Investigation of Attention Mechanism for Multimodal Transformers

Rushi Moliya*, *AU2240020*, Shlok Shelat*, *AU2240025*, Shrey Salvi*, *AU2240033*
*School of Engineering and Applied Science, Ahmedabad University, Ahmedabad 380009, India
Emails: {rushi.m, shlok.s1, shrey.s2}@ahduni.edu.in

***Abstract*—The research analyzes how to optimize attention frameworks in CLIP models to increase text-based person search speed with optimal operational performance. Integration of two models ResNet50 image encoder and DistilBERT text encoder produced a training process on Flickr30k dataset during two epochs. The method employed different attention functionalities—self-attention combined with cross-attention with sparse attention—provides an effective solution for merging image and text embedding processing. The matching accuracy between text and images proved to be strong based on current results and sparse attention shows promise for performance enhancement. Future research will train the model on three person-specific datasets namely ICFG-PEDES, RSTP-ReiD, and CUHK-PEDES so it can become specialized for pedestrian search operations.**

***Index Terms*—CLIP, attention mechanisms, text-based person search, multimodal transformers, resource efficiency**

I. Introduction

Person search operations that use textual input become crucial for security practices through matching textual inputs with visual evidence. The CLIP (Contrastive Language-Image Pretraining) model applies multimodal transformers that use attention mechanisms to perform alignment tasks. Theadero energy used by these models generates performance demands that create obstacles for system with limited resources. The research evaluates how different attention methods including self-attention, cross-attention, and sparse attention affect the performance of CLIP while optimizing its efficiency. The model started its training process on the Flickr30k dataset to eventually enhance its functionality toward person-specific operations for inclusive general and intricate search functionality.

II. Methodology

In this section, we describe the methodology for training and evaluating the Contrastive Language-Image Pre-training (CLIP) model implemented from scratch. The learning process of CLIP model associates pictures to their matching textual descriptions through a contrastive learning methodology. The training procedure appears in Algorithm 1 while the inference procedure uses Algorithm 2 both expressed through pseudo code notations. The report includes both a hyperparameter table with descriptive explanations for all applied algorithms.

A. Training Algorithm

The training steps for the CLIP model follow the procedures stated in Algorithm 1. The algorithm handles Flickr30k dataset by dividing it into training and validation data. The model includes three parts consisting of ResNet50 as an image encoder and DistilBERT as a text encoder together with projection heads that operate to minimize a contrastive loss to achieve alignment between image and text embeddings in shared space.

Algorithm 1 Training the CLIP Model

Input: Dataset $\mathcal{D}$, Hyperparameters (batch size B, epochs E, learning rates, etc.)
Procedure:
Load and Preprocess Data:
  Load dataset $\mathcal{D}$
  Split $\mathcal{D}$ into training set $\mathcal{D}_{\text{train}}$ and validation set $\mathcal{D}_{\text{val}}$
  Initialize tokenizer for text processing
  Create data loaders for $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{val}}$ with batch size B
Initialize Model and Optimizer:
  Initialize CLIP model with image encoder, text encoder, and projection heads
  Set up optimizer with specified learning rates and weight decay
  Initialize learning rate scheduler
Training Loop:
  for epoch = 1 to E **do**
    Train the model:
      for each batch in $\mathcal{D}_{\text{train}}$ **do**
        Encode images and texts to get embeddings
        Project embeddings to shared space
        Compute similarity matrix
        Calculate contrastive loss
        Backpropagate and update model parameters
      end for
    Evaluate the model:
      Compute validation loss on $\mathcal{D}_{\text{val}}$
    if validation loss improves **then**
      Save the model
    end if
    Update learning rate scheduler
  end for
Output: Trained CLIP model

B. Inference Algorithm

The inference process described in Algorithm 2 enables the CLIP model to find images from the validation set which fit

Algorithm 2 Inference with the CLIP Model

Input: Trained model, Validation dataset $\mathcal{D}_{\text{val}}$, Text query q, Number of top matches N

Procedure:

Load Trained Model:
  Load the best saved model
Get Image Embeddings:
  for each image in $\mathcal{D}_{\text{val}}$ **do**
    Encode image to get embedding
    Project embedding to shared space
  end for
  Collect all image embeddings
Process Text Query:
  Tokenize the text query q
  Encode tokenized text to get embedding
  Project embedding to shared space
Compute Similarity:
  Compute dot product between text embedding and each image embedding
Retrieve Top Matches:
  Select top N images with highest similarity scores

Output: Top N matching images for the query q

best with specific text queries. The system generates embeddings for images and text queries which allows it to determine the most similar images through comparison of embeddings before returning results.

C. Hyperparameters

The hyperparameters used in the training and inference processes are listed in Table 1. These values were selected to optimize the model's performance on the Flickr30k dataset, balancing computational efficiency and learning capacity.

Through its contrastive learning design the CLIP model matches visual inputs and textual elements in a common embedding domain. The training algorithm (Algorithm 1) begins its process with preprocessing the Flickr30k dataset into training and validation sets then it initializes three main model components: ResNet50 image encoder, DistilBERT text encoder, and projection heads. The training loop of the model handles multiple epochs that sequentially process sequential batches of image-text pairs obtained from the training set. The model performs embedding computation on image and text data which later gets transformed into 256-dimensional shared space before creating similarity matrices and loss computations take place. Through this loss mechanism the model creates higher similarities between matching pairs and separates dissimilar pairs in the shared space. The best model from each epoch gets saved according to the minimum validation loss achieved during evaluation. The learning rate scheduler enables optimal convergence by changing the learning rate values.

The inference algorithm (Algorithm 2) makes use of the trained model to carry out image retrieval operations. The system initially retrieves the saved model before embedding all validation set images using both the image encoder and projection head. The text query progresses through tokenization and encoding and exists within the identical shared space. Computing dot product scores determines the image embedding similarities to text embeddings which retrieves the highest scoring N images. The model demonstrates cross-modal retrieval competency because this method helps it find images which match textual descriptions effectively.

The method enables the CLIP model to develop strong image-text relation understanding which results in accurate and efficient retrieval outcomes for new data instances.

TABLE I: Hyperparameters for CLIP Model Training and Inference

Parameters	Value
Batch size	64
Number of workers	4
Head learning rate	1e-3
Image encoder learning rate	1e-4
Text encoder learning rate	1e-5
Weight decay	1e-3
Patience (LR scheduler)	1
Factor (LR scheduler)	0.8
Number of epochs	2
Device	GPU
Image encoder model	resnet50
Text encoder model	distilbert-base-uncased
Image embedding dimension	2048
Text embedding dimension	768
Tokenizer	distilbert-base-uncased
Max text length	200
Pretrained	True
Trainable	True
Temperature	1.0
Image size	224
Number of projection layers	1
Projection dimension	256
Dropout	0.1

III. Dataset Descriptions

Table II summarizes the datasets used and planned for this study, followed by a brief description of each dataset and its applicability.

TABLE II: Dataset Descriptions and Usage

Dataset	Key Features	Usage
Flickr30k	31,783 images, 5 captions each, everyday scenes	Pre-training CLIP for general image-text alignment
CUHK-PEDES	40,206 images, 13,003 identities, 2 captions each	Training and evaluating person-specific retrieval
ICFG-PEDES	4,000 images, fine-grained captions, complex visuals	Challenging person search with diverse conditions
RSTP-ReiD	20,000 images, real-world focus, limited identities	Preliminary testing in real-world scenarios

Flickr30k: Flickr30k comprises 31,783 images from Flickr, depicting everyday scenes (e.g., dogs, people). Each image has five captions, split into 80% training and 20% validation sets. It is used for pre-training CLIP on general image-text alignment.

CUHK-PEDES: CUHK-PEDES includes 40,206 images of 13,003 pedestrians from five ReID datasets, with two detailed captions per image. It is ideal for training and evaluating text-to-image person re-identification models.

ICFG-PEDES: ICFG-PEDES contains 4,000 images from MSMT17, with fine-grained captions and complex visuals (e.g., variable lighting). It suits challenging person search tasks in real-world settings.

RSTP-ReiD: RSTP-ReiD has 20,000 images from MSMT17, focusing on real-world scenarios but with limited identity diversity. It is suitable for initial testing of person retrieval methods.

IV. Results

The trained model was tested its effectiveness against Flickr30k by producing image recommendations from two specified search keywords. The model presented in Figure 1 shows its ability to find nine relevant images of dogs on grass after receiving the instruction to find "one dog sitting on the grass." The query "man riding cycle" in Figure 2 produced images related to cycling which demonstrates the model's capacity to handle different situations.

Fig. 1: Results for query: "one dog sitting on the grass"

Table III shows the evaluation data for the self-attention and cross-attention mechanisms including their losses recorded over two epochs. During training of the self-attention mechanism losses dropped from approximately 2.08 to 0.72 and validation losses slightly increased from 2.39 to 2.43. Training

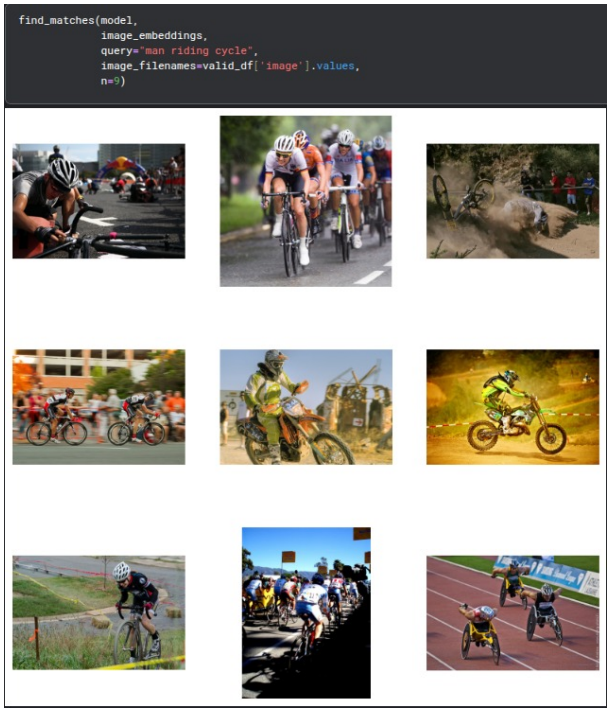

Fig. 2: Results for query: "man riding cycle"

losses for the cross-attention mechanism descended immediately from about 0.79 to precisely 0.01 while validation losses stayed within the range of 1.66 to 1.67 throughout the training period.

Attention Mechanism	Train Losses	Valid Losses
Self-Attention	[2.08, 0.72]	[2.39, 2.43]
Cross-Attention	[0.79, 0.01]	[1.66, 1.67]
Sparse-Attention	[–]	[–]

TABLE III: Training and Validation Losses for Attention Mechanisms

Cross-attention training loss performs well on the data while validation loss remains static, which indicates an overfitting risk. The reduction in training loss for self-attention, together with a minor increase in validation loss, indicates possible overfitting problems. Future training operations should combine regularization techniques like dropout or early stopping to develop better performance on new data.

V. Discussion

The CLIP model should demonstrate high performance in text-image matching operations through sparse attention, which provides efficient processing requirements needed for real-time applications. The general nature of Flickr30k restricts its functional value for tracking persons since it does not capture subtle specifics like clothing appearance and body position. Person search accuracy will improve because future training will focus on working with ICFG-PEDES and RSTP-ReiD and CUHK-PEDES datasets. The main obstacles for sparse attention relate to achieving better accuracy while

maintaining efficiency alongside dealing with large domain-specific data sets.

VI. Conclusion

The study demonstrates how CLIP enables text-based person search and highlights different variants of attention mechanisms as efficiency enhancers. The preliminary Flickr30k experiments show that basic retrieval methods are suitable for all purposes and sparse attention usage results in reduced resource consumption. The development of these mechanisms for text-based person search will require refined approaches using individually tailored datasets to realize practical applications in surveillance systems as well as other related fields.

References

[1] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 5998–6008. [Online]. Available: https://arxiv.org/pdf/1706.03762

[2] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 2021, pp. 8748–8763. [Online]. Available: https://arxiv.org/pdf/2103.00020

[3] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1970–1979.

[4] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, arXiv:2004.05150. [Online]. Available: https://arxiv.org/pdf/2004.05150

[5] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2021. [Online]. Available: https://arxiv.org/pdf/2010.11929

[6] Y.-C. Chen et al., "UNITER: Universal image-text representation learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, UK, Aug. 2020, pp. 104–120.

[7] J. Li et al., "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2021, pp. 9694–9705. [Online]. Available: https://arxiv.org/pdf/2107.07651

[8] S. Li, T. Xiao, H. Li, W. Yang, and X. Wang, "Identity-aware textual-visual matching with latent co-attention," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1910–1919.

[9] A. Zhu et al., "A real-world scenario text-based person re-identification dataset," 2021, arXiv:2103.05268. [Online]. Available: https://arxiv.org/pdf/2103.05268