UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning

Wei Li[*], **Can Gao**[*], **Guocheng Niu**[*], **Xinyan Xiao**[*],
Hao Liu, Jiachen Liu, Hua Wu, Haifeng Wang
Baidu Inc., Beijing, China
{liwei85,gaocan01,niuguocheng,xiaoxinyan,
liuhao24,liujiachen,wu_hua,wanghaifeng}@baidu.com

Abstract

Existed pre-training methods either focus on single-modal tasks or multi-modal tasks, and cannot effectively adapt to each other. They can only utilize single-modal data (i.e. text or image) or limited multi-modal data (i.e. image-text pairs). In this work, we propose a unified-modal pre-training architecture, namely UNIMO, which can effectively adapt to both single-modal and multi-modal understanding and generation tasks. Large scale of free text corpus and image collections can be utilized to improve the capability of visual and textual understanding, and cross-modal contrastive learning (CMCL) is leveraged to align the textual and visual information into a unified semantic space over a corpus of image-text pairs. As the non-paired single-modal data is very rich, our model can utilize much larger scale of data to learn more generalizable representations. Moreover, the textual knowledge and visual knowledge can enhance each other in the unified semantic space. The experimental results show that UNIMO significantly improves the performance of several single-modal and multi-modal downstream tasks.

1 Introduction

Large-scale pre-training has draw much attention in both the community of Compute Vision (CV) and Natural Language Processing (NLP) due to its strong capability of generalization and efficient usage of large-scale data. Firstly in CV, a series of models were designed and pre-trained on the large-scale dataset ImageNet, such as AlexNet (Krizhevsky et al., 2017), VGG (Simonyan and Zisserman, 2014) and ResNet (He et al., 2016), which effectively improved the capability of image recognition for numerous tasks. Recent years have witnessed the burst of pre-training in NLP, such as

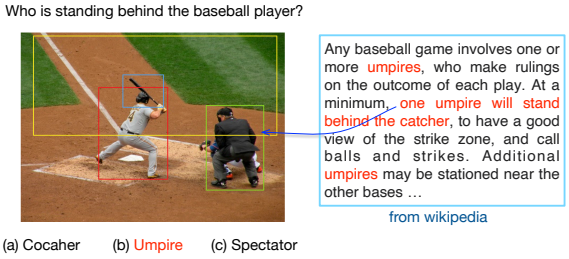

Figure 1: An illustrative example for the necessary of the unified-modal learning. We can only determine the correct answer of the visual question based on the textual background information.

BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019) and BART (Lewis et al., 2019), which greatly improve the capability of language understanding and generation. However, the above researches towards the single-modal learning and can only be used in single-modal (i.e. only text or image) scenarios. In order to adapt to multi-modal scenarios, a series of multi-modal pre-training methods were proposed and pre-trained on the corpus of image-text pairs, such as ViLBERT (Lu et al., 2019), VisualBERT (Li et al., 2019b) and UNITER (Chen et al., 2020), which greatly improve the ability to process multi-modal information. However, these models can only utilize the limited corpus of image-text pairs and cannot be effectively adapted to single-modal scenarios (Lin et al., 2020).

A smarter AI system should be able to process different modalities of information effectively. There are large scale of data in different modalities on the Web, mainly textual and visual information. The textual knowledge and the visual knowledge usually can enhance and complement with each other. As the example shown in Figure 1, it's difficult to answer the question correctly only with the visual information in the image. However, if we

[*] These authors contribute equally to this study and are listed with random order.

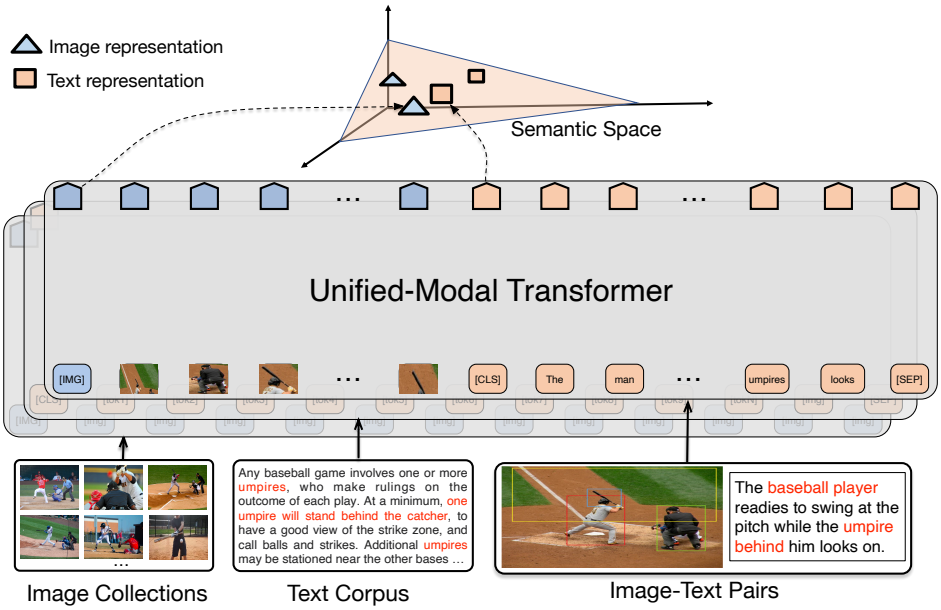

Figure 2: Illustration of the unified-modal pre-training architecture. Both image collections and text corpus can be effectively utilized for representation learning. The visual and textual information are aligned and unified into the same semantic space via CMCL on a corpus of image-text pairs.

connect the visual information to the textual information which describes the background of a baseball game, it's very easy to determine the correct answer. Also, the visual information can make it easier to understand the scene described by the text. The research in neuroscience by Van Ackeren et al. (2018) reveals that the parts of the human brain responsible for vision can learn to process other kinds of information, including touch and sound. Inspired by the research, we propose to design a unified-modal architecture UNIMO which can process multi-scene and multi-modal data input, including textual, visual and vision-and-language data, as shown in Figure 2.

The greatest challenge to unify different modalities is to align and unify them into the same semantic space which are generalizable to different modalities of data. Existed cross-modal pre-training methods try to learn cross-modal representations based on only limited image-text pairs by simple image-text matching and masked language modeling (Chen et al., 2020). They can only learn specific representations for image-text pairs, which are not generalizable for single-modal scenarios. So their performance will drop dramatically when applied to language tasks (Lin et al., 2020). In this work, UNIMO learns visual representations and textual representations in similar ways, and unify them into the same semantic space via cross-modal contrastive learning (CMCL) based on a large-scale corpus of image collections, text corpus and image-text pairs.

UNIMO effectively utilizes the large-scale of text corpus and image collections to learn general textual and visual representations. The CMCL aligns the visual representation and textual representation, and unifies them into the same semantic space based on image-text pairs. To facilitate different levels of semantic alignment between vision and language, we propose to utilize a series of text rewriting techniques to improve the diversity of cross-modal information. As shown in Figure 3, we utilize back-translation to generate several positive examples for an image-text pair. Also, to enhance the detail semantic alignment between text and image, we further parse the caption to scene graph (Wang et al., 2018) and randomly replace either the objects, attributes or relations in the caption to generate various negative samples. Sentence-level retrieval and replacement is also utilized to enhance the sentence-level alignment. In this way, our model can effectively unify different levels of visual and textual representations into the same semantic space.

The unified-modal architecture mainly has the following advantages compared with previous methods:

- We can utilize large scale of non-paired text corpus and image collections on the Web to

learn more generalizable textual and visual representations, and improve the capability of vision and language understanding and generation.

- Our model can be effectively fine-tuned for both single-modal and multi-modal understanding and generation downstream tasks.
- The visual knowledge and textual knowledge can enhance each other to achieve better performance on several single-modal and multi-modal tasks than previous methods.

2 Related Work

Existing researches on pre-training can be mainly classified into two categories: single-modal pre-training and multi-modal pre-training. The single-modal pre-training methods only focus on single-modal tasks, while the multi-modal pre-training methods only focus on multi-modal tasks.

Single-Modal Pre-training The single-modal pre-training methods mainly consists of visual pre-training methods and language pre-training methods. Most visual pre-training methods are based on the multi-layer CNN architecture such as VGG (Simonyan and Zisserman, 2014) and ResNet (He et al., 2016), and trained on the ImageNet dataset. These pre-trained models only focus on visual tasks (e.g. image classification etc.), however, they cannot be used in textual or multi-modal (i.e. with both text and image) tasks. The language pre-training methods are also more and more popular in NLP models, which are based on the multi-layer Transformer architecture, such as GPT (Radford et al.), BERT (Devlin et al., 2018), XLNET (Yang et al., 2019) and BART (Lewis et al., 2019). All of them are trained on large-scale corpus by language modeling, which learn contextualized token representations by either predicting tokens based on their context for language understanding or predicting tokens auto-regressively for language generation. However, they can only be used on textual tasks. They cannot deal with multi-modal tasks with both image and text, such as visual question answering (VQA), image-text retrieval and image captioning.

Multi-Modal Pre-training Recently, multi-modal pre-training methods have been more and more popular for solving the multi-modal tasks. All of them are trained on a corpus of image-text pairs, such as ViLBERT (Lu et al., 2019), VisualBERT (Li et al., 2019b), VL-BERT (Su et al., 2019), Unicoder-VL (Li et al., 2019a) and UNITER (Chen et al., 2020). Based on the multi-layer Transformer network, they all employ the BERT-like objectives to learn multi-modal representations from a concatenated-sequence of vision features and language embeddings. Their architectures can be mainly classified into two categories: single-stream and two-stream. The two-stream methods, such as ViLBERT, utilize two single-modal Transformer to process visual features and language embeddings respectively, and then learn their interactions based on a cross-modal Transformer. The single-stream methods directly utilize a single Transformer network to model both the visual features and the language embeddings. VisualBERT, VL-BERT, Unicoder-VL and UNITER all utilize the single-stream architecture, which validate that fusing cross-modal information early and freely by a single-stream network can achieve better performance. All existed multi-modal pre-training methods only focus on multi-modal tasks with both vision and language inputs. However, they cannot be effectively adapted to single-modal tasks. Their performance will drop dramatically when fine-tuned on language tasks (Lin et al., 2020). Moreover, they can only utilize the limited corpus of image-text pairs. By contrast, our unified-modal pre-training method UNIMO can employ large volumes of text corpus and image collections to enhance each other, and can be effectively adapted to both textual and multi-modal scenarios. UNIMO also achieves the best performance on multi-modal tasks including VQA, image caption and visual entailment.

3 UNIMO

Humans perceive the world through many modalities, such as sound, vision and language. Even though any individual modality might be incomplete or noisy, important information are still perceivable since they tend to be shared or enhance each other. With this motivation, we propose a unified-modal pre-training method UNIMO to learn representations that capture modality-invariant information at the semantic level. Different from previous methods, UNIMO can learn from different modalities of data, including images, texts and image-text pairs, thus achieving more robust and generalizable representations for both

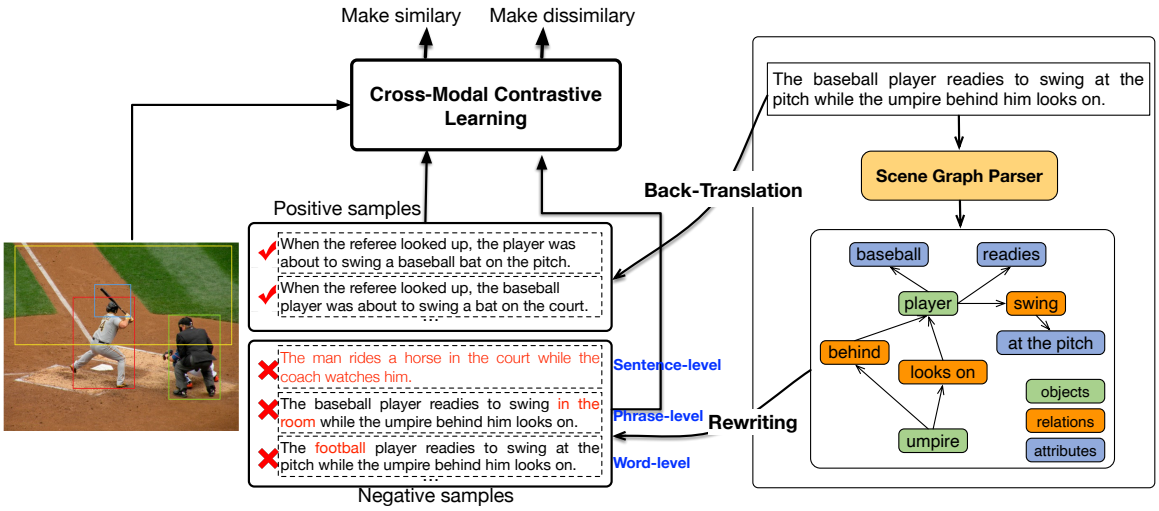

Figure 3: Illustration of the CMCL. Both back-translation and text rewriting based on scene graph are utilized to create positive and hard negative examples.

textual and visual input.

As shown in Figure 2, UNIMO employs multi-layer self-attention Transformers to learn unified semantic representations for both textual and visual data. For a textual input W, it is first split into a sequence of subwords $W = \{[CLS], w_1, ..., w_n, [SEP]\}$ by Word-Pieces, and then the self-attention mechanism is leveraged to learn contextual token representations $\{h_{[CLS]}, h_{w_1}, ..., h_{w_n}, h_{[SEP]}\}$. The special tokens $[CLS]$ and $[SEP]$ denote the start and end of the textual sequence, respectively. Similarly, for an image V, it is first converted to a sequence of region features $V = \{[IMG], v_1, ..., v_t\}$ ($[IMG]$ denotes the representation of the entire image), and then the self-attention mechanism is leveraged to learn contextual region representations $\{h_{[IMG]}, h_{v_1}, ..., h_{v_t}\}$. Similar to previous work (Chen et al., 2020), we use Faster R-CNN (Ren et al., 2016) to detect the salient image regions and extract the visual features (pooled ROI features) for each region.

Based on large volumes of image collections $\{V\}$ and text corpus $\{W\}$, UNIMO learns generalizable visual and textual representations in similar ways by masked prediction. The visual and textual representations are unified into the same semantic space via CMCL based on image-text pairs $\{(V, W)\}$. To effectively unify different levels of visual and textual representations, a series of novel text rewriting techniques are utilized to enhance the CMCL process. Joint visual learning on image collections, language learning on text corpus and cross-modal learning on image-text pairs not only improve the capability of visual and language understanding and generation, but also enable the textual knowledge and language knowledge enhance each other in the unified semantic space.

3.1 Cross-Modal Contrastive Learning

The greatest challenge to unify different modalities is to align and unify their representations at different levels. For the example shown in Figure 2, the model not only needs to connect the scene shown in the whole image to an article describing a baseball game, but also needs to align the two men and their location relationship in the image with "baseball player", "umpire" and "behind" in the text, respectively. Several existing cross-modal pre-training methods try to align visual and textual representations by simply image-text matching (Li et al., 2019a; Chen et al., 2020) based on a limited corpus of image-text pairs. They randomly sample a negative image or text from the same training batch for each image-text pair, and utilize a classifier to determine whether the image and text are matching. As the randomly sampled negative text or image are usually very different from the origin text or image, they can only learn very coarse alignment between textual and visual representations. In this work, we propose a novel CMCL method to align and unify different levels of textual and visual representations into the same semantic space.

For an image-text pair (V, W), its visual features and textual tokens are concatenated as a sequence $\{[IMG], v_1, ..., v_t, [CLS], w_1, ..., w_n, [SEP]\}$.

Then the sequence is feed into the multi-layer Transformer network to learn cross-modal contextual representations for both the textual tokens and image regions. We extract the representation of $[IMG]$ and $[CLS]$ as the semantic representations of the whole image V and text W, respectively. Then the representations of the image and text are transformed into the same embedding space by an FC layer and cosine similarity between them is calculated to measure their distance $d(V, W)$. The main idea is to let the representations of the paired image and text near in the representation space while the non-paired far away. To facilitate semantic alignment between vision and language at different levels, we design several novel text rewriting techniques to rewrite the origin caption of an image either from word, phrase or sentence level, as shown in Figure 3. In this way, we can create large volumes of positive and negative examples from existed image-text pairs. Based on these positive and negative image-text pairs, the following contrastive loss is utilized to learn detailed semantic alignments across vision and language:

$$\mathcal{L}_{CMCL} = -log\frac{exp(d(V, W^+)/\tau)}{\sum_{W' \in \{W^+, W'\}} exp(d(V, W')/\tau)} \quad (1)$$

where W^+ and W^- denotes positive and negative examples for image V, respectively. τ denotes the temperature parameter.

Text Rewriting. To enhance multi-granularity of semantic alignment between image and text, we rewrite the caption of an image at different levels, including sentence-level, phrase-level and word-level. For sentence-level rewriting, we utilize the back-translation techniques (Edunov et al., 2018) to obtain several positive samples for each image-text pair. Specifically, each caption of an image is translated into another language and then translated back to the origin language. In this way, several similar captions can be obtained for an image. Furthermore, for each image-text pair, the most similar captions of other images are retrieved based on TF-IDF similarity. The retrieved results are very similar with the origin caption but doesn't accurately describe the corresponding image, so they can be used as hard negative samples to enhance the sentence-level alignment between image and text.

For phrase-level and word-level rewriting, we first parse the image caption into scene graph (Wang et al., 2018), then randomly replacing the object, attribute or relation nodes of the scene graph with a different object, attribute or relation from the corresponding vocabularies. As shown in Figure 3, the image caption can be parsed into a scene graph, then the objects ("umpire", "player"), attributes ("baseball", "at the pitch", "redies") and relations ("behind", "looks on", "swing') can be extracted from the scene graph and randomly replaced with other objects, attributes or relations. Instead of randomly sampling negative samples as previous methods, text rewriting can generate large volumes of hard negative samples. In this way, we can help the model to learn more detailed semantic alignment from different levels between image and text. To the best of our knowledge, this is the first work to explore CMCL to unify visual and textual semantic space.

Besides cross-modal learning on image-text pairs, UNIMO also learns from large scale corpus of image collections via visual learning and text corpora via language learning.

3.2 Visual Learning

In order to leverage the self-attention mechanism to learn contextual visual representations, UNIMO takes the visual regions of images as inputs. As the self-attention mechanism in Transformer is order-less, we also encode the location features of each region vis a 5-dimensional vector $(\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H}, \frac{(y_2-y_1)(x_2-x_1)}{WH})$, where (x_1, y_1) and (x_2, y_2) denote the coordinates of the bottom-left and top-right corner of the region while W and H denote the width and height of the input image. Both the visual and location features are fused to obtain the final sequence of region embeddings $V = \{[IMG], v_1, ..., v_t\}$. Finally, the region sequence is fed into the multi-layer Transformer to learn contextualized visual representations $\{h_{[IMG]}, h_{v_1}, ..., h_{v_t}\}$.

Similar to the masked language modeling in BERT, we also sample image regions and mask their visual features with a probability of 15%. The visual features of the masked region are replaced by zeros. As the regions from an image usually are highly overlapped with each other, we choose to mask all regions that have a high proportion of mutual intersection to avoid information leakage. Similar to Lin et al. (2020), we randomly choose regions as masking anchors and mask the regions whose IoUs with the anchors are larger than 0.3.

The model is trained to reconstruct the masked regions v_m given the remaining regions $v_{\setminus m}$:

$$\mathcal{L}_V = \mathbb{E}_{v \in D} f_\theta(v_m | v_{\setminus m}) \quad (2)$$

As the visual features are high-dimensional and continuous, we utilize both feature regression and region classification objective to learn better visual representations. The feature regression learns to regress the contextualized visual representations h_{v_i} to its visual features v_i, which can be formulated as: $f_\theta(v_m|v_{\setminus m}) = \sum_{i=1}^{M} \|r(h_{v_i}) - v_i\|^2$, where r indicates an FC layer to convert h_{v_i} into a vector of the same dimension as v_i. The region classification learns to recognize the object semantic class of each masked region based on its contextualized visual representation h_{v_i}. An FC layer is utilized to compute the scores for K object classes $s(h_{v_i})$, which further goes through a $softmax$ function to obtain the normalized distribution. The final objective minimizes the cross-entropy (CE) loss between the predicted distribution and the object detection output $c(v_i)$ from Faster R-CNN: $f_\theta(v_m|v_{\setminus m}) = \sum_{i=1}^{M} CE(softmax(s(h_{v_i})), c(v_i))$.

3.3 Language Learning

To learn general language representations for both language understanding and generation tasks, our model is trained as a unified encoder-decoder model with two types of language modeling tasks: bidirectional prediction and sequence-to-sequence (Seq2Seq) generation. The unified modeling is achieved by utilizing specific self-attention masks to control what context the prediction conditions on, inspired by Dong et al. (2019). To improve the language learning process, we firstly detect semantic complete phrases from text, such as name entities by syntactic parsing, and then treat them as a whole in the following masking strategies. Different from previous work, we always sample a sequence of complete words or phrases instead of subword tokens, for both bidirectional prediction and Seq2Seq generation.

Bidirectional prediction. Given a sequence of tokens $W = \{[CLS], w_1, ..., w_n, [SEP]\}$, we iteratively sampling spans of text until totally 15% tokens have been selected. We sample the span length from a geometric distribution $l \sim Geo(p)$, where p is set as 0.2, similar to SpanBERT (Joshi et al., 2020). All tokens in the selected spans are replaced with either a special $[MASK]$ token, a random token or the origin token with probability 80%, 10% and 10%, respectively. The goal is to predict these masked tokens w_m based on their surrounding context $w_{\setminus m}$, by minimizing the negative log-likelihood:

$$\mathcal{L}_{Bidirectional} = -\mathbb{E}_{w \in D} log P_\theta(w_m | w_{\setminus m}) \quad (3)$$

Seq2Seq generation. For the Seq2Seq generation task, we iteratively sample fragments from the token sequence until the 25% budget has been spent. For each iterate, we first sample a fragment length from a uniform distribution $l \sim U(4, 32)$, and then sample a fragment with the specified length. Every selected fragment $\{w_i, ..., w_j\}$ is further appended with two special tokens $[CLS]$ and $[SEP]$ (i.e. $\{[CLS], w_i, ..., w_j, [SEP]\}$), which denotes the begin and end of the fragment. All selected fragments are removed from the text and concatenated as the target sequence T while the remaining parts are concatenated as the source sequence S. The model is trained to generate the target sequence auto-regressively condition on the source sequence:

$$\mathcal{L}_{Seq2Seq} = -\mathbb{E}_{(S,T) \in D} log P_\theta(T|S) \quad (4)$$

where $P_\theta(T|S) = \prod_{j=1}^{|T|} P_\theta(T_j | T_{<j}, S)$. During pre-training, we alternate between the bidirectional prediction objective and the Seq2Seq generation objective uniformly.

4 Experimental Settings

In this section, we introduce the pre-training and evaluation details, including the pre-training dataset, pre-training and fine-tuning experimental settings.

4.1 Pre-training Dataset

Our pre-training datasets consist of three types: text corpus, image collections and image-text pairs. The statistics of them are shown in Table 1. The text corpus include two large-scale corpora: BookWiki and OpenWebText. BookWiki is composed of English Wikipedia and BookCorpus (Zhu et al., 2015), and OpenWebText is an open recreation of the WebText corpora. The image collections are crawled from the Web, which doesn't contain textual descriptions. The image-text pairs are composed of four existing multi-modal datasets: COCO Captions, Visual Genome (VG), Conceptual Captions (CC) and SBU Captions.

Type	Image-Text Pairs				Images	Text Corpus	
Dataset	COCO	VG	CC	SBU		BookWiki	OpenWebText
Train	533K	5.06M	3.0M	990K	300K	16G	38G
Val	25K	106K	14K	10K			

Table 1: Statistics of the image-text pairs, image collections and text corpus for pre-training.

Task	Image Src.	#Images (#Text)			
		Train	Val	Test	
				test-std	test-dev
VQA	COCO	83K(444K)	41K(214K)	81K(107K)	81K(448K)
Image Caption	COCO	113.2K	5K	5K	-
Visual Entailment	Flickr30K	529.5K	17.9K	17.9K	-

Table 2: Statistics of the datasets for multi-modal downstream tasks.

4.2 Implementation Detail

Our model has 12 layers of Transformer block, where each block has 768 hidden units and 12 self-attention heads. The maximum sequence length of text tokens and image region features are set as 512 and 36, respectively. We pre-train a base version of UNIMO, initialized with RoBERTa base. UNIMO is trained for at least 50W steps. An Adam optimizer with initial learning rate 5e-5 and a learning rate linear decay schedule is utilized.

For visual learning, we adopt Faster R-CNN (Ren et al., 2016) pre-trained on the Visual-Genome dataset to select salient image regions and extract region features from images. The regions with class detection probability exceeds a confidence threshold of 0.2 are selected and 36 boxes are kept. For CMCL, we utilize back-translation to create two positive samples and apply rewriting to obtain 100 hard negative samples for each image-text pair.

4.3 Finetuning Tasks

We fine-tune our model on two categories of downstream tasks: (1) single-modal language understanding and generation tasks; (2) multi-modal vision-language understanding and generation tasks. The single-modal generation tasks include: abstractive summarization on the CNN/DailyMail (CNNDM) dataset (Hermann et al., 2015) and and question generation on the SQuAD dataset (Rajpurkar et al., 2016). The single-modal understanding tasks include: sentiment classification on the SST-2 dataset (Socher et al., 2013) and natural language inference on the MNLI dataset (Williams et al., 2017). The multi-modal tasks include:

- **VQA** requires the model to answer natural language questions by selecting the correct answer from a multi-choice list based on an image. We conduct experiments on the widely-used VQA v2.0 dataset. Similar to previous work, both training and validation sets are used for training for the results on both the test-std and test-dev splits.

- **Image Caption** requires the model to generate a natural language description of an image. We report our results on the Microsoft COCO Captions dataset (Chen et al., 2015). Following Karpathy's (Karpathy and Fei-Fei, 2015) split, the dataset contains 113.2k/5k/5k images for train/val/test splits respectively.

- **Visual Entailment (SNLI-VE)** is a task derived from Flickr30K images and Stanford Natural Language Inference (SNLI) dataset. The task is to determine the logical relationship (i.e. "Entailment", "Neutral" and "Contradiction") between a natural language statement and an image.

The statistics of the datasets for above multimodal-tasks are described in Table 2.

5 Results and Analysis

In this section, we first introduce the evaluation results on both the multi-modal tasks and single-modal tasks to show the adaptability of our model to different scenarios. Then, we further analyze and validate the effectiveness of the unified modeling of different modalities.

Model	VQA		CoCo Caption		SNLI-VE	
	test-dev	test-std	BLUE4	CIDEr	Val	Test
ViLBERT-base	70.55	70.92	-	-	-	-
VL-BERT-base	71.79	72.22	-	-	-	-
VLP-base	70.5	70.7	36.5	116.9	-	-
UNITER-base	72.70	72.91	-	-	78.59	78.28
Oscar-base	73.16	73.61	36.5	123.7	-	-
Villa-base	73.59	73.67	-	-	79.47	79.03
UNIMO-base	**73.79**	**74.02**	**38.6**	**124.1**	**80.00**	**79.10**

Table 3: Evaluation results on the multi-modal downstream tasks.

Model	SST-2	MNLI	CNNDM			SQuAD		
	Acc	Acc-(m/mm)	R-1	R-2	R-L	R-1	R-2	R-L
RoBERTa-base	95.3	87.2/86.5	42.06	19.93	39.22	21.47	24.14	50.36
UniLM-large	94.5	87.0/85.9	43.33	20.21	40.51	22.12	25.06	51.07
UNIMO-base	**95.4**	**86.8/86.1**	**42.42**	**20.12**	**39.61**	**22.26**	**24.77**	**51.17**
w/o single-modal	81.4	59.2/64.1	41.06	19.01	38.23	17.09	21.04	46.47

Table 4: Comparison on the single-modal downstream tasks. R-1, R-2, R-L denotes ROUGE-1, ROUGE-2 and ROUGE-L, respectively. "w/o single-modal" denotes removing the single-modal learning process on the single-modal data from UNIMO, which is similar to UNITER-base (Chen et al., 2020).

5.1 Multi-Modal tasks

The evaluate results on the multi-modal tasks are shown in Table 3. We compare with most of the existed multi-modal pre-training models, including ViLBERT (Lu et al., 2019), VL-BERT (Su et al., 2019), VLP (Zhou et al.), UNITER (Chen et al., 2020), Oscar (Li et al., 2020) and Villa (Gan et al., 2020).

The results show that our UNIMO achieves new state-of-the-art results across all the benchmarks (with base-size model). Specifically, on VQA task, UNIMO outperforms previous SOTA method Villa. On the image caption task, UNIMO outperforms the best performing model Oscar by 2.1 BLUE4 score. On the visual entailment, UNIMO also outperforms all previous models and achieves the best performance.

Existed pre-training methods for the vision-language tasks can only utilize the corpus of image-text pairs. Besides the corpus of image-text pairs, our UNIMO can learn from large scale of single-modal images and texts. The results demonstrate that our UNIMO achieves better performance than previous methods on both the multi-modal understanding and generation tasks.

5.2 Single-Modal tasks

The evaluation results on the single-modal tasks are shown in Table 4, which demonstrate that UNIMO achieves better or comparable performance than existed pre-trained language models on both the language understanding and language generation tasks. Specifically, UniLM (Dong et al., 2019) is a unified pre-trained language model for both natural language understanding and generation tasks. Our base version UNIMO-base outperforms the UniLM-large on several language understanding and generation tasks, including sentiment classification on the SST-2 dataset and question generation on the SQuAD dataset. After removing the single-modal learning on the text corpus and image collections (i.e. "w/o single-modal"), our model is just a multi-modal pre-training method which is similar to UNITER (Chen et al., 2020). As shown in 4, the performance of the model on all language understanding and generation tasks drop dramatically, which demonstrate that multi-modal pre-training method on image-text pairs cannot adapt to single-modal tasks effectively. However, UNIMO not only achieves better performance on multi-modal tasks, but also perform very well on the single-modal tasks for both language understanding and language generation. The results demonstrate the

Model	SST-2	MNLI	CNNDM		
	Acc	Acc-(m/mm)	R-1	R-2	R-L
UNIMO-base	**95.4**	86.8/86.1	**42.42**	**20.12**	**39.61**
w/o pairs&images	94.7	87.0/86.2	42.26	20.09	39.41

Table 5: Analyzing the effectiveness of visual knowledge to language tasks.

Model	CoCo Caption		SNLI-VE
	BLUE4	CIDEr	Val
UNIMO-base	**38.6**	**124.1**	**80.00**
w/o texts	38.3	123.2	79.52

Table 6: Analyzing the effectiveness of textual knowledge to multi-modal tasks.

adaptability of our unified-modal pre-training architecture to different scenarios.

5.3 Vision Enhance Text

The above experimental results have shown the effectiveness of our UNIMO method on both the multi-modal tasks and single-modal tasks. To further validate that the visual knowledge in the image collections and image-text pairs can enhance the language tasks, we remove the images and image-text pairs from the pre-training dataset (i.e. w/o pairs&images) and compare their performance on the single-modal language tasks. After removing the images and image-text pairs, our model is trained by only the language learning objectives, which are similar to previous pre-trained language models BERT and UniLM. Table 5 summarizes the comparison results, which demonstrate that after removing the visual data, the performance of our model drop obviously on several language tasks. The results reveal that the visual knowledge can enhance the language tasks by learning more robust and generalizable representations on the unified semantic space.

5.4 Text Enhance Vision

In this section, we further analyze the influence of textual knowledge to the multi-modal tasks by removing the language learning process on the text corpus (i.e. w/o texts), and comparing their performance on the multi-modal tasks. Table 6 summarizes the comparison results, which show that the model performance decline consistently on both the multi-modal understanding and generation tasks. The results demonstrate that the textual knowledge in the text corpus are useful for the vision-language tasks by enhance the process of cross-modal learning.

6 Conclusion

In this work, we propose a unified-modal pre-training architecture UNIMO, which can leverage large-scale of non-paired text corpus and image collections for cross-modal learning. Our model can effectively adapt to both single-modal and multi-modal understanding and generation tasks. Based on the unified-modal architecture, the textual knowledge and visual knowledge can enhance each other in the unified semantic space. Our UNIMO model outperforms previous methods on both the multi-modal and single-modal downstream tasks. In the future, we will utilize larger scale of image collections and text corpus for unified-modal learning, and extend UNIMO to other modalities of data such as video, audio and so on.

References

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

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

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

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, pages 13063–13075.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *arXiv preprint arXiv:2006.06195*.

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

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28:1693–1701.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2019a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *arXiv preprint arXiv:1908.06066*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019b. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

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

Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. 2020. Interbert: Vision-and-language interaction for multi-modal pretraining. *arXiv preprint arXiv:2003.13198*.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Markus Johannes Van Ackeren, Francesca M Barbero, Stefania Mattioni, Roberto Bottini, and Olivier Collignon. 2018. Neuronal populations in the occipital cortex of the blind synchronize to the temporal dynamics of speech. *ELife*, 7:e31640.

Yu-Siang Wang, Chenxi Liu, Xiaohui Zeng, and Alan Yuille. 2018. Scene graph parsing as dependency parsing. *arXiv preprint arXiv:1803.09189*.

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

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5753–5763.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. 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*, pages 19–27.