COSE474-2024F: Final Project Proposal
Optimizing MobileCLIP: Architectural Modifications for Efficient Image-Text Processing

Hyunjun Chun

1. Introduction

CLIP(Contrastive Language-Image Pretraining) is a pre-trained Text-Image model. It aligns the vector of image and text to the common latent space. Due to this feature of the model, CLIP is zero-shot. CLIP is now used in various field such as image-captioning and image-searching. However CLIP has clear difficulty to be deployed on mobile device due to large size and high latency. A few studies such as TinyCLIP, MiniCLIP, EfficientCLIP and MobileCLIP were conducted to solve the difficulty. In this study, I will explore the ways to make CLIP model available on mobile devices, by modifying encoder layer of MobileCLIP, which sets a new state-of-the-art latency-accuracy tradeoff for zero-shot classification.

2. Problem definition & challenges

CLIP is based on transformer architecture, which requires a lot of parameters. Furthermore, being a multi-modal model, CLIP should calculate both text and image inputs and map the vectors on to the common latent space, which results in high latency. Therefore there is a challenge to deploy CLIP on mobile devices.

3. Related Works

CLIP. Extract image vector and text vector with ViT and transformer each. Aligning image and text vector in latent space by contrastive learning. (Radford et al., 2021) Common architectures for accomplishing vision tasks are purely convolutional, transformer based and hybrid of convolution and transformer.

MobileCLIP. Multimodal reinforced training by using DataCompDR. Text-encoder is built with 1D Conv and Self-Attention. Image-encoder is built with Fast-ViT. MobileCLIP use convolution-transformer hybrid to build encoders, which requires smaller size than pure transformer even though the transformer is pruned.

4. Datasets

For evaluating Zero-shot Classification score, ImageNet-V2, ImageNet-A, ImageNet-O, ImageNet-R and ObjectNet are used. For training, use DataCompDR dataset, which also contribute to high performance.

5. State-of-the-art methods and baselines

MobileCLIP-S2 variant is 2.3× faster and more accurate compared to ViT-B/16 CLIP. (Vasu et al., 2024)

6. Schedule

1. Select Benchmark (11/1)
2. Implement text-encoder. (11/8)
3. Implement image-encoder. (11/15)
4. Train the model with contrastive learning. (11/22)
5. Evaluate the model with ImageNet datasets. (11/29)

References

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision.

Vasu, P. K. A., Pouransari, H., Faghri, F., Vemulapalli, R., and Tuzel, O. (2024). Mobileclip: Fast image-text models through multi-modal reinforced training.