CS 512 Project Proposal: Implementation of DeiT - Data-Efficient Image Transformer

Name: Aravind Balaji Srinivasan
Student ID: [A20563386]

Name: Vignesh Ram Ramesh Kutti
Student ID: [A20548747]

Paper Information
Title: Training data-efficient image transformers & distillation through attention

Authors: Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, Hervé Jégou

Publication: International Conference on Machine Learning (ICML) in 2021.

Link: https://arxiv.org/pdf/2012.12877

Problem Statement
While transformers have shown tremendous success in natural language processing (NLP), their application to image classification has been data-intensive. Vision Transformers (ViT) need large-scale datasets (like ImageNet-21k) to achieve competitive performance, which is not feasible in many real-world scenarios. DeiT addresses this challenge by introducing a distillation mechanism that reduces the dependency on vast datasets, enabling efficient training even on smaller datasets. The project aims to implement DeiT and evaluate its performance using various datasets, comparing it to traditional convolutional neural networks (CNNs).

Approach
The project will involve:
1. Implementing the DeiT architecture using the PyTorch framework.
2. Training the model on image classification tasks using datasets like CIFAR-10, CIFAR-100, and ImageNet-1k.
3. Distillation Process: Using a CNN model (e.g., ResNet) as the teacher for knowledge distillation to guide the DeiT model during training.
4. Evaluating the Performance: The performance of DeiT will be compared to other models,

especially CNNs and ViTs, on classification accuracy, training time, and data efficiency.
5. Experimentation: We will test how different teacher models influence DeiT's performance and explore possible modifications to the distillation token to improve the efficiency further.

Data

We will be using a subset of the datasets due to the limited computational power and time.
The following datasets will be used for training and evaluation:
- CIFAR-10: A dataset of 60,000 32x32 color images in 10 classes.
- CIFAR-100: A dataset similar to CIFAR-10 but with 100 classes.

Link: https://www.cs.toronto.edu/~kriz/cifar.html
- ImageNet-1k: A large-scale dataset with 1,000 classes, commonly used for image classification benchmarks.

Link: https://www.image-net.org/download.php

References

1. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2020). Training data-efficient image transformers & distillation through attention. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2012.12877

2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2010.11929

3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1706.03762

Team Member Responsibilities

We will work together on all the tasks required to complete this project.

Aravind Balaji Srinivasan:

- Dataset analysis and training.
- Implementing Pytorch model.
- Comparison of results from the models.

Vignesh Ram Ramesh Kutti:

- Testing and Validation
- Implementation of Pytorch for the model.
- Reports and analysis of results.