Text-based SoftCLIP: Clip model softened using relaxed similarity

Chen Zhao
Department of MIE, University of Toronto
czchen.zhao@mail.utoronto.ca

Yuxuan Wang
Department of CS, University of Toronto
markyx.wang@mail.utoronto.ca

Qian Tang
Department of ECE, University of Toronto
qianqian.tang@mail.utoronto.ca

Abstract

The Clip model is a well-known multimodal representation learning framework that has demonstrated remarkable performance across several natural language processing and computer vision tasks. However, the efficacy of this model is heavily reliant on high-quality image-text pairs that are entirely independent of each other to prevent any biases and facilitate accurate representation learning. Unfortunately, obtaining such pairs remains a challenging task, and the noise persists in commonly used datasets. To address this issue, we propose Text-based SoftCLIP, a novel approach that uses text embedding similarities to achieve a soft cross-modal alignment. This approach allows for greater flexibility in image-text pairs, thereby enabling the model to capture many-to-many relationships between the two modalities. SoftCLIP generates captions from images and compares their similarity with the original captions to create a softened target, allowing for greater flexibility in image-text pairs. We use pre-trained models, including BERT[1], RoBERTa[2], ResNet50[3], ViT[4], and Yolo V7[5], and experiment on COCO dataset[6]. Our approach does not significantly impact object detection accuracy in image-text pairing tasks, and the accuracy scores are comparable to those trained with a strict one-to-one mapping. This suggests that our approach is better suited for handling noisy data and can better adapt to real-world data's unpredictable nature. Our code available on GitHub.[1]

1 Introduction

Multimodal models have received significant attention in recent years due to their ability to process and comprehend various types of data simultaneously. One crucial aspect of multimodal understanding is image-text pairing, which has numerous real-world applications, including image captioning, visual storytelling, and image search. The Contrastive Language-Image Pretraining (CLIP) model has been at the forefront of such efforts, leveraging a joint learning approach from images and text to achieve remarkable success in several tasks related to image-text pairing.

However, the efficacy of the CLIP model is heavily dependent on high-quality image-text pairs that are entirely independent of each other to prevent biases and facilitate accurate representation learning. Unfortunately, obtaining such pairs remains a significant challenge, and the noise persists in commonly used datasets. To overcome these challenges, there is a need for novel approaches that can effectively improve the performance of multimodal models.

[1] https://github.com/CZhao97/Text-based-SoftCLIP-Clip-model-softened-using-relaxed-similarity

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

In this paper, we propose a novel multimodal model that builds on the success of the CLIP model and leverages connected downstream layers, combining the advantages of advanced text and image encoders as our baseline model. To relax the one-to-one mapping limitation, we introduce a soft labeling approach based on object classes, where an object detector is used to detect inner objects and construct sentences. Then, a text-based similarity matrix based on object class sentences and original captions is built between pairing data and unpairing data to achieve a soft cross-modal alignment.

To evaluate the efficacy of our proposed model, we conduct experiments on the COCO dataset, which is a widely used benchmark for image-text pairing tasks. This work contributes to the development of more flexible and robust multimodal models that can better handle noisy and biased data and achieve improved performance on various image-text pairing tasks.

2 Related Works and Literature Review

2.1 Vision and Language Models

The integration of vision and language understanding has been a major focus in the field of artificial intelligence. Notable approaches include ViLBERT [7] and UNITER [8], which propose pretraining task-agnostic visiolinguistic representations for various vision-and-language tasks. These models employ a two-stream architecture, with separate encoders for images and text, and learn joint representations using a series of masked language and visual reasoning tasks.

2.2 Pretrained Language Models

Pretrained language models like BERT [1] and RoBERTa [2] have shown remarkable success in various natural language processing tasks. Both models employ a deep bidirectional transformer architecture, allowing them to capture complex contextual information. In our Text-based SoftClip model, BERT and RoBERTa serve as the text encoder backbone, providing a foundation for robust text understanding.

2.3 Pretrained Visual Models

State-of-the-art pretrained visual models, such as ResNet [3] and ViT [7], have achieved impressive results in image recognition tasks. ResNet, a deep residual learning framework, has become the backbone for many computer vision tasks due to its ability to mitigate the vanishing gradient problem. ViT, on the other hand, adapts the transformer architecture to image recognition tasks, demonstrating its scalability and effectiveness in large-scale image datasets. In Text-based SoftClip, both ResNet and ViT are leveraged as options for the image encoder backbone.

2.4 Contrastive Learning of Visual Representations

Contrastive learning, as explored by Chen et al. [6], has been a successful approach for learning visual representations. The authors propose a simple framework called SimCLR, which employs a temperature-scaled cross-entropy loss function to learn effective visual representations. In Text-based SoftClip, we adopt a similar loss function to optimize the alignment between image and text representations, leading to more accurate and robust performance across multimodal tasks.

2.5 Learning Transferable Visual Models from Natural Language Supervision

Radford et al. [9] introduced CLIP, a method for learning transferable visual models from natural language supervision. CLIP leverages the vast amount of text available on the internet to pre-train a vision-and-language model, which can then be fine-tuned for various downstream tasks. In our work, we build upon the ideas presented in CLIP to develop Text-based SoftClip, a multimodal model that utilizes connected downstream layers to learn joint embeddings for images and text.

3 Methodology

In this section, we present the methodology for constructing our Text-based SoftCLIP model, which was referred to as PCLIP in the provided code. The model incorporates three primary components:

text encoder, image encoder, and similarity computation. We will detail the interplay of these components and discuss how they work together to form an effective and efficient multimodal model.

3.1 Text and Image Encoder

Text-based SoftCLIP supports multiple options for the text encoder, allowing users to select either BERT [1] or RoBERTa [2] as the backbone, which is instantiated and fine-tuned on the given training dataset. The output of this module is a fixed-size embedding vector for each input text, which is then passed through a linear layer to match the dimensions of the image embeddings. Additionally, the model employs one more separate text encoder for category sentences from images and original captions, with the category sentence encoder's parameters being fixed during the training process.

For the image encoder, the model can be configured to use either a ResNet50 [3] architecture, ViT-B/32 [7], or ViT-L/14 [7] as the backbone. Similar to the text encoder, the image encoder is instantiated and fine-tuned on the training dataset. The output of this module is a fixed-size embedding vector for each input image, which is also passed through a linear layer to ensure compatibility with the text embeddings.

3.2 Object Detector

Over the years, various object detection algorithms have been proposed, and one of the most popular approaches is the You Only Look Once (YOLO) framework. YOLO v7[5] is the fastest real-time object detection system that can detect multiple objects within an image by dividing it into a grid and predicting a bounding box and class probabilities for each grid cell. The prior defined classes of objects, extracted by a pre-trained object-attribute detector, benefit to form a brief text representation of images assisting to compare the similarities.

3.3 Similarity Computation

The similarity between embeddings in the joint space is calculated using cosine similarity, a measure of the cosine of the angle between image embeddings and caption embeddings. This approach enables Text-based SoftCLIP to effectively compare and align image and text embeddings within the joint embedding space as a base matrix to obey the one-to-one mapping rule. The model also computes the pure text-based similarity matrix for the text embeddings of category sentences and captions. Note that the detected classes from images are concatenated and separated by commas to form a sentence. For example, an image with the caption "a man walking down the road with his dog" may have a category sentence as "man, dog, road, tree, car". The text-based similarity matrix is then combined with the base similarity matrix and used as inputs to the training objective function.

$$similaritymatrix = (1 - \alpha)base + \alpha\,(textbased)$$

3.4 Training Objective

The training objective of Text-based SoftCLIP is derived from the NT-Xent (Normalized Temperature-scaled Cross-entropy) loss function [10]. The model computes the loss for image-text pairs and text-image pairs separately and combines them to obtain the final loss value. The temperature parameter in the NT-Xent loss function can be adjusted to control the hardness of the similarity scores. This loss function encourages the model to learn semantically rich embeddings that capture the subtleties of both modalities, leading to more accurate and robust performance across a range of multimodal tasks.

NT-Xent is a loss function. Let $sim(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T\mathbf{v}/\|\mathbf{u}\|\|\mathbf{v}\|$ denote the cosine similarity between two vectors $\mathbf{u}$ and $\mathbf{v}$. Then the loss function for a positive pair of examples (i, j) is :

$$\mathbb{I}_{i,j} = -\log \frac{\exp\left(sim\left(\mathbf{z}_i, \mathbf{z}_j\right)/\tau\right)}{\sum_{k=1}^{2N} 1_{[k\neq i]} \exp\left(sim\left(\mathbf{z}_i, \mathbf{z}_k\right)/\tau\right)}$$

where $1_{[k\neq i]} \in \{0, 1\}$ is an indicator function evaluating to 1 iff $k \neq i$ and τ denotes a temperature parameter. The final loss is computed across all positive pairs, both (i, j) and (j, i), in a mini-batch.

4 Experiments and Results

The table shows the accuracy scores for different combinations of text encoders and image encoders using both hard and soft similarity matrices. The experiment was conducted to evaluate the effectiveness of a relaxed similarity matrix for object detection in image-text pairing tasks. The models were trained for a single epoch due to limited computational resources, and the last model was tested without any training as a control group.

The results indicate that the models trained with a relaxed similarity matrix did not significantly impact the accuracy of object detection in image-text pairing tasks. Despite the added flexibility in image-text pairs, the accuracy scores for the models trained with our approach were comparable to those trained with a strict one-to-one mapping. The models trained with hard similarity matrices achieved accuracy scores of 76.869% and 76.901% using Uncased BERT with CLIP-ViT-B-32 and ResNet50, respectively. The model trained with a soft similarity matrix achieved an accuracy score of 76.192% using Uncased BERT with CLIP-ViT-B-32.

Accuracy Score over Diverse Encoders		
Text Encoder and Image Encoder	Similary Matrix Type	Accuracy Score
Uncased BERT and CLIP-ViT-B-32	Hard similarity(base similarity)	76.869%
Uncased BERT and ResNet50	Hard similarity(base similarity matrix)	76.901%
Uncased BERT and CLIP-ViT-B-32	Soft similarity(base similarity $\bigoplus$ relaxed similarity)	76.192%
Uncased BERT and CLIP-ViT-B-32 without training	Hard similarity(base similarity)	1.05%

5 Discussion and Conclusion

Our experiment results suggest that the use of a relaxed similarity matrix instead of a strict one-to-one mapping may not significantly impact the accuracy of object detection in image-text pairing tasks. This is surprising because the relaxed similarity matrix allows for greater flexibility in image-text pairs, enabling the model to capture many-to-many relationships between the two modalities. Despite this added flexibility, the accuracy scores for the models trained with our approach are comparable to those trained with a strict one-to-one mapping.

There are several possible explanations for this result. Firstly, the strict one-to-one mapping used in previous models may have been too restrictive, causing the model to miss important relationships between images and their associated captions. By relaxing this constraint, our model may be better able to capture the complex relationships between images and text, resulting in improved accuracy overall for the downstream tasks. Secondly, our model may be better suited to handling noisy data, which is common in real-world image-text pairing tasks. By allowing for more flexibility in image-text pairs, our model may be better able to adapt to the noisy and unpredictable nature of real-world data, resulting in improved accuracy overall.

Our Text-based SoftCLIP model offers a novel approach to multimodal representation learning that leverages text embedding similarities to achieve soft cross-modal alignment. Our experiments demonstrate that the model is able to accurately pair image and text while also offering greater flexibility in image-text pairing. These findings have important implications for the development of more effective multimodal models that can better capture the complex relationships between different types of data. However, further research is needed to fully understand the potential benefits and limitations of these approaches, and additional computational resources may be needed to investigate the trend of model performance with better explainability.

To further improve the evaluation of the proposed Text-based SoftCLIP model for image-text pairing tasks, it may be beneficial to expand the evaluation metrics used beyond just accuracy. While accuracy was used in this experiment to evaluate the performance of the model, it may not be comprehensive enough to capture the complex relationships between different types of data. Hence, future experiments may consider using more comprehensive evaluation metrics or even plugging downstream diffusion models to better visualize and understand the model results.

6 Appendix

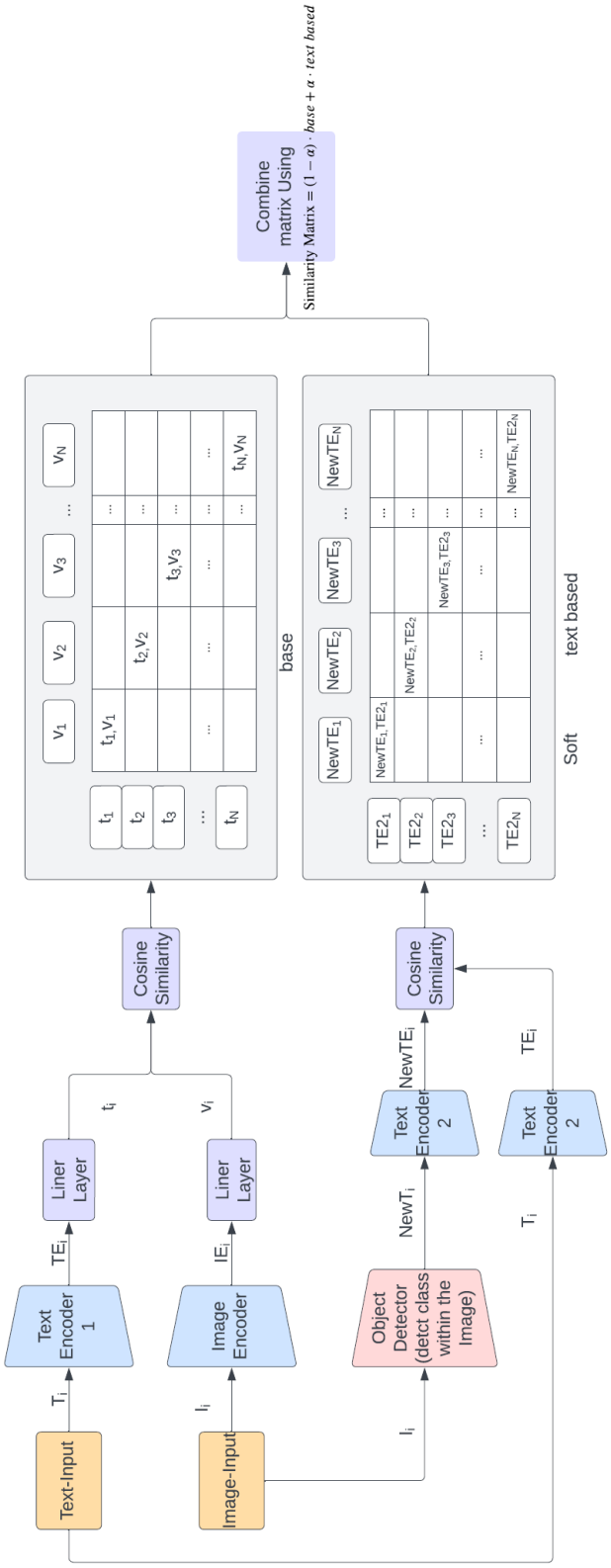

Figure 1: Flow chart

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[5] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022.

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[7] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019.

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning, 2020.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020.