Make ViT Faster with Multi-Exit Architecture

Hyogon Ryu, 20233477

Problem Definition

- Although ViT has achieved very high performance, it faces challenges in both training and inference due to the high computation cost of its structure.
- In the case of inference, this is not a significant issue for devices with high computation power, but it becomes a critical problem at the edge.
- Regarding training, as the size of ViT-based architectures continues to increase, efforts are being made to find more efficient ways to train them.

Goal

- Incorporating the concept of multi-exit, commonly used in knowledge distillation with CNN architectures, into Vision Transformers enables faster inference and efficient training.
- Multi-exit allows for early predictions and intermediate exit points in the network, improving both inference speed and training efficiency.

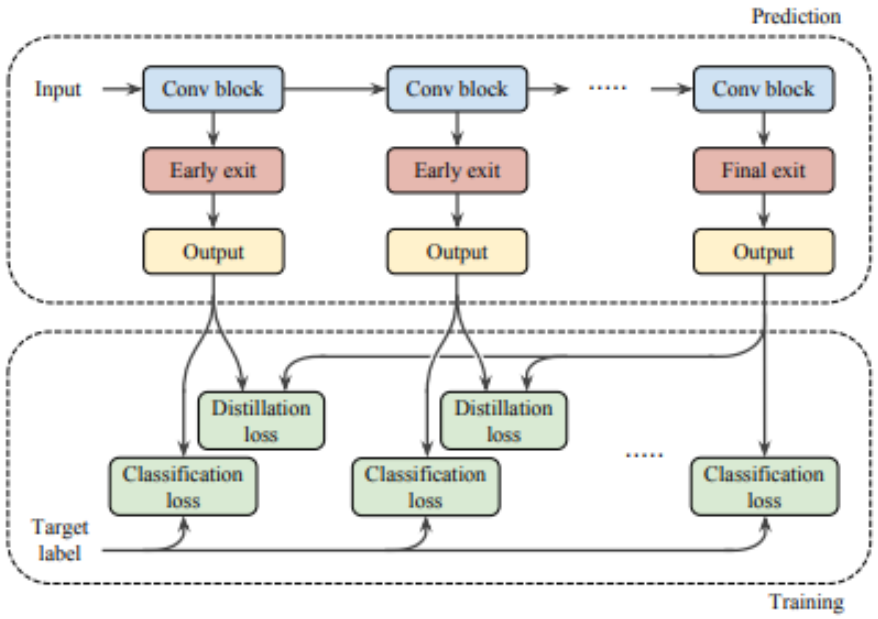

Figure 1: Illustration of the proposed method: distillation-based training (bottom) for a multi-exit a architecture (top).

Dataset

- CIFAR-10 / CIFAR-100 / ImageNet-1k Validation / and etc.

Evaluation Metric

- There are two considerable points in this project.
  1. Accuracy
  2. Efficiency
- Metric:
  - Accuracy: accuracy, top-k accuracy
  - Efficiency: flops, inference time(im/s), training time(epochs to converge), memory usage.

Other Important considerations

- Ablation study: Analysis of Attention map.
  - This project is not only for the engineering. I guess that multi-exit architecture will give the clue for understanding the attention mechanism and transformer architecture.
  - So I'll observe the attention map and if there are some important or surprise results on it, then I'll note that.