Where Does the Performance Improvement Come From? - A Reproducibility Concern about Image-Text Retrieval

Jun Rao[1*], Fei Wang[2*], Liang Ding[3†], Shuhan Qi[1†], Yibing Zhan[3], Weifeng Liu[2], Dacheng Tao[3]
[1]Harbin Institute of Technology, Shenzhen [2]China University of Petroleum (East China)
[3]JD Explore Academy
dingliang1@jd.com,shuhanqi@cs.hitsz.edu.cn

ABSTRACT

This paper seeks to provide the information retrieval community with some reflections on the current improvements of retrieval learning through the analysis of the reproducibility aspects of image-text retrieval models. For the latter part of the past decade, image-text retrieval has gradually become a major research direction in the field of information retrieval because of the growth of multi-modal data. Many researchers use benchmark datasets like MS-COCO and Flickr30k to train and assess the performance of image-text retrieval algorithms. Research in the past has mostly focused on performance, with several state-of-the-art methods being proposed in various ways. According to their claims, these approaches achieve better modal interactions and thus better multimodal representations with greater precision. In contrast to those previous works, we focus on the repeatability of the approaches and the overall examination of the elements that lead to improved performance by pretrained and nonpretrained models in retrieving images and text.

To be more specific, we first examine the related reproducibility concerns and why the focus is on image-text retrieval tasks, and then we systematically summarize the current paradigm of image-text retrieval models and the stated contributions of those approaches. Second, we analyze various aspects of the reproduction of pretrained and nonpretrained retrieval models. Based on this, we conducted ablation experiments and obtained some influencing factors that affect retrieval recall more than the improvement claimed in the original paper. Finally, we also present some reflections and issues that should be considered by the retrieval community in the future. Our code is freely available at https://github.com/WangFei-2019/Image-text-Retrieval.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Specialized information retrieval**; **Multimedia and multimodal retrieval**;

∗ Equal contributions from both authors. Work was done when Jun and Fei were interning at JD Explore Academy.
† Corresponding Author.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/XXXXXXX.XXXXXXX

KEYWORDS

Image-text retrieval, Network reliability, Reproducibility

ACM Reference Format:
Jun Rao[1*], Fei Wang[2*], Liang Ding[3†], Shuhan Qi[1†], Yibing Zhan[3], Weifeng Liu[2], Dacheng Tao[3]. 2018. Where Does the Performance Improvement Come From? - A Reproducibility Concern about Image-Text Retrieval. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

As technology progresses, the content of information retrieval has evolved from a single-modality to a multimodal approach [11]. The continuous development of social platforms has resulted in an increase in the quantity of multimedia data on the internet, such as images and text. Finding similar content within such massive quantities of multimedia data has become a real issue in the industry. [9]. Due to the requirements of these practical applications, developing an effective image and text retrieval system has become a significant area of research in the field of information retrieval. The specific goal is to provide a flexible retrieval experience that indexes semantically relevant instances from one modality to another.

Image-text retrieval has been intensively investigated in recent years and can be divided into two categories according to whether using pretrained models. On the one hand, *visual-and-language pretraining* (VLP) based on pretrain-finetune paradigm has achieved state-of-the-art results on a range of downstream tasks such as image retrieval, visual question answering, and visual reasoning (e.g. Chen et al. [4], Lu et al. [25]). Most of these VLP models extend BERT [5] to learn representations grounded in both visual and textual contexts. These VLP models mainly differ in designing the pretraining tasks, modality interaction, and the quantity of pretraining data [13]. However, although these VLP models have been proposed and reported state-of-the-art results on various downstream tasks, there is still little research on what factors affect the final downstream task. To address this gap, we focus on the image-text retrieval task and attempt to compare these VLPs to the best of our ability, exploring salient factors that may affect retrieval results. Additionally, while reproducing these VLP models, we raised concerns and thought about the reproducibility of the results.

On the other hand, the current *nonpretrained image-text retrieval models* are also a research hotspot because it generally requires significantly fewer parameters compared to their pretrained counterparts. The price for such efficiency and flexibility is performance. Numerous methods [6, 8, 17, 19, 39, 43] have been proposed in recent years, most of which claim to achieve better modal interactions and thus better multimodal representations. It is relatively easy to

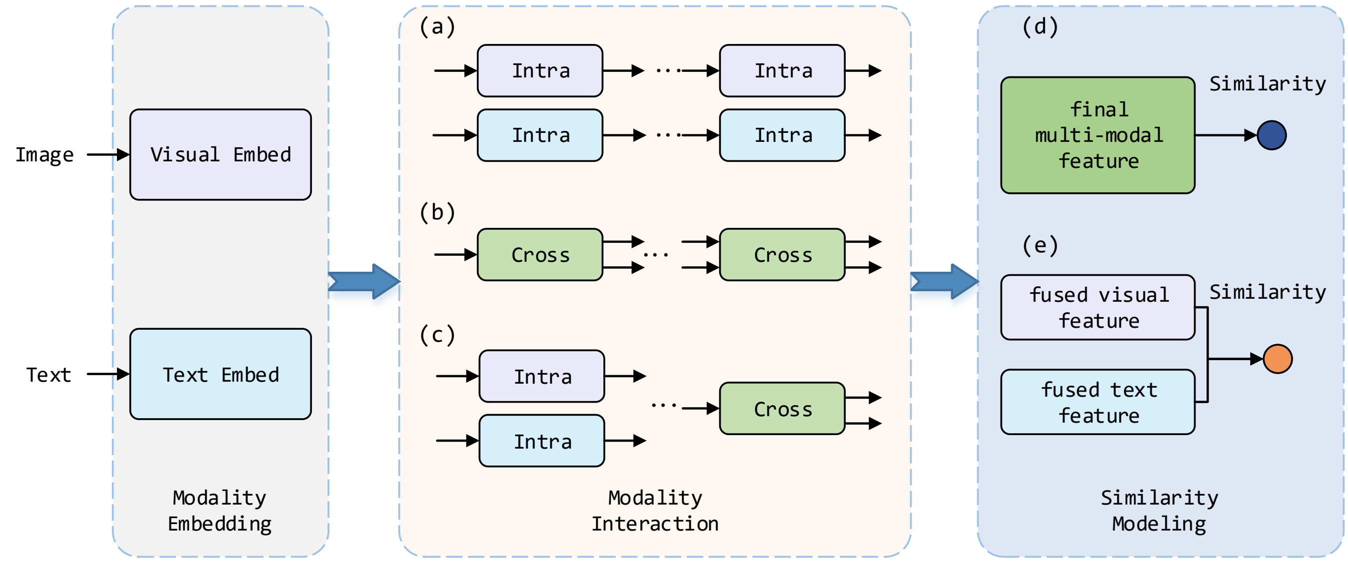

Figure 1: Overview image-text retrieval framework

disentangle the factors that influence these nonpretrained models compared to pretrained models. We, therefore, chose a group of open-source methods, tried our best to reproduce the results of the original paper, and perform them with different experimental setups to obtain new findings and the key factors that may influence the results.

We conducted experiments on two of the most widely used large-scale datasets, Flickr30k and MS-COCO. We tried 5 pretrained retrieval models and 7 nonpretrained retrieval models and reproduced these methods as closely as possible according to the paper description and the provided codes. We ran three separate experiments using different random seeds for each model on both datasets and took the final mean as the results reported in our table. Surprisingly, just different initializations along with hard samples and some seemingly trivial details can lead to huge differences in model performance. Thus, using nonpretrained methods, we tried to run another ten experiments with different random seeds from the front on the Flickr30k dataset and draw the violin figure proving it.

In summary, our contributions in this paper are as follows:

- We give a comprehensive overview of image-text retrieval learning methods, including modality embedding, modality interaction and similarity modeling, and a family of retrieval methods with pretrained and nonpretrained.
- We conduct a series of controlled studies in two benchmark datasets, raise some concerns about the reproducibility of the settings of pretrained models and find that the improvements of nonpretrained models may come from hyperparameters, hard negative sampling strategies and modality interaction types.
- We discuss the conjectures, give recommendations and provide insightful guidance in the information retrieval area.

2 WHY CONCERN REPRODUCIBILITY IN IMAGE-TEXT RETRIEVAL?

2.1 What is image-text retrieval?

From 2018 to the present, many research papers related to cross-modal retrieval have been presented at major conferences, such as CVPR, ICCV, ECCV, MM, SIGIR, ICML, etc. And some easy-to-practice and effective methods [1, 4, 16, 17, 25] have been widely used in practical commercial applications. With the growth of the Internet, the forms of multimodal data, such as photos, texts, audios, and videos have expanded rapidly, with images and texts being the two most common modalities. As a result, how to retrieve these two fundamental modalities of vision and text is crucial and inspiring for cross-modal retrieval.

A major challenge of image-text retrieval is the need to model the semantic information of different modalities and align the semantic information of different modalities. Many current image-text retrieval methods encode the features to a semantic space through modality-independent encoders and then perform modal fusion to obtain the corresponding fusion features. Finally, the features are converted by a head pooler into similarity to measure the similarity of the image and text. Following the completion of learning, the features of database items are calculated and indexed so that the retrieval system can efficiently perform retrieval similarity calculations to return the retrieval ranking results to the user.

2.2 Why concern reproducibility?

A remarkable series [1, 5, 10, 29, 36] of empirical successes in academia and industry [9] has accompanied and nourished the rapid increase in academic research on image-text retrieval.

Recent research, in contrast, appears to have slowed, and the proposed method is not a straightforward framework for generalizability. Instead, through complicated module and model ensembles, extra parameter settings are provided to achieve performance benefits on datasets. These approaches are not very generic or useful, and it is difficult to maintain their effectiveness when circumstances change. However, proposals that are eventually embraced by the information retrieval community and practitioners are those that steadily increase performance across a wide range of "real-world" situations. An influential method should be highly generalizable and capable of many different parameter settings, such as transformers [36] and residual networks [10]. As a result, it is crucial to determine which approaches are reproducible and can be generalized in different settings and environments.

3 A UNIFIED FRAMEWORK OF IMAGE-TEXT RETRIEVAL

As shown in Figure 1, we summarize the general process of the current image-text retrieval model and roughly divide each component of the retrieval model into three blocks, namely, modality embedding, modality interaction and similarity calculation. In the following subsections, we describe the three key components and provide an architectural overview of image-text retrieval.

3.1 Modality embedding

The majority of work is devoted to enhancing the model's capability through modifying visual features, while text features are rarely considered. Most researchers have previously concentrated on visual features, thinking them to be the bottleneck affecting the retrieval model. We believe, however, that learning how to use text features is critical. Additionally, we demonstrate a series of visual and textual feature advancements.

3.1.1 Visual representations.

Region feature. Region features are dominantly utilized among image-text retrieval models. They are pretrained on the Visual Genome (VG) dataset processed by Anderson et al. [1] to obtain an off-the-shelf object detector, such as Faster R-CNN [29].

The region feature extractor model can be varied by using different detection architectures, such as FPN [22] and C4, [1] or using different CNN backbones, such as ResNet101 [25, 33, 34] and ResNet152 [20, 21]. Although features can greatly affect retrieval performance, previous work seems to be very tolerant of visual embedding. Even if the encoders are different, many methods still only compare the final retrieval accuracy and ***do not mention the effect*** of visual embedding on their own model. Moreover, the number of regions also has a great impact on the final result. However, some methods [21] use more regions, resulting in unfair comparisons.

Grid feature and patch projection. These two types of features are mostly used in the pretrained image-text retrieval model and are rarely used in the nonpretrained model because of their worse retrieval performance. However, once pretrained with a large amount of image-text pairs, these two types of features seem to be effective and meaningful. The grid feature was first proposed in the VQA task by Jiang et al. [14] to reduce the slow region selection operation. It is also extracted through the pretrained CNN model. Compared with regional features, such features do not need regional selection, so they are faster in practice. Patch Projection [7] was first adopted in image-text retrieval by ViLT [16]. Compared with the previous two types of features, this feature is more direct and faster with less parameter consumption, without region selection or pretrained CNN. However, in practice, the performance of these two types of features is still a certain margin reduced compared with regional features.

3.1.2 Textual representations.

Different from visual representation, text representation does not seem to have great differences. Most methods use the powerful pretrained language model Bert [5] to get text representation, and some methods [6, 8, 17, 19, 24, 28] also use GRU [2, 31].

3.2 Modality interaction

Most nonpretrained models claim that their contribution includes better modality interaction. Modality interactions can be roughly divided into two basic types, as shown in Figure 1. Mode (a) is self-interaction, which usually uses the attention mechanism to interact with the features in the model. The second method, as shown in (b), is the interaction between modalities. Usually, different modal features aggregate and share features through different mechanisms, such as graph attention networks [37], self-attention [3] and cross attention [25]. The third method (c) is the combination of the first two. Better retrieval results can be obtained through artificially defined feature interactions. Although the third method can achieve better retrieval performance through parameter adjustment, this method is too dependent on experience, lacks a certain generalization ability, and has worse robustness and generalization.

3.3 Similarity modeling

Similarity modeling can be roughly divided into two categories, as shown in Fig. 1, (d) and (e). The first category (d) obtains the joint representation of the image-text pair after multimodal interaction and usually appends a fully connected (FC) layer to obtain the similarity followed by softmax to predict a two-class probability p^{itm}. Many VLP models include the image-text matching (ITM) loss, which predicts whether an image and text pair match:

$$\mathcal{L}_{\mathrm{itm}} = \mathbb{E}_{(I,T)\sim D}\mathrm{H}\left(y^{\mathrm{itm}}, p^{\mathrm{itm}}(I,T)\right), \tag{1}$$

where y^{itm} is a 2-dimensional one-hot vector representing the ground-truth label and H is the cross-entropy. In general, this calculation method is usually used in the pretrained retrieval model.

The second method (e) obtains the representation of each modality and obtains the similarity between unimodal by directly calculating the similarity or learning the similarity function. This kind of similarity modeling method usually adopts contrastive image-text matching losses, which have been successful in self-supervised representation learning [35]. For each image and text, one method for calculating the softmax-normalized image-to-text and text-to-image similarity is as follows:

$$p_m^{\mathrm{i2t}}(I) = \frac{\exp\left(s\left(I, T_m\right)/\tau\right)}{\sum_{m=1}^{M}\exp\left(s\left(I, T_m\right)/\tau\right)} \tag{2}$$

$$p_m^{\mathrm{t2i}}(T) = \frac{\exp\left(s\left(T, I_m\right)/\tau\right)}{\sum_{m=1}^{M}\exp\left(s\left(T, I_m\right)/\tau\right)}, \tag{3}$$

where τ is the temperature parameter. Let $y^{\mathrm{i2t}}(I)$ and $y^{\mathrm{t2i}}(T)$ denote the ground truth, where the probability equals 1 if matched otherwise 0. Then, the contrastive loss can be defined as follows:

$$\mathcal{L}_{\mathrm{itc}} = \frac{1}{2}\mathbb{E}_{(I,T)\sim D}\left[\mathrm{H}\left(y^{\mathrm{i2t}}(I), p^{\mathrm{i2t}}(I)\right) + \mathrm{H}\left(y^{\mathrm{t2i}}(T), p^{\mathrm{t2i}}(T)\right)\right] \quad (4)$$

Similar to the contrastive loss, another basic form for updating similarity is the bidirectional ranking loss as shown in Equation 5:

$$\begin{aligned}\mathcal{L}(\boldsymbol{I},\boldsymbol{T}) = \sum\left[\mu - s(\boldsymbol{I},\boldsymbol{T}) + s\left(\boldsymbol{I},\boldsymbol{T}^{-}\right)\right]_{+} \\ + \sum\left[\mu - s(\boldsymbol{I},\boldsymbol{T}) + s\left(\boldsymbol{I}^{-},\boldsymbol{T}\right)\right]_{+}\end{aligned} \quad (5)$$

Compared to a pairwise loss, VSE++ [8] uses hard negative mining to give us more flexibility in the embedding and can be easier to optimize:

$$\begin{aligned}\mathcal{L}(\boldsymbol{I},\boldsymbol{T}) = \max\left[\mu - s(\boldsymbol{I},\boldsymbol{T}) + s\left(\boldsymbol{I},\boldsymbol{T}^{-}\right)\right]_{+} \\ + \max\left[\mu - s(\boldsymbol{I},\boldsymbol{T}) + s\left(\boldsymbol{I}^{-},\boldsymbol{T}\right)\right]_{+},\end{aligned} \quad (6)$$

where $[x]+ = max(x, 0)$ is a clip function, $s(,)$ indicates the similarity prediction function and μ is a positive constant, which we term the margin. In Equation 6, compared to a pairwise loss (Equation 5), this loss addresses about the rank of the points with respect to a query rather than their exact distance, while another considers the sum of the violations for each negative sample.

In general, the idea of these functions is to raise the relevance score between an image and its matching text while decreasing the relevance score between an image and its irrelevant words. In terms of reproducibility, how to select training losses and samples can seriously affect the final retrieval result.

4 MATERIALS AND METHODS

4.1 Datasets

MS-COCO [23] and Flickr30k [41] have been used as benchmark datasets in most methods. The MS-COCO and Flickr30k datasets contain 123,287 and 31,783 images, respectively, and each image has five corresponding sentence descriptions. Most of the methods claim to split by Karpathy and Fei-Fei [15], using 1,000 images for validation and 1,000 images for testing in Flickr30k dataset and 113,287 images for training and 5,000 images for validation and 5,000 images for testing in the MS-COCO dataset. However, in the process of reproduction, we found that almost all methods add the data of the validation set to the training data to report higher test set results. Additionally, because MS-COCO's 5k test is too time-consuming, some methods use a test that averages 5-fold of 1K test images from 5k images, also known as the 1k test. However, how to divide and whether this result is the implementation of the best dividing result is still unclear. Therefore, for a fair comparison later, we adopt a unified division method and average multiple measurements to obtain more stable results.

4.2 Evaluation metrics

At the test time, the result performance for image-text retrieval is reported by recall at K (R@K) which represents the ranking proportion of ground-truth queries within the top K. R@1, R@5 and R@10 are our evaluation metrics.

4.3 Models

4.3.1 pretrained models.

The pretrained language model has gained considerable interest from the natural language processing, computer vision, and information retrieval communities because it can use self-supervised learning through unified pre-training and performs well on many downstream tasks. The visual-and-language pretraining (VLP) models achieve better performance in different downstream tasks. Most of the VLP models are pretrained on the image-text pairs of Google Conceptual Captions (GCC) [32], SBU Captions (SBU) [27], Microsoft COCO (MS-COCO) [23] and VG datasets.

The existing VLPs are often aimed at many downstream tasks, resulting in many VLPs that may not have tested the results of the image-text retrieval task. Therefore, it is meaningless to make comparisons with these missing data methods. We selected these VLPs based on the availability of full image-text retrieval test results and the influence of their citation count. We chose these five models: ViLBERT [25], PixelBERT [12], Unicoder-VL [18], UNITER [4] and ViLT [16]. We summarize these VLPs in Table 2. These models use similar text encoders (BERT) and similar visual encoders (ROIs), but use different pre-training tasks and modality interaction architectures. On the downstream image-text retrieval tasks, these models all use the ITM loss, as shown in Equation 1.

ViLBERT [25] introduced a cross-attention mechanism to fuse the features of the visual flow and the text flow and obtained fused visual features and text features, respectively. This modality interaction method belongs to (c) in Figure 1, which is also the most important contribution of this paper. This work is the originator of the pretrained visual-and-language model and has received nearly 1,000 citations. However, this work only reports the image retrieval score of Flickr30k, and other numerical reports related to image-text retrieval are missing.

PixelBERT [12] feeds the text and image with CNN embeddings into the transformer together, which indicates the single-stream framework and belongs to type (b) in Figure 1. It uses a multimodal transformer to align visual-and-language information and became the standard fusion method for the subsequent single-stream model. In addition, it reports the complete image-text retrieval results but lacks the code and details of the implementation.

Unicoder-VL [18] uses a larger pre-training dataset (CC3M), and uses the contrast loss with the hardest negatives (Equation 6) to optimize the image-text retrieval task for the first time. Additionally, the code for this work is not open source, nor does it provide specific checkpoints.

PixelBERT and **Unicoder-VL** lack code and details, it is basically impossible to reproduce the pretraining and downstream task results. However, due to the influence and inspiration of these two works, we still consider these two methods in the subsequent discussion.

The model architectures of **UNITER [4]** and Unicoder-VL are basically the same, belonging to type (b) in Figure 1. The difference is that UNITER uses a better combination of pretraining tasks and larger pretraining datasets. Additionally, it completely tests the results on the image-text retrieval datasets. Although this work provides training checkpoints and open-source code, it is nearly impossible to reproduce due to the hard negative mining time limit. This work models similarity (type d) using ITM loss (Equation 1) as in previous work but with hard negative mining, which may improve nearly 5̃10 points in R@1.

Table 1: Terms and notations in our work.

Notation	Description
MLM	masked language modeling
ITM	image-text matching
MIM	masked visual-feature classification
WRA	word region alignment
BS	batch size
LR	learning rate
CC	conceptual captions
SBU Captions	SBU
HNM	hard negative mining [30]
DA	data augmentation
FP16	use float16 operations

ViLT[16] is one of the simplest VLP models. It is mostly the same as UNITER, using the same architecture type (b) and pretraining tasks, and using uniformly input image patch and text encoding into the transformer to obtain competitive results in the image-text retrieval task. Similar to UNITER, its similarity modeling (d) also uses ITM loss (Equation 1). This method is easier to reproduce due to its simplicity and less extra setup.

4.3.2 Nonpretrained models. Direct comparison of nonpretrained models and VLP is not fair due to the use of more data and longer training time. In the case of limited resources, it is also necessary to study nonpretrained models. The main improvement of the nonpretrained models claiming is mainly the modality interaction and similarity modeling in Figure 1. Therefore, we use 7 nonpretrained models with open source code for experimental comparison to determine the extent to which these assumptions hold. We show the differences in the architecture of these nonpretrained models in Table 4. The claimed contributions of the individual models are further explained next.

VSE++ [8] includes the hard negative mining technique in the ranking loss, which contributed significantly to the improvement as they claim. Additionally, unlike many later works, their visual encoding uses CNN, and text encoding uses GRU. VSE++ uses self-attention model to obtain the weighted modal encoding of type (a), maps visual and text features to a representation space, and obtains the similarity of the two modalities through the dot product of (e).

SCAN [17] employs a stacked cross-attention model to predict similarity by taking into account the dense paired cross-modal interaction. Different from VSE++ [8], SCAN uses regions of interest (ROIs), to obtain the visual embedding. Then, SCAN uses the attention between modalities to obtain the fused modal information through type (b) and obtains the final global image-text matching score by means of (d), as shown in Figure 1.

CAMP [39] claims that with an adaptive gating method, they not only consider comprehensive and fine-grained cross-modal interactions but also properly handle negative pairings and irrelevant information, which are classified into categories (c). They regard their method as an efficient method for exploring the interactions between images and sentences before calculating similarities.

VSRN [19] proposes an interpretable and straightforward reasoning model by constructing visual representations that capture significant items and semantic concepts in a picture. This approach focuses on interactions within visual modalities belonging to type (a). This proves that the modal information of vision has not been fully exploited.

SAEM [40] employs self-attention embeddings to take advantage of fragment relations in pictures or texts and aggregate fragment information into visual and textual embeddings. The modality interaction can be classified into category (a). Similar to the four previous works, the basic loss used in the similarity modeling is a contrastive loss (Equation 6). Furthermore, SAEM [40] adds hard negative mining on angular loss [38].

CAMERA [28] does not use a pair of image-text data for training but adds image-text joint training for multiview descriptions, and selects content information through an attention module, which also takes advantage of intra-modal interactions and inter-modal interactions (c). Although CAMERA also used a contrastive loss similar to previous works, CAMERA introduces a diversity regularization term that causes a difference in the loss term. This causes additional parameter adjustments and increases the difficulty for subsequent improvement exploration.

SGRAF [6] designs the SGR module for graph reasoning, and designs the SAF to filter useless information, using type (c) to conduct modality interaction and obtain better semantic alignment. It also uses a contrastive loss with the hardest negative (Equation 6), using the (d) method to model similarity.

Although these models claim well performance in their corresponding papers, there are still some concerns and areas for improvement. First, unlike VLPs, modality embeddings are rarely studied in nonpretrained models. Most methods use 36 visual regions to encode the image input and the text encoder GRU. Second, these methods lack generalization and have high parameter sensitivity. SCAN [17], VSRN [19] and CAMERA [28] report ensemble results. By doing so, these methods get better reporting results but reduce the generalization and ease of use of the method. Moreover, they rely too much on the granularity of feature encoding and essentially filter and weight modal features of different granularities through handcrafted fine-grained interaction modules. Finally, some key elements of the models are poorly described in the original paper, but these elements are often the key to influencing the results of the model.

5 ANALYSIS

5.1 Experimental Setup

To allow readers to intuitively understand some of the settings of the following tables, we list the relevant symbols as shown in Table 1. In the following experiments, we take some tables of the experimental setup of the method, as shown in Table 2 and Table 4.

5.2 Pretrained Models

We compare existing pretrained image-text retrieval models and give their detailed settings and parameter comparisons as shown in Table 2. We analyze the ability of the retrieval model, the impact of the factor, and the reproducibility from the following two perspectives: the quantity of pretraining data and additional settings.

Table 2: Comparisons with existing VLP methods and details in image-text retrieval. † represents methods that cannot reproduce results due to lack of code and training details. ‡ represents that the reproduced results are very close to the results of the paper.

Method	Params	Architecture	Visual Tokens	Pre-train Datasets	Pre-train Tasks	Flickr30k		COCO		BS	warmup %	Loss	tricks	code
						epoch	LR	epoch	LR					
ViLBERT (paper [25]/reproduction [42])	221M	one single-modal Transformer (language) + one cross-modal Transformer (with restricted attention pattern)	image RoI	CC	1) MLM 2) ITM 3) MIM	20/17	4e-5	-/17	4e-5	64	0.1	cross entropy	1) HNM 2) FP16	PyTorch
PixelBERT† [12]	142M	single cross-modal Transformer	CNN	MS-COCO VG	1) MLM 2) ITM	10	1e-4	4	1e-4	512	-	cross entropy	1) HNM 2) DA 3) FP16	no
Unicoder-VL† [18]	110M	single cross-modal Transformer	image RoI	CC	1) MLM 2) ITM 3) MIM	-	5e-5	-	5e-5	192	0.1	contrastive loss	1) HNM 2) FP16	no
UNITER (paper [4]/reproduction [42])	110M	single cross-modal Transformer	image RoI	CC SBU MS-COCO VG	1) MLM 2) ITM 3) MIM 4) WRA	5000 steps/15	5e-5 / 4e-5	5000 steps/15	5e-5 / 4e-5	8/64	0.1	cross entropy	1) HNM 2) FP16	PyTorch
ViLT‡ [16]	111M	single cross-modal Transformer	image patch	CC SBU MS-COCO VG	1) MLM 2) ITM	15	1e-4	10	1e-4	256	0.1	cross entropy	1) DA 2) FP16	PyTorch

Table 3: Comparisons with existing VLP methods and their results in image-text retrieval. - represents the results of the original paper were not given.

Method	Flickr30k (1K)						MS-COCO (5K)					
	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
ViLBERT (paper/reproduction)	58.2/59.1	84.9/85.7	91.5/92.0	-/76.8	-/93.7	-/97.6	-/38.6	-/68.2	-/79.0	-/53.5	-/79.7	-/87.9
PixelBERT (R50/X152)	59.8/71.5	85.5/92.1	91.6/95.8	75.7/87	94.7/98.9	97.1/99.5	41.1/50.1	69.7/77.6	80.5/86.2	53.4/63.6	80.4/87.5	88.5/93.6
Unicoder-VL	71.5	90.9	94.9	86.2	96.3	99	46.7	76	85.3	62.3	87.1	92.8
UNITER-Base (paper/reproduction)	72.52/62.9	92.36/87.2	96.08/92.7	85.9/78.3	97.1/93.3	98.8/96.5	50.33/37.8	78.52/67.3	87.16/78.0	64.4/52.8	87.4/79.7	93.08/87.8
ViLT-DA (paper/reproduction)	62.2/62.3	87.6/87.6	93.2/93.5	83.7/82.9	97.2/98.1	98.1/98.1	42.6/42.2	72.8/73.2	83.4/84.0	62.9/62.7	87.1/87.5	92.7/93.0

Table 4: Comparisons with existing nonpretrained methods and details in image-text retrieval. The data corresponding to the column where LR located is the initial learning rate/the epoch when the learning rate changes/the Change rate. The previous addition of *each* means that the change will occur after the specified number of epochs.

Method		Flickr30k			MS-COCO			Visual Encoder	Text Encoder	Framework	Loss	Params	Cites
		Epoch	BS	LR	Epoch	BS	LR						
VSE++ [8]		30	128	0.0002/15/×0.1	30	128	0.0002/15/×0.1	CNN	GRU	a	contrastive loss	67M	610
SCAN [17]		30	128	0.0002/15/×0.1	20	128	0.0005/10/×0.1	image RoI	Bi-GRU	b	contrastive loss	9M	475
CAMP [39]		300	128	0.0002/15/×0.1	-	-	-	image RoI	Bi-GRU	c	contrastive loss	30M	86
VSRN [19]		30	128	0.0002/15/×0.1	30	128	0.0002/15/×0.1	image RoI	Bi-LSTM	a	a hinge-based triplet ranking loss, log-likelihood loss	140M	162
SGRAF [6]	SGR	40	128	0.0002/30/×0.1	20	128	0.0002/10/×0.1	image RoI	Bi-GRU	c	contrastive loss	19M	16
	SAF	30	128	0.0002/20/×0.1	20	128	0.0002/10/×0.1					18M	
SAEM [40]		30	64	0.0001/*each*10/×0.1	30	64	0.0001/*each*10/×0.1	image RoI	BERT	a	contrastive loss and angular loss	114M	40
CAMERA		30	128	0.0001/*each*10/×0.1	40	128	0.0001/*each*20/×0.1	image RoI	BERT	a	contrastive loss and diversity regularization	156M	15

Although the VLP model is hard to make a fair comparison due to the abovementioned differences in section 4.3, we attempt to obtain some insightful conclusions considering reproducibility and practical improvement by comparing several models with the most similar settings.

5.2.1 *Concerning of pre-training.*

From the perspective of pretrained data, the pretrained data of the 5 models are divided into 3 categories. PixelBERT [12] only uses data in the field such as MS-COCO and VG, while VilBERT [25] and Unicoder-VL [18] only use CC, while ViLT [16] and UNITER [4] pretrained on both in-domain and out-of-domain datasets.

As noted in Jia et al. [13], it holds true that downstream tasks such as image-text retrieval perform better with more pretraining data. This conclusion was also confirmed in the original ablation experiments of many papers, but this has provked our concern. Even if the paper provides the original pretraining code, it will not be reproduced due to lack of detail and excessive expense.

Even if researchers have the resources for reproduction, they are unwilling to devote too much money and resource consumption within a limited time and unexplained details. Instead of reproducing these pretraining models, they directly use the provided checkpoints in the pretraining stage, this is also true of small institutions and schools and independent researchers. However, it is unclear whether the provided checkpoints required some skills in the pretraining stage, whether they included the data of downstream tasks and whether they used additional manual annotation.

5.2.2 Concerning of additional settings.

Considering additional settings is a determination of whether data augmentation and hard negative sample mining are adopted.

As shown in Table 2 and Table 3, we focus on the image-text retrieval task and use the checkpoints provided by the original paper to make a certain comparison. At this time, the training rounds of most models are the same as the warmup strategy. However, the difference is the use of the technique and batch size. The main gap in batch size is the use of hard negative mining, and we analyze this factor below.

We refer to Zhang et al. [42] and add hard negative mining according to the description of ViLBERT's original paper by using the settings shown in Table 3. We train a hard negative from among the 100 closest neighbors of the target image. The results of the two datasets exceed the original paper, and some values missing from the original paper are reported.

For PixelBERT and Unicode-VL, even if they have a large number of references and great influence, we still cannot obtain comparable results. On the one hand, due to the lack of pretrained checkpoints, we cannot obtain the results of the pretraining stage. On the other hand, due to the lack of details, it is also unknown how much influences the fuzziness of training rounds and hyperparametric settings, as well as the sampling method of hard samples and data enhancement. For PixelBERT, there is a lack of training details, such as hyperparameter settings in the pretraining stage and retrieval stage. Unicode-VL adopts the hard negative method of Robinson et al. [30], but due to the lack of details, we cannot reproduce this method.

For UNITER, the hard negative mining method provides open-source code, but after practice, we find that this method is too time-consuming. In the original paper, the authors carried out the forward propagation of the network through the network model at a certain time, obtained M negative samples, and then took the most difficult N samples as the hard negative samples. On MS-COCO, its M setting is 399 and N setting is 31. Even with 16 A100 GPUs, and ignoring the time for backpropagation and sorting negative samples, forward propagation on UNITER-base (111M) is approaching 125 hours. This is impossible for researchers with limited resources to reproduce in a short time.

Therefore, we did not reproduce the hard negative mining results of UNITER. We loaded the released pretrained model of UNITER-Base and fine-tune it on MS-COCO and Flicker30K to obtain new results, as shown in Table 3.

UNITER-Base and ViLT differ only in visual-feature embedding and pretraining tasks. However, it can be seen that only part of the results of the reproduced UNITER-Base is close to ViLT, and there is a significant decline in most indicators, such as Flickr30k TR@1, TR@5 and TR@10 and the MS-COCO dataset. It can be seen that hard negative samples have a great impact on the image-text retrieval model.

Compared with ViLBERT (221M) and UNITER-Base (110M), their modal embedding and similarity modeling methods are similar, but modal interaction, the parameters, and the amount of pretrained data are different.

Due to the combination of a larger parameter quantity, modality interaction of the cross-attention mechanism and hard negative mining, even if UNITER-Base adopts more pretrained data, ViLBERT is still on many retrieval indicators (MS-COCO in R@1 , 5 and 10) that slightly exceed the UNITER-Base. *This leads us to consider whether the modal interaction style and the use of hard samples are also key factors in performance improvement.*

For ViLT, we almost reported the value close to the original paper by loading the pretrained checkpoints provided by authors, but the training went through longer training rounds due to the uncertainty enhanced by random data augmentation.

It was found that the improvement of these VLP models may not only come from the pretraining and architecture design but also have their own tricks and unknown details.

Table 5: Comparisons with existing nonpretrained methods and their results in image-text retrieval on Flicker30K.The three values of */*/* in the table represent the results provided by the paper/reproduced results/results of removing difficult samples.

Method		Flickr30k (1K)					
		IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
VSE++ [8]		43.7/44.4/31.7	71.9/73.1/61.4	82.1/83.1/73.1	32.3/32.6/26.2	60.9/61.2/57.4	72.1/79.5/70.1
SCAN [17]	SCAN t-i LSE	61.1/58.9/45.4	85.4/85.5/78.9	91.5/91.5/87.7	43.3/41.3/35	71.9/69.8/65.1	80.9/79.4/76.4
	SCAN t-i AVG	61.8/63.0/47.5	87.5/88.3/80.2	93.7/93.9/89.0	45.8/44.5/35.8	74.4/73.9/67.0	83.0/81.7/77.8
	SCAN i-t LSE	67.7/66.3/46.5	88.9/88.4/77	94/93.7/86.2	44.0/42.0/33.9	74.2/72.2/64.0	82.6/81.1/74.8
	SCAN i-t AVG	67.9/67.7/46	89/88.7/77.5	94.4/94.6/87	43.9/44.5/34.1	74.2/73.5/65.9	82.8/82.3/76.4
CAMP [39]		68.1/31.9/20.4	89.7/59.9/45.4	95.2/73.0/58.3	51.5/24.7/14.7	77.1/51.0/38.1	85.3/63.0/58.0
VSRN [19]		71.3/67.7/54.7	90.6/88.2/82.5	96/93.8/90.2	54.7/49.1/40.6	81.8/75.8/72.0	88.2/84.1/81.5
SGRAF [6]	SAF	73.7/73.5/74.5	93.3/90.5/92.5	96.3/95.5/96.8	56.1/53.1/56.9	81.5/78.7/82.4	88.0/85.4/89.1
	SGR	75.2/74.6/73.4	93.3/93.1/93.1	96.6/96.5/97.2	56.2/56.1/54.9	81.0/80.4/81.4	86.5/87.3/88.1
SAEM [40]		69.1/50.4/40.3	91.0/76.5/70.4	95.1/85.7/79.3	52.4/34.1/28.5	81.1/62.4/57.6	88.1/72.3/68.5
CAMERA [28]		76.5/38.8/50.1	95.1/61.6/79.2	97.2/71.6/87.8	58.9/23.7/35.5	84.7/47.9/65.7	90.2/56.9/76.4

5.3 NonPretrained Models

Compared with pretrained models for image-text retrieval, nonpretrained models tend to design complex modality interactions. A specially designed alignment architecture can better refer to the information of the two modalities. Moreover, the improvement of the nonpretrained image-text retrieval methods also provided many inspirations for VLPs. To drive the development of the image-text retrieval community, the nonpretrained method is also an important factor because of lower calculation consumption with fewer parameters. By gathering information from a wealth of related papers, we found that few methods were reproduced, and the results in the

Table 6: Comparisons with existing nonpretrained methods and their results in image-text retrieval on MS-COCO. The three values of */*/* in the table represent the results provided by the paper/reproduced results/results of removing difficult samples. † represents methods that can not reproduce results due to lack of code and training details.

Method		MS-COCO (5K)						MS-COCO (1K)					
		IR@1	IR@5	IR@10	TR@1	TR@5	TR@10	IR@1	IR@5	IR@10	TR@1	TR@5	TR@10
VSE++ [8]		49/44.8/21.3	79.8/75.2/45.9	88.4/84.6/59.1	37.1/32.6/16.5	72.2/66.7/42.4	83.8/79.1/56.7	-/67.8/35.4	-/90.9/65.2	-/96.1/75.9	-/56.9/25.2	-/87.6/54.9	-/93.2/67.4
SCAN [17]	SCAN t-i LSE	-/39.1/28.5	-/71/59.0	-/82.5/73.4	-/27.3/20.0	-/56.6/46.9	-/69.6/61.3	67.5/64.4/54.4	92.9/91.9/86.0	97.6/96.9/94.5	53/49.5/40.7	85.4/83/76.1	92.9/91.2/86.7
	SCAN t-i AVG	-/45.1/30.4	-/75.7/62.5	-/86.4/76.5	-/32.9/23.9	-/62.2/51.7	-/74.4/65.1	70.9/69.4/57.1	94.5/93.8/88.6	97.8/97.5/95.8	56.4/54.9/45.4	87/85.8/79	93.9/93.2/88.5
	SCAN i-t LSE	46.4/41.0/23.3	77.4/73.6/54.4	87.2/84.5/68.9	34.4/30.6/17.2	63.7/60.9/44.5	75.7/73.5/59.2	68.4/65.9/49.0	93.9/93.3/83.0	98.0/98.0/92.9	54.8/53.1/38.1	86.1/85.2/75.4	93.3/92.7/87.3
	SCAN i-t AVG	-/43.4/30.4	-/75.0/62.7	-/86.6/76.6	-/32.9/22.0	-/62.8/51.7	-/75/65.7	69.2/68.1/55.9	93.2/93.7/88.8	97.5/97.5/95.7	54.4/55.5/44.5	86/86.3/79.6	93.6/93.5/89.5
CAMP† [39]		50.1/-/-	82.1/-/-	89.7/-/-	39/-/-	68.9/-/-	80.2/-/-	72.3/-/-	94.8/-/-	98.3/-/-	58.5/-/-	87.9/-/-	95/-/-
VSRN [19]		53/48/41.3	81.1/77.6/73.4	89.4/87.3/84.4	40.5/35.9/31.8	70.6/66.4/63.4	70.6/77.7/75.7	76.2/71.1/66.7	94.8/94.1/92.4	98.2/97.6/96.8	62.8/58.5/55.4	89.7/87.2/86.4	95.1/93.5/92.9
SGRAF [6]	SAF	53.3/54.8/51.9	-/82.4/81.6	90.1/90.4/89.8	39.8/38.8/38.7	-/67.7/68.1	80.2/79/79.3	76.1/75.9/75.5	95.4/95.5/95.3	98.3/98.3/98.1	61.8/60.5/60.7	89.4/88.5/89.0	95.3/94.7/95.1
	SGR	56.9/55.1/51.3	-/82.7/81.2	90.5/90.7/89.5	40.2/39.1/38.7	-/68.5/68.1	79.8/79.5/79.3	78/76.6/74.9	95.8/95.8/95.5	98.2/98.4/98.2	61.4/61.0/60.3	89.3/89.2/89.0	95.4/95.1/94.9
SAEM [40]†		-/34.6/28.8	-/63.9/57.6	-/76.5/70.4	-/25/20.8	-/53.3/47.0	-/66.9/60.9	71.2/-/-	94.1/-/-	97.7/-/-	57.8/-/-	88.6/-/-	94.9/-/-
CAMERA [28]		53.1/38.3/32.0	81.3/68.1/62.4	89.8/80/75.2	39.0/27.5/24.4	70.5/57.3/53.4	81.5/69.9/66.8	75.9/62.9/57.4	95.5/89.2/87.0	98.6/95.1/93.8	62.3/49.7/64.4	90.1/82.1/81.4	95.2/90.8/90.8

original papers were directly cited. In this way, it is difficult for the community to know what lessons from prior work held up, and it is difficult for future researchers to obtain better ideas for them. We compare some existing, widely cited nonpretrained image-text matching models to show factors that affect the performance of nonpretrained methods. Table 4 shows some settings, losses and parameter number of those models.

5.3.1 About the experiment and code.

Reproducing the paper results is not a trivial task, especially for the image-text retrieval task. We found a lot of noteworthy points, which have great implications for the development of the image retrieval community. In the code of the papers we collected, most of them were built under the older torch framework version and python 2. We simply modified the code without changing the function of the program, so that the code can run on Tesla A100 with torch1.8.0 and CUDA11.

SCAN [17], SGRAF [6], VSRN [19], CAMERA, and SAEM [40] methods provide accurate training sentences, which we reproduce relatively easily. However, in the Github issues, some researchers raised the problem that the recurrence result of SAEM [40] is too low, which is consistent with our confusion. In addition, SAEM [40] does not provide a test code on MS-COCO 1,000 test.

The VSE++ [8] contains many tricks, including using the validation set for training, using a single random crop or center crop when preprocessing images, and whether to finetune the image feature extraction part. Some of the methods do not provide the complete results on Flickr30k and MS-COCO. We removed all tricks, fixed the feature extractor, used only training set images for training and adopted random cropping.

CAMP [39] only provides detailed parameters of Flickr30k, which makes it impossible for us to reproduce the results on the MS-COCO dataset. Moreover, the author provides a two-stage training process. Consistent with the issue mentioned in the Github issue, we were also unable to reproduce the valuable second-stage results, so we only provide the first-stage results for comparison.

We also found that in the five-fold cross-validation of the MS-COCO 1,000 test, the selection of data was not random. This also reduces the credibility of all methods. Therefore, a unified code framework and reasonable testing methods are some of the important factors to promote the orderly development of the image-text retrieval community. This helps to rule out some ineffective methods and is more conducive to future research work for researchers.

5.3.2 Stability analysis.

The better modal interactions claimed by these nonpretrained models are not true, and the real reason may be from parameter tuning.

Research that advances the development of image-text retrieval should have sufficient stability. This means that existing methods can more easily be reproduced and improved upon by future researchers. Because many top conferences and journals have no requirements with source code, we collected some methods for providing source code and reproduced them. To our surprise, even if the source code method is provided, it is difficult to reproduce these methods. The reproduced results are shown in Table 5 and Table 5. We show the results presented in the paper, our reproduced results, and the results after removing the hard sample method.

The method proposed in the paper has low robustness. The lack of random seeds in the code made it harder to reproduce, so we resurfaced the code after adding random seeds. VSE++ [8], SCAN [17], SGRAF [6] obtained similar results as in the original paper, but CAMP [39], SAEM [40], CAMERA obtained results much lower than those obtained in the original paper. This shows that the stability of the latter improved method is worse. We suspect that the authors have used the best value among multiple results as the presentation result and used the ensemble method. In addition, this also shows that many improvement methods in the image and text retrieval community are not effective. And to a certain extent, the redundancy of the original model is increased. For example, CAMERA was proposed in 2020, and the author mentioned in the README.md file in the code that the method is based on VSRN [19] and SAEM [40] proposed in 2019. Disappointingly, our replicated results show that CAMERA [28] is the worst performer.

Compared with the performance on the Flickr30k dataset, most methods have a more stable performance on the MS-COCO dataset. And this basically corresponds to the performance on Flickr30k. We guess this is due to the fact that the latter has far more training data than the former. This is not bad news. It shows from the side that *data is still a focus of image and text retrieval tasks, and more data helps the model learn a distribution closer to the overall data.* Because of this, we conjecture that the model's learning of the data distribution may have a large impact on the model's performance. We, therefore, supplemented the experiments to remove hard samples.

The image-text retrieval community should pay more attention to evaluation metrics. And more experimental results should be provided. On Flickr30k, we rerun several more experiments on

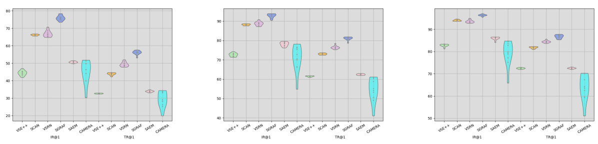

Figure 2: Fine-tuning variance in nonpretrained models on Flickr30k. Each model is fine-tuned 10 times with different random seeds. For SCAN/SGRAF, we chose the i-t ANG/SGR result to show.

several methods. It can be clearly seen that the fluctuation range of the experimental results is particularly large, even exceeding the number of points for improvement. The result is shown in Figure 2.

We also have a more interesting finding that the interaction between modalities is more conducive to the stability of the model. With the exception of CAMP [39], which we cannot accurately reproduce, most methods for stabilizing aberration are the structure shown in Figure1 (a). We are sure that the authors of the paper got the results they gave. Methods with modal interactions reproduce results closer to these results. This suggests that researchers in the image retrieval community can focus more on the fusion of vision and text information. This would also greatly contribute to methods using VLP.

5.3.3 Impact of hard samples.

We found that in all methods, hard samples were used to obtain better results. We try to remove hard samples to explore the role of hard samples in image-text retrieval. Table 5 and Table 6 shows the result after removing the hard samples.

The main purpose of adding hard samples to the loss function is to make the spatial distribution of hard samples farther away from the correct samples and obtain a more sparse spatial distribution. In this way, when the image/text features are mapped to the same spatial distribution, the surrounding negative samples can be reduced. From the results presented in Table 4, the results of all methods drop significantly with removing the hard items from the loss. Therefore, we believe that when image features and text features cannot be sufficiently aligned, a reasonable control over the spatial distribution of features can maximize the performance improvement on image-text retrieval tasks. This encourages researchers to pay more attention to hard samples and the spatial distribution of samples.

6 CONJECTURES AND RECOMMENDATIONS

As discussed above, we were surprised to find that so few of the architectural modifications of modality interaction and similarity modeling produced improvements even if we use approximately the same settings of the original paper but removed some tricks and used different random seeds. There are various possible explanations as to why our results bore out the way they did:

1. Not tuning hyperparameters handicapped other methods. In our reproduction, we found that the random seed has a very large influence on the experimental results and can even differ by 10 points on IR@1. Therefore these modification methods are not reasonably hyperparameter-agnostic and stable with their modification of the modality interaction. Furthermore, if parameter sensitivity is the key to the problem, then taking a good initialization as the final result does not represent the contribution of one's own method.

2. Modality embedding is important. In conclusion, more fine-grained modal coding and a stronger modal encoder will bring stable performance improvement for image-text retrieval.

3. Modality interaction methods may be critical. For many VLPs, we can use most of their methods as simple transformers as modal interactions. Although the cross-attention mechanism of ViLBERT has a marginal performance improvement, it introduces double the parameters, resulting in greater computational consumption. For many nonpretrained models, most put the direction of improvement on modal interaction. To our surprise, in addition to improving the performance of the model to a certain extent, the multimodal interaction can also make the model more stable.

4. Differences in training data. In the VLP results, we found that the addition of data can significantly improve the final results but has worse reproducibility. Meanwhile, when we reproduced the nonpretrained model, the original text of the specific operation of data enhancement and the description of the code is too vague, so the data comparison is not necessarily carried out under the exact same settings. Moreover, we found that for the larger dataset MS-COCO and the harder 5K test, the results of each method run are more stable. In the process of reproduction, we did not use the validation set data, so this quantitative data may have caused a certain degree of decline in the reproduction results.

5. How to make better use of training samples is also the key. Hard samples are widely used in both VLPs and nonpretrained models. Through ablation experiments on hard negative mining in VLPs experiments, we found that more hard samples can greatly enhance the retrieval results. In the nonpretrained results, we found that the use of the most difficult samples was not the same as that of VLPs. These methods only use the most similar negative samples and one positive sample in a batch for optimization, resulting in less utilization of the characteristics of the data. Additionally, due to the parameter sensitivity of these methods, under different random seeds, the model similarity modeling ability is different. This in turn magnifies the effect of hard samples.

Given this sober take, we propose some suggestions to improve the robustness and generalizability of future image-text retrieval research. First, when proposing a new method, the random seed used and the results of multiple runs and specific details should be given, not just a nonreproducible code and some vague expressions like Figure 2. Second, we should not pay too much attention to the tuning of models and parameters, and we should focus on the characteristics of the mining data. Hard samples can bring a more stable improvement to the model [30], but such methods are rarely used in the field of image-text retrieval. Based on the findings of this recurring result, we believe that in the future, we should focus on such methods of stable improvement rather than those that need to be sensitive to parameters. Finally, best-practice results reporting should include the mean and standard deviation across numerous trials or at the very least avoid cherry-picking the best run [26]. We hope that future work in the field of information retrieval should pay more attention to the reproducibilities and capabilities of the model, rather than tuning parameters and stacking tricks for better performance.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*. Computer Vision Foundation / IEEE Computer Society, 6077–6086.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*.

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *ECCV (30) (Lecture Notes in Computer Science, Vol. 12375)*. Springer, 104–120.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*. Association for Computational Linguistics, 4171–4186.

[6] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity Reasoning and Filtration for Image-Text Matching. In *AAAI*. AAAI Press, 1218–1226.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net.

[8] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *BMVC*. BMVA Press, 12.

[9] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. 2020. FashionBERT: Text and Image Matching with Adaptive Loss for Cross-modal Retrieval. In *SIGIR*. ACM, 2251–2260.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. IEEE Computer Society, 770–778.

[11] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. 2019. Scalable Deep Multimodal Learning for Cross-Modal Retrieval. In *SIGIR*. ACM, 635–644.

[12] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers. *CoRR* abs/2004.00849 (2020).

[13] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *ICML (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 4904–4916.

[14] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik G. Learned-Miller, and Xinlei Chen. 2020. In Defense of Grid Features for Visual Question Answering. In *CVPR*. Computer Vision Foundation / IEEE, 10264–10273.

[15] Andrej Karpathy and Li Fei-Fei. 2017. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 4 (2017), 664–676.

[16] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *ICML*.

[17] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *ECCV*. 201–216.

[18] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training. In *AAAI*. AAAI Press, 11336–11344.

[19] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual Semantic Reasoning for Image-Text Matching. In *ICCV*. IEEE, 4653–4661.

[20] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *CoRR* abs/1908.03557 (2019).

[21] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *ECCV (30) (Lecture Notes in Computer Science, Vol. 12375)*. Springer, 121–137.

[22] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. 2017. Feature Pyramid Networks for Object Detection. In *CVPR*. IEEE Computer Society, 936–944.

[23] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*, Vol. 8693. Springer, 740–755.

[24] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. 2020. Graph Structured Network for Image-Text Matching. In *CVPR*. Computer Vision Foundation / IEEE, 10918–10927.

[25] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*.

[26] Sharan Narang, Hyung Won Chung, Yi Tay, Liam Fedus, Thibault Févry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, Yanqi Zhou, Wei Li, Nan Ding, Jake Marcus, Adam Roberts, and Colin Raffel. 2021. Do Transformer Modifications Transfer Across Implementations and Applications?. In *EMNLP (1)*. Association for Computational Linguistics, 5758–5773.

[27] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *NIPS*. 1143–1151.

[28] Leigang Qu, Meng Liu, Da Cao, Liqiang Nie, and Qi Tian. 2020. Context-Aware Multi-View Summarization Network for Image-Text Matching. In *ACM Multimedia*. ACM, 1047–1055.

[29] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 6 (2017), 1137–1149.

[30] Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive Learning with Hard Negative Samples. In *ICLR*. OpenReview.net.

[31] Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* 45, 11 (1997), 2673–2681.

[32] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *ACL (1)*. Association for Computational Linguistics, 2556–2565.

[33] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *ICLR*. OpenReview.net.

[34] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP*. Association for Computational Linguistics, 5099–5110.

[35] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR* abs/1807.03748 (2018).

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NeurIPS*.

[37] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *ICLR (Poster)*. OpenReview.net.

[38] Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. 2017. Deep Metric Learning with Angular Loss. In *ICCV*. IEEE Computer Society, 2612–2620.

[39] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2019. CAMP: Cross-Modal Adaptive Message Passing for Text-Image Retrieval. In *The IEEE International Conference on Computer Vision (ICCV)*.

[40] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. 2019. Learning Fragment Self-Attention Embeddings for Image-Text Matching. In *ACM Multimedia*. ACM, 2088–2096.

[41] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguistics* 2 (2014), 67–78.

[42] Bowen Zhang, Hexiang Hu, Vihan Jain, Eugene Ie, and Fei Sha. 2020. Learning to Represent Image and Text with Denotation Graph. In *EMNLP (1)*. Association for Computational Linguistics, 823–839.

[43] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z. Li. 2020. Context-Aware Attention Network for Image-Text Retrieval. In *CVPR*. IEEE, 3533–3542.