DEEP LEARNING ASSIGNMENT 2

Pei-Chun Chen

Department of Statistics, National Cheng Kung University
r26121073gs.ncku.edu.tw

ABSTRACT

This project is centered on developing a unique convolutional module designed to maintain consistent spatial dimensions while accommodating various input channel sizes. The emphasis is on enhancing the module's ability to handle diverse input configurations effectively. The goal is to create a CNN with two to four layers that performs comparably to the ResNet34 model on the ImageNet-mini dataset, thereby simplifying the handling of different input types and setting a new standard for efficiency in convolutional network design.

Keywords- ImageNet-mini dataset, Dynamic Convolution, Advanced Residual Networks, Self-Attention Mechanisms.

1. TASK1:DESIGNING A CONVOLUTION MODULE FOR VARIABLE INPUT CHANNELS

1.1. Description

"Design a special convolutional module that is spatial size invariant and can handle an arbitrary number of input channels. You only need to design this special module, not every layer of the CNN. After designing, explain the design principles, references, additional costs (such as FLOPS or #PARAMS), and compare with naive models. To simulate the practicality of your method, use the ImageNet-mini dataset for training, and during the inference process, test images with various channel combinations (such as RGB, RG, GB, R, G, B, etc.) and compare the performance."

1.2. Dataset and Preprocessing

The CustomImageNetDataset class was designed to facilitate the loading and preprocessing of ImageNet images, tailored to our requirements. Images are dynamically resized and cropped to a uniform dimension (96x96 pixels) to ensure consistency in input data size. This standardization is crucial for maintaining the efficiency of the convolution operations across differently sourced images.

1.3. Network Architecture

The cornerstone of our approach is the DynamicConv2D module, which innovatively adjusts convolutional weights based on the number of input channels. This module replaces the first layer of a traditional ResNet18 model, providing the network with the capability to process images with varying channel dimensions without loss of information fidelity. The network concludes with a fully connected layer, resized to output 50 classes, aligning with the ImageNet-mini dataset specifications.

1.4. Data Augmentation

To simulate real-world scenarios where images may come with different channel compositions, the SelectiveChannelDrop augmentation method was employed. This method randomly omits or blackens certain image channels (R, G, B), forcing the network to learn from incomplete data and thus, enhancing its generalization capabilities. Other augmentations include random resized cropping, flipping, rotation, and perspective distortions, each adding variability to the training process and further robustifying the model against overfitting.

1.5. Training

Training was conducted over 30 epochs with a batch size of 64, using SGD as the optimizer. Learning rates were adjusted at predetermined epochs (50%, 75%, 85% of the total epochs) to refine the learning process as the training progressed. These strategic adjustments were crucial for the convergence of the model on the more challenging, variably augmented dataset.

1.6. Results and Discussion

The following results demonstrate the performance of ResNet18 with its first layer modified to DynamicConv2D, tested with 1, 2, and 3 input channels(RGB, RG, GB, R, G, B) . Metrics including accuracy, precision, recall, F1-score, and computational efficiency (FLOPS) are presented in table 1. As expected, a higher number of channels yields better performance.

Next, comparing the modified ResNet18 with DynamicConv2D in the first layer against the original ResNet18 (as shown in Table 2), it is evident that the dynamic convolution design enhances the model's ability to learn more effective feature representations. This experiment confirms that higher overall performance can be achieved, underscoring the importance of specific adaptations for particular datasets.

Table 1: Performance and Computational Capacity Across Various Input Channel Configurations

RGB	RG R	RB G	GB B
Accuracy 57.11%	44% 17.56%	40.22% 15.56%	37.78% 12.67%
Precision 58.52%	51.4% 22.61%	46.48% 16.53%	44.57% 14.57%
Recall 57.11%	44% 17.56%	40.22% 15.56%	37.78% 12.67%
f1-score 56%	42.5% 15.33%	38.41% 12.3%	35.56% 10.32%
FLOPS 333585408	326360064 319134720	326360064 319134720	326360064 319134720

Table 2: Comparative F1-Score of Three Classification Models Utilizing Four Feature Extraction Methods

	ResNet18(DtnamicConv2D)	ResNet18
Accuracy	57.11%	51.11%
Precision	58.52%	52.57%
Recall	57.11%	51.11%
f1-score	56%	49.78%
FLOPS	333585408	333585408

2. TASK2:DESIGNING A TWO-LAYER NETWORK FOR IMAGE CLASSIFICATION

2.1. Description

"Design a (2-4)-layer CNN, Transformer, or RNN network that can achieve 90% performance of ResNet34 on ImageNet-mini (i.e., with no more than 10% performance loss). There are no restrictions on parameter count or FLOPS, but the maximum number of input and output layers is limited to 4-6. Explain the design principles, references, and provide experimental results. We suggest you DO NOT use pre-trained models for ResNet34."

2.2. Network Architecture:
Detailed Explanation of the Three-Layer Design

The EnhancedResidualAttentionModel is architecturally crafted to optimize image processing through a structured three-layer design, each serving a critical function in the model's ability to enhance feature recognition and classification accuracy. Here's how each layer contributes to the model:

Initial Convolutional Layer: This first layer serves as the entry point for raw image data, preparing the features for deeper analysis:

- **Convolutional Layer (conv1):**Equipped with a 7x7 kernel, this layer captures initial spatial hierarchies and patterns such as edges and textures by processing raw pixel data. The stride of 2 and padding of 3 help in maintaining a balance between feature resolution and computational efficiency.
- **Batch Normalization and ReLU:**Stability and non-linearity are introduced immediately after initial convolution, helping in normalizing the data flow for the subsequent layers and adding non-linearity to enable the capture of complex patterns.

Residual Block with Attention Module: Residual Block with Attention Module

- **Residual Block (residual_block1):**Implements the basic building block of a ResNet, enhancing the model's ability to learn from data without being hindered by the vanishing gradient problem. This block adapts the features for deeper attention processing.
- **Attention Module (attention_module1):**Focuses on crucial features by applying a trainable attention mechanism. This module assesses the importance of different features and dynamically adjusts the neural focus, enhancing important features while diminishing less relevant ones.

Advanced Processing and Output Layer: These modules are placed strategically to The final layer aggregates and classifies the information:

- **Second Residual Block (residual_block2):**Processes the attended features further, increasing the abstraction level. This block prepares the features for final classification, ensuring that the most relevant information is emphasized.
- **Pooling and Output:**
  - **Adaptive Average Pooling (mpool2):**Reduces the spatial dimensions to a single vector per feature map, summarizing the essential information in a compact form.

– **Fully Connected Layer (fc):**Transforms the pooled features into final class predictions, directly mapping the deep network features to the output classes.

The architecture culminates in a sequence of batch normalization, ReLU activation, and adaptive average pooling, leading up to a fully connected layer that classifies the images into 50 categories based on the learned features.

2.3. Comparative Framework

We compare the EnhancedResidualAttentionModel directly with the standard ResNet34. Both models are trained under identical conditions on the ImageNet-mini dataset, using cross-entropy loss, stochastic gradient descent (SGD) with momentum, and scheduled learning rate adjustments.

2.4. Results and Discussion

As illustrated in Table 3, the EnhancedResidualAttentionModel that I have designed outperforms the non-pretrained ResNet34. A critical aspect of the design is the integration of an attention mechanism, specifically self-attention, which is pivotal in image processing tasks. Self-attention allows the model to focus more effectively on the important regions of an image, which is particularly beneficial in handling visual tasks.

Additionally, the architecture considers a balance between network depth and complexity. The moderate depth and enhanced feature extraction strategy employed in my design appear to be key contributors to its superior performance. This optimized combination facilitates a more effective learning process, enabling better generalization and recognition accuracy in complex visual tasks.

This demonstrates the importance of tailored architectural enhancements in improving model efficacy, particularly when addressing specific challenges inherent in image classification.

Table 3: Comparative F1-Score of Three Classification Models Utilizing Four Feature Extraction Methods

	3-layer ResAttentionNet	ResNet34
Accuracy	62%	59.78%
Precision	62.73%	59.99%
Recall	62%	59.78%
f1-score	61.23%	58.98%
FLOPS	309248256	612876800

3. VISUALIZATION

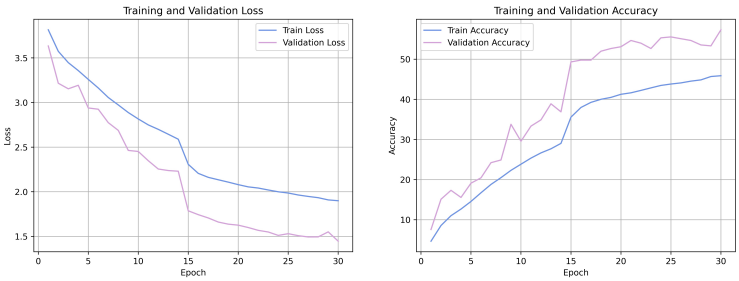

Fig. 1: ResNet18(DynamicConv2D): loss, accuracy curve

Fig. 2: ResNet18: loss, accuracy curve

Fig. 3: 3-layer ResAttentionNet: loss, accuracy curve

Fig. 4: ResNet34: loss, accuracy curve