Revising Image-Text Retrieval via Multi-Modal Entailment

Anonymous ACL submission

Abstract

An outstanding image-text retrieval model depends on high-quality labeled data. While the builders of existing image-text retrieval datasets strive to ensure that the caption matches the linked image, they cannot prevent a caption from fitting other images. We observe that such a many-to-many matching phenomenon is quite common in the widely-used retrieval datasets, where one caption can describe up to 178 images. These large matching-lost data not only confuse the model in training but also weaken the evaluation accuracy. Inspired by visual and textual entailment tasks, we propose a multi-modal entailment classifier to determine whether a sentence is entailed by an image plus its linked captions. Subsequently, we revise the image-text retrieval datasets by adding these entailed captions as additional weak labels of an image and develop a universal variable learning rate strategy to teach a retrieval model to distinguish the entailed captions from other negative samples. In experiments, we manually annotate an entailment-corrected image-text retrieval dataset for evaluation. The results demonstrate that the proposed entailment classifier achieves about 78% accuracy and consistently improves the performance of image-text retrieval baselines.

1 Introduction

Image-text retrieval aims to retrieve items through visual or semantic information. It contains two subtasks: image retrieval and text retrieval, depending on which modality is used as the retrieved target. Image-text retrieval has been widely adopted in various applications, such as the retrieval of commodity pictures given textual descriptions. Most image-text retrieval approaches (Li et al., 2019c,a; Tan and Bansal, 2019; Li et al., 2019b; Su et al., 2020) focus on mapping features of image and text modalities into a common semantic space. Notably, recent studies (Li et al., 2020; Chen et al., 2020; Jia et al., 2021; Radford et al., 2021; Li et al., 2021a) have shown that Vision-and-Language Pre-training (VLP) can effectively learn general representations and achieves high performance on this task.

Figure 1: Examples of images and texts from MSCOCO dataset. While all of captions can describe the two images, only image-text pairs with the same color are marked as positive pairs.

Image-text retrieval relies on curated training datasets that are usually expensive and sometimes even require expert knowledge to acquire. Common image-text retrieval datasets, including Flickr8K (Rashtchian et al., 2010), Flickr30K (Young et al., 2014), Multi30k (Elliott et al., 2016) and MSCOCO (Lin et al., 2015), are constructed through manually writing a few descriptive captions for each image using crowdsourcing. Therefore, it is only ensured that the image and its descriptive captions are matched when annotated. However, the possible associations between an image and other captions in the dataset are not fully considered. Taking Figure 1 as an example, two images depicting the same scene have their different text descriptions, which can also be used to describe each other. Such a many-to-many matching phenomenon is quite common in retrieval datasets. For example, in MSCOCO, we find that 89 captions can describe one image while this number amazingly reaches 178 on the text side (refer to Section 5 for more details). Unfortunately, the cross-matched image-text pairs with similar semantics are typically regarded as negative examples. As we know, treating semantically matched image-text pairs as negative in training will increase their distance in vector space and thus reduce the quality of representation learning. Meanwhile, marking them as errors in evaluation leads to a significant false negative rate.

This paper proposes an automatic solution to handle the many-to-many matching problem in the retrieval datasets. Our solution recognizes this kind of relationship and utilizes the relationship in training. We argue that if an image and its descriptive captions entail the meaning of a sentence, this sentence should be able to describe the image. Inspired by the tasks of visual entailment (Xie et al., 2019a) and textual entailment (Glockner et al., 2018), we propose a multi-modal entailment classifier to recognize the entailment relationship between a caption and an image combined with its descriptive captions. To fully utilize the external textual and visual entailment data, our entailment model supports various forms of input, including text-text, image-text, and image&text-text. We modify existing models (Li et al., 2021a; Devlin et al., 2019) to conduct textual entailment and visual entailment, and combine the hidden states of textual/visual modules to produce the final multi-modal entailment result. Next, we use this entailment model to find the entailed image-text pairs in the retrieval datasets. During training, we treat these entailed pairs as additional weak positive samples and set a small learning rate for them. This learning strategy can be used for any retrieval model without changing its internal structure.

In order to verify the proposed entailment model, we manually annotated an entailment-corrected dataset containing 2k image-text pair samples from MSCOCO and Flickr30K. Results show that our entailment classifier achieves about 78% accuracy. Moreover, trained on image-text pairs revised by our entailment classifier, the retrieval models uniformly achieve a performance improvement in both retrieval and entailment evaluations.

The contributions of this paper can be summarized as follows:

- We utilize multi-modal entailment to handle the many-to-many matching problem in image-text retrieval datasets and annotate an entailment-corrected dataset for evaluation[1].
- We propose a strong multi-modal entailment classifier to determine the entailed image-text pairs in the retrieval datasets automatically.
- We develop a universal entailment-enhanced learning strategy to consistently to improve retrieval models' matching performance consistently.

[1]Code and the dataset will be released in the final version.

2 Related Work

2.1 Image-Text Retrieval Datasets

Early image-text datasets include Flickr8K (Rashtchian et al., 2010) and Flickr30K (Young et al., 2014). Inspired by them, Lin et al. (2015) builds a larger Microsoft Common Objects in COntext (MSCOCO) Caption dataset. A number of datasets subsequently emerge such as Multi30k (Elliott et al., 2016), Conceptual Captions (Sharma et al., 2018) and RedCaps (Desai et al., 2021). Notably, Conceptual Captions and RedCaps are built through web crawling, while others are constructed by manually writing a few descriptive captions for each image using crowd-sourcing. All these datasets only ensure relationships between images and texts created for them and ignore possible associations of external image-text pairs.

Some recent works have been aware of this problem and attempted to introduce many-to-many correspondences for image-text datasets. Criss-Crossed Caption (CxC) (Parekh et al., 2021) and Extended COCO Validation (ECCV) (Chun et al., 2022) datasets are built through manually annotating sampled MSCOCO image-text pairs with similarity scores or categories. However, due to expensive labor costs and unscalable annotations, it is challenging to construct a large-scale dataset for training. Moreover, the human similarity score does not entirely fit the retrieval task, and even image-text pairs with high scores cannot always be taken as positive samples. For example, in the CxC dataset, the caption "A couple of birds that are walking on some sand." matches the image with a single seagull.

2.2 Textual Entailment and Visual Entailment

Textual entailment (Dagan et al., 2005), often used as a benchmark to measure the ability of language understanding (Dagan et al., 2005; Bowman et al., 2015a), has been a hot research topic in the NLP area. In the last few years, with the advancement of deep learning, the study of textual entailment is gradually being carried out on some large-scale data such as SNLI (Bowman et al., 2015b), SciTaiL (Khot et al., 2018), MNLI (Williams et al., 2017), and XNLI (Conneau et al., 2018). In addition, textual entailment in the context of the few-shot scenario has also been much studied, like UFO-ENTAIL (Yin et al., 2020).

Inspired by textual entailment, Xie et al. (2019b) proposes visual entailment task to determine the entailment between a given image and text pair. They annotate a dataset SNLI-VE by linking SNLI to Flickr30K. In recent studies, it has often been treated as a downstream task of Vision-and-Language Pre-training(VLP) model (Huang et al., 2021; Li et al., 2021b; Wang et al., 2021, 2022). In addition, Ilharco et al. (2021) proposes a multi-modal entailment dataset, but the dataset is not well adapted to our multi-modal entailment model.

3 Multi-Modal Entailment Classifier

The proposed multi-modal entailment classifier is used to recognize whether a sentence is entailed by an image plus its captions. We utilize the classifier to construct the entailment-revised retrieval dataset for training automatically. Figure 2 shows the model structure. It contains a visual entailment module and a textual entailment module and combines the hidden states of the two modules to predict the final multi-modal entailment category. Our model supports three types of input premises: an image, text, and a combination of image and text. Note that to be adaptable to downstream image-text retrieval tasks, we only classify the relationship into entailment or non-entailment, rather than the traditional entailment task with three categories: entailment, neutral, and contradiction. In the following description we use x^{p_v} and x^{p_t} for the image and text in premise, x^h for the text hypothesis and $y \in \{0, 1\}$ for the target where 1 means entailment and 0 means non-entailment. This section will illustrate how our model conducts the three types of entailment data.

3.1 Textual Entailment

In textual entailment, both the premise and hypothesis are textual sentences, namely the input $= (x^{p_t}, x^h)$. We define this form of the task as text-text and adopt BERT (Devlin et al., 2019) as our backbone model.

Following the common practice, we pack two sentences x^{p_t} and x^h together as $([cls], x^{p_t}, [sep], x^h)$, where $[cls]$ and $[sep]$ are two special tags. Next, the packed texts are fed into the BERT model to get the entire representation:

$$h^t = BERT(x^{p_t}, x^h). \quad (1)$$

Like Choi et al. (2021), we just use the hidden state at the sentence tag ($[cls]$) to represent the entire input. On top of h_t, we add a simple multi-layer perceptron (MLP) classifier with two hidden layers to predict the final label:

$$p(\hat{y}|x^{p_t}, x^h) = softmax(MLP(h^t)). \quad (2)$$

where we adopt ReLU (Glorot et al., 2011) as the activation function for MLP. Notably, we use softmax rather than sigmoid for this binary classification task as we compare the two methods, and the results show that softmax is 1.8% higher than sigmoid.

3.2 Visual Entailment

In visual entailment, the premise is an image x_{p_v}, and the task form is defined as image-text. We adopt the structure of the state-of-art image-text retrieval model ALBEF (Li et al., 2021b) to encode x_{p_v} and x_h, namely:

$$h^v = ALBEF(x^{p_v}, x^h). \quad (3)$$

ALBEF consists of a 12-layer visual transformer (ViT) (Dosovitskiy et al., 2020) as the image encoder and a 6-layer transformer for both text encoder and multi-modal encoder. The cross-attention mechanism in a multi-modal encoder achieves an alignment between visual and textual modals. Similar to textual entailment, after a simple multi-layer perceptron with two hidden layers, we can get a distribution of prediction $\hat{y}$.

$$p(\hat{y}|x^{p_v}, x^h) = softmax(MLP(h^v)). \quad (4)$$

Referring to the practice of Liang et al. (2022) in ViT, we develop an image augment method to increment negative samples. Concretely, ViT will split an image into patches and encode them by self-attention mechanism (Vaswani et al., 2017). Intuitively, patches with higher attention scores should represent more significant regions and play a critical role in recognizing entailment relationships. For images of positive samples, we mask their partial patches with the highest score according to the attention matrix in ViT. Through this augment, original image-text pairs will become non-entailment

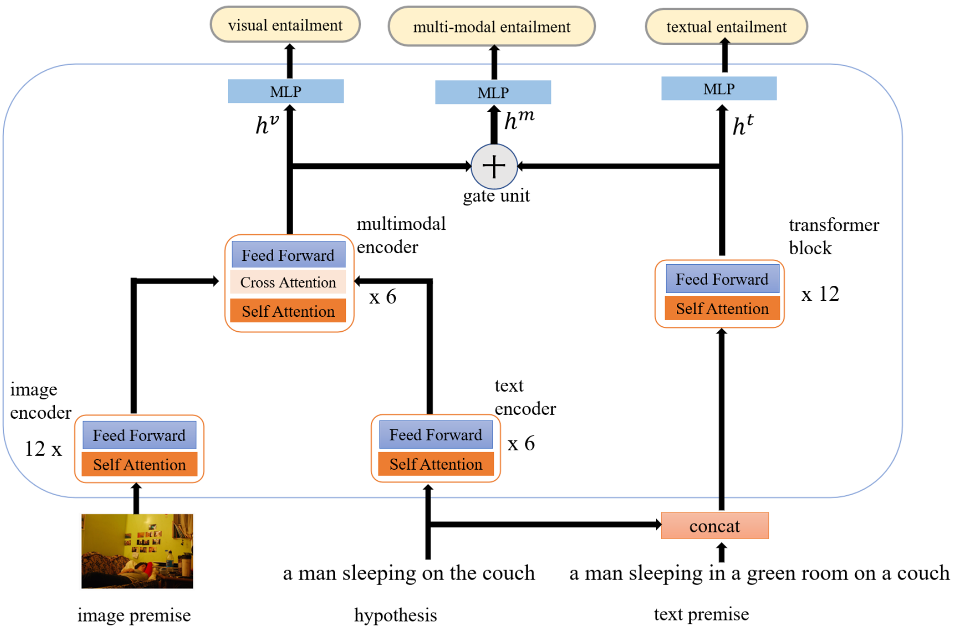

Figure 2: Illustration of our multi-modal entailment classifier. It consists of a visual entailment module and a textual entailment module. The result of multi-modal entailment is obtained by combining the hidden states of visual and textual entailment through a gate unit.

and supply negative samples. In the experiments, the masking ratio is a hyper-parameter we set as 0.4, and in each batch, we select up to 4 images for mask augment.

3.3 Multi-Modal Entailment

In textual entailment and visual entailment, the premise is just uni-modal. However, we actually need to check whether a sentence is entailed by an image plus its captions, and we define the form of the task when the premise input of our task is multi-modal as image&text-text. In this section, we want to combine textual and visual entailment for multi-modal entailment. The data pairs are defined as $(x^{p_v} + x^{p_t}, x^h)$. Briefly, we merge the captions of the same image to form x^{p_t}. Inspired by Xu et al. (2021), we want to build a gate unit to combine visual entailment and textual entailment to make a comprehensive judgment. Given the hidden states h^t and h^v computed in the above textual entailment and visual entailment modules, we propose a gate unit to merge them into multi-modal hidden states:

$$g^t = \sigma(W^t h^t + b^t), \tag{5}$$
$$g^v = \sigma(W^v h^v + b^v), \tag{6}$$
$$h^m = g^t \cdot h^t + g^v \cdot h^v. \tag{7}$$

where W^t, b^t, W^v, b^v are learnable parameters and σ is sigmoid function. Finally, the classification is done by a multi-layer perceptron classifier with two hidden layers:

$$p(\hat{y}|x^{p_v}, x^{p_t}, x^h) = softmax(MLP(h^m)). \tag{8}$$

We have tried to merge x^{p_v}, x^{p_t} and x^h directly using a multi-modal encoder instead of a gate unit, but this can easily cause memory overflow and make it impossible to separate visual and textual entailment.

3.4 Joint Learning

The learning process is driven by optimizing three objectives, corresponding to visual entailment $\mathcal{L}_v$, textual entailment $\mathcal{L}_t$ and multi-modal entailment $\mathcal{L}_m$ respectively.

$$\mathcal{L}_t = -\sum_i \log p(\hat{y}_i = y_i | x_i^{p_t}, x_i^h), \tag{9}$$
$$\mathcal{L}_v = -\sum_i \log p(\hat{y}_i = y_i | x_i^{p_v}, x_i^h), \tag{10}$$
$$\mathcal{L}_m = -\sum_i \log p(\hat{y}_i = y_i | x_i^{p_v} + x_i^{p_t}, x_i^h). \tag{11}$$

To facilitate training, we unify the input form of the model as the multi-modal task. To achieve this goal, we fill plain black images for textual entailment and empty premise strings for visual entailment. Meanwhile, we introduce three binary indicators $\theta_v, \theta_t, \theta_m$ to accumulate the related losses for back-propagation:

$$\mathcal{L}_{\text{all}} = \theta_t \mathcal{L}_t + \theta_v \mathcal{L}_v + \theta_m \mathcal{L}_m. \quad (12)$$

For textual entailment, only $\theta_t = 1$ and for visual entailment, only $\theta_v = 1$, while all the losses are used in multi-modal entailment.

4 Entailment-Enhanced Training for Retrieval Models

With the proposed multi-modal entailment classifier, we automatically detect the entailed image-text pairs in image-text retrieval datasets. Subsequently, we use entailed pairs in the following two aspects. On the one hand, current image-text retrieval models usually adopt negative sampling (Li et al., 2021a; Radford et al., 2021; Chen et al., 2020) to enforce dissimilar representations between non-golden image-text pairs. In the training process, we optimize negative sampling method by preventing sentences being selected as negative samples of entailed images. On the other hand, we regard these extra entailed image-text pairs as weak positives and propose a universal variable learning rate strategy to handle them. Specifically, assume that the learning rate of the golden positive examples during training is λ. Then we apply a smaller learning rate $\lambda^{'}$ to weak positives, where $\lambda^{'} = \alpha\lambda$ and $\alpha \in (0, 1)$ is a hyper-parameter.

In subsequent experiments, we empirically set α to 0.3. Considering the learning rate cannot be distinguished within the same batch, we assemble weak positives into an additional batch immediately after each normal batch. We preferentially select weak positives according to images in normal batch.

Through these two methods above, semantically related images and texts can be close to each other, without introducing too much noise in training. While optional methods include contrastive learning (Gutmann and Hyvärinen, 2010) and applying different weights on training loss for weak positives, they all need to modify models specifically and are not as universal as our strategy. Our experiments show that our methods can effectively enhance the entailment degree of the retrieval models, while keeping the retrieval performance.

5 Entailment-Corrected Dataset Annotation

We manually annotate an entailment-corrected dataset to evaluate the effects of our multi-modal entailment model. We select images and texts from the MSCOCO and Flikr30K test datasets to improve their diversity.

	Flickr30K	MSCOCO
Total pairs	1000	1000
Entailment	699	307

Table 1: Statistics of the entailment-corrected dataset.

Since most of the image-text pairs in retrieval datasets are semantically irrelevant and have no entailment relationship, we use a fine-tuned retrieval model ALBEF to get the top-30 text retrieval results as annotation candidates. After sampling images in the candidates, we randomly select one text for every image. In this way, the assembled image-text pairs usually hold high semantic association. We also add a small part of random image-text pairs to ensure the diversity of our dataset.

Seven graduate students are arranged for annotation. They must make an inference for the hypothesis sentence according to the given premise. To better use multi-modal information for entailment relationship classification, every premise in our dataset includes both image and its linked ground truth captions. More details of our dataset are shown in Appendix A.1. A hypothesis sentence can be regarded as entailment with its premise only if it meets the following two points: **(1)** This hypothesis sentence must clearly describe the content of the image premise without ambiguity. **(2)** This hypothesis sentence can be inferred from premise texts and cannot be contradictory to them all. All pairs not meeting the above conditions are regarded as negative examples. Testing on 30 identical samples, the Kappa score (Falotico and Quatto, 2015) of annotators reaches about 0.8, indicating high consistency. Finally, we get 1k labeled image-text pairs for Flickr30K and 1k for MSCOCO. Statistics about our dataset are shown in Table 1.

In addition, we use the same method to annotate some typical examples in the original MSCOCO testset. Surprisingly, we found that one plain caption "A picture of something and it appears like food" can match accord with up to 178 images with food, and the image with a person who is playing a baseball game can be depicted by according up to 89 captions. More details of these datasets are described in Appendix A.2. These huge numbers demonstrate the universality of the many-to-many matching phenomenon.

	Task	Dataset	Count
Train	TE	XNLI (Conneau et al., 2018)	400.2k
		MRPC (Dolan and Brockett, 2005)	5.8k
		RTE (Bentivogli et al., 2009)	2.7k
		STS-B (Cer et al., 2017)	7.2k
		QQP (Chen et al., 2017)	404.2k
		TS (Kauchak, 2013)	167.6k
	VE	SNLI-VE (Xie et al., 2019b)	529.5k
		Image Masking	132.3k
	MME	SNLI-VE	529.5k
		CXC (Parekh et al., 2021)	39.5k
		ECCV (Chun et al., 2022)	26.4k
		Image Masking	148.8k
Dev		SNLI-VE	17.8k
Test		Annotated Dataset	2k

Table 2: Statistics of datasets used in the multi-modal entailment task. TE, VE, and MME denote textual entailment, visual entailment, and multi-modal entailment, respectively.

Model	accuracy	precision	recall	$f_{0.5}$
Only TE	71.1	65.0	90.9	68.9
Only VE	72.3	66.9	87.5	70.2
OFA	73.3	67.4	89.6	70.9
Ours	78.1	80.2	74.3	**78.9**
w/o Image Masking	78.4	77.7	79.4	78.0
w/o VE Data	66.4	62.5	81.9	65.6
w/o TE Data	77.7	74.2	84.6	76.1
w/o BERT	76.5	72.4	85.1	74.6

Table 3: Performance (%) of different entailment models tested on our annotated dataset. w/o BERT means using a text encoder from ALBEF in the textual entailment.

6 Experiment

In this section, we present experimental results for our multi-modal entailment classifier and the proposed entailment-enhanced training for various retrieval models.

6.1 Datasets

Multi-Modal Entailment The datasets we used for textual entailment, visual entailment, multi-modal entailment are listed in Table 2. More details of these datasets are described in Appendix A.3. For visual entailment, we perform image data augment by masking critical patchs of images, as described in Section 3.2.

Image-Text Retrieval We consider two widely-used datasets for image-text retrieval tasks: MSCOCO and Flickr30K. Specifically, we adopt both datasets' widely used Karpathy split (Karpathy and Fei-Fei, 2015). The MSCOCO contains 113/5k/5k for train/validation/test, and the Flickr30K contains 29k/1k/1k images for train/validation/test. We present experimental results on MSCOCO 5K and Flickr 1K testsets.

6.2 Baseline Models

Multi-Modal Entailment We adopt BERT (Devlin et al., 2019) and ALBEF (Li et al., 2021a) as the backbone structure of textual entailment and visual entailment. Therefore we test the performance using each module. In addition, we introduce OFA (Wang et al., 2022), a state-of-the-art visual entailment classifier, as a comparison baseline.

Image-Text Retrieval We compare our variable learning rate strategy with some competitive image-text retrieval models, including ALBEF, CLIP (Radford et al., 2021) and UNITER (Chen et al., 2020). More details of these baseline models are described in Appendix A.4.

6.3 Evaluation Metrics

Multi-Modal Entailment The accuracy, precision, and recall of our annotated dataset are reported as the evaluation metrics, which are commonly used in the entailment task. Particularly, following the Zhao et al. (2018), We put more weight on precision and apply $F_{0.5}$ as our final evaluation metric.

Image-Text Retrieval As the common practice (Karpathy and Fei-Fei, 2015), we report the **Recall@K** (R@K) as evaluating metrics, which measures the fraction of times a correct item was found among the top K results. For text-retrieval (TR) and image-retrieval (IR), we report TR@1/5/10 and IR@1/5/10, respectively.

To quantitatively measure the relevance between retrieved texts and the query images, we propose a novel metric called **Entail@K** (E@k). E@K measures the averaged entailment ratio in the top-k retrieved items:

$$Entail@K = \frac{1}{K}\sum_{i=1}^{K} e_i(x), \tag{13}$$

where the binary indicator $e_i(x)$ equals 1 If and only if the i-th retrieved text is ground truth or has an entailment relationship with the query image x. Higher E@k values mean that the retrieved texts have a stronger descriptive and semantic association with the query images.

For the image-text pairs included in our entailment-corrected dataset, the relationship can be obtained directly. For the rest pairs, we use two ways to get their entailment labels. On the one hand, we sample some images and manually annotate the entailment relationship of their retrieval

Method	Flickr30K / MSCOCO					
	TR@1.	TR@5.	TR@10.	IR@1.	IR@5.	IR@10.
ALBEF	95.2 / 77.4	98.9 / 93.9	100.0 / 97.1	85.3 / 61.2	97.3 / 84.6	98.7 / 91.0
ALBEF$^{\#}$	+0.1 / +0.2	+0.6 / +0.2	-0.2 / +0.2	+0.2 / -0.3	+0.1 / -0.1	0.0 / -0.1
CLIP	89.2 / 64.5	97.4 / 85.9	99.4 / 92.2	74.4 / 47.4	93.5 / 74.4	96.7 / 83.4
CLIP$^{\#}$	**+1.6** / **+2.0**	**+1.4** / **+1.1**	+0.5 / +0.5	**+3.1** / **+1.5**	**+2.1** / **+1.4**	+0.9 / **+1.0**
UNITER	84.2 / 64.7	97.1 / 88.2	98.7 / 93.5	70.8 / 49.1	91.7 / 77.4	95.5 / 86.0
UNITER$^{\#}$	**-1.0** / +0.4	+0.1 / **+1.3**	+0.1 / 0.0	+0.4 / **+1.3**	+0.6 / +0.1	+0.7 / +0.9

Table 4: Performance (%) of different image-text retrieval models finetuned on Flickr30K and MSCOCO. The scores before and after the symbol "/" represent the evaluation results on original Flickr30K and MSCOCO testsets, respectively. "$\#$" denotes the model is trained with our entailment-enhanced strategy. The changes ≥ 1.0 are shown in **bold**.

results with the same rules as Section 5. On the other hand, we use our trained multi-modal entailment model to infer the relationship between image x and i-th text. The manual method is more accurate but requires too much cost, while the automatic way can quickly evaluate all the datasets.

In subsequent experiments, we randomly selected 50 common images with their retrieved top-10 texts from the text-retrieval results on both testset of Flickr30K and MSCOCO for manual annotation. We denote these manual entailment results with **E@M**.

Method	E@10	E@30	E@M
ALBEF	63.9	44.1	76.7
ALBEF$^{\#}$	**66.0**	**46.0**	**78.0**
CLIP	58.2	41.5	67.4
CLIP$^{\#}$	**60.9**	**43.2**	**75.5**
UNITER	44.9	27.2	73.1
UNITER$^{\#}$	**46.9**	**28.5**	**76.4**

Table 5: Performance of E@k on different retrieval models. E@M stands for evaluation by manually annotated 50 common samples. E@10/30 are averaged scores over Flickr30K and MSCOCO testsets.

6.4 Implementation Details

We mix the textual, visual, and multi-modal entailment data and train them together indiscriminately for our multi-modal entailment model. We found that this mixing strategy is much better than training separately. We trained the multi-modal entailment model with five epochs on 8 Amax-5000 GPUs with a batch size of 96. We use the AdamW (Loshchilov and Hutter, 2019) optimizer with a weight decay of 0.02 and initial learning rate 2e-5.

For image-text retrieval, due to models' scales, we set different batch sizes and initial learning rates for different models (i.e., 96/2e-5 for ALBEF, 1536/1e-5 for CLIP, 96/5e-5 for UNITER). We use the AdamW optimizer with a weight decay of 0.02.

6.5 Main Results

6.5.1 Results on entailment

The results of the entailment experiments are shown in Table 3. As can be seen, our multi-modal entailment model all the other baselines to a large extent. For instance, the $f_{0.5}$ is more than 8% larger than the state-of-the-art visual entailment model OFA. The results demonstrate our proposed multi-modal entailment model is more competitive than the traditional textual and visual entailment models. Meanwhile, the precision of annotated dataset has improved dramatically, which guarantees the possibility that the model will be used for automatic detection. In addition, we conduct a series of ablation experiments for training data. As can be seen, removing any training data will degrade the f-score, while the labeled visual entailment data seem more critical. A possible reason is that the visual entailment datasets fit the multi-modal entailment task well. We use the text encoder from ALBEF as a comparison, and the results show that the $f_{0.5}$ was about 4.3% higher using BERT. Overall, both the textual and visual entailment modules are helpful, making an essential contribution to our model in learning more about multi-modal interactions.

6.5.2 Entailment-enhanced training strategy

Table 4 shows the results of different retrieval methods with or without applying our variable learning rate strategy on two benchmarks, Flickr30K and MSCOCO, respectively. Although we focus on the improvement of many-to-many matching recognition, we find that our entailment-enhanced

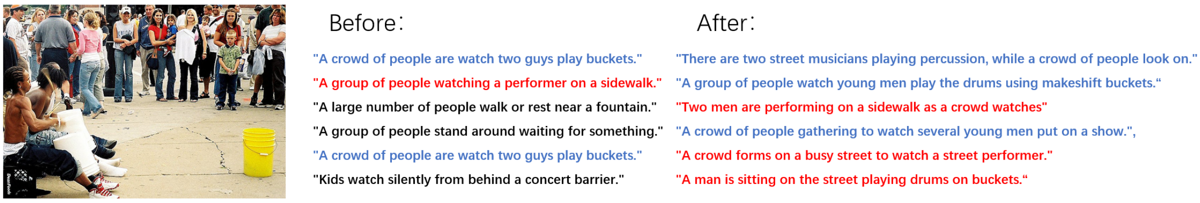

Figure 3: Comparison of examples of retrieval results before and after applying our entailment-enhanced learning strategy. Blue: original positives. Red: manually annotated entailment samples. Black: irrelevant samples.

training could also often improve the retrieval performance. Especially for CLIP's IR@1 score on Flickr30K raises more than 3% with our learning strategy. Therefore, we believe our entailment-enhanced training indeed helps the retrieval models find appropriate positive and negative image-text pairs.

In addition, we demonstrate the entailment performance of different retrieval models in Table 5. As can be seen, after applying our entailment-enhanced training strategy, all models' entailment performance obviously improves on both automatic and manual evaluations. Notably, CLIP$^{\#}$ significantly exceeds CLIP by more than 8% in terms of E@M. The results reveal the effectiveness of our strategy in refining the entailment degree for retrieval models universally.

6.5.3 Recall rate is unreliable

The recall rate is unreliable because it only focuses on the ranking of the golden captions and ignores others, which cannot fully reflect the retrieval quality. We have manually analyzed several cases where the recall rate declines after applying an entailment-enhanced strategy and found that most of their retrieval results have entailment relationship. On the contrary, the increase of the content rate may occupy the position of original golden captions, resulting in a decrease in the number of the top-K retrieval results. This is why the recall rate results in Table 4 fluctuate.

6.6 Case Study

Multi-Modal Entailment During annotating the entailment performance, we find that our multi-modal entailment model has achieved satisfactory performance in most cases. However, there is still room for improvement in a few cases. Error types include the following: (i) Identification of the number of objects is disturbed; (ii) Wrong recognition of gender; (iii) For scenes with multiple objects, the model may only focus on the main objects and put less attention on others. More details of error cases are shown in Appendix A.5. In the future, we could use data augmentation on the text side to reduce these mistakes, thus enhancing the robustness of the proposed model.

Entailment-Enhanced Retrieval As for the retrieval results, we find that applying entailment-enhanced training could usually make the retrieved captions more relevant and reasonable. As shown in Figure 3, before applying entailment-enhanced strategy, many inappropriate descriptions exist in the retrieval results, such as "near a fountain" and "concert barrier". Besides, vague words like "waiting for something" will also reduce the retrieval quality. After training with our strategy, the number of entailed captions has increased to 3, while original positives also increased by one. In addition, the retrieval results describe the image from multiple aspects. For instance, the caption "two men are performing on a sidewalk as a crowd watches" indicates the number of performers in the picture, while "a man is sitting on the street playing drums on buckets" concretely describes what is happening in the scene.

7 Conclusion

In this paper, we propose to apply multi-modal entailment to handle the frequent many-to-many matching problem in image-text retrieval datasets. Our solution recognizes the relationship and utilizes the relationship in training. Automatic and manual experiments reveal that the proposed method can consistently improve the matching performance of retrieval models. In the future, we plan to extend our multi-modal entailment model to the video-text retrieval task. Besides, we are devoted to handling the typical entailment errors mentioned in Section 6.6.

References

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015b. A large annotated corpus for learning natural language inference. *CoRR*, abs/1508.05326.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. Number: arXiv:1909.11740 arXiv:1909.11740 [cs].

Zihang Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2017. Quora question pairs.

Hyunjin Choi, Judong Kim, Seongho Joe, and Youngjune Gwon. 2021. Evaluation of BERT and ALBERT sentence embedding performance on downstream NLP tasks. *CoRR*, abs/2101.10642.

Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang, and Seong Joon Oh. 2022. ECCV Caption: Correcting False Negatives by Collecting Machine-and-Human-verified Image-Caption Associations for MS-COCO. Number: arXiv:2204.03359 arXiv:2204.03359 [cs].

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *MLCW*.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. 2021. Redcaps: web-curated image-text data created by the people, for the people. *CoRR*, abs/2111.11431.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Number: arXiv:1810.04805 arXiv:1810.04805 [cs].

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German Image Descriptions. Number: arXiv:1605.00459 arXiv:1605.00459 [cs].

Rosa Falotico and Piero Quatto. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. *CoRR*, abs/1805.02266.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 315–323, Fort Lauderdale, FL, USA. PMLR.

M. Gutmann and A Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *Journal of Machine Learning Research*, 9:297–304.

Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cesar Ilharco, Afsaneh Shirazi, Arjun Gopalan, Arsha Nagrani, Blaz Bratanic, Chris Bregler, Christina Funk, Felipe Ferreira, Gabriel Barcik, Gabriel Ilharco, Georg Osang, Jannis Bulian, Jared Frank, Lucas Smaira, Qin Cao, Ricardo Marino, Roma Patel, Thomas Leung, and Vaiva Imbrasaite. 2021. Recognizing multimodal entailment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 29–30, Online. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. Number: arXiv:2102.05918 arXiv:2102.05918 [cs].

A. Karpathy and L. Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Computer Vision & Pattern Recognition*.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria. Association for Computational Linguistics.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *AAAI*.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. 2019a. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training. Number: arXiv:1908.06066 arXiv:1908.06066 [cs].

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021a. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. Number: arXiv:2107.07651 arXiv:2107.07651 [cs].

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021b. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019b. Visual Semantic Reasoning for Image-Text Matching. Number: arXiv:1909.02701 arXiv:1909.02701 [cs].

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019c. VisualBERT: A Simple and Performant Baseline for Vision and Language. Number: arXiv:1908.03557 arXiv:1908.03557 [cs].

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. Number: arXiv:2004.06165 arXiv:2004.06165 [cs].

Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. 2022. Not All Patches are What You Need: Expediting Vision Transformers via Token Reorganizations. Number: arXiv:2202.07800 arXiv:2202.07800 [cs].

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. Number: arXiv:1405.0312 arXiv:1405.0312 [cs].

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. Number: arXiv:1711.05101 arXiv:1711.05101 [cs, math].

Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. 2020. Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO. *CoRR*, abs/2004.15020.

Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. 2021. Crisscrossed Captions: Extended Intramodal and Intermodal Semantic Similarity Judgments for MS-COCO. Number: arXiv:2004.15020 arXiv:2004.15020 [cs].

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. Number: arXiv:2103.00020 arXiv:2103.00020 [cs].

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147, Los Angeles. Association for Computational Linguistics.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. Number: arXiv:1908.08530 arXiv:1908.08530 [cs].

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. Number: arXiv:1908.07490 arXiv:1908.07490 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. Number: arXiv:1706.03762 arXiv:1706.03762 [cs].

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *CoRR*, abs/2108.10904.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *CoRR*, abs/1704.05426.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019a. Visual Entailment: A Novel Task for Fine-Grained Image Understanding. Number: arXiv:1901.06706 arXiv:1901.06706 [cs].

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019b. Visual entailment: A novel task for fine-grained image understanding. *CoRR*, abs/1901.06706.

Heng-Da Xu, Zhongli Li, Qingyu Zhou, Chao Li, Zizhen Wang, Yunbo Cao, Heyan Huang, and Xian-Ling Mao. 2021. Read, listen, and see: Leveraging multimodal information helps Chinese spell checking. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 716–728, Online. Association for Computational Linguistics.

Wenpeng Yin, Nazneen Fatema Rajani, Dragomir R. Radev, Richard Socher, and Caiming Xiong. 2020. Universal natural language processing with limited annotations: Try few-shot textual entailment as a start. *CoRR*, abs/2010.02584.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. 2018. Overview of the NLPCC 2018 Shared Task: Grammatical Error Correction. In Min Zhang, Vincent Ng, Dongyan Zhao, Sujian Li, and Hongying Zan, editors, *Natural Language Processing and Chinese Computing*, volume 11109, pages 439–445. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

A Example Appendix

A.1 Examples of Entailment-Corrected Dataset

Examples of our entailment-corrected dataset are shown in Figure 4. Every image corresponds to five golden captions and one hypothesis text.

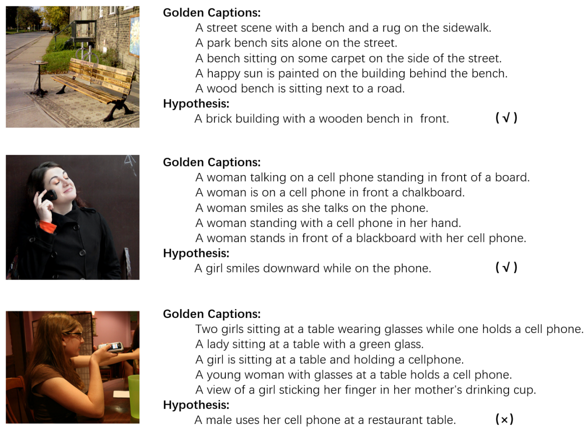

Figure 4: Examples in our entailment-corrected dataset. Symbol "$\surd$" represents the entailment relationship between premise and hypothesis, and symbol "$\times$" is the opposite.

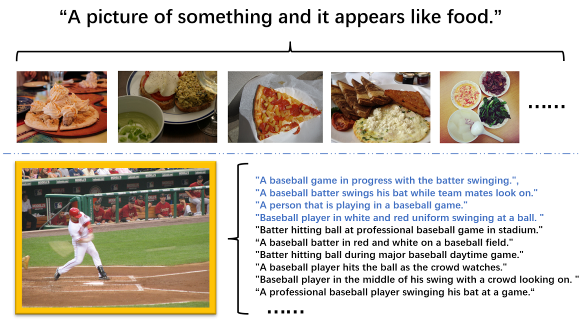

Figure 5: Typical examples about how many items that one image or caption can match. Blue: original positives.

A.2 Maximum Match

In addition, we use the same method to annotate some typical examples in the original MSCOCO testset. As shown in Figure 5, we found that one plain caption “A picture of something and it appears like food” can match accord with up to 178 images with food, and the image with a person who is playing a baseball game can be depicted by according up to 89 captions. These huge numbers demonstrate the universality of the many-to-many matching phenomenon. We also find contradictions even in the original golden image-text pairs. For example, different annotators describe a child in the same picture as a boy and a girl.

A.3 Datasets For Multi-modal Entailment

We constructed a training dataset for multi-modal entailment by integrating Visual entailment, Textual entailment, and Natural Language Understanding (NLU) datasets, the components of which are shown below:

SNLI-VE SNLI-VE is a visual entailment dataset that is constructed based on Flickr30K and SNLI.

CrissCrossed Caption (CxC) Parekh et al. (2020) annotate the dataset CrissCrossed Caption (CxC) based on MSCOCO to enhance the dataset of cross-modal correlations: image-image,image-text,text-text.

XNLI XNLI is a significant dataset in natural language understanding. It contains 15 languages, and each piece of data consists of two sentences named promise and hypothesis, respectively, intending to predict the relationship between a given two sentences: entailment, contradiction, or neutral.

Extended COCO Validation (ECCV) Similar to CxC, Extended COCO Validation (ECCV) (Chun et al., 2022) is a caption dataset containing 1,261 image queries (originally 5,000) but with 17.9 positive captions per image query on average (originally 5). It also contains 1,332 caption queries (originally 25,000) with 8.5 positive images per caption (originally 1).

MRPC Microsoft Research Paraphrase Corpus consists of sentence pairs automatically extracted from online news sources, with human annotations for whether the sentences in the pair are semantically equivalent (Dolan and Brockett, 2005). We transform the semantic similarity discriminant in sentence pairs into an entailment discriminant.

RTE Recognizing Textual Entailment is a binary entailment task similar to XNLI but with much less training data (Bentivogli et al., 2009).

STS-B The Semantic Textual Similarity Benchmark is a collection of sentence pairs drawn from news headlines and other sources (Cer et al., 2017). They were annotated with a score from 1 to 5 denoting how similar the two sentences are in terms of semantic meaning.

QQP Quora Question Pairs is a binary classification task that aims to determine if two questions asked on Quora are semantically equivalent (Chen et al., 2017).

Text Simplification(TS) The text simplification task is to transform a complex sentence into a clean and clear sentence, which makes it more convenient to read and communicate (Kauchak, 2013). To translate the data into the form of an entailment task, we consider the existence of entailment relations between pairs of sentences in the text simplification task.

Since the labels of STS-B and CXC datasets are scores ranging from 0 to 5, we use three as a threshold and thus transform them to be usable for our task.

A.4 Baseline Models For Image-Text Retrieval

ALBEF (Li et al., 2021a) model combines a ViT as a visual encoder and stacked 6-layer transformer blocks as text encoders. In the image-text retrieval task, ALBEF first aligns the unimodal image and text representation before fusing them with a multi-modal encoder.

CLIP (Radford et al., 2021) performs pre-training on massive noisy image-text data using a contrastive loss. CLIP officially provides a variety of image encoders. In our experiment, we choose the official ViT-B/32 as our image encoder for quickly training and evaluation.

UNITER (Chen et al., 2020) leverage a transformer-based architecture to learn universal representations from image and text features. We choose UNITER-base as our pre-train model.

A.5 Error Cases in Model

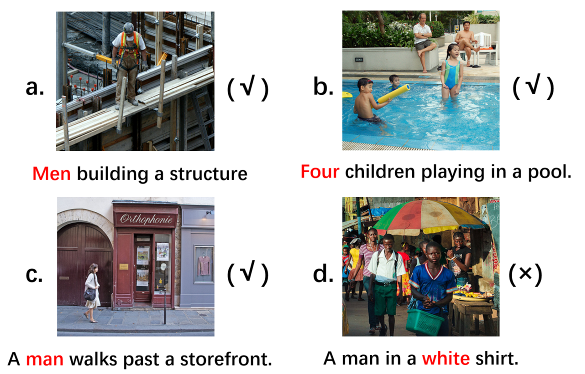

Figure 6: Typical error cases of our multi-modal entailment model inference. The entailment relationship inferred by the model is remarked as the symbol "$\surd$" and the symbol "$\times$" on the contrary.

During annotating the entailment performance, we find that our multi-modal entailment model has achieved satisfactory performance in most cases. However, there is still room for improvement in a few cases. Error cases shown in Figure 6 represent the following typical mistakes occurred occasionally: (i) Identification of the number of objects is disturbed. In regions (a) and (b), the model does not accurately measure the number of people, like 'Men' and 'Man'; (ii) Wrong recognition of gender. In region (c), the person depicted in the photo is a woman; (iii) For scenes with multiple objects, the model may only focus on the main objects and put less attention on others. In region (d), we try to replace "A man in a white shirt." with "A woman in a green shirt." and find the inference result to be entailment. However, in manual annotation, we usually also focus on secondary characters and scenes; In the future, we could use data augmentation on the text side to reduce these mistakes, thus enhancing the robustness of the proposed model.