Multi-Modal Representation Learning with Text-Driven Soft Masks

Jaeyoo Park[1]   Bohyung Han[1,2]

Computer Vision Laboratory, ECE[1] & IPAI[2], Seoul National University

{bellos1203, bhhan}@snu.ac.kr

Abstract

We propose a visual-linguistic representation learning approach within a self-supervised learning framework by introducing a new operation, loss, and data augmentation strategy. First, we generate diverse features for the image-text matching (ITM) task via soft-masking the regions in an image, which are most relevant to a certain word in the corresponding caption, instead of completely removing them. Since our framework relies only on image-caption pairs with no fine-grained annotations, we identify the relevant regions to each word by computing the word-conditional visual attention using multi-modal encoder. Second, we encourage the model to focus more on hard but diverse examples by proposing a focal loss for the image-text contrastive learning (ITC) objective, which alleviates the inherent limitations of overfitting and bias issues. Last, we perform multi-modal data augmentations for self-supervised learning via mining various examples by masking texts and rendering distortions on images. We show that the combination of these three innovations is effective for learning a pretrained model, leading to outstanding performance on multiple vision-language downstream tasks.

1. Introduction

Vision-language representation learning aims at optimizing a joint embedding model for the data in both the image and text domains, where the learned representations are transferred to solve various vision-language downstream tasks including image-text retrieval [20, 27, 32], visual question answering [1, 3, 29, 44], visual reasoning [37, 41], etc. With the introduction of large-scale image-text datasets [18, 27, 31, 36] and the recent advances in language models [2, 7, 40], research for vision-language representation learning [4, 17, 21, 24, 28] has been actively exploring self-supervision tasks such as image-text matching (ITM), masked language modeling (MLM), and image-text contrastive learning (ITC).

During the pretraining stage, the model solves the tasks based only on image-caption pairs with no fine-grained annotations, *i.e.,* annotations of region-word relations. Thus,

Text : A zebra next to a deer

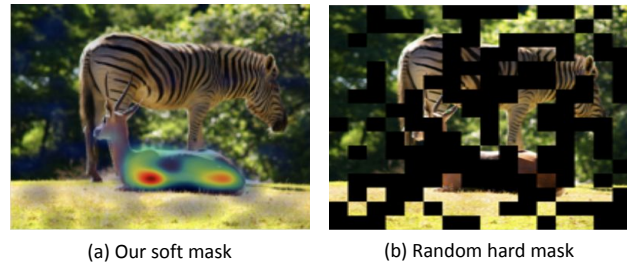(a) Our soft mask   (b) Random hard mask

Figure 1. Illustration of the difference between the proposed soft feature masking and the random hard feature masking. The random hard feature masking has a risk of completely removing the important regions while our soft feature masking strategy still preserves the regions.

the model focuses on the most discriminative regions in images, which typically correspond to the primary objects in the scenes. Hence, the model lacks a comprehensive understanding of various attributes observed in images despite detailed descriptions in captions. To alleviate this limitation, we propose the following three techniques for diversifying observations, eventually resulting in performance improvement of learned representations.

First, we devise a soft masking technique on visual features for training, which suppresses activations at the discriminative parts in an input image with respect to a certain word in its caption. The proposed soft feature masking makes a model focus on the regions that are initially paid less attention to and synthesizes diverse samples during the training procedure. To be specific, we first utilize the cross-attention map of image and text features derived by the multi-modal fusion encoder and identify salient regions in the image by computing the word-specific Grad-CAM [35] of the image-text matching score. Then we feed a softly-masked image feature to the fusion encoder for learning to match the features of both modalities. While the most straightforward way to augment diverse visual features is to employ discrete hard masks as recent trends [12, 42], which aim to reconstruct images, our approach suppresses information in images using the soft masks with real-valued

weights instead of completely removing the content relevant to input text. Figure 1 illustrates the difference between the proposed soft mask and the random hard (discrete) mask.

Second, we also encourage the model to focus more on hard examples by adopting a focal version of the ITC loss. The focal loss [26] is originally designed for object detection to prevent overfitting and class imbalance problems, and a similar idea turns out to be effective for our contrastive learning task. This is partly because our task still has overfitting issues for easy examples and is also biased toward negative examples. Since large-scale image-caption corpora are composed of multiple datasets with large domain gaps, a model distinguishes a lot of samples from different datasets very easily [6]. Hence, we propose the focal ITC loss for our model to achieve more balanced training.

In addition to the two components discussed above, we perform multi-modal data augmentations, which include diverse strong augmentations proposed in [43] and binary masks to captions for the ITM task. Strong data augmentations have been believed to harm vision-language models [16, 39] by breaking the semantic consistency of the image-text pair, but we argue that this does not hold for the large-scale vision-language representation pretraining, as also shown in recent studies [15, 25]. Note that the use of binary masks in ITM also homogenizes text inputs with the masked language model (MLM) task.

The organization of this paper is as follows. We first discuss related works about vision-language representation learning in Section 2 and provide the preliminaries about vision-language representation learning in Section 3. Section 4 discusses the vision-language representation learning approach equipped with our ideas. We demonstrate the performance of the proposed approach on multiple downstream tasks in Section 5 and conclude this paper in Section 6.

2. Related Works

Vision-language representation learning approaches are mainly split into the methods equipped with the pretrained object detector [4, 10, 24, 28, 38, 45] and the ones with a detector-free visual encoder [9, 13, 17, 21, 43]. Among lots of literature about the topic, below we provide the ones that are most related to ours.

Following the studies for various vision-language tasks such as visual question answering [1, 3, 44] and image-text matching [19, 22], research in the early day adopted a bottom-up strategy with a pretrained object detector, *e.g.*, Faster R-CNN [34], as a visual encoder. ViLBERT [28] and LXMERT [38] directly extend BERT [7] by constructing co-attention-based frameworks with two separate modality-specifc streams. There exist some approaches deploying single-stream transformers as fusion encoders. UNITER [4] proposes word-region alignment loss and investigates various mask modeling objectives. Meanwhile, OSCAR [24]

incorporates object tags from detection results as anchors to reduce the gap between image and text embedding spaces. Based on [4], [10] adopts an adversarial training strategy to improve generalization capability, although being yet time-consuming. Recently, [45] studies the effect of an improved object detector and provides rich object embeddings.

However, approaches incorporating fixed pretrained object detectors as their visual encoders have several drawbacks. First, the visual encoders cannot be trained during training, due to the high computation cost. Although detector-based visual encoders provide rich embeddings for the pre-defined object categories, they suffer from the lack of information about the objects that are not seen during the training of the detection model. Second, the visual embeddings are learned from bounding box regions focusing on objects. Hence, the reasoning process fails to learn the global context from the outside of the bounding boxes, such as the background.

To address such drawbacks, recent works attempt to build models with detector-free visual encoders. For example, VilT [17] feeds the linear projections of grid-patch embeddings within an image, but shows lower performance due to the lack of the deep visual encoder. SOHO [13] discretizes the embedding of an input image using an end-to-end trainable CNN-based visual encoder with a visual dictionary. ALBEF [21] adopts a Vision Transformer [8] and proposes to align the intermediate features of two modalities before being fused, using contrastive objective. TCL [43] further learns to maximize the mutual information between local regions and global embeddings in each modality, based on the features from [21]. CODIS [9] proposes a multimodal codebook learning, and bridges the gap between two modalities. While our work shares the basic architecture with [21] and [43], its main goal is to take advantage of softly masked visual features and incorporate various regularization techniques through the construction of diversified examples.

3. Preliminaries

This section discusses the details of baseline architectures and self-supervised learning tasks, which are widely used for vision-language representation learning.

3.1. Multi-Modal Representation Learning

The model consists of a visual encoder f_v, a text encoder f_t, and a multi-modal encoder f_{mm}. Each encoder has a paired momentum encoder and the model stores historical [CLS] tokens for training. An input image I is first patchified into a sequence of small grid patch tokens, $\boldsymbol{I}$. We encode a set of image patch tokens, $\boldsymbol{I}$, into a sequence of embedding tokens using an image encoder, f_v, and obtain $\boldsymbol{V}_{\text{emb}} = \{\boldsymbol{v}_0, \boldsymbol{v}_1, ..., \boldsymbol{v}_N\}$, where $\boldsymbol{v}_0$ denotes the [CLS] token feature and N is number of the patches in the input im-
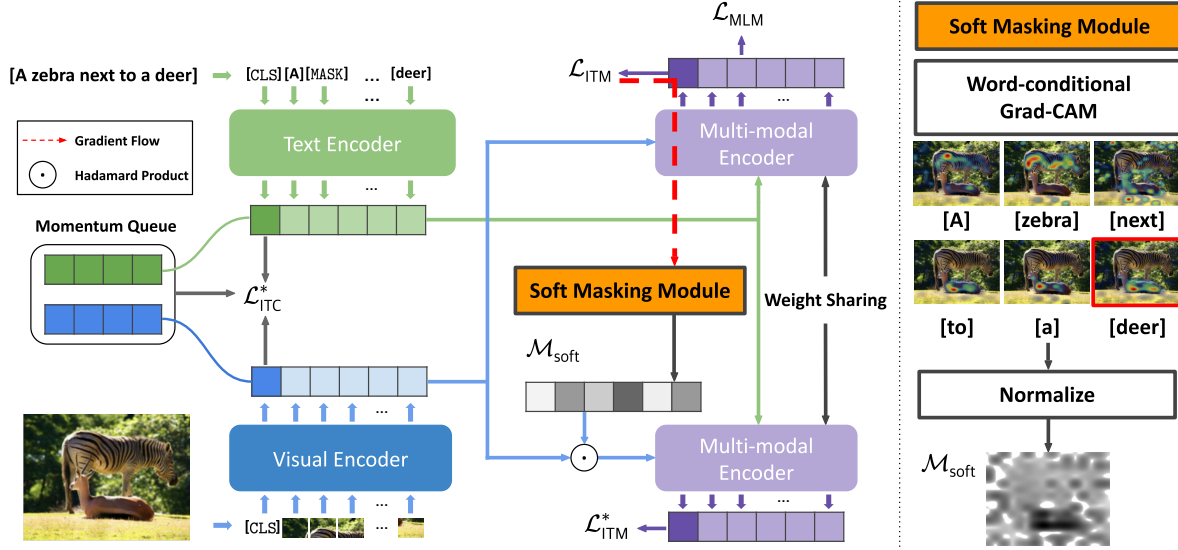
Figure 2. Illustration of the proposed framework. Given the image and the masked text data, we encode them to obtain the sequences of unimodal embedding tokens from both modalities. Each encoder has a paired momentum model updated by the moving-average, which we omit in the figure for simplicity. For the unimodal level, the model optimizes ITC objective to align the embedding space of each modality before fusing them (Section 3). We feed the sequences of embedding tokens to the multi-modal encoder to aggregate the information of both modalities and learn to solve ITM and MLM tasks (Section 3). We propose to additionally employ soft masked visual features for learning complement attributes from the image (Section 4.1). We generate the soft masks from the Grad-CAM of the positive matching score with respect to the cross-attention map of the multi-modal encoder. We randomly select one of the word-wise Grad-CAM and use the normalized Grad-CAM as a soft mask for the visual feature. Our generated soft-masked visual feature works as a hard but diversified sample by softly-masking the important regions while not perfectly removing them. Moreover, we introduce the focal version of ITC, where the model focuses more on the hard examples (Section 4.2).

age. Likewise, we tokenize an input text T into a sequence of word tokens $\boldsymbol{T}$, and the text encoder embeds $\boldsymbol{T}$ into a sequence of text features $\boldsymbol{T}_{\text{emb}} = \{\boldsymbol{t}_0, \boldsymbol{t}_1, ..., \boldsymbol{t}_L\}$, where $\boldsymbol{t}_0$ denotes the [CLS] token feature and L is the length of the input text. The image features $\boldsymbol{V}_{\text{emb}}$ are fused with the text features $\boldsymbol{T}_{\text{emb}}$ through a cross-attention at each layer of the multi-modal encoder. The multi-modal encoder, f_{mm}, outputs the joint embedding tokens of two modalities, $\boldsymbol{M}_{\text{emb}} = \{\boldsymbol{m}_0, \boldsymbol{m}_1, ..., \boldsymbol{m}_L\}$, where $\boldsymbol{m}_0$ denotes the [CLS] token feature. Note that $|\boldsymbol{T}_{\text{emb}}| = |\boldsymbol{M}_{\text{emb}}| = L + 1$, since the multi-modal encoder takes $\boldsymbol{T}_{\text{emb}}$ as its query and fuses $\boldsymbol{V}_{\text{emb}}$ and $\boldsymbol{T}_{\text{emb}}$ via a cross-attention mechanism. Based on this configuration, we perform multi-modal self-supervised representation learning by using the following objectives.

3.2. Self-Supervised Learning Objectives

We discuss three tasks based on self-supervision, which are widely used for multi-modal representation learning.

Image-Text Contrastive Learning (ITC) ITC aligns two modality-specific representations, *i.e.*, [CLS] tokens denoted by $\boldsymbol{v}_0$ and $\boldsymbol{t}_0$ for an image and a text, before their aggregations using a multi-modal fusion encoder. In other words, ITC enforces an image-text pair, $(\boldsymbol{v}_0, \boldsymbol{t}_0)$, to be close if they are matched, and to be far away otherwise. Fol-

lowing ALBEF [21], we adopt two momentum queues for storing recent [CLS] token features from two unimodal encoders, *i.e.*, $\boldsymbol{v}_0$ and $\boldsymbol{t}_0$, and use them as additional examples for contrastive learning. Formally, the similarity between I and T, $s(I, T)$, is defined as

$$s(I, T) = h_v(\boldsymbol{v}_0)^T h_t(\boldsymbol{t}_0), \quad (1)$$

where $h_v(\cdot)$ and $h_t(\cdot)$ are linear projection layers. Then, for each I and T, the image-to-text similarity score for the ith example is defined as

$$p_{v2t}^{(i)} = \frac{\exp(s(I_i, T_i)/\tau)}{\sum_{j=1}^{B+Q} \exp(s(I_i, T_j)/\tau)}, \quad (2)$$

where τ is a learnable temperature, B denotes the size of mini-batch, and Q is the size of the momentum queue. Note that $\{T_j | j = B + 1, ..., B + Q\}$ is a set of the negative text samples stored in the momentum queue. We also define $p_{t2v}^{(i)}$, text-to-image similarity score for the ith data, in the same way. The ITC loss, which is the sum of InfoNCE losses [30] in two directions, is then given by

$$\mathcal{L}_{\text{ITC}} = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log p_{v2t}^{(i)} + \log p_{t2v}^{(i)} \right]. \quad (3)$$

Image-Text Matching (ITM) The ITM task aims to classify whether a given image and text pair is matched or not. The classification score of the ith image-text pair in a mini-batch, $\boldsymbol{q}_{\text{ITM}}^{(i)} \in \mathbb{R}^2$, is obtained by feeding the final joint embedding feature, $\boldsymbol{m}_0$, to a binary linear classifier $h_{itm}(\cdot)$, which is given by

$$\boldsymbol{q}_{\text{ITM}}^{(i)} = h_{itm}(\boldsymbol{m}_0^{(i)}) = h_{itm}(f_{mm}(\boldsymbol{V}_{\text{emb}}^{(i)}, \boldsymbol{T}_{emb}^{(i)})[0]). \quad (4)$$

The ITM loss is defined as

$$\mathcal{L}_{\text{ITM}} = \frac{1}{S} \sum_{i=1}^{S} H\left(\boldsymbol{y}_{\text{ITM}}^{(i)}, \boldsymbol{q}_{\text{ITM}}^{(i)}\right), \quad (5)$$

where $\boldsymbol{y}_{\text{ITM}}^{(i)}$ is a one-hot vector representing the ground-truth label, S denotes the number of image-text pairs, and $H(\cdot, \cdot)$ stands for the cross-entropy function.

Masked Language Modeling (MLM) MLM aims to predict the masked text by employing both the image and the contextualized text. Following BERT [7], the text tokens of the ith text from the image-text pair, $\boldsymbol{T}^{(\mathrm{i})}$, are randomly masked out with a probability of 15%, and then the special token [MASK] replaces the selected tokens, which results in $\hat{\boldsymbol{T}}^{(i)}$. Given the output of multi-modal encoder, the probability of a masked token over the predefined vocabulary, $\boldsymbol{q}_{\text{MLM}}^{(i)} \in \mathbb{R}^{30522}$ is obtained by feeding the joint embedding feature of the masked text token, $\boldsymbol{m}_{\text{mask}}^{(i)} \in \boldsymbol{M}_{\text{emb}}^{(i)}$, to a linear classifier $h_{mlm}(\cdot)$ as

$$\boldsymbol{q}_{\text{MLM}}^{(i)} = h_{mlm}(\boldsymbol{m}_{\text{mask}}^{(i)}). \quad (6)$$

Then, the MLM objective is defined as follows:

$$\mathcal{L}_{\text{MLM}} = \frac{1}{S} \sum_{i=1}^{S} \mathbb{E}_{(\boldsymbol{I}, \hat{\boldsymbol{T}}) \sim \mathcal{D}} \left[H\left(\boldsymbol{y}_{\text{MLM}}^{(i)}, \boldsymbol{q}_{\text{MLM}}^{(i)}\right) \right]. \quad (7)$$

4. Method

This section describes the proposed approach, including text-driven soft feature masking, focal image-text contrastive learning, and multi-modal data augmentation. Figure 2 illustrates the overall procedure and network architecture of our approach.

4.1. Text-Driven Soft Feature Masking

Our model is pretrained based only on image-caption pairs and is prone to overfit discriminative regions in images as often observed in image recognition tasks. Hence, the model may not be able to understand the details of input images despite the availability of matched captions.

To tackle this issue, we propose to learn multi-modal representations by matching the masked image embedding

and the corresponding text embeddings. Contrary to a common hard masking strategy, we generate soft masks by using the concept of Grad-CAM [35], which highlights the relevant parts in an image on the output of the learned model. Because it is critical to maintain crucial information in an input image for accurate matching between images and texts while augmenting hard examples by partly masking informative regions, we believe that soft masking based on a word-conditional Grad-CAM is a reasonable idea and adopt it as a new operation in our algorithm. As shown in the right-hand side of Figure 2, the word-conditional Grad-CAM effectively captures the discriminative regions, which correspond to the objects in general, even for stop-words.

Specifically, for the ith image-text pair in a mini-batch, we first identify salient regions in the image by computing the Grad-CAM of the initial image-text matching score, $\boldsymbol{q}_{\text{ITM}}^{(i)}$, using the cross-attention map of the image and text embeddings given by the multi-modal fusion encoder. Let $A_k^{(i)} \in \mathbb{R}^{(L+1) \times (N+1)}$ denote the cross-attention map in the kth transformer block of the multi-modal encoder. We obtain the Grad-CAM denoted by $A_{\text{GCAM}}^{(i)} \in \mathbb{R}^{(L+1) \times (N+1)}$ of the positive matching score with respect to $A_k^{(i)}$ as

$$A_{\text{GCAM}}^{(i)} = \frac{1}{K} \sum_{k=1}^{K} \text{ReLU}\left(\frac{\partial q_{\text{ITM}}^{+(i)}}{\partial A_k^{(i)}} \odot A_k^{(i)} \right), \quad (8)$$

where $q_{\text{ITM}}^{+(i)}$ denotes the positive matching score of the ITM task, which is equal to $\boldsymbol{q}_{\text{ITM}}^{(i)}[1]$. Note that we aggregate the Grad-CAM maps from all cross-attention layers to obtain a more accurate Grad-CAM.

Among $L + 1$ word tokens, including the [CLS] token, we randomly sample a single word index from a uniform distribution and obtain the corresponding word-conditional Grad-CAM. The rationale behind the random selection is to boost stochasticity and obtain attention for diverse regions during training. Let i_w be the index of the sampled word. Then, the text-driven soft mask for an image embedding, $\mathcal{M}_{\text{soft}}^{(i)} \in \mathbb{R}^{N+1}$, is given by

$$\mathcal{M}_{\text{soft}}^{(i)} = \mathbb{1} - \hat{A}_{\text{GCAM}}^{(i)}[i_w], \quad (9)$$

where $\hat{A}_{\text{GCAM}}^{(i)}[i_w]$ is the normalized value of $A_{\text{GCAM}}^{(i)}[i_w]$ with min-max clipping. We compute the masked embedding $\hat{\boldsymbol{V}}_{\text{emb}}^{(i)}$, which is given by

$$\hat{\boldsymbol{V}}_{\text{emb}}^{(i)} = \mathcal{M}_{\text{soft}}^{(i)} \odot \boldsymbol{V}_{\text{emb}}^{(i)}, \quad (10)$$

and feedforward it to the multi-modal encoder to obtain new multi-modal joint embedding, $\hat{\boldsymbol{M}}_{\text{emb}}^{(i)}$. By classifying $\hat{\boldsymbol{m}}_0^{(i)}$ with h_{itm}, we obtain the prediction scores for the masked soft feature as

$$\hat{\boldsymbol{q}}_{\text{ITM}}^{(i)} = h_{itm}(\hat{\boldsymbol{m}}_0^{(i)}) = h_{itm}\left(f_{mm}(\hat{\boldsymbol{V}}_{\text{emb}}^{(i)}, \boldsymbol{T}_{\text{emb}}^{(i)})[0] \right). \quad (11)$$

Then, the ITM loss of the soft masked feature is defined as

$$\mathcal{L}_{\text{ITM}}^* = \frac{1}{S'} \sum_{i=1}^{S'} H(\boldsymbol{y}_{\text{ITM}}^{(i)}, \hat{\boldsymbol{q}}_{\text{ITM}}^{(i)}), \tag{12}$$

where $\boldsymbol{y}_{\text{ITM}}^{(i)}$ is the same ground-truth vector used in (5) and S' is the number of positive pairs in a mini-batch.

4.2. Focal Image-Text Contrastive Learning

To boost the regularization effect from hard examples, we design a variant of the ITC loss, which is similar to the focal loss [26]. We define the focal ITC loss as

$$\mathcal{L}_{\text{ITC}}^* = \tag{13}$$
$$-\frac{1}{2B} \sum_{i=1}^{B} \left[(1 - p_{v2t}^{(i)})^\gamma \log p_{v2t}^{(i)} + (1 - p_{t2v}^{(i)})^\gamma \log p_{t2v}^{(i)} \right],$$

where γ is a modulating factor to reduce weights for easy samples and is set to 2 following the original version of the focal loss. The proposed focal ITC loss alleviates the overfitting to easy examples while handling the class imbalance issue effectively. Since large-scale image-caption corpora are composed of multiple datasets with large domain gaps, a model distinguishes a lot of samples from different datasets easily, as also revealed in [6]. Hence, in this sense, the focal ITC loss is particularly useful for our training scheme.

4.3. Multi-Modal Data Augmentation

We perform multi-modal data augmentations (MM-DAs), which include diverse strong augmentations proposed in [43] and binary masking to captions for the ITM task. We propose to generate hard positive samples by applying strong augmentations on both modalities. We crop images randomly to resolution 256×256 and apply RandAugment [5]. We also perform color distortions such as random color jittering, random grayscale conversion, and random Gaussian blur. On the text side, we replace $\boldsymbol{T}$ with $\hat{\boldsymbol{T}}$. As illustrated in Section 5.5, such a simple augmentation strategy turns out to improve performance despite the preconception that it deteriorates semantic relations on vision-language tasks. Note that the data augmentation also affects the ITC. We reduce the computational cost of the whole framework because we now perform forward the text encoder just once unlike the previous works [21, 43][1]. Such the saved cost allows us to compute another forward pass for soft masking and optimizing $\mathcal{L}_{\text{ITM}}^*$ in (12). The analysis on computation cost is presented in Section 5.6.

---

[1]While the papers described their models as if they learn ITM with augmented images and masked texts, we found that their ITM takes text without masks and they have to feedforward text features twice with and without masks because MLM adopts masked text as its input.

4.4. Training Objective

The final objective function of our algorithm, denoted by $\mathcal{L}_{\text{Final}}$, is given by

$$\mathcal{L}_{\text{Final}} = \mathcal{L}_{\text{ITM}} + \mathcal{L}_{\text{ITC}}^* + \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{ITM}}^*. \tag{14}$$

4.5. Discussion

Our work introduces the aforementioned three components to utilize diversely augmented image-text pairs and facilitate large-scale multi-modal pretraining. We employ a novel text-driven soft feature masking strategy to populate examples with diverse aspects, where the discriminative parts of an image are softly masked, at the intermediate feature level. By feeding the features covered by soft masks, the model learns to see complementary regions without completely sacrificing major information whereas conventional hard masking often loses critical contents. Next, we encourage the model to pay more attention to the hard examples, by adopting a focal version of the image-text contrastive learning objective. Last, we augment data in the input level and make the model learn with hard image-text pairs. Although employing imperfectly aligned image-text pairs is counter-intuitive and not explored much, our ideas align with recent studies [15, 25] that show perfect modality alignment does not guarantee performance improvement in downstream task.

5. Experiments

This section presents the experimental results of our approach, referred to as SoftMask++, on various downstream tasks. We also demonstrate the effectiveness of our ideas via several ablation studies.

5.1. Pretraining Datasets

We utilize large-scale image-text corpora for pretraining, which include MS-COCO [27], Visual Genome [18], SBU Captions [31], and Conceptual Captions [36]. The combined dataset consists of 4M images and 5.1M image-text pairs as in the one used for previous works [4, 9, 21, 43].

5.2. Downstream Tasks

We briefly describe several vision-language downstream tasks used for the evaluation of our pretraining approach. Refer to the supplementary document for more details.

Image-Text Retrieval (ITR) We assess the effectiveness of our method on Image-Text Retrieval task, which consists of two subtasks, which are image-to-text retrieval (TR) and text-to-image retrieval (IR). We evaluate the proposed approach on the Flickr30K [32] and MS-COCO [27] benchmarks using two different settings following the previous protocol [9, 21, 43]. In the first setting, we fine-tune the pretrained model using the training data in each dataset. We

Table 1. Performance comparison of fine-tuned image-text retrieval on the Flickr30K and MS-COCO datasets. Our approach is named as SoftMask++: SoftMask together with Focal ITC and MMDA. For completeness, we also provide the results of ALIGN [14], which uses 1.8B image-text pairs (1.2B unique images) for pretraining. The bold-faced numbers indicate the best performance.

| Method | #Img | Flickr30K (1K) | | | | | | MS-COCO (5K) | | | | | |
| | | Text Retrieval | | | Image Retrieval | | | Text Retrieval | | | Image Retrieval | | |
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UNITER [4]	4M	87.3	98.0	99.2	75.6	94.1	96.8	65.7	88.6	93.8	52.9	79.9	88.0
VILLA [10]	4M	87.9	97.5	98.8	76.3	94.2	96.8	—	—	—	—	—	—
OSCAR [24]	4M	—	—	—	—	—	—	70.0	91.1	95.5	54.0	80.8	88.5
ViLT [17]	4M	83.5	96.7	98.6	64.4	88.7	93.8	61.5	86.3	92.7	42.7	72.9	83.1
UNIMO [23]	4M	89.7	98.4	99.1	74.7	93.5	96.1	—	—	—	—	—	—
SOHO [13]	200K	86.5	98.1	99.3	72.5	92.7	96.1	66.4	88.2	93.8	50.6	78.0	86.7
ALBEF [21]	4M	94.3	99.4	99.8	82.8	96.7	98.4	73.1	91.4	96.0	56.8	81.5	89.2
TCL [43]	4M	94.9	99.5	99.8	84.0	96.7	**98.5**	75.6	92.8	**96.7**	59.0	83.2	89.9
CODIS [9]	4M	95.1	99.4	**99.9**	83.3	96.1	97.8	75.3	92.6	96.6	58.7	82.8	89.7
VinVL [45]	6M	—	—	—	—	—	—	75.4	92.9	96.2	58.8	83.5	90.3
SoftMask++ (ours)	4M	**95.4**	**99.7**	**99.9**	**84.6**	**96.8**	**98.5**	**76.6**	**93.5**	96.6	**60.2**	**83.7**	**90.5**
ALIGN [14]	1.2B	95.3	99.8	100.0	84.9	97.4	98.6	77.0	93.5	96.9	59.9	83.3	89.8

Table 2. Performance comparison of zero-shot image-text retrieval on the Flickr30K and MS-COCO datasets. Our approach is named as SoftMask++: SoftMask together with Focal ITC and MMDA. For completeness, we also provide the results of ALIGN [21], which uses 1.8B image-text pairs (1.2B unique images) for pretraining. The bold-faced numbers indicate the best performance.

| Method | #Img | Flickr30K (1K) | | | | | | MS-COCO (5K) | | | | | |
| | | Text Retrieval | | | Image Retrieval | | | Text Retrieval | | | Image Retrieval | | |
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UNITER [4]	4M	80.7	95.7	98.0	66.2	88.4	92.9	64.1	87.7	93.3	48.8	76.7	85.8
ViLT [17]	4M	73.2	93.6	96.5	55.0	82.5	89.8	56.5	82.6	89.6	40.4	70.0	81.1
CLIP [33]	400M	88.0	98.7	99.4	68.7	90.6	95.2	58.4	81.5	88.1	37.8	62.4	72.2
ALIGN [14]	1.2B	88.6	98.7	99.7	75.7	93.8	96.8	58.6	83.0	89.7	45.6	69.8	78.6
ALBEF [21]	4M	90.5	98.8	99.7	76.8	93.7	96.7	68.7	89.5	94.7	50.1	76.4	84.5
TCL [43]	4M	93.0	99.1	99.6	79.6	**95.1**	97.4	71.4	90.8	95.4	53.5	79.0	87.1
CODIS [9]	4M	91.7	**99.3**	**99.8**	79.7	94.8	97.3	71.5	91.1	95.5	53.9	79.5	87.1
SoftMask++ (ours)	4M	**93.4**	**99.3**	**99.8**	**80.1**	94.9	**97.7**	**72.3**	**91.5**	**95.7**	**54.1**	**79.8**	**87.3**

also report the zero-shot performance, where the pre-trained model is directly evaluated on the test data without fine-tuning. Note that, for zero-shot evaluation on Flickr30K, we exploit the model fine-tuned on MS-COCO, following the convention [9, 21, 43].

Visual Entailment (VE) This is a fine-grained visual reasoning task that aims to predict whether an image semantically entails, contradicts, or is neutral to a text. We use the SNLI-VE [41] dataset, which consists of 30K, 1K, and 1K images as training, validation, and testing sets, respectively. We predict the probability by feeding the joint embedding, m_0, to a linear classifier.

Natural Language Visual Reasoning (NLVR) This task learns to reason about whether the text description is relevant to a pair of input images. Since the task requires two images as inputs, we modify our model by duplicating transformer blocks following [21, 43] and perform an additional pretraining step to accommodate the multi-modal encoder for the image pair. We evaluate our model on the NLVR^2 dataset, which consists of 86K, 7K, and 7K examples for training, development, and test splits, respectively.

Visual Question Answering (VQA) Visual question answering (VQA) is a task to find an answer for a question about an input image, which requires understanding local and global context of an image as well as a question. We conduct experiments on VQA2.0 dataset [11], which is based on the images collected from the MS-COCO dataset. We formulate the VQA task as an answer generation problem, following [21, 43]. The answer decoder is initialized with the pretrained weights from the multi-modal encoder and is fine-tuned on the training datasets to generate answers from the pre-defined candidates.

5.3. Results on Image-Text Retrieval

Table 1 presents the overall results of the proposed algorithm and other baselines on Flickr30K and MS-COCO for the fine-tuning setting, where our approach outperforms the baseline methods in most cases, especially for R@1.

Table 3. Performance comparison on various vision-language tasks, including VE, NLVR^2, and VQA. The bold-faced numbers indicate the best performance. Note that O.D. denotes object detector and the methods with the remark exploit the pretrained object detector as visual encoders, and D.F. represents the detector-free models.

	Method	#Img	SNLI-VE		NLVR^2		VQA	
			val	test	dev	test-P	dev	std
O.D.	OSCAR [24]	4M	—	—	78.1	78.4	73.2	73.4
	UNITER [4]	4M	78.6	78.3	77.2	77.9	72.7	72.9
	UNIMO [23]	4M	80.0	79.1	—	—	73.3	74.0
	VILLA [10]	4M	79.5	79.0	78.4	79.3	73.6	73.7
	VinVL [45]	6M	—	—	82.1	83.1	75.9	76.1
D.F.	ViLT [17]	4M	—	—	75.7	76.1	71.3	—
	ALBEF [21]	4M	80.1	80.3	80.2	80.5	74.5	74.7
	TCL [43]	4M	80.5	80.3	80.5	81.3	74.9	74.9
	CODIS [9]	4M	80.5	80.4	80.5	80.8	**75.0**	74.9
	SoftMask++ (ours)	4M	**80.9**	**80.6**	**80.6**	**81.6**	**75.0**	**75.1**

Table 4. Ablation study results of the image-text retrieval task on the MS-COCO dataset. The baseline algorithm without all the three components, corresponding to the first row of this table, is the same as ALBEF [21]. The bold-faced numbers indicate the best performance in each column.

SoftMask	Focal ITC	MMDA	TR@1	IR@1
			73.10	56.80
✓			75.74	58.24
	✓		75.66	58.85
		✓	75.26	58.85
✓	✓		76.06	59.54
✓		✓	75.98	59.44
	✓	✓	76.16	59.62
✓	✓	✓	**76.62**	**60.15**

Table 5. Image-text retrieval results on MS-COCO by varying masking strategy on visual features. The bold-faced numbers indicate the best performance.

Method	TR@1	IR@1
RandMask ($p = 0.3$)	75.82	59.52
RandMask ($p = 0.5$)	76.22	59.74
SoftMask (Ours)	**76.62**	**60.15**

For completeness, we provide the results from ALIGN [14], which uses 1.8B image-text pairs (1.2B unique images) for pretraining. The results show that our model successfully transfer knowledge for the image-text retrieval task.

Table 2 illustrates the overall results of the proposed algorithm and other baselines in a zero-shot setting on the Flickr30K and MS-COCO datasets. Our approach outperforms other methods in most experimental settings and also surpasses ALIGN [14], which employs about $300\times$ times more pretraining images. The results illustrate that the proposed framework learns better generalizable representations than the compared methods.

5.4. Results on Other Downstream Tasks

Table 3 summarizes the results of the proposed algorithm on the VE, NLVR, and VQA tasks. Among the methods without pretrained object detectors, our algorithm achieves state-of-the-art performance for all evaluation metrics. Since the evaluated tasks employ different types of objective functions for training, the results show the generalizability of our multi-modal representation learning technique. Another important feature of these tasks is that they require fine-grained reasoning about images and texts. The results demonstrate that the representations learned by the proposed soft masking approach and the other regularization techniques are effective for fine-grained recognition. Note that VinVL [45] exploits about 6M images, which is larger than ours, to train the object detector before multi-modal representation learning.

5.5. Ablation Study

We perform several ablation studies on the image-text retrieval task with the MS-COCO dataset and report R@1 performance to analyze the effectiveness of each compo-

nent of our approach. From the results in Table 4, we have the following observations. First, our soft masking strategy enhances the overall transferability, by generating diverse visual features as we hypothesized. Second, the proposed focal ITC loss boosts performance by large margins. These observations support our claim that concentrating on the hard examples and focusing less on easy ones alleviate overfitting while handling the class imbalance issue effectively. Last, multi-modal data augmentation (MMDA) delivers strong regularization effect, indicating that adopting augmented examples with diverse aspects for vision-language pretraining has a positive effect.

Soft vs. Random Masking Table 5 illustrates the comparison between our text-driven soft masking (SoftMask) and the random hard masking (RandMask) strategy. For RandMask, we randomly remove regions in an image embedding with pre-defined ratios of 0.3 and 0.5. SoftMask outperforms the hard RandMask, which is mainly because soft masking manages to populate diverse and, more importantly, semantically meaningful visual embeddings with varying rates given by Grad-CAM.

SoftMask for MLM We conduct additional experiments to investigate whether our soft-masked visual features are also beneficial for the MLM task. Table 6 presents the results from the MLM loss with the soft-masked features, where we observe that SoftMask is not particularly effective for MLM and fails to improve performance in downstream

(a) A dog on a beach carrying a frisbee.

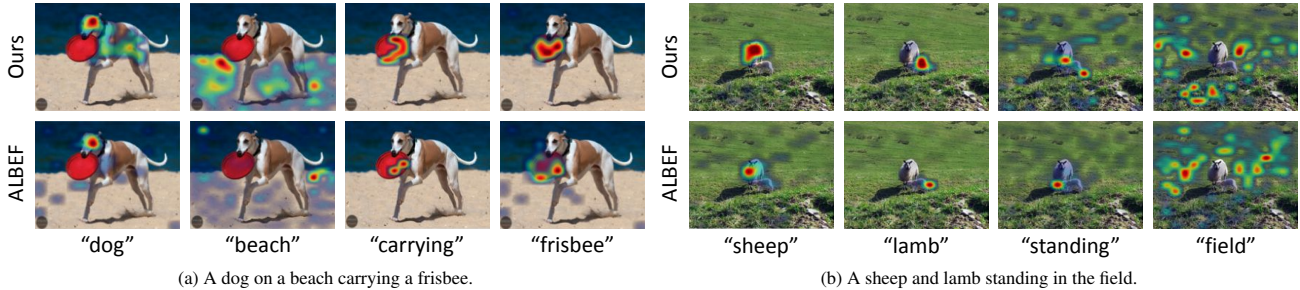(b) A sheep and lamb standing in the field.

Figure 3. Word-conditional Grad-CAM visualization of our method and ALBEF [21] after pretraining.

Table 6. Performance comparisons in the Image-text retrieval task on the MS-COCO dataset by applying SoftMask to the ITM and MLM pretraining tasks. Our default setting adopts SoftMask only for ITM. The bold-faced numbers indicate the best performance.

$\mathcal{L}^*_{ITM}$	$\mathcal{L}^*_{MLM}$	TR@1	IR@1
		76.16	59.62
$\checkmark$		**76.62**	**60.15**
$\checkmark$	$\checkmark$	76.21	59.84

Table 7. Computation cost per a single pretraining iteration compared to ALBEF [21]. We measure computational cost on a single NVIDIA Quadro RTX GPU with 128 batch size per GPU.

Methods	Time (sec/it)	GPU Mem. (GB/GPU)
ALBEF [21]	1.94	41.5
Ours w/o SoftMask	1.91	38.7
SoftMask++ (ours)	1.98	41.4

tasks according to our experiments. This is partly because the reconstruction task is not well addressed when the signals from both modalities are imperfect. Meanwhile, ITM enjoys the regularization effect since the task depends more on global information in both modalities.

5.6. Computational Complexity

Two minor modifications of our backbone network, AL-BEF [21] affect computational cost. First, we unify the forward path in the ITM and MLM tasks by using the same masked text input and reduce computation even with performance improvements. Second, our soft masking strategy incurs an additional computation for computing Grad-CAM and feedforwarding soft-masked features through the fusion encoder. Table 7 presents the computational analysis compared to the ALBEF [21] baseline. On a single NVIDIA Quadro RTX GPU, ALBEF requires 1.94 seconds per iteration while our algorithm requires 1.98s with batch size 128.

5.7. Qualitative Results

Figure 3 illustrates qualitative comparisons between our algorithm and ALBEF [21], where we provide word-conditional Grad-CAM visualizations of two image-caption pairs from the MS-COCO dataset. The followings are the lessons from the observation of the qualitative results. First, our model provides more accurate and comprehensive coverage of the objects, which are relevant to the corresponding caption. For instance, ALBEF only highlights the head of the dog, which is the most discriminative part, while our model is also activated on its body. Second, our model obtains more accurate activations for action attributes. For the action "standing", ALBEF only highlights the feet of the

sheep while ours recognizes the feet of both animals. Last, our model may be biased toward certain objects or background as, for example, Grad-CAM of "standing" fires on fields, if the most salient parts are heavily masked.

6. Conclusion

In this paper, we presented a novel visual-linguistic representation learning framework based on the explainable soft feature masking strategy and the regularizations via diversifications. We proposed a novel language-driven soft feature masking strategy to populate visual features with various aspects, where the most discriminative parts of an image are softly masked based on the stochastically estimated contribution of each local region to the matching score. We also introduced a focal loss for the image-text contrastive learning objective, which addresses the inherent limitations of the overfitting and bias issues. Last, we manipulate multi-modal data augmentation strategies for constructing more diversified samples by applying text maskings and rendering distortions on images. Our algorithm achieves outstanding performance compared to existing vision-language pretraining methods on various vision-language downstream tasks.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 1, 2

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 1

[3] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *CVPR*, 2019. 1, 2

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. In *ECCV*, 2020. 1, 2, 5, 6, 7

[5] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *NeurIPS*, 2020. 5

[6] Quan Cui, Boyan Zhou, Yu Guo, Weidong Yin, Hao Wu, Osamu Yoshie, and Yubo Chen. Contrastive vision-language pre-training with limited resources. In *ECCV*, 2022. 2, 5

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 1, 2, 4

[8] 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. In *ICLR*, 2021. 2

[9] Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Multi-modal alignment using representation codebook. In *CVPR*, 2022. 2, 5, 6, 7

[10] 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, 6, 7

[11] 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

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1

[13] 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, 6

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 6, 7

[15] Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Understanding and constructing latent modality structures in multi-modal representation learning. *arXiv preprint arXiv:2303.05952*, 2023. 2, 5

[16] Yash Kant, Abhinav Moudgil, Dhruv Batra, Devi Parikh, and Harsh Agrawal. Contrast and classify: Training robust vqa models. In *ICCV*, 2021. 2

[17] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 1, 2, 6, 7

[18] 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*, 123(1):32–73, 2017. 1, 5

[19] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, 2018. 2

[20] Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval with text feedback. In *CVPR*, 2021. 1

[21] 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, 3, 5, 6, 7, 8

[22] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, 2019. 2

[23] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *ACL*, 2021. 6, 7

[24] 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. 1, 2, 6, 7

[25] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022. 2, 5

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 2, 5

[27] 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. 1, 5

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 1, 2

[29] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *CVPR*, 2016. 1

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[31] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 1, 5

[32] 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. 1, 5

[33] 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. In *ICML*, 2021. 6

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2

[35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 1, 4

[36] 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, 5

[37] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *ACL*, 2019. 1

[38] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 2

[39] Ruixue Tang, Chao Ma, Wei Emma Zhang, Qi Wu, and Xiaokang Yang. Semantic equivalent adversarial data augmentation for visual question answering. In *ECCV*, 2020. 2

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 1

[41] 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. 1, 6

[42] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022. 1

[43] 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. 2, 5, 6, 7

[44] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019. 1, 2

[45] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, 2021. 2, 6, 7