Final CMPSC 448
Work by: Anik Shikapuri
PSU ID: 983401307
PSU email: aps6773

Convolutional Neural Network (CNN) & Transformer-Based System Image Classification:

CNN Report:

Task and Data Preprocessing:
The primary task was to classify images from the CIFAR-10 dataset using a Convolutional Neural Network (CNN). This dataset includes 60,000 32x32 color images across 10 different classes. The preprocessing involved normalizing the images to ensure consistency in the data format. Each color channel in the images was standardized to have a mean of 0.5 and a standard deviation of 0.5. This normalization is crucial as it simplifies the model's learning process, making it easier for the CNN to extract and learn features from the images.

Implementation & Architectures of Learning Systems:
The CNN was implemented using the PyTorch framework, known for its flexibility and efficiency in building deep learning models. The architecture consisted of two convolutional layers with ReLU activation functions and max pooling, followed by three fully connected layers. The convolutional layers are key in feature extraction, while the fully connected layers interpret these features to classify the images. The model was structured to effectively handle the complexity of the image data and was designed to generalize well on unseen data.

Training Details Systems:
The training of the CNN involved several carefully chosen hyperparameters. It was trained for five epochs to balance learning and overfitting, with a batch size of 4 to process a small set of images at each step, and a learning rate of 0.001 to ensure steady model convergence. The training loop included a forward pass for predictions, a backward pass using backpropagation for updating the model based on the loss calculated via the Cross-Entropy Loss function, and optimization using Stochastic Gradient Descent (SGD). Progress was monitored at regular intervals to adjust the training process if necessary.

Results, Observations & Conclusions:
Post-training, the model's performance was evaluated on a separate test set from the CIFAR-10 dataset. The observed results showed the model's capability to classify images with a considerable degree of accuracy. The accuracy measurements, both overall and for individual classes, indicated that the CNN had successfully learned to distinguish between different image categories. The conclusion drawn from this project is that the CNN, with its specific architecture

and training regimen, was effective in addressing the image classification task, demonstrating the practical utility of such models in real-world applications. However, the results were not extremely accurate and had a network accuracy of 51.3%. On the other hand, it also only took around 10 minutes to execute.

Challenges/Obstacles and Solutions:
One major challenge was avoiding overfitting while ensuring sufficient learning. To tackle this, a balanced number of training epochs and a moderate learning rate were chosen. Another challenge was ensuring efficient training with large image datasets. Utilizing CUDA-enabled GPUs significantly sped up the training process. There were also considerations regarding the choice of architecture; ensuring the CNN was neither too simple nor too complex for the task at hand. This was addressed by experimenting with different layers and parameters, ultimately settling on an architecture that provided a good balance between complexity and performance.

Transformer-Based System (ViT) Report:

Task and Data Preprocessing:
The task remains the classification of images from the CIFAR-10 dataset. In a Transformer-based approach, particularly with Vision Transformers, the data preprocessing includes an additional step of splitting images into patches. These patches are then linearly embedded, akin to tokens in NLP tasks, to be processed by the Transformer.

Implementation and Architectures of Learning Systems:
ViT, a relatively recent innovation in computer vision, marks a significant departure from conventional convolutional approaches. It applies the Transformer architecture, originally designed for NLP tasks, directly to sequences of image patches. In my implementation, the ViT model processes these image patches through self-attention mechanisms. This approach enables the model to weigh different parts of the image differently, thereby capturing a more global context. This is in contrast to the localized receptive fields typical of CNNs. By focusing on the relationships between various patches of an image, ViT can offer a more comprehensive understanding of the visual data, potentially leading to more nuanced and accurate image classification results.

Training Details: The training process for a ViT would be similar in terms of utilizing an optimizer and a loss function. However, due to the different nature of the architecture, the training might involve tuning specific parameters unique to Transformers, such as attention heads and the size of the feed-forward networks within each Transformer block. The computational demand could also be different, often requiring more resources compared to traditional CNNs. I experienced a much longer training time with

the Transformer Based System. It takes around 4-5 hours to run. This can be due to hyperparameter tuning, specifically my need to increase the number of workers.

Results, Observations, and Conclusions:
The transformer based system took much longer to run. This was because of some slight changes in the hyperparameter tuning. It was also because in order to implement ViT I needed to resize the images and normalize them which took longer since it was a large dataset. But, the results of this model were much stronger with the network accuracy being 89.61%.

Challenges and Solutions:
A significant challenge in implementing ViT models is their reliance on large amounts of data and computational resources. Unlike CNNs, which can capture local features effectively with fewer parameters, ViTs often require extensive training datasets to generalize well. This issue can be mitigated by using pre-trained models or techniques like data augmentation. Another challenge is the interpretability of ViT models, as the workings of self-attention mechanisms are less intuitive than the filters in CNNs. My biggest challenge was dealing with the time it takes to train the model. To see my result faster I would run my code on the cluster through Penn State Roar Collab. This was much faster than using my personal machine.

Conclusion:

The CNN model, characterized by its custom architecture, represents a more traditional approach in the field of image processing and machine learning. Its structure, comprising a series of convolutional layers followed by pooling and fully connected layers, is tailored for feature extraction and classification from image data. However, the simplicity of this model, while beneficial for educational and computational simplicity, becomes a limitation when dealing with the diverse and complex nature of the CIFAR10 dataset. The model's performance, with an accuracy of around 45%, reflects this limitation. Its inability to resize or augment input data further restricts its capacity to capture and learn from the finer nuances present in the dataset.

In contrast, the Transformer-based system model, exemplified by the pre-trained ResNet18, showcases the advancements and sophistication in modern deep learning techniques. The use of a pre-trained network provides a significant advantage, leveraging a vast repository of learned features and patterns that can be fine-tuned to specific tasks like CIFAR10 classification. This approach not only enhances the model's ability to discern and classify complex image data but also reflects in its substantially higher accuracy of 89%. Furthermore, the preprocessing step of resizing images to

128x128 pixels allows the network to examine images at a resolution that reveals more detailed features, thus contributing to its superior performance.

This comparative analysis underlines the transformative impact of advanced neural network architectures and pre-training in machine learning. While the CNN model offers a foundational understanding of image classification, the Transformer-based system model, with its pre-trained network and sophisticated data processing, sets a new benchmark in terms of accuracy and efficiency. The evolution from a basic CNN to a Transformer-based model encapsulates the rapid progression in the field of artificial intelligence, heralding a new era where the depth of learning and the complexity of models redefine the boundaries of computational capability and application.