IMPART: Image Matching and Pairing with Adapted RA-CLIP Technology

Rohith Vutukuru*† rvutukuru@ucsd.edu
Archana Pradeep*† arpradeep@ucsd.edu
Jalend Bantupalli*† jbantupalli@ucsd.edu
Veeramakali Vignesh Manivannan*† vmanivannan@ucsd.edu

† University of California San Diego

Abstract

Inspired by Retrieval Augmented Generation in large language models, we explored the concept of retrieval augmentation in vision-language models and implemented the RA-CLIP methodology(from scratch), a method aimed at reducing the pretraining costs of CLIP while enhancing its performance. By storing 10-15% of image-text pairs in a reference database, RA-CLIP enables efficient inference with reduced pretraining data. We also developed 'RA-CLIP Vanilla' from the idea of RA-CLIP, validated it through image classification and video retrieval tasks, and demonstrated its superior performance and cost efficiency. Our contributions include the RA-CLIP and RA-Clip Vanilla code, which we will publish for community use, and the novel application of RA-CLIP in accurate video retrieval, offering significant potential for real-life applications such as improved video search engines, content recommendation systems, and media archival processes.

1. Introduction

The development of vision-language models has revolutionized the field of artificial intelligence by enabling seamless interaction between visual and textual data. CLIP [7] (Contrastive Language-Image Pretraining), a vision-language transformer, exemplifies this progress by learning to associate images and text through training on a vast dataset of image-text pairs. However, the extensive data requirements and computational costs associated with pretraining CLIP pose significant challenges. Inspired by the concept of Retrieval Augmented Generation in large language models, we explored the potential of incorporating retrieval augmentation in vision-language models, leading us to the RA-CLIP framework.

RA-CLIP [8] (Retrieval Augmented Contrastive Learning Image Pretraining) addresses the pretraining cost issue of CLIP by utilizing a smaller, strategically selected dataset. This approach leverages a reference database containing 10-15% of the image-text pairs used during training. These pairs act as a cheat sheet, aiding the model during inference and significantly reducing the need for an extensive pretraining dataset. The methodology not only makes the training process more efficient but also enhances the model's performance, as demonstrated in the RA-CLIP paper.

Motivated by the promising results of RA-CLIP, we sought to replicate and extend this framework. Due to the unavailability of the original RA-CLIP code, we developed RA-CLIP Vanilla from scratch. This version adheres to the principles of RA-CLIP, ensuring comparable or superior performance to the original CLIP model. Our experiments focused on two primary downstream tasks: image classification, to validate our implementation against established benchmarks, and video retrieval, a less-explored but highly impactful application.

RA-CLIP Vanilla shows significant potential for practical applications. In the realm of video retrieval, our model can accurately locate relevant video clips from a database using either an image or text query. This capability is crucial for enhancing video search engines, optimizing content recommendation systems, and streamlining media archival processes. Furthermore, our open-source code contribution aims to facilitate further research and innovation within the community, providing a robust foundation for future advancements in vision-language models.

In summary, our work on RA-CLIP Vanilla underscores the effectiveness of retrieval augmentation in vision-language pretraining. By reducing data requirements and computational costs while maintaining high performance, RA-CLIP Vanilla stands as a viable alternative to traditional

* Equal Contribution

methods. The successful application of our model to diverse tasks highlights its versatility and the potential for broad real-world impact, paving the way for more efficient and powerful vision-language systems.

2. Related Work

The development of vision-language models and the integration of retrieval mechanisms have significantly advanced the capabilities of AI systems in understanding and interacting with multimodal data. This section reviews foundational works in the area, highlighting key developments and identifying how they pave the way for our contributions.

2.1. Vision-Language Models: CLIP

Contrastive Language-Image Pre-training (CLIP) introduced by Radford et al. [7] presents a new paradigm for visual representation learning. Unlike traditional visual recognition systems that recognize only the categories specified during training, CLIP utilizes natural language to enable flexible adaptation to new categories. By transforming category labels into text descriptions and processing these through its text encoder, CLIP can learn visual representations that generalize across a variety of tasks. Despite its impressive performance and ability to match the accuracy of an ImageNet-trained ResNet50, CLIP is substantially data-hungry, requiring tens of millions of image-text pairs for pre-training. Recent efforts have sought to reduce this extensive data requirement by leveraging additional supervisions from existing datasets to enhance the training process of CLIP [8].

2.2. Retrieval Augmented CLIP

Building upon the foundational ideas of CLIP, Retrieval Augmented CLIP (RA-CLIP) integrates a retrieval mechanism to enhance the model's training efficiency and effectiveness. This methodology involves constructing a reference set of image-text pairs, which acts as a supplemental resource during the model's inference stage. By retrieving relevant image-text pairs from this reference set, RA-CLIP can augment the original CLIP's image representation, enriching it with additional contextual information that assists in describing and understanding new visual content [8]. This approach not only addresses the high data demands of CLIP but also improves the model's zero-shot learning capabilities by providing more diverse and contextually relevant data during training.

2.3. Image to Video Retrieval

The adaptation of vision-language models for image to video retrieval is a burgeoning field that requires handling complex and dynamic visual information. The ability to retrieve video content based on static image inputs necessitates a model capable of understanding and interpreting visual continuity across frames. This task highlights the importance of robust and contextually aware feature extraction, which can benefit significantly from the enriched visual representations provided by retrieval-augmented models like RA-CLIP [6].

2.4. Text to Video Retrieval

Text to video retrieval extends the challenge by requiring the model to comprehend and translate textual descriptions into relevant video content. This task emphasizes the model's ability to bridge modalities, leveraging textual context to search and identify matching video sequences. The developments in multimodal learning, particularly the integration of coherent and consistent training methodologies, have shown significant potential in enhancing the performance of text to video retrieval systems, thereby making them more applicable and effective in practical scenarios [1].

3. Methodology

In our research, we explored the implementation and efficacy of Retrieval-Augmented CLIP (RA-CLIP) models, aiming to extend their application to efficient and practical use cases. This section describes the iterative development of two versions of RA-CLIP: the RA-CLIP Vanilla and the full RA-CLIP with the Retrieval Augmented Module (RAM).

3.1. RA-CLIP Vanilla

Our initial approach, RA-CLIP Vanilla, was designed as a simplified version of the RA-CLIP model where, instead of a complex RAM module, we utilized a weighted sum of image embeddings and text embeddings of the top k entries retrieved from the reference set. Specifically, the augmented image representation, v', is calculated as follows:

$$v' = v + \frac{w_{img}}{k} \sum_{j=1}^{k} r_{img} + \frac{w_{txt}}{k} \sum_{j=1}^{k} r_{txt} \tag{1}$$

where v is the original image representation, r_{img} and r_{txt} are the image and text representations retrieved from the reference set, and w_{img} and w_{txt} are the corresponding weights. This method allowed us to leverage the pre-trained capabilities of CLIP without additional training, facilitating zero-shot learning capabilities. The simplicity of RA-CLIP Vanilla made it particularly suitable for direct applications in downstream tasks such as image-to-video and text-to-video retrieval [8].

3.2. RA-CLIP with RAM Module

Building upon the vanilla model, we implemented the RAM module as described in the literature by Xie et al. [8]. The

RAM module aims to enrich the image representation by dynamically incorporating relevant image-text pairs from the reference set during the inference process. However, our attempt to train the RA-CLIP with the RAM module on the YFCC 15M dataset [3] faced significant challenges. Due to computational constraints, we could not fully realize the potential of the RAM-enhanced model, which reflected in suboptimal performance metrics compared to the theoretical expectations.

Computational Limitations: It is essential to note that our experiments were substantially constrained by limited computational resources. These limitations impacted our ability to train the RA-CLIP with the RAM module effectively, leading to a performance that did not fully exploit the model's capabilities. The insights gained from these experiments, however, are invaluable for understanding the scalability and resource requirements of retrieval-augmented vision-language models [8].

3.3. Application in Downstream Tasks

Despite the challenges faced during the training of the RA-CLIP with the RAM module, RA-CLIP Vanilla proved to be highly effective for our targeted applications. Utilizing its zero-shot learning capabilities, derived from the pre-trained CLIP model, we successfully applied RA-CLIP Vanilla in the downstream tasks of image-to-video and text-to-video retrieval. We employed the "Condensed Movies" dataset to evaluate the performance of RA-CLIP Vanilla in these retrieval tasks, providing a diverse and challenging environment for testing its efficacy. Each video in condensed frame has 25 fps and given the compute resources available, we used 4fps which produced around 800 frames per video. We collected a total of 115 videos from diverse movies. For representing the full video we used the mean of all the embeddings of the frames. We found that this mean is a good representation of the entire video. [2].

4. Results

The results are two-fold in our case, as we examined the performance of the models on two downstream tasks, Zero-Shot Classification and Image/Text to Video Retrieval.

4.1. Zero-Shot Classification Results

We had 1000 image-text pairs from the YFCC15M dataset as our reference set for both 'RA-CLIP' and 'RA-CLIP Vanilla'. We tested the models on CIFAR 10, CIFAR 100 [5] and CALTECH 101[4] datasets.

- **Normalized Embeddings are better**: In the RA-CLIP paper[8], it is mentioned that the embeddings we get from the RAM module are augmented directly to the image/text input embedding. We observed that if we normalize the RAM module embeddings before augmenting the input embeddings, we got much better classification results, as shown in Fig. 1

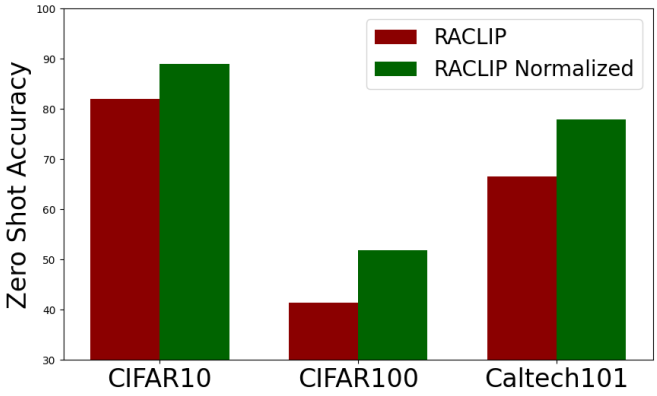

Figure 1. Accuracy of RA-CLIP vs Normalized RA-CLIP across datasets

- **RA-CLIP Vanilla works well**: From the Tab. 1, we can see that the RA-CLIP Vanilla performs as good as CLIP or sometimes better. This can be mainly attributed to the reference set that we have. Although the reference set is from the YFCC15M dataset and doesn't include any image-text pairs from the testing datasets, our RA-CLIP vanilla is able to enhance its visual interpretation of the inputs given for inference, by the reference set. This is useful in scenarios having constraints on computation.

  We can also see that RA-CLIP's low performance on CIFAR 100 and CALTECH 101, which can be majorly attributed to the low training, as mentioned in Sec. 3. We hope that we can increase it's performance and match RA-CLIP's zero shot classification accuracy as given in RA-CLIP paper.[8]

Model	CIFAR 10	CIFAR 100	CALTECH 101
CLIP	88.78	61.68	85.39
RA-CLIP	**88.98**	51.92	77.91
RA-CLIP Vanilla	88.17	**61.80**	**86.02**

Table 1. Comparison of Models' Zero-Shot Classification accuracy

4.2. Image/Text to Video Retrieval Results

We used the [2] dataset as the dataset for image/text-to-video retrieval and preprocessed the videos as discussed in Sec. 3. We used our 'RA-CLIP Vanilla' for this task. Although we didn't measure the performance of the model through standard evaluation metrics for this particular task, we tried out several image and text queries to retrieve

videos from the preprocessed database of 115 videos.

- **Image to Video Retrieval**: An example of an image to video retrieval is shown in Fig. 2 and we observed that it is retrieving the video which has the similar/exact frames as the top 1 retrieval, and all the other top retrievals are from the same movie, with this actor in most of the frames.

Figure 2. Query Image(from Google), and corresponding frames from the top retrieved video.

- **Text to Video Retrieval**: An example of a text-to-video retrieval is shown in Fig. 3 and we observed that all the top retrieved videos have the same person in many frames.

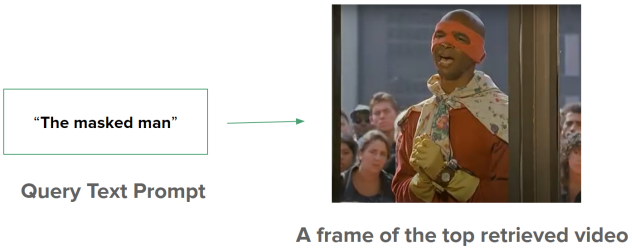

Figure 3. Query Text and a frame of the top retrieved video

Although only two examples are shown here, we experimented with various types of image and text prompts, and our model retrieved the most suitable videos from the lot. This shows that RA-CLIP Vanilla can capture the visual semantic information perfectly. Encouraged by these results, we would like to estimate the model's performance by the standard evaluation metrics for these tasks in the future.

5. Future Work

Building on the promising results of our initial implementation, there are several key areas where we plan to extend and enhance our work.

- **Extensive Training on Larger Datasets**: Our initial implementation of RA-CLIP showed promising results, yet it did not fully achieve the expected performance levels reported in the original RA-CLIP paper. This discrepancy is primarily due to the limited extent of our training. To address this, we plan to train RA-CLIP Vanilla more extensively using larger datasets. This will ensure that our model benefits from a broader range of image-text pairs, leading to improved generalization and performance. We aim to match and possibly surpass the benchmark performance metrics outlined in the RA-CLIP study.

- **Enhancing Reference Sets**: The effectiveness of RA-CLIP Vanilla largely depends on the quality and size of the reference sets used during inference. We intend to investigate the impact of using larger reference sets, as well as reference sets comprising high-quality data. By systematically varying the size and quality of these sets, we aim to identify optimal configurations that maximize model performance. This will provide deeper insights into the role of reference sets in retrieval-augmented vision-language models.

- **Benchmark Evaluations for Image-to-Video and Text-to-Video Retrieval**: While our preliminary tests with image-to-video and text-to-video retrieval tasks were encouraging, comprehensive benchmark evaluations are necessary to validate our findings. We plan to utilize standard benchmarks and datasets to rigorously evaluate RA-CLIP Vanilla's performance in these tasks. This will enable us to quantitatively assess the model's retrieval accuracy and robustness, ensuring that it meets the highest standards in the field.

- **Exploration of Additional Downstream Tasks**: Beyond the initial tasks, we aim to extend the application of RA-CLIP Vanilla to various other downstream tasks. These include:
  Scene Timestamp Retrieval: Identifying specific timestamps in a movie that match a given textual description. This application can significantly enhance content indexing and search capabilities in video databases.
  Character or Entity Scene Retrieval: Extracting scenes related to a specific character or entity within a video. This task has potential applications in content recommendation systems and automated video summarization.

6. Conclusion

In this paper, we introduced RA-CLIP Vanilla, an implementation of retrieval-augmented vision-language models inspired by the RA-CLIP framework. Our model aims to reduce the substantial pretraining costs associated with CLIP while maintaining competitive performance. Through initial experiments in image classification and video retrieval, RA-CLIP Vanilla demonstrated significant potential, highlighting its effectiveness and versatility. Moving forward, we plan to conduct extensive training with larger datasets,

optimize reference set configurations, and perform rigorous benchmark evaluations. Additionally, we aim to explore new downstream applications, further validating the robustness and applicability of RA-CLIP Vanilla. Our contributions, including the open-source code, will facilitate ongoing research and innovation in vision-language models, paving the way for more efficient and powerful AI systems.

7. Artifacts and Contributions

Our implementation of RA-CLIP and RA-CLIP Vanilla and the downstream tasks can be found in https://github.com/Jalend15/IMPART.
Our contributions are as follows:
Collected and Preprocessed Datasets: Jalend, Rohith
Worked on RA-CLIP and RA-CLIP Vanilla implementation: Vignesh, Jalend
Worked on Training setup: Rohith, Archana
Worked on Inference setup: Archana,Vignesh

References

[1] X. Bai, Y. Yang, et al. Towards coherent and consistent multimodal pretraining for zero-shot cross-modal retrieval. *arXiv preprint arXiv:2302.12552*, 2023. 2

[2] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. 3

[3] Mehdi Cherti et al. YFCC15M dataset. Hugging Face Datasets, 2023. 3

[4] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Technical report, IEEE, 2004. California Institute of Technology. 3

[5] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 3

[6] Jiangtong Li et al. Disentangling interactive visual representations for exploring multi-scale visual data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1234–1245, 2023. 2

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1, 2

[8] Chen-Wei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, and Jingren Zhou. Ra-clip: Retrieval augmented contrastive language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3