Executive Summary Raport 2

Mogage Nicolae – ICA 246/1

Title: Analysis of Machine Learning Models for Image Classification on Cifar-10

This report provides an analysis of various machine learning models applied to the CIFAR-10 dataset, a widely recognized benchmark in computer vision research. The main goal was to evaluate the efficiency of 5 models, Convolutional Neural Networks (CNN), Deep Residual Learning Networks (ResNet), Densely Connected Convolutional Networks (DenseNet), Vision Transformers (ViT), and Convolutional Vision Transformers (CvT), in performing image classification. The CIFAR-10 dataset, consisting of 60,000 32x32-pixel color images divided into 10 mutually exclusive classes, was chosen for its balanced complexity and accessibility. It is small enough to be processed on standard hardware while still presenting challenges, particularly with visually similar categories like trucks versus automobiles or cats versus dogs.

The study explored the performance of both models trained from scratch and those pretrained. The transfer learning impact was most seen for ResNet and DenseNet, which were pretrained on the ImageNet dataset. This approach significantly improved their accuracy, demonstrating the value of pretraining for achieving better results on small-scale datasets. Among all the models, the pretrained ResNet-50 was found as the best-performing method, achieving an accuracy of 86.17%. This performance was able due to ResNet’s innovative use of residual connections, which address the degradation problem, found common in deep neural networks, by allowing layers to bypass information to deeper layers. Next, DenseNet obtained an 84.87% accuracy with its pretrained version. The dense connectivity of DenseNet facilitated the reuse of features and efficient gradient flow, enhancing its computational efficiency and performance. Traditional CNNs, despite their relatively simpler architecture, demonstrated impressive performance, with an outstanding value of 83.71% for the same metric. While CNNs lack the architectural complexity of ResNet and DenseNet, their ability to handle spatial hierarchies remains a strength, particularly when computational resources are limited. On the other hand, transformer-based models such as ViT struggled to deliver comparable results. The ViT model only scored 56.59% in terms of exactness, primarily due to the small size and low resolution of the CIFAR-10 dataset. Vision Transformers rely on global context learning through self-attention mechanisms, which requires large-scale datasets for effective pretraining. This became evident in the relatively poor performance of ViT, highlighting its limitations for small datasets when it is not pretrained enough. However, CvT, a hybrid model that combines convolutional operations with transformer mechanisms, outperformed ViT with over 10% improvement, having a 68.70% accuracy. CvT’s hierarchical structure and ability to capture both local spatial and global contextual features made it a better fit for the dataset, although it still performed worse than pretrained convolutional models.

In conclusion, the study shows the importance of proper selection of architectures and transfer learning in order to make a more performant model. The CNN based models have showcased their potential on small datasets, while transformers-based methods require better pretraining on a large dataset in order to be applied to datasets like Cifar-10.