COOKIE: Contrastive Cross-Modal Knowledge Sharing Pre-training for Vision-Language Representation

Keyu Wen[1] Jin Xia[1] Yuanyuan Huang[1] Linyang Li[2] Jiayan Xu[1] Jie Shao[1]

[1]ByteDance AI Lab [2]Fudan University

Abstract

There has been a recent surge of interest in cross-modal pre-training. However, existed approaches pre-train a one-stream model to learn joint vision-language representation, which suffers from calculation explosion when conducting cross-modal retrieval. In this work, we propose the Contrastive Cross-Modal Knowledge Sharing Pre-training (COOKIE) method to learn universal text-image representations. There are two key designs in it, one is the weight-sharing transformer on top of the visual and textual encoders to align text and image semantically, the other is three kinds of contrastive learning designed for sharing knowledge between different modalities. Cross-modal knowledge sharing greatly promotes the learning of unimodal representation. Experiments on multi-modal matching tasks including cross-modal retrieval, text matching, and image retrieval show the effectiveness and efficiency of our pre-training framework. Our COOKIE fine-tuned on cross-modal datasets MSCOCO, Flickr30K, and MSRVTT achieves new state-of-the-art results while using only 3/1000 inference time comparing to one-stream models. There are also 5.7% and 3.9% improvements in the task of image retrieval and text matching. Source code will be available at https://github.com/kywen1119/COOKIE.

1. Introduction

Cross-modal pre-training has significantly advanced the progress of representation learning in vision-language field. It aims at narrowing the heterogeneous gap between vision and language [27, 11, 30]. Recent vision-language pre-training(VLP) methods utilize large-scale image-text pairs to learn the unified representation of visual and textual inputs, which greatly improve the performance of V+L tasks such as cross-modal retrieval [19, 48, 12], image captioning [46, 15] and visual question answering [2, 1]. In this paper we focus on multi-modal retrieval tasks including cross-modal retrieval (image-text matching and video-text matching) and single-modal matching (text matching and image retrieval).

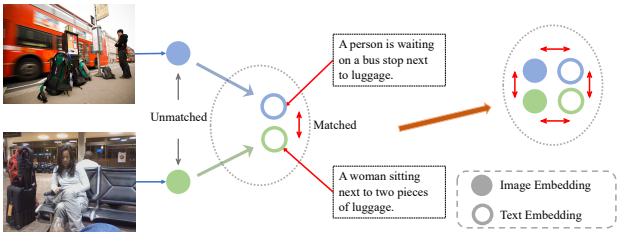


Figure 1: Illustration of cross-modal knowledge sharing. Images with similar semantics are sometimes different in structure, subject, background, and style, which leads to inaccurate matching. By matching with the semantics of the corresponding texts, the distance of their embeddings in the common space is narrowed.

Humans do not perceive the world with just one sense. It is the same for model pre-training: simply using single-modal supervision seems not enough. As illustrated in Fig. 1, two images having the same semantic meaning could look totally different. In this case, we need to resort to cross-modal pre-training. One-stream VLP methods were recently used for cross-modal pre-training. They use multi-layer transformers [45] as the joint encoder. The input is the concatenation of visual tokens and textual tokens. However, there are two obvious shortcomings for such methods: a) Two-stage visual feature extraction with Faster R-CNN [41] is time-consuming and may lose some global information, as discussed in [48]. b) One-stream methods need to process the concatenation of image and text tokens. Such calculation will lead to inference time explosion for retrieval tasks. Double-stream methods are also commonly used for cross-modal pre-training. They use a visual path and a textual path to encode images and texts separately. This leads to high efficiency but very limited performance for cross-modal retrieval. There are two obvious constraints: a) The lack of cross-modal interactions weakens the semantic alignment of images and texts. b) Simple supervision from cross-modal contrastive learning(CCL) loses the knowledge

that the single-modal encoders have learned from the original images or texts.

In this work, we propose **COOKIE**: **C**ontrastive Cr**o**ss-M**o**dal **K**nowledge Shar**i**ng Pr**e**-training, a novel framework designed for multi-modal retrieval tasks. Our COOKIE framework is able to leverage the advantages of both one-stream VLP methods and the double-stream methods while avoiding their aforementioned disadvantages. There are mainly two designs in our framework: The double-stream visual semantic embedding structure with weight-sharing transformer encoder(WS-TE) and the cross-modal and single-modal contrastive learning methods.

The former design, the double-stream visual semantic embedding structure with WS-TE, speeds up cross-modal training and testing while strengthening the semantic alignment of images and texts. More specifically, COOKIE is designed in a double-stream fashion, thus the inference time explosion caused by one-stream methods is avoided. In the visual stream, the feature is extracted by ResNet instead of Faster-RCNN. In this way, we avoid the huge computation cost while keeping the global visual information. To address the absence of cross-modal interactions which previous double-stream methods lack, a weight-sharing transformer encoder(WS-TE) is designed to force the model to pay more attention to tokens with the same semantic meanings, which guarantees refined vision-language alignment.

Secondly, our COOKIE is optimized by three kinds of contrastive learning: cross-modal contrastive learning, single-modal visual contrastive learning(VCL) and textual contrastive learning(TCL). Compared with single-modal methods [51, 17, 49], cross-modal contrastive pre-training shares knowledge of pre-trained image encoders and text encoders, e.g. ResNet and BERT. An explanation of cross-modal knowledge sharing can be seen in Fig. 1. The two pictures have the same semantic meaning "a person is waiting with luggage", but are quite different due to camera angle and background. With the help of cross-modal contrastive learning, the image embeddings are drawn closer to each other by the lead of text embeddings. Meanwhile, we don't expect the single-modal encoders to lose too much information learned from large-scale unimodal pretraining. Thus, VCL and TCL are added to maintain the single-modal knowledge learned from original images and texts. Our single-modal objectives differ from structure preserving losses [47, 43]. By manually searching for positive within-modal pairs, they promote the alignment of cross-modal semantics. While our design is much simpler and more effective due to automatically generated pairs. Further, our VCL and TCL also allow the visual and textual encoder to retain the ability to capture within-modal similarity, which helps single-modal retrieval tasks.

To summarize, we make the following contributions.

- We propose a new cross-modal pre-training paradigm COOKIE. With specially designed weight-sharing transformer encoder(WS-TE), COOKIE provides both efficiency from its double-stream structure and comparable effectiveness of one-stream methods.
- Three pre-training objectives including cross-modal contrastive learning(CCL) and single-modal contrastive learning(VCL and TCL) are designed for cross-modal knowledge sharing which promotes multi-modal retrieval.
- The proposed COOKIE outperforms previous methods on multi-modal matching tasks including image-text matching, video-text matching, text matching and image retrieval. Specifically, our COOKIE achieves comparable results with sota method Oscar [30] using only 3/1000 inference time on Flickr30K and MSCOCO. COOKIE increases $R@1$ of MSRVTT from 16.0 to 20.0. For single-modal matching tasks, our model has 3.9% and 5.7% performance gains on text matching and image retrieval respectively.

2. Related Work

2.1. Multi-modal retrieval and matching

Multi-modal retrieval and matching tasks include cross-modal matching and single-modal retrieval and matching. In this paper we mainly discuss four of them: image-text matching, video-text matching, image retrieval and text matching, which can furthest prove the effectiveness and efficiency of our pre-training framework. Initially, CCA [44] creates a paradigm for cross-modal retrieval, that is, to map images and texts to a common subspace and measure their similarity. Recently Faghri et al. [19] proposed a hinge-based hard triplet loss, which acts as a baseline for later methods. SCAN [25] utilizes object detection methods like Faster R-CNN [41] to extract regional visual features. Image retrieval requires finding the most relevant images given the image query [18]. Semantic text similarity(STS) [5] is a classic text matching task, aiming to measure the similarity of two given sentences.

2.2. Cross-modal pre-training

Inspired by single-modal self-supervised pre-training, like visual contrastive learning [8, 9, 10] and textual masked language modeling [39, 17], cross-modal encoders can also be pre-trained with large-scale image-text pairs for better performance, which can be divided into two kinds. Inputs are usually visual regional features and word embeddings. One kind of them [34] applies two transformers for images and texts and one unified transformer in a later stage, while the other kind directly takes the concatenation of region features and word embeddings as input and process it with one deep transformer. Our method doesn't belong to them. For

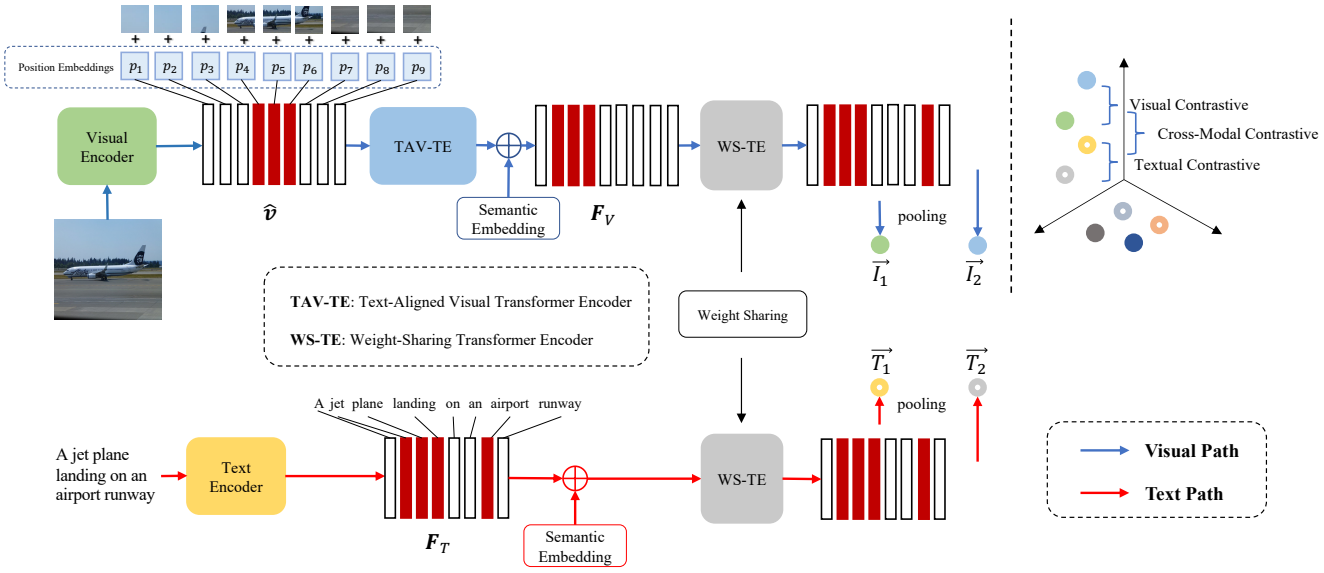


Figure 2: An overview of the proposed COOKIE. It consists of two paths, the visual path and the textual path. The visual path contains a CNN to extract the patch features, a text-aligned visual transformer and a weight-sharing transformer. The textual path has a BERT encoder and the same weight-sharing transformer. We design three contrastive learning objectives.

retrieval efficiency, we directly use the raw images and sentences as input and process them in a two-stream way with one weight-sharing transformer.

2.3. Contrastive Learning

We conduct three kinds of contrastive learning: visual contrastive learning(VCL), textual contrastive learning(TCL) and cross-modal contrastive learning(CCL). Contrastive learning plays an important part in representation learning as well as content-based retrieval. For VCL, the goal is to minimize the distance between learned representations of the raw image and the augmented image, as in [8, 9, 36, 10]. Wu et al. [49] proved the effectiveness of TCL based on BERT [17]. CCL learns the common subspace of visual and textual modalities as discussed in Section 3.2. For example, researchers [29, 52] utilize CCL to do single-modal or multi-modal understanding and generation tasks. All of them are carefully designed for efficient and effective multi-modal retrieval and matching tasks.

3. Pre-Training

In this section, we detail our contrastive cross-modal knowledge sharing pre-training for vision-language representation. In Section 3.1, we describe the model architecture which consists of an image encoder, a text encoder, a text-aligned visual transformer encoder and a weight-sharing transformer encoder. In Section 3.2, we introduce cross-modal contrastive pre-training aiming at cross-modal alignment and knowledge transferring. In Section 3.3, single-modal contrastive pre-training is detailed.

3.1. Overall Structure

Our structure is shown in Fig. 2. Previous vision-language pre-training methods [34, 11, 30] take the concatenation of image regional features extracted by Faster R-CNN [41] and textual word embeddings as input and process it with transformer-based models [17]. Different from them, we directly use ResNet [21, 35] and BERT [17] to process images and texts separately. To be specific, given an image-text pair (V, C), the goal is to learn the individual embedding $\vec{I}$ and $\vec{T}$, which can be used for multi-modal retrieval.

Visual Representation Learning We directly use ResNet for visual feature extraction. Previous VLP methods utilize bottom-up and top-down(BUTD) attention to extract regional features, which results in the two-stage training and inference process. Our end-to-end way guarantees the efficiency and takes more global features into account than BUTD methods. We remove the last fully-connected layer of ResNet [21] or ResNeXt [35] and flat the output feature before pooling, which results in visual patch features $\mathbf{v} = \{v_1, v_2, ..., v_n\} \in \mathbb{R}^{D_V}$, where n is the patch number and D_V is the visual feature dimension. A fully-connected layer is followed. To learn the relative positions of features in the image, we add position embeddings. The output visual features are $\hat{\mathbf{v}} = \{\hat{v_1}, \hat{v_2}, ..., \hat{v_n}\} \in \mathbb{R}^{D}$.

$$\hat{v_i} = v_i W_V + b_V + p_i, \tag{1}$$

where $W_V \in \mathbb{R}^{D_V \times D}, b_V, p_i \in \mathbb{R}^D$. W_V and b_V are the FC parameters and p_i is the position embedding for patch i.

Textual Representation Learning We take the output of the last layer of the pre-trained BERT-base model [17] as the text features. The textual features are denoted as $\mathbf{t} = \{t_1, t_2, ..., t_m\} \in \mathbb{R}^{D_T}$, where m is the number of words and D_T is the dimension of word feature. An FC encodes word features into the same space with image features. The output textual features are $\mathbf{F}_T = \{f_{T_1}, f_{T_2}, ..., f_{T_m}\} \in \mathbb{R}^D$.

$$f_{T_i} = t_i W_T + b_T + s_T, \tag{2}$$

where $W_T \in \mathbb{R}^{D_T \times D}, b_T \in \mathbb{R}^D$ in Eq. 2 are the FC parameters. Same as [22], we add a textual semantic embedding vector s_T to the features.

Text-Aligned Visual Transformer Since the convolution used in CNN is a local operator while the transformer layer used in BERT is a global operator, visual features extracted from CNNs could have different distribution from textual features. To align the distribution of visual features and textual features, we add a text-aligned visual transformer encoder(TAV-TE). TAV-TE provides a global attention calculation for the image side. The transformer encoder(TE) in this paper follows the standard definition [45]. We add a visual semantic embedding vector s_V to the features.

$$\mathbf{F}_V = TE_{TAV}(\hat{\mathbf{v}}) + s_V. \tag{3}$$

Weight-Sharing Transformer To prompt images and texts to focus on the same semantics, we add a weight-sharing transformer encoder(WS-TE) on top of the network. WS-TE contains a multi-head self-attention process and a feed-forward network, which makes input tokens pay more attention to the salient areas. Originally in CNNs, sharing weights between convolution kernels not only reduces parameters but also enables translation equivariant [26]. That is, the network extracts the same features no matter how the image translates. Similarly for images and texts, parameter sharing enables the self-attention layer to give close attention values for analogous semantics of images and texts. As our goal is to align visual and textual representations, if the similar semantics of images and texts are given greater weights, the final representation will also be better aligned.

$$\vec{I} = Pooling(TE_{WS}(\mathbf{F}_V)), \vec{T} = Pooling(TE_{WS}(\mathbf{F}_T)). \tag{4}$$

3.2. Cross-Modal Contrastive Learning

Cross-modal contrastive learning plays a key role in cross-modal retrieval. It learns a common subspace for images and texts where they are semantically aligned. At the same time, such a learning process enables cross-modal knowledge transferring, namely from image to language understanding and vice versa.

The image and text encoders together with weight-shared TE are optimized with InfoNCE loss [36], which is widely used in contrastive learning. For L_{i2t}, the positive sample is the matched text and the negative samples are the remaining texts in the mini batch. For L_{t2i}, vice versa.

$$L_{InfoNCE}(q, k) = -\frac{1}{N}\sum_{i=1}^{N} log\frac{exp(q \cdot k^+/\tau)}{exp(q \cdot k^+/\tau) + \sum_{j=1}^{N-1} exp(q \cdot k^-/\tau)}, \tag{5}$$

$$L_{i2t} = L_{InfoNCE}(\vec{I}, \vec{T}), \tag{6}$$

$$L_{t2i} = L_{InfoNCE}(\vec{T}, \vec{I}). \tag{7}$$

Here N is the size of mini batch. $+$ and $-$ refer to the positive sample and the negative sample respectively. τ is a temperature hyper-parameter.

3.3. Single Modal Contrastive Learning

Cross-modal contrastive learning promotes knowledge sharing between image encoder and text encoder. However, we don't expect the encoder to lose too much information learned from single-modal data. Thus we design visual contrastive learning and textual contrastive learning to maintain the single-modal encoder's ability to process its own modal data.

Visual Contrastive Learning Image self-supervised learning can effectively improve the deep neural network's ability to understand images [8, 9, 10]. In our framework, we utilize visual contrastive learning to enhance the image encoder's understanding of images while accepting knowledge from text. Two augmentations of the raw image act as the input and the goal is to draw the two learned representations closer. Specifically, we directly minimize the distance between the positive pairs while maximizing the distance of the negative pairs. Given a raw image V, the image encoder together with the weight-sharing TE is denoted as $\mathbf{E}_V$. We optimize the visual InfoNCE loss.

$$\vec{I_1} = \mathbf{E}_V(aug_{v_1}(V)), \vec{I_2} = \mathbf{E}_V(aug_{v_2}(V)), \tag{8}$$

$$L_i = L_{InfoNCE}(\vec{I_1}, \vec{I_2}), \tag{9}$$

where $aug_v(\cdot)$ denotes image augmentation. For our method, the image augmentation includes randomly cropping, flipping, color jitter, gaussian blur and color dropping.

Methods	Image-to-Text			Text-to-Image				Image-to-Text			Text-to-Image			
	1K Test Set							5K Test Set						
	R@1	R@5	R@10	R@1	R@5	R@10	Rsum	R@1	R@5	R@10	R@1	R@5	R@10	Rsum
Double-Stream Methods														
VSE++	64.6	90.0	95.7	52.0	84.3	92.0	478.6	41.3	71.1	81.2	30.3	59.4	72.4	355.7
DSRAN*	80.6	96.7	98.7	64.5	90.8	95.8	527.1	57.9	85.3	92.0	41.7	72.7	82.8	432.4
GPO^{R101}	78.0	95.8	98.5	62.6	90.6	96.0	521.5	56.2	83.7	90.9	40.8	70.6	81.5	423.7
GPO^{X101*}	85.6	98.0	99.4	73.1	94.3	97.7	548.1	68.1	90.2	95.2	52.7	80.2	88.3	474.8
One-Stream Pre-training Methods														
Pixel-BERT^{X152}	84.9	97.7	99.3	71.6	93.7	97.4	544.6	63.6	87.5	93.6	50.1	77.6	86.2	458.6
Uniter^{b}	-	-	-	-	-	-	-	63.3	87.0	93.1	48.4	76.7	85.9	454.4
Uniter^{l}	-	-	-	-	-	-	-	66.6	89.4	94.3	51.7	78.4	86.9	467.3
Oscar^{b}	**88.4**	**99.1**	**99.8**	**75.7**	**95.2**	**98.3**	**556.6**	70.0	**91.1**	**95.5**	54.0	80.0	**88.5**	479.1
Double-Stream Pre-training Methods														
COOKIE^{R101}	81.3	96.2	98.7	67.5	91.5	96.1	531.3	61.7	86.7	92.3	46.6	75.2	84.1	446.6
COOKIE^{X101}	87.3	98.1	99.6	73.5	94.0	97.5	550.0	69.2	89.6	94.4	52.4	79.6	87.1	472.3
COOKIE^{X101*}	**88.4**	98.5	**99.8**	75.2	94.7	97.5	554.1	**71.6**	90.9	95.4	**54.5**	**81.0**	88.2	**481.6**

Table 1: Results on image-text matching task with MS-COCO dataset. We record results on both $1K$ and $5K$ test set. Here $R101$, $X101$ and $X152$ refer to ResNet101, ResNeXt101 and ResNeXt152. b and l mean base and large models for Uniter and Oscar. $*$ represents model ensemble. The best results are in bold, while the suboptimal values are underlined.

Textual Contrastive Learning For texts, self-supervised learning always consists of masked language modeling(MLM) [17] instead of contrastive learning. However, Wu et al. [49] proved the effectiveness of contrastive learning in sentence representation learning. In our model, randomly masking, substituting, deleting are used for textual augmentation. Such random operations can enhance the robustness of the model. The text encoder retains attention to the semantic features of the sentence while accepting the knowledge from the image. Same for images, we optimize the text encoder together with weight-sharing TE(denoted as $\mathbf{E}_T$) with InfoNCE loss. Given a raw sentence C,

$$\vec{T_1} = \mathbf{E}_T(aug_{t_1}(C)), \vec{T_2} = \mathbf{E}_T(aug_{t_2}(C)), \tag{10}$$

$$L_t = L_{InfoNCE}(\vec{T_1}, \vec{T_2}), \tag{11}$$

Here $aug_t(\cdot)$ in Eq. 10 denotes text augmentation.

The overall pre-training objective of COOKIE is defined below.

$$L_{Pre-training} = L_{i2t} + L_{t2i} + L_i + L_t. \tag{12}$$

4. Experiments

4.1. Pre-training Configurations

Pre-training Corpus For our COOKIE, we use the public available image-text datasets Conceptual-Captions(CC) [42], SBU captions [37], MSCOCO [31], Flickr30K [38], VQA2.0 [20] and GQA [23]. This results in the total size of 3.9 million images and 5.9 million image-text pairs.

Implementations We select ResNet50, ResNet101 [21] or ResNeXt101 [35] as the image encoder and BERT-base [17] as the text encoder. All images are reshaped to 512×512, if not otherwise specified. The dimensions of output features of the image encoder and text encoder D_V and D_T are 2048 and 768, respectively. The dimension of the cross-modal space D is set to 1024. The number of image patches n is $16 \times 16 = 256$ while the number of words m is set to 50. The weight-sharing TE has 2 layers and the TAV-TE has 1 layer. The intermediate size and multi-head number are set to 1024 and 8.

We pre-train the model with AdamW [33] for two stages. During the first stage, for stability, the model is merely trained with L_{i2t} and L_{t2i} for 30 epochs with the batch size set to 576. At the second stage, we use the full $L_{Pre-training}$ to supervise the training for 10 epochs with the batch size set to 288. The learning rate is 2e-5 initially and decays by 10 times after half of the total epochs for each stage. It is noticed that LR for ResNeXt101 is one-tenth of the global LR. Experiments are conducted with Tensorflow v2.2 on 48 Tesla V100 GPUs.

4.2. Downstream Matching Tasks

Our COOKIE is designed for multi-modal matching tasks including image-text matching, video-text matching, text matching and content based image retrieval. All these

Methods	Image-to-Text			Text-to-Image			
	R@1	R@5	R@10	R@1	R@5	R@10	Rsum
Double-Stream Methods							
VSE++	52.9	80.5	87.2	39.6	70.1	79.5	409.8
DSRAN*	80.5	95.5	97.9	59.2	86.0	91.9	511.0
GPO^{R101}	77.9	93.7	97.4	57.5	83.4	90.2	500.2
GPO^{X101*}	88.7	**98.9**	**99.8**	**76.1**	94.5	97.1	555.1
One-Stream Pre-training Methods							
Pixel-BERT^{X152}	87.0	**98.9**	99.5	71.5	92.1	95.8	544.8
Uniter^{b}	85.9	97.1	98.8	72.5	92.4	96.1	542.8
Uniter^{l}	87.3	98.0	99.2	75.6	94.1	96.7	550.9
Double-Stream Pre-training Methods							
COOKIE^{R101}	84.7	96.9	98.3	68.3	91.1	95.2	534.5
COOKIE^{X101}	**89.0**	**98.9**	**99.8**	75.6	94.5	97.1	554.9
COOKIE^{X101*}	**89.0**	**98.9**	99.7	75.6	**94.6**	**97.2**	**555.3**

Table 2: Results on image-text matching task with Flickr30K dataset.

Methods	Video-to-Text			Text-to-Video			
	R@1	R@5	R@10	R@1	R@5	R@10	Rsum
VSE++	14.4	34.1	45.6	8.3	24.0	34.1	160.5
HGR	15.0	36.7	48.8	9.2	26.2	36.5	172.4
GPO	16.0	38.6	50.2	8.7	25.3	35.9	174.7
COOKIE($mean$)	19.4	40.7	51.5	**9.8**	27.6	38.6	187.6
COOKIE(gpo)	**20.0**	**42.0**	**54.9**	**9.8**	**28.3**	**39.6**	**194.6**

Table 3: Results on video-text matching task with MSRVTT dataset. Here $mean$ refers to mean pooling and gpo means the same pooling strategy with GPO [6].

tasks ask the quality of the learned representation as well as the inference speed, which are well addressed by our pre-training. All tasks use BERT-base [17] as the text encoder. The statistics of the downstream datasets and implementation details for finetuning on downstream tasks can be seen in Appendix.

Image-Text Matching Image-text matching(ITM) is a fundamental task in cross-modal representation learning, which requires semantic consistency of visual and textual representations. ITM includes image-to-text retrieval and text-to-image retrieval. Same as traditional double-stream methods [19, 28, 48], a hinged hard triplet loss supervises the fine-tuning process. We conduct experiments on two widely used datasets MSCOCO [31] and Flickr30K [38] and use the same train-dev-test split with [24]. We record recall at K ($R@K$) together with $Rsum$. ResNet101 [21] and ResNeXt101 [35] are used. We compare our COOKIE with double-stream methods without pre-training [19, 48, 6] as well as one-stream methods with pre-training [22, 11, 30]. Results can be seen in Table 1 and Table 2.

Methods	STS12	STS13	STS14	STS15	STS16	STSB
BERT	28.8	50.8	43.9	57.6	58.7	46.1
RoBERTa	47.4	37.5	47.9	55.1	57.6	71.9
MACD	-	-	-	-	-	71.8
CLEAR	49.0	48.9	57.4	63.6	65.6	72.5
COOKIE	**63.2**	**68.0**	**68.0**	**72.4**	**68.1**	**75.3**

(a) Results on STS task. Mean values of Pearson and Spearman are recorded.

Methods	MSCOCO			NUS-WIDE		
	16bit	32bit	64bit	16bit	32bit	64bit
HashNet	0.745	0.773	0.788	0.757	0.775	0.790
DCH	0.759	0.801	0.825	0.773	0.795	0.818
CSQ	0.796	0.838	0.861	0.810	0.825	0.839
COOKIE	**0.811**	**0.884**	**0.910**	**0.822**	**0.852**	**0.855**

(b) Results on image retrieval task with MSCOCO and NUSWIDE datasets.

Table 4: Experimental results for single-modal matching tasks including (a) text matching and (b) image retrieval.

Video-Text Matching Similar to ITM, video-text matching(VTM) ranks the sentence features by similarity with the video query or vice versa. We do experiments on MSRVTT dataset [50]. For fair comparison, we use the same video features extracted by ResNet152 [21] pre-trained on ImageNet [16]. Thus for VTM no image encoder is used. Conditioned on it, we take the output of our pre-trained BERT as the text features. The final visual and textual representations $\vec{I}$ and $\vec{T}$ come from processing the frame and text features using a pooling strategy. We use either mean-pooling or g-pooling [6]. The objective function is the InfoNCE loss [36]. We use the same split as [7, 6] and compare our results with them. Results are recorded in Table 3.

Text Matching Text Matching(TM) is a single-modal matching task regarding only texts. We focus on semantic text similarity(STS) [5], a classic task to evaluate text representation learning by calculating the similarity of two input sentences. As illustrated in [40], directly computing the cosine similarity of two text representations is much more efficient than processing the concatenation of two sentences as did in [17]. Having two sentence embeddings $\vec{T}_1$ and $\vec{T}_2$, we compute their cosine similarity using the pre-trained models, which is an unsupervised process. The labels of STS are decimals in range 0-5. We conduct experiments on the widely used STS12-16 and STSbenchmark datasets [5] and report the mean values of Pearson and Spearman metrics. Results are in Table 4a. We compare with methods only pre-trained on texts [17, 32, 49] and MACD [14] which is pre-trained using multi-modal data.

Image Retrieval Image retrieval(IR) has great practical value in real life [18]. To mostly utilize the knowl-

(a) **Pre-training Tasks**

$\mathcal{CCL}$	$\mathcal{VCL}$	$\mathcal{TCL}$	ITM	IR	TM
			526.9	0.861	46.1
✓			545.4	0.898	72.4
✓	✓		547.8	0.909	72.1
✓		✓	548.1	0.899	75.1
✓	✓	✓	**550.0**	**0.910**	**75.3**

(b) **Weight-sharing & TAV TE**

model	ITM
baseline	536.6
FC Layers(w/ weight sharing)	536.8
WS-TE(w/o weight sharing)	541.6
WS-TE(w/ weight sharing)	547.5
WS-TE(w/o weight sharing)+TAV-TE	543.8
WS-TE(w/ weight sharing)+TAV-TE	**550.0**

(c) **Num of Layers of WS-TE**

w/ pre-training		**w/o pre-training**	
num	ITM	num	ITM
0×	536.6	0×	528.7
1×	547.8	1×	**534.6**
2×	**550.0**	2×	526.9
4×	539.8	4×	510.4
8×	500.7	8×	250.8

(d) **Pre-training Corpus**

dataset	ITM	IR	TM
MSCOCO(11w)	530.4	0.895	71.9
Flickr30K(3w)	531.7	0.886	69.5
CC(2.8M)	540.5	0.902	74.4
CC+SBU(3.6M)	544.3	0.905	**75.3**
CC+SBU+COCO+F30K(4.2M)	547.6	**0.910**	75.1
CC+SBU+COCO+F30K+VQA+GQA(5.9M)	**550.0**	0.908	74.9

(e) **Visual Backbones**

	w/ pre-training			**w/o pre-training**		
model	ITM	IR	TM	ITM	IR	TM
ResNet50	526.4	0.910	74.8	512.8	0.861	46.1
ResNet101	531.3	0.911	**75.3**	516.5	0.872	46.1
ResNeXt101	**550.0**	**0.932**	75.0	**526.9**	**0.911**	46.1

Table 5: **Ablation Experiments.** Here $\mathcal{CCL}$, $\mathcal{VCL}$ and $\mathcal{TCL}$ are cross-modal, visual and textual contrastive learning. ITM, IR and TM refer to image-text matching, image retrieval and text matching tasks. We record $Rsum$ of MSCOCO 1k test set for ITM, MAP@5000 of MSCOCO-64bit test set for IR and mean value of Pearson and Spearman of STS-B test set for TM.

edge learned from our pre-training, we do experiments on MSCOCO [31] and NUS-WIDE [13] datasets which require more understanding of the semantic meanings of the entire picture than the matching of key points. We use CSQ [51], the current sota method on these benchmarks, as our baseline and substitute the image encoder with our pre-trained image encoder which contains much more information learned from cross-modal data. For fair comparison, ResNet50 [21] is used and the image size is set to 224. MAP@5000 is recorded. We compare COOKIE with HashNet [4], DCH [3], and CSQ. Results are in Table 4b.

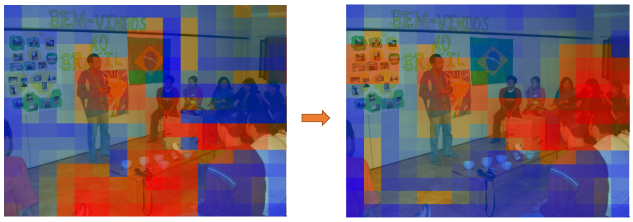
Performance Comparison with SoTA Our contrastive cross-modal knowledge sharing pre-training learns universal multi-modal representations for downstream matching tasks. Specifically, for cross-modal retrieval, COOKIE sets new sota results for Flickr30K and MSRVTT and achieves comparable results on MSCOCO with Oscar [30] consuming only 3/1000 inference time. For image-text matching task, comparing to traditional double-stream methods including GPO [6] with ResNeXt101 [35], our pre-training structure significantly improves performances, as seen in Table 1 and Table 2. When compared with two-stage pre-training methods coupled with Faster R-CNN [41] like Uniter [11] and Oscar, our COOKIE not only has the advantage of speed, but we also use less pre-training data (5.9M vs 6.5M&9.6M). Our model also outperforms Pixel-BERT [22] which uses ResNeXt-152. In Table 3, our image-text pre-training greatly promotes video-text matching on MSRVTT dataset, increasing R@1 of V2T from 16.0 to 20.0 and 9.2 to 9.8 for T2V.

As for single-modal matching tasks, COOKIE also sets new sota results, proving the effectiveness of our cross-modal knowledge sharing. For text matching, as seen in Table 4a, there is 3.9% performance gain on STS-B and more obvious improvements on five datasets for STS12 to STS16. It is noticed that BERT, RoBERTa and CLEAR are all trained with mere texts. Our cross-modal pre-training successfully shares visual semantics with the text encoder. COOKIE also outperforms MACD [14] which uses similar cross-modal pre-training. Concurrently for image retrieval, we obtain 5.7% and 1.9% improvement on MSCOCO-64bit and NUSWIDE-64bit respectively, as seen in Table 4b. All the performance growths come from contrastive cross-modal knowledge sharing.

4.3. Ablation Study

We conduct several ablation studies to explore the performance of COOKIE under various model settings. Results are in Table 5. For pre-training, the default encoders are ResNeXt101 [35], BERT-base [17], the TAV-TE and the 2-layer WS-TE. Models are trained with 3 losses using the full pre-training dataset. For image-text matching(ITM), the default model setting is the same as the one for pre-training. For image retrieval(IR) and text matching(TM), the visual backbone is substituted with ResNet50 and ResNet101 respectively. As VTM is similar to ITM, we select ITM as a representative of cross-modal retrieval.

Effectiveness of Three Contrastive Losses We propose three contrastive losses to supervise pre-training. Cross-modal contrastive learning(CCL) is designed for bridging the heterogeneous gap between the two modalities, while visual and textual contrastive learning(VCL & TCL) help retain the knowledge that the single-modal encoder originally learned from the respective modality. As seen in Table 5a, ITM benefits more from CCL, while IR and TM rely on both CCL and single-modal contrastive learning.



(a) w/o WS-TE (b) w/ WS-TE

Figure 3: Illustration of the effect of weight-sharing transformer encoder. With the WS-TE, images and texts concentrate on the same semantics.

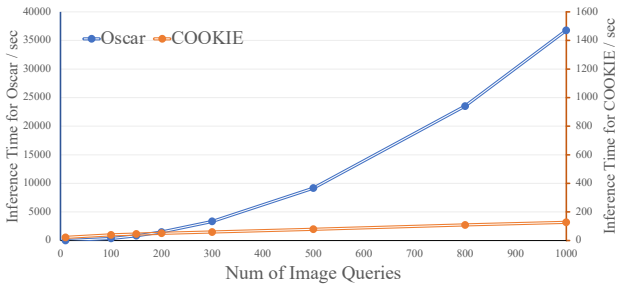


Figure 4: Comparison of inference time against Oscar.

Transformer Encoders Two special transformer encoders are applied. Text-aligned-visual transformer encoder(TAV-TE) transfers visual features' distribution before the weight-sharing transformer encoder(WS-TE). WS-TE is a crucial module that ensures the image and text pay attention to the same semantics. These two designs are both crucial to cross-modal alignment, as seen in Table 5b.

Depth of WS-TE A critical point for designing the transformer encoder is the depth of it. Thus we explore the optimal number of layers of WS-TE in Table 5c for better attention. Without pre-training, the best number is 1. The pre-training stage enables deeper network, resulting in the best number of 2. As the WS-TE is trained from scratch, size of pre-training data determines the max depth of the network. For the image-text data we have, the depth of 2 layers may currently be a bottleneck.

Size of Pre-training Corpus The size of image-text pairs plays a key role in cross-modal pre-training [30]. Uniter [11] utilizes $9.6M$ pairs and Oscar uses $6.5M$. We record results on three tasks with different sizes of pre-training corpus in Table 5d. It is noticed that VQA and GQA datasets are composed of image-question pairs, which leads to little improvement. And the growth of data size can't bring significant improvement in single-modal tasks.

Different Visual Backbones To prove the robustness of our COOKIE, we substitute the visual backbones with or without pre-training. In Table 5e, as expected, the stronger the visual backbone, the stronger COOKIE is for ITM and IR. However for TM, better visual encoders don't bring performance gain. We infer it's because if the visual encoder is too powerful, the text encoder will lose too much original information. For texts, we only use BERT-base model due to the limitation of computing resources. We leave the pre-training with BERT-Large model for future work.

4.4. Analysis

Analysis of Weight-Sharing Transformer We design a weight-sharing transformer encoder(WS-TE) at the end of the network. Although images and texts have cross-modal heterogeneous gap, the process of weight-sharing attention constrains the two paths to focus on the tokens with the same semantics. We visualize the attention learned by the WS-TE in Figure 3. Same as [28, 48], considering the final representation should pay more attention to salient objects in the images or texts, we compute similarities of the final representation $\vec{I}$ or $\vec{T}$ between the tokens after the WS-TE. In this way, every area has a similarity score with the final representation. Then we rank the scores and the areas with higher ranks are marked brighter in the figures. We mark the top-5 words for texts. As we can see from the figure, without WS-TE(the figure on left), the image and the sentence pay attention to different semantics. In the text, "man", "jacket" and "people" are salient, while in the image, more attention is given to irrelevant "flag" and "table". With the WS-TE(the figure on right), images and texts are prone to emphasize the same semantics, which are "man" and "group of people".

Analysis of Inference Time COOKIE is a two-stream method without cross-modal interaction, thus it greatly speeds up the task of image-text retrieval. We conduct experiments on Flickr30K test set and record the inference time(feature extraction plus similarity computing). As shown in Fig. 4, one-stream methods like Oscar [30] have a $O(n^2)$ time complexity against $O(n)$ of our model.

5. Conclusion

In this paper, we propose a new Crontrastive Cross-Modal Knowledge Sharing Pre-training(COOKIE) to learn universal separate vision and language representations for downstream matching tasks. We design a weight-sharing transformer encoder to better align visual and textual semantics and pre-train the model with cross-modal contrastive learning together with single-modal contrastive learning using 5.9M image-text pairs. COOKIE sets new state-of-the-art results on single-modal matching tasks and at the same time reaches comparable results on cross-modal retrieval with only 3/1000 inference time.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[3] Yue Cao, Mingsheng Long, Bin Liu, and Jianmin Wang. Deep cauchy hashing for hamming space retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1229–1237, 2018.

[4] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Philip S Yu. Hashnet: Deep learning to hash by continuation. In *Proceedings of the IEEE international conference on computer vision*, pages 5608–5617, 2017.

[5] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 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, 2017.

[6] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15789–15798, 2021.

[7] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10638–10647, 2020.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[9] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[10] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[11] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020.

[12] Qingrong Cheng and Xiaodong Gu. Bridging multimedia heterogeneity gap via graph representation learning for cross-modal retrieval. *Neural Networks*, 134:143–162, 2021.

[13] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009.

[14] Wanyun Cui, Guangyu Zheng, and Wei Wang. Unsupervised natural language inference via decoupled multimodal contrastive learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5511–5520, 2020.

[15] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. Learning to evaluate image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5804–5812, 2018.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[18] Shiv Ram Dubey. A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[19] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[20] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

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

[23] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[24] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[25] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

[26] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 991–999, 2015.

[27] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020.

[28] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4654–4662, 2019.

[29] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

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

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

[32] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

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

[34] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13–23, 2019.

[35] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181–196, 2018.

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

[37] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*, 2011.

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

[39] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[40] Nils Reimers, Iryna Gurevych, Nils Reimers, Iryna Gurevych, Nandan Thakur, Nils Reimers, Johannes Daxenberger, Iryna Gurevych, Nils Reimers, Iryna Gurevych, et al. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[42] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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, 2018.

[43] Christopher Thomas and Adriana Kovashka. Preserving semantic neighborhoods for robust cross-modal retrieval. In *European Conference on Computer Vision*, pages 317–335, 2020.

[44] Bruce Thompson. Canonical correlation analysis. *Encyclopedia of statistics in behavioral science*, 2005.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[46] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[47] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016.

[48] Keyu Wen, Xiaodong Gu, and Qingrong Cheng. Learning dual semantic relations with graph attention for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[49] Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*, 2020.

[50] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[51] Li Yuan, Tao Wang, Xiaopeng Zhang, Francis EH Tay, Zequn Jie, Wei Liu, and Jiashi Feng. Central similarity quantization for efficient image and video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3083–3092, 2020.

[52] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 833–842, 2021.