VLMixer: Unpaired Vision-Language Pre-training via Cross-Modal CutMix

Teng Wang [1 2] **Wenhao Jiang** [3] **Zhichao Lu** [1] **Feng Zheng** [1] **Ran Cheng** [1] **Chengguo Yin** [3] **Ping Luo** [2]

Abstract

Existing vision-language pre-training (VLP) methods primarily rely on paired image-text datasets, which are either annotated by enormous human labors, or crawled from the internet followed by elaborate data cleaning techniques. To reduce the dependency on well-aligned image-text pairs, it is promising to directly leverage the large-scale text-only and image-only corpora. This paper proposes a data augmentation method, namely cross-modal CutMix (CMC), for implicit cross-modal alignment learning in unpaired VLP. Specifically, CMC transforms natural sentences from the textual view into a multi-modal view, where visually-grounded words in a sentence are randomly replaced by diverse image patches with similar semantics. There are several appealing proprieties of the proposed CMC. First, it enhances the data diversity while keeping the semantic meaning intact for tackling problems where the aligned data are scarce; Second, by attaching cross-modal noise on uni-modal data, it guides models to learn token-level interactions across modalities for better denoising. Furthermore, we present a new unpaired VLP method, dubbed as VLMixer, that integrates CMC with contrastive learning to pull together the uni-modal and multi-modal views for better instance-level alignments among different modalities. Extensive experiments on five downstream tasks show that VLMixer could surpass previous state-of-the-art unpaired VLP methods. Project page: https://github.com/ttengwang/VLMixer

1. Introduction

Vision-language pre-training (VLP) has received increasing attention and brought real benefits to a large variety of downstream tasks in the recent past (Tan & Bansal, 2019; Li et al., 2019b; Lu et al., 2019; Chen et al., 2019; Li et al., 2020b; Cao et al., 2020; Hu et al., 2020; Li et al., 2020a; Zhang et al., 2021; Radford et al., 2021; Kim et al., 2021; Li et al., 2021a; Jia et al., 2021). The success of existing VLP models mainly comes from manually-labeled and well-aligned image captioning datasets, such as COCO (Lin et al., 2014) and Visual Genome (Krishna et al., 2017b), and high-capacity transformer models (Vaswani et al., 2017) with effective pre-training objectives for discovering the cross-modal alignments. In mainstream VLP methods, the modeling of cross-modal alignment has been proved to be effective in achieving promising performance for several downstream tasks (Cao et al., 2020). At a global level, image-text matching losses (Li et al., 2020b; Chen et al., 2019; Zhang et al., 2021) are designed to guide the model to judge whether the input image and sentence are aligned. With the warranty of the instance-level alignment, the self-attention layers could further excavate the fine-grained interactions between input tokens across two modalities in an implicit manner.

Although promising performance has been reported, the improvements of these methods that pre-trained on well-aligned datasets have gradually reached saturation due to the cost of annotating large-scale datasets. The following works alleviate this issue by introducing weakly-aligned image-caption pairs, which contain noisy annotations but are easy to access and scale-up. Unpaired vision-language pre-training (Li et al., 2021b) further relieves the reliance on paired image-caption data, aiming to learn multi-modal

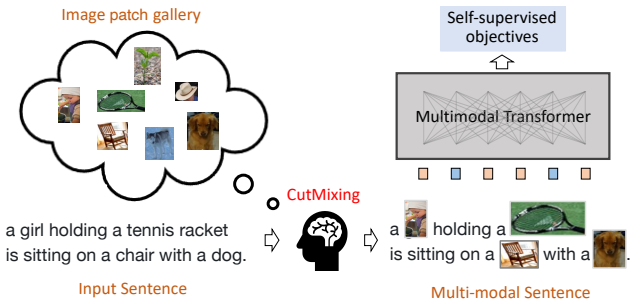

Figure 1. Illustration of the cross-modal CutMix (CMC). By randomly replacing the grounded words in a sentence with visual tokens, we obtain diverse "multi-modal sentences" without changing the semantics but injecting cross-modal noises.

[1]Department of Computer Science and Engineering, Southern University of Science and Technology [2]Department of Computer Science, The University of Hong Kong [3]Data Platform, Tencent. Correspondence to: Feng Zheng <f.zheng@ieee.org>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

representation from the standalone image and text corpus.

Without explicit annotations of the cross-modal correspondence, unpaired VLP faces the challenge of distinguishing the alignment degree between an image and a text effectively. Previous work (Li et al., 2021b) utilizes a shared encoder to learn a joint representation space, meanwhile introducing image tags as an intermediate representation to bridge the two modalities. We argue that, image tags are not reliable representations for complex images, as the permutation-invariant nature and the lack of syntactic structure make them unrecognizable for visual relationships between objects. This further hurts downstream tasks that heavily rely on fine-grained alignments between images and texts, such as NLVR^2 (Suhr et al., 2018) and image-text retrieval.

For fine-grained alignments across modalities, we propose the cross-modal CutMix (CMC) to construct a new representation, "multi-modal sentence", to connect images and texts, which not only preserves the linguistic nature of a sentence but also links to the visual elements in images. A natural sentence can be transformed into its multi-modal view by replacing some grounded words with the image patches of the same semantic meaning[1]. To this end, we create a visual patch gallery with diverse visual patterns from the image-only datasets, where high-quality visual patches are detected and tagged by a concept detector. As shown in Fig. 1, the input sentence after cutmixing not only preserves the syntactic and semantic information but also introduces the visual tokens as the cross-modality noise. Together with the mask-then-predict training objectives, it is promising for the model to learn cross-modal interactions among input tokens and token-level alignment between "grounded words" and image patches.

Furthermore, we propose a contrastive learning framework to fully exploit the instance-level alignments between modalities. For an input sentence, CMC could produce a multi-modal view of the sentence, which has the same semantics as the language view. The contrastive supervision is then adopted to pull together the semantic-similar instances with different views and push away semantically different instances from the anchor. By distinguishing the positive samples from negative samples, the model could judge the alignment between inputs with different modalities.

Our key contributions are summarized as follows:

- We propose cross-modal CutMix to construct a multi-modal representation to bridge the images and texts, guiding the model to learn cross-modal alignment at the token level.

[1]We assume that the text corpus shares a proportion of visual concepts with the image dataset, as it is unpractical to align arbitrary uni-modal datasets with semantic disparity, such as aligning cooking images with a corpus of mathematical terms.

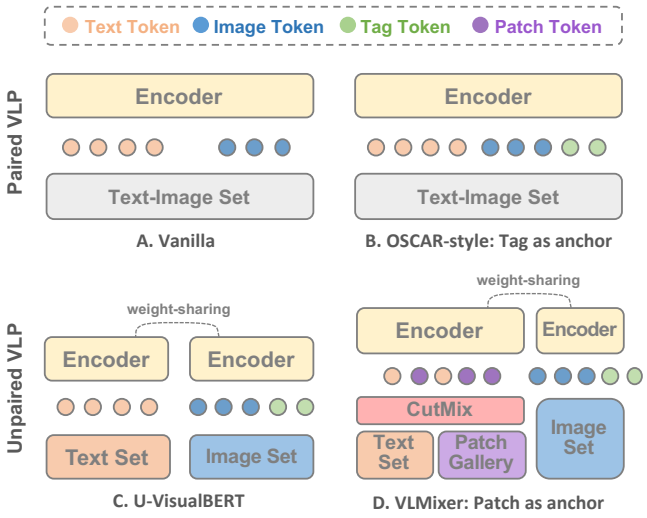

Figure 2. Comparison between existing methods and our framework in model structure and token construction. (A) Vanilla-style methods (Chen et al., 2019; Tan & Bansal, 2019; Li et al., 2019b) directly concatenate the visual tokens (object or grid features) with paired language tokens as inputs. (B) Oscar-style methods (Li et al., 2020b; Zhang et al., 2021) utilize the image tags extracted by an object detector, serving as the anchor points that existed in both visual and text data to bridge two modalities for better alignment learning. (C) U-VisualBERT (Li et al., 2021b) extends oscar-style inputs into unpaired VLP and employs two separate branches to process text and image data. (D) VLMixer injects visual patches into the texts to form a "multi-modal sentence", which is considered an intermediate representation to bridge the two modalities, since it keeps the syntactic structure of the original sentence meanwhile linking to the diverse visual patterns.

- We propose cross-modal contrastive learning upon CMC to facilitate instance-level alignments between unpaired images and texts, where semantically similar instances are pulled closer and dissimilar instances are pushed away.

- Extensive experiments on diverse downstream tasks show that our approach achieves superior performance over previous unpaired VLP methods.

2. Related Work

Paired vision-language pre-training. Benefiting from the soaring performance of transformers (Vaswani et al., 2017) on representation learning in both computer vision and natural language processing (Dosovitskiy et al., 2020; Devlin et al., 2019), there is a surging interest in the field of joint pre-training (Tan & Bansal, 2019; Li et al., 2019b; 2020b) of parallel visual and language data. According to the learning objective, prior works can be divided into two categories, single-stream and dual-stream. Single-stream models (Tan & Bansal, 2019; Li et al., 2019b; 2020b; Chen

et al., 2019; Kim et al., 2021) aim to learn the joint representations of two modalities by a cross-modal encoder, which could handle very well the down-stream vision-language tasks with fine-level interactions and reasoning. Dual-stream models (Radford et al., 2021) learn separate representations for each modality by two independent uni-modal encoders, supervised by a constraint on the similarity between representations. It is suitable for downstream tasks requiring coarse-level cross-modal matching (e.g., image-text retrieval) and tasks where a single modality is presented, such as image and text classification.

Unpaired vision-language pre-training. Before the emergence of transformer-based VLP, image encoder and language encoder in traditional methods (Yu et al., 2019) are pre-trained separately based on uni-modal datasets. However, there are no special designs for learning cross-modality alignments during pre-training, indicating that all knowledge about alignment is learned from the fine-tuning stage, where only a few manually-labeled image-text pairs are available. Li et al. (2021b) proposed the task of unpaired VLP, aiming to discover the complex interactions and semantic alignments between modalities in uni-modal pre-training datasets. Their method shares the encoders across modalities, forcing the samples in different modalities to be projected into the same space and thus encouraging alignments. Given an image, it concatenates the image regions together with their detector tags as aligned multi-modal inputs. Given a sentence, it directly considers the uni-modal subwords as input tokens. The pre-training objective is to reconstruct the masked inputs. We argue that, this scheme lacks the interactions between visual regions and linguistic cues (like quantifiers and words indicating relationships between two visual entities), resulting in a gap between pre-training and downstream tasks. Moreover, it lacks the ability to distinguish the alignment degree between visual and text data as no explicit matching supervision exists. Compared with Li et al. (2021b), the most salient difference of VLMixer is that we use a cross-modal augmentation to construct semantic-invariant cross-modal inputs for 1) aligning the multi-modal and uni-modal view of the original sentence by contrastive learning; 2) effectively fusing the visual tokens and non-grounded linguistic tokens. Fig. 2 summarizes mainstream paired and unpaired VLP methods[2].

Unpaired image captioning. Unpaired image captioning focuses on training a useful image-to-text translation model without parallel image-text training data. Similar to unpaired VLP, the key component of this task is the cross-modal alignment. Existing literature designs different types of intermediate signals for connecting two modalities. Gu et al. (2018) explored the pivot language to connect the source image and target language. Feng et al. (2019) explored an adversarial training framework including a concept detector and a sentence discriminator with three types of well-designed adversarial rewards, where concept words serve as the anchor points to bridge and align images and texts. Gu et al. (2019) regarded the scene graph as an intermediate representation of each modality and trained a cycle-consistency adversarial method that maps scene graph features from the image to the text modality.

[2]In this paper, unpaired VLP aims for cross-modal learning given image-only and text-only corpora. We classify some methods which rely on image-label pairs for pre-training a visual backbone into unpaired VLP since they use discrete concept categories instead of semantic-rich natural language with syntactic structure.

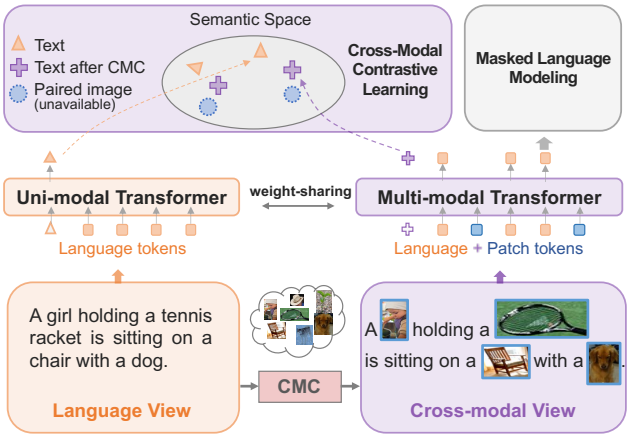

Figure 3. Visually-aided language pre-training. Given a sentence sample, we randomly wipe off some concept words in the sentence and then paste the visual patches with the same concept labels to obtain mixed sentences, serving as the cross-modal view of the original sentence. Two objectives are proposed for cross-modal learning: First, masked language modeling aims to learn the denoising representation, which encourages the token-level alignment between two modalities; Then, cross-modal contrastive learning guides the model to judge the instance-level alignment between the two views. Unlike contrastive learning used in paired VLP methods (Li et al., 2020a; 2021a), paired images are not available in our setting. The proposed contrast between text and text after CMC can be regarded as a proxy task of text-image contrast in paired VLP.

Data augmentation. Data augmentation contains techniques for improving data diversity without collecting more data. It have been widely applied to several modalities, such as images (DeVries & Taylor, 2017; Yun et al., 2019), texts (Wei & Zou, 2019), and audios (Wei & Zou, 2019). Our method is inspired by CutMix (Yun et al., 2019) in vision tasks, which randomly removes image patches by overlaying salient patches from other images. The resultant image serves as an intermediate representation to bridge two images with different semantics. This paper constructs "multi-modal sentences" to bridge the visual and linguistic modality, which could produce diverse multi-modal data without altering the semantics.

3. VLMixer Pre-training

VLMixer contains two parallel pre-training branches, visually-aided language pre-training (VALP) and tag-aided visual pre-training (TAVP). In VALP, given a sentence sampled from the text-only dataset, we adopt cross-modal cutmix (CMC) to obtain a multi-modal view of the sentence and performs two learning objective on it, masked language modeling for reconstructing the masked inputs and contrastive learning for learning cross-modal alignments. In TAVP, given an image sampled from the image-only dataset, we follow (Li et al., 2020b) to consider image tags and detected objects as the inputs for masked tag modeling. In the following, we introduce the cross-modal cutmix in subsection 3.1, the VALP and TAVP branches in subsections 3.2 and 3.3, respectively.

3.1. Cross-Modal CutMix

The inputs in paired VLP (Tan & Bansal, 2019; Li et al., 2020b) share a similar format with downstream tasks for fine-tuning: the mixed multi-modal sequence of both visual tokens and text tokens with consistent semantics. However, unpaired VLP without explicit alignments brings difficulties in constructing such a multi-modal input. Directly combining a text with a random image not only loses the cross-modal alignment but also introduces too much noise, which may overwhelm the interactions between intra-modal tokens. This section proposes cross-modal CutMix to construct diverse multi-modal sequences to mitigate the discrepancy between the pre-training and fine-tuning stages.

Patch gallery. We first collect a visual patch gallery of high-quality object regions with their concept labels from the image-only dataset. To this end, an off-the-shelf concept detector (e.g., Faster RCNN (Ren et al., 2015)) is utilized to detect salient regions x_i and predict their concept labels w_i^{con} and corresponding confidences c_i^{con}. We denote the concept vocabulary as $\mathcal{C}$. Besides concepts of the current object, we also record "contextual concepts", i.e., the concepts of other regions occurred in the same image, denoted as $\{(w_{ij}^{\text{ctx}}, c_{ij}^{\text{ctx}})\}$, where w_{ij}^{ctx} and c_{ij}^{ctx} represents the j-th contextual concept and its confidence score. The visual patches with their concepts are visually-grounded, serving as anchoring points to connect the images and sentences. We denote the patch gallery as:

$$\mathcal{G} = \left\{\left(w_i^{\text{con}}, c_i^{\text{con}}, \left\{\left(w_{ij}^{\text{ctx}}, c_{ij}^{\text{ctx}}\right)\right\}\right)\right\}. \quad (1)$$

CutMix visual patches into sentences. Given a sentence $\mathbf{T} = \{w_n\}_{n=1}^N$ sampled from the text corpus D^T, our goal is to construct a multi-modal sequence while preserving the high-level semantics. For each (sub-)word in the sentence meanwhile appearing in the concept vocabulary $w_n \in \mathcal{C}$, we randomly replace it with a visual patch from the gallery with a probability of r_{cmc}. The target visual patch is sampled from all patches with a concept label of w_n. We note that the sampled patches should not only accurately match the global semantics of the sentences, but also have diverse patterns for enhancing the generalization ability. This drives us to take the influence of the global semantics of the sentence into consideration. We design a context-aware sampling according to the following probability distribution. For a concept (sub-)word w_n in $\mathbf{T}$, we calculate the probability of being chosen of all the items in the patch gallery. We sample a patch x_q with $q \sim \text{Norm}(\{p_i\})$ from the gallery, and p_i is defined as:

$$p_i = \begin{cases} c_i^{\text{con}} + \frac{r_{\text{ctx}}}{|G_i|} \sum_{w_{ij}^{\text{ctx}} \in G_i} c_{ij}^{\text{ctx}}, & \text{if } w_i^{\text{con}} = w_n \\ 0, & \text{otherwise} \end{cases}, \quad (2)$$

where $G_i = \mathbf{T} \cap \{w_{ij}^{\text{ctx}}\}$ represents the intersection between the sentence and contextual concepts of x_i, $\text{Norm}(\cdot)$ normalizes the confidences $\{p_i\}$ as a probability distribution. r_{ctx} controls the importance of contextual concepts for sampling. The resultant sequence $\mathbf{S}$ after CMC can be represented as a mixture of multi-modal tokens, like $\mathbf{S} = \{w_1, x_{q_2}, w_3, x_{q_4}, ..., w_N\}$, where x_{q_i} represents the sampled patch for the i-th (sub-)word.

K-shot CMC. Considering that a single patch only reveals a partial view of the concept (sub-)word, we propose the K-shot CMC which collects diverse patches as multiple views of this concept. Specifically, we replace w_n with a set of patches that may come from different sources, by repeating the sampling process K times. Thus, the resultant multi-modal tokens $\mathbf{S}$ becomes $\{w_1, x_{q_2^{(1)}}, ..., x_{q_2^{(K)}}, w_3, x_{q_4^{(1)}}, ..., x_{q_4^{(K)}}, ..., w_N\}$.

3.2. Visually-Aided Language Pre-training

VALP focuses on cross-modal learning from the text corpus with the assistance of the visual patch gallery. Different from U-VisualBERT (Li et al., 2021b) which only adopts the uni-modal representation learning for text-only data, we construct the multi-modal inputs for effectively exploiting the multi-modal fusion by masked language modeling and cross-modal alignments by contrastive learning. The detailed illustration of VALP is shown in Fig. 3.

The input sentence $\mathbf{T}$ is firstly converted into a sequence of subwords $\{[CLS], w_1, w_2, ..., w_N, [SEP]\}$ by lower-case byte pair encoding (BPE) (Sennrich et al., 2015), where $[CLS]$ and $[SEP]$ denote the start and the end token of the subword sequence, respectively. We use cross-modal CutMix to obtain the cross-modal view $\mathbf{S}$. The representation of each patch token in $\mathbf{S}$ is the regional features produced by the concept detector. Then $\mathbf{S}$ is fed into a transformer encoder (Vaswani et al., 2017) to learn cross-modal interactions by attention layers. The output feature vector of

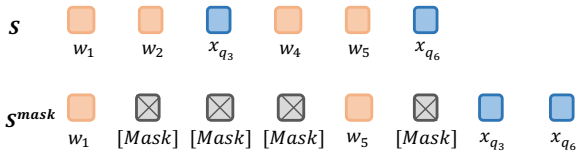

Figure 4. Masking strategy on the cross-modal view.

$[CLS]$ is regarded as the global representation of $\mathbf{S}$.

Masked language modeling (MLM). We use a masking strategy analogy to BERT (Devlin et al., 2019). We randomly mask each language token in $\mathbf{S}$ with a probability of 15%. For each patch token, we add mask tokens into the sequence to indicate the position occurring CMC replacement. These mask tokens gather informative contextual features to recover the corrupted concept word at the same position. We denote the masked input as $\mathbf{S}^{\text{mask}}$. We provide an example in Fig. 4 to illustrate the difference between $\mathbf{S}$ and $\mathbf{S}^{\text{mask}}$.

The goal of MLM is to reconstruct the original text from the two types of corruptions, i.e., cross-modal noise introduced by CMC and corruption from the masking mechanism. Thus, the model could effectively aggregate the contextual information and learn token-level alignments between visual and language tokens. The MLM objective is to minimize the negative log-likelihood of the reconstructed sequence $\hat{\mathbf{S}}$:

$$\mathcal{L}_{\text{mlm}} = -\mathbb{E}_{\mathbf{T}\sim\mathcal{D}^{\text{T}}} \log(\mathbf{T}|\mathbf{S}^{\text{mask}}) = \text{CE}(\hat{\mathbf{S}}, \mathbf{T}), \quad (3)$$

where $\text{CE}(\cdot,\cdot)$ represents the cross-entropy loss.

Cross-modal contrastive learning (CMCL). A common practice for paired VLP is the image-text matching task (Chen et al., 2019; Li et al., 2020b), where positive/negative samples, i.e., paired/unpaired inputs are constructed and the model is trained to distinguish whether the input image and text have similar semantics. Obviously, constructing such positive pairs requires well-aligned data and thus becomes unavailable for unpaired VLP. Despite the difficulties of finding a semantic-similar image for a given text, we propose to construct an intermediate representation by CMC that matches the meaning of the text.

Given a training batch including a random set of the texts, we pair them with their CMC augmentation for contrastive learning, represented as $\{(\mathbf{T}_1, \mathbf{S}_1), \cdots, (\mathbf{T}_M, \mathbf{S}_M)\}$. For the anchor instance $\mathbf{T}_m$, we choose $\mathbf{S}_m$ as the positive instance and the remaining pairs in the batch as negative instances $\{\mathbf{S}_l\}_{l\neq m}$. The contrastive loss is calculated by:

$$\mathcal{L}_{\text{cl}} = -\sum_{m=1}^{M} \log \frac{\exp\left(f\left(\mathbf{T}_m^{\text{mask}}, \mathbf{S}_m^{\text{mask}}\right)/\tau\right)}{\sum_{l=1}^{M} \exp\left(f\left(\mathbf{T}_m^{\text{mask}}, \mathbf{S}_l^{\text{mask}}\right)/\tau\right)}, \quad (4)$$

where $\mathbf{T}_m^{\text{mask}}$ and $\mathbf{S}_l^{\text{mask}}$ are the masked sequences of $\mathbf{T}_m$ and $\mathbf{S}_l$, $f(\mathbf{T}_m^{\text{mask}}, \mathbf{S}_l^{\text{mask}})$ represents the cosine similarity between the output features on the $[CLS]$ tokens for $\mathbf{T}_m^{\text{mask}}$ and $\mathbf{S}_l^{\text{mask}}$. τ is the temperature ratio.

Algorithm 1 Unpaired VLP via CMC

Input: image set D^I, text set D^T.
Output: pre-trained transformer.
Construct a patch gallery $\mathcal{G}$ from D^I
for $iter := 1$ **to** max_iter **do**
  Sample a mini-batch of sentences $\mathbf{T}$ from D^T.
  Sample a mini-batch of images $\mathbf{I}$ from D^I.
  Obtain $\mathbf{S}$ by performing CMC on $\mathbf{T}$.
  Obtain image tokens with tags $\mathbf{Q} = (\mathbf{I}, \texttt{Det}(\mathbf{I}))$.
  Obtain $\mathbf{T}^{\text{mask}}$, $\mathbf{S}^{\text{mask}}$, $\mathbf{Q}^{\text{mask}}$ by random masking.
  $\hat{\mathbf{T}} = \texttt{Transformer}\left(\mathbf{T}^{\text{mask}}\right)$
  $\hat{\mathbf{S}} = \texttt{Transformer}\left(\mathbf{S}^{\text{mask}}\right)$
  $\hat{\mathbf{Q}} = \texttt{Transformer}\left(\mathbf{Q}^{\text{mask}}\right)$
  Compute $\mathcal{L}_{\text{total}}$ with (6) and update model parameters.
end for

Note that our method differs from existing contrastive learning methods (Radford et al., 2021; Li et al., 2021a) for paired VLP for two reasons: 1) The paired image used in their model are unavailable in our setting; 2) The proposed contrast between uni-modal sample and multi-modal sample encourages multi-modal fusion, compared with the contrast between two uni-modal samples.

3.3. Tag-Aided Visual Pre-training

TAVP mainly focuses on the exploitation of multi-modal knowledge from the visual-only data. Inspired by Li et al. (2021b), we use image tags (concepts) as anchor points to connect vision and language, since they are detected from images but also play an important role in language learning. Specifically, given an image $\mathbf{I}$ from the image set $\mathcal{D}^{\text{I}}$, a pre-trained concept detector is utilized to predict a number of image regions and their tags. The region token and the tag token are concatenated as the multi-modal representation of the image $\mathbf{Q} = (\mathbf{I}, \texttt{Det}(\mathbf{I}))$, where $\texttt{Det}(\mathbf{I})$ represents the sequence of tag tokens.

We adopt the mask-then-predict pre-training on image and tag tokens similar to OSCAR (Li et al., 2020b). Each tag token is randomly masked with a probability of 15%. Afterwards, we input the masked input $\mathbf{Q}^{\text{mask}}$ into the transformer to calculate the reconstruction loss:

$$\mathcal{L}_{\text{mtm}} = -\mathbb{E}_{\mathbf{I}\sim\mathcal{D}^{\text{I}}} \log(\mathbf{Q}|\mathbf{Q}^{\text{mask}}) = \text{CE}(\hat{\mathbf{Q}}, \mathbf{Q}). \quad (5)$$

3.4. Training Objective

The overall training objective is defined as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{cl}} + \mathcal{L}_{\text{mtm}}, \quad (6)$$

which is the summation of the masked language modeling loss, contrastive loss, and the masked tag modeling loss. At each iteration, we sample a mini-batch of images and a mini-batch of texts for loss calculation. The detailed pre-training process is summarized in Algorithm 1.

4. Experiments

For fair comparisons, we first follow the standard practice in unpaired vision-language (VL) tasks (Li et al., 2021b; Feng et al., 2019) to evaluate the model performance on the paired VL datasets without the alignment information. Next, we show that VLMixer could benefit from large-scale images or texts collected independently from different sources. Finally, we conduct ablation studies on important design choices to show the effectiveness of VLMixer.

4.1. Datasets

We use a variety of datasets covering diverse visual and language patterns. Specifically, three kinds of pre-training datasets are taken into account: image-text pairs, image-only collections and text-only corpora. The paired VL datasets contain COCO Captions (Lin et al., 2014), Visual Genome (Krishna et al., 2017b), Conceptual Captions 3M (Sharma et al., 2018), SBU Captions (Ordonez et al., 2011), Flickr30K (Plummer et al., 2015), and GQA (Hudson & Manning, 2019), totally 4.1M images and 5.6M captions. For additional image-only data, we use Open Images(Kuznetsova et al., 2018) containing 1.7M images. The text-only corpus comes from three sources: 1) human-annotated captions from existing video captioning datasets, i.e., MSVD (Chen & Dolan, 2011), MSRVTT (Xu et al., 2016), VATEX (Wang et al., 2019), and ActivityNet Captions (Krishna et al., 2017a); 2) Auto-crawled captions from a online stock photography website Shutterstock, provided by Feng et al. (2019); 3) General text segments from BookCorpus (Zhu et al., 2015). The total size of text-only instances is around 15.5M. Detailed statistics of the pre-training corpus are provided in Table 1.

Dataset	Images	Texts	Text Domain
COCO (train)	112K	560K	Image Caption
Conceptual Captions (train)	3M	3M	Image Caption
SBU Caption (all)	840K	840K	Image Caption
Flickr30k (train)	29K	145K	Image Caption
VQA (train)	83K	445K	Question
GQA (train)	79K	1.0M	Question
VG-QA (train)	87K	931K	Question
MSVD (train)	-	48K	Video Caption
MSRVTT (train)	-	130K	Video Caption
VATEX (train)	-	260K	Video Caption
ActivityNet Captions (train)	-	36K	Video Caption
Shutterstock (all)	-	1M	Caption
BookCorpus	-	14M	General Text
OpenImages (od train)	1.67M	-	-

Table 1. Pre-training dataset statistics. We use several large-scale image and text datasets with diverse language patterns.

4.2. Experimental Setting

Implementation details. We use a Base Transformer with 12 layers of transformer block and a hidden size of 768 as the backbone. For position embedding, we adopt the learnable position embedding for language/tag tokens and use a linear projection of spatial positions for patch/image tokens. To reduce the computation cost, we restrict the max token length in TAVP and VALP to 100 and 80, respectively. For VALP, we first collect a subset of object regions within the image dataset as the patch gallery by filtering regions with high confidence. The object regions are detected by an off-the-shelf concept detector ResNeXt-152 C4 provided by Zhang et al. (2021). The size of the concept vocabulary is 1600. For each sentence, we adopt K-shot CMC to enhance the diversity of sampled patch tokens with K=15. The replacing probability r_{cmc} in CMC is set to 0.5 and the context weight r_{ctx} is set to 0.5. Note that we followed the standard practices in unpaired VL tasks to prevent selecting patches in the paired image. To reduce the noisy data, we drop the sentences shorter than five words as there is a high probability that it does not contain concept words. The temperature ratio τ in CMCL is set to 0.1.

Pre-training and fine-tuning. We initialize VLMixer from the parameters of $\text{BERT}_{\text{base}}$, and pre-train the model on unpaired image and text data for a maximum of 300k steps. An Adam optimizer is adopted with an initial learning rate of 5e-5 and a mini-batch size of 1024. The warm-up rate is set to 10%. Mixed precision training is used to accelerate the training stage. The training time on the full pre-training data is around six days on 16 Telsa A100 GPUs.

After pre-training, we adapt the weights of VLMixer to five downstream tasks, i.e., VQA (Goyal et al., 2017), NLVR^2 (Suhr et al., 2018), image retrieval (COCO 5K), text retrieval (COCO 5K), and GQA (Hudson & Manning, 2019). We follow the fine-tuning strategy and evaluation metrics in Zhang et al. (2021) for downstream tasks.

4.3. Comparison with State-of-the-Art Methods

We compared VLMixer with the following methods in the setting of unpaired pre-training: 1) **U-VisualBERT** is a pioneer work in unpaired VLP. It uses a parallel pre-training scheme for each modality and utilizes tags as anchor points to connect images and texts. 2) $\textbf{VinVL}_{\text{unpaired}}$: We modified the paired VLP method VinVL to fit the unpaired setting

Method	Pre-training Data		VQA	NLVR^2		Text Retrieval			Image Retrieval			GQA
	Image	Text	Test-Dev	Dev	Test	R@1	R@5	R@10	R@1	R@5	R@10	Test-Dev
Paired VLP												
UnicoderVL$_{\text{base}}$ (Li et al., 2019a)			-	-	-	62.3	87.1	92.8	46.7	76.0	85.3	-
UNITER$_{\text{base}}$ (Chen et al., 2019)			72.27	77.14	77.87	63.3	87.0	93.1	48.4	76.7	85.9	-
OSCAR$_{\text{base}}$ (Li et al., 2020b)			73.16	78.07	78.36	70.0	91.1	95.5	54.0	80.8	88.5	61.58
VILT$_{\text{base}}$ (Kim et al., 2021)			71.26	75.70	76.13	61.5	86.3	92.7	42.7	72.9	83.1	-
VinVL$_{\text{base}}$ (Zhang et al., 2021)			75.95	82.05	83.08	74.6	92.6	96.3	58.1	83.2	90.1	65.05
ALBEF (Li et al., 2021a)			75.84	82.55	83.14	77.6	94.3	97.2	60.7	84.3	90.5	-
Unpaired VLP												
BERT$_{\text{base}}$ (Devlin et al., 2019)	None	None	64.85	51.30	51.34	57.44	84.00	91.58	44.03	74.12	84.06	50.20
VinVL$_{\text{unpaired}}$ (Zhang et al., 2021)	COCO	COCO	71.78	71.14	72.01	61.92	86.90	93.08	46.90	76.18	85.53	62.24
U-VisualBERT (Li et al., 2021b)*	COCO	COCO	72.41	-	-	-	-	-	-	-	-	-
VLMixer	COCO	COCO	**72.60**	**72.71**	**73.08**	**62.69**	**87.35**	**93.64**	**47.95**	**77.06**	**86.22**	**63.13**
U-VisualBERT (Li et al., 2021b)	CC3M	CC3M+BC	70.74	71.74	71.02	-	-	-	-	-	-	-
VinVL$_{\text{unpaired}}$ (Zhang et al., 2021)	CC3M	CC3M	72.20	68.96	68.94	62.08	86.04	93.00	47.29	76.15	85.53	63.12
VLMixer	CC3M	CC3M	72.66	74.31	73.86	62.20	86.32	92.80	47.44	76.22	85.41	62.65
VLMixer	Full	Full	**72.89**	**76.61**	**77.01**	**64.76**	**88.56**	**94.22**	**50.06**	**78.36**	**86.91**	**63.25**

Table 2. Comparison with state-of-the-art unpaired VLP methods. We report the performance on COCO and CC3M for a fair comparison with previous state-of-the-art methods. "Full" data means we leverage all image data and text data introduced in subsection 4.1. "CC3M" and "BC" denote the conceptual captions 3M and the BookCorpus datasets. * denotes the results of our re-trained model with the VinVL object features. We also list the performance of paired VLP methods for reference.

by simply considering a text with randomly sampled images as inputs for mask-then-predict learning. The image-text matching loss is disabled. 3) **BERT**$_{\text{base}}$ is the standard BERT base model pre-trained on text datasets.

The performance comparison is shown in Table 2. We test the performance of VLMixer based on pre-training corpora with three scales, COCO, CC3M, and the full corpus. Our method achieves a better performance on most downstream tasks than other methods under a similar size of pre-training data. Compared with paired VLP, we achieve comparable performance with UNITER$_{\text{base}}$ and VILT$_{\text{base}}$, showing that pre-training on large-scale easy-to-collect unpaired data has great potential to benefit the vision-language tasks.

As images and captions in COCO/CC3M datasets come from the same source, it is natural to ask what if image and text sets are not fully aligned. Then we conduct pre-training on full corpora, which contains rich images and diverse language patterns that are collected from different sources. The language data contains image caption, video caption, question, and general text, while the image data are usually for common image recognition. The superior performance of VLMixer on the full pre-training data shows that our model could effectively learn useful cross-modal interactions from large-scale images or texts independently collected from different sources.

4.4. Ablation Studies

Main results. The ablation study of the proposed method is shown in Table 3. "MLM+CMC" achieves a considerable performance-boosting over "MLM", which means the introduction of patch tokens significantly boosts the learning of cross-modal interactions. We also notice that "MLM+CMC" performs better than TAVP, which shows the importance of the syntactic information introduced by multi-modal sentences. When incorporating tag-aided visual pre-training (TAVP), the overall performance could obtain further improvement. It is an interesting phenomenon that our best model, which uses CMC and CMCL together, could achieve better performance than the paired pre-training model on VQA. We conjecture that the reason may lie in that mixing the language tokens with patch tokens could largely increase the data diversity, which could benefit the model to achieve good generalization performance.

Cross-modal CutMix. The number of shared concepts between patch gallery and text corpus matters, since it reflects the global alignment between an image dataset and a text dataset. In Fig. 5, we test the influence of the number of shared concepts by constructing six subsets of COCO Captions with an increasing number of concepts. We see that with the increase of concept number, the NLVR^2 performance gradually improves. The reason for the slight decrease at 1600 concepts may lie in too many concepts

VALP			TAVP	VQA	NLVR^2		Text Retrieval			Image Retrieval		
MLM	CMC	CMCL		Test-Dev	Dev	Test	R@1	R@5	R@10	R@1	R@5	R@10
			√	71.16	70.52	69.23	60.18	85.50	91.72	45.87	75.39	84.96
√				71.50	50.89	52.16	49.32	78.02	87.72	38.04	69.62	80.92
√			√	72.00	72.52	72.20	59.30	85.36	91.76	45.78	74.94	84.60
√	√			71.52	71.13	70.99	60.40	85.72	92.92	46.92	75.86	85.31
√	√		√	71.84	**73.19**	72.81	60.54	86.24	92.44	47.29	76.43	85.61
√	√	√	√	**72.60**$_{\pm 0.10}$	72.71$_{\pm 0.61}$	**73.08**$_{\pm 0.26}$	**62.69**$_{\pm 0.51}$	**87.35**$_{\pm 0.19}$	**93.64**$_{\pm 0.14}$	**47.95**$_{\pm 0.21}$	**77.06**$_{\pm 0.13}$	**86.22**$_{\pm 0.08}$
Paired Pre-training				72.39	75.28	75.54	65.10	88.82	94.38	50.23	78.49	87.13

Table 3. Ablation studies of pre-training objectives. All models are pre-trained on COCO without alignment information except in the last row. For paired pre-training, we feed the concatenation of image, tag, and language tokens into the transformer and use image-text matching loss with masked token modeling loss as the training objectives, as in (Li et al., 2020b). For the final model, we run three times to report the mean and standard deviation.

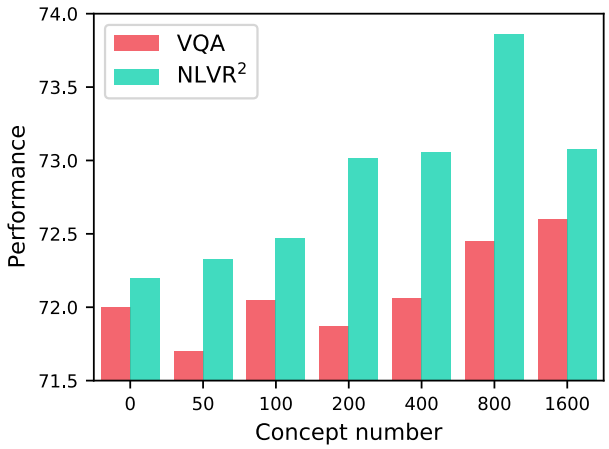

Figure 5. The downstream performance using different number of concepts in the patch gallery. We modulate the number of shared concepts by controlling the concept number of the patch gallery.

introducing inevitable more misrecognition, especially for long-tailed concept classes. We conclude that the model could benefit from an increasing number of shared concepts, showing that VLMixer effectively exploits the multi-modal interactions and makes better usage of the concepts.

We also verify the effectiveness of two techniques, K-shot CMC and context-aware sampling. For K-shot CMC, a larger K performs better. A small K causes an imbalance between language and patch tokens, resulting in pre-training being dominated by language tokens, which in turn restricts cross-modal learning. For context-aware sampling, combining confidences from concept tokens and context tokens performs better than removing the context tokens, indicating that context tokens are informative for accurate sampling.

Contrastive learning. To effectively learn the cross-modal alignment, we propose the contrast between uni-modal textual sentences and the multi-modal sentences after CMC. We compare this design with the previous data augmentation method used in contrastive learning. Specifically, we implement two text data augmentation methods: 1) Crop $k\%$. We randomly crop the sentence and keep a continuous segment with a length of 100-$k\%$. 2) Delete $k\%$. We randomly remove $k\%$ words from the sentence. The performance of the ablated methods is shown in Table 5. Compared with the model without contrastive learning, our method could improve both VQA and NLVR^2, while text-text contrast can not achieve consistent improvement on two tasks. Our method is superior to text-text contrast. The reason may be that our method encourages the cross-modal fusion of inputs where two modalities are semantically consistent, and discourages that where two modalities are semantically incompatible.

Method	VQA	NLVR^2	
	Test-Dev	Dev	Test
w/o K-shot CMC (K=1)	71.88	71.77	72.25
w/o context-aware sampling (r_{ctx}=0)	71.86	72.07	72.48
VLMixer	**72.60**	**72.71**	**73.08**

Table 4. Ablation of Cross-modal CutMix.

5. Conclusions

This paper presents a new method named VLMixer for unpaired vision language pre-training. Different from traditional methods that use tags as the anchor to bridge the two modalities, we propose to construct the cross-modal view of the textual sentences by cross-modal CutMix. By doing so, the diversity of multi-modal data could be increased to a large extent without altering the semantics. Furthermore, to enable better alignment learning at the instance level, we build the contrastive learning objective on multi-modal sentences to pull together semantically similar instances and

Method	VQA	NLVR^2	
	Test-Dev	Dev	Test
Cross-modal contrast	72.60	72.71	73.08
w/o contrastive learning	72.00	72.52	72.20
Text-text contrast: crop 10%	72.04	71.82	72.74
Text-text contrast: crop 20%	71.79	72.97	72.15
Text-text contrast: crop 30%	72.52	70.15	70.17
Text-text contrast: delete 10%	72.70	71.54	71.42
Text-text contrast: delete 20%	71.71	71.60	71.51
Text-text contrast: delete 30%	71.77	72.31	72.27

Table 5. Ablation study of the contrastive learning methods and data augmentations. All models are pre-trained on COCO.

push away semantically dissimilar instances. Experiments on five downstream tasks show that our method achieves state-of-the-art performance in unpaired VLP. The ablation studies of the design choices in CMC and contrastive learning verify the effectiveness of the proposed model.

6. Acknowledgement

This work is supported by the National Natural Science Foundation of China under Grant No. 61972188, No. 62122035, No. 61906081, No. 62106097, and the China Postdoctoral Science Foundation (2021M691424). Ping Luo is supported by the General Research Fund of HK No. 27208720 and 17212120.

References

Cao, J., Gan, Z., Cheng, Y., Yu, L., Chen, Y.-C., and Liu, J. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision*, pp. 565–580. Springer, 2020.

Chen, D. and Dolan, W. B. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 190–200, 2011.

Chen, Y.-C., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019.

DeVries, T. and Taylor, G. W. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

Feng, Y., Ma, L., Liu, W., and Luo, J. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4125–4134, 2019.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.

Gu, J., Joty, S., Cai, J., and Wang, G. Unpaired image captioning by language pivoting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 503–519, 2018.

Gu, J., Joty, S., Cai, J., Zhao, H., Yang, X., and Wang, G. Unpaired image captioning via scene graph alignments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10323–10332, 2019.

Hu, X., Yin, X., Lin, K., Wang, L., Zhang, L., Gao, J., and Liu, Z. Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training. *arXiv preprint arXiv:2009.13682*, 2020.

Hudson, D. A. and Manning, C. D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

Kim, W., Son, B., and Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*, 2021.

Krishna, R., Hata, K., Ren, F., Fei-Fei, L., and Carlos Niebles, J. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pp. 706–715, 2017a.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017b.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018.

Li, G., Duan, N., Fang, Y., Jiang, D., and Zhou, M. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. *arXiv preprint arXiv:1908.06066*, 2019a.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021a.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019b.

Li, L. H., You, H., Wang, Z., Zareian, A., Chang, S.-F., and Chang, K.-W. Unsupervised vision-and-language pre-training without parallel images and captions. *NAACL*, 2021b.

Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., and Wang, H. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020a.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., and Gao, J. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pp. 121–137. Springer, 2020b.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

Lu, J., Batra, D., Parikh, D., and Lee, S. VilBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

Ordonez, V., Kulkarni, G., and Berg, T. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24:1143–1151, 2011.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*, 2018.

Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., and Artzi, Y. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

Tan, H. and Bansal, M. LXMERT: Learning cross-modality encoder representations from transformers. *EMNLP*, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *NeurIPS*, 2017.

Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.-F., and Wang, W. Y. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4581–4591, 2019.

Wei, J. and Zou, K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

Xu, J., Mei, T., Yao, T., and Rui, Y. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

Yu, Z., Yu, J., Cui, Y., Tao, D., and Tian, Q. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6281–6290, 2019.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032, 2019.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5579–5588, 2021.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pp. 19–27, 2015.