UNIMO-2: End-to-End Unified Vision-Language Grounded 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

Vision-Language Pre-training (VLP) has achieved impressive performance on various cross-modal downstream tasks. However, most existing methods can only learn from aligned image-caption data and rely heavily on expensive regional features, which greatly limits their scalability and performance. In this paper, we propose an end-to-end unified-modal pre-training framework, namely UNIMO-2, for joint learning on both aligned image-caption data and unaligned image-only and text-only corpus. We build a unified Transformer model to jointly learn visual representations, textual representations and semantic alignment between images and texts. In particular, we propose to conduct grounded learning on both images and texts via a sharing grounded space, which helps bridge unaligned images and texts, and align the visual and textual semantic spaces on different types of corpora. The experiments show that our grounded learning method can improve textual and visual semantic alignment for improving performance on various cross-modal tasks. Moreover, benefiting from effective joint modeling of different types of corpora, our model also achieves impressive performance on single-modal visual and textual tasks. Our code and models are public at the UNIMO project page https://unimo-ptm.github.io/.

1 Introduction

Large-scale pre-training has drawn much attention in the community of Computer Vision (CV), Natural Language Processing (NLP) and Multi-Modal (MM) due to its strong capability of generalization and efficient usage of large-scale data. However, in the existing literature, the work on vision, language and vision-language representation learning are mostly studied separately with different training data sources. In the vision domain, pre-training on large-scale image corpus such as ImageNet (Deng et al., 2009), OpenImages (Kuznetsova et al., 2020) and JFT-300M (Dosovitskiy et al., 2020) has proven to be critical for learning transferring visual representation for various downstream tasks. In NLP, pre-training on easily-accessible unannotated text corpora greatly improves the capabilities of language understanding and generation (Devlin et al., 2019; Liu et al., 2019; Yang et al., 2019). Pre-training has also become the de-facto approach in vision-language modeling (Lu et al., 2019; Chen et al., 2020c; Li et al., 2020, 2019a; Yu et al., 2020). However, existing VLP methods require a massive amount of aligned image-text pairs which are costly to collect and hard to scale up. The large volumes of image corpus in CV and text corpus in NLP cannot be effectively utilized. Thus, the scalability and performance upper limit of existing VLP methods are largely restricted. As they only learn joint vision-language representations on image-text pairs, they are also difficult to be effectively adapted to visual and textual tasks (Li et al., 2021b; Lin et al., 2020).

To address the limitations, we propose a new end-to-end unified-modal pre-training framework, namely UNIMO-2, for joint learning on various types of corpora, including images, texts, and image-caption pairs. Specifically, we build a unified Transformer model to jointly learn visual representations, textual representations, and cross-modal alignment from the three types of corpora. Both the visual and textual representations are learned end-to-end from raw images and textual sentences. Combining a large number of unaligned images and texts is not only expected to improve the performance of joint vision-language tasks, but also improve the scalability of adapting to single-modal visual and textual tasks. However, it is challenging to bridge unaligned images and texts and effectively align the visual and textual semantic spaces on different types of corpora.

Only a few works have attempted to bridge

3187

unaligned images and texts by leveraging object tags from an pre-trained object detector as "anchor points" (Li et al., 2021a,b). However, they all rely heavily on expensive object-centric visual feature extraction, thus facing the problems of limited visual expressive power and computation inefficiency. In this paper, in order to bridge the unpaired image and text corpora and align the visual and textual semantic spaces end-to-end, we propose to conduct grounded learning on images, texts, and image-text pairs via a sharing grounded space. Specifically, we introduce a grounded dictionary shared by images and texts, which represents vision-language grounded semantics. To learn the grounded dictionary, we apply vector quantization on both visual and textual representations to group image patches and text tokens with similar semantics into grounded tokens. Furthermore, we design a Grounded Transformer architecture to let the visual and textual information exchanged by the grounded tokens, which not only facilitates grounded dictionary learning, but also improves cross-modal alignment. Our grounded learning method can help bridge the textual and visual semantic spaces on unpaired image and text corpora to improve cross-modal fusion on different types of corpora.

We evaluate UNIMO-2 on a variety of representative vision-language understanding and generation tasks, including image/text retrieval, visual question answering, visual reasoning and image caption. On all these tasks, UNIMO-2 obtains obvious improvements compared to the baselines that only learn on aligned image-caption data or without our grounded learning component. Moreover, we also evaluate our model on single-modal textual tasks such as natural language inference and visual tasks such as image classification (Deng et al., 2009). The results show that our model has also achieved very impressive performance on these tasks, which proves the strong scalability and adaptability of our model.

UNIMO-2 has the following advantages compared with previous methods:

- UNIMO-2 can jointly learn from both aligned and unaligned image and text corpora end-to-end, effectively alleviating the limitations of corpus, and learning more generalized visual and textual representations on large volumes of different types of corpus.

- Benefiting from utilizing different types of corpora, UNIMO-2 has better scalability for different types of tasks, including both cross-modal tasks and single-modal tasks.

- Our grounded learning method can help align textual and visual semantic spaces more effectively, thereby greatly improving the performance of various cross-modal tasks. In particular, the performance of zero-shot image/text retrieval even outperforms CLIP pre-trained on an order of magnitude larger pair corpus.

2 Related Work

Vision-Language Pre-training Recent years have witnessed rapid progress in vision-and-language pretraining (VLP) (Li et al., 2019b; Lu et al., 2019; Chen et al., 2020c; Li et al., 2019a, 2020; Yu et al., 2020). Most existing mainstream VLP models adopt a two-stage training method, which firstly extracts region-based visual features using a pre-trained object detection model, and then combines the derived object-centric region features of images and text embeddings as the input of Transformer (Vaswani et al., 2017) for cross-modal pre-training. These methods rely heavily on an off-the-shelf object detector like Faster R-CNN (Ren et al., 2016) typically pretrained on the Visual Genome dataset (Anderson et al., 2018). As the visual representation is not optimized towards a more generic cross-modal understanding and extracting region features with an object detection model is so time-consuming, they face the problems of limited visual expressive power and computation inefficiency, which makes them less scalable.

Some recent work has also explored VLP without object detection modules (Xu et al., 2021; Kim et al., 2021; Huang et al., 2021; Wang et al., 2021). They either utilize grid features from pretrained CNNs or patch features following ViT (Dosovitskiy et al., 2020), however they only use limited image-caption pairs for cross-modal pretraining and thus their scalability and performance are limited. Only a few works have explored utilizing unaligned images and texts for vision-language pre-training, including our previous work UNIMO (Li et al., 2021b) and U-VisualBERT (Li et al., 2021a). However, they all rely on pre-extraction of region-based visual features or object tags by time-consuming object detection. How to bridge unpaired visual and textual corpora end-to-end without using object detection remains challenging.
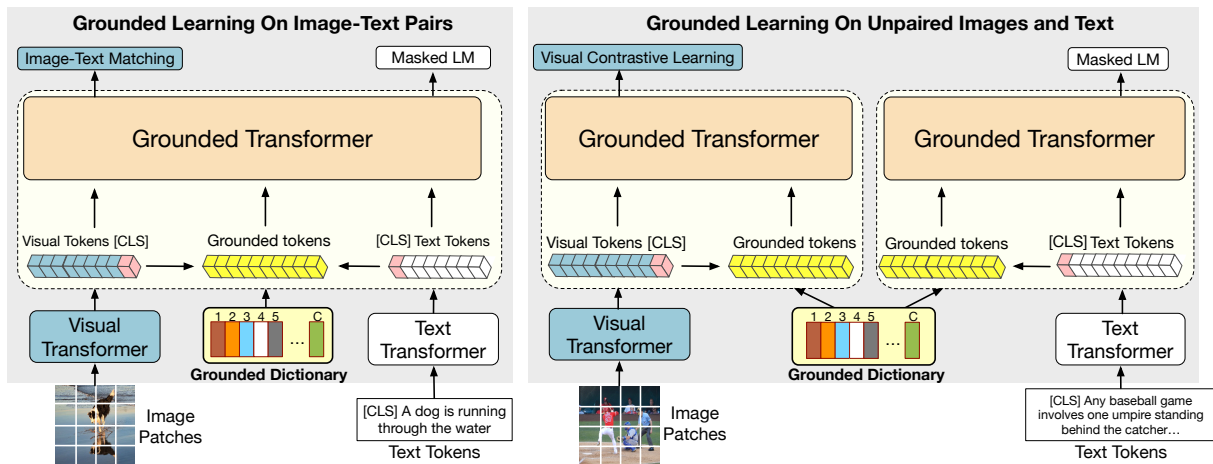
Figure 1: Illustration of our UNIMO-2 framework. The left part shows the architecture of learning on image-text pairs, which produces grounded tokens based on the sharing semantics in images and texts. The right part shows the architecture of learning on unpaired images and texts, which produces grounded tokens from image representations or text representations, respectively. As they share the same grounded dictionary, the grounded tokens act as "anchor points" to bridge the gap between images and texts.

Grounded Learning Language grounding is an active field aiming at enriching textual representations with visual information, which has been shown to improve performance on a variety of core NLP tasks (Bruni et al., 2014; Baroni, 2016; Kiela, 2017). Kiela et al. (2018) investigate grounded sentence representations by training a sentence encoder to predict the image features of a given caption. Tan and Bansal (2020) propose a vokenization method that maps language tokens to their related images. These works all enrich the language representation with visual information by learning a projection of text representations to corresponding images (Chrupała et al., 2015). Recently, Huang et al. (2021) propose an end-to-end VLP method that aggregates visual features from a CNN encoder into visual tokens with a visual dictionary. Liu et al. (2021) propose to improve cross-modal retrieval tasks by incorporating a shared discretized embedding space, which is utilized to compute matching scores between different modalities to complement the representations from individual encoders. These works all rely on image-text pairs to learn cross-modal representations and only focus on joint vision-language tasks. By contrast, our work for the first time proposes to jointly model both aligned and unaligned images and texts by end-to-end learning a shared grounded semantic space, which can improve modality alignment between both aligned and unaligned images and texts.

3 Approach

The overall architecture of our model is shown in Figure 1. UNIMO-2 is an end-to-end framework, which consists of a trainable Transformer-based visual encoder, a Transformer-based text encoder, a grounded dictionary (GD) embedding module, and a multi-layer Grounded Transformer for modality fusion. The visual encoder takes an image as input by splitting it into small sizes of patches, and produces the high-level visual representations for all patches, similar to ViT (Dosovitskiy et al., 2020). The text encoder encodes textual tokens to produce high-level token representations. Based on the high-level representations of patches and tokens, we design a GD embedding module to group similar vision-language representations into grounded tokens with a shared grounded dictionary. The Grounded Transformer is further adopted to fuse features from vision and language modalities through interacting with the common grounded tokens. UNIMO-2 can be end-to-end pre-trained by joint Masked Language Modeling (MLM) on text, Image-Text Matching (ITM) on image-text pairs and Visual Contrastive Learning (VCL) on images. UNIMO-2 can also be easily adapted to various tasks including visual, textual and cross-modal tasks.

3.1 End-to-End Grounded Learning

Human acquire much of their knowledge through grounded learning – visual concepts can be acquired through language, and language acquisi-

3189

tion emerges through visual interaction (Jones et al., 1991; Perfetti, 1998; Fincher-Kiefer, 2001; Andrews et al., 2009; Riordan and Jones, 2011). Inspired by this type of grounded learning, we propose to learn a sharing semantic space (i.e. grounded space) between images and texts to better align fine-grained visual and textual semantics. Specifically, based on the high-level visual representations of patches $V = \{v_1, \ldots, v_M\}$ and textual representations of tokens $T = \{t_1, \ldots, t_N\}$, we introduce a grounded dictionary to group similar visual and textual representations into the same grounded token. The grounded features not only help align the visual and textual semantics in aligned image-caption data, but also act as "anchor points" to help bridge the unaligned images and texts, as shown in Figure 1.

Grounded Dictionary Learning We define a grounded dictionary (GD) as a matrix $G \in \mathbb{R}^{C \times D}$ which contains C embedding vectors with D-dim. The embedding vector for the jth grounded token is denoted as $g_j \in \mathbb{R}^D, j \in 1, 2, \ldots, C$. Vector Quantization (VQ) is widely used to group continuous embeddings into groups of discrete latent variables (Oord et al., 2017; Liu et al., 2021; Huang et al., 2021). For example, each patch or token can be mapped to a grounded token by finding its nearest neighbor in the GD, as in Oord et al. (2017).

Most existing VLP methods implicitly assume that there is a one-to-one correspondence hypothesis between the visual and textual modalities of image-text pairs. However, this hypothesis does not hold in reality as most image-text pairs on the Web are noisy or only have weak correlation. To tackle this issue, instead of mapping each patch or token representation to a grounded token, we only detect the most significant sharing semantics between image and text. We propose to find the top-K most significant grounded tokens for both the textual and visual input. Specifically, let x_{ij} denotes the similarity between embedding vectors of visual token v_i and grounded token g_j, which is computed by:

$$x_{ij} = \sigma(\eta * v_i^T g_j) \qquad (1)$$

where σ denotes the *sigmoid* function, and η denotes a learnable temperature parameter. Similarly, y_{kj} denotes the similarity between embedding vectors of textual token t_k and grounded token g_j.

For image-text pairs, the accumulated score of

the grounded token g_j is computed as:

$$s_j = \sum_{i=1}^{M} x_{ij} + \sum_{k=1}^{N} y_{kj} \qquad (2)$$

We obtain the top-K most significant grounded tokens with the largest accumulated scores: $g_1, \ldots, g_K = Top_K\{s_1, \ldots, s_C\}$, where K is a hyper-parameter. Note that, if we set $K = M + N$, then it is similar that each patch or token is mapped to a grounded token, which will increase the computation cost and introduce noisy information into the grounded learning process. So, we set K much smaller than $M + N$ to obtain the most significant and sharing grounded tokens, which can help align fine-grained visual and textual representations while eliminating the noisy or unrelated information in image-text pairs. For unpaired images or text, the accumulated score of each grounded token g_j is $s_j = \sum_{i=1}^{M} x_{ij}$ or $s_j = \sum_{k=1}^{N} y_{kj}$, and the top-$K$ grounded tokens can be obtained similarly.

The grounded dictionary is randomly initialized, and further updated end-to-end while pre-training. As the Top_K function is non-differentiable, we import a *grounding loss* to help learn the grounded dictionary. Specifically, we propose a revised form of the Vector Quantisation (VQ) algorithm (Oord et al., 2017), which uses the l_2 error to move the embedding vectors g_i towards the mapped patch or token representations, as shown in the first term of Equation 3. For simplicity, here we take image input as an example. Since the volume of the embedding space is dimensionless, it can grow arbitrarily if the embeddings g_i do not train as fast as the visual and textual encoder parameters. To make sure the encoder commits to an embedding and its output does not grow, we add a commitment loss, the second term in Equation 3. Thus, the total grounding loss becomes:

$$\mathcal{L}_{GD} = \sum_{i=1}^{M} \| sg(v_i) - \sum_{j} \frac{x_{ij}}{\sum_k x_{ik}} g_j \|_2^2 \\ + \beta \sum_{j=1}^{K} \| sg(g_j) - \sum_{i} \frac{x_{ij}}{s_j} v_i \|_2^2 \qquad (3)$$

where $sg(.)$ denotes the stop-gradient operator that is defined as identity at forward computation time and has zero partial derivatives, and β denotes a weight parameter.

The grounded dictionary faces a cold-start problem for unpaired images and texts. So we apply
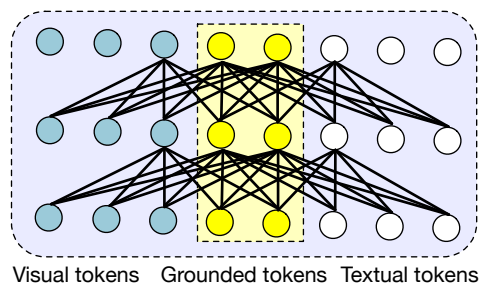
Figure 2: The self-attention architecture of Grounded Transformer. Cross-modal information is exchanged through the grounded tokens.

curriculum learning on different types of corpora. Specifically, we first only train on image-text pairs for 20 epoches to obtain a usable grounded embedding space, then further train on all three types of corpus to help bridge unpaired images and texts. To show what the GD has learned, we have visualized some grounded tokens in Appendix A.

Grounded Transformer After obtaining the grounded tokens, we append them with the visual tokens and textual tokens as input to our Grounded Transformer for cross-modal fusion. Specifically, we propose to bridge visual and textual representations by grounded tokens. As shown in Figure 2, the cross-modal information can only be exchanged by grounded tokens, which also push the grounded tokens to capture the most significant sharing semantics between images and texts. In this way, our model is more robust on weak correlation image-text pairs by modeling cross-modal interaction through common grounded tokens. Furthermore, the novel self-attention architecture can improve the computation efficiency compared to the standard pairwise self-attention mechanism.

For unpaired images and texts, the Grounded Transformer also models the fusion of visual tokens or textual tokens with the grounded tokens. As the grounded dictionary captures common visual and textual semantics, it also helps learn cross-modal representations on unpaired images and texts.

3.2 Pre-training On Different Corpus

Based on the outputs of the Grounded Transformer, we adopt Masked Language Modeling (MLM) and Image-Text Matching (ITM) pre-training tasks on image-text pairs. Furthermore, we also apply MLM on text corpus and Visual Constrastive Learning (VCL) on images.

Masked Language Modeling We iteratively sample 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 original token with probability 80%, 10% and 10%, respectively. The goal is to predict these masked tokens based on their surrounding context and all visual features. The MLM task is also applied on text-only corpus, which predicts masked tokens only based on the surrounding tokens.

Image-Text Matching To enhance the cross-modal matching, we adopt ITM task for pre-training as in previous works (Chen et al., 2020c). We apply a binary classifier on the concatenated embedding features of the "[CLS]" token in text and the "[CLS]" token in image by Grounded Transformer to predict whether the input image and text are matched or not.

Visual Contrastive Learning UNIMO-2 learns representations on unpaired images by maximizing agreement between differently augmented views of the same image while minimizing similarities between different images via a contrastive loss in the latent space, similar to SimCLR (Chen et al., 2020a). We apply stochastic data argumentation module that transforms an image randomly resulted in two corelated views as a positive pair, and random images in the same minibatch as negative pairs. We combine augmentations of random cropping, random rotating and random color distortion followed by resizing back to the original size.

3.3 Transferring To Different Tasks

Our model can be effectively finetuned on different types of tasks, including cross-modal tasks, visual tasks and textual tasks. For cross-modal tasks, the model architecture is the same as the pre-training architecture on image-text pairs, as shown in the left part of Figure 1. Grounded tokens are produced based on both the visual and textual representations to facilitate cross-modal understanding and generation. For visual tasks, the model architecture is the same as the pre-training architecture on images, as shown in the middle part of Figure 1. Grounded tokens are obtained based on the visual representations from the Visual Transformer. As the grounded tokens contain sharing semantics between images and texts, UNIMO-2 can learn language-grounded

image representations for visual tasks. Similarly, for textual tasks the model architecture is the same as the pre-training architecture on text, as shown in the right part of Figure 1. Grounded tokens are obtained based on the textual representations from the Text Transformer. Also, the sharing grounded space helps learn grounded text representations to facilitate textual tasks.

4 Experimental Settings

Pretraining Dataset Our pre-training datasets consist of three types: text corpus, image corpus and image-text pairs. The text corpus includes two large-scale corpora: BookWiki and OpenWebText, which are part of the training dataset of RoBERTa (Liu et al., 2019). The image corpus are images without textual descriptions, including a subset of OpenImages (Krasin et al., 2017) and ImageNet-21k (Deng et al., 2009). Each image in these datasets contains a textual label. The image-text pairs are composed of four existing multi-modal datasets: COCO (Lin et al., 2014), Visual Genome (VG) (Krishna et al., 2017), Conceptual Captions (CC) (Sharma et al., 2018) and SBU Captions (Ordonez et al., 2011), which have also been widely used in previous VLP models. The detail statistics are shown in the appendix. We also transform the label of each image to a sentence by prompts (e.g. "a photo of [label]") to create pseudo image-text pairs from the OpenImages and ImageNet-21k datasets for pretraining.

Implementation Detail UNIMO-2 consists of 12 layers of Visual Transformer, 12 layers of Text Transformer, and 12 layers of Grounded Transformer. The Visual Transformer is initialized by ViT-B/16. The Text Transformer and Grounded Transformer are both initialized by RoBERTa-Base. The maximum sequence length of text tokens are set as 512. An Adam optimizer with initial learning rate 5e-5 and a learning rate linear decay schedule is utilized.

For the visual encoder, our model receives the raw image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ and maps it into flattened $1D$ sequence of patches $\mathbf{x}_p \in \mathbb{R}^{\frac{HW}{P^2} \times D}$ as input for the transformer, where D is the fixed hidden size of the transformer layers and P is the patch size. During pretraining, we utilize the 224×224 resolution with a fixed patch size of 16×16, resulting in a patch sequence of length 14×14 as visual tokens. During fine-tuning, we increase the image resolution to 384×384 and interpolate the

positional encoding of image patches following (Dosovitskiy et al., 2020). For the grounded embedding module, the grounded dictionary size C is set as 2048, and the number of grounded tokens K during pre-training and finetuning are both set as 100 that is much smaller than the max number of patches and tokens for pre-training (i.e. 709) and finetuning (i.e. 1089). We set $\beta = 0.25$ in all our experiments and the results did not vary obviously for values ranging from 0.1 to 1.0. We have compared different grounding settings in detail in Appendix A.

Finetuning Tasks To show the scalability of our model, we fine-tune it on three types of downstream tasks: (1) joint vision-language cross-modal tasks, (2) visual tasks, and (3) textual tasks. The cross-modal tasks include: visual question answering (VQA) on the VQA v2.0 dataset (Goyal et al., 2017), image caption on the Microsoft COCO Captions dataset (Chen et al., 2015), visual entailment on the SNLI-VE dataset (Xie et al., 2019) and image-text retrieval on Flickr30k datasets (Young et al., 2014). The visual tasks include image classification on the ImageNet-1k dataset (Krizhevsky et al., 2012). The textual tasks include sentiment classification on the SST-2 dataset (Socher et al., 2013), natural language inference on the MNLI dataset (Williams et al., 2018), linguistic acceptability analysis on the CoLA dataset (Warstadt et al., 2019) and semantic similarity analysis on the STS-B dataset (Cer et al., 2017). The detail statistics of the datasets and hyper-parameter settings for the above tasks are described in Appendix B.

5 Results and Analysis

We compare UNIMO-2 to a variety of state-of-the-art models on cross-modal, visual and textual tasks.

5.1 Cross-Modal Tasks

The evaluation results on the joint vision-language cross-modal tasks are shown in Table 1. We compare with most of the existed VLP models, including regional feature-based models ViLBERT (Lu et al., 2019), UNITER (Chen et al., 2020c), Oscar (Li et al., 2020), Villa (Gan et al., 2020) and UNIMO (Li et al., 2021b), and end-to-end models ViLT (Kim et al., 2021), E2E-VLP (Xu et al., 2021), SOHO (Huang et al., 2021) and CLIP (Radford et al., 2021).The results show that UNIMO-2 achieves the best results against most benchmarks, outperforming both the base and large sizes of other

Model	ZS-IR R@1/R@5	ZS-TR R@1/R@5	IR R@1/R@5	TR R@1/R@5	SNLI-VE Val / Test	VQA test-dev / std	Caption B@4 / C
Region-based Models Pretrained on Image-Text Pairs of CC, SBU, COCO and VG.							
ViLBERT	31.86/61.12	-	58.20/84.90	-	-	70.55/70.92	-
UNITER-Base	66.16/88.40	80.70/95.70	72.52/92.36	85.90/97.10	78.59/78.28	72.70/72.91	-
Villa-Base	-	-	74.74/92.86	86.60/97.90	79.47/79.03	73.59/73.67	-
Oscar-Base	-	-	-	-	-	73.16/73.44	36.5/123.7
UNIMO-Base	62.44/86.16	77.40/95.10	74.66/93.40	89.70/98.40	80.00/79.10	73.79/74.02	38.8/124.4
UNITER-Large	68.74/89.20	83.60/95.70	75.56/94.08	87.30/98.00	79.39/79.38	73.82/74.02	-
Villa-Large	-	-	76.26/94.24	87.90/97.50	80.18/80.02	74.69/74.87	-
Oscar-Large	-	-	-	-	-	73.61/73.82	37.4/127.8
UNIMO-Large	72.14/91.14	85.80/96.80	78.04/94.24	89.40/98.90	81.11/80.63	75.06/75.27	39.6/127.7
End-to-End Models Pretrained on Image-Text Pairs of CC, SBU, COCO and VG. † denotes 400 Million pairs.							
ViLT	51.3/79.9	69.7/91.0	62.2/87.6	83.7/97.2	-	70.94/-	-
E2E-VLP	-	-	73.58/92.42	86.24/97.50	-	73.25/73.67	36.2/117.3
SOHO	-	-	72.5/92.7	86.5/98.1	**85.00/84.95**	73.25/73.47	-
CLIP†	68.7/90.6	88.0/**98.7**	-	-	-	-	-
Our Baseline	65.11/87.44	78.80/94.38	78.52/94.02	91.62/98.72	80.37/80.43	75.69/75.87	38.5/128.4
UNIMO-2	**72.70/91.18**	**88.46**/96.84	**80.14/95.58**	**92.01/99.31**	81.97/81.48	**76.31/76.42**	**39.7/131.2**

Table 1: Evaluation results on cross-modal tasks. ZS denotes zero-shot performance. IR and TR represents image-retrieval and text-retrieval, respectively. B@4 and C denotes metrics of BLUE4 and CIDEr, respectively. "Our Baseline" is similar to UNIMO-2, except that the grounded embedding module in UNIMO-2 is removed. It is trained on the same corpus and experimental settings with UNIMO-2.

Model	Acc@1 Zero-Shot	Acc@1 Finetuned
SimCLRv2 (Chen et al., 2020b)	-	80.5
CLIP-ViT(B/16)	**68.6**	80.2
Our Baseline	58.2	80.7
UNIMO-2	66.3	**80.8**

Table 2: Evaluation results on visual tasks, compared to state-of-the-art representation learning methods. We report both the zero-shot and finetuned top-1 accuracy on ImageNet-1k. The finetuned result of CLIP-ViT is linear probe performance.

Model	SST-2 Acc	MNLI Acc-(m/mm)	CoLA Mat	STS-B Per
BERT	92.7	84.4 / -	-	-
RoBERTa	94.8	-	63.6	-
UniLM	94.5	87.0/85.9	61.1	87.7
UNITER	89.7	80.8/-	37.4	-
VilBERT	90.4	79.9/-	36.1	-
UNIMO	**95.1**	86.8/86.7	**65.4**	91.0
Our Baseline	94.1	87.1/86.9	60.6	91.0
UNIMO-2	94.7	**87.5/87.5**	62.1	**91.2**

Table 3: Evaluation results on textual tasks. Mat and Per denote Matthews correlation coefficient and Pearson correlation coefficient, respectively. All the results are evaluated on the dev set.

VLP models. Particularly, UNIMO-2 achieves very good performance on the task of zero-shot image/text retrieval, even outperforming CLIP (Radford et al., 2021) that pre-trained on an order of magnitude larger corpus. The results demonstrate that UNIMO-2 can obtain better cross-modal representations based on joint end-to-end grounded learning on different types of corpus.

Furthermore, the performance of "Our Baseline" that just removes the grounded embedding module in UNIMO-2 drop obviously on all tasks, which demonstrates the effectiveness of our grounded learning method for cross-modal alignment. Especially, on the zero-shot image retrieval and text retrieval tasks, UNIMO-2 obtains 7.59 R@1 and 9.66 R@1 absolute gains compared to "Our Baseline". The results demonstrate that our grounded learning method can help align the visual and textual semantic space on different types of corpora to obtain more effective cross-modal representations.

5.2 Visual Tasks

UNIMO-2 can also be effectively adapted to visual tasks such as image classification. As UNIMO-2 learns effective cross-modal representations, it can classify images without finetuning. Specifically, the target labels of images can be transformed into pseudo image descriptions, such as "a photo of [label]". Then the zero-shot image-to-text retrieval method can be used to obtain the label for each image, similar to CLIP (Radford et al., 2021). Both the zero-shot and finetuned performance is compared to several state-of-the-art representation learning methods. The results in Table 2 show that UNIMO-2 can achieve comparable performance with CLIP that pretrained on billions of image-text pairs, on both the zero-shot and supervised settings. Moreover, UNIMO-2 obviously outperforms

	Model	ZS-IR R@1	ZS-TR R@1	IR R@1	TR R@1	COCO Caption B@4 / C	ZS-ImageNet Acc@1	MNLI m/mm
	UNIMO-2	**72.70**	**88.46**	**80.14**	**92.01**	**39.7 / 131.2**	66.3	**87.5/87.5**
GD	w/o GD (P)	65.11	78.80	78.52	91.62	38.5 / 128.4	58.2	87.1/86.9
	w/o GD (I)	40.22	31.76	74.08	88.26	39.0 / 127.4	21.3	87.5/87.3
	w/o G.T.	70.10	85.01	78.84	91.12	39.6 / 130.1	**66.4**	87.1/86.8
	1-to-1 Map	66.06	80.97	77.61	90.43	38.7 / 127.4	66.3	87.0/86.9
Corpus	w/o Text	70.00	85.50	78.90	90.24	39.0 / 128.7	65.0	84.9/85.0
	w/o Images	69.17	84.81	77.65	90.34	39.4 / 129.5	42.2	87.1/87.0
	w/o Both	70.06	84.12	78.17	91.32	39.3 / 129.3	43.0	85.9/85.7

Table 4: Ablation study on the effectiveness of our unified end-to-end grounded learning architecture.

"Our Baseline" on the zero-shot setting, achieving 8.1 Acc@1 absolute gains. The results demonstrate that UNIMO-2 also learns generalized visual representations through unified-modal learning on different types of corpora.

5.3 Textual Tasks

To show the effectiveness of UNIMO-2 on textual tasks, we further compare with both VLP models including UNITER, VilBERT and UNIMO, and pre-trained language models including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and UniLM (Dong et al., 2019). The comparison results in Table 3 demonstrate that UNIMO-2 achieves much better performance than existing VLP models including UNITER and VilBERT, and achieves comparable performance than existed PLMs such as RoBERTa. UNIMO-2 also outperforms "Our Baseline" on all textual tasks.

The above results demonstrate the adaptability and scalability of our unified end-to-end VLP architecture for joint learning on both aligned and unaligned images and texts. In all, UNIMO-2 not only achieves excellent performance on cross-modal tasks, but also performs very well on visual and textual tasks, which validates the superiority of our unified-modal learning architecture.

5.4 Analysis

Effectiveness of Grounded Learning We further validate the effectiveness of our grounded learning component by ablation study. "w/o GD (P)" denotes removing the grounded learning component during both pre-training and inference in order to validate its effectiveness for unified learning on different types of corpus. "w/o GD (I)" denotes keeping the grounded learning component during pre-training, but removing it during inference, in order to validate the effectiveness of the grounded representations to downstream tasks. "1-to-1 Map" denotes mapping each patch or token to

a grounded token by finding its nearest neighbor in the grounded dictionary, similar to the vector quantization method in (Oord et al., 2017). We compare their performance on three types of tasks, as shown in the top part of Table 4. The results demonstrate that our grounded learning (GD) method is essential to the end-to-end joint learning from different types of corpus, which can help bridge unaligned images and texts and improve vision-language semantic alignment. The learned grounded representations is also critical to both the cross-modal and single-modal downstream tasks. We further validate the effectiveness of our Grounded Transformer by replacing it with a traditional Transformer, denoted as "w/o G.T.". The results show that the performance of cross-modal tasks drop obviously compared to UNIMO-2, which demonstrate the effectiveness of our Grounded Transformer architecture.

Effectiveness of Unaligned Images and Texts To further validate the effectiveness of unaligned images and texts to cross-modal learning, we compare the performance of UNIMO-2 on different pre-training datasets. Specifically, we compare the performance of UNIMO-2 by either removing the text cropus (i.e. "w/o Text"), the image corpus (i.e. "w/o Images") or removing them both (i.e. "w/o Both"). The comparison results are shown in the bottom part of Table 4, which show that either removing text corpus or image corpus will consistently reduce the performance of all three types of tasks, including cross-modal, visual and textual tasks. It is worth noting that the performance of the image/text retrieval tasks drop obviously when either removing the text-only cropus or image-only corpus, which demonstrate that unaligned corpus is also useful to cross-modal tasks. UNIMO-2 can effectively leverage unaligned images and texts to improve cross-modal learning.

6 Conclusion

In this work, we propose UNIMO-2, an end-to-end unified-modal pre-training framework that can learn from both aligned and unaligned image and text corpora. Our proposed grounded learning method can help bridge unpaired images and texts and align the textual and visual semantic spaces more effectively. Benefiting from effectively utilizing different types of corpora, UNIMO-2 has better scalability for different types of tasks. Experiments show that UNIMO-2 greatly improves the performance of various cross-modal tasks and also achieves very impressive performance on visual and textual tasks. The results also show that it is promising to further uniformly improve the performance of cross-modal, visual and textual tasks by utilizing larger scales of unpaired images and texts.

Acknowledgments

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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.

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological review*, 116(3):463.

Marco Baroni. 2016. Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1):3–13.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49:1–47.

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

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

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. 2020b. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*.

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. 2020c. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer.

Grzegorz Chrupała, Ákos Kádár, and Afra Alishahi. 2015. Learning language through pictures. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 112–118, Beijing, China. Association for Computational Linguistics.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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

Rebecca Fincher-Kiefer. 2001. Perceptual components of situation models. *Memory & Cognition*, 29(2):336–343.

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*.

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

Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. *arXiv preprint arXiv:2104.03135*.

Susan S Jones, Linda B Smith, and Barbara Landau. 1991. Object properties and knowledge in early lexical learning. *Child development*, 62(3):499–516.

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.

Douwe Kiela. 2017. Deep embodiment: grounding semantics in perceptual modalities. Technical report, University of Cambridge, Computer Laboratory.

Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. 2018. Learning visually grounded sentence representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 408–418, New Orleans, Louisiana. Association for Computational Linguistics.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*.

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. 2017. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):2–3.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981.

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*.

Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. 2021a. Unsupervised vision-and-language pre-training without parallel images and captions. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5339–5350, Online. Association for Computational Linguistics.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021b. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2592–2607, Online. Association for Computational Linguistics.

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*.

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

Alexander H Liu, SouYoung Jin, Cheng-I Jeff Lai, Andrew Rouditchenko, Aude Oliva, and James Glass. 2021. Cross-modal discrete representation learning. *arXiv preprint arXiv:2106.05438*.

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.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*.

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

Charles A Perfetti. 1998. The limits of co-occurrence: Tools and theories in language research.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

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.

Brian Riordan and Michael N Jones. 2011. Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2):303–345.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew 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, Seattle, Washington, USA. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages

2066–2080, Online. Association for Computational Linguistics.

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

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

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Cola: The corpus of linguistic acceptability (with added annotations).

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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

Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. 2021. E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 503–513, Online. Association for Computational Linguistics.

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.

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

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernievil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*.

A Grounded Learning Analysis

Visualization of Grounded Dictionary To show the semantics of the grounded dictionary learned by UNIMO-2, we visualize the image patches and textual tokens that are grouped in each grounded token. We map each patch or token into a grounded token with which has the largest similarity between their representations by Equation 1. For each grounded token, the patches and tokens that have the largest similarity scores are selected and visualized. Several examples are shown in Figure 3, which demonstrate that each grounded token captures meaningful and consistent vision-language grounded semantics.

Parameter Analysis In all our experiments, we utilize the default grounding settings that the grounded dictionary (GD) size C is set as 2048 and the number of grounded tokens K is set as 100. We further compare different grounding settings to explore the properties of the grounded semantic space for cross-modal learning. Specifically, we validate the performance of grounded learning with different grounded dictionary (GD) size C from $\{1024, 2048, 4096, 8192\}$ and different number of grounded tokens K from $\{10, 20, 50, 100\}$. When comparing different GD size C, we set K as 100. We also keep $C = 2048$ when comparing different settings of K. Furthermore, we also compare our method with the simplest Vector Quantization (VQ) method that maps each visual or textual token to a grounded token by finding its nearest neighbor in the grounded dictionary, namely "1-to-1 map". The number of grounded tokens for "1-to-1 map" is depended on the total number of image patches and textual tokens, which is 709 (i.e. 197 + 512) during pre-training and 1089 (i.e. 577 + 512) during finetuning.

For time efficiency, we only pre-train UNIMO-2 on the corpus of image-text pairs for 10 epoches under the above settings, and then compare their performance on two representative cross-modal tasks, including zero-shot image/text retrieval and image caption, to validate their effectiveness on cross-modal alignment. The comparison results are shown in Table 5, which demonstrate that our grounded learning method achieves better performance on the two representative cross-modal tasks when the GD size C is set as 4096 or the number of grounded tokens K is set as 50. Too large C will increase the difficulty of learning while too

small C may restrict the volume of grounded semantic space. Similarly, too small K will lose sharing semantics between images and texts while too large K will introduce noisy information. Although different settings have different behavior, the performance of our grounded learning method is relatively stable. In particular, the "1-to-1 map" method achieves much worse results than our grounded learning method under different settings, which validates the effectiveness of our grounded learning method on cross-modal alignment. Furthermore, our grounded learning method is much more efficient in computation than "1-to-1 map" as the number of grounded tokens is much smaller, which largely reduce the sequence length during cross-modal fusion.

B Experimental Settings

Pretraining Datasets The pre-training datasets consist of text corpus, image collections and image-text pairs. The detail statistics of them are shown in Table 6.

**Finetuning Tasks** The multi-modal finetuning 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 (Goyal et al., 2017), which is built based on the COCO (Chen et al., 2015) images. 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 evaluated on the SLNI-VE dataset (Xie et al., 2019) which was 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.
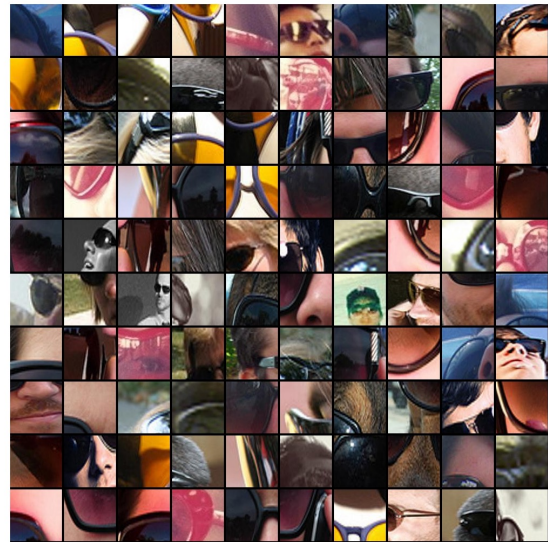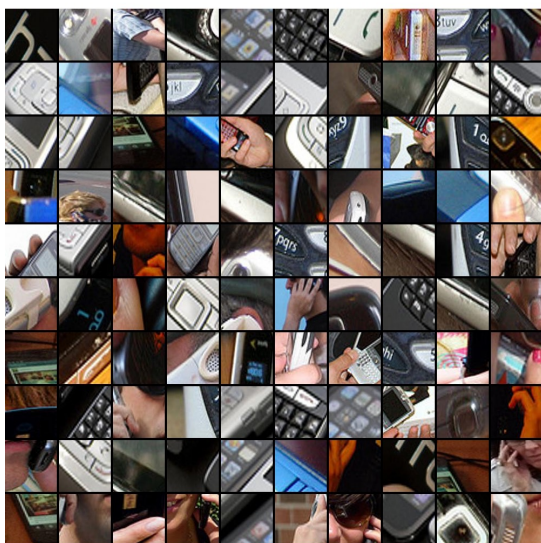
ID=67: cake, birthday, candles, happy, party

ID=74: airplane, flight, flying, plane, aircraft
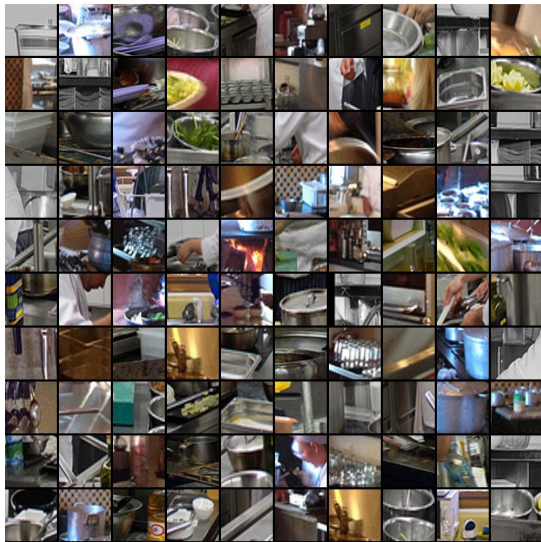
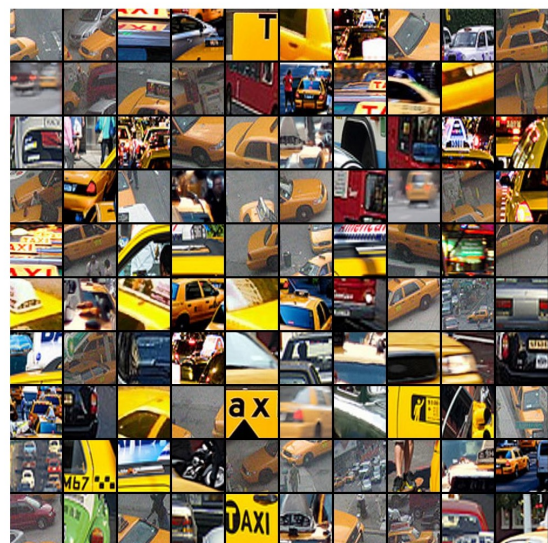ID=153: kitchen, cook, chefs, food, pans

ID=680: glasses, sun, wearing, beach, sunny

ID=885: phone, cell, mobile, cellphone, talking

ID=1211: taxi, transports, city, traffic, street

Figure 3: Visualization of the grounded dictionary learned by UNIMO-2, which groups consistent semantics of image patches and textual tokens. Each grounded token reflects an abstraction of vision-language grounded semantics.

Model		ZeroShot-IR R@1 / R@5 / R@10	ZeroShot-TR R@1 / R@5 / R@10	COCO Caption B@4 / M / C / S
GD Size C	1024	58.52 / 82.19 / 88.92	71.10 / 90.14 / 95.17	37.58 / 29.18 / 123.53 / 22.23
	2048	60.32 / 84.02 / 89.72	75.84 / 91.91 / 95.56	37.62 / 29.12 / 123.38 / 22.16
	4096	**64.10 / 86.41 / 91.79**	**77.91 / 94.38 / 96.75**	**38.07 / 29.20** / 124.18 / 22.20
	8192	61.20 / 85.29 / 90.73	75.84 / 92.50 / 96.15	37.86 / 29.03 / **124.23 / 22.33**
Top-K	10	57.79 / 82.66 / 89.47	69.92 / 91.42 / 95.56	37.36 / 28.92 / 122.93 / 22.15
	20	61.46 / 85.07 / 90.75	74.46 / 93.10 / **97.34**	37.90 / 28.81 / 123.68 / 22.03
	50	**63.49 / 86.13 / 91.54**	**77.32 / 93.10** / 96.65	**38.38 / 29.17 / 125.31 / 22.39**
	100	60.32 / 84.02 / 89.72	75.84 / 91.91 / 95.56	37.62 / 29.12 / 123.38 / 22.16
1-to-1 Map		56.51 / 81.54 / 88.19	71.99 / 90.43 / 94.58	35.62 / 27.97 / 117.92 / 21.38

Table 5: Parameter analysis for grounded learning. The top part validates the influence of GD size C, and the middle part compares the performance of different number of grounded tokens K used during learning. The bottom part shows the effectiveness of our grounded learning method compared with the existing VQ method.

Type	Image-Text Pairs				Unaligned Images		Unaligned Text	
Dataset	COCO	VG	CC	SBU	ImageNet21K	Open Images	BookWiki	OpenWebText
#Images	113K	108K	3.01M	867K	14M	1.7M		
#Texts	567K	5.41M	3.01M	867K			16G	38G

Table 6: Statistics of the aligned image-text pairs, and unaligned images and texts 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	-
Image-Text Retrieval	Flickr30K	29K(145K)	1K(5K)	1K(5K)	-

Table 7: Statistics of the datasets for the cross-modal downstream tasks.

Hyper-params	Textual Tasks	Visual Tasks
Learning Rate	{1e-5, 2e-5, 3e-5}	{1e-4, 3e-4, 5e-4}
Batch Size	{16, 32}	512
Epochs	10	10
Warmup Raito	0.06	0.06
Weight Decay	0.01	0.01

Table 8: Hyper-parameters for fine-tuning on visual and textual tasks.

VLP methods are shown in Table 10.

- **Image-Text Retrieval** is evaluated on the Flickr30k dataset (Young et al., 2014), which contains two sub-tasks: image retrieval (Flickr30k-IR) and text retrieval (Flickr30k-TR), depending on which modality is used as the retrieved target. We report the top-K retrieval results on the test sets, including R@1, R@5 and R@10.

The statistics of the datasets for the above multimodal-tasks are described in Table 7. The hyper-parameters for finetuning all the downstream tasks, including both the single-modal tasks and cross-modal tasks are shown in Table 8 and 9, respectively. The full evaluation results (including R@1, R@5 and R@10) on Image/Text Retrieval tasks and comparison with other state-of-the-art

Hyper-parameters	Image-Text Retrieval	SNLI-VE	VQA	COCO Caption
Batch Size	32	64	256	32
Epoch	40	10	12	10
Learning Rate	5e-6 for epoch=[0,24] 5e-7 for epoch=[24,32] 5e-8 for epoch=[32,40]	1e-5	4e-5 for epoch=[0,5] 4e-6 for epoch=[6,8] 4e-7 for epoch=[9,12]	1e-5
Warmup Ratio	-	0.06	-	0.06
Weight Decay	0.01	0.0	0.0	0.01

Table 9: Hyper-parameters for fine-tuning on cross-modal tasks .

Model	ZeroShot-IR R@1 / R@5 / R@10	ZeroShot-TR R@1 / R@5 / R@10	Finetuned-IR R@1 / R@5 / R@10	Finetuned-TR R@1 / R@5 / R@10
ViLBERT-base	31.86 / 61.12 / 72.80	-	58.20 / 84.90 / 91.52	-
UNITER-base	66.16 / 88.40 / 92.94	80.70 / 95.70 / 98.00	72.52 / 92.36 / 96.08	85.90 / 97.10 / 98.80
Villa-base	-	-	74.74 / 92.86 / 95.82	86.60 / 97.90 / 99.20
UNIMO-base	62.44 / 86.16 / 91.68	77.40 / 95.10 / 97.80	74.66 / 93.40 / 96.08	89.70 / 98.40 / 99.10
UNITER-large	68.74 / 89.20 / 93.86	83.60 / 95.70 / 97.70	75.56 / 94.08 / 96.76	87.30 / 98.00 / 99.20
Villa-large	-	-	76.26 / 94.24 / 96.84	87.90 / 97.50 / 98.80
UNIMO-large	72.14 / 91.14 / 94.98	85.80 / 96.80 / 98.80	78.04 / 94.24 / 97.12	89.40 / 98.90 / 99.80
ViLT	51.3 / 79.9 / 81.9	69.7 / 91.0 / 96.0	62.2 / 87.6 / 93.2	83.7 / 97.2 / 98.1
E2E-VLP	-	-	73.58 / 92.42 / 96.03	86.24 / 97.50 / 98.92
SOHO	-	-	72.5 / 92.7 / 96.1	86.5 / 98.1 / 99.3
CLIP	68.7 / 90.6 / **95.2**	88.0 / **98.7 / 99.4**	-	-
Our Baseline	65.11 / 87.44 / 92.62	78.80 / 94.38 / 97.63	78.52 / 94.02 / 96.63	91.62 / 98.72 / 99.51
UNIMO-2	**72.70 / 91.18** / 94.60	**88.46** / 96.84 / 98.92	**80.14 / 95.58 / 97.75**	**92.01 / 99.31 / 99.51**

Table 10: Full evaluation results on the Flickr30k retrieval tasks.