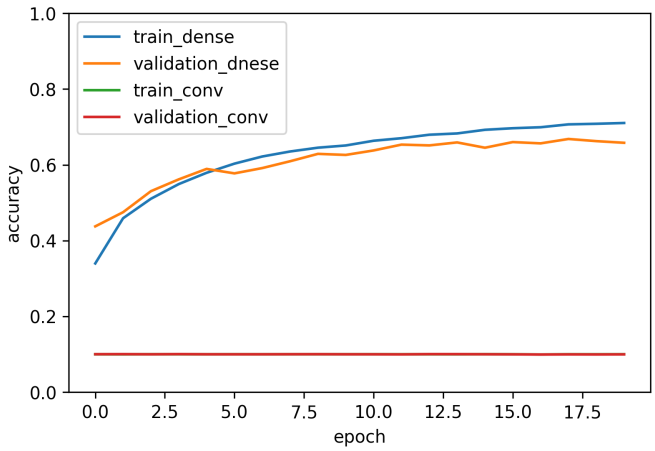

FIG. 4. Accuracy with dense and convolutional outputs on Cifar10 dataset

It can be observed that while the training and validation loss of dense and convolutional output are both descending, the accuracy indicates that the convolutional-based output layer fails to capture useful information to make proper judgements. Hence, all the experiments conducted later in our project used a dense layer at the output.

B. Training Process

Initially, we apply image resizing and data augmentation to the data pre-processing process, aiming for better performance. All the parameters in the processing process have been set to the same, to avoid any possible disturbances. Subsequently, we train different datasets using the improved FishNets and other models. In order to enhance the efficiency of our training, we create a bash file allowing training in parallel to best facilitate the process. Finally, we calculate the loss and top-1 accuracy at every epoch after several repeating experiments.

IV. EXPERIMENTS AND RESULTS

As mentioned in the above section, we made some modifications to the architecture of networks in [1] to reduce the capacity and enhance the performance of the original model. In this section, we focus on two types of modified FishNets named FishNet55 and FishNet77, conduct experiments to reproduce some results in [1], and compare their performances with neural networks proposed in other papers.

A. FishNet with Different Capacity

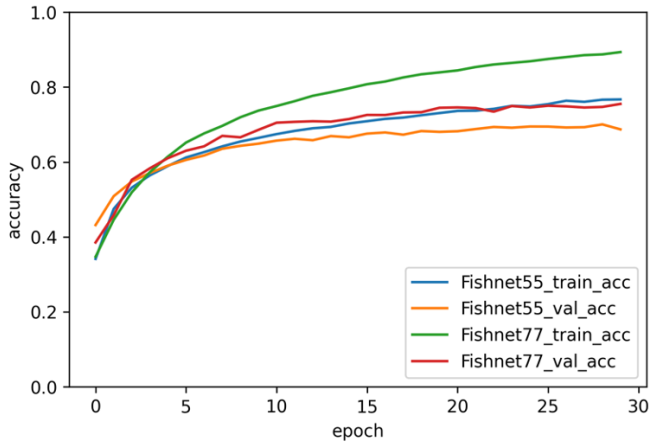

FIG. 5 Training and test accuracy of FishNet55 and FishNet77 with Cifar-10 dataset

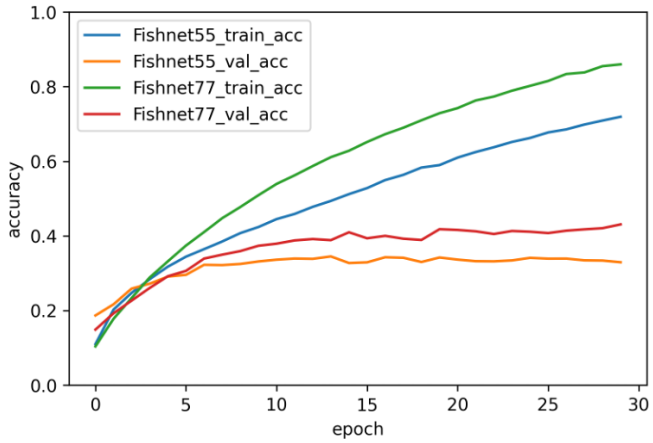

FIG. 6 Training and test accuracy of FishNet55 and FishNet77 with Cifar-100 dataset

Firstly, in order to compare the performance of FishNet55 and FishNet77 in training and testing tasks, we choose two datasets for experiments, Cifar-10 and Cifar-100. The Cifar-10 dataset has color images in 10 classes and the Cifar-100 dataset consists of 100 classes of RGB images. The results of Cifar-10 and Cifar-100 correspond to FIG.5 and FIG.6 respectively. The training and test accuracy of FishNet55 are plotted as blue and orange solid

lines respectively, while those of FishNet77 are shown as green and red solid lines.

It could be observed from FIG.5 and FIG.6 that in the training process, whatever the dataset is, the accuracy of FishNet77 is higher than that of FishNet55, and the difference tends to be larger with the increase of the number of epochs. These features also apply to the test results, where the performance of FishNet55 is always worse than FishNet77. This is consistent with our intuition that networks with bigger capacity invariantly produce better performance in not easy tasks. It could also be seen that the test accuracy of FishNets with Cifar-100 achieves around 30%, in comparison to the accuracy around 70% with Cifar-10, in spite of the fact that both of the training accuracy can get close to 80%. This indicates that it is easier to be underfitting in a harder task, which is also in harmony with the conclusion we learnt in neural networks and deep learning.

B. Comparison with Other Networks

Subsequently, for the purpose to obtain more concise results in the performance of FishNet55, FishNet77 and other neural networks. We choose two conventional models, a one-hidden layer MLP, which contains 256 nodes in its hidden linear layer, and a two-hidden layer CNN, which has a 2D convolutional layer with the kernel size of 5 and a linear layer with 520 nodes. Additionally, after literature search in modern popular networks, we found that MobileNet is an efficient model which balances the depth of the architecture and the computational complexity [7]. As presented in TABLE 4, we evaluate their performance using an easy dataset Fashion-MNIST and a harder dataset Cifar-10. After 10 times of training and testing, the mathematical average of testing accuracy, training times and number of parameters are shown in every column in the table.

TABLE 4. Comparison in Different Networks

	Fashion-Mnist		Cifar10		Parameters
	Accuracy (%)	Time (s/epc)	Accuracy (%)	Time (s/epc)	
MLP	86.5	1.3	10.1	9.11	2,411,274
CNN	86.0	1.4	37.3	10.13	3,675,534
MobileNet	87.4	6.19	62.0	23.29	4,253,864
Fishnet55	88.6	11.36	70.1	48.30	118,752
Fishnet77	90.1	24.76	75.5	67.43	2,317,636

It could be concluded that for Fashion-MNIST, the accuracy of MobileNet is slightly higher than conventional models such as MLP and CNN. And the accuracies of FishNet55 and FishNet77 are even higher than that of MobileNet, reaching about 90% for FishNet77. This conclusion also applies to Cifar-10, where the advantage of FishNets is much more apparent. It could be seen from the table that the test accuracy of FishNets could achieve above 70% while the simple MLP could not make any useful predictions, indicating that one hidden-layer is too shallow to succeed in the classification task for Cifar-10. Furthermore, it should be noticed that the number of parameters of FishNet55 is much smaller than that of MLP, CNN and MobileNet, which means that it achieves better performance than other models with a smaller model capacity. And this is also true for FishNet77. Therefore, FishNet outweighs other neural networks in view of both classification accuracy and computational cost, supporting the effectiveness claimed in [1]. However, we also observe the phenomenon that the training times of FishNets are not as short as those of MLP and CNN. And with sufficient study upon this, we attempt to explain it in the way that the computations for matrix operations are less efficient than that for kernels.

V. DISCUSSION & CONCLUSION

In this project we reconstructed the basic building blocks and network structure of FishiNet, performed necessary modifications, conducted experiments under different model capacity and compared the performance of FishNet with other popular models on simple image classification tasks. Here we try to gain some insights on the model structure and experiments conducted.

A. Choice of Output Layer

A possible explanation of why the original convolutional output layer performs so badly in our experiments is the insufficient number of output channels. The original network pools from a layer with size (7, 7, 1600) and convolves it to the final 1000 output channels, which is large enough to extract meaningful information for predictions. However, when we implemented FishNet55 on Cifar10, the capacity of the output layer was dramatically reduced to (7, 7, 200) to 10 final outputs. We believe this will potentially result in failure of convergence in the final layer, while the details are to be further studied.

B. Impact of Model Capacity

Upon experiments conducted in section four, it could be concluded that for complicated tasks such as Cifar-10, FishNet models with bigger capacity could produce better accuracy in classification tasks.

C. Advantages

It can be observed from our experiments that FishNet surpasses other widely used shallow or deep neural networks in validation accuracy and computational cost, supporting the effectiveness claimed in [1]. However, we also found out that the training times of FishNets are longer than those of MLP and CNN. And with sufficient study upon this, we have attempted to explain it, but further details still need to be specified.

References

[1] Sun, S., Pang, J., Shi, J., Yi, S., & Ouyang, W. (2019). Fishnet: A versatile backbone for image, region, and pixel level prediction.

[2] D.O. Hebb(1949) The Organization of Behavior: A Neuropsychological Theory.

[3] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review, 65(6), 386–408. https://doi.org/10.1037/h0042519.

[4] A Krizhevsky, I Sutskever, GE Hinton. (2012) Imagenet classification with deep convolutional neural networks - Advances in neural information processing systems.

[5] He K., Zhang X., Ren S., Sun J. (2016) Identity Mappings in Deep Residual Networks. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9908. Springer, Cham. https://doi-org.ezproxy.cul.columbia.edu/10.1007/978-3-319-46493-0_38.

[6] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. Science, 313(5786), 504–507. http://www.jstor.org/stable/3846811.

[7] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Contributions:

Member	Major Contributions
Chengbo Zang	FishNet model construction and adjustments.
Yuqing Cao	Image preprocessing, experiments design and analysis.
Fengyang Shang	Model training and parameter tuning.
All	Code debugging, validations and final report.