DiCLAM for CIFAR-10 Image Classification: Multi-Fusion Deep-Learning Neural Network

1. Introduction and Overview

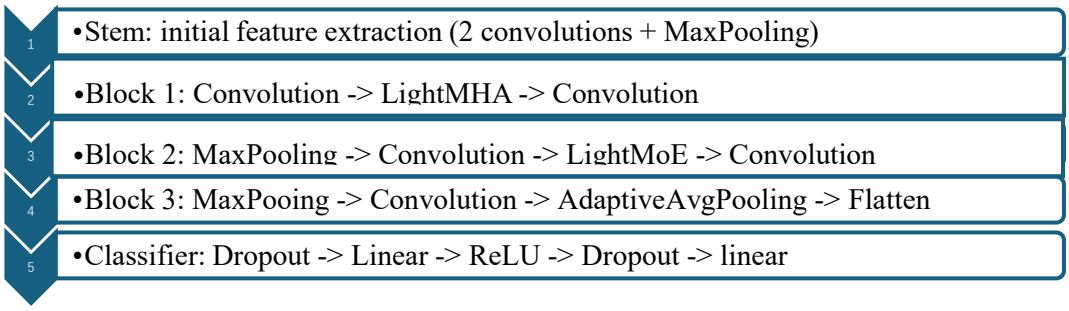
This report proposes a Distilled Convolutional Neural Network (CNN) framework with Light-Multi-Head Attention (LightMHA) and Light-Mixture of Experts (LightMoE) (DiCLAM) for the CIFAR-10 image classification, which requires an efficient and accurate model constraining its running time within 10 minutes. And the final performance of DiCLAM is inspiring, the total prediction accuracy is 91.36%, and the three typical benchmarks of Machine Learning (ML) model performance – precision, recall and f1-score are 0.9137, 0.9136 and 0.9136 respectively. The seamless integration of these deep learning blocks possibly illustrates the feasibility and high performance of light-distilled models in classification as well as the potential for practical application value in daily production and life.

2. Dataset

The CIFAR-10 dataset is a traditional dataset and benchmark for image classification tasks, containing 60,000 32x32 color images across 10 classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck). This report deploys 50,000 images for training and 10,000 images for testing.

3. Methodology

This report introduces a distilled CNN major architecture, which consists of three layers: a stem layer for initial feature extraction; a convolutional layer containing three convolutional blocks for feature refinement and enhancement and a fully-connected-layers-made classifier for final classification. Particularly, in the first block of the convolutional layer, this report innovatively applies a LightMHA for attention mechanisms-based feature refinement, but different from the traditional MHA, LightMHA replaces the fully connected layers with the simple 1×1 convolution to generate the query (q), key (k), value (v) these three matrices but still with residual and normalization layers. And this distillation could prominently decline computational complexity while maintaining the advantages of MHA. Another highlight is the LightMoE block, for the same reason before, the MoE block also has simplified. In this report, the number of expert networks is reduced to two, which is comprised of two 1×1 convolution and one Rectified Linear Unit (ReLU). Besides, in the gating network, average mean pooling as well as 1×1 convolution are implemented to produce gating weights. (see Figure 1)



Moreover, in the pre-processing stage, this report adopts the most popular pre-process method, random flipping, cropping and normalization to achieve data augmentation for better training. In the training stage, the author draws on the Pareto Rule (80/20), splitting 40,000 images and 10,000 images for cross validation, then applying the best hyperparameters to final 50,000 training and 10,000 testing. Additionally, to address common issues during training, this model not only imports a time supervisor to stop the training when close to time limits or beyond loss convergence but also deploys cross-entropy, gradient clipping, AdamW and OneCycleLR to avoid overfitting related

difficulties. Finally, the authors use confusion matrix, loss curve over epoch and training curve to visualize the whole training and testing process, then final prediction results are stored as a CSV file to check.

4. Results and Analysis *(The model evaluates validation after every three epoch)*

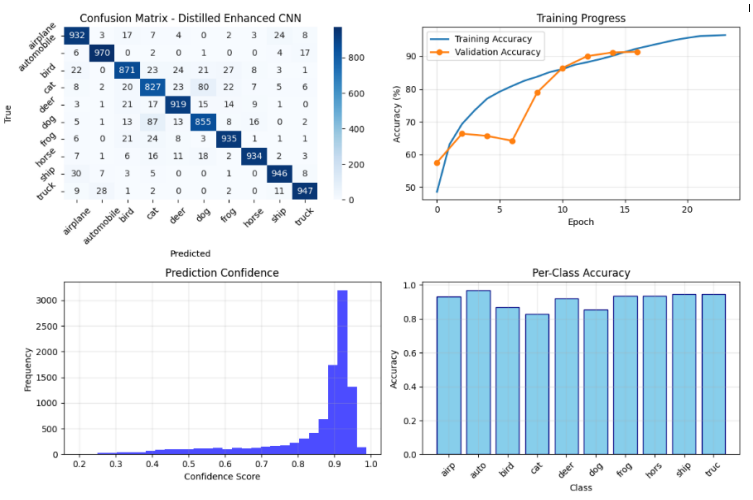


Fig 2 confusion matrix, loss curve over epoch, training curve

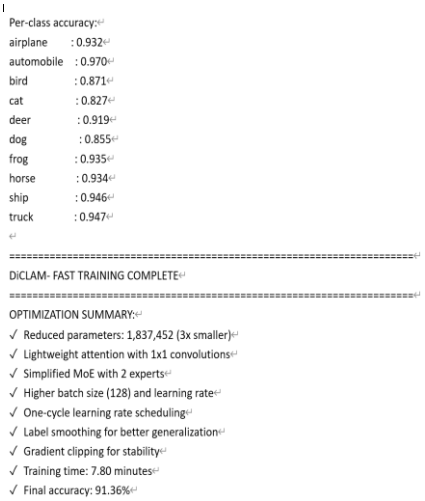


Figure 3 accuracy, training time, precision, recall and f1-score

DiCLAM has proven to be remarkably effective on the CIFAR-10 dataset, reaching an impressive test accuracy of 91.36% in just 7.80 minutes of training, also, three evaluation benchmarks, precision, recall and f1-score are 0.9137, 0.9136 and 0.9136 respectively. DiCLAM's efficiency is revealed in varied aspects, including a significant reduction in parameter:1,837,452 vs 5,512,356 in full model, and the application of advanced techniques like LightMHA and LightMoE block. (see Figure 2 and Figure 3)

Despite these strengths, however, the model shows signs of difficulty in accurately classifying visually similar classes such as 'cat' and 'dog', as evidenced by the confusion matrix and the relatively lower per-class accuracy for these categories. This might be triggered by less specialized processes and tuning of the 'cat' and 'dog' images both pre-processing and training (pre and post) stage.

5. Conclusion and Recommendations

The DiCLAM model from this report successfully achieves the classification and time limits on CIFAR-10 dataset, which are 91.36% accuracy rate in 7.80 minutes. Nevertheless, given that the certain limitation of the DiCLAM, several strategies could be conducted in future work: augmenting the training data to introduce more variability, meticulously tuning hyperparameters to optimize learning dynamics, experimenting with LSTM, GRU, finetuned pre-trained model or even self-supervised learning to capture more nuanced patterns and save computational resource while remaining highly accurate model performance. These enhancements are likely to promote the model's accuracy and generalization capabilities, especially for classes that present greater challenges.

Data Availability

Data will be made available on request.