MAFA: Managing False Negatives for Vision-Language Pre-training

Jaeseok Byun[1*] Dohoon Kim[1*] Taesup Moon[1,2†]
[1] Department of ECE, Seoul National University
[2] Department of ASRI/INMC/IPAI/AIIS, Seoul National University
{wotjr3868, dohoon.kim, tsmoon}@snu.ac.kr

Abstract

*We consider a critical issue of false negatives in Vision-Language Pre-training (VLP), a challenge that arises from the inherent many-to-many correspondence of image-text pairs in large-scale web-crawled datasets. The presence of false negatives can impede achieving optimal performance and even lead to a significant performance drop. To address this challenge, we propose MAFA (**MA**naging **FA**lse negatives), which consists of two pivotal components building upon the recently developed GRouped mIni-baTch sampling (GRIT) strategy: 1) an efficient connection mining process that identifies and converts false negatives into positives, and 2) label smoothing for the image-text contrastive (ITC) loss. Our comprehensive experiments verify the effectiveness of MAFA across multiple downstream tasks, emphasizing the crucial role of addressing false negatives in VLP, potentially even surpassing the importance of addressing false positives. In addition, the compatibility of MAFA with the recent BLIP-family model is also demonstrated. Code is available at https://github.com/jaeseokbyun/MAFA.*

1. Introduction

With large-scale web-crawled datasets [3, 50–52], majorities of vision-language pre-training (VLP) models are trained in a self-supervised learning manner using the combination of several pre-tasks and losses [2, 33, 34, 63, 65]: *e.g.,* masked language modeling (MLM), image-text contrastive (ITC), and image-text matching (ITM) losses. Despite their promising results, one of the pressing challenges they face is the presence of *noisy* captions in image-text pairs that often provide incomplete or even incorrect descriptions [9, 13, 41, 44, 47, 59, 64]. Several recent works have focused on addressing such issue of noisy correspondence in image-text pairs [11, 18, 19, 21, 34, 47]. Notably, BLIP [34] introduced a caption refinement process by leveraging an image captioning model and a filter to generate synthetic clean captions and remove noisy captions. Such process can be seen as correcting the *false positives* that were injected by the noisy captions.

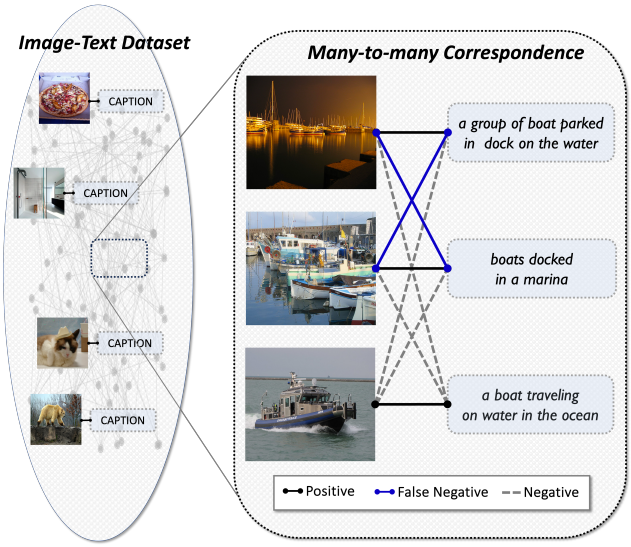

Figure 1. Examples of positives, negatives, and false negatives among image-text pairs.

Contrastively, we note that there is another type of challenge for VLP that stems from the nature of many-to-many correspondence of image-text pairs. Namely, it is common for an image (resp. text) to have *additional* positive connections (blue lines in Figure 1) with another texts (resp. images), which are paired with their corresponding images (resp. texts). This is due to the fact that the existing image-text datasets are constructed by only collecting paired image-text instances, hence the information regarding non-paired but semantically close image-text combination can be missed. Consequently, for each image (resp. text), the text (resp. image) that is given as the pair with the image (resp. text) is treated as the only positive sample during pre-training, while the other texts (resp. images) are all treated as negatives. This setup inevitably leads to the prevalence of *false negatives* during computing ITC and ITM losses and con-

*Equal contribution
†Corresponding author

fuses the learning process. A naive solution would be to identify missing positive connections by examining all possible image-text combinations in the dataset. However, it is cleaerly infeasible for both manual and model-based evaluations due to prohibitive complexity. For example, even for a dataset of moderate size, *e.g.,* containing 5 million image-text pairs, the number of combinations that need to be examined is $\binom{5M}{2}$, which amounts to approximately 12 trillion.

We note such issue of false negatives has been more or less overlooked in recent studies [2, 34, 63], which incorporated the in-batch hard negative sampling for the ITM task as a standard tool for VLP. This sampling technique, initially proposed in ALBEF [33], involves selecting hard negative samples from the mini-batch based on the image-text similarity scores computed from the ITC task. More recently, GRIT-VLP [2] significantly improved the performance by proposing an improved hard negative sampling by first grouping the similar image-text pairs together before forming a mini-batch. Ideally, if all semantically close image-text pairs were correctly labeled, hard negative mining could effectively identify only informative hard negatives. However, in a typical VLP setting where such information is absent, the hard negatives in fact frequently become false negatives, resulting in sub-optimal model performance.

To address this challenge of false negatives, particularly prevalent when hard negative sampling is in action, we have implemented two significant enhancements. Firstly, we devise an Efficient Connection Mining (ECM) process that identifies missing positive connections between non-paired but semantically close images and texts. Rather than reviewing all possible combinations, ECM strategically extracts the plausible candidates which are selected as hard negatives. These candidates are inspected by a pre-trained discriminator, which determines their potential to be converted into positives. The candidates identified as positives by the discriminator are then incorporated as additional positives for calculating ITC, ITM, and MLM losses during the training process. Secondly, we introduce Smoothed ITC (S-ITC) which is based on the principle of label smoothing [56]. This approach is specifically designed to mitigate the over-penalization of false negative samples within grouped mini-batches, without incurring any additional memory or computational overhead.

Our experimental results demonstrate that the proposed method, dubbed as MAFA (**MA**naging **FA**lse negatives), can substantially improve the VLP performance. For example, a model trained with MAFA on a standard 4M dataset (*i.e.,* 4M-Noisy) can almost achieve the performance of a baseline model trained on a much larger 14M dataset, without exploiting any additional information such as bounding boxes or object tags. Our systematic ablation analyses demonstrate that such performance enhancement primarily results from mitigating effect of false negatives. Another finding from our experiments is that converting false negatives into additional positives is more advantageous than merely eliminating them. Moreover, we also demonstrate that the impact of addressing the false negative issue is orthogonal to and may outweigh that of addressing the false positive issue in VLP, which is done by comparing and combining MAFA with the BLIP [34] framework. Finally, we show MAFA is also compatible with recently proposed BLIP-2 [35], underscoring the generality of our method in VLP.

2. Related Work

Vision-language pre-training (VLP). Initial VLP models [5, 6, 14, 20, 22, 29, 37, 39, 40, 54, 60, 66] which utilized a single multi-modal encoder, primarily employed random negative sampling during the ITM task. Recently, ALBEF [33] incorporated the ITC loss and in-batch hard negative sampling strategy for ITM by leveraging image-text contrastive similarity scores. Subsequently, the in-batch hard negative sampling strategy for ITM became an implicit rule for the BLIP-family models [1, 2, 26, 27, 34, 35, 63, 65] which adopt both ITC and ITM as training objectives. While the significance of hard negative sampling for the ITM task has been highlighted in GRIT-VLP [2], limited attention has been given to addressing the issue of false negatives arising from the hard negative samples. Existing studies have primarily focused on tackling false negatives only in the context of contrastive learning [2, 8, 33, 53] with particular emphasis on the vision domain [4, 7, 24, 49]. To that end, we highlight the need for effective strategies to address false negatives in VLP and demonstrate that false negatives can be managed.

Label smoothing. Label smoothing [56] is a widely adopted technique for improving generalization in various classification tasks. It converts the one-hot target labels into soft labels by mixing them with uniform distribution. This simple technique has demonstrated its efficacy in both visual [31, 36, 43] and language domains [15, 32]. Its benefits have also led to its incorporation as a supplementary technique to enhance the fine-tuning of image-text models like CLIP [48] for image classification tasks [16, 25, 61]. However, the application of label smoothing within VLP and its ability to address false negatives have not been thoroughly explored. Recently, some studies [2, 33] have introduced model-generated soft labels in the VLP domain. However, we show that such soft labels are insufficient for effectively addressing false negatives, justifying the need to incorporate label smoothing to the contrastive loss when the hard negative sampling is employed.

2.1. GRIT-VLP [2]

GRIT-VLP uses ITC, ITM with in-batch hard negative sampling, and MLM as the objectives proposed in ALBEF, ex-

cept the utilization of the momentum encoder in the pre-training. However, it significantly extends ALBEF by implementing two key components as follows:

(a) GRouped mIni-baTch (GRIT) sampling aims to construct mini-batches containing highly similar example groups. This facilitates the selection of informative hard negative samples during in-batch hard negative sampling. To avoid excessive memory or computational overhead, the procedure of constructing grouped mini-batches for the next epoch is performed concurrently with the loss calculation at each epoch. For this, additional queues are used to collect and search for the most similar examples, one by one, in which the similarity is measured by the ITC scores. These queues serve as the search space and are significantly larger than the mini-batch size B. Thus, the size of the queue, denoted as search space M, controls the level of hardness in selecting hard negative samples.

(b) ITC with consistency loss attempts to address the issue of over-penalization in ITC that arises when GRIT is combined. Contrary to ALBEF, similar examples are gathered in the GRIT-sampled mini-batch. Thus, when one-hot labels are used for ITC, they result in equal penalization of all negatives, and it has been observed that the representations of similar samples may unintentionally drift apart. To mitigate this, GRIT-VLP incorporates soft pseudo-targets generated from the same pre-trained model as a mean of regularization.

3. Motivation

In order to quantify the tendency of the number of false negatives in the ITM task, we report a quantitative analysis result in Table 1. We estimated the number of false negative pairs during a single epoch while training the ITM task, employing two distinct mini-batch sampling strategies: random sampling and GRIT sampling. These strategies were evaluated on both the original 4M dataset (4M-Noisy) and the BLIP-generated clean dataset (4M-Clean). Given the infeasibility of manually examining every negative pair to determine whether it is a false negative, we utilized a strong ITM model pre-trained on a large-scale 129M dataset from BLIP — *i.e.,* a negative pair is regarded as *false* negative if the strong ITM model predicts it is "matched". While the ITM model does not always classify false negatives with perfect accuracy, its reliability is deemed adequate for approximating the trend in false negative counts. More details and analyses regarding counting the number of false negatives can be found in the Supplementary Material (S.M).

From the table, we observe that GRIT sampling exhibits significantly more false negatives than random sampling, as mentioned in the Introduction. The reason is that GRIT sampling generates challenging in-batch hard negatives, which are beneficial for learning fine-grained representations, but they also often end up being false negatives. Moreover, we observe this trend exacerbates in the 4M-Clean dataset.

Table 1. Estimated number of false negatives (FN) for random sampling and GRIT sampling. The FNs are counted for each anchor image and text separately. The ratio (%) represents the estimated proportion of FNs with respect to the total number of negative pairs used in ITM during a single epoch. We set the batch size B as 96 for both samplings and $M = 4800$ for GRIT sampling.

Dataset	Sampling	FN w.r.t. image	FN w.r.t. text
4M-Noisy	Random	127,130 (2.5%)	118,080 (2.3%)
	GRIT	817,991 (16.4%)	8111,145 (16.2%)
4M-Clean	Random	153,006 (3.1%)	148,729 (3.0%)
	GRIT	1,114,851 (23.2%)	1,096,485 (22.2%)

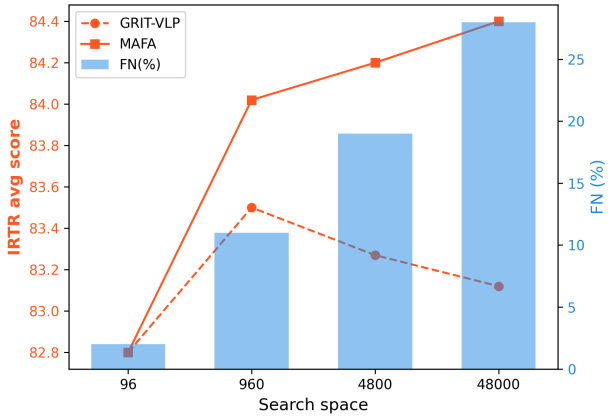

Figure 2. Comparison of IRTR average scores and false negatives (%) in the 4M-Noisy dataset for GRIT-VLP and MAFA across different search spaces (M) when applying GRIT sampling. Here, IRTR average score is defined as the average image-text retrieval accuracy across (TR/R@1, TR/R@5, TR/R@10, IR/R@1, IR/R@5, IR/R@10) on COCO 5k test set. For all models, we set the batch size B as 96. Thus, when $M = 96$, GRIT sampling becomes equivalent to random sampling.

In Figure 2, we examine the impact of an increasing number of false negatives on the downstream performance of GRIT-VLP. Specifically, we measured the average IRTR score of GRIT-VLP across different search space (queue) sizes M, while keeping the batch size B constant. We first clearly observe that the number of false negatives rises as M increases. This is expected since expanding the search space for GRIT sampling leads to more similar examples being grouped together in a mini-batch, thereby generating more false negatives. In terms of the downstream performance of GRIT-VLP, we notice a decline when the value of M exceeds a certain threshold ($M = 960$). We attribute this decline to the introduction of "noise" caused by the increasing presence of false negatives, which subsequently hampers the effectiveness of hard negative sampling in GRIT-VLP. Based on this analysis, we anticipate that effectively addressing the issue of false negatives while leveraging the potential of hard negative samples will be crucial for enhancing VLP models even further. In S.M, we further explore the impact of varying batch sizes on the occurrence of false negatives

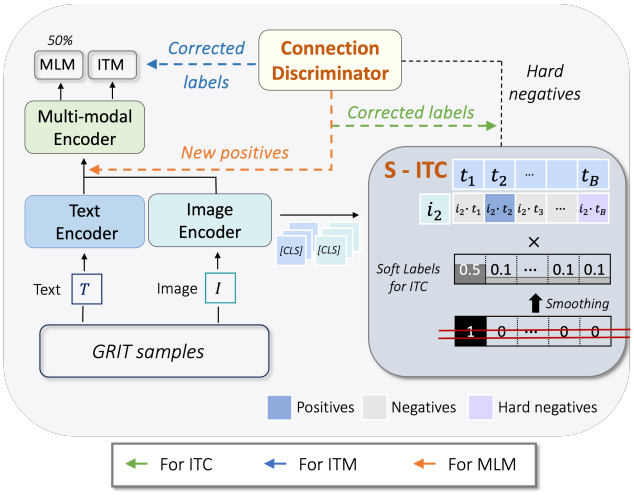

Figure 3. Overall framework of MAFA.

under the random sampling scenario, which highlights the significance of handling false negatives even in the typical VLP setting (large batch size under random sampling).

In response to this challenge, we propose MAFA, which effectively addresses the issue of false negatives and improves the downstream performance significantly compared to GRIT-VLP. The preview of the performance of MAFA is also shown in Figure 2 — it is evident that the IRTR score of MAFA continues to improve despite an increase in the number of false negatives.

4. Main Method: MAFA

Our MAFA consists of two integral components: we first present the intuition and details of the Efficient Connection Mining (ECM), then delineate the Smoothed ITC (S-ITC).

4.1. Efficient Connection Mining (ECM)

As outlined in the Introduction, the issue of false negatives originates from *missing* positive connections within a paired dataset, a challenge that is computationally infeasible to address naively. To tackle this, ECM process is strategically designed to exclusively examine hard negatives with significantly higher likelihoods of being false negatives. Namely, as described in Figure 4 and Algorithm 1 in S.M, the training model selects the hardest negatives for each anchor based on ITC similarity in the GRIT-sampled mini-batches. Once hard negatives are selected, a separate pre-trained ITM model, *Connection Discriminator (Con-D)*, is employed to determine whether these hard negatives are true (hard) negatives or false negatives. If *Con-D* assigns a probability to a candidate image-text pair of being positive higher than a threshold τ (which was set to 0.8), that candidate is adopted as a new positive pair in the pre-training losses.

We note that ECM process can be seamlessly integrated with the training process of BLIP-family models (ALBEF [33], BLIP [34], BLIP-2 [35]) as well by adopting GRIT sampling and *Con-D*, given that ITC and ITM are used as their training objectives. Moreover, due to the inherent randomness of mini-batches, *Con-D* encounters a variety of hard negatives in each batch and epoch, which enables ECM to create diverse positive connections during training.

Now, we will elaborate on the details of the three pre-training losses (ITC, ITM, and MLM) utilized in our model, and then explain how false negatives identified by ECM are integrated into these losses. Briefly, for ITC and ITM, the labels for identified false negatives are revised from negatives to positives. Moreover, these new positives are additionally used as inputs for MLM.

[ITC with ECM] In ITC, to measure the similarity between images and texts, the [CLS] tokens from the unimodal encoders are utilized, as illustrated in Figure 3. We denote the cosine similarity between image i and text t as $s(i,t) = g_{\mathrm{I}}(i^{cls})^T g_{\mathrm{T}}(t^{cls})$, where $g_{\mathrm{I}}(\cdot)$ and $g_{\mathrm{T}}(\cdot)$ are linear projections for [CLS] tokens of image and text embeddings, respectively. The objective of ITC is to maximize the similarity of positive pairs while minimizing that of negative pairs; hence, the loss becomes

$$\mathcal{L}_{\mathrm{ITC}} = \frac{1}{2}\mathbb{E}_{(i,t)\sim D}\Big[\mathrm{CE}\Big(\boldsymbol{y}^{\mathrm{I2T}}(i), \boldsymbol{p}^{\mathrm{I2T}}(i)\Big) + \mathrm{CE}\Big(\boldsymbol{y}^{\mathrm{T2I}}(t), \boldsymbol{p}^{\mathrm{T2I}}(t)\Big)\Big], \quad (1)$$

in which $\boldsymbol{y}^{\mathrm{I2T}}(i)$ and $\boldsymbol{y}^{\mathrm{T2I}}(t)$ stand for the one-hot vectors for the correct sample pairs for image i and text t, respectively. $\mathrm{CE}(\cdot)$ denotes the cross-entropy loss. The softmax-normalized image-to-text and text-to-image similarities between image i and text t, $\boldsymbol{p}_t^{\mathrm{I2T}}(i)$ and $\boldsymbol{p}_i^{\mathrm{T2I}}(t)$, are defined as

$$\boldsymbol{p}_t^{\mathrm{I2T}}(i) = \frac{e^{s(i,t)/\tau}}{\sum_{k=1}^{N} e^{s(i,t_k)/\tau}}, \boldsymbol{p}_i^{\mathrm{T2I}}(t) = \frac{e^{s(i,t)/\tau}}{\sum_{k=1}^{N} e^{s(i_k,t)/\tau}}, \quad (2)$$

in which τ is the temperature and N is the number of considered texts and images.

To incorporate missing positive connections identified by *Con-D*, the one-hot label $\boldsymbol{y}$ is adjusted to $\tilde{\boldsymbol{y}}^{\mathrm{ITC}}$. For example, for an anchor image i, a single text t_k is picked by in-batch hard negative sampling. Then, if t_k is recognized as a new positive by *Con-D*, the k-th element of the one-hot vector $\boldsymbol{y}^{\mathrm{I2T}}(i)$ changes from 0 to 0.5. Simultaneously, the original label value of 1 in $\boldsymbol{y}^{\mathrm{I2T}}(i)$ becomes 0.5, ensuring that the sum remains 1. If a positive connection is not newly established, the label remains unchanged. The above process is applied identically for text t and image i_k. Now, we denote the new ITC loss equipped with $\tilde{\boldsymbol{y}}^{\mathrm{ITC}}$ as $\mathcal{L}_{\mathrm{ITC}}^{ECM}$.

[ITM with ECM] ITM task aims to predict whether the provided pair is matched or not. Similar to ITC, the labels for ITM are revised based on the missing positive connections identified by *Con-D*. We employ a re-sampling strategy for

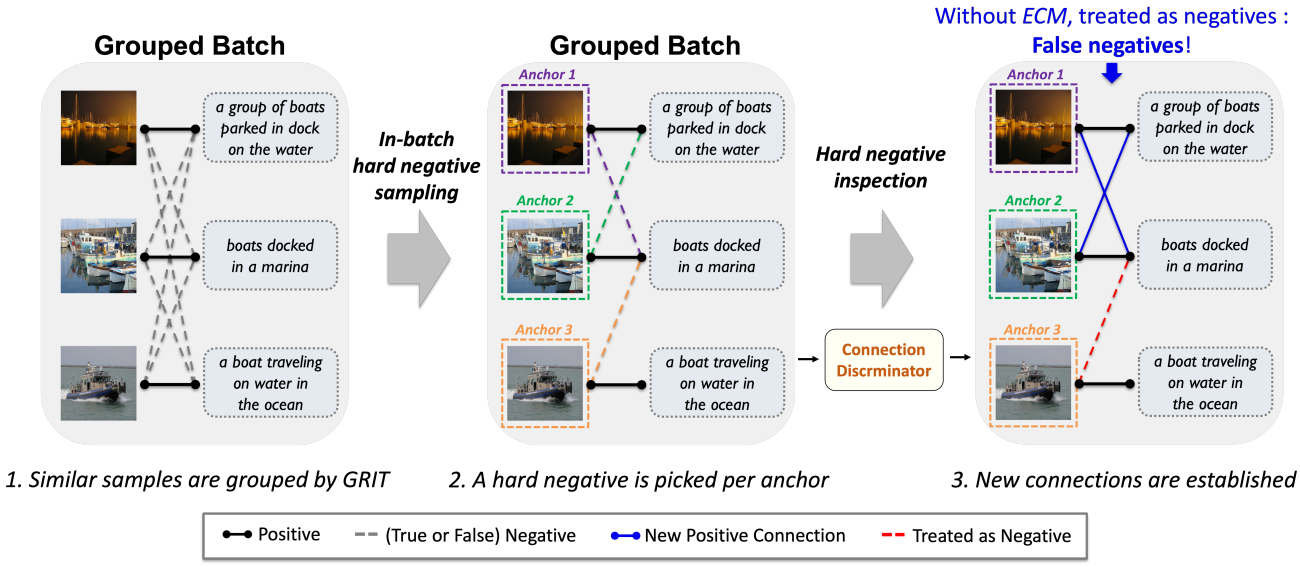

Figure 4. Efficient Connection Mining (ECM).

the *ambiguous* samples, which are uncertain whether they are false negatives or not (those with a probability of being positive between 0.5 and 0.8). These ambiguous samples are discarded, and the second hardest text (or image) is re-sampled for the anchor image (or text) to obtain a more certain negative for ITM. Thus, the form of the ITM loss is as follows:

$$\mathcal{L}_{\text{ITM}}^{ECM} = \mathbb{E}_{(i,t)\sim D}\Big[\text{CE}\Big(\tilde{\boldsymbol{y}}^{\text{ITM}}, \boldsymbol{p}^{\text{ITM}}(i,t)\Big)\Big], \tag{3}$$

in which $\tilde{\boldsymbol{y}}^{\text{ITM}}$ represents the corrected one-hot label from *Con-D*.

[MLM with ECM] For MLM, the model is asked to predict masked tokens in the caption using unmasked text tokens and visual information. In addition to the original positive pairs in the dataset, new positive pairs detected by *Con-D* are additionally used:

$$\mathcal{L}_{\text{MLM}}^{ECM} = \mathbb{E}_{(i,t)\sim D\cup D_{\text{ECM}}}\Big[\text{CE}\Big(\boldsymbol{y}^{\text{MLM}}, \boldsymbol{p}^{\text{MLM}}(i, t^{\text{mask}})\Big)\Big], \tag{4}$$

in which D_{ECM} denotes the set of newly constructed pairs, $\boldsymbol{y}^{\text{MLM}}$ represents the one-hot label for the masked token, and $\boldsymbol{p}^{\text{MLM}}(i, t^{\text{mask}})$ indicates the model's prediction for the masked token.

Remark: *Con-D* is pre-trained with the following pre-training objectives: S-ITC, MLM, and ITM with GRIT sampling. Then, it is fine-tuned on the Karpathy training split of MS-COCO [38], and the output of ITM head of *Con-D* serves as the probability for a candidate image-text pair to be a positive pair. The additional computation overhead of ECM during training depends only on the number of samples provided to *Con-D* and the number of new positives given to the multi-modal encoder of the training model for MLM. Since only labels are corrected for ITC and ITM, it does not require additional forward passes for the model being trained. Thus, despite the inclusion of additional forwarding passes for ECM, the extra overhead introduced by ECM is relatively low compared to the momentum distillation technique employed in ALBEF and BLIP. Detailed information regarding the computational cost is described in S.M.

4.2. Smoothed ITC (S-ITC)

To overcome the challenge of false negatives in ITC under GRIT sampling, we additionally introduce a computation-free approach named S-ITC, which employs label smoothing to contrastive loss, which has not been extensively explored in VLP. Specifically, we take the following loss form:

$$\begin{aligned}\mathcal{L}_{\text{S-ITC}} = \frac{1}{2}\mathbb{E}_{(i,t)\sim D}\Big[&\text{CE}\Big((1-\alpha)\boldsymbol{y}^{\text{I2T}}(i) + \frac{\alpha}{N}\mathbf{1}, \boldsymbol{p}^{\text{I2T}}(i)\Big) \\ &+ \text{CE}\Big((1-\alpha)\boldsymbol{y}^{\text{T2I}}(t) + \frac{\alpha}{N}\mathbf{1}, \boldsymbol{p}^{\text{T2I}}(t)\Big)\Big],\end{aligned} \tag{5}$$

in which α represents a mixing parameter, and $\mathbf{1}$ denotes all-one vector.

We emphasize that label smoothing has not been widely adopted in typical VLP settings since it has not been very effective. As we show in Table 6 (Section 5), performance is significantly degraded when S-ITC is applied under the random sampling scenario. This decline is largely due to the detrimental effect of providing soft labels for the examples in the randomly sampled batch where true negatives are prevalent. In contrast, under GRIT sampling where each mini-batch is predominantly composed of samples that are likely to be false negatives, we observe that S-ITC, which ensures relatively high soft labels for all negatives, becomes highly effective.

There also have been other attempts to address the issue of false negatives in ITC, such as momentum distilla-

Table 2. Values of soft labels assigned to samples in ITC for different methods. The batch size B is set to 96, and the queue size Q is set to 48000. The soft labels were computed in the last epoch of the training.

Method	Sum of soft labels		
	Top $1 \sim 5$	Top $6 \sim B$	Top $B+1 \sim B+Q$
S-ITC	0.5260	0.4740	·
Consistency Loss	0.9822	0.0178	·
Momentum Distillation	0.6746	0.0009	0.3245

tion [33] and consistency loss [2]. Here, we explain only I2T-related terms for simplicity; T2I-related terms are similarly computed. Momentum distillation replaces $\boldsymbol{y}_t^{\text{I2T}}(i)$ by $\boldsymbol{y}_t^{\text{MD}}(i) = (1-\alpha)\boldsymbol{y}_t^{\text{I2T}}(i) + \alpha sg[\tilde{\boldsymbol{p}}_t^{\text{I2T}}(i)]$, where $sg[\cdot]$ is the stop gradient operator, and $\tilde{\boldsymbol{p}}$ denotes the probability obtained from the momentum encoder. Here, N is equal to batch-size $B+Q$ since the model is accompanied by a queue of size Q that stores embeddings to provide additional negatives. However, this approach suffers from inefficiency due to the additional forwarding of the momentum model, and it results in increasing the model size. In consistency loss, $\boldsymbol{y}_t^{\text{I2T}}(i)$ is substituted with $\boldsymbol{y}_t^{\text{CS}}(i) = (1-\alpha)\boldsymbol{y}_t^{\text{I2T}}(i) + \alpha sg[\boldsymbol{p}_i^{\text{T2I}}(t)]$, where $\boldsymbol{p}_i^{\text{T2I}}(t)$ is computed by the model itself. Here, N is the same as B since it does not involve a queue.

However, as shown in Table 6 (in Section 5), the effectiveness of momentum distillation and consistency loss is limited. To explore the reason behind this, we examine the soft labels from the above methods in the GRIT sampling scenario as reported in Table 2, aiming to uncover the distribution shapes of the soft labels. The values in the table are obtained through the following process: the soft labels are sorted in descending order, and then averaged across all samples. Further details on the computation process are described in S.M. We observe that momentum distillation continues to assign almost zero labels to negative samples, which are likely to be false negatives under GRIT sampling. This result may stem from the large number of negatives in the queue, which prevents each negative sample from receiving non-negligible labels. On the other hand, consistency loss assigns comparatively higher soft labels (0.0178) than momentum distillation (0.0009) but overly concentrates on a few pairs, resulting in negligible labels for most negatives. In S.M, we provide an analysis that this phenomenon cannot be resolved by merely tuning α.

Given that both consistency loss and the momentum distillation fail to achieve the intended objective of assigning non-negligible soft labels to the majority of negatives, we argue that S-ITC, which explicitly assigns higher soft labels to all negatives, can be a simple but effective solution. In S.M, we include an analysis of its robustness to α.

Consequently, as illustrated in Figure 3 and Algorithm 1 in S.M, we adopt the following pre-training objective:

$$\mathcal{L} = \mathcal{L}_{\text{S-ITC}}^{ECM} + \mathcal{L}_{\text{MLM}}^{ECM} + \mathcal{L}_{\text{ITM}}^{ECM}, \tag{6}$$

where $\mathcal{L}_{\text{S-ITC}}^{ECM}$ represents the integrated ITC loss of $\mathcal{L}_{\text{ITC}}^{ECM}$ and $\mathcal{L}_{\text{S-ITC}}$, which adopts the target labels as $(1-\alpha)\tilde{\boldsymbol{y}}^{\text{ITC}} + \frac{\alpha}{N}\mathbf{1}$.

5. Experimental Results

5.1. Data and experimental settings

During our training process, we utilize four datasets (MS-COCO [38], Visual Genome [30], Conceptual Captions [52], and SBU Captions [45]) with a total of 4M unique images (5M image-text pairs), as proposed by ALBEF [33] and UNITER [5]. We refer to this collective dataset as the "4M-Noisy" dataset due to a significant number of captions that offer either incomplete or incorrect descriptions, which can be seen as false positives. To analyze the impact of our approach in handling false negatives relative to the effect of removing false positives, we construct an additional same-sized training set named "4M-Clean" which is composed of clean image-text pairs, refined by the BLIP captioner [34]. Note that all the models are pre-trained with the "4M-Noisy" unless specifically stated as "4M-Clean" in our results table below. Further details on constructing the 4M-Clean dataset are in S.M.

Following ALBEF, we adopt our image encoder as a 12-layer Vision Transformer [12] with 86 million parameters, pre-trained on ImageNet-1k [57]. Both the text and multimodal encoders utilize a 6-layer Transformer [58], initializing the former with the first 6 layers and the latter with the last 6 layers of BERT-base model (123.7M parameters) [10]. We use the same data augmentation method used in ALBEF and train our model for 20 epochs using 4 NVIDIA A100 GPUs, but excluding the momentum encoder in ALBEF. For *Con-D*, we use the exact same model architecture as the training model. Unless otherwise noted, we set $B = 96$ and $M = 4800$ for GRIT sampling, and for all other hyperparameter settings, we follow GRIT-VLP [2]. More details on the dataset, training, and hyperparameters are in S.M.

5.2. Downstream vision and language tasks

After completing the pre-training phase, we proceed to fine-tune our model on three downstream vision and language tasks: image-text retrieval (IRTR) [38], visual question answering (VQA) [17], and natural language for visual reasoning (NLVR2) [55]. For IRTR, we utilize the MS-COCO [38] and Flickr30K (F30K) [46] datasets, with F30K being re-splitted according to [28]. Following BLIP [34], we exclude the SNLI-VE dataset [62] due to reported noise in the data. Our fine-tuning and evaluation process mostly follows that of GRIT-VLP. More details of downstream tasks are in S.M.

5.3. Comparison with baselines

In Table 3, we observe that our approach consistently outperforms other baselines in multiple downstream tasks (IRTR, VQA, NLVR2). Notably, MAFA even surpasses ALBEF

Table 3. Comparison with various methods on downstream vision-language tasks. **Bold** denotes the best result among models trained with 4M dataset. * refers to the reproduced models by the authors. Methods without explicit designation are trained on 4M-Noisy dataset.

Method	Pre-train # Images	COCO R@1		Flickr R@1		NLVR2		VQA	
		TR	IR	TR	IR	dev	test-P	test-dev	test-std
UNITER [5]	4M	65.7	52.9	87.3	75.6	77.18	77.85	72.70	72.91
VILLA [14]	4M	-	-	87.9	76.3	78.39	79.30	73.59	73.67
OSCAR [37]	4M	70.0	54.0	-	-	78.07	78.36	73.16	73.44
ALBEF [33]	4M	73.1	56.8	94.3	82.8	80.24	80.50	74.54	74.70
TCL [63]	4M	75.6	59.0	94.9	84.0	80.54	81.33	74.90	74.92
BLIP* (4M-Clean) [34]	4M	75.5	58.9	94.3	82.6	79.70	80.87	75.50	75.76
GRIT-VLP* [2]	4M	76.6	59.6	95.5	82.9	81.40	81.23	75.26	75.32
MAFA	4M	78.0	61.2	96.1	**84.9**	82.52	82.08	75.55	75.75
MAFA (4M-Clean)	4M	**79.4**	**61.6**	**96.2**	84.6	**82.66**	**82.16**	**75.91**	**75.93**
ALBEF	14M	77.6	60.7	95.9	85.6	82.55	83.14	75.85	76.04
BLIP	14M	80.6	63.1	96.6	87.2	82.67	82.30	77.54	77.62

Table 4. Ablation study on the proposed method. **Bold** denotes the best result among models trained with 4M-Noisy, 4M-Clean dataset, respectively.

Pre-train dataset	MAFA		COCO R@1		Flickr R@1		NLVR2		VQA	
	S-ITC	ECM	TR	IR	TR	IR	dev	test-P	test-dev	test-std
4M-Noisy	✗	✗	76.6	59.6	95.5	82.9	81.40	81.23	75.26	75.32
	✗	✓	77.4	60.2	95.5	83.3	82.03	81.76	75.39	75.52
	✓	✗	77.5	60.5	96.1	84.2	81.74	81.33	75.42	75.51
	✓	✓	**78.0**	**61.2**	**96.1**	**84.9**	**82.52**	**82.08**	**75.55**	**75.75**
4M-Clean	✗	✗	77.7	60.7	95.2	84.2	81.44	81.39	75.50	75.57
	✓	✓	**79.4**	**61.6**	**96.2**	**84.6**	**82.66**	**82.16**	**75.91**	**75.93**

(14M) and competes with BLIP (14M) on certain metrics, despite being trained on a significantly smaller dataset. Specifically, MAFA achieves significant improvements over GRIT-VLP, with a substantial margin of $+1.4\%$ IR/R@1, $+1.6\%$ TR/R@1 on MS-COCO and $+1.1\%$ on NLVR2 dev. These results clearly show the significance of addressing false negatives when leveraging hard negative mining. Additionally, we believe that the comparison between BLIP (4M-Clean) and our MAFA shows that the effectiveness of managing false negatives may surpass the impact of mitigating false positives. Furthermore, the enhanced performance of MAFA (4M-Clean) over MAFA shows the synergistic effect of addressing both false positives and negatives.

5.4. Ablation studies

Table 4 presents the effectiveness of two proposed components: efficient connection mining (ECM) and smoothed ITC (S-ITC). Here, all model variants adopt GRIT sampling, with row 1 representing the original GRIT-VLP model. The results clearly demonstrate that applying either the S-ITC (row 3) or the ECM (row 2) individually leads to performance improvements compared to a model that does not consider false negatives (row 1). By combining both S-ITC and ECM in our final model (row 4), we observe significant performance enhancements on the 4M-Noisy dataset. This tendency is validated again in the 4M-Clean dataset, confirming the consistent effectiveness of the proposed components (row 6). Moreover, by comparing the performance gap between MAFA trained on the noisy dataset (row 4) and GRIT-VLP trained on the clean dataset (row 5), we reaffirm that addressing false negatives outweighs the impact of handling false positives. Beyond the 4M dataset, we present additional results across a broader range of data scales (1M, 2M, and 14M) in S.M, demonstrating the robustness of MAFA with respect to data scale variations.

Table 5. Analysis of the effect of MAFA with GRIT sampling. "ECM-E" denotes eliminating false negatives rather than using them as positives.

Method		COCO R@1		NLVR2		VQA	
GRIT	MAFA	IR	TR	dev	test-P	test-dev	test-std
✗	✗	74.4	57.6	79.75	79.94	74.49	74.67
✗	✓	74.3	57.8	81.20	81.03	74.61	74.78
✓	✗	76.6	59.6	81.40	81.23	75.26	75.32
✓	✓(ECM-E)	77.1	61.1	82.33	81.95	75.50	75.54
✓	✓	**78.0**	**61.2**	**82.52**	**82.08**	**75.55**	**75.75**

Table 5 provides an additional comparative analysis on the effectiveness of MAFA, based on whether GRIT sampling and ECM are either applied or not. Here, row 1 denotes the ALBEF model without momentum distillation. Since S-ITC is ineffective under random sampling (as we show in Table 6 below), S-ITC is excluded when GRIT sampling is not utilized (row2). We observe that MAFA enhances the

performance for both random and GRIT sampling. However, the effect of MAFA is much more vivid for GRIT sampling (row 4, 5), underscoring the critical role of managing false negatives in hard negative sampling. Moreover, our experiments reveal that converting false negatives into additional positives (row 5) is considerably more beneficial than merely removing them (row 4), which highlights the effect of leveraging new positive connections constructed by the model within the dataset itself.

Furthermore, in Table 6, we provide a comparative analysis on S-ITC, which supports our discussion in Section 4.2; S-ITC brings out a unique synergy only when combined with GRIT-sampling. Namely, in random sampling, we observe that S-ITC rather detrimentally affects performance (row 2). Conversely, under GRIT sampling, we verify that assigning relatively high nonzero labels to most negatives enhances performance. Namely, consistency loss (row 4), which assigns relatively higher soft labels to samples in a batch, outperforms momentum distillation (row 5). S-ITC significantly outperforms the other two variants, which highlights the importance of assigning substantial labels to the majority of negatives, rather than just a few.

5.5. Compatibility of MAFA with BLIP-2 [35]

In Tables 7 and 8, we demonstrate that our MAFA can be successfully integrated with the recent BLIP-2 [35], which is quite a successful vision-language pre-training framework. As described in S.M., the *stage-1* of BLIP-2, which adopts ITC, ITM, and (auto-regressive) LM losses as objectives, closely resembles the pre-training procedures of both BLIP and ALBEF. Thus, MAFA can be effortlessly incorporated into *stage-1* of BLIP-2, following the identical way described in Section 4.

In Table 7, we observe the performance of BLIP-2+GRIT is significantly degraded (row 2), which indicates that solely applying GRIT sampling leads to a failure of learning. We believe this performance degradation primarily stems from more frequent occurrences of false negatives in BLIP-2. In BLIP-2, due to the significantly enhanced capacity of the model, GRIT sampling, which mines hard negatives based on contrastive similarities calculated from the training model, includes a substantially higher number of false negatives in each batch. The integration of MAFA with BLIP-2 leads to enhanced performance, highlighting the importance of managing false negatives to increase the stability of the training process.

We further explore whether the integration of MAFA in *stage-1* leads to improved generative learning capabilities after additional *stage-2* training where the model is connected to a frozen LLM and pre-trained only with LM loss. We evaluate the performance of models in various zero-shot visual question answering benchmark datasets including GQA [23], OKVQA [42], and VQA [17]. Moreover, we assess the zero-shot image captioning ability on the Karpathy test split of MS-COCO [38]. Table 8 shows that MAFA significantly improves zero-shot performance across various VQA and image captioning tasks. This result not only underscores the compatibility of MAFA with BLIP-2 but also emphasizes that the exclusive integration of MAFA in *stage-1* is also beneficial in generative learning capability (*stage-2*) as well. More detailed results, including those from fine-tuned image captioning and an analysis on how extra positive examples from ECM influence the BLIP-2 *stage-2* performance, are provided in S.M.

Table 6. Comparison of soft-labeling methods for ITC.

Method		COCO R@1		Flickr R@1	
GRIT	Soft labeling	TR	IR	TR	IR
✗	✗	**74.4**	**57.6**	**93.5**	81.7
	Momentum Distillation	74.2	57.4	**93.5**	**81.9**
	S-ITC	73.5	56.1	92.9	79.9
✓	Consistency Loss	76.6	59.6	95.5	82.9
	Momentum Distillation	76.1	58.9	94.4	82.7
	S-ITC	**77.5**	**60.5**	**96.1**	**84.2**

Table 7. Fine-tuned IRTR results with BLIP-2 framework on MS-COCO datasets.

Model	TR			IR		
	R@1	R@5	R@10	R@1	R@5	R@10
BLIP-2	82.6	96.3	98.2	66.1	86.8	**92.0**
BLIP-2 + GRIT	65.9	89.1	95.0	52.5	79.4	87.3
BLIP-2 + MAFA	**83.7**	**96.6**	**98.4**	**66.7**	**86.8**	91.9

Table 8. Zero-shot visual question answering and image captioning results with BLIP-2 framework.

Model	VQAv2	OK-VQA	GQA	COCO zero-shot Karpathy test	
	val	test	test-dev	BLEU@4	CIDEr
BLIP-2	46.6	23.8	29.1	35.6	118.8
BLIP-2 + MAFA	**50.8**	**29.0**	**31.8**	**37.6**	**125.4**

6. Concluding Remarks

We introduce MAFA, a novel approach equipped with two key components (ECM, S-ITC), specifically designed to tackle the prevalent issue of false negatives in VLP. Our comprehensive experiments demonstrate that addressing false negatives plays a crucial role in VLP. Moreover, we believe that the concept of converting false negatives into additional positives paves the way for future research that leverages the inherent missing positive connections within a dataset.

Acknowledgment

This work was supported in part by the National Research Foundation of Korea (NRF) grant [No.2021R1A2C2007884] and by Institute of Information & communications Technology Planning & Evaluation (IITP) grants [No.2021-0-01343, No.2021-0-02068, No.2022-0-00113, No.2022-0-00959] funded by the Korean government (MSIT). It was also supported by SNU-Naver Hyperscale AI Center.

References

[1] Junyu Bi, Daixuan Cheng, Ping Yao, Bochen Pang, Yuefeng Zhan, Chuanguang Yang, Yujing Wang, Hao Sun, Weiwei Deng, and Qi Zhang. VL-Match: Enhancing vision-language pretraining with token-level and instance-level matching. In *ICCV*, 2023. 2

[2] Jaeseok Byun, Taebaek Hwang, Jianlong Fu, and Taesup Moon. GRIT-VLP: Grouped mini-batch sampling for efficient vision and language pre-training. In *ECCV*, 2022. 1, 2, 6, 7

[3] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 1

[4] Tsai-Shien Chen, Wei-Chih Hung, Hung-Yu Tseng, Shao-Yi Chien, and Ming-Hsuan Yang. Incremental false negative detection for contrastive learning. *arXiv preprint arXiv:2106.03719*, 2021. 2

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *ECCV*, 2020. 2, 6, 7

[6] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, 2021. 2

[7] Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *NeurIPS*, 2020. 2

[8] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *CVPR*, 2021. 2

[9] Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang, and Seong Joon Oh. ECCV Caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for ms-coco. In *ECCV*, 2022. 1

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 6

[11] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *AAAI*, 2021. 1

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image is Worth 16x16 Words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[13] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. DataComp: In search of the next generation of multimodal datasets. In *NeurIPS*, 2024. 1

[14] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020. 2, 7

[15] Asish Ghoshal, Xilun Chen, Sonal Gupta, Luke Zettlemoyer, and Yashar Mehdad. Learning better structured representations using low-rank adaptive label smoothing. In *ICLR*, 2021. 2

[16] Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *NeurIPS*, 2023. 2

[17] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 6, 8

[18] Peng Hu, Xi Peng, Hongyuan Zhu, Liangli Zhen, and Jie Lin. Learning cross-modal retrieval with noisy labels. In *CVPR*, 2021. 1

[19] Runhui Huang, Yanxin Long, Jianhua Han, Hang Xu, Xiwen Liang, Chunjing Xu, and Xiaodan Liang. NLIP: Noise-robust language-image pre-training. *arXiv preprint arXiv:2212.07086*, 2022. 1

[20] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 2

[21] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. In *NeurIPS*, 2021. 1

[22] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing Out of tHe bOx: End-to-end pre-training for vision-language representation learning. In *CVPR*, 2021. 2

[23] Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 8

[24] Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *WACV*, 2022. 2

[25] Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. Patching open-vocabulary models by interpolating weights. In *NeurIPS*, 2022. 2

[26] Yiren Jian, Chongyang Gao, and Soroush Vosoughi. Bootstrapping vision-language learning with decoupled language pre-training. *arXiv preprint arXiv:2307.07063*, 2023. 2

[27] Chaoya Jiang, Haiyang Xu, Wei Ye, Qinghao Ye, Chenliang Li, Ming Yan, Bin Bi, Shikun Zhang, Fei Huang, and Songfang Huang. BUS: Efficient and effective vision-language pre-training with bottom-up patch summarization. In *ICCV*, 2023. 2

[28] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 6

[29] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*, 2021. 2

[30] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual Genome:

Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017. 6

[31] Ujwal Krothapalli and A Lynn Abbott. Adaptive label smoothing. *arXiv preprint arXiv:2009.06432*, 2020. 2

[32] Dongkyu Lee, Ka Chun Cheung, and Nevin L Zhang. Adaptive label smoothing with self-knowledge in natural language generation. In *EMNLP*, 2022. 2

[33] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before Fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 1, 2, 4, 6, 7

[34] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 1, 2, 4, 6, 7

[35] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 4, 8

[36] Weizhi Li, Gautam Dasarathy, and Visar Berisha. Regularization via structural label smoothing. In *AISTATS*, 2020. 2

[37] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 2, 7

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 5, 6, 8

[39] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 2

[40] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020. 2

[41] Pratyush Maini, Sachin Goyal, Zachary C Lipton, J Zico Kolter, and Aditi Raghunathan. T-MARS: Improving visual representations by circumventing text feature learning. *arXiv preprint arXiv:2307.03132*, 2023. 1

[42] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019. 8

[43] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *NeurIPS*, 2019. 2

[44] Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. In *NeurIPS*, 2024. 1

[45] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2Text: Describing images using 1 million captioned photographs. In *NeurIPS*, 2011. 6

[46] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 6

[47] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. Filtering, distillation, and hard negatives for vision-language pre-training. *arXiv preprint arXiv:2301.02280*, 2023. 1

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

[49] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020. 2

[50] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1

[51] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[52] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 1, 6

[53] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *CVPR*, 2019. 2

[54] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2

[55] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018. 6

[56] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 2

[57] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 6

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 6

[59] Alex Jinpeng Wang, Kevin Qinghong Lin, David Junhao Zhang, Stan Weixian Lei, and Mike Zheng Shou. Too Large; data reduction for vision-language pre-training. In *ICCV*, 2023. 1

[60] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[61] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *CVPR*, 2022. 2

[62] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual Entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019. 6

[63] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *CVPR*, 2022. 1, 2, 7

[64] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *ICLR*, 2022. 1

[65] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-Grained Vision Language Pre-Training: Aligning texts with visual concepts. In *ICML*, 2022. 1, 2

[66] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, 2020. 2