SemVLP: Vision-Language Pre-training by Aligning Semantics at Multiple Levels

Anonymous authors
Paper under double-blind review

Abstract

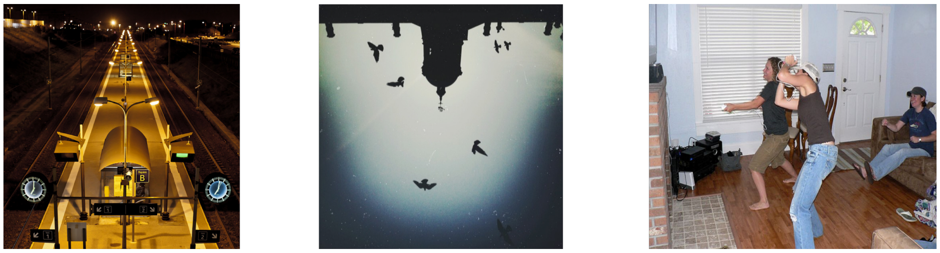
Vision-language pre-training (VLP) on large-scale image-text pairs has recently witnessed rapid progress for learning cross-modal representations. Existing pre-training methods either directly concatenate image representation and text representation at a feature level as input to a single-stream Transformer, or use a two-stream cross-modal Transformer to align the image-text representation at a high-level semantic space. In real-world image-text data, we observe that it is easy for some of the image-text pairs to align simple semantics on both modalities, while others may be related after higher-level abstraction. Therefore, in this paper, we propose a new pre-training method SemVLP, which jointly aligns both the low-level and high-level semantics between image and text representations. The model is pre-trained iteratively with two prevalent fashions: single-stream pre-training to align low-level semantics and two-stream pre-training to align high-level semantics, by employing a shared Transformer network with a pluggable cross-modal attention module. An extensive set of experiments have been conducted on four well-established vision-language understanding tasks to demonstrate the effectiveness of the proposed SemVLP in aligning cross-modal representations towards different semantic granularities.

1 Introduction

Inspired by the success of pre-trained language models in various NLP tasks, recent studies (Lu et al., 2019; Li et al., 2019; Su et al., 2019; Tan & Bansal, 2019; Chen et al., b; Li et al., 2020; Yu et al., 2020) on vision-language pre-training (VLP) have demonstrated the state-of-the-art results in a variety of Vision-and-Language (V+L) tasks, which learn the semantic alignment between the different modalities by harnessing from large-scale image-text pairs.

Existing VLP models basically follow the multi-layer Transformer architecture (Vaswani et al., 2017) such as BERT (Devlin et al., 2018) and use the self-attention mechanism to learn the image-text semantic alignment on large-scale cross-modal data. In terms of the granularity of the cross-modal alignment, there are roughly two mainstreams which use different model architectures to align the cross-modal representations. The single-stream Transformer architecture (Li et al., 2019; Su et al., 2019) assumes that the underlying semantics behind the two modalities is simple and clear, and thus simply concatenates image-region features and text features in low-level semantic space for early fusion in a straightforward manner. The two-stream Transformer architecture (Lu et al., 2019; Tan & Bansal, 2019; Yu et al., 2020) first uses a single-modal Transformer to learn high-level abstraction of image and sentence representation respectively, and then combines the two modalities together with a cross-modal Transformer.

The semantic gap between different modalities has always been treated as one of the most significant problems in cross-modality research. Previous methods bridge the semantic gap either at low-level feature space by using techniques such as cross-media hashing (Song et al., 2013) and canonical correlation analysis (CCA) (Hardoon et al., 2004), or by aligning cross-modal representations at the concept space of high-level semantics, such as multi-modal topic models (Wang et al., 2014) and multi-modal autoencoders (Feng et al., 2014). In real-world image-text data, we observe that it is easy for some of the image-text pairs to align simple semantics on both modalities, while others may be related after higher-level abstraction. As shown in Figure 1, the captions of T1 are more focused on the overview of the image with coarse-level semantics, while T2 are more detailed descriptions



T1: A view of two streets during the night time.
T2: A man sits alone on a train platform at night.

T1: Upside down picture of a building surrounded by birds.
T2: An image of a building with a steeple and birds flying overhead reflected in the water.

T1: A group of women playing video games together.
T2: Two people using an interactive gaming system while a person observes them from a couch.

Figure 1: Examples of images with two different caption text pieces from the MS COCO caption dataset, where some captions are more fine-grained than the others that are more abstract.

that emphasize on the specific parts of the images. The semantic granularity spans different levels for different captions of the same images. It is essential to explicitly consider aligning semantics at multiple levels for deeply understanding the real world image-text data.

In this paper, we introduce a new VLP pre-training architecture, SemVLP, which jointly aligns the image and text representation at multiple semantic levels. Specifically, different from the prevalent single-stream and two-stream Transformer architectures, we use a shared Transformer network with a pluggable cross-modal attention module for both the low-level and high-level semantic alignments, as shown in Figure 2. For low-level semantic alignment, we directly concatenate image-region features and text features as input to the shared Transformer network for single-stream pre-training. For high-level semantic alignment, we introduce a novel two-stream Transformer network to align more abstract semantics by separately encoding the image and text parts with the shared Transformer, where a cross-modal attention module is further added to allow cross-modal fusion. The pre-training procedure is conducted iteratively to align the real-world image-text data at both semantic levels. During the iterative pre-training phase, the shared Transformer network is forced to align the semantics at multiple levels, which enables the trained model to adapt to diverse image-text pairs. In this way, we take advantages of both single-stream architecture and two-stream architecture for cross-modal fusion, where the parameters are shared to allow for different pre-training styles that regularize with each other.

To demonstrate the effectiveness of SemVLP, we evaluate it on various of vision-language tasks, (1) visual question answering (VQA 2.0 (Antol et al., 2015)), (2) natural language visual reasoning (NLVR2 (Suhr et al., 2018)), (3) visual reasoning in the real world (GQA (Hudson & Manning, 2019b)), and (4) image-text/text-image retrieval (Flickr30K (Young et al., 2014)). On all these tasks, SemVLP obtains significant improvements compared to those methods that align semantics at a single fixed level, where our 12-layer SemVLP model outperforms all the previous single-stream and two-stream architectures with the same model size.

The main contributions of this work can be summarized as follows: (i) We introduce SemVLP, a new VLP method to learn generic image-text representations for V+L understanding tasks. (ii) We propose a new pre-training framework that aligns cross-modal semantics at multiple levels, which can take advantages of both single-stream and two-stream architectures. (iii) We present extensive experiments and analysis to validate the effectiveness of the proposed SemVLP model, which can obtain superior performance with a 12-layer Transformer backbone on four V+L understanding tasks.

2 SemVLP Pre-training

In this section, we will first introduce the model architecture of SemVLP. Then we will illustrate how we align the low-level and high-level semantics with SemVLP model. Finally, the pre-training tasks and strategy will be introduced.

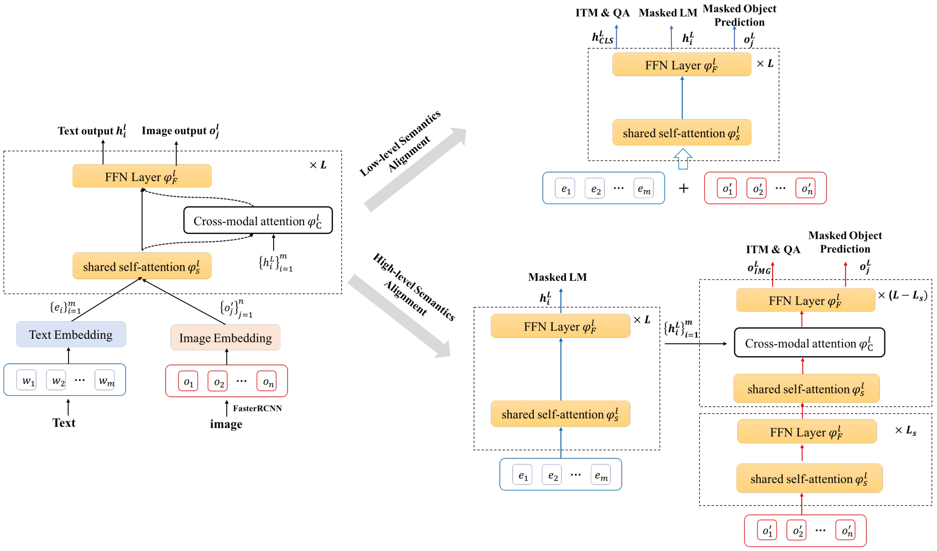


Figure 2: The overall framework of SemVLP. We use a shared Transformer network and a pluggable cross-modal attention module to support both the low-level and high-level semantic alignments in an iterative learning framework.

2.1 Model Architecture

The architecture overview of SemVLP is shown in Figure 2. Inspired by the idea of sharing the encoder and decoder in Transformers for neural machine translation (Xia et al., 2019), we base our architecture on a shared bidirectional Transformer encoder, where a pluggable cross-modal attention module is further used to align the semantics at multiple levels. By sharing the model parameters, SemVLP can be flexible to switch between single-stream and two-stream pre-training architectures, with the input text and image encoded in different semantic levels. In this way, we cast both the typical pre-training architectures into a more compact one in that there is only one copy of parameter set, which is applicable to both the low-level and high-level semantic alignment. We iteratively pre-train on the two settings towards better understanding of the real-world image-text pairs.

2.1.1 Input Embeddings

The input to our SemVLP model is an image and its related sentence (e.g. caption text). Each image is represented as a sequence of objects $\{o_1, ..., o_n\}$, and each sentence is represented as a sequence of words $\{w_1, ..., w_m\}$. After cross-modal fusion and alignment at multiple semantic levels, our SemVLP model is able to generate language representations, image representations and cross-modal representations from the image-text inputs. Given the sequence of words and objects, we first introduce the methods to embed the inputs to the feature space.

Sentence Embeddings We adopt the same method as BERT (Devlin et al., 2018), which uses WordPiece tokenizer to tokenize the input sentence into sub-word tokens. The sequence of input tokens is as $\{[CLS], w_1, ..., w_m, [SEP]\}$, where $[CLS]$ and $[SEP]$ are special tokens in BERT. The final embedding e_i for each token is generated by combining the original word embedding, segment embedding and position embedding.

Image Embeddings We use a pre-trained object detector Faster R-CNN (Ren et al., 2015) to extract the object-level image features from the image, where each object o_j is represented as a 2048-dimensional feature vector f_j. To capture the spatial information of the object, we also encode the box-level location features for each object via a 4-dimensional vector $l_j = (\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H})$, where (x_1, y_1) and (x_2, y_2) denote the coordinate of the bottom-left and top-right corner while W and H are the width and height of the input image. We concatenate f_j and l_j to form a position-sensitive object feature vector, which is further transformed into o_j' using a linear projection to ensure that it has the same vector dimension as that of word embeddings. Similar to special token $[CLS]$ in sentence embeddings, we also add a special feature $[IMG]$ to denote the representation of the entire image and add it to the beginning of the input object sequence.

2.2 Shared Transformer Encoder

Given the embeddings of the words for the sentence $\{e_i\}_{i=1}^{m}$ and the image regions $\{o'_j\}_{j=1}^{n}$, we use a shared bidirectional Transformer encoder with a pluggable cross-modal attention module to better learn the cross-modal representations at multiple semantic levels, as shown in Figure 2. The full encoder is a stacked model with L blocks, where the l'th block consists of a self-attention module φ_S^l, a nonlinear function φ_F^l and a pluggable cross-modal attention module φ_C^l, where superscript l represents the layer id. Next we will introduce the details of our method by combining different encoder modules to align the cross-modal representations at multiple semantic levels.

2.2.1 Low-level Semantic Alignment

To align low-level semantics, we directly concatenate the image and text embedding features as input to the single-stream mode of SemVLP, which consists of the shared self-attention module and nonlinear FFN layer. Specifically, we initialize $S^0 = \{o'_1, o'_2, ..., o'_n, e_1, e_2, ..., e_m\}$. The encoding process can be formulated as follows:

$$\begin{aligned} s_i^l &= \varphi_F^l(\varphi_S^l(s_i^{l-1}, S^{l-1})) \\ S^l &= \{s_1^l, s_2^l, ..., s_{n+m}^l\} = \{o_1^l, o_2^l, ..., o_n^l, h_1^l, h_2^l, ..., h_m^l\} \end{aligned} \tag{1}$$

where $\{h_i^l\}$ and $\{o_j^l\}$ are the text and object representation of layer l, respectively. In this way, we can get full interaction between the image and text representations at a low-level embedding space. Eventually, we obtain $O^L = \{o_1^L, o_2^L, ..., o_n^L\}$ and $H^L = \{h_1^L, h_2^L, ..., h_m^L\}$, the representations of all the object outputs and text outputs of the last layer in the SemVLP encoder.

2.2.2 High-level Semantic Alignment

For high-level semantic alignment, we adopt the two-stream mode of SemVLP, where text and image objects are separately encoded first and then fuse at a high-level semantic space. It consists of the shared self-attention module, cross-modal attention module and nonlinear FFN layer. To make it possible to separately encode the text and image representations with the SemVLP model, we adopt a two-encoder architecture shown in the bottom-right part of Figure 2 by tying all the parameters of self-attention module and FFN layer of the text encoder and image encoder, where a cross-modal attention module is further used to fuse the cross-modal representations. Different from the previous Transformer encoder-decoder architecture which introduces the cross-attention module to all blocks of the decoder, we only introduce the cross-modal attention module at the upper parts of the blocks, so as to better fuse the cross-modal representations at high-level semantic space. Specifically, we initialize $H^0 = \{e_1, e_2, ..., e_m\}$ and $O^0 = \{o'_1, o'_2, ..., o'_n\}$. The encoding process of two-stream mode can be formulated as follows:

$$\begin{aligned} h_j^l &= \varphi_F^l(\varphi_S^l(h_j^{l-1}, H^{l-1})) \\ o_j^l &= \varphi_F^l(\varphi_S^l(o_j^{l-1}, O^{l-1})), \quad s.t. \quad l <= L_s \\ o_{j+1}^l &= \varphi_F^l(\varphi_C^l(\varphi_S^l(o_{j+1}^{l-1}, O^{l-1}), H^L)), \quad s.t. \quad l > L_s \end{aligned} \tag{2}$$

where L_s indicates the layer index that cross-modal attention is introduced. Eventually, we can obtain the output representations of image objects and text, $O^L = \{o_1^L, o_2^L, ..., o_n^L\}$ and $H^L = \{h_1^L, h_2^L, ..., h_m^L\}$. With O^L and H^L, we could use a simple network with a softmax layer to conduct the subsequent pre-training tasks.

2.3 Joint Training

In this section, we first introduce the pre-training tasks used in our method, then the training strategy in terms of different semantic alignments will be introduced.

2.3.1 PRE-TRAINING TASKS

We follow LXMERT (Tan & Bansal, 2019) and use three-types of pre-training tasks: i.e., language task, vision task and cross-modality tasks.

Masked LM Prediction The task setup is basically the same as in BERT (Devlin et al., 2018), we randomly mask 15% tokens in the text and the model is asked to predict these masked words with the output text representations H^L. For different pre-training modes, the masked words will be predicted either with the help of visual modality so as to resolve ambiguity (single-stream mode), or from text modality alone so as to increase task difficulty (two-stream mode).

Masked Object Prediction Similarly, we pretrain the vision side by randomly masking objects, i.e., the object features are masked with zeros. We randomly mask 15% image objects and ask the model to predict properties of these masked objects with the output object representations O^L. To capture more object-level semantics, we follow the object prediction task in LXMERT (Tan & Bansal, 2019) and perform two sub-tasks: ROI-Feature Regression and Detected Label Classification. We take the detected labels output by Faster R-CNN (Ren et al., 2015) as the ground-truth labels for prediction.

Image-Text Matching (ITM) The task setup is almost the same as in LXMERT (Tan & Bansal, 2019), that we randomly sample 50% mismatched image-text pairs and 50% matched pairs, and train an classifier to predict whether an image and a sentence match each other on the representation $\mathbf{h}^L_{CLS}$ (single-stream mode) and $\mathbf{o}^L_{IMG}$ (two-stream mode). One difference is that we do not enforce the masked LM prediction and Object Prediction loss when sampling a mismatched image-text pair.

Image Question Answering (QA) We also cast the image question answering task as a classification problem and pre-train the model with image QA data as in LXMERT (Tan & Bansal, 2019), which leads to a better cross-modality representation. We build the classifier on top of the representation $\mathbf{h}^L_{CLS}$ for single-stream mode and on that of $\mathbf{o}^L_{IMG}$ for two-stream mode.

2.3.2 PRE-TRAINING STRATEGY

For both low-level and high-level semantic alignments, SemVLP is pre-trained with multiple pre-training tasks and we add all these task losses with equal weights. To jointly align semantics at multiple levels, given a mini-batch of image-text pairs, 50% of the time we update the model with low-level semantic alignment, while 50% of the time we update it with high-level semantic alignment. In this way, for every update of SemVLP, the model is pre-trained at multiple semantic levels, so as to better model the diverse image-text data.

3 EXPERIMENTS

3.1 PRE-TRAINING SETUP

Pre-training Data We use the same in-domain data as in LXMERT (Tan & Bansal, 2019) for pre-training. It consists of the image caption data from MS COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017), and image question answering data from VQA v2.0 (Antol et al., 2015), GQA balanced version (Hudson & Manning, 2019b) and VG-QA (Zhu et al., 2016). The total amount of the dataset is 9.18M image-and-sentence pairs on 180K distinct images. Besides, we also use additional out-of-domain data from Conceptual Captions (Sharma et al., 2018) and SBU Captions (Ordonez et al., 2011) for model pre-training, which consists of about 4M image-text pairs on 4M images.

Implementation Details The maximum sequence length for the sentence is set as 20. We use Faster R-CNN (Ren et al., 2015) (with ResNet-101 backbone (He et al., 2016)) pre-trained on Visual Genome dataset (Krishna et al., 2017) to detect the objects and extract the region features. We consistently keep 100 objects for each image to maximize the pre-training compute utilization by avoiding padding. For the model architecture, we pre-train a 12-layer SemVLP-base model with hidden size of 768, where we initialize it with the parameters from StructBERT base model (Wang et al., 2019). We set $L_s = 6$, which obtains the best performances on the development set of the downstream tasks, at a proper semantic level for cross-modal fusion [1]. We train SemVLP model with a total batch size of 256 for 40 epochs on 4 V100 GPUs. The Adam optimizer with initial learning rate of 1e-4 and a learning rate linear decay schedule is utilized.

3.2 Downstream Tasks

3.2.1 Visual Question Answering (VQA)

The VQA task requires the model to answer natural language questions given an image. We conduct experiments on the widely-used VQA v2.0 dataset (Antol et al., 2015), which contains 204K images and 1.1M questions about these images. Following (Anderson et al., 2018), we treat VQA as a multi-label classification task, which picks the corresponding answer from a shared set consisting of 3,129 answers. We use the hidden state of h^L_{CLS} (single-stream mode) or o^L_{IMG} (two-stream mode) to map the representation into 3,129 possible answers with an additional MLP layer. The model is optimized with a binary cross-entropy loss on the soft target scores. We fine-tune SemVLP model on VQA training data for 3 epochs with a batch size of 32. We use the BERT Adam optimizer with an initial learning rate of 5e-5. At inference, we simply use a Softmax function for prediction, and we choose the architecture mode with better performance on development set for final evaluation.

3.2.2 Image-Text Retrieval

The image-text retrieval task consists of two sub-tasks: image retrieval and text retrieval, depending on which modality is used as the retrieval target. We conduct experiments on Flickr30K dataset (Young et al., 2014), which contains 31,000 images collected from Flickr website and each image has 5 captions. We follow the same split in (Lee et al., 2018) for training and evaluation.

During fine-tuning, we follow the method in UNITER (Chen et al., b) and formulate it as a ranking problem. We use the hidden state of h^L_{CLS} (single-stream mode) or o^L_{IMG} (two-stream mode) to compute the similarity scores for the sampled positive and negative pairs, and maximize the margin between them through triplet loss. We fine-tune our model with a batch size of 64 and a learning rate of 5e-5 for 4 epochs. Moreover, we also use hard negatives sampling method in (Chen et al., b) for further improving the performance.

3.2.3 Natural Language Visual Reasoning for Real (NLVR2)

NLVR2 (Suhr et al., 2018) is a challenging task for visual reasoning. The goal is to determine whether a natural language statement is true about a pair of images, where it consists of 86K/7K/7K data for training/development/test. Since each data example in NLVR2 has two natural images img_0, img_1 and one language statement s, we use SemVLP to encode the two image-statement pairs (img_0, s) and (img_1, s), then train a classifier based on the concatenation of the two outputs as in LXMERT (Tan & Bansal, 2019). We fine-tune SemVLP with a batch size of 32 and a learning rate of 5e-5 for 3 epochs.

3.2.4 Visual Reasoning in the Real World (GQA)

GQA is an image question answering task, which emphasize on the reasoning capability of the model to answer a question. We conduct experiments on the public GQA 2019 dataset (Hudson & Manning, 2019b). For each question, the model picks the proper answer from a shared set of 1,852 candidate answers. We follow the two-stage fine-tuning method in OSCAR Li et al. (2020), where SemVLP model is first fine-tuned on unbalanced “all-split” for 2 epochs, and then fine-tune on the “balanced-split” for 2 epochs with batch size of 32 and learning rate of 5e-6.

3.3 Results on Downstream Tasks

We compare our pre-training SemVLP-base model against other state-of-the-art single-stream and two-stream cross-modal pre-training models. To account for parameter efficiency, we also list two types of models with different model sizes: (i) the VLP models of similar size to BERT base. (ii)

[1] We only introduce the cross-modal attention from text space to image space due to the superior performance in our framework, where the modeling of vision modality is emphasized.

Models		VQA		IR-Flickr30K			TR-Flickr30K		
		Test-dev	Test-std	R@1	R@5	R@10	R@1	R@5	R@10
Single-stream	VisualBERT	70.80	71.00	-	-	-	-	-	-
	Unicoder-VL-base	-	-	71.50	90.90	94.90	86.20	96.30	99.00
	VLBERT-large	71.79	72.22	-	-	-	-	-	-
	UNITER-base	72.70	72.91	72.52	92.36	96.08	85.90	97.10	98.80
	UNITER-large	73.82	74.02	75.56	**94.08**	**96.76**	87.30	98.00	99.20
	OSCAR-base	73.16	73.61	-	-	-	-	-	-
	OSCAR-large	73.44	73.82	-	-	-	-	-	-
	PixelBERT-r50	71.35	71.42	59.80	85.50	91.60	75.70	94.70	97.10
	PixelBERT-x152	74.45	74.55	71.50	92.10	95.80	87.00	**98.90**	**99.50**
Two-stream	ViLBERT-base	70.55	70.92	58.20	84.90	91.52	-	-	-
	LXMERT	72.42	72.54	-	-	-	-	-	-
	ERNIE-ViL-base	72.62	72.85	74.44	92.72	95.94	86.70	97.80	99.00
	ERNIE-ViL-large	**74.75**	**74.93**	**76.70**	93.58	96.44	**88.10**	98.00	99.20
Our Model	SemVLP-base	74.52	74.68	74.10	92.43	96.12	86.40	97.30	99.00

Table 1: Evaluation Results on VQA and Flickr30K.

Models	LXMERT	MMN	12-in-1	NSM	OSCAR	SemVLP-base
		(Chen et al., a)	(Lu et al., 2020)	(Hudson & Manning, 2019a)		
Test-dev	60.00	-	-	-	61.58	**62.87**
Test-std	60.33	60.83	60.65	63.17	61.62	**63.62**

Table 2: Evaluation Results on GQA.

Models	VisualBERT	LXMERT	UNITER-base	UNITER-large	OSCAR-base	OSCAR-large	SemVLP-base
Dev	67.40	74.90	77.14	78.40	78.07	**79.12**	79.00
Test-P	67.00	74.50	77.87	79.50	78.36	**80.37**	79.55

Table 3: Evaluation Results on NLVR2.

the VLP models that have similar size to BERT large. The results on four downstream V+L tasks are shown in Table 1,2,3 respectively. We can see that the proposed SemVLP model using a 12-layer base backbone can achieve performances comparable to those of the previous state-of-the-art methods on almost all the tasks, and even outperform many large VLP models (e.g., VLBERT-large (Su et al., 2019) and UNITER-large (Chen et al., b)) on VQA and NLVR2 tasks. In all the VLP models of similar size to BERT base, our SemVLP model consistently outperforms other strong VLP base models (e.g., LXMERT (Tan & Bansal, 2019), UNITER-base (Chen et al., b)) on most tasks, and often by a significantly large margin. It demonstrates that the proposed SemVLP model is highly parameter-efficient, partially because it is pre-trained to align cross-modal semantics at multiple semantic levels, which makes the learning of semantic alignments more robust toward the diverse image-text pairs.

3.4 Pre-training on Different Semantic Levels

To validate the effectiveness of aligning cross-modal semantics at multiple levels, we conduct in-depth analysis on pre-training at different semantic levels with various architectures.

Analysis on Various Pre-training Architectures We first examine the importance of pre-training at multiple semantic levels by conducting ablation study on the pre-training fashions. Specifically, we pre-train the SemVLP model with only one type of semantic alignment each time and test the performance on the downstream tasks. All the pre-training settings are kept the same as in the original SemVLP pre-training. As shown in Table 4, both the low-level semantic alignment and high-level semantic alignment play important roles in pre-training the full SemVLP model, and removing each task will consistently decrease the final downstream task performance. The single-stream architecture is used to align low-level semantics, while the two-stream architecture helps align semantics at a higher-level semantic space. By iterative training with a shared set of Transformer parameters, the proposed SemVLP model can take the advantage of both the single-stream architecture and two-stream architecture towards more robust vision-language pre-training.

Models	VQA		GQA	NLVR2	
	Test-dev	Test-std	Test-dev	Dev	Test-P
Baseline 1 (only low level)	73.72	73.91	61.82	78.02	78.25
Baseline 2 (only high level)	73.48	73.68	61.68	77.81	78.06
SemVLP	**74.52**	**74.68**	**62.87**	**79.00**	**79.55**

Table 4: Ablation study on pre-training fashions. Baseline 1 and Baseline 2 denote pre-training with only low-level semantic alignment or with only high-level semantic alignment, respectively.

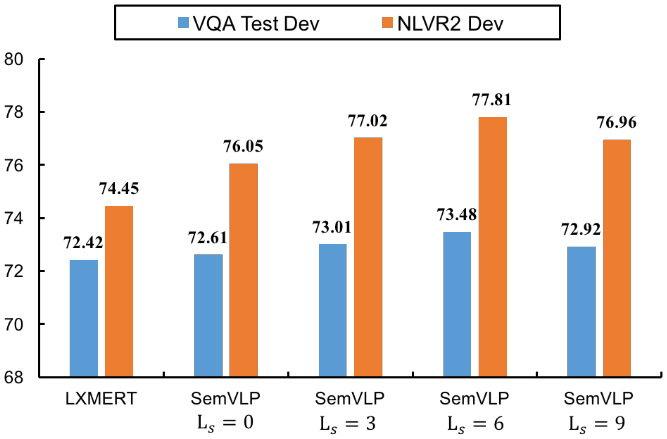


Figure 3: Results w.r.t different two-stream architectures for aligning high-level semantics on VQA and NLVR2 development set (best performance achieved at the semantic level of $L_s = 6$).

Analysis on Different High-level Semantic Alignments There are many different ways for high-level semantic alignment, now we further analyze the advantage of our architecture to conduct high-level semantic alignment and the specific "point" to conduct cross-modal fusion with the cross-modal attention module. Therefore, we pre-train the SemVLP model with only high-level semantic alignment and examine in which layer to introduce the cross-modal attention module by setting different L_s. The pre-training details are kept the same as in the original SemVLP pre-training. We test the performance on VQA and NLVR2 tasks, and the results are shown in Figure 3. We can see that by introducing the cross-modal attention module at proper layers, the two-stream mode of SemVLP method obtains significantly better performance than the previous two-stream model LXMERT. The best performance is obtained when $L_s = 6$, where the separated image/text encoding and cross-modal attention is equally emphasized. It again demonstrates the importance of aligning cross-modal representations at a proper semantic level.

4 Related Work

Pre-training methods have substantially advanced the performance of massive Computer Vision (CV) and Natural Language Processing (NLP) tasks. Pre-training methods in CV mainly employ ImageNet as generic feature representation or data augmentation as supervision information (Girshick et al., 2014; Long et al., 2015; Gidaris et al., 2018). Pre-training methods in NLP mainly rely on language model for text understanding (Devlin et al., 2018; Lan et al., 2019; Wang et al., 2019) and generation (Dong et al., 2019; Lewis et al., 2019).

Recently, researchers begin to focus on pre-training for vision-language Tasks. There are mainly two broad direction to learn cross-modal representations. The first line uses a single-stream transformer architecture to model both image and text representations (Li et al., 2019; Su et al., 2019), while the other method employs two-stream transformers to align the cross-modal representations (Lu et al., 2019; Tan & Bansal, 2019; Yu et al., 2020). Furthermore, other works focus on designing different pre-training tasks to learn better cross-modal representations (Chen et al., b; Yu et al., 2020).

5 Conclusion

In this paper, we propose a new pre-training method SemVLP to learn the joint representation of vision and language. Different from the existing VLP methods relying on a fixed-level semantic alignment, we introduce to align cross-modal semantics at multiple levels, by assembling a shared Transformer encoder and a pluggable cross-modal attention module in different ways. Experiment results on various downstream V+L tasks demonstrate the effectiveness of our method for understanding the diverse semantics behind the real-world image-text data.

References

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

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

Wenhu Chen, Zhe Gan, Linjie Li, Yu Cheng, William Wang, and Jingjing Liu. Meta module network for compositional visual reasoning. *arXiv preprint arXiv:1910.03230*, a.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. b.

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

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 7–16, 2014.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR 2018*, 2018.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

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

Drew Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. In *Advances in Neural Information Processing Systems*, pp. 5903–5916, 2019a.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019b.

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

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2019.

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

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

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

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. *arXiv preprint arXiv:2004.06165*, 2020.

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

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

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

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10437–10446, 2020.

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

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

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

Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, and Heng Tao Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pp. 785–796, 2013.

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

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

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

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. Structbert: Incorporating language structures into pre-training for deep language understanding. *arXiv preprint arXiv:1908.04577*, 2019.

Yanfei Wang, Fei Wu, Jun Song, Xi Li, and Yueting Zhuang. Multi-modal mutual topic reinforce modeling for cross-media retrieval. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 307–316, 2014.

Yingce Xia, Tianyu He, Xu Tan, Fei Tian, Di He, and Tao Qin. Tied transformers: Neural machine translation with shared encoder and decoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5466–5473, 2019.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 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, 2014.

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

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4995–5004, 2016.