**Journal of Computing and Electronic Information Management**
ISSN: 2413-1660 | Vol. 14, No. 2, 2024

# Rethinking Semantic Contrastive Learning and Content Fusion in Multimodal Retrieval

**Baishun Shi *, Zhenghu Li**

School of Computer Science and Technology, Qingdao University, Qingdao, Shandong, 266000, China

* **Corresponding author**: Baishun Shi (Email: 422587267@qq.com)

**Abstract:** In the domain of image-text retrieval (ITR), recent advancements have enabled fine-grained (FG) instance-level retrieval through large-scale visual language pre-training (VLP). While these methods have achieved high accuracy, they have also led to an increase in computational complexity. primary challenges in cross-modal retrieval involve the induction of isomorphic knowledge and the association of heterogeneous knowledge. Homogeneous knowledge comprises elements with identical dimensions, whereas the interrelation of heterogeneous knowledge necessitates a prior unification of internal elements. Traditional cross-modal methods typically extract features from various modalities and engage in joint training. However, experimental results indicate that performance discrepancies among different modal networks can adversely affect overall generalization capability. Current state-of-the-art visual systems aim to minimize constrained supervisory signals to enhance the model's generalization performance. Although end-to-end models can simplify training, they often result in an exponential increase in data volume, which can be unmanageable for the average user. Our research demonstrates that pre-training with a singular focus can efficiently and scalably learn semantic features. This novel model is conceptually straightforward and can be implemented using existing, mature modules. In terms of performance, each module maintains a singular responsibility, significantly improving both the model's parameter count and training speed.

**Keywords:** Image-Text matching; Cross-Modal retrieval; Image-Text contrastive learning; Zero-Shot retrieval.

## 1. Introduction

Cross-modal learning involves simulating the mapping functions between inputs of different modalities [1]. Text- to-image retrieval is a significant cross-modal task that has garnered considerable attention in recent years [26,29,30]. Its primary objective is to provide users with cross-modal content that extends beyond mere text. Image- text retrieval (ITR) is a bidirectional retrieval task designed to extract relevant samples from one modality based on user expressions in another modality [1]. This task bridges the gap between visual and textual information, enhancing the richness, relevance, and accessibility of search results [4]. By simultaneously considering both images and text, ITR systems offer users a more intuitive means of exploring retrieved information. Although visual-language models (VLMs) have achieved state-of-the-art performance in this domain [2,22,18,23,15,27], there remains a need for refinement in the datasets and evaluation methods used to assess their performance. This necessity arises from two primary challenges, which will be elaborated upon in the following sections.

Despite its fundamental importance, image-text retrieval encounters the challenge of bridging the gap between the distinct modalities of text and images, which gives rise to a series of specific research questions. Image-text retrieval (ITR) generally comprises two sub-tasks: image-to-text (i2t) retrieval and text-to-image (t2i) retrieval. For each sub-task, the primary issue that image-text retrieval must address is how to enhance the understanding and alignment of information across varying modalities.

Image Text Retrieval (ITR) has significant potential in the search domain and is a critical area of research. Recent advancements in deep learning models for language and vision have led to notable successes in ITR [12,22]. For example, the introduction of BERT [9] has propelled the development of transformer-based cross-modal pre-training paradigms, which have been adapted for downstream ITR tasks, thereby accelerating their progress. As Transformers [25] gain widespread application in the image domain and large pre-trained models emerge, questions have arisen regarding the feasibility of employing a single encoder to directly encode both modalities. This approach aims to align them within a shared semantic space without the need for separate encoding of text and images prior to alignment. However, this architecture presents efficiency challenges, rendering it less suitable for large-scale image retrieval scenarios [5].

Current methods for text-to-image retrieval can be classified into two categories: 1) The first category, known as the single-stream structure [8, 10], employs a cross- attention mechanism to model fine-grained interactions. Recently developed large visual-language models, such as BLIP [19], have demonstrated significant capabilities in accurately ranking a small set of images. However, the single-stream structure has efficiency limitations, rendering it less suitable for large-scale image retrieval scenarios. 2) The second category is the dual-stream structure. In this framework, images and text, being two distinctly different modalities, are typically encoded separately before their features are mapped to a common semantic space for similarity computation [3, 30]. For instance, CLIP [15] independently maps visual and textual samples to a joint embedding space to calculate cross-modal similarity. To enhance efficiency, this framework sacrifices some accuracy, enabling it to perform well when retrieving relevant images from a large collection. In practice, retrieving a specific number of images from a large dataset is a preliminary and essential step for accurately ranking a small subset of images. Thus, we concentrate on this fundamental step to achieve effective cross-modal retrieval while

maintaining high efficiency.

In this study, we examine the methodologies associated with dual-stream structures and common space mapping. However, we have observed that the performance of this structure does not scale proportionately with the increase in the number of parameters. Upon further investigation, we have identified that this issue arises from the challenges associated with varying spatial mappings [28]. Notably, there are significant structural differences among different modalities, which are often evident in the variance of their dimensions. During the training process, the model must not only account for mappings within the same modality but is also influenced by heterogeneous data. Therefore, how can we mitigate the effects of heterogeneous data? We propose to re-evaluate cross-modal retrieval from a straightforward perspective: 1) first, identify the mapping space for each modality; and 2) project the features of disparate modalities into a common space.

## 2. Related work

Cross-modal retrieval, also known as text-image matching, is a fundamental task in multimedia processing. Based on the interaction methods between modalities, cross-modal retrieval can be categorized into single-stream structures and dual-stream frameworks.

One of the earliest and most influential papers in the field of image-text retrieval is VSE++ [11], which employs a conventional dual-stream architecture. Its Image Encoder utilizes VGG19 [24] and ResNet152 [17], while the Text Encoder is based on GRU [6]. The features generated by the two encoders are mapped to a shared semantic space through a linear layer, after which cosine similarity is employed to compute image-text similarity. A significant contribution of this work is the introduction of the hardest negative triplet loss during the training phase. This approach calculates triplet loss exclusively for the samples that are most similar to the target within the mini-batch, rather than considering all samples outside the target. Experimental results indicate that the max method outperforms the traditional sum method.

Following VSE++, another seminal paper on image-text retrieval, SCAN [20] also utilizes a dual-stream architecture model. In contrast to VSE++, SCAN achieves a more refined alignment between images and text. Initially, the model inputs image features encoded by ResNet into an object detector (Faster R-CNN) to identify and encode the target regions within the image. Subsequently, a two-stage attention mechanism is implemented, leveraging both object-level and word-level features to compute the similarity between images and text.

Subsequent non-pretrained image-text retrieval models essentially build upon the foundational concepts of VSE++ and SCAN, with an emphasis on refining the techniques for image-text alignment. Most of these models utilize a dual-stream architecture, and their primary advancements can be classified into three key areas: enhancing image encoding methods, optimizing attention mechanisms for image-text alignment, and incorporating external knowledge.

The widely utilized CLIP model exemplifies a dual- stream architecture, comprising two distinct image encoders: Vision Transformer (ViT) and ResNet, while the text encoder is based on the Generative Pre-trained Transformer (GPT) framework. During the pre-training phase, the model focuses exclusively on a single objective: image-text contrastive learning (ITC). The optimization goal encompasses two components: maximizing the cosine similarity score of matching image-text pairs within small batches and concurrently minimizing the scores of non-matching pairs. Although the pre-training task itself is relatively straightforward, the vast scale of the training dataset—comprising 400 million image-text pairs—yields remarkable zero-shot performance across various tasks.

The single-stream and dual-stream models each possess distinct advantages and disadvantages, complicating the determination of which structure is more effective. The single-stream model facilitates training and can effectively align diverse modalities, such as images and text, through intricate structures. However, in practical applications, it necessitates paired image-text inputs for inference, resulting in substantial computational demands during online processing. In contrast, the dual-stream model separates the encoders for images and text, permitting the offline computation of image features prior to inference, thereby enhancing response speed. Nevertheless, its approach to modeling image-text alignment is often relatively simplistic, typically relying on direct cosine similarity calculations of image-text features following a basic mapping.

Our study distinguishes itself from previous research by implementing semantic unification training before joint training. Once the semantic space is established, we proceed with mapping training. This methodology enables each component to fulfill its specific role, thereby significantly enhancing computational efficiency and reducing the volume of reports generated. Our contributions can be summarized as follows:

- Intra-modal semantic unification contrast training
- Heterogeneous modal correlation mapping.

## 3. The Proposed Method

This model is a visual-language pre-training framework distinguished by its streamlined architecture, which employs a minimal visual embedding pipeline and adheres to a dual-stream approach. In designing the model, we closely followed the original Transformer structure. One advantage of this intentionally simple configuration is that it facilitates the near out-of-the-box application of scalable natural language processing Transformer architectures and their efficient implementations.

The proposed multimodal model comprises three stages: 1) Text Semantic Alignment Learning: This stage focuses on acquiring semantic knowledge within the same modality, mapping identical semantics to a unified space. 2) Image Semantic Contrastive Learning: This stage applies the same processing techniques used for text to images, enabling the model to learn expressive features of similar image types and map them into the image space. 3) Semantic Shared Space Learning: This stage is designed to incorporate shareable semantic relationships into feature encoding, thereby facilitating the learning of more universally similar embeddings. Figure 1 shows an overview of the overall process of the model.

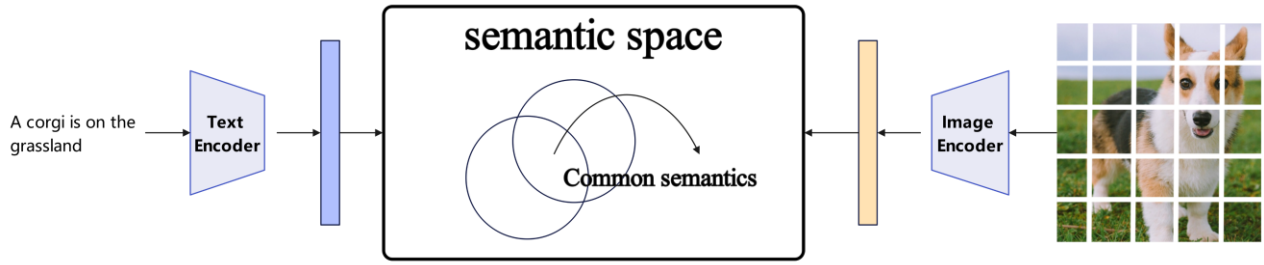


**Figure 1.** Overview flowchart of the model

## 3.1. Text Semantic Contrastive Learning

An overview of the model for textual semantic contrastive learning is presented in Figure 2. Initially, we categorize text sets with identical semantics as the positive sample set, while texts with differing semantics are classified as negative samples. The standard Transformer processes a one-dimensional sequence of token embeddings as input.

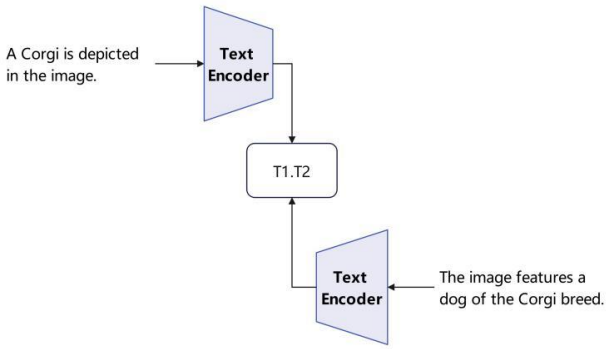


**Figure 2.** Contrastive learning with the same text semantics.

Similar to the [class] token in BERT, we introduce a learnable embedding ($z_0$= $x_{class}$) at the beginning of the sequence of embedded patches. The state of this embedding at the output of the Transformer encoder ($z_0$) is employed as the image representation T.

Positional embeddings are incorporated into patch embeddings to preserve positional information. We utilize standard learnable one-dimensional positional embeddings, as our observations indicate that more advanced two-dimensional perceptual positional embeddings do not yield significant performance improvements. The resulting sequence of embedding vectors is then fed into the encoder.

Our objective is to learn a shared embedding across identical modalities. To accomplish this, we propose utilizing the distance between the image representations generated by the two models as a learning metric, as illustrated in the following formula.

$$F_i = \text{text\_encoder}(T_i). \quad (1)$$

$$L=\text{Sum}_i\ (\ |\text{Euclidean Distance}(F_i, F_{i+1})+\text{dot}(F_i, F_{i+1})|). \quad (2)$$

After training, text features with similar semantics will cluster at proximate distances, indicating that features within this range share comparable meanings. Analogous to operations on sets, different semantic ranges may also demonstrate intersections, unions, and other relationships.

## 3.2. Image Semantic Contrastive Learning

An overview of the model is shown in Figure 3. First, we use a collection of images of the same type as the positive sample set, while images of different types are used as the negative samples. The standard Transformer takes a one-dimensional sequence of token embeddings as input.

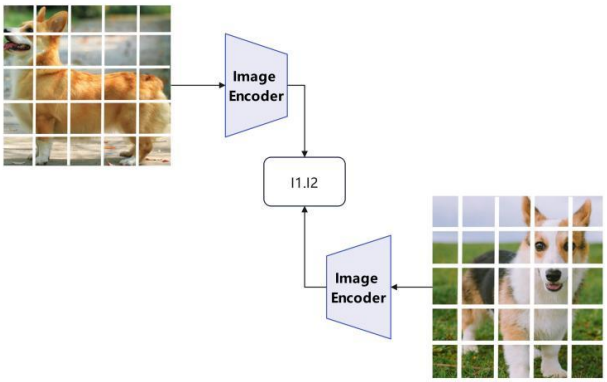


**Figure 3.** Contrastive learning with the same image semantics.

To process two-dimensional images, we divide the images into fixed-size patches, linearly embed each patch, add positional embeddings, and input the resulting sequence of vectors into a standard Transformer encoder. For classification purposes, we enhance this approach by appending a learnable "classification token" to the sequence. The Transformer maintains a constant latent vector size (D) across all layers; consequently, we flatten the patches and map them to (D) dimensions through a trainable linear transformation. The output of this mapping is termed "patch embeddings."

Our objective is to learn a shared embedding across identical modalities. To accomplish this, we propose utilizing the distance between the image representations generated by the two models as a learning metric, as illustrated in the following formula.

$$F_i = \text{image\_encoder}(M_i). \quad (3)$$

$$L=\text{Sum}_i\ (\ |\text{Euclidean Distance}(F_i, F_{i+1})+\text{dot}(F_i, F_{i+1})|). \quad (4)$$

Following training, the outputs of image features with identical semantics will cluster in close proximity. This clustering indicates that the features within this range exhibit similar image representations. Likewise, semantic features from varying ranges can also undergo operations such as intersection and union.

## 3.3. Semantic Shared Space

Through supervised signals derived from natural language processing, we can train visual models that demonstrate effective transfer learning. The input to the semantic shared space comprises pairs of images and text. For instance, if an image depicts a dog, the corresponding textual description should convey the same meaning.

By employing semantic contrastive learning for text, we

can generate a semantic set for the text; similarly, through semantic contrastive learning for images, we can derive a semantic set for the images. This module aims to establish mappings for the features of each modality. The semantic relationships among different regions across modalities are analogous, yet their expression dimensions differ. This is akin to observing lower dimensions from a higher- dimensional perspective, necessitating the establishment of relationships between different dimensions. The module is illustrated in Figure 4.

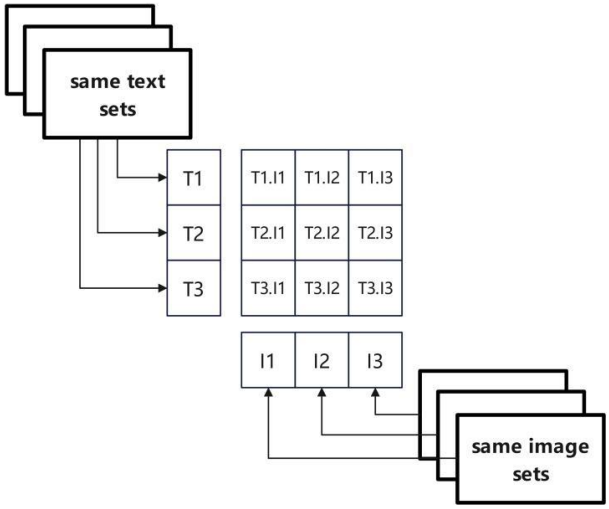


**Figure 4.** Contrastive learning with the same feature semantics.

The most straightforward approach involves reducing the distance between all feature pairs using a linear layer, thereby creating a mapping function. This method offers the advantage of allowing for retraining of the mapping at any time. Additionally, due to the principle of single responsibility, this approach significantly reduces the number of parameters involved.

# 4. Experiments

In this section, we conducted comprehensive experiments on three real-world datasets to assess their effectiveness.

## 4.1. Models

In this study, we selected four pre-trained large visual-language (VL) models for comparison. These models demonstrate state-of-the-art performance across a range of visual-language tasks, particularly excelling in information retrieval (IR) tasks.

1) ALIGN [13] is a visual language model (VLM) that utilizes a dual-encoder architecture based on contrastive learning, designed to mitigate the substantial dependence of contemporary visual and visual-language representation learning on specific training datasets. The ALIGN model harnesses large-scale noisy datasets for training, employing a contrastive learning mechanism that equips it with cross-modal retrieval capabilities and enhances its representation learning efficacy. Additionally, it adopts a straightforward training scheme, positioning it as a significant milestone in the domain of visual and visual-language representation learning.

2) The AltCLIP [035] model is a multimodal representation framework designed to enhance linguistic capabilities by modifying the language encoder within the CLIP (Contrastive Language-Image Pre-training) model. It employs a two-stage training scheme. In the teacher-student learning phase, the AltCLIP model uses CLIP's text encoder as the teacher encoder, while the XLM-R model, which is pre-trained on multilingual data, functions as the student encoder. This teacher-student learning strategy enables the AltCLIP model to extract knowledge from CLIP and adjust the output dimensions of XLM-R to align with those of the CLIP text encoder. During the contrastive learning phase, the AltCLIP model utilizes a relatively small number of Chinese and English text-image pairs for training. This process further enhances the consistency between text and image representations, thereby facilitating multilingual understanding of text and images.

3) The CLIP [15] (Contrastive Language-Image Pretraining) model, developed by OpenAI in 2021, is a multimodal pretraining framework. Its fundamental objective is to establish a joint representation space for images and text through contrastive learning. The CLIP model is pretrained on a large-scale dataset comprising image-text pairs, which enables it to understand the intricate relationships between images and text. This capability allows the CLIP model to exhibit strong generalization performance and zero-shot learning abilities across a variety of downstream tasks.

4) The GroupViT (Grouping Vision Transformer) model is an innovative visual transformer designed to perform semantic segmentation tasks using text supervision. Developed by researchers at NVIDIA NVlabs, this model mitigates the dependency on extensive pixel-level annotated data typically required in traditional semantic segmentation. By integrating a grouping mechanism and text supervision, GroupViT efficiently executes semantic segmentation without relying on pixel-level annotations.

## 4.2. Datasets

We utilized three public datasets as benchmarks:

1) The COCO Captions dataset [21] comprises 123,287 images sourced from Microsoft's Common Objects in Context (COCO) dataset, with each image accompanied by five human-generated captions. After excluding infrequent words, the average length of the captions is 8.7 words. This dataset is partitioned into 82,783 training images, 5,000 validation images, and 5,000 test images. We employed the partitioning method proposed by Karpathy [14] for processing this dataset.

2) Flickr30K [16] comprises 31,000 images sourced from the Flickr website, with each image accompanied by five textual descriptions. The dataset is partitioned into three subsets: 1,000 images for validation, 1,000 images for testing, and the remaining images for training. Following the methodology proposed by Karpathy [39], the dataset is organized into 29,783 training images, 1,000 validation images, and 1,000 testing images.

3) NUS-WIDE [7] is a multi-label dataset comprising 9,648 samples distributed across 81 categories. Each category contains 1,000 single-label images, culminating in a total of 10,000 images that constitute the NUS-WIDE-10K dataset. This dataset is randomly partitioned into three subsets: the training set, validation set, and test set, with sample sizes of 8,000, 1,000, and 1,000, respectively.

## 4.3. Evaluation Metric

We selected the widely utilized K-value recall rate (R@K) metric, which is the predominant evaluation metric in information retrieval (IR). R@K denotes the recall rate at the K-th position in a ranked list, defined as the proportion of

correctly matched items among the top K retrieval results, as outlined below:

$$R@k = \frac{number\ of\ relevant\ items\ in\ top\ k}{total\ relevant\ items} \quad (5)$$

Recall is particularly well-suited for instance-level retrieval, as it assesses the identification of specific individual items. In contrast, for category-level retrieval, we employ mean Average Precision at k (mAP@k) as the standard evaluation metric. The definition of precision at k (P@k) is provided below:

$$P@k = \frac{number\ of\ relevant\ items\ in\ top\ k}{total\ retrieved\ items} \quad (6)$$

### 4.4. Retrieval Results

As shown in Table I, our model outperforms all baseline methods. Compared to the results of the previously best-performing model, our model achieves a significant improvement on the Flickr30k dataset, with an average Recall at 1 (R@1) increase of 4.2%, an average Recall at 5 (R@5) increase of 3.1%, and an average Recall at 10 (R@10) increase of 2.3%. Additionally, on the COCO dataset, our model demonstrates an average R@1 increase of 4.2%, an average R@5 increase of 2.6%, and an average R@10 increase of 2.0%. These improvements can be primarily attributed to the precise delineation of individual responsibilities within our model, which effectively identifies the shared semantics between text and speech while capturing the subtle differences between them.

**Table 1.** Models performance on i2t

| Model | R@1 | R@5 | R@10 |
|---|---|---|---|
| ALIGN | 60.41 | 42.21 | 54.42 |
| AltCLIP | 58.24 | 40.66 | 53.01 |
| CLIP | 50.05 | 33.66 | 35.89 |
| GroupViT | 34.38 | 24.88 | 35.77 |

**Table 2.** Models performance on t2i

| Model | R@1 | R@5 | R@10 |
|---|---|---|---|
| ALIGN | 22.93 | 42.15 | 50.23 |
| AltCLIP | 22.45 | 41.86 | 50.09 |
| CLIP | 13.15 | 33.11 | 42.06 |
| GroupViT | 8.99 | 18.36 | 26.56 |

### 4.5. Zero-Shot

To evaluate the generalization capability of our framework, we conducted zero-shot retrieval on the Flickr8K test set using models trained on the COCO dataset. This investigation represents the first exploration of the semantic generalization ability of text-image retrieval models, as noted by the authors. The results are summarized in Table II, where "Supervised" denotes models trained on the Flickr8K training set. Notably, the model trained on the COCO dataset significantly outperformed the model trained on the Flickr8K dataset, demonstrating the exceptional generalization ability of our framework. This superior performance can be attributed both to the high quality of the COCO dataset and to the unification of knowledge, which enhances the model's capacity for accurate mappings. Consequently, our model exhibits strong scalability. We firmly assert that training on larger corpora will further improve its generalization capabilities.

**Table 3.** Recall scores for zero-shot performance on i2t

| Method | R@1 | R@5 | R@10 |
|---|---|---|---|
| Supervised | 61.41 | 50.21 | 70.42 |
| Zero-Shot | 66.24 | 61.66 | 72.01 |

## 5. Conclusions

In this study, we address the issue of single responsibility, focusing on two primary concerns: the unification of homogeneous knowledge and the mapping of relationships among heterogeneous modalities. We selected four state-of-the-art visual language models (VLMs)—AltCLIP, ALIGN, CLIP, and GroupViT—for our experiments, which comprised three distinct phases. First, we trained the models using refined and fully accurate textual semantics in conjunction with images of single objects. Next, we trained the models with incomplete sentences paired with images of single objects. Finally, we trained the models using semantically ambiguous text alongside images of multiple objects.

We found that the models trained on the first group of precise data can generalize well to the other two ambiguous datasets, significantly enhancing the performance of visual language models (VLMs) in image-text retrieval (ITR) tasks. This improvement has been observed across different datasets and tasks. Specifically, the fine-grained nature of the datasets positively impacts the performance of VLMs in ITR tasks. Fine-grained datasets consistently yield higher performance for all selected models. This finding underscores the importance of semantic coherence within the same modality. Therefore, we recommend that future benchmarking efforts focus on developing more fine-grained datasets, such as MS-COCO-FG and Flickr30k-FG, to better assess the ability of VLMs to capture subtle semantic differences.

Furthermore, our findings indicate that applying perturbations to the dataset generally results in a decline in model performance, underscoring the sensitivity of visual language models (VLMs) to variations in input data. In contrast, models evaluated on more fine-grained datasets exhibited relatively smaller performance declines, thereby reinforcing the significance of dataset granularity. This observation emphasizes the importance of dataset curation and evaluation methods in ensuring the robustness and generalization capabilities of visual language models. Overall, the results of this study make substantial contributions to the establishment of more reliable benchmarks and evaluation practices for visual language models in the field of information retrieval (IR).

Several limitations of this study are noteworthy, including the requirement for a backup model to serve as a reference during the training phase. This necessity leads to suboptimal utilization of machine performance throughout the training process. While a backup model is not essential during the inference phase, we aim to maximize its utility in future training endeavors.

## References


[1] Baltrušaitis, T., Ahuja, C., & Morency, L. (2017). Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41, 423-443.

[2] Chen, Z., Liu, G., Zhang, B., Ye, F., Yang, Q., & Wu, L.Y. (2022). AltCLIP: Altering the Language Encoder in CLIP for Extended Language Capabilities. ArXiv, abs/2211.06679.

[3] Chen, J., Hu, H., Wu, H., Jiang, Y., & Wang, C.L. (2020). Learning the Best Pooling Strategy for Visual Semantic Embedding. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15784-15793.

[4] Cao, M., Li, S., Li, J., Nie, L., & Zhang, M. (2022). Image-text Retrieval: A Survey on Recent Research and Development. ArXiv, abs/2203.14713.

[5] Chen, Y., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2019). UNITER: UNiversal Image-TExt Representation Learning. European Conference on Computer Vision.

[6] Chung, J., Gülçehre, Ç., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. ArXiv, abs/1412.3555.

[7] Chua, T., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009). NUS-WIDE: a real-world web image database from National University of Singapore. ACM International Conference on Image and Video Retrieval.

[8] Chen, H., Ding, G., Liu, X., Lin, Z., Liu, J., & Han, J. (2020). IMRAM: Iterative Matching With Recurrent Attention Memory for Cross-Modal Image-Text Retrieval. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12652-12660.

[9] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics.

[10] Diao, H., Zhang, Y., Ma, L., & Lu, H. (2021). Similarity Reasoning and Filtration for Image-Text Matching. ArXiv, abs/2101.01368.

[11] Faghri, F., Fleet, D.J., Kiros, J.R., & Fidler, S. (2017). VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. British Machine Vision Conference.

[12] Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., & Mikolov, T. (2013). DeViSE: A Deep Visual-Semantic Embedding Model. Neural Information Processing Systems.

[13] Jia, C., Yang, Y., Xia, Y., Chen, Y., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., & Duerig, T. (2021). Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. ArXiv, abs/2102.05918.

[14] Karpathy, A., & Fei-Fei, L. (2014). Deep visual-semantic alignments for generating image descriptions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3128-3137.

[15] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. International Conference on Machine Learning.

[16] Huiskes, M.J., & Lew, M.S. (2008). The MIR flickr retrieval evaluation. Multimedia Information Retrieval.

[17] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778.

[18] Li, X., Yin, X., Li, C., Hu, X., Zhang, P., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., & Gao, J. (2020). Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. ArXiv, abs/2004.06165.

[19] Li, J., Li, D., Xiong, C., & Hoi, S.C. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. International Conference on Machine Learning.

[20] Lee, K., Chen, X., Hua, G., Hu, H., & He, X. (2018). Stacked Cross Attention for Image-Text Matching. ArXiv, abs/1803.08024.

[21] Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C.L. (2014). Microsoft COCO: Common Objects in Context. European Conference on Computer Vision.

[22] Li, J., Selvaraju, R.R., Gotmare, A., Joty, S.R., Xiong, C., & Hoi, S.C. (2021). Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. Neural Information Processing Systems.

[23] Mu, N., Kirillov, A., Wagner, D.A., & Xie, S. (2021). SLIP: Self-supervision meets Language-Image Pre-training. ArXiv, abs/2112.12750.

[24] Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR, abs/1409.1556.

[25] Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. Neural Information Processing Systems.

[26] Wang, B., Yang, Y., Xu, X., Hanjalic, A., & Shen, H.T. (2017). Adversarial Cross-Modal Retrieval. Proceedings of the 25th ACM international conference on Multimedia.

[27] Xu, J., Mello, S.D., Liu, S., Byeon, W., Breuel, T., Kautz, J., & Wang, X. (2022). GroupViT: Semantic Segmentation Emerges from Text Supervision. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 18113-18123.

[28] Yang, Y., Ye, H., Zhan, D., & Jiang, Y. (2015). Auxiliary Information Regularized Machine for Multiple Modality Feature Learning. International Joint Conference on Artificial Intelligence.

[29] Zhen, L., Hu, P., Wang, X., & Peng, D. (2019). Deep Supervised Cross-Modal Retrieval. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10386-10395.

[30] Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Xu, M., & Shen, Y. (2017). Dual-path Convolutional Image-Text Embeddings with Instance Loss. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 16, 1 – 23.