Pre-training Cross-Modal Retrieval by Expansive Lexicon-Patch Alignment

Yang Yiyuan[1], **Guodong Long**[1], **Michael Blumenstein**[1], **Xiubo Geng**[2], **Chongyang Tao**[2], **Tao Shen**[2], **Daxin Jiang**[2*]
[1]University of Technology Sydney, Australia
[2]Microsoft Corporation
Yiyuan.Yang-1@student.uts.edu.au, {Guodong.Long, Michael.Blumenstein}@uts.edu.au
{xigeng,chongyang.tao,shentao,djiang}@microsoft.com

Abstract

Recent large-scale vision-language pre-training depends on image-text global alignment by contrastive learning and is further boosted by fine-grained alignment in a weakly contrastive manner for cross-modal retrieval. Nonetheless, besides semantic matching learned by contrastive learning, cross-modal retrieval also largely relies on object matching between modalities. This necessitates fine-grained categorical discriminative learning, which however suffers from scarce data in full-supervised scenarios and information asymmetry in weakly-supervised scenarios when applied to cross-modal retrieval. To address these issues, we propose expansive lexicon-patch alignment (ELA) to align image patches with a vocabulary rather than only the words explicitly in the text for annotation-free alignment and information augmentation, thus enabling more effective fine-grained categorical discriminative learning for cross-modal retrieval. Experimental results show that ELA could effectively learn representative fine-grained information and outperform state-of-the-art methods on cross-modal retrieval.

Keywords: Cross-modal retrieval, Multi-modal alignment, Lexicon-based representation, Open vocabulary

1. Introduction

With the surge of multimedia data online, the need has shifted from just text searches to multi-modal searches. Cross-modal retrieval, which uses one form of content (e.g., image or text) to find related content in another, aims to bridge this modality gap. This technique has applications ranging from recipe suggestions (Carvalho et al., 2018) to visual searches (Bain et al., 2021) and story creation (Fan et al., 2018).

Recently, inspired by the success of self-supervised learning in intra-modal tasks, large-scale pre-training has attracted surging attention in the vision-language (VL) community. The most prevalent pre-training methods, e.g., CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), train over hundreds of million pseudo image-text pairs via contrastive learning for global alignment. These methods are proposed to benefit general-purpose VL tasks by fine-tuning but appear to be sub-optimal in cross-modal retrieval due to a lack of fine-grained alignments between modalities (Yao et al., 2021; Yang et al., 2022). To mitigate this, many works, e.g., ViLT (Kim et al., 2021), FILIP (Yao et al., 2021) and ALBEF (Li et al., 2021), are proposed to complement the global alignment with fine-grained ones by weakly-supervised contrastive learning.

However, the above works mainly depend on contrastive learning, leading to insufficient categorical information and thus sub-optimal performance

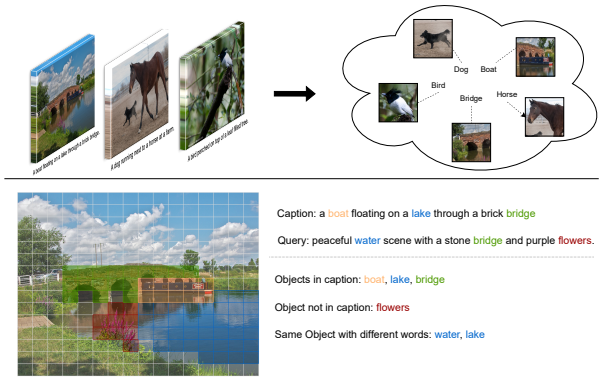


Figure 1: An illustration of cross-modal retrieval. Cross-modal retrieval depends on objects matching with referred words like dog and horse. As shown in the figure, the pre-train caption can only capture a portion of the information in the image (such as the boat, lake, and bridge), leaving out other objects (such as flowers) and synonyms (such as water) that may be included in a query.

in cross-modal retrieval. This is because compared to heavy semantic-matching dependency in pure text retrieval (Karpukhin et al., 2020), cross-modal retrieval relies primarily on matching multiple objects (e.g., horse, dog in Figure 1) between mentions in a text and concepts in an image. Therefore, fine-grained categorical discriminative learning (Dosovitskiy et al., 2020), which aims at learning to distinguish between different categories for

* Corresponding Authors.

LREC-COLING 2024, pages 12977–12987

improving image understanding, is highlighted in cross-modal retrieval tasks.

Despite promising, human-annotated categorical supervisions in a fine-grained fashion (Kazemzadeh et al., 2014) are too scarce to pre-train a VL model. Additionally, weakly fine-grained categorical supervision methods (Zareian et al., 2021), which rely on noun phrases in captions for patch-level discriminative learning, suffer from information asymmetry problems. As illustrated in Figure 1, the short text cannot exhaustively describe the image, making categorical learning less effective.

To address these issues, we propose a vision-language pre-training framework, dubbed Expansive Lexicon-patch Alignment (ELA), for cross-modal retrieval. Unlike previous work reliant on contrastive learning and fine-grained alignments between the paired image and texts, ELA introduces fine-grained categorical discriminative learning by expansive lexicon-patch alignment to align patches with all relevant words in a vocabulary, expanding beyond just paired text. Specifically, built upon the popular backbone ALBEF (Li et al., 2021), we first propose *a discriminative patch-to-lexicon head* in section 3.1 to map encoded patch embeddings to the vocabulary space of language, producing distributions of image patches on vocabulary. After getting these distributions of patches, to achieve a fine-grained cross-modal alignment using distributions from the text head, we present *an expansive lexicon-patch alignment module* in section 3.2 to align the distributions of patches and words in the whole vocabulary space. Lastly, to make the alignment more precise, we propose *a cross-modal lexicon-distillation module* in section 3.3 to distill cross-modal-aware lexicon distribution into the masked language modeling (MLM) head.

As such, the fine-grained alignment in our framework is not restricted to leveraging words/phrases that explicitly appear in the paired text, but also the expansive lexicons (full of synonyms and coordinate terms) in light of contextualization of the text. This is achieved by our carefully designed fine-grained cross-modal alignment in the entire language vocabulary space, which could greatly alleviate the information asymmetry problem above. Extensive experiments demonstrate that ELA achieves state-of-the-art performance on cross-modal retrieval by learning fine-grained information with expansive lexicon-patch alignment.

2. Related Work

Contrastive Vision-language Pre-training. Existing vision-language pre-training (VLP) relies on large-scale pre-train data and learns information by self-supervised contrastive learning. According to the granularity of learning, they can be categorized as global contrastive models or fine-grained contrastive models. Unlike global contrastive models such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), fine-grained contrastive VLP models either rely on the detector to extract objects for image objects and text words matching like UNITER (Chen et al., 2020c), or use the paired caption text as weak supervision to align patches and words for annotation-free, e.g. ViLT (Kim et al., 2021) and FILIP (Yao et al., 2021). Additionally, deep fusion models like ALBEF (Li et al., 2021), TCL (Yang et al., 2022) and X-VLM (Zeng et al., 2021) introduce a cross-encoder for better interactions between different modalities. Despite introducing an object detection task as fine-grained discriminative learning, X-VLM (Zeng et al., 2021) still relies on partially annotated data in pre-training. Given their lack of categorical information, these contrastive VLP models might struggle in cross-modal retrieval, which motivates this work.

Categorical Discriminative Learning. Categorical discriminative learning enhances image understanding by aiming at learning generalized image representation for recognition. One simple task is image classification, labeling images with predefined categories. Convolutional neural network (CNN) (Szegedy et al., 2015; He et al., 2016) is the first proposed network for image classification by learning multi-scale features of images with deep layers. With the NLP success of the Transformer (Vaswani et al., 2017), models like ViT (Dosovitskiy et al., 2020) and Swin-transformer (Liu et al., 2021b) adopted its architecture for images. However, these approaches above mainly focus on global category learning, which may fail in multiple object detection in one image. Therefore, fine-grained categorical learning such as visual grounding is needed. Since cross-modal retrieval heavily relies on the category matching between image and text, some research works (Peng and Qi, 2019; Zhen et al., 2019) use categorical discriminative learning for supervised cross-modal retrieval. However, they only consider global categorical learning and require human-annotated data. To mitigate this, we add fine-grained categorical discriminative learning in pre-training and take paired text as weak supervision.

Visual Grounding. Visual grounding involves locating an object in an image based on a provided noun phrase. Traditional classical methods (Liu et al., 2020; Yang et al., 2020) require human annotations which can be costly and error-prone. To address this, weakly supervised visual grounding methods (Datta et al., 2019; Akbari et al., 2019; Liu et al., 2021a) utilize noun phrases from paired

captions as labels. Inspired by this, Zareian et al. (2021) expanded object detection to the open vocabulary through a weak-supervised visual grounding pre-training model to learn the matched objects and words. However, directly applying this objective suffers from data scarcity in full-supervised scenarios and information asymmetry in weakly-supervised scenarios. Thus, we propose ELA to align the image patch and text word in the lexicon space and expand the candidate words for abundant image information learning.

3. Pre-training Cross-Modal Retrieval

Vision-Language Pre-training Task. Given a pseudo image-text pair, i.e., $< I, T >$, the visual-language pre-training task focuses on capturing semantics underlying both the image and text by a cross-modal neural architecture. It endeavors to not only identify and reinforce the semantic congruence within the given image-text pair but also alienate pairs that are semantically irrelevant.

Built on ALBEF (Li et al., 2021) backbone, which consists of an image encoder, a text encoder, and a cross-modal encoder, our pre-training method proposes to leverage a lexicon-centric alignment between modalities, as shown in Figure 2.

3.1. Discriminative Patch-to-Lexicon Head

Previous works in learning cross-modal alignment depend heavily on contrastive learning due to its powerful generalization ability in cross-modal representation learning (Radford et al., 2021; Jia et al., 2021). In particular, fine-grained contrastive learning between image patches and text words has been proven effective in a weakly supervised manner for cross-modal retrieval (Yao et al., 2021). Nonetheless, a recent work (Zeng and Mao, 2022) found that contrastive learning can be incompetent to learn the alignment for cross-modal retrieval tasks because the key to this task lies in object matching across modalities. Therefore, it is imperative to integrate categorical discriminative learning into the pre-training framework for cross-modal object matching.

Despite the inherent challenges associated with pre-defining real-world object categories for all visual patches, a feasible solution is resorting to the textual vocabulary at the text side, which provides an exhaustive categorical definition by a close set of lexicons or subwords. Such a solution is aligned with open-vocab visual grounding (Zareian et al., 2021). But in contrast to the conventional visual grounding that detects bounding boxes of objects, our method applies categorical learning to each individual patch, enabling a more general alignment in representation learning.

Formally, given an image, we first split it into a sequence of flattened patches, i.e., $I = [a_1, \ldots, a_N]$, to satisfy the input format of the Transformer encoder. Then, the flattened patches are fed into a Transformer encoder for patch-level contextualized representation. Instead of taking the original patches of the image as input, we apply a masking perturbation (30% of original tokens are masked), where the indices of masked patches are denoted as $\mathbb{M}^{(\text{img})}$. This approach shares a similar inspiration with mask-regularized pre-training (Gao and Callan, 2021), which inherently bolsters the robustness of the model, enhancing its resilience to perturbations. We denotes the masked image as $\bar{I} = [\bar{a}_1, \ldots, \bar{a}_N]$ and the image encoding is written as

$$\boldsymbol{V} = f^{\text{img}}(\bar{I}) := \text{Transfm}([\bar{a}_1, \ldots, \bar{a}_N]; \theta^{(\text{img})}), \quad (1)$$

where N is the number of patches, $\theta^{(\text{img})}$ parameterizes this image encoder, and $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N] \in \mathbb{R}^{d \times N}$ denotes the resulting patch embeddings. Next, to facilitate lexicon-centric alignment with the language counterpart, we introduce a discriminative patch-to-lexicon (P2L) head. This P2L head is designed to map each patch-level embedding $\boldsymbol{v}_i$ into a probability distribution over language vocabulary $\mathbb{V}$, i.e.,

$$\boldsymbol{P}^{(\text{p2l})} = \sigma(\boldsymbol{W}^{(\text{we})}(\boldsymbol{W}_{\theta^{(\text{p2l})}} \boldsymbol{V} + b_{\theta^{(\text{p2l})}})), \quad (2)$$

where $\sigma(\cdot)$ is a softmax along with $|\mathbb{V}|$, $\boldsymbol{W}^{(\text{we})} \in \mathbb{R}^{|\mathbb{V}| \times d}$ denotes word embedding weight matrix, $\boldsymbol{W}_{\theta^{(\text{p2l})}}$ and $b_{\theta^{(\text{p2l})}}$ denotes a $\theta^{(\text{p2l})}$-parameterized linear layer, and the $\boldsymbol{P}^{(\text{p2l})} = [\boldsymbol{p}_1^{(\text{p2l})}, \ldots, \boldsymbol{p}_N^{(\text{p2l})}] \in \mathbb{R}^{|\mathbb{V}| \times N}$. With the integration of our proposed patch-to-lexicon module, the alignment of an image with a corresponding text through $\boldsymbol{P}^{(\text{p2l})}$ becomes intrinsically feasible, and this method will be fully explained in the next section.

3.2. Expansive Lexicon-Patch Alignment

Given the distribution $\boldsymbol{p}_i^{(\text{p2l})} \in \boldsymbol{P}^{(\text{img})}$ that maps each visual patch into lexicons, a straightforward learning strategy might be to assign a pseudo label to each patch and subsequently apply discriminative learning. Conventionally, these pseudo labels are derived from explicit words $[w_1, \ldots, w_M]$, in the paired text T, where M denotes the number of words in T.

However, the short length of text T often fails to exhaustively describe the rich contents of the image I while typically concentrating on a constricted viewpoint. Such a constraint could cause an information asymmetry problem, wherein a pre-trained model may be prone to learn image representations biased to the text distribution of a specific corpus.

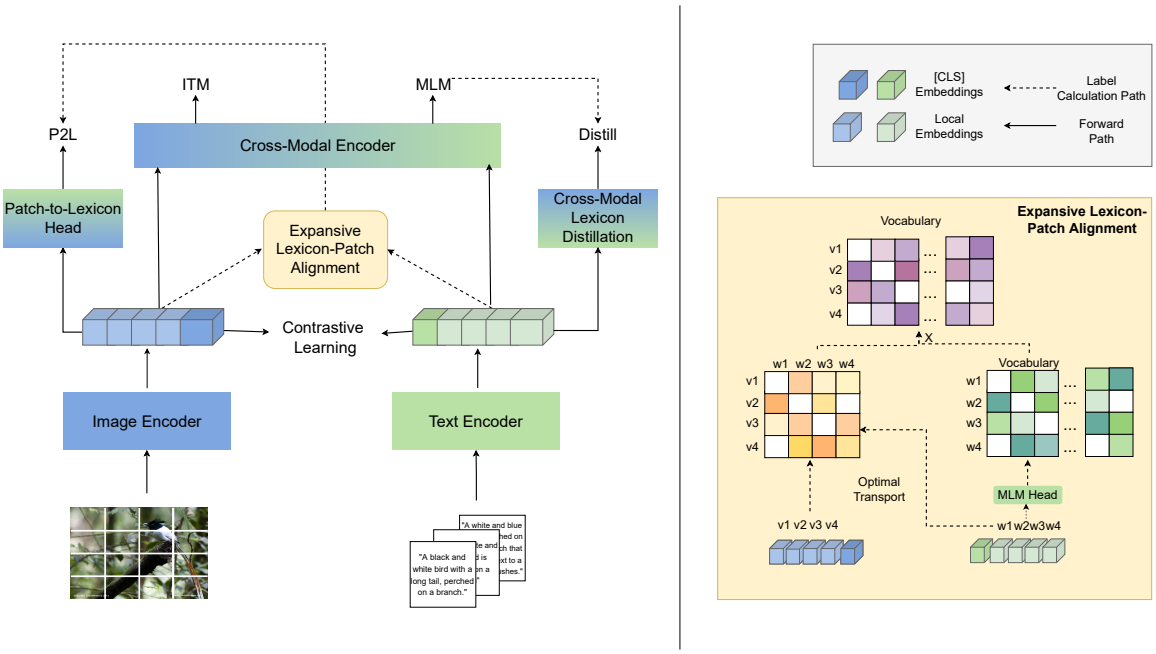


Figure 2: The overall architecture of ELA. It consists of an image encoder, a text encoder, a cross-modal encoder, a patch-to-lexicon head and a cross-modal lexicon distillation module. As shown on the right side, we use optimal transport to get the similarity scores of each image patch and text word, and use MLM head to get the lexicon distribution of each word as an expansion. By multiplying them together, the final lexicon distributions of each patch are obtained as labels of expansive fine-grained discriminative learning.

Thereby, inspired by recent advances in lexicon expansion (Formal et al., 2021; Shen et al., 2022; Nogueira et al., 2019) for text retrieval, we propose an expansive lexicon-patch alignment between the patch-to-lexicon (P2L) head at the image side and the masked language modeling (MLM) head at the text side. Instead of a hard pseudo label for each patch, we assign each patch with a soft distribution over $\mathbb{V}$ by deeply exploiting rich lexicon correlations pre-trained by MLM.

First, similar to the image side, a text encoder is applied to the input text T, for contextualized embeddings. That is

$$\boldsymbol{U}=f^{\text{txt}}(T):=\text{Transfm}([w_1,\ldots,w_M];\theta^{(\text{txt})}). \quad (3)$$

Then, an MLM head is applied to $\boldsymbol{U}$ by

$$\boldsymbol{P}^{(\text{w2l})}=\sigma(\boldsymbol{W}^{(\text{we})}\text{Transfm-Layer}(\boldsymbol{U};\theta^{(\text{w2l})})), \quad (4)$$

where $\boldsymbol{P}^{(\text{w2l})}=[\boldsymbol{p}_1^{(\text{w2l})},\ldots,\boldsymbol{p}_M^{(\text{w2l})}]\in\mathbb{R}^{|\mathbb{V}|\times M}$ and each $\boldsymbol{p}_j^{(\text{w2l})}$ denotes contextualization-based lexicon expansions for w_j, full of synonyms and coordinate terms from MLM pre-training.

Next, to associate $\boldsymbol{P}^{(\text{w2l})}$ to $\boldsymbol{P}^{(\text{p2l})}$ for cross-modal alignment learning, we follow the prevalent weakly label assigning method, optimal transport (OT) by Xie et al. (2020), which has been proven effective in many works (Chen et al., 2020b,a; Kim et al., 2021). Instead of directly using $\boldsymbol{P}^{(\text{w2l})}$ and $\boldsymbol{P}^{(\text{p2l})}$ for OT, we propose to use the hidden states, $\boldsymbol{V}$ and $\boldsymbol{U}$, to calculate each distance from a word w_j to a patch a_i, i.e.,

$$c_{i,j}=\text{OptimalTrans}(\boldsymbol{V},\boldsymbol{U}). \quad (5)$$

Here, we use the inexact proximal point method optimal transport algorithm (IPOT) (Xie et al., 2020) to calculate the cost from w_j to a_i. Thereby, we can easily derive a soft label for each patch by

$$\hat{\boldsymbol{p}}_i^{(\text{p2l})}=\frac{\sum_j(1-c_{i,j})\cdot\boldsymbol{p}_j^{(\text{w2l})}}{\sum_j(1-c_{i,j})}. \quad (6)$$

Lastly, The loss function for our expansive lexicon-patch alignment is defined as

$$\mathcal{L}_{ela}=\sum\nolimits_{i\in\mathbb{M}^{(\text{img})}}\text{KL-Div}(\boldsymbol{p}_i^{(\text{p2l})},\hat{\boldsymbol{p}}_i^{(\text{p2l})}). \quad (7)$$

Note that we cut off the gradient backward to $\boldsymbol{p}_i^{(\text{p2l})}$ as its back-end is updated with main modules.

Remark. Compared to hidden states-based OT loss to directly optimize cross-modal alignment, ELA depends on the OT over hidden states to calculate cross-modal weights merely and then integrates the weights for our expansive lexicon-patch alignment. With masking perturbation, our ELA, similar to momentum-based pre-training (He et al., 2020), can be seen as a robust boosting mechanism via transferring lexicon-augmented knowledge in MLM into the image encoder.

3.3. Cross-Modal Lexicon-Distillation

As detailed in expansive lexicon-patch alignment, a precise distribution $\boldsymbol{p}_j^{(\text{w2l})}$ is critical to $\hat{\boldsymbol{p}}_i^{(\text{p2l})}$, therefore serving as an essential foundation for the efficacy of our model. In order to further improve the accuracy of $\boldsymbol{p}_j^{(\text{w2l})}$, we employ a heterogeneous knowledge distillation, where the teacher is a powerful cross-modal encoder to provide image-aware language modeling results. Such image-aware information can subsequently equip the MLM head on top of the text encoder to expand its vocabulary with more visual object-related lexicons.

Formally, we present a cross-modal encoder defined over a combination of the patch embeddings $\boldsymbol{V}$ and word embeddings $\boldsymbol{U}$, i.e.,

$$\begin{aligned}[\tilde{\boldsymbol{V}}, \tilde{\boldsymbol{U}}] &= f^{\text{cross}}(I, T) \\ &:= \text{Transfm-Layer}([\boldsymbol{V}, \boldsymbol{U}]; \theta^{(\text{xm})}), \end{aligned} \tag{8}$$

where $\tilde{\boldsymbol{U}} = [\tilde{\boldsymbol{u}}_1, \ldots, \tilde{\boldsymbol{u}}_M]$ are image-aware word representations, corresponding to $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M]$. Then, we apply word embedding matrix to $\tilde{\boldsymbol{U}}$ for image-aware distribution, i.e.,

$$\boldsymbol{P}^{(\text{mlm})} = \sigma(\boldsymbol{W}^{(\text{we})} \tilde{\boldsymbol{U}}). \tag{9}$$

Here, $\boldsymbol{P}^{(\text{mlm})} = [\boldsymbol{p}_1^{(\text{mlm})}, \ldots, \boldsymbol{p}_M^{(\text{mlm})}]$ denotes the image-aware prediction distributions for masked language modeling. Lastly, a position-wise distillation loss on masked words in T is defined as

$$\mathcal{L}_{distill} = \sum\nolimits_{j \in \mathbb{M}^{(\text{txt})}} \text{KL-Div}(\boldsymbol{p}_j^{(\text{w2l})}, \boldsymbol{p}_j^{(\text{mlm})}). \tag{10}$$

Since the cross-modal encoder enables fine-grained interaction between modalities, masked language modeling can perform more precise expansive predictions based on the visibility of image semantics. Distilling such precise expansive lexicon information makes the text encoder aware of its counterpart image – thus improving the feasibility of our expansive lexicon-patch alignment objective.

3.4. Pre-training Objectives

In addition to the above two learning objectives, the other three widely-used pre-training objectives in VLP are adopted in our model: Cross-Modal Alignment (CMA) $\mathcal{L}_{cma}$, Image-Text Matching (ITM) $\mathcal{L}_{itm}$, Mask Language Modeling (MLM) $\mathcal{L}_{mlm}$. CMA aligns paired images and texts based on global information by contrastive learning and is regarded as preliminary overall alignment before further fusion. ITM is used to further fuse image and text representations by determining whether the input image-text pair is matched or not, where one negative pair is sampled in batch for each input image and text in training. MLM is another way to enhance the interactions between images and text, which aims at predicting the masked tokens for better representation learning in accordance with the context of text and image. Therefore, the overall loss for our pre-training is:

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{ela} + \mathcal{L}_{distill} \\ &+ \mathcal{L}_{cma} + \mathcal{L}_{itm} + \mathcal{L}_{mlm} \end{aligned} \tag{11}$$

Model	MSCOCO(5K)					
	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
UNIT	64.1	87.7	93.3	48.8	76.7	85.8
ViLT	56.5	82.6	89.6	40.4	70.0	81.1
CLIP	58.4	81.5	88.1	37.8	62.4	72.2
ALBEF(4M)	68.7	89.5	94.7	50.1	76.4	84.5
TCL	71.4	90.8	95.4	53.5	79.0	87.1
ELA	**72.2**	**91.3**	**95.4**	**54.6**	**80.3**	**88.1**

Table 1: Zero-shot performance on COCO compared with previous work, and R@1, R@5 and R@10 of text-retrieval and image-retrieval results are reported.

4. Experiment

Data and Metrics. For pre-training, we use COCO (Lin et al., 2014), Visual Genome (VG) (Krishna et al., 2017), Conceptual Captions (CC) (Sharma et al., 2018), and SBU Captions (Ordonez et al., 2011), which consist of 4.0M images and 5.1M image-text pairs in total. Following most image-text retrieval methods, we evaluate the pre-trained model on COCO (Lin et al., 2014) and Flickr30K (Plummer et al., 2015) under the settings of fine-tuning and zero-shot, and the recalled results are used for calculating the metrics. For the setting of fine-tuning, we just keep the image encoder, text encoder and fusion encoder for simplicity. At the same time, only CMA and ITM are left to calculate the retrieval results. CMA is first used to select a wide range of pre-ranking retrieval results and then ITM is to rerank the selected results as the final ranking results. For the setting of zero-shot, test results are directly obtained by evaluating the pre-trained model on test data.

Implementation Details. We implement[1] our model based on the framework of ALBEF (Li et al., 2021), where the image encoder is ViT-B/16 with 12 layers, the text encoder is a 6-layer transformer and the fusion encoder is also a 6-layer transformer. All of them are initialized by TCL (Yang et al., 2022) for efficient training. The text head for distill learning is a two-layer transformer with random initialization. Both the text distillation classification head and the P2W classification head share the same word embedding in the text encoder. We set the

[1] https://github.com/Lydia-yang/ELA

Model	MSCOCO(5K)					
	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
UNIT	65.7	88.6	93.8	52.9	79.9	88.0
ViLT	61.5	86.3	92.7	42.7	72.9	83.1
ALBEF(4M)	73.1	91.4	96.0	56.8	81.5	89.2
TCL	75.6	92.8	96.7	59.0	**83.2**	**89.9**
ELA	**76.4**	**92.8**	**96.8**	**59.0**	83.1	89.8

Table 2: Fine-tune performance on COCO compared with previous work, and R@1, R@5 and R@10 of text-retrieval and image-retrieval results are reported.

Model	Flickr30K(1K)					
	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
UNIT	80.7	95.7	98.0	66.2	88.4	92.9
ViLT	73.2	93.6	96.5	55.0	82.5	89.8
CLIP	88.0	98.7	99.4	68.7	90.6	95.2
ALBEF(4M)	90.5	98.8	99.7	76.8	93.7	96.7
TCL	**93.0**	99.1	99.6	**79.6**	95.1	97.4
ELA	92.1	**99.3**	**99.8**	79.4	**95.2**	**97.4**

Table 3: Zero-shot performance on Flickr30K compared with previous work, and R@1, R@5 and R@10 of text-retrieval and image-retrieval results are reported.

Model	Flickr30K(1K)					
	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
UNIT	87.3	98.0	99.2	75.6	94.1	96.8
ViLT	83.5	96.7	98.6	64.4	88.7	93.8
ALBEF(4M)	94.3	99.4	99.8	82.8	96.7	98.4
TCL	94.9	99.5	99.8	**84.0**	96.7	**98.5**
ELA	**95.4**	**99.6**	**99.9**	83.9	**97.0**	98.1

Table 4: Fine-tune performance on Flickr30K compared with previous work, and R@1, R@5 and R@10 of text-retrieval and image-retrieval results are reported.

Model	VQA		SNLI-VE	
	test-dev	test-std	val	test
UNIT	72.70	72.91	78.59	78.28
ViLT	71.26	-	-	-
ALBEF(4M)	74.54	74.70	80.14	80.30
TCL	74.90	74.92	**80.51**	80.29
ELA	**74.92**	**74.97**	79.96	**80.41**

Table 5: Performance on other vision-language tasks.

image mask possibility to 0.30. During the pre-training, we take a batch size of 32 and train on 8 NVIDIA V100 GPUs for 5 epochs. The learning rate is initialized as 1e-5 which drops to 1e-6 gradually with AdamW optimizer (Loshchilov and Hutter, 2017). During pre-training, the input image is randomly cropped to 256 × 256 resolution, and applied RandAugment4 (Cubuk et al., 2020) for data augmentation. During downstream fine-tuning, we increase the image resolution to 384 × 384 and interpolate the positional encodings of image patches following work (Yang et al., 2022).

4.1. Evaluation Results

Since our proposed model mainly focuses on cross-modal retrieval tasks, we test ELA on the widely-used cross-modal retrieval datasets MSCOCO and Flickr30K, and compare with previous contrastive VLP models using pre-training data of the same magnitude under both zero-shot setting and full fine-tuning setting. Detailed evaluation results are discussed as follows.

Zero-shot Retrieval Results on MSCOCO. As illustrated in Table 1, compared with previous contrastive VLP models, our model achieves the best performance, which proves that adding categorical discriminative learning benefits cross-modal retrieval tasks. Compared with the global contrastive model CLIP, our model improves 13.8% of R@1 on text retrieval and 16.8% of R@1 on image retrieval respectively, revealing the necessity of fine-grained information learning. Compared with fine-grained contrastive models UNITER, our model could improve the performance by a large margin by using captions as weak supervision. For Other fine-grained contrastive models like ALBEF and TCL, our model achieves the best performance, which suggests the importance of adding categorical discriminative learning for learning matched objects and words. In general, the improvement suggests that our method is more efficient and could learn more generalized and transferable representations.

Fine-tuned Retrieval Results on MSCOCO. To evaluate our model in downstream tasks, we also conduct fine-tuned experiments. As shown in Table 2. Our model could perform better than most contrastive VLP models, including the global contrastive model CLIP and fine-grained contrastive models UNIT, VILT and ALBEF. For the most related work TCL, our model could improve the performance on text retrieval with comparable results on image retrieval, that is because TCL uses all losses of pre-training as fine-tuning losses and intra-modal contrastive loss would help to improve the performance by augmenting data with more negative examples while our model just uses ITM and CMA with small batch size. Overall, our designed fine-grained categorical discriminative learning could learn more suitable representations for cross-modal retrieval.

Zero-shot Retrieval Results on Flickr30K. We also evaluate our model on a smaller data Flickr30K which only contains around 30K images in total, and the results are displayed in Table 3. From Table 3, we can see that our model still per-

Module	MSCOCO(5K)					
	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Baseline	71.4	90.8	95.4	53.5	79.0	87.1
+P2L(w/o exp)	71.9	91.2	95.5	54.5	79.8	87.6
+P2L	72.3	91.1	95.6	54.2	79.8	87.6
+P2L+distll	72.2	91.3	95.4	54.6	80.3	88.1

Table 6: Ablation study of ELA on MSCOO in the zero-shot setting. We use results in TCL as our baseline here.

Model	MSCOCO(5K)					
	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
X-VLM	70.8	92.1	**96.5**	55.6	**82.7**	**90.0**
ELA	72.2	91.3	95.4	54.6	80.3	88.1
ELA-L	**74.9**	**92.7**	96.4	**56.6**	81.2	88.5

Table 7: Ablation study of the model size. Zero-shot performances are reported. ELA-L represents a larger model with the same encoder structures as X-VLM.

forms best on R@5 and R@10 compared with the previous contrastive VLP models, which further proves that our proposed model could learn more general representations by adding fine-grained discriminative learning.

Fine-tuned Retrieval Results on Flickr30K. Since Flickr30K is simpler than MSCOCO, the whole level of retrieval results is higher than MSCOCO and leaves less space to improve by fine-tuning. As shown in Table 4, our model still performs best in text retrieval and maintains comparable results in image retrieval, which demonstrates the importance of learning in an expansive alignment for improvement.

Evaluation Results on Other Vision-Language Tasks. To further evaluate our model, we also test ELA on two other vision-language tasks, VQA (Goyal et al., 2017) and SNLI-VE (Xie et al., 2019). As shown in Table 5, our model ELA achieves the best results across three out of four criteria, indicating that adding expansive lexicon-patch alignment could not only help with object matching in cross-modal retrieval tasks but also benefit the multimodal feature fusion in other general vision-language tasks.

4.2. Ablation Study

To learn the effectiveness of each component in our model, we conduct ablation studies on MSCOCO in the zero-shot setting. Since we use TCL to initialize our model, TCL is regarded as our baseline. For patch-to-lexicon loss, we consider two kinds of settings, one is with word expansion referred to P2L and the other is without word expansion referred to P2L (w/o exp). As shown in Table 6, all components are conducive to improving the performance in cross-modal retrieval. Among them, adding patch-to-lexicon loss brings the most benefit, which demonstrates the effectiveness of fine-grained categorical discriminative learning in cross-modal retrieval. Especially, the expansive patch-to-lexicon loss could further improve the performance of text retrieval by exploiting related words in the vocabulary. By blending more accurate information for expansion, the cross-modal lexicon distillation can also contribute to the improvement of the model's performance by increasing the image retrieval results. Overall, the combination of all these proposed modules could achieve the best performance.

4.3. Insight to Model Structure

As the performance of deep network relies on the model size and data size, we also implement our method based on a stronger baseline model X-VLM (Zeng et al., 2021), where the image encoder is Swin-transformer (Liu et al., 2021b) with larger parameters. As shown in Table 7, with a stronger image encoder, the performance of ELA-L could be further improved compared with ELA with a small size. In addition, ELA-L has better performance than X-VLM, especially increasing 4.1% on R@1 of text retrieval. These experiment results prove that our proposed modules are beneficial for cross-modal retrieval and a larger and stronger baseline can further improve the performance.

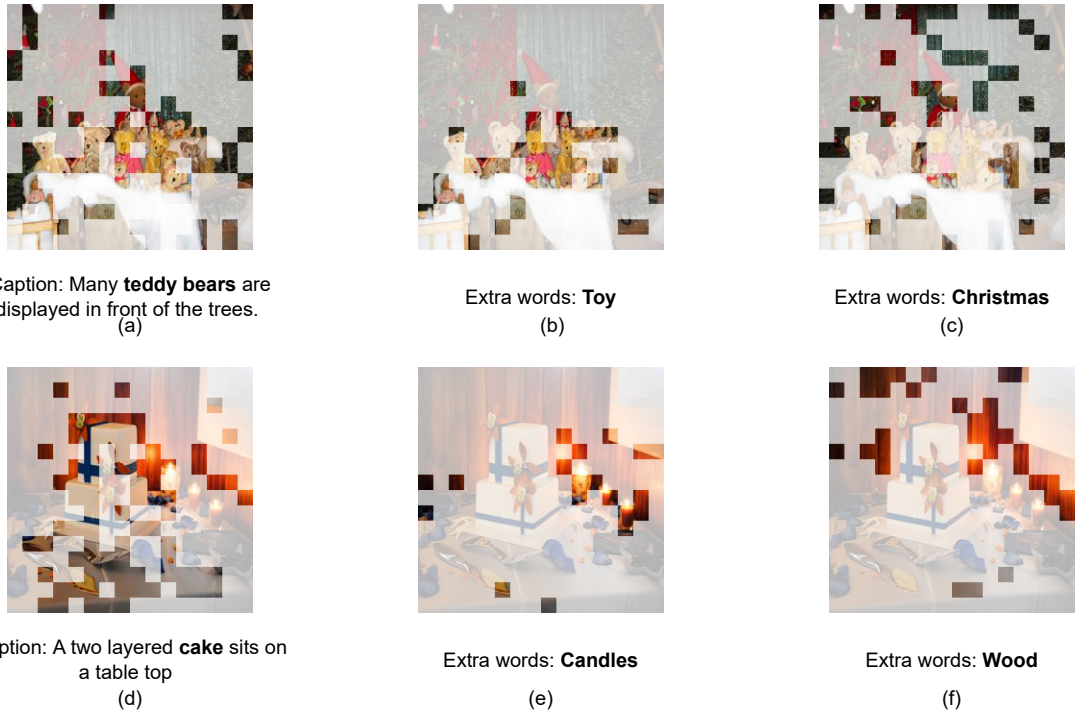
Since our model is based on the framework of ALBEF, which contains a bi-encoder (image encoder and text encoder) to retrieve the pre-ranking results and a cross-modal encoder to rerank the final results, it is necessary to investigate both the pre-ranking retrieval results and rerank results for more comprehensive study. For input data, only the image encoder and the text encoder are used to get image embeddings and text embeddings respectively for pre-ranking retrieval calculation based on CMA, and results are shown in Table 8. We compare our model with TCL sharing the same framework of ALBEF with better performance on MSCOCO under the setting of zero-shot and fine-tuning. In Table 8, we can see that our model could perform better than TCL both in the zero-shot setting and fine-tune setting. These improvements illustrate that our proposed modules could help to enhance the image encoder and text encoder by learning fine-grained discriminative information, and the cross-encoder could be further improved based on these enhanced embeddings.

4.4. Visualization and Discussion

To gain an explicit understanding of our model, we visualize the ability of our model to match re-

Model	Zero-shot on MSCOCO						Fine-tune on MSCOCO					
	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
TCL	57.0	84.1	91.3	42.2	70.8	81.0	65.4	88.5	94.1	49.0	77.0	85.4
ELA	**59.0**	**84.9**	**92.0**	**43.7**	**72.9**	**82.6**	**66.4**	**89.4**	**94.3**	**50.9**	**78.4**	**86.9**

Table 8: Pre-ranking retrieval results of COCO in zero-shot and fine-tune settings. Both TCL and ELA use the image encoder and the text encoder to calculate the Pre-ranking retrieval results for comparison.



Caption: Many **teddy bears** are displayed in front of the trees.
(a)

Extra words: **Toy**
(b)

Extra words: **Christmas**
(c)

Caption: A two layered **cake** sits on a table top
(d)

Extra words: **Candles**
(e)

Extra words: **Wood**
(f)

Figure 3: Examples of fine-grained alignment of ELA. The paired image and caption are first calculated based on the OT to get the related patches with words in the caption, then the image is fed into the patch-to-lexicon head for related word predictions. The attended image patches are highlighted by opaque patches while others are masked. For each example image, we present one result of predictions of objects in the caption and the other two results of predictions of words' absence in the caption.

lated words and objects. To be more specific, for a paired image and text, we first encode them respectively by the image encoder and the text encoder to get the image embeddings of each patch and text embeddings of each text. By calculating OT, the similarity scores of each patch and word are obtained, and the top words are selected as categorical classification results in the caption of each patch. To get the categorical classification result in vocabulary, we pass the image embeddings to the patch-to-lexicon head to get the lexicon distribution of each image patch and select the top words as the expansive categorical classification results.

As shown in Figure 3, the well-matched patch and word in the caption suggest that our model could learn fine-grained information via OT. For example, in image (a) of Figure 3, the related image patches of objects "teddy bears" in the caption are mostly covered, same with the object "cake" in image (b). In addition, as shown in Figure 3, words that are absent in the caption but related to the patches can be predicted based on the content of the image patch, which proves the effectiveness of expansive lexicon-patch alignment. As shown in image (b), the synonym "toy" of " teddy bears" is predicted by our model and image patches are attended to correctly. More importantly, our model could predict related words based on the whole image like image (c). According to the decorates and colors (green and red) with the Christmas atmosphere in the image, the word "Christmas" is predicted. By learning expansive lexicon-patch alignment, our model is equipped with the ability to predict objects not mentioned in the caption. In the image (e) and image (f), the object "wood" and the object "candle" are predicted in the image.

5. Conclusion

In conclusion, we propose a vision-language pre-training framework, expansive lexicon-patch alignment, for cross-modal retrieval. Compared with previous contrastive VLP models lacking categor-

ical discriminative information, our model learns fine-grained patch-to-word alignment by transferring image patches in the lexicon space for lexicon-centric alignment. Experimental results show that ELA could outperform existing methods by evaluating the cross-modal retrieval benchmark.

Limitations

The proposed expansive lexicon-patch alignment to pre-train cross-modal retriever is limited by i) *applicable scenarios*: the proposed ELA is focused only on cross-modal retrieval tasks by alleviating the information asymmetry problem in pseudo image-text pairs. ii) *data and model scales*: due to limited computation resources, the proposed framework is only evaluated on limited pre-training data over a base-scaled model.

6. Bibliographical References

Hassan Akbari, Svebor Karaman, Surabhi Bhargava, Brian Chen, Carl Vondrick, and Shih-Fu Chang. 2019. Multi-level multimodal common semantic space for image-phrase grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12476–12486.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. 2018. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 35–44.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Liqun Chen, Ke Bai, Chenyang Tao, Yizhe Zhang, Guoyin Wang, Wenlin Wang, Ricardo Henao, and Lawrence Carin. 2020a. Sequence generation with optimal-transport-enhanced reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7512–7520.

Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. 2020b. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, pages 1542–1553. PMLR.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020c. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703.

Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. 2019. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2601–2610.

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

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

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*.

Luyu Gao and Jamie Callan. 2021. Unsupervised corpus aware language model pre-training for dense passage retrieval. *arXiv preprint arXiv:2108.05540*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Yongfei Liu, Bo Wan, Lin Ma, and Xuming He. 2021a. Relation-aware instance refinement for weakly supervised visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5612–5621.

Yongfei Liu, Bo Wan, Xiaodan Zhu, and Xuming He. 2020. Learning cross-modal context graph for visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11645–11652.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

Yuxin Peng and Jinwei Qi. 2019. Cm-gans: Cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(1):1–24.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

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 ACL*.

Tao Shen, Xiubo Geng, Chongyang Tao, Can Xu, Kai Zhang, and Daxin Jiang. 2022. Unifier: A unified retriever for large-scale retrieval. *arXiv preprint arXiv:2205.11194*.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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

Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. 2020. A fast proximal point method for computing exact wasserstein distance. In *Uncertainty in artificial intelligence*, pages 433–453. PMLR.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 2022. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680.

Sibei Yang, Guanbin Li, and Yizhou Yu. 2020. Graph-structured referring expression reasoning in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9952–9961.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. 2021. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402.

Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*.

Zhixiong Zeng and Wenji Mao. 2022. A comprehensive empirical study of vision-language pre-trained model for supervised cross-modal retrieval. *arXiv preprint arXiv:2201.02772*.

Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10394–10403.