COSE474-2024F: Final Project
Text-to-Image Retrieval Using Soft Prompt Tuning on CLIP

Author: Hania Akhlaqi, Korea University, hania.akhlaqi@gmail.com

1. Introduction

Text-to-image retrieval is a critical task in multimedia information retrieval, with applications in areas such as content-based search engines, digital asset management, and accessibility tools. Recent advances in multimodal models like OpenAI's Contrastive Language–Image Pretraining (CLIP) have significantly improved performance in these tasks by jointly learning text and image representations. However, while CLIP excels in zero-shot scenarios, its performance on domain-specific datasets remains limited due to a lack of specialized adaptation. The challenge lies in enhancing its performance for specific tasks without the computational overhead of fine-tuning the entire model.

Soft prompt tuning emerges as a promising solution, introducing learnable embeddings that adapt the pre-trained model to new tasks. Unlike traditional prompt engineering, which relies on hand-crafted prompts, soft prompts are trainable vectors that are seamlessly integrated into the input sequence. This approach retains the efficiency of pre-trained models while enhancing their domain-specific capabilities, making it particularly suitable for resource-constrained settings.

The problem addressed in this project is to improve the text-to-image retrieval performance of CLIP on the Flickr30K dataset by leveraging soft prompt tuning. Flickr30K, a widely used benchmark for multimodal retrieval tasks, includes captions that are rich in detail and complexity. These captions often require models to achieve fine-grained text-image alignment, meaning the ability to precisely associate specific textual descriptions with their corresponding visual elements in an image. While CLIP performs well in general retrieval tasks, its default configuration struggles with these nuanced relationships, such as recognizing and accurately aligning detailed attributes, actions, or spatial arrangements described in the captions. This project investigates whether soft prompt tuning can improve CLIP's ability to align text and image embeddings, enhancing its performance on such challenging tasks.

2. Methods

Soft prompt tuning is a novel adaptation technique that directly modifies the input space by introducing learnable embeddings, contrasting with conventional fine-tuning, which alters the model's weights. This lightweight adaptation strategy allows efficient and targeted improvements, especially in scenarios where computational resources are limited or full dataset training is impractical.

The significance of this project lies in its exploration of soft prompt tuning's applicability to text-to-image retrieval. Prior works, such as (Lester et al., 2021) Lester et al. (2021) on soft prompting in language models and (Radford et al., 2021)Radford et al. (2021) on CLIP's zero-shot capabilities, highlight the potential of pre-trained models and lightweight adaptation methods. However, their combination in the context of multimodal retrieval remains underexplored.

The main challenges in this project involve ensuring that the soft prompts are compatible with CLIP's existing architecture and finding the right balance between fine-tuning the model to be domain-specific while maintaining its ability to generalize to new, unseen data. Since CLIP relies on a contrastive learning approach, it requires precise adjustments of hyperparameters, such as temperature and batch size, to optimize the alignment between the textual and visual modalities. If these hyperparameters are not carefully tuned, the model may fail to learn an effective alignment between the text and image embeddings, resulting in poor retrieval performance.

In this project, I addressed these challenges by adapting soft prompts to fit seamlessly into CLIP's architecture without disrupting its pre-trained structure. I introduced learnable embeddings as soft prompts, allowing CLIP to refine its ability to match text and images without modifying the model's core weights. To ensure compatibility with CLIP's contrastive learning setup, I carefully tuned hyperparameters such as temperature and batch size, optimizing them through iterative experiments to achieve the best performance for text-to-image retrieval on the Flickr30K dataset. By using soft prompt tuning, I was able to make targeted improvements while preserving CLIP's ability to generalize across diverse queries and images.

Algorithm 1 Text-to-Image Retrieval Using Soft Prompt Tuning

Input: Pre-trained CLIP model, Dataset D (images I, captions T), Prompt length P, Epochs E, Learning rate LR
Output: Soft prompt-tuned model

1. Initialize CLIP model with frozen weights.
2. Create learnable prompt embeddings of shape $(P, \text{embedding\_dim})$.
3. **For** epoch in range(E):
   (a) **For** each batch in D:
      i. Tokenize captions T into text embeddings.
      ii. Preprocess images I into image embeddings.
      iii. Concatenate prompt embeddings with text embeddings.
      iv. Compute text and image features via CLIP.
      v. Calculate contrastive loss.
      vi. Backpropagate and update prompt embeddings.
4. Evaluate the tuned model on the test set.
5. Compute retrieval metrics (Accuracy, Recall@k).

As you can see in Algorithm 1, soft prompt embeddings were appended to CLIP's text encoder input. These embeddings were initialized randomly and trained using a contrastive loss function, which aims to maximize the similarity between paired text and image embeddings while minimizing the similarity for unpaired samples. The rest of the CLIP model was frozen to retain its pre-trained capabilities. The training process involved iteratively updating the soft prompts while evaluating performance on the Flickr30K validation set.

The system utilized the NVIDIA L4 GPU on Google Colab for training. Images were resized to 224×224, and text captions were tokenized using CLIP's pre-trained tokenizer. Hyperparameters, including the learning rate (1e-4) and batch size (32), were selected based on some initial experiments and iterations.

3. Experiments

3.1. Datasets and Preprocessing

The Flickr30K dataset, with 30,000 images and 5 captions per image, was used for evaluation. It poses significant challenges due to its diversity and the fine-grained nature of the captions. Standard preprocessing techniques were applied, including image resizing and normalization to match CLIP's input requirements. Text data was tokenized using CLIP's tokenizer, and padding ensured uniform input lengths.

3.2. Computing Resources

Google Colab with the following configuration was used:

- GPU: NVIDIA L4
- System RAM: 53 GB
- GPU RAM: 22.5 GB
- Disk Space: 112.6 GB

The training utilized the PyTorch framework with the Google Compute Engine backend, which allowed efficient data loading and model optimization.

3.3. Training and Evaluation Setup

The training process focused solely on optimizing the soft prompt embeddings, while the rest of the CLIP model remained frozen. The optimization objective was a contrastive loss function that encouraged high similarity between paired text and image embeddings. Evaluation metrics included accuracy and Recall@5, which measure the model's ability to retrieve relevant images for a given query.

3.4. Results and analysis

Table 1. Performance Comparison on Flickr30K (Text-to-Image Retrieval)

Method	Test Accuracy (%)	Recall@5
Baseline (CLIP Paper)	-	0.65
Soft Prompt Tuning	29.8	0.6773
SOTA (CLIP Weights)	-	80.7

The visual results demonstrate that our model is capable of effectively retrieving images that closely match the provided text descriptions. However, it faces limitations in more challenging cases.

Our tuned model achieved a Test Accuracy of 29.8% and a Recall@5 of 0.6773, improving upon CLIP's baseline Recall@5 of 0.65 reported in (Radford et al., 2021) However, it falls short of the SOTA Recall@5 of 80.7 achieved using publicly available CLIP weights (AndresPMD, 2021). This

gap reflects the limitations of soft prompt tuning, which, while computationally efficient, is less impactful than the more extensive fine-tuning and architectural enhancements used in SOTA approaches. Nonetheless, the results demonstrate the potential of soft prompt tuning to enhance CLIP's baseline performance for domain-specific tasks.

3.5. Discussion

The success of soft prompt tuning lies in its ability to achieve domain-specific adaptation with minimal computational costs. This method enables better alignment of text and image embeddings, which is particularly advantageous in resource-constrained scenarios. However, the results also highlight its limitations. The modest gains in metrics like Recall@5 compared to SOTA demonstrate that soft prompt tuning alone may not fully exploit CLIP's potential for complex datasets like Flickr30K.

Several factors contribute to these observations:

SOTA methods often incorporate task-specific layers or pretraining on augmented datasets, enabling more robust feature learning. Soft prompt tuning, in contrast, makes minimal changes to the input space.
The added non-linear probe in SOTA models enables them to capture intricate relationships between text and images more effectively than lightweight tuning approaches.
Optimization Challenges: Soft prompt tuning relies on tuning hyperparameters like learning rate, temperature, and batch size. While these were optimized in this project, they might still require further refinement to achieve closer performance to SOTA.
Overall, while soft prompt tuning is a promising adaptation technique, its performance in this study underscores the need for complementary strategies to achieve competitive results on benchmarks like Flickr30K. This finding suggests future directions for combining soft prompt tuning with more robust feature extraction methods or pretraining pipelines.

4. Future Direction

One key limitation of this project lies in the static selection of hyperparameters such as learning rate, batch size, and temperature for contrastive loss. These parameters were chosen manually, which may not have been optimal for fully leveraging the potential of soft prompt tuning. Future work could explore dynamic hyperparameter optimization techniques, such as Bayesian optimization or reinforcement learning-based tuning, to adaptively adjust these values during training. Additionally, methods like learning rate scheduling or adaptive temperature scaling could help the model achieve better convergence and improved modality alignment. Incorporating such dynamic strategies may enhance performance and narrow the gap between the current results and state-of-the-art performance.

References

AndresPMD. Clip based image text matching. https://github.com/AndresPMD/Clip_CMR, 2021. Accessed: 2024-12-10.

Lester, B., Al-Rfou, R., and Constant, N. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. URL https://arxiv.org/abs/2104.08691.

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. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. URL https://arxiv.org/abs/2103.00020.