Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks

Xiujun Li[♡♠], Xi Yin[♡], Chunyuan Li[♡], Pengchuan Zhang[♡], Xiaowei Hu[♡], Lei Zhang[♡], Lijuan Wang[♡], Houdong Hu[♡], Li Dong[♡], Furu Wei[♡], Yejin Choi[♠], and Jianfeng Gao[♡]

[♡]Microsoft Corporation [♠]University of Washington

Abstract. Large-scale pre-training methods of learning cross-modal representations on image-text pairs are becoming popular for vision-language tasks. While existing methods simply concatenate image region features and text features as input to the model to be pre-trained and use self-attention to learn image-text semantic alignments in a brute force manner, in this paper, we propose a new learning method Oscar[1], which uses object tags detected in images as *anchor points* to significantly ease the learning of alignments. Our method is motivated by the observation that the salient objects in an image can be accurately detected, and are often mentioned in the paired text. We pre-train an Oscar model on the public corpus of 6.5 million text-image pairs, and fine-tune it on downstream tasks, creating new state-of-the-arts on six well-established vision-language understanding and generation tasks.[2]

Keywords: Object Semantics, Vision-and-Language, Pre-training

1 Introduction

Learning cross-modal representations is fundamental to a wide range of vision-language (V+L) tasks, such as visual question answering, image-text retrieval, image captioning. Recent studies [21, 36, 5, 33, 19, 18, 41] on vision-language pre-training (VLP) have shown that it can effectively learn generic representations from massive image-text pairs, and that fine-tuning VLP models on task-specific data achieves state-of-the-art (SoTA) results on well-established V+L tasks.

These VLP models are based on multi-layer Transformers [37]. To pre-train such models, existing methods simply concatenate image region features and text features as input and resort to the self-attention mechanism to learn semantic alignments between image regions and text in a brute force manner. However, the lack of explicit alignment information between the image regions and text poses alignment modeling a weakly-supervised learning task. In addition, visual regions are often over-sampled [2], noisy and ambiguous, which makes the task even more challenging.

[1] **O**bject-**S**emanti**c**s **A**ligned P**r**e-training

[2] The code and pre-trained models are released: https://github.com/microsoft/Oscar

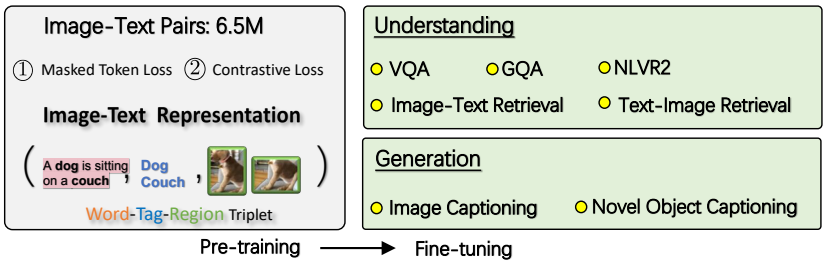


Fig. 1: OSCAR pipeline. The model takes a triple as input, is pre-trained with two losses (a masked token loss over words & tags, and a contrastive loss between tags and others), and fine-tuned for 5 understanding and 2 generation tasks (detailed in Sec. 4).

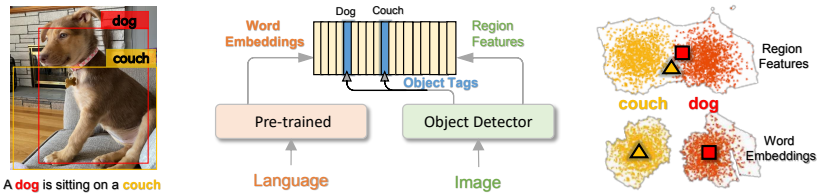
In this study, we show that the learning of cross-modal representations can be significantly improved by introducing object tags detected in images as *anchor points* to ease the learning of semantic alignments between images and texts. We propose a new VLP method OSCAR, where we define the training samples as triples, each consisting of a word sequence, a set of object tags, and a set of image region features. Our method is motivated by the observation that the salient objects in an image can be accurately detected by modern object detectors [27], and that these objects are often mentioned in the paired text. For example, on the MS COCO dataset [20], the percentages that an image and its paired text share at least 1, 2, 3 objects are 49.7%, 22.2%, 12.9%, respectively. Our OSCAR model is pre-trained on a large-scale V+L dataset composed of 6.5 million pairs, and is fine-tuned and evaluated on seven V+L understanding and generation tasks. The overall setting is illustrated in Fig 1.

Although the use of anchor points for alignment modeling has been explored in natural language processing *e.g.,* [3], to the best of our knowledge, this work is the first that explores the idea for VLP. There have been previous works that use object or image tags in V+L tasks for the sake of enhancing the feature representation of image regions, rather than for learning image-text alignments. For example, Zhou *et al.* [41] uses the object prediction probability as a soft label and concatenate it with its corresponding region features. Wu *et al.* [38] and You *et al.* [39] introduce image-level labels or attributes to improve image-level visual representations.

The main contributions of this work can be summarized as follows: (*i*) We introduce OSCAR, a powerful VLP method to learn generic image-text representations for V+L understanding and generation tasks. (*ii*) We have developed an OSCAR model that achieves new SoTA on multiple V+L benchmarks, outperforming existing approaches by a significant margin; (*iii*) We present extensive experiments and analysis to provide insights on the effectiveness of using object tags as anchor points for cross-modal representation learning and downstream tasks.

2 Background

The training data for many V+L tasks consists of image-text pairs, as shown in Fig. 2(a). We denote a dataset of size N by $\mathcal{D} = \{(\mathbf{I}_i, \boldsymbol{w}_i)\}_{i=1}^{N}$, with image $\mathbf{I}$ and



(a) Image-text pair (b) Objects as anchor points (c) Semantics spaces

Fig. 2: Illustration on the process that OSCAR represents an image-text pair into semantic space via dictionary look up. (a) An example of input image-text pair (b) The object tags are used as anchor points to align image regions with word embeddings of pre-trained language models. (c) The word semantic space is more representative than image region features. In this example, `dog` and `couch` are similar in the visual feature space due to the overlap regions, but distinctive in the word embedding space.

text sequence $\boldsymbol{w}$. The goal of pre-training is to learn cross-modal representations of image-text pairs in a self-supervised manner, which can be adapted to serve various down-stream tasks via fine-tuning.

VLP typically employs multi-layer self-attention Transformers [37] to learn cross-modal *contextualized* representations, based on the *singular* embedding of each modality. Hence, the success of VLP fundamentally relies on the quality of the input singular embeddings. Existing VLP methods take visual region features $\boldsymbol{v} = \{v_1, \cdots, v_K\}$ of an image and word embeddings $\boldsymbol{w} = \{w_1, \cdots, w_T\}$ of its paired text as input, and relies on the self-attention mechanism to learn image-text alignments and produce cross-modal contextual representations.

Though intuitive and effective, existing VLP methods suffer from two issues: (*i*) *Ambiguity*. The visual region features are usually extracted from over-sampled regions [2] via Faster R-CNN object detectors [27], which inevitably results in overlaps among image regions at different positions. This renders ambiguities for the extracted visual embeddings. For example, in Fig. 2(a) the region features for `dog` and `couch` are not easily distinguishable, as their regions heavily overlap. (*ii*) *Lack of grounding*. VLP is naturally a weakly-supervised learning problem because there is no explicitly labeled alignments between regions or objects in an image and words or phrases in text. However, we can see that salient objects such as `dog` and `couch` are presented in both image and its paired text as in Fig. 2(a), and can be used as anchor points for learning semantic alignments between image regions and textual units as in Fig. 2(b). In this paper we propose a new VLP method that utilizes these anchor points to address the aforementioned issues.

3 Oscar Pre-training

Humans perceive the world through many channels. Even though any individual channel might be incomplete or noisy, important factors are still perceivable since they tend to be shared among multiple channels (*e.g.,* `dog` can be described visually and verbally, as in Fig. 2). With this motivation, we propose a new

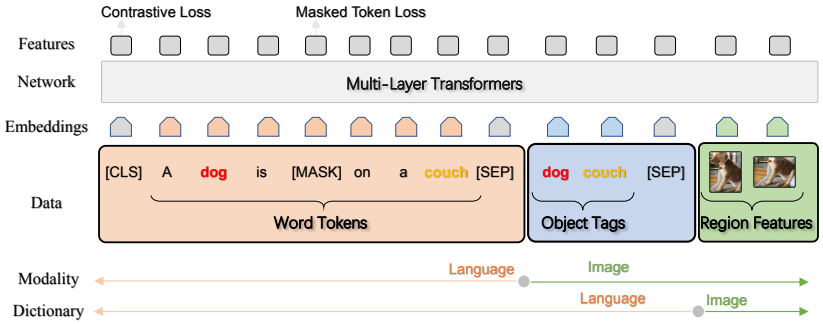


Fig. 3: Illustration of Oscar. We represent the image-text pair as a triple [word tokens , object tags , region features], where the object tags (*e.g.,* "`dog`" or "`couch`") are proposed to align the cross-domain semantics; when removed, Oscar reduces to previous VLP methods. The input triple can be understood from two perspectives: a *modality* view and a *dictionary* view.

VLP method Oscar to learn representations that capture channel-invariant (or modality-invariant) factors at the semantic level. Oscar differs from existing VLP in the way that the input image-text pairs are represented and the pre-training objective, as outlined in Fig. 3.

Input Oscar represents each input image-text pair as a Word-Tag-Image triple $(\boldsymbol{w}, \boldsymbol{q}, \boldsymbol{v})$, where $\boldsymbol{w}$ is the sequence of word embeddings of the text, $\boldsymbol{q}$ is the word embedding sequence of the object tags (in text) detected from the image, and $\boldsymbol{v}$ is the set of region vectors of the image.

Existing VLP methods represent each input pair as $(\boldsymbol{w}, \boldsymbol{v})$. Oscar introduces $\boldsymbol{q}$ as anchor points to ease the learning of image-text alignment. This is motivated by the observation that in training data, important objects in an image are often also *presented* in the image-paired text, using either the same words as object tags or different but semantically similar or related words. Since the alignments between $\boldsymbol{q}$ and $\boldsymbol{w}$, both in text, are relatively easy to identified by using pre-trained BERT models [6], which are used as initialization for VLP in Oscar, the image regions from which the object tags are detected are likely to have higher attention weights than other regions, when queried by the semantically related words in the text. This alignment learning process is conceptually illustrated in Fig. 2(b). The process can also be interpreted as learning to ground the image objects, which might be ambiguously represented in the vision space such as `dog` and `couch` in Fig. 2(a), in distinctive entities represented in the language space, as illustrated in Fig. 2(c).

Specifically, $\boldsymbol{v}$ and $\boldsymbol{q}$ are generated as follows. Given an image with K regions of objects (normally over-sampled and noisy), Faster R-CNN [27] is used to extract the visual semantics of each region as (v', z), where region feature $v' \in \mathbb{R}^P$ is a P-dimensional vector (*i.e.,* $P = 2048$), and region position z a R-dimensional

vector (*i.e.,* $R = 4$ or 6)[3]. We concatenate v' and z to form a position-sensitive region feature vector, which is further transformed into v using a linear projection to ensure that it has the same vector dimension as that of word embeddings. Meanwhile, the same Faster R-CNN is used to detect a set of high precision object tags. $\boldsymbol{q}$ is the sequence of word embeddings of the object tags.

Pre-Training Objective The OSCAR input can be viewed from two different perspectives as

$$\boldsymbol{x} \triangleq [\underbrace{\boldsymbol{w}}_{\text{language}}, \underbrace{\boldsymbol{q}, \boldsymbol{v}}_{\text{image}}] = [\underbrace{\boldsymbol{w}, \boldsymbol{q}}_{\text{language}}, \underbrace{\boldsymbol{v}}_{\text{image}}] \triangleq \boldsymbol{x}' \tag{1}$$

where $\boldsymbol{x}$ is the *modality* view to distinguish the representations between a text and an image; while $\boldsymbol{x}'$ is the *dictionary* view[4] to distinguish the two different semantic spaces, in which the input is represented. The two-view perspective allows us to design a novel pre-training objective.

A Dictionary View: Masked Token Loss. The use of different dictionaries determines the semantic spaces utilized to represent different sub-sequences. Specifically, the object tags and word tokens share the same linguistic semantic space, while the image region features lie in the visual semantic space. We define the *discrete token sequence* as $\boldsymbol{h} \triangleq [\boldsymbol{w}, \boldsymbol{q}]$, and apply the Masked Token Loss (MTL) for pre-training. At each iteration, we randomly mask each input token in $\boldsymbol{h}$ with probability 15%, and replace the masked one h_i with a special token [MASK]. The goal of training is to predict these masked tokens based on their surrounding tokens $\boldsymbol{h}_{\setminus i}$ and all image features $\boldsymbol{v}$ by minimizing the negative log-likelihood:

$$\mathcal{L}_{\text{MTL}} = -\mathbb{E}_{(\boldsymbol{v},\boldsymbol{h})\sim\mathcal{D}} \log p(h_i|\boldsymbol{h}_{\setminus i}, \boldsymbol{v}) \tag{2}$$

This is similar to masked language model used by BERT. The masked word or tag needs to be recovered from its surroundings, with additional image information attended to help ground the learned word embeddings in the vision context.

A Modality View: Contrastive Loss. For each input triple, we group $\boldsymbol{h}' \triangleq [\boldsymbol{q}, \boldsymbol{v}]$ to represent the image modality, and consider $\boldsymbol{w}$ as the language modality. We then sample a set of "polluted" image representations by replacing $\boldsymbol{q}$ with probability 50% with a different tag sequence randomly sampled from the dataset $\mathcal{D}$. Since the encoder output on the special token [CLS] is the fused vision-language representation of $(\boldsymbol{h}', \boldsymbol{w})$, we apply a fully-connected (FC) layer on the top of it as a binary classifier $f(.)$ to predict whether the pair contains the original

[3] It includes coordinates of top-left & bottom-right corners, and/or height & width.

[4] A semantic space can be viewed a vector space defined by a dictionary, which maps an input to a vector representation in the semantic space. For example, BERT can be viewed as a dictionary that defines a linguistic semantic space. BERT maps an input word or word sequence into a feature vector in the semantic space.

image representation ($y = 1$) or any polluted ones ($y = 0$). The contrastive loss is defined as

$$\mathcal{L}_{\mathrm{C}} = -\mathbb{E}_{(\boldsymbol{h}', \boldsymbol{w}) \sim \mathcal{D}} \log p(y | f(\boldsymbol{h}', \boldsymbol{w})). \tag{3}$$

During the cross-modal pre-training, we utilize object tags as the proxy of images to adjust the word embedding space of BERT, where a text is similar to its paired image (or more specifically, the object tags detected from the image), and dissimilar to the polluted ones.

The full pre-training objective of OSCAR is:

$$\mathcal{L}_{\text{Pre-training}} = \mathcal{L}_{\mathrm{MTL}} + \mathcal{L}_{\mathrm{C}}. \tag{4}$$

Discussion. Although other loss function designs can be considered as pre-training objectives, we perform experiments with these two losses for two reasons: (*i*) Each loss provides a representative learning signal from its own perspective. We deliberately keep a clear and simple form for the joint loss to study the effectiveness of the proposed dictionary and modality views, respectively. (*ii*) Though the overall loss is much simpler than those of existing VLP methods, it yields superior performance in our experiments.

Pre-training Corpus We have built the pre-training corpus based on the existing V+L datasets, including COCO [20], Conceptual Captions (CC) [30], SBU captions [25], flicker30k [40], GQA [12] *etc.*. In total, the unique image set is 4.1 million, and the corpus consists of 6.5 million text-tag-image triples. The detail is in Appendix.

Implementation Details We pre-train two model variants, denoted as OSCAR$_{\mathrm{B}}$ and OSCAR$_{\mathrm{L}}$, initialized with parameters $\boldsymbol{\theta}_{\mathrm{BERT}}$ of BERT base ($H = 768$) and large ($H = 1024$), respectively, where H is the hidden size. To ensure that the image region features have the same input embedding size as BERT, we transform the position-sensitive region features using a linear projection via matrix $\mathbf{W}$. The trainable parameters are $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{\mathrm{BERT}}, \mathbf{W}\}$. The AdamW Optimizer is used. OSCAR$_{\mathrm{B}}$ is trained for at least 1.0M steps, with learning rate $5e^{-5}$ and batch size 768. OSCAR$_{\mathrm{L}}$ is trained for at least 900k steps, with learning rate $1e^{-5}$ and batch size 512. The sequence length of discrete tokens $\boldsymbol{h}$ and region features $\boldsymbol{v}$ are 35 and 50, respectively.

4 Adapting to V+L Tasks

We adapt the pre-trained models to seven downstream V+L tasks, including five understanding tasks and two generation tasks. Each task poses different challenges for adaptation. We introduce the tasks and our fine-tuning strategy in this section, and leave the detailed description of datasets and evaluation metrics to Appendix.

Image-Text Retrieval heavily relies on the joint representations. There are two sub-tasks: *image retrieval* and *text retrieval*, depending on which modality is used as the retrieved target. During training, we formulate it as a binary classification problem. Given an aligned image-text pair, we randomly select a different image or a different caption to form an unaligned pair. The final representation of [CLS] is used as the input to the classifier to predict whether the given pair is aligned or not. We did not use ranking losses [13, 17], as we found that the binary classification loss works better, similarly as reported in [26]. In the testing stage, the probability score is used to rank the given image-text pairs of a query. Following [18], we report the top-K retrieval results on both the 1K and 5K COCO test sets.

Image Captioning requires the model to generate a natural language description of the content of an image. To enable sentence generation, we fine-tune OSCAR using the seq2seq objective. The input samples are processed to triples consisting of image region features, captions, and object tags, in the same way as that during the pre-training. We randomly mask out 15% of the caption tokens and use the corresponding output representations to perform classification to predict the token ids. Similar to VLP [41], the self-attention mask is constrained such that a caption token can only attend to the tokens before its position to simulate a uni-directional generation process. Note that all caption tokens will have full attentions to image regions and object tags but not the other way around.

During inference, we first encode the image regions, object tags, and a special token [CLS] as input. Then the model starts the generation by feeding in a [MASK] token and sampling a token from the vocabulary based on the likelihood output. Next, the [MASK] token in the previous input sequence is replaced with the sampled token and a new [MASK] is appended for the next word prediction. The generation process terminates when the model outputs the [STOP] token. We use beam search (*i.e.,* beam size = 5) [2] in our experiments and report our results on the COCO image captioning dataset.

Novel Object Captioning (NoCaps) [1] extends the image captioning task, and provides a benchmark with images from the Open Images dataset [16] to test models' capability of describing novel objects which are not seen in the training corpus. Following the restriction guideline of NoCaps, we use the predicted Visual Genome and Open Images labels to form tag sequences, and train OSCAR on COCO without the initialization of pre-training.

VQA [8] requires the model to answer natural language questions based on an image. Given an image and a question, the task is to select the correct answer from a multi-choice list. Here we conduct experiments on the widely-used VQA v2.0 dataset [8], which is built based on the MSCOCO [20] image corpus. The dataset is split into training (83k images and 444k questions), validation (41k images and 214k questions), and test (81k images and 448k questions) sets. Following [2], for each question, the model picks the corresponding answer from a shared set consisting of 3,129 answers.

When fine-tuning on the VQA task, we construct one input sequence, which contains the concatenation of a given question, object tags and region features, and then the [CLS] output from OSCAR is fed to a task-specific linear classifier for answer prediction. We treat VQA as a multi-label classification problem [2] – assigning a soft target score to each answer based on its relevancy to the human answer responses, and then we fine-tune the model by minimizing the cross-entropy loss computed using the predicted scores and the soft target scores. At inference, we simply use a Softmax function for prediction.

GQA [12] is similar to VQA, except that GQA tests the reasoning capability of the model to answer a question. We conduct experiments on the public GQA dataset [12]. For each question, the model chooses an answer from a shared set of $1,852$ candidate answers. We develop two fine-tuned models using OSCAR$_B$. One is similar to that of VQA. The other, denoted as OSCAR$^*_B$ in Table 2(d), is first fine-tuned on unbalanced "all-split" for 5 epochs, and then fine-tuned on the "balanced-split" for 2 epochs, as suggested in [4].

Natural Language Visual Reasoning for Real (NLVR2) [34] takes a pair of images and a natural language, and the goal is to determine whether the natural language statement is true about the image pair. When fine-tuning on the NLVR2 task, we first construct two input sequences, each containing the concatenation of the given sentence (the natural language description) and one image, and then two [CLS] outputs from OSCAR are concatenated as the joint input for a binary classifier, implemented by an MLP[5].

5 Experimental Results & Analysis

5.1 Performance Comparison with SoTA

To account for parameter efficiency, we compare OSCAR against three types of SoTA's: (*i*) SoTA$_S$ indicates the best performance achieved by small models prior to the Transformer-based VLP models. (*ii*) SoTA$_B$ indicates the best performance achieved by VLP models of similar size to BERT base. (*iii*) SoTA$_L$ indicates the best performance yielded by models that have a similar size to BERT large. To the best of our knowledge, UNITER [5] is the only model of BERT large size.

Table 1 summarizes the overall results on all tasks[6]. For all the tables in this paper, **Blue** indicates the best result for a task, and gray background indicates results produced by OSCAR. As shown in the table, our base model outperforms previous large models on most tasks, often by a significantly large margin. It demonstrates that the proposed OSCAR is highly parameter-efficient, partially because the use of object tags as anchor points significantly eases the learning of semantic alignments between images and texts. Note that OSCAR is pre-trained

[5] This is not necessarily the best fine-tuning choice for NLVR2, please refer to the *Pair-biattn* finetuning in UNITER [5] for a better choice, which introduces a multi-head attention layer to look back the concatenated text-image sequences.

[6] All the (single-model) SoTAs are from the published results.

Table 1: Overall results on six tasks. Δ indicates the improvement over SoTA. SoTA with subscript S, B, L indicates performance achieved by small models, VLP of similar size to BERT base and large model, respectively. Most results are from [5], except that image captioning results are from[10, 41], NoCaps results are from [1], VQA results are from [36].

Task	Image Retrieval			Text Retrieval			Image Captioning				NoCaps		VQA	NLVR2
	R@1	R@5	R@10	R@1	R@5	R@10	B@4	M	C	S	C	S	test-std	test-P
SoTA_S	39.2	68.0	81.3	56.6	84.5	92.0	38.9	29.2	129.8	22.4	61.5	9.2	70.90	53.50
SoTA_B	48.4	76.7	85.9	63.3	87.0	93.1	39.5	29.3	129.3	23.2	73.1	11.2	72.54	78.87
SoTA_L	51.7	78.4	86.9	66.6	89.4	94.3	–	–	–	–	–	–	73.40	79.50
OSCAR_B	**54.0**	**80.8**	**88.5**	**70.0**	**91.1**	**95.5**	**40.5**	**29.7**	**137.6**	22.8	**78.8**	**11.7**	**73.44**	78.36
OSCAR_L	**57.5**	**82.8**	**89.8**	**73.5**	**92.2**	**96.0**	**41.7**	**30.6**	**140.0**	**24.5**	**80.9**	**11.3**	**73.82**	**80.37**
Δ	**5.8** ↑	**4.4** ↑	**2.9** ↑	**6.9** ↑	**2.8** ↑	**1.7** ↑	**2.2** ↑	**1.3** ↑	**10.7** ↑	**1.3** ↑	**7.8** ↑	**0.5** ↑	**0.42** ↑	**0.87** ↑

on 6.5 million pairs, which is less than 9.6 million pairs used for UNITER pre-training and 9.18 million pairs for LXMERT.

We report the detailed comparison on each task in Table 2. (*i*) VLP methods dominate empirical performance across many V+L tasks, compared with small models. OSCAR outperforms all existing VLP methods on all seven tasks, and achieves new SoTA on six of them. On GQA, neural state machine (NSM) [11] relies on a strong structural prior, which can also be incorporated into OSCAR for improvement in the future. (*ii*) 12-in-1 is a recently proposed multi-task learning model [22] for V+L, implemented on BERT base. We see that OSCAR_B outperforms 12-in-1 on almost all the tasks, except on Test-P of NLVR2. Given that our method is based on single task fine-tuning, the result demonstrates the effectiveness of our proposed pre-training scheme. (*iii*) overall, OSCAR is the best performer on both understanding and generation tasks. On the captioning task, we further fine-tune OSCAR with self-critical sequence training (SCST) [29] to improve sequence-level learning. The only comparable VLP method for captioning is [41]. The results in Table 2 (e) show that OSCAR yields a much better performance, *e.g.,* improving BLEU@4 and CIDEr by more than 2 and 10 points, respectively. (*iv*) The NoCaps guideline requires to only use the COCO captioning training set. Hence, we initialize with BERT, and train OSCAR on the COCO training set. Constrained beam search (CBS) is used. The results in Table 2 (f) show that the variants of OSCAR consistently outperform the previous SoTA method UpDown [1]. The gap is much larger on the near-domain or out-of-domain cases, demonstrating the strong generalization ability of OSCAR.

5.2 Qualitative Studies

We visualize the learned semantic feature space of image-text pairs of the COCO test set on a 2D map using *t*-SNE [23]. For each image region and word token, we pass it through the model, and use its last-layer output as features. Pre-trained models with and without object tags are compared. The results in Fig 4 reveal some interesting findings. (*i*) *Intra-class.* With the aid of object tags, the distance of the same object between two modalities is substantially reduced. For

Table 2: Detailed results on V+L tasks.

Method	Size	Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
		1K Test Set						5K Test Set					
DVSA [14]	-	38.4	69.9	80.5	27.4	60.2	74.8	-	-	-	-	-	-
VSE++ [7]	-	64.7	-	95.9	52.0	-	92.0	41.3	-	81.2	30.3	-	72.4
DPC [46]	-	65.6	89.8	95.5	47.1	79.9	90.0	41.2	70.5	81.1	25.3	53.4	66.4
CAMP [42]	-	72.3	94.8	98.3	58.5	87.9	95.0	50.1	82.1	89.7	39.0	68.9	80.2
SCAN [18]	-	72.7	94.8	98.4	58.8	88.4	94.8	50.4	82.2	90.0	38.6	69.3	80.4
SCG [33]	-	76.6	96.3	99.2	61.4	88.9	95.1	56.6	84.5	92.0	39.2	68.0	81.3
PFAN [41]	-	76.5	96.3	99.0	61.6	89.6	95.2	-	-	-	-	-	-
Unicoder-VL [19]	B	84.3	97.3	99.3	69.7	93.5	97.2	62.3	87.1	92.8	46.7	76.0	85.3
12-in-1 [24]	B	-	-	-	65.2	91.0	96.2	-	-	-	-	-	-
UNITER [5]	B	-	-	-	-	-	-	63.3	87.0	93.1	48.4	76.7	85.9
UNITER [5]	L	-	-	-	-	-	-	66.6	89.4	94.3	51.7	78.4	86.9
OSCAR	B	88.4	**99.1**	**99.8**	75.7	95.2	98.3	70.0	91.1	95.5	54.0	80.8	88.5
	L	**89.8**	98.8	99.7	**78.2**	**95.8**	**98.3**	**73.5**	**92.2**	**96.0**	**57.5**	**82.8**	**89.8**

(a) Image-text retrieval

Method	ViLBERT	VL-BERT	VisualBERT	LXMERT	12-in-1	UNITER_B	UNITER_L	OSCAR_B	OSCAR_L
Test-dev	70.63	70.50	70.80	72.42	73.15	72.27	**73.24**	73.16	**73.61**
Test-std	70.92	70.83	71.00	72.54	–	72.46	73.40	**73.44**	**73.82**

(b) VQA

Method	MAC	VisualBERT	LXMERT	12-in-1	UNITER_B	UNITER_L	OSCAR_B	OSCAR_L
Dev	50.8	67.40	74.90	–	77.14	**78.40**	78.07	**79.12**
Test-P	51.4	67.00	74.50	78.87	77.87	**79.50**	78.36	**80.37**

(c) NLVR2

Method	Test-dev	Test-std
LXMERT [39]	60.00	60.33
MMN [4]	–	60.83
12-in-1 [24]	–	60.65
NSM [12]	–	63.17
OSCAR_B	61.19	61.23
OSCAR_B*	**61.58**	**61.62**

(d) GQA

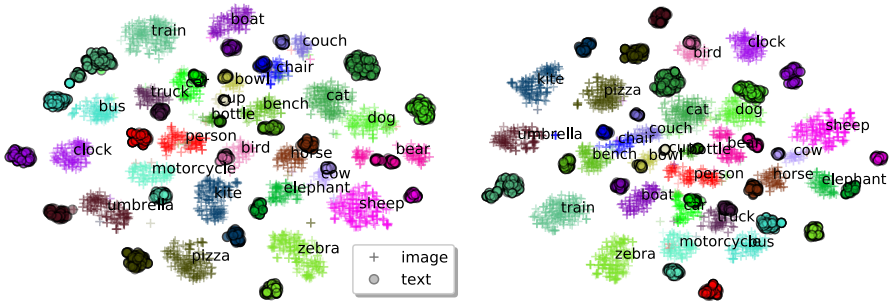
Method	cross-entropy optimization				CIDEr optimization			
	B@4	M	C	S	B@4	M	C	S
BUTD [2]	36.2	27.0	113.5	20.3	36.3	27.7	120.1	21.4
VLP [47]	36.5	28.4	117.7	21.3	39.5	29.3	129.3	23.2
AoANet [11]	37.2	28.4	119.8	21.3	38.9	29.2	129.8	22.4
OSCAR_B	36.5	**30.3**	**123.7**	**23.1**	**40.5**	**29.7**	**137.6**	**22.8**
OSCAR_L	**37.4**	**30.7**	**127.8**	**23.5**	**41.7**	**30.6**	**140.0**	**24.5**

(e) Image captioning on COCO

Method	in-domain		near-domain		out-of-domain		overall	
	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE	CIDEr	SPICE
UpDown [1]	78.1	11.6	57.7	10.3	31.3	8.3	55.3	10.1
UpDown + CBS [1]	80.0	12.0	73.6	11.3	66.4	9.7	73.1	11.1
UpDown + ELMo + CBS [1]	79.3	**12.4**	73.8	11.4	71.7	9.9	74.3	11.2
OSCAR_B	79.6	12.3	66.1	11.5	45.3	9.7	63.8	11.2
OSCAR_B + CBS	80.0	12.1	80.4	**12.2**	75.3	**10.6**	79.3	**11.9**
OSCAR_B + SCST + CBS	**83.4**	12.0	**81.6**	12.0	**77.6**	**10.6**	**81.1**	11.7
OSCAR_L	79.9	**12.4**	68.2	11.8	45.1	9.4	65.2	11.4
OSCAR_L + CBS	78.8	12.2	78.9	**12.1**	77.4	10.5	78.6	**11.8**
OSCAR_L + SCST + CBS	**85.4**	11.9	**84.0**	11.7	**80.3**	10.0	**83.4**	11.4

(f) Evaluation on NoCaps Val. Models are trained on COCO only without pre-training.

example, the visual and textual representations for `person` (or `zebra`) in OSCAR is much closer than that in the baseline method. (*ii*) *Inter-class.* Object classes of related semantics are getting closer (but still distinguishable) after adding tags, while there are some mixtures in the baseline, such as animal (`person`, `zebra`, `sheep`, `bird`), furniture (`chair`, `couch`, `bench`), and transportation (`bus`,



(a) OSCAR (b) Baseline (No tags)

Fig. 4: 2D visualization using t-SNE. The points from the same object class share the same color. Please refer Appendix for full visualization.

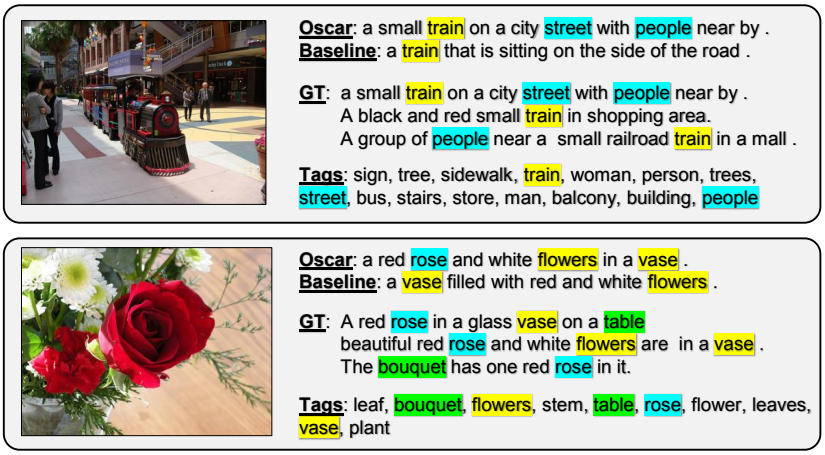


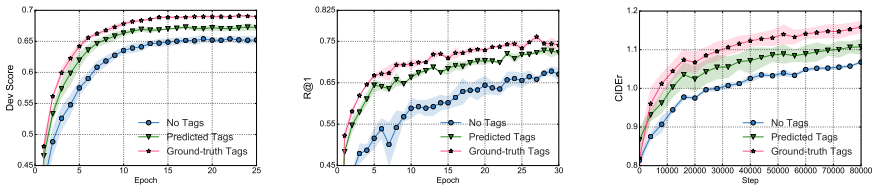
Fig. 5: Examples of image captioning. Objects are colored, based on their appearance against the groud-truth (GT): all, OSCAR & tags, tags only.

`train`, `truck`, `motorcycle`, `car`). This verifies the importance of object tags in alignment learning: it plays the role of anchor points in linking and regularizing the cross-modal feature learning.

We compare generated captions of different models in Fig. 5. The baseline method is VLP without object tags. We see that OSCAR generates more detailed descriptions of images than the baseline, due to the use of the accurate and diverse object tags detected by Faster R-CNN. They are the anchor points in the word embedding space, guiding the text generation process.

5.3 Ablation Analysis

We perform ablation experiments over a number of design choices of OSCAR in both pre-training and fine-tuning to better understand their relative impor-



(a) VQA (b) Image Retrieval R@1 (c) Image Captioning

Fig. 6: The learning curves of fine-tuning downstream tasks with different object tags. Each curve is with 3 runs.

tance to four representative downstream tasks. All the ablation experiments are conducted on the base model.

The Effect of Object Tags To study the effect of object tags, we experiment three different settings: (*i*) *Baseline (No Tags)*: this reduces the models to their previous VLP counterparts, where no tag information is exploited. (*ii*) *Predicted Tags*: we use an off-the-shelf object detector (trained on COCO dataset) to predict object tags. (*iii*) *Ground-truth Tags*: The ground-truth tags from COCO dataset are utilized to serve as a performance "upper bound" for our method. The experiments are conducted with the same BERT base model on three representative tasks, including VQA, image retrieval, and image captioning. As shown in Fig. 6, the learning curves for fine-tuning with object tags converges significantly faster and better than the VLP method without tags on all tasks. On the VQA and retrieval tasks, training using tags only takes half of the training time to achieve the final performance of the baseline, showing that Oscar is a more practical and efficient scheme for VLP. With more accurate object detectors developed in the future, Oscar can achieve even better performance, closing the gap demonstrated by using the ground-truth tags.

Attention Interaction To further understand the interaction among the text, object tags and object regions, we conduct fine-tuning experiments by varying the attention masks for image-text retrieval. The default setting uses full attentions across all modalities. We then enable certain part of the attention masks. All models are initialized from BERT base without pre-training. Table 3 reports the performance on the COCO 1K test set. By comparing the results of using full attention and partial attention $\boldsymbol{w}$-$\boldsymbol{v}$, we see that it is beneficial to add object tags. Moreover, region features are more informative than object tags ($\boldsymbol{w}$-$\boldsymbol{v}$, vs. $\boldsymbol{v}$-$\boldsymbol{q}$) in representing an image. This suggests that tags yield minor improvement when used as features; a more promising way is to use them as anchor points, as done in Oscar.

Table 3: Retrieval results on the COCO 1K test set, with different types of attention interactions.

Attention			Text R.		Image R.	
$\boldsymbol{w}$-$\boldsymbol{v}$	$\boldsymbol{w}$-$\boldsymbol{q}$	$\boldsymbol{v}$-$\boldsymbol{q}$	R@1	R@5	R@1	R@5
✓	✓	✓	77.3	95.6	65.2	91.5
✓			75.4	94.8	64.2	91.4
	✓		32.3	57.6	25.7	60.1

Table 4: Results with various pre-training schemes.

Pre-train	VQA	Text Retrieval			Image Retrieval			Image Captioning			
	dev	R@1	R@5	R@10	R@1	R@5	R@10	B@4	M	C	S
BASELINE (NO TAGS)	70.93	84.4	98.1	99.5	73.1	94.5	97.9	34.5	29.1	115.6	21.9
OSCAR^{VG}	71.70	88.4	99.1	99.8	75.7	95.2	98.3	36.4	30.3	123.4	23.0
OSCAR^{OI}	71.15	85.9	97.9	99.5	72.9	94.3	97.6	35.3	29.6	119.5	22.6

Object Tags in Pre-training To study the impact of different object tag sets in pre-trained models, we pre-train two variants: OSCAR^{VG} and OSCAR^{OI} utilizes object tags produced by the object detector trained on the visual genome (VG) dataset [15] and the open images (OI) dataset [16], respectively. In this ablation, all the models are pre-trained for 589k steps. The results are shown in Table 4, where BASELINE (NO TAGS) is also listed for comparison. It is clear that the OSCAR scheme of using object tags as anchor points improves the baseline, regardless of which set of object tags is used. VG tags performs slightly better than OI. We hypothesize that the object detector trained on VG has a more diverse set of objects, although the object detector trained on OI has a higher precision.

6 Related Work

Vision-Language Pre-training There is a growing interest in pre-training generic models to solve a variety of V+L problems, such as visual question-answering (VQA), image-text retrieval and image captioning *etc.* The existing methods [35, 36, 21, 5, 41, 33, 18, 9] employ BERT-like objectives [6] to learn cross-modal representations from a concatenated-sequence of visual region features and language token embeddings. They heavily rely on the self-attention mechanism of Transformers to learn joint representations that are appropriately contextualized in both modalities. For example, early efforts such as [21, 36] propose a two-stream and three-stream Transformer-based framework with co-attention to fuse the two modalities, respectively. Chen *et al.* [5] conduct comprehensive studies on the effects of different pre-training objectives for the learned generic representations. Zhou *et al.* [41] propose the first unified model to deal with both understanding and generation tasks, using only VQA and image captioning as the downstream tasks. In this paper, the OSCAR models have been applied to a wider range of downstream tasks, including both understanding and generation tasks, and have achieved new SoTA in most of them. Compared to existing VLP methods, the most salient difference of the proposed OSCAR is the use of object tags for aligning elements in two modalities. It alleviates the challenge of VLP models having to figure out the cross-modal semantic alignment from scratch, and thus improves the learning efficiency. In fact, our base model already outperforms the existing large VLP models on most V+L tasks.

Object Tags Anderson *et al.* [2] introduce the bottom-up mechanism to represent an image as a set of visual regions via Faster R-CNN [27], each with

an associated feature vector. It enables attention to be computed at the object level, and has quickly become the de facto standard for fine-grained image understanding tasks. In this paper, we propose to use object tags to align the object-region features in [2] in the pre-trained linguistic semantic space. The idea of utilizing object tags has been explored for image understanding [38, 39, 41]. Based on grid-wise region features of CNNs, Wu *et al.* [38] employ the predicted object tags only as the input to LSTM for image captioning, while You *et al.* [39] consider both tags and region features. Based on salient regions proposed by object detectors, Zhou *et al.* [41] concatenate the object prediction probability vector with region features as the visual input for VLP. Unfortunately, the tags in these works are not simultaneously associated with both object regions and word embeddings of text, resulting in a lack of grounding. Our construction of object tags with their corresponding region features & word embeddings yields more complete and informative representations for objects, particularly when the linguistic entity embeddings are pre-trained, as described next.

Multimodal Embeddings It has been shown that V+L tasks can benefit from a shared embedding space to align the inter-modal correspondences between images and text. Early attempts from Socher *et al.* [31] project words and image regions into a common space using kernelized canonical correlation analysis, and achieve good results for annotation and segmentation. Similar ideas are employed for image captioning [13] and text-based image retrieval [28]. In particular, the seminal work DeViSE [7] proposes to identify visual objects using semantic information gleaned from un-annotated text. This semantic information is exploited to make predictions of image labels that are not observed during training, and improves zero-shot predictions dramatically across thousands of novel labels that have never been seen by the vision model. The idea has been extended in [32, 14, 24], showing that leveraging pre-trained linguistic knowledge is highly effective for aligning semantics and improving sample efficiency in cross-modal transfer learning. Inspired by this line of research, we revisit the idea and propose to leverage the rich semantics from the learned word embeddings in the era of neural language model pre-training. Indeed, our results on novel object captioning demonstrate that Oscar helps improve the generalizability of the pre-trained models.

7 Conclusion

In this paper, we have presented a new pre-training method Oscar, which uses object tags as anchor points to align the image and language modalities in a shared semantic space. We validate the schema by pre-training Oscar models on a public corpus with 6.5 million text-image pairs. The pre-trained models archive new state-of-the-arts on six established V+L understanding and generation tasks.

References

1. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: nocaps: novel object captioning at scale. In: ICCV (2019)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018)
3. Brown, P.F., Lai, J.C., Mercer, R.L.: Aligning sentences in parallel corpora. In: Proceedings of the 29th annual meeting on Association for Computational Linguistics (1991)
4. Chen, W., Gan, Z., Li, L., Cheng, Y., Wang, W., Liu, J.: Meta module network for compositional visual reasoning. arXiv preprint arXiv:1910.03230 (2019)
5. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Learning universal image-text representations. arXiv preprint arXiv:1909.11740 (2019)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL (2019)
7. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: DeViSE: A deep visual-semantic embedding model. In: NeurIPS (2013)
8. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In: CVPR (2017)
9. Hao, W., Li, C., Li, X., Carin, L., Gao, J.: Towards learning a generic agent for vision-and-language navigation via pre-training. CVPR (2020)
10. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: ICCV (2019)
11. Hudson, D., Manning, C.D.: Learning by abstraction: The neural state machine. In: NeurIPS (2019)
12. Hudson, D.A., Manning, C.D.: GQA: A new dataset for real-world visual reasoning and compositional question answering. arXiv preprint arXiv:1902.09506 (2019)
13. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
14. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)
15. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)
16. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv preprint arXiv:1811.00982 (2018)
17. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: ECCV (2018)
18. Li, G., Duan, N., Fang, Y., Jiang, D., Zhou, M.: Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. arXiv preprint arXiv:1908.06066 (2019)
19. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)

20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014)
21. Lu, J., Batra, D., Parikh, D., Lee, S.: VilBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: NeurIPS (2019)
22. Lu, J., Goswami, V., Rohrbach, M., Parikh, D., Lee, S.: 12-in-1: Multi-Task vision and language representation learning. arXiv preprint arXiv:1912.02315 (2019)
23. Maaten, L.v.d., Hinton, G.: Visualizing data using t-SNE. Journal of machine learning research (2008)
24. Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G.S., Dean, J.: Zero-shot learning by convex combination of semantic embeddings. arXiv preprint arXiv:1312.5650 (2013)
25. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2text: Describing images using 1 million captioned photographs. In: NeurIPS (2011)
26. Qi, D., Su, L., Song, J., Cui, E., Bharti, T., Sacheti, A.: Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. arXiv preprint arXiv:2001.07966 (2020)
27. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015)
28. Ren, Z., Jin, H., Lin, Z., Fang, C., Yuille, A.: Joint image-text representation by gaussian visual-semantic embedding. In: Multimedia (2016)
29. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: CVPR (2017)
30. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Annual Meeting of the Association for Computational Linguistics (2018)
31. Socher, R., Fei-Fei, L.: Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In: CVPR (2010)
32. Socher, R., Ganjoo, M., Manning, C.D., Ng, A.: Zero-shot learning through cross-modal transfer. In: NeurIPS (2013)
33. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: VL-BERT: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530 (2019)
34. Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., Artzi, Y.: A corpus for reasoning about natural language grounded in photographs. arXiv preprint arXiv:1811.00491 (2018)
35. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: VideoBERT: A joint model for video and language representation learning. ICCV (2019)
36. Tan, H., Bansal, M.: LXMERT: Learning cross-modality encoder representations from transformers. EMNLP (2019)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
38. Wu, Q., Shen, C., Liu, L., Dick, A., Van Den Hengel, A.: What value do explicit high level concepts have in vision to language problems? In: CVPR (2016)
39. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: CVPR (2016)
40. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: 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)
41. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J.J., Gao, J.: Unified vision-language pre-training for image captioning and VQA. AAAI (2020)