Deep Learning Assignment 2

1st Ming-Xuan Wu
Institute of Data Science
National Cheng Kung University
Tainan, Taiwan
Email: RE6124019@gs.ncku.edu.tw

***Abstract*—This study uses the mini-ImageNet dataset for image classification, focusing on two main objectives. The first is to design a convolution module that can handle variable input channels while maintaining spatial size invariance, thereby improving model performance and flexibility over different channel combinations. The second goal is to create a two- to four-layer network, whether a CNN, Transformer, or RNN, that achieves 90% of the performance of ResNet34 on the ImageNet-mini dataset, with no more than a 10% loss in performance. To achieve these goals, we implement dynamic convolution for the first objective, which allows the model to adapt to different input channels and increases flexibility. For the second objective, we use a channel-wise attention mechanism to outperform ResNet34 with a shallower network. Our proposed methods significantly improve the handling of variable input channels and achieve high classification performance with fewer layers. Experimental results validate the effectiveness of our approaches and highlight their potential for practical applications in image classification.**

***Index Terms*—Deep Learning, Image Classification, Dynamic Convolution, Attention mechanism.**

I. Introduction

Since the advent of deep learning, it has become an integral part of many fields, particularly computer vision and speech recognition, and forms the backbone of many leading systems. Experiments on standard benchmark datasets such as TIMIT for speech recognition and ImageNet and CIFAR-10 for image recognition have shown that deep learning significantly improves recognition accuracy, underscoring its critical importance in advancing these technologies. However, most conventional CNNs are designed for fixed input channels, which limits their flexibility in practical applications where input channels may vary. To address this limitation, this study draws on the concept of dynamic convolution to design a unique convolutional network at the input layer of the network architecture, enabling seamless training with inputs from different channels.

Since 2012, many groundbreaking neural networks such as AlexNet, VGG, and ResNet have been proposed, significantly advancing the field of deep learning. More recently, approaches such as Vision Transformers (ViT), inspired by the success of transformers in natural language processing, have also shown remarkable performance improvements. In particular, ResNet addressed the problem of network overfitting through its residual learning framework, allowing the training of very deep networks. However, the increased depth of these networks inevitably leads to a significant increase in training parameters and computation time. To address this issue, the current task aims to achieve performance beyond that of ResNet34 using a shallow network. This study proposes to use an attention mechanism inspired by the ghost module of GhostNet to surpass the performance of ResNet34 with a more efficient and less complex network architecture.

II. Related Work

A. Dynamic Convolution

Dynamic convolution has emerged as a powerful technique for generating adaptive convolution kernels based on input features. Wu et al. (2020) introduced a dynamic convolution method that employs attention mechanisms over convolution kernels to enhance feature extraction. This approach allows the model to dynamically adjust its filters in response to different input features, thereby improving its ability to capture relevant patterns in the data. [1].

B. ResNet

The introduction of Residual Networks (ResNet) by He et al. (2015) constituted a pivotal advancement in the evolution of deep learning architectures for image recognition. ResNet addressed the issue of vanishing gradients by incorporating residual connections, paving the way for the training of considerably deeper networks [2].

C. MobileNet

MobileNets, as proposed by Howard et al. (2017), are designed for the efficient execution of deep neural networks on mobile and embedded devices. By utilizing depthwise separable convolutions, MobileNets significantly reduce the number of parameters and computational cost, making them suitable for resource-constrained environments [3].

D. GhostNet

In their 2020 paper, Han et al. introduce a novel approach to reducing the computational cost of neural networks by generating more feature maps from inexpensive operations. This is achieved by employing a series of linear transformations and inexpensive operations to produce additional feature maps, thus maintaining the representational power while reducing the complexity. [4].

E. Attention Mechanisms

The concept of attention mechanisms, popularized by the seminal work "Attention Is All You Need" by Vaswani et al. (2023), has revolutionized various fields, including natural language processing and computer vision. Attention mechanisms enable models to focus on relevant parts of the input sequence, thereby improving the efficiency and accuracy of feature extraction. [5].

III. Methodology

This section outlines the methodologies employed to achieve the objectives of the study. The methodologies are divided into two main tasks, each of which focuses on a specific aspect of the research.

A. Convolution Module for Variable Input Channels

The initial objective is to devise a convolutional module that can accommodate variable input channels while maintaining spatial size invariance. This design is intended to enhance the model's flexibility and performance across different channel combinations.

1) DynamicConv2D Implementation: The DynamicConv2D class represents a custom PyTorch module for dynamic convolution operations. It adjusts convolution kernel weights based on the input channels, thereby enabling it to handle varying input channel numbers. The initialization method defines the weights and biases of the convolution kernels. During forward propagation, the module dynamically adjusts the weights according to the current number of input channels. For instance, if the input has two channels, it randomly selects and updates weights for those two channels. This operation is particularly useful in scenarios with varying input channels. For further details, please refer to Algorithm 1.

Algorithm 1 Dynamic Convolution Operation

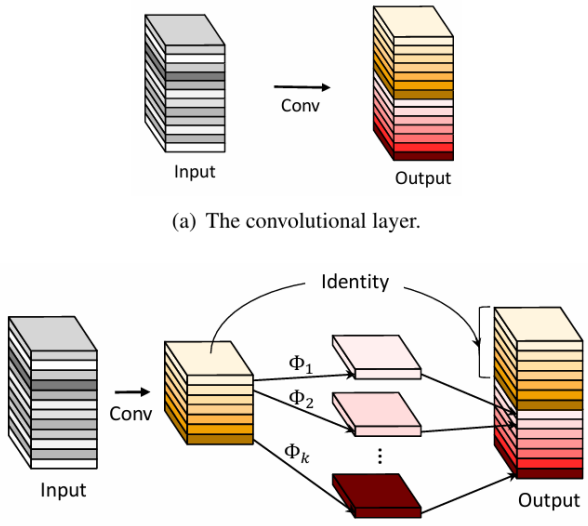
Input: Input tensor x of shape (N, C_{in}, H, W)
Input: Weight tensor W of shape $(C_{\text{out}}, C_{\text{in}}, K, K)$
Input: Bias vector $\mathbf{b}$ of length C_{out}
Input: Stride stride and padding padding
Output: Output tensor y of shape $(N, C_{\text{out}}, H_{\text{out}}, W_{\text{out}})$
$C_{\text{current}} \leftarrow$ Number of input channels in x
$\mathcal{C} \leftarrow$ Randomly select C_{current} channels from $\{1, 2, \ldots, C_{\text{in}}\}$
$W_{\text{selected}} \leftarrow W[:, \mathcal{C}, :, :]$
$y \leftarrow \text{Conv2D}(x, W_{\text{selected}}, \mathbf{b}, \text{stride}, \text{padding})$
Return y

2) Integration with MobileNetV3: To illustrate the practicality of the DynamicConv2D module, we modified the input convolutional layer of MobileNetV3 to utilize DynamicConv2D, thereby enabling it to accommodate various input channel sizes. The model was initialized with pre-trained weights from MobileNetV3.

B. Two-Layer Network for Image Classification

The second objective of this study is to develop a shallow neural network for image classification. The methodology employed is based on the GhostNet approach, which incorporates the ghost module, ghost bottleneck, and attention mechanisms to enhance the efficiency and performance of the network.

1) Ghost Module: The Ghost module represents a novel convolution module that is designed to produce a greater number of feature maps with a reduced number of parameters. This module divides a conventional convolutional layer into two distinct parts. The initial part employs standard convolution to generate a limited set of intrinsic feature maps, which are subsequently augmented by a series of inexpensive linear operations to generate additional feature maps, referred to as "ghost features." This approach results in a reduction in the number of parameters and computational complexity while maintaining the size of the output feature maps. Figure 1 illustrates the structure of the Ghost module.



(a) The convolutional layer.

Fig. 1: The Ghost module

2) Ghost Bottleneck: The Ghost Bottleneck is a bespoke bottleneck structure designed for use with lightweight convolutional neural networks (CNNs). This bottleneck structure is comprised primarily of two stacked Ghost Modules. The first Ghost Module serves as an expansion layer, increasing the number of channels. The second Ghost Module reduces the number of channels, matching the shortcut path. Shortcut connections are established between the input and output of the two Ghost Modules. Batch Normalization (BN) and ReLU activation functions are applied after each layer, with the exception of the second Ghost Module, where ReLU is not utilized. Figure 2 illustrates the structure of the Ghost Bottleneck.

3) Network Design: In order to achieve the objective, a two-layer neural network was designed which incorporates the aforementioned Ghost Module and Ghost Bottleneck. This design not only efficiently utilises computational resources but also maintains high classification performance.

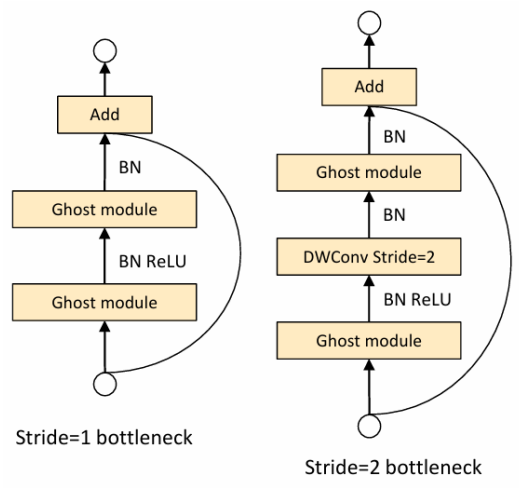


Fig. 2: The Ghost module

IV. Experiments

The experimental results of this study will include the performance of the training, validation, and testing phases, which will be evaluated using various metrics and visualized through a confusion matrix. The final results will be presented using 10 epochs. This choice was made because, after experimenting with 30 and 50 epochs, overfitting phenomena were observed, as evidenced in Figure 3.

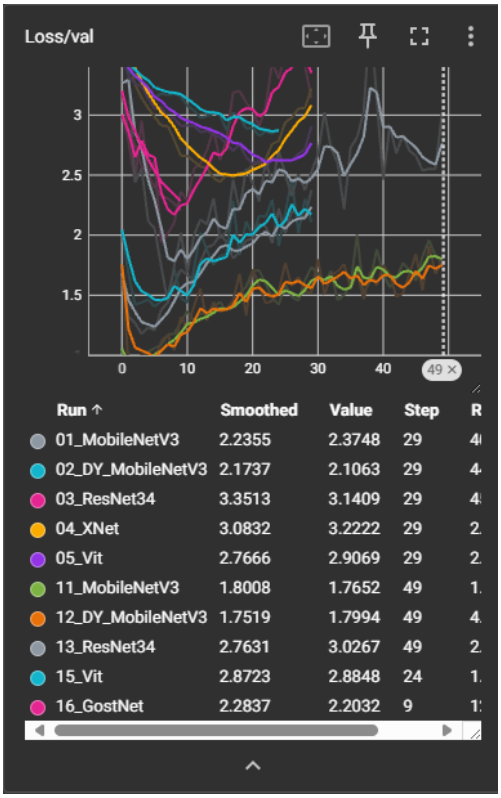


Fig. 3: Overfitting observed in train and validation phases

A. Dynamic Convolution

The initial convolutional layer of MobileNet was replaced with a DynamicConv2D module to enable inputs with varying numbers of channels. This allows the model to handle inputs with different channel configurations, including RGB, RG, RB, GB, R, G, and B. The sole purpose of using DynamicConv2D is to accommodate different input channels and ensure proper training without the use of additional techniques to enhance performance. Consequently, the computational complexity is increased due to the consideration of different channel inputs.

A series of experiments were conducted to evaluate the accuracy, precision, recall, F1 score, and inference time of DY_MobileNet with different input configurations. Additionally, the performance of DY_MobileNet was compared with that of the original MobileNet model. All experiments were conducted with a fixed image size of 128x128, and the models were trained for 10 epochs.

The experiments presented in Table I demonstrate that the performance of the model with three channels is superior to that of the model with two or one channel, as anticipated. The results for the two-channel combinations are nearly identical, as are those for single channels. A comparison of MobileNet and DY_MobileNet reveals that while DY_MobileNet effectively handles various input channels, it incurs higher computational costs due to increased FLOPS from diverse channel input considerations. Notwithstanding, DY_MobileNet's adaptability in accommodating disparate input channels without the necessity for additional techniques serves to illustrate its versatility for a multitude of applications.

B. Two-Layer Network

The objective of this experiment was to design a shallow neural network, referred to as GhostNet (Figure 4), for image classification. The goal was to achieve at least 90% of the performance of ResNet34. To enhance the network's efficiency and performance, we employed GhostNet's methodology, which incorporates the Ghost Module and Ghost Bottleneck with attention mechanisms to optimize resource utilization.

A series of experiments were conducted to evaluate the accuracy, precision, recall, F1 score, and inference time of the two-layer network, comparing it to ResNet34. All experiments utilized a fixed image size of 128x128, and the models were trained for 30 epochs with image augmentations to ensure robustness.

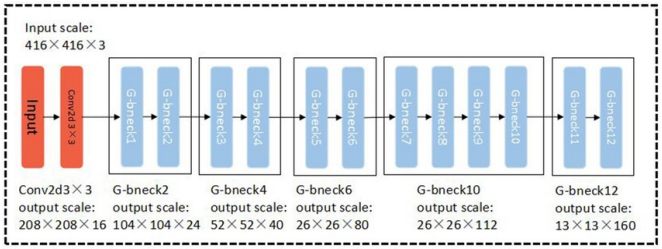


Fig. 4: GhostNet network structure

TABLE I: Comparison of MobileNet and DY_MobileNet Performance with Different Input Channels

Model	**MobileNet**	**DY_MobileNet**						
Channel	**RGB**	**RGB**	**RG**	**RB**	**GB**	**R**	**G**	**B**
Accuracy	0.6533	0.6378	0.5844	0.6222	0.5978	0.4244	0.4444	0.4489
Precision	0.6869	0.6578	0.6273	0.6480	0.6297	0.5129	0.5224	0.5298
Recall	0.6533	0.6378	0.5844	0.6222	0.5978	0.4244	0.4444	0.4489
F1 Score	0.6537	0.6331	0.5806	0.6142	0.5851	0.4073	0.4335	0.4256
Inference Time (s)	0.0083	0.0192	0.0060	0.0070	0.0065	0.0080	0.0085	0.0080
GFLOPS	17.43	58.33						
Parameters (M)	2.43	1.50						

TABLE II: **Performance Comparison of ResNet34 and GhostNet**

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Inference Time (s)**	**GFLOPS**	**Parameters (M)**
ResNet34	0.4000	0.5037	0.4000	0.3966	0.0060	960.86	21.21
GhostNet	0.3489	0.3658	0.3489	0.3406	0.0135	0.5129	0.5224
ResNet34-Pretrain	0.6467	0.6767	0.6467	0.6478	0.0060	960.86	21.21
GhostNet-Pretrain	**0.7489**	**0.7664**	**0.7489**	**0.7413**	0.0130	41.90	3.97

The experiments presented in Table II demonstrate several key insights. Firstly, ResNet34, which employs a deeper network architecture, generally achieves robust performance metrics. However, GhostNet, which employs a shallower architecture, achieves comparable results, particularly when pretrained. This highlights the efficiency of GhostNet in utilizing fewer layers to obtain similar performance levels. Without pretraining, GhostNet's performance does not lag behind ResNet34 by more than 10%. This indicates that even without extensive training, GhostNet's shallow architecture with attention mechanisms can approximate the performance of deeper networks. Furthermore, the significant improvement in GhostNet-Pretrain's metrics, surpassing ResNet34-Pretrain, underscores the effectiveness of incorporating attention mechanisms. Pretraining enables GhostNet to fully leverage its efficient architecture, resulting in higher accuracy, precision, recall, and F1 scores compared to ResNet34. Despite GhostNet's higher inference time due to its complexity in handling diverse input channels, its flexibility and enhanced performance metrics in the pretrained scenario validate the benefits of using a shallower network with attention mechanisms. This finding suggests that shallow networks, such as GhostNet, when combined with attention mechanisms and pretraining, can be highly effective for image classification tasks. They may therefore represent a more resource-efficient alternative to deeper networks, such as ResNet34.

V. Conclusion

In this study, we designed and evaluated two novel approaches for enhancing image classification performance on the ImageNet-mini dataset. Our first approach involved developing a convolution module capable of handling variable input channels while maintaining spatial size invariance. By integrating dynamic convolution (DynamicConv2D) into the input layer of MobileNetV3, we demonstrated the module's ability to adapt to different input channel configurations. The experimental results demonstrated that while DY_MobileNet incurred higher computational costs due to its flexibility, it successfully accommodated diverse input channels without additional techniques, highlighting its versatility for various practical applications.Second approach aimed to create a shallow neural network that could achieve at least 90% of the performance of ResNet34 with fewer layers. We employed the GhostNet methodology, incorporating Ghost Modules, Ghost Bottlenecks, and attention mechanisms. The results demonstrated that GhostNet, particularly when pretrained, not only achieved comparable performance to ResNet34 but also surpassed it in certain metrics. This highlights the efficiency of GhostNet's architecture in utilizing fewer parameters and computational resources while maintaining high classification accuracy.

References

[1] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11030–11039, 2020.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[3] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[4] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1580–1589, 2020.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.