CAP-6411 Computer Vision Systems Assignment 4- CLIP

Karthik Ramasamy
University of Central Florida
ka234388@ucf.edu

1 Introduction

In this assignment, I have fine-tune three models—CLIP, and SGLIP—on the Human Action Recognition (HAR) dataset using the same train/test split with 10 epochs. where the model CLIP leverages large-scale vision–language pre-training, while and SGLIP is designed for efficiency and trainned with extremely large batch sizes 32k. I have compared these models in terms of accuracy, training speed, and memory consumption. I have also tried A-CLIP and add their results where the core network is CLIP, with the saliency-based masking which resembles A-CLIP strategies.

2 Dataset and Training

The HAR dataset contains 15 classes (e.g., calling, cycling, eating, running), totaling 12.6k images with a fixed train/test split. Images were resized to 224 × 224 before input trained with 10 epochs also performed:
Augmentation: Standard image normalization and random crops/flips.
Loss: Cross-entropy.
Optimization: AdamW with cosine LR schedule.
Precision: Mixed precision training (AMP).
Batch sizes: CLIP, SgLIP

3 Results Comparison

3.1 Clip Fine-tuning

Table 1 shows the training and validation performance across 10 epochs. We observe steady improvements in both training and validation accuracy, with the best validation accuracy reaching at epoch 10.

Table 1: CLIP Fine-Tuning Results

Epoch	Train Acc	Val Acc	Time	Peak Mem
1	0.118	0.173	444.7	1940
5	0.250	0.264	389.9	1940
10	**0.377**	**0.322**	**385.0**	1940

Fine-tuning CLIP shows the loss is decreased and accuracy is improved steadily across epochs which is showing that CLIP's vision encoder is adaptable to downstream classification tasks.

3.2 SGLIP Fine-tuning

The core of our model is the pre-trained vision transformer from Hugging Face. This model is a Vision Transformer (ViT). The model was fine-tuned for 10 epochs using the following setup:

- **Loss Function:** Cross-Entropy loss, suitable for multi-class classification.
- **Optimizer:** AdamW with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4}.

3.2.1 Results

Table 2: SgLIP Full Fine-Tuning

Epoch	Train Acc	Val Acc	Time (s)	Peak Mem (MB)
1	0.615	0.827	284.5	3582.7
2	**0.881**	**0.848**	**283.9**	**3581.7**
5	0.952	0.827	283.3	3581.7
10	0.974	0.817	284.7	3581.7

The model demonstrated strong and rapid convergence, achieving its best validation accuracy of **84.76%** at epoch 2. The training process was stable, with consistent improvements in training accuracy and a steady decrease in loss over the 10 epochs. Table 2 summarizes the key performance metrics recorded during the training run. The peak memory usage remained consistent at approximately 3.5 GB, and the average time per epoch was around 284 seconds.

The results clearly indicate that the SgLIP model is a highly effective vision backbone. Its ability to achieve **high accuracy** after only two epochs of fine-tuning highlights the quality of its pre-trained representations and its adaptability to downstream classification tasks like human action recognition.

3.3 A-CLIP Fine-Tuning

To further improve performance, I have experimented with A-CLIP, a variant of CLIP and dual loss objectives (contrastive + classification).

- **Optimizer:** AdamW, LR 1×10^{-4}, weight decay 1×10^{-4}
- **Batch Size:** Effective BS: 4096 (micro 64 $\times$ accum 64)

Table 3 summarizes the performance across 10 epochs. We observe rapid convergence: validation accuracy rose

from 24.7% in epoch 1 to 79.3% by epoch 10. Contrastive and classification losses decreased in parallel, demonstrating effective joint optimization for this model.

Table 3: A-CLIP Fine-Tuning Results

Epoch	Train Acc	Val Acc	Time (s)	Peak Mem (MB)
1	0.130	0.248	69.0	**4859**
5	0.628	0.594	69.0	3531
10	**0.850**	**0.793**	**69.1**	3531

Compared to the standard CLIP baseline (best validation accuracy 32.1%), A-CLIP achieved a significant boost, reaching **79.3%** validation accuracy after 10 epochs. This highlights the benefit of using a larger embedding space and joint optimization with contrastive and classification losses. The results suggest that A-CLIP is better suited for small-scale human action recognition tasks where discriminative fine-grained features are required. **But this A-Clip Model is not available directly even though i tried to get it via the training pipeline applies an "A-CLIP-style" approach by masking low-saliency tokens in the image patches before classification. This masking method, where only the most salient tokens are kept, is inspired by works such as A-CLIP.**