

Unified Visual-Semantic Embeddings: Bridging Vision and Language with Structured Meaning Representations

Hao Wu^{1,3,4,6,*†}, Jiayuan Mao^{5,6,*†}, Yufeng Zhang^{2,6,†}, Yuning Jiang⁶, Lei Li⁶, Weiwei Sun^{1,3,4}, Wei-Ying Ma⁶

¹School of Computer Science, ²School of Economics, Fudan University

³Systems and Shanghai Key Laboratory of Data Science, Fudan University

⁴Shanghai Institute of Intelligent Electronics & Systems

⁵ITCS, Institute for Interdisciplinary Information Sciences, Tsinghua University, ⁶Bytedance AI Lab

{wuha05688, zhangyf, wwsun}@fudan.edu.cn, mjj14@mails.tsinghua.edu.cn,

{jiangyuning, lileilab, maweiyiying}@bytedance.com

Abstract

We propose the Unified Visual-Semantic Embeddings (Unified VSE) for learning a joint space of visual representation and textual semantics. The model unifies the embeddings of concepts at different levels: objects, attributes, relations, and full scenes. We view the sentential semantics as a combination of different semantic components such as objects and relations; their embeddings are aligned with different image regions. A contrastive learning approach is proposed for the effective learning of this fine-grained alignment from only image-caption pairs. We also present a simple yet effective approach that enforces the coverage of caption embeddings on the semantic components that appear in the sentence. We demonstrate that the Unified VSE outperforms baselines on cross-modal retrieval tasks; the enforcement of the semantic coverage improves the model’s robustness in defending text-domain adversarial attacks. Moreover, our model empowers the use of visual cues to accurately resolve word dependencies in novel sentences.

1. Introduction

We study the problem of establishing accurate and generalizable alignments between visual concepts and textual semantics efficiently, based upon rich but few, paired but noisy, or even biased visual-textual inputs (*e.g.*, image-caption pairs). Consider the image-caption pair A shown in Fig. 1: “A white clock on the wall is above a wooden table”. The alignments are formed at multiple levels: This short sentence can be decomposed into a rich set of semantic components [3]:

*indicates equal contribution.

†Work was done when HW, JM and YZ were intern researchers at the Bytedance AI Lab.

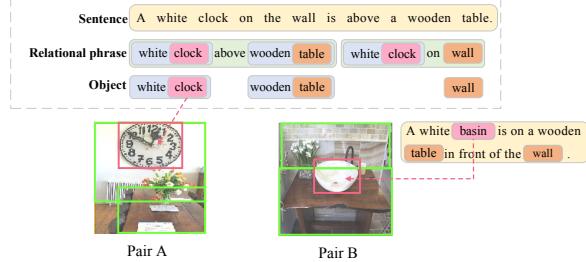


Figure 1. Two exemplar image-caption pairs. Humans are able to establish accurate and generalizable alignments between vision and language, at different levels: objects, relations and full sentences. Pair A and B form a pair of contrastive example for the concepts clock and basin.

objects (clock, table and wall) and relations (clock above table, and clock on wall). These components are linked with different parts of the scene.

This motivates our work to introduce *Unified Visual-Semantic Embeddings* (Unified VSE for short). Shown in Fig. 2, Unified VSE bridges visual and textual representation in a joint embedding space that unifies the embeddings for objects (noun phrases vs. visual objects), attributes (prenominal phrases vs. visual attributes), relations (verbs or prepositional phrases vs. visual relations) and scenes (sentence vs. image).

There are two major challenges in establishing such a factorized alignment. First, the link between the textual description of an object and the corresponding image region is ambiguous: A visual scene consists of multiple objects, and thus it is unclear to the learner which object should be aligned with the description. Second, it could be problematic to directly learn a neural network that combines various semantic components in a caption and form an encoding for the full sentence, with the training objective to maximize the

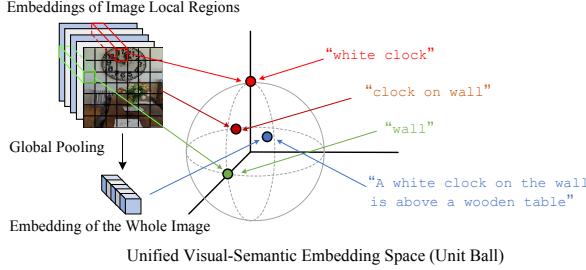


Figure 2. We build a visual-semantic embedding space, which unifies the embeddings for objects, attributes, relations and full scenes.

cross-modal retrieval performance in the training set (*e.g.*, in [49, 30, 40]). As reported by [40], because of the inevitable bias in the dataset (*e.g.*, two objects may co-occur with each other in most cases, see the table and the wall in Fig. 1 as an example), the learned sentence encoders usually pay attention to only part of the sentence. As a result, they are vulnerable to text-domain adversarial attacks: Adversarial captions constructed from original captions by adding small perturbations (*e.g.*, by changing *wall* to be *shelf*) can easily fool the model [40, 39].

We resolve the aforementioned challenges by a natural combination of two ideas: *cross-situational learning* and the enforcement of *semantic coverage* that regularizes the encoder. Cross-situational learning, or learning from contrastive examples [12], uses contrastive examples in the dataset to resolve the referential ambiguity of objects: Looking at both Pair A and B in Fig. 1, we know that *Clock* should refer to an object that occurs only in scene A but not B. Meanwhile, to alleviate the biases of datasets such as object co-occurrence, we present an effective approach that enforces the *semantic coverage*: The meaning of a caption is a composition of all semantic components in the sentence [3]. Reflectively, the embedding of a caption should have a coverage of all semantic components, while changing any of them should affect the global caption embedding.

Conceptually and empirically, Unified VSE makes the following three contributions.

First, the explicit factorization of the visual-semantic embedding space enables us to build a fine-grained correspondence between visual and textual data, which further benefits a set of downstream visual-textual tasks. We achieve this through a contrastive example mining technique that uniformly applies to different semantic components, in contrast to the sentence or image-level contrastive samples used by existing visual-semantic learning [49, 30, 11]. Unified VSE consistently outperforms pre-existing approaches on a diverse set of retrieval-based tasks.

Second, we propose a caption encoder that ensures a coverage of all semantic components appeared in the sentence. We show that this regularization helps our model to learn a robust semantic representation for captions. It effectively defends adversarial attacks on the text domain.

Furthermore, we show how our learned embeddings can provide visual cues to assist the parsing of novel sentences, including determining content word dependencies and labelling semantic roles for certain verbs. It ends up that our model can build reliable connections between vision and language using given semantic cues and in return, bootstrap the acquisition of language.

2. Related work

Visual semantic embedding. Visual semantic embedding [13] is a common technique for learning a joint representation of vision and language. The embedding space empowers a set of cross-modal tasks such as image captioning [43, 48, 8] and visual question answering [4, 47].

A fundamental technique proposed in [13] for aligning two modalities is to use the pairwise ranking to learn a distance metric from similar and dissimilar cross-modal pairs [44, 35, 23, 9, 28, 24]. As a representative, VSE++ [11] uses the online hard negative mining (OHEM) strategy [41] for data sampling and shows the performance gain. VSE-C [40], based on VSE++, enhances the robustness of the learned visual-semantic embeddings by incorporating rule-generated textual adversarial samples as hard negatives during training. In this paper, we present a contrastive learning approach based on semantic components.

There are multiple VSE approaches that also use linguistically-aware techniques for the sentence encoding and learning. Hierarchical multimodal LSTM (HM-LSTM) [33] and [46], as two examples, both leverage the constituency parsing tree. Multimodal-CNN (m-CNN) [30] and CSE [49] apply convolutional neural networks to the caption and extract the a hierarchical representation of sentences. Our model differs with them in two aspects. First, Unified VSE is built upon a factorized semantic space instead of the syntactic knowledge. Second, we employ a contrastive example mining approach that uniformly applies to different semantic components. It substantially improves the learned embeddings, while the related works use only sentence-level contrastive examples.

The learning of object-level alignment in unified VSE is also related to [19, 21, 36], where the authors incorporate pre-trained object detectors for the semantic alignment. [10] propose a selective pooling technique for the aggregation of object features. Compared with them, Unified VSE presents a more general approach that embeds concepts of different levels, while still requiring no extra supervisions.

Structured representation for vision and language. We connect visual and textual representations in a structured embedding space. The design of its structure is partially motivated by the papers on relational visual representations (scene graphs) [29, 18, 17], where a scene is represented by a set of objects and their relations. Compared with them, our

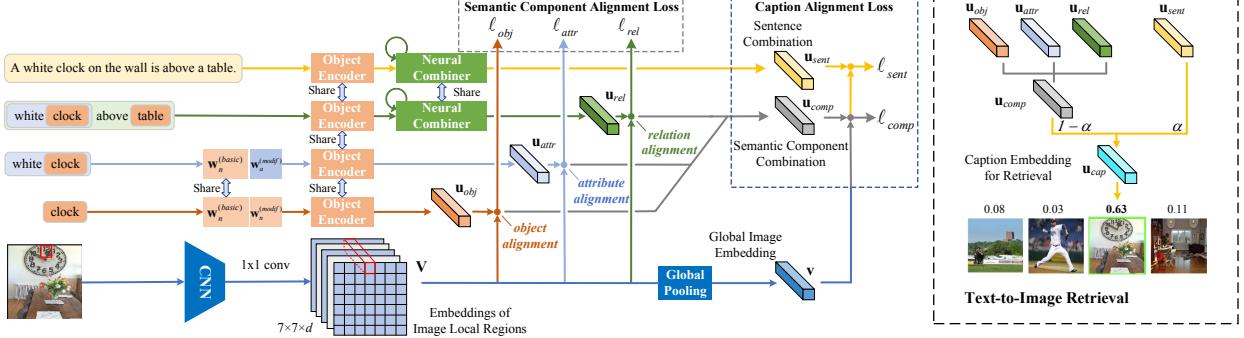


Figure 3. **Left:** the architecture of Unified VSE. The semantic component alignment is learned from contrastive examples sampled from factorized semantic space. The model also learns a caption encoder that combines the semantic components and aligns the caption with the corresponding image. **Right:** An exemplar computation graph for retrieving images from texts. The presence of \mathbf{u}_{comp} in the caption encoding enforces the coverage of all semantic components. See Sec. 3.2 for details.

model does not rely on labelled graphs during training.

Researchers have designed various types of representations [5, 32] as well as different models [26, 50] for translating natural language sentences into structured representations. In this paper, we present how the usage of such semantic parsing into visual-semantic embedding facilitates the learning of the embedding space. Moreover, we present how the learned VSE can, in return, helps the parser to resolve parsing ambiguities using visual cues.

3. Unified Visual-Semantic Embeddings

We now describe the overall architecture and training paradigm for the proposed *Unified Visual-Semantic Embeddings*. Shown in Fig. 3, given an image-caption pair, we first parse the caption into a structured meaning representation, composed by a set of semantic components: object nouns, prenominal modifiers, and relational dependencies. We encode different types of semantic components with type-specific encoders. A caption encoder combines the embedding of the semantic components into a caption semantic embedding. Jointly, we encode images with a convolutional neural network (CNN) into the same, *unified VSE* space. The distance between the image embedding and the sentential embedding measures the semantic similarity between the image and the caption.

We employ a multi-task learning approach for the joint learning of embeddings for semantic components (as the “basis” of the VSE space) as well as the caption encoder (as the combiner of semantic components).

3.1. Visual-Semantic Embedding: A Revisit

We begin the section with an introduction to the two-stream VSE approach. It jointly learns the embedding spaces of two modalities: vision and language, and aligns them using parallel image-text pairs (*e.g.*, image and captions from the MS-COCO dataset [27]).

Let $\mathbf{v} \in \mathbb{R}^d$ be the representation of the image and $\mathbf{u} \in \mathbb{R}^d$ be the representation of a caption matching this

image, both encoded by neural modules. To archive the alignment, a bidirectional margin-based ranking loss has been widely applied [11, 49, 15]. Formally, for an image (caption) embedding \mathbf{v} (\mathbf{u}), denote the embedding of its matched caption (image) as \mathbf{u}^+ (\mathbf{v}^+). A negative (unmatched) caption (image) is sampled whose embedding is denoted as \mathbf{u}^- (\mathbf{v}^-). We define the bidirectional ranking loss ℓ_{sent} between captions and images as:

$$\ell_{sent} = \sum_{\mathbf{u}} F_{\mathbf{v}^-} (|\delta + s(\mathbf{u}, \mathbf{v}^-) - s(\mathbf{u}, \mathbf{v}^+)|_+) + \sum_{\mathbf{v}} F_{\mathbf{u}^-} (|\delta + s(\mathbf{u}^-, \mathbf{v}) - s(\mathbf{u}^+, \mathbf{v})|_+) \quad (1)$$

, where δ is a predefined margin, $|x|_+ = \max(x, 0)$ is the traditional ranking loss and $F_x(\cdot) = \max_x(\cdot)$ denotes the hard negative mining strategy [11, 41]. $s(\cdot, \cdot)$ is a similarity function between two embeddings and is usually implemented as cosine similarity [11, 40, 49].

3.2. Semantic Encodings

The encoding of a caption is made up of three steps. As an example, consider the caption shown in Fig. 3, “A white clock on the wall is above a wooden table”. 1) We extract a structured meaning representation as a collection of three types of semantic components: object (clock, wall, table), attribute-object dependencies (white clock, wooden table) and relational dependencies (clock above table, clock on wall). 2) We encode each component as well as the full sentence with type-specific encoders into the unified VSE space. 3) We compose the embedding of the caption by combining semantic components.

Semantic parsing. We implement a semantic parser ¹ of image captions based on [38]. Given the input sentence, the parser first performs a syntactic dependency parsing. A set of rules is applied to the dependency tree and extracts object entities appeared in the sentence, adjectives that modify the

¹<https://github.com/vacancy/SceneGraphParser>

object nouns, subjects/objects of the verbs and prepositional phrases. For simplicity, we consider only single-word nouns for objects and single-word adjectives for object attributes.

Encoding objects and attributes. We use an unified object encoder ϕ for nouns and adjective-noun pairs. For each word w in the vocabulary, we initialize a basic semantic embedding $\mathbf{w}^{(basic)} \in \mathbb{R}^{d_{basic}}$ and a modifier semantic embedding $\mathbf{w}^{(modif)} \in \mathbb{R}^{d_{modif}}$.

For a single noun word w_n (*e.g.*, `clock`), we define its embedding \mathbf{w}_n as $\mathbf{w}_n^{(basic)} \oplus \mathbf{w}_n^{(modif)}$, where \oplus means the concatenation of vectors. For an (adjective, noun) pair (w_a, w_n) (*e.g.*, (`white`, `clock`)), its embedding $\mathbf{w}_{a,n}$ is defined as $\mathbf{w}_n^{(basic)} \oplus \mathbf{w}_a^{(modif)}$ where $\mathbf{w}_a^{(modif)}$ encodes the attribute information. In implementation, the basic semantic embedding is initialized from GloVe [34]. The modifier semantic embeddings (both $\mathbf{w}_n^{(modif)}$ and $\mathbf{w}_a^{(modif)}$) are randomly initialized and jointly learned. $\mathbf{w}_n^{(modif)}$ can be regarded as an intrinsic modifier for each nouns.

To fuse the embeddings of basic and modifier semantics, we employ a gated fusion function:

$$\begin{aligned}\phi(\mathbf{w}_n) &= \text{Norm}(\sigma(\mathbf{W}_1 \mathbf{w}_n + \mathbf{b}_1)) \tanh(\mathbf{W}_2 \mathbf{w}_n + \mathbf{b}_2), \\ \phi(\mathbf{w}_{a,n}) &= \text{Norm}(\sigma(\mathbf{W}_1 \mathbf{w}_{a,n} + \mathbf{b}_1) \tanh(\mathbf{W}_2 \mathbf{w}_{a,n} + \mathbf{b}_2)).\end{aligned}$$

Throughout the text, σ denotes the sigmoid function: $\sigma(x) = 1/(1 + \exp(-x))$, and Norm denotes the L2 normalization, *i.e.*, $\text{Norm}(\mathbf{w}) = \mathbf{w}/\|\mathbf{w}\|_2$. One may interpret ϕ as a GRU cell [7] taking no historical state.

Encoding relations and full sentence. Since relations and sentences are the composed based on objects, we encode them with a neural combiner ψ , which takes the embeddings of word-level semantics encoded by ϕ as input. In practice, we implement ψ as an uni-directional GRU [7], and pick the L2-normalized last state as the output.

To obtain a visual-semantic embedding for a relational triple (w_s, w_r, w_o) (*e.g.*, (`clock`, `above`, `table`)), we first extract the word embeddings for the subject, relational word and the object using ϕ . We then feed the encoded word embeddings in the same order into ψ and takes the L2-normalized last state of the GRU cell. Mathematically, $\mathbf{u}_{rel} = \psi(w_s, w_r, w_o) = \psi(\{\phi(\mathbf{w}_s), \phi(\mathbf{w}_r), \phi(\mathbf{w}_o)\})$.

The embedding of a sentence \mathbf{u}_{sent} is computed over the word sequence w_1, w_2, \dots, w_k of the caption:

$$\mathbf{u}_{sent} = \psi(\{\phi(\mathbf{w}_1), \phi(\mathbf{w}_2), \dots, \phi(\mathbf{w}_k)\}),$$

where for any word x , $\phi(\mathbf{w}_x) = \phi(\mathbf{w}_x^{(basic)} \oplus \mathbf{w}_x^{(modif)})$

Note that we share the weights of the encoders ψ and ϕ among the encoding processes of all semantic levels. This allows our encoders of various types of components to bootstrap the learning of each other.

Combining all of the components. A straight-forward implementation of the caption encoder is to directly use the

sentence embedding \mathbf{u}_{sent} , as it has already combined the semantics of components in a contextually-weighted manner [25]. However, it has been revealed in [40] that such combination is vulnerable to adversarial attacks: Because of the biases in the dataset, the combiner ψ usually focuses on only a small set of semantic components appeared in the caption.

We alleviate such biases by enforcing the coverage of the semantic components appeared in the sentence. Specifically, to form the caption embedding \mathbf{u}_{cap} , the sentence embedding \mathbf{u}_{sent} is combined with an explicit bag-of-components embedding \mathbf{u}_{comp} , as illustrated in Fig. 3 (right). Mathematically, we define \mathbf{u}_{comp} is computed by the aggregation of all components in the sentence:

$$\mathbf{u}_{comp} = \text{Norm}(\Phi(\{\mathbf{u}_{obj}\} \cup \{\mathbf{u}_{attr}\} \cup \{\mathbf{u}_{rel}\})),$$

where $\Phi(\cdot)$ is the aggregation function of semantic components. Then the caption is encoded as: $\mathbf{u}_{cap} = \alpha \mathbf{u}_{sent} + (1 - \alpha) \mathbf{u}_{comp}$, where $0 \leq \alpha \leq 1$ is a scalar weight. The presence of \mathbf{u}_{comp} disallows the ignorance of any of the components in the final caption embedding \mathbf{u}_{cap} .

3.3. Image Encodings

We use CNN to encode the input RGB image into the unified VSE space. Specifically, we choose a ResNet-152 model [14] pretrained on ImageNet [37] as the image encoder. We apply a layer of 1×1 convolution on top of the last convolutional layer (*i.e.*, `conv5_3`) and obtain a convolutional feature map of shape $7 \times 7 \times d$ for each image. d denotes the dimension of the unified VSE space.

The feature map, denoted as $\mathbf{V} \in \mathbb{R}^{7 \times 7 \times d}$, can be viewed as the embeddings of 7×7 local regions in the image. The embedding \mathbf{v} for the whole image is defined as the aggregation $\Psi(\cdot)$ of the embeddings at all regions through a global spatial pooling operator.

3.4. Learning Paradigm

In this section, we present how to align vision and language into the unified space using contrastive learning on different semantic levels. The training pipeline is illustrated in Fig. 3. We start from the generation of contrastive examples for different semantic components.

Negative example sampling. It has been discussed in [40] that to explore a large compositional space of semantics, directly sampling negative captions from a human-built dataset (*e.g.*, MS-COCO captions) is not sufficient. In this paper, instead of manually define rules that augment the training data as in [40], we address this problem by sampling contrastive negative examples in the explicitly factorized semantic space. The generation does not require manually labelled data, and can be easily applied to any datasets. For a specific caption, we generate the following four types of contrastive negative samples.

- **Nouns.** We sample negative noun words from all nouns that do not appear in the caption.²
- **Attribute-noun pairs.** We sample negative pairs by randomly substituting the adjective by another adjective or substituting the noun.
- **Relational triples.** We sample negative triples by randomly substituting the subject, or the relation, or the object. Moreover, we also sample the whole relational triples of captions in the dataset which describe other images, as the negative triples.
- **Sentences.** We sample negative sentences from the whole dataset. Meanwhile, following [13, 11], we also sample negative images from the whole dataset as contrastive images.

The key motivation behind our visual-semantic alignment is that: an object appears in a local region of the image, while the aggregation of all local regions should be aligned with the full semantics of a caption.

Local region-level alignment. In detail, we propose a relevance-weighted alignment mechanism for linking textual object descriptors and local image regions. As shown in Fig. 4, consider the embedding of a positive textual object descriptor \mathbf{u}_o^+ , a negative textual object descriptor \mathbf{u}_o^- and the set image local region embeddings \mathbf{V}_i where $i \in 7 \times 7$ extracted from the image. We generate a relevance map $\mathbf{M} \in \mathbb{R}^{7 \times 7}$ with $\mathbf{M}_{i,i} \in 7 \times 7$ representing the relevance between \mathbf{u}_o^+ and \mathbf{V}_i , computed as Eq. (2). We compute the loss for noun and (adjective, noun) pairs by:

$$\mathbf{M}_i = \frac{\exp(s(\mathbf{u}_o^+, \mathbf{V}_i))}{\sum_j \exp(s(\mathbf{u}_o^+, \mathbf{V}_j))} \quad (2)$$

$$\ell_{obj} = \sum_{i \in 7 \times 7} (\mathbf{M}_i \cdot |\delta + s(\mathbf{u}_o^-, \mathbf{V}_i) - s(\mathbf{u}_o^+, \mathbf{V}_i)|_+) \quad (3)$$

The intuition behind the definition is that, we explicitly try to align the embedding at each image region with \mathbf{u}_o^+ . The losses are weighted by the matching score, thus reinforce the correspondence between \mathbf{u}_o^+ and the matched region. This technique is related to multi-instance learning [45].

Global image-level alignment. For relational triples \mathbf{u}_{rel} , semantic components aggregations \mathbf{u}_{comp} and sentences \mathbf{u}_{sent} , their semantics usually cover multiple objects. Thus, we align them with the full image embedding \mathbf{v} via bidirectional ranking losses as Eq. (1)³. The alignment loss is denoted as ℓ_{rel} , ℓ_{comp} and ℓ_{sent} , respectively.

We want to highlight that, during training, we separately align the two type of semantic representations of the caption, *i.e.*, \mathbf{u}_{sent} and \mathbf{u}_{comp} , with the image. This differs from the inference-time computation of the caption. Recall that α can be viewed as a factor that balances the training objective and

²For the MS-COCO dataset, in all 5 captions associated with the same image. This also applies to other components.

³Only textual negative samples are used for ℓ_{rel} .

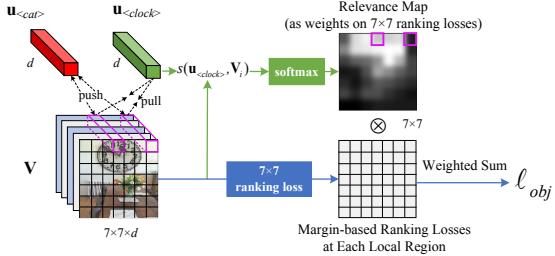


Figure 4. An illustration of our relevance-weighted alignment mechanism. The relevance map shows the similarity of each region with the object embedding \mathbf{u}_{clock} . We weight the alignment loss with the map to reinforce the correspondence between the \mathbf{u}_{clock} and its matched region.

the enforcement of semantic coverage. This allows us to flexibly adjust α during inference.

3.5. Implementation details

We use $d = 1024$ as the dimension of the unified VSE space like [11, 40, 49]. We train the model by minimizing the alignment losses in a multi-task learning way.

$$\ell = \ell_{sent} + \eta_c \ell_{comp} + \eta_o \ell_{obj} + \eta_a \ell_{attr} + \eta_r \ell_{rel} \quad (4)$$

In the first 2 epochs, we set η_c , η_o and η_a to 0.5 and η_r to 0 for learning single-object level representations. Then we turn up η_r to 1.0 to make the model learn relational semantics. To make the comparison with related works fair, we always fix the weights of the ResNet. We use the Adam [22] optimizer with learning rate at 0.001. For model details, please refer to our supplementary material.

4. Experiments

We evaluate our model on the MS-COCO [27] dataset. It contains 82,783 training images with each image annotated by 5 captions. We use the common 1K validation and test split from [19]. We also report the performance on a 5K test split for comparison with [49, 11, 42].

We begin this section with the evaluation of traditional cross-modal retrieval. Next, we validate the effectiveness of enforcing the semantic coverage of caption embeddings by comparing models on cross-modal retrieval tasks with adversarial examples. We then propose a unified text-to-image retrieval task to support the contrastive learning on various semantic components. We end this section with an application of using visual cues to facilitate the semantic parsing of novel sentences. Due to the limitation of the text length, for mode details on data processing, metrics and model implementation, we refer the readers to our supplementary material.

4.1. Overall Evaluation on Cross-Modal Retrieval.

We first show the performance of image-to-sentence and sentence-to-image retrieval tasks to evaluate learned visual-semantic embeddings. We report the R@1 (recall@1), R@5,

Task	Image-to-sentence Retrieval				Sentence-to-image Retrieval				
Metric	R@1	R@5	R@10	Med. r	R@1	R@5	R@10	Med. r	rsum
1K testing split (5,000 captions)									
m-RNN [31]	41.0	73.0	83.5	2	29.0	42.2	77.0	3	345.7
DVSA [20]	38.4	69.9	80.5	1	27.4	60.2	74.8	3	351.2
MNLM [24]	43.4	75.7	85.8	-	31.0	66.7	79.9	-	382.5
m-CNN [30]	42.8	73.1	84.1	3	32.6	68.6	82.8	3	384.0
HM-LSTM[33]	43.9	-	87.8	2	36.1	-	86.7	3	-
Order-embedding [42]	46.7	-	88.9	2	37.9	-	85.9	2	-
VSE-C [40, 1]	48.0	81.0	89.2	2	39.7	72.9	83.2	2	414
DeepSP[44]	50.1	79.7	89.2	-	39.6	75.2	86.9	-	420.7
2WayNet [9]	55.8	75.2	-	-	39.7	63.3	-	-	-
sm-LSTM [15]	53.2	83.1	91.5	1	40.7	75.8	87.4	2	431.8
RRF-Net[28]	56.4	85.3	91.5	-	43.9	78.1	88.6	-	443.8
VSE++ [11, 2]	57.7	86.0	94.0	1	42.8	77.2	87.4	2	445.1
CSE[49]	56.3	84.4	92.2	1	45.7	81.2	90.6	2	450.4
UniVSE (Ours)	64.3	89.2	94.8	1	48.3	81.7	91.2	2	469.5
5K testing split (25,000 captions)									
Order-embedding [42]	23.3	-	65.0	5	18.0	-	57.6	7	-
VSE-C[11, 1]	22.3	51.1	65.1	5	18.7	43.8	56.7	7	257.7
CSE[49]	27.9	57.1	70.4	4	22.2	50.2	64.4	5	292.2
VSE++[11, 2]	31.7	60.9	72.7	3	22.1	49.0	62.7	6	299.1
UniVSE (Ours)	36.1	66.4	77.7	3	25.4	53.0	66.2	5	324.8

Table 1. Results of cross-modal retrieval task on MS-COCO dataset (1K and 5K testing split). All listed baselines and our models fix weights of the image encoders. For fair comparison, we do not include [10] and [16] that finetunes the image encoder or adds extra training data.

Metric	Object attack				Attribute attack				Relation attack				total sum
	R@1	R@5	R@10	rsum	R@1	R@5	R@10	rsum	R@1	R@5	R@10	rsum	
VSE++	32.3	69.6	81.4	183.3	19.8	59.4	76.0	155.2	26.1	66.8	78.7	171.6	510.1
VSE-C	41.1	76.0	85.6	202.7	26.7	61.0	74.3	162.0	35.5	71.1	81.5	188.1	552.8
UniVSE ($\mathbf{u}_{sent} + \mathbf{u}_{comp}$)	45.3	78.3	87.3	210.9	35.3	71.5	83.1	189.9	39.0	76.5	86.7	202.2	603.0
UniVSE (\mathbf{u}_{sent})	40.7	76.4	85.5	202.6	30.0	70.5	80.6	181.1	32.6	72.6	83.5	188.7	572.4
UniVSE ($\mathbf{u}_{sent} + \mathbf{u}_{obj}$)	42.9	77.2	85.6	205.7	30.1	69.0	79.8	178.9	34.0	71.2	83.6	188.8	573.4
UniVSE ($\mathbf{u}_{sent} + \mathbf{u}_{attr}$)	40.1	73.9	83.3	197.3	37.4	72.0	81.9	191.3	30.5	70.0	81.9	182.4	571.0
UniVSE ($\mathbf{u}_{sent} + \mathbf{u}_{rel}$)	45.4	77.1	85.5	208.0	29.2	68.1	78.5	175.8	42.8	77.5	85.6	205.9	589.7

Table 2. Results on image-to-sentence retrieval task with text-domain adversarial attacks. For each caption, we generate 5 adversarial fake captions which do not match the images. Thus, the models need to retrieve 5 positive captions from 30,000 candidate captions.

R@10, and the median retrieval rank as in [11, 40, 49, 15]. To summarize the performance, we compute *rsum* as the summation of R@1, R@5, and R@10.

Shown in Table 1, Unified VSE outperforms other baselines with various model architecture and training techniques [11, 49, 28, 40, 15]. This validates the effectiveness learning visual-semantic embeddings in the explicitly factorized visual-semantic embedding space. We also include the results under more challenging 5K test split. The gap between Unified VSE and other models gets further enlarged across all metrics.

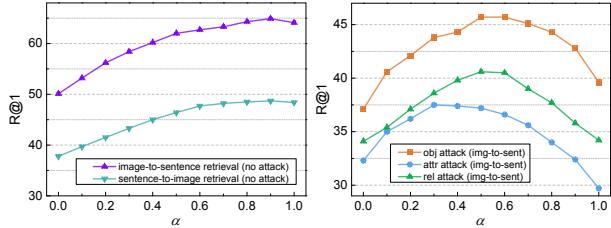
4.2. Retrieval under text-domain adversarial attack

Recent works [40, 39] have raised their concerns on the robustness of the learned visual-semantic embeddings. They show that existing models are vulnerable to text-domain adversarial attacks (*i.e.*, using adversarial captions) and can be easily fooled. This is closely related to the bias in small datasets over a large, compositional semantic space [40]. To prove the robustness of the learned unified VSE, we further

conduct experiments on the image-to-sentence retrieval task with text-domain adversarial attacks. Following [40], we first design several types of adversarial captions by adding perturbations to existing captions.

- Object attack:** Randomly replace / append by an irrelevant one in the original caption.
- Attribute attack:** Randomly replace / add an irrelevant attribute modifier for one object in the original caption.
- Relational attack:** 1) Randomly replace the subject/relation/object word by an irrelevant one. 2) Randomly select an entity as a subject/object and add an irrelevant relational word and object/subject.

We include VSE++ and VSE-C as the baselines and show the results in Table 2 where different columns represent different types of attacks. VSE++ performs worst as it is only optimized for the retrieval performance on the dataset. Its sentence encoder is insensitive to a small perturbation in the text. VSE-C explicitly generates the adversarial captions based on human-designed rules as hard negative examples during training, which makes it relatively robust to those



(a) Normal cross-modal retrieval (5,000 captions) (b) Adversarial attacked image-to-sentence retrieval (30,000 captions)

Figure 5. The performance of UniVSE on cross-modal retrieval tasks with different combination weight α . Our model can effectively defend adversarial attacks, with no sacrifice for the performance on other tasks by choosing a reasonable α (thus we set $\alpha = 0.75$ in all other experiments).

adversarial attacks. Unified VSE shows strong robustness across all types of adversarial attacks.

It is worth noting that VSE-C shows inferior performances in the normal retrieval tasks without adversarial captions (see Table 1), even compared with VSE++. Considering that VSE-C shares the exactly the same model architecture as VSE++, we can conclude that directly adding adversarial captions during training, although improves models’ robustness, may sacrifice the performance on other tasks. In contrast, the ability of Unified VSE to defend adversarial texts comes almost for free: we present *zero* adversarial captions during training. Unified VSE builds fine-grained semantic alignments via the contrastive learning of semantic components. It uses the explicit aggregation of the components \mathbf{u}_{comp} to alleviate the dataset biases.

Ablation study: semantic components. We now delve into the effectiveness of different semantic components by choosing different combinations of components for the caption embedding. Shown in Table 2, we use different subsets of the semantic components to form the bag-of-component embeddings \mathbf{u}_{comp} . For example, in UniVSE_{obj}, only object nouns are selected and aggregated as \mathbf{u}_{comp} .

The results demonstrate the effectiveness of the enforcement of semantic coverage: even if the semantic components have got fine-grained alignment with visual concepts, directly using \mathbf{u}_{sent} as the caption encoding still degenerates the robustness against adversarial examples. Consistent with the intuition, enforcing of coverage of a certain type of components (*e.g.*, objects) helps the model to defend the adversarial attacks of the same type (*e.g.*, defending adversarial attacks of nouns). Combining all components leads to the best performance.

Choice of the combination factor: α . We study the choice of α by conducting experiments on both normal retrieval tasks and the adversarial one. Fig 4.2 shows the R@1 performance under the normal/adversarial retrieval scenario w.r.t. different choices of α . We observe that the \mathbf{u}_{comp} term contributes little on the normal retrieval tasks but largely on tasks

Task	obj	attr	rel	obj (det)	sum
VSE++	29.95	26.64	27.54	50.57	134.70
VSE-C	27.48	28.76	26.55	46.20	128.99
UniVSE _{all}	39.49	33.43	39.13	58.37	170.42
UniVSE _{obj}	39.71	33.37	34.38	56.84	164.30
UniVSE _{attr}	31.31	37.51	34.73	52.26	155.81
UniVSE _{rel}	37.55	32.70	39.57	59.12	168.94

Table 3. The mAP performance on the unified text-to-image retrieval task. Please refer to the text for details.

with adversarial attacks. Recall that α can be viewed as a factor that balances the training objective and the enforcement of semantic coverage. By choosing α from a reasonable range (0.6 to 0.8), our model can effectively defend adversarial attacks, with no sacrifice for the overall performance.

4.3. Unified Text-to-Image Retrieval

We extend the word-to-scene retrieval used by [40] into a general *unified text-to-image retrieval* task. In this task, models receive queries of different semantic levels, including single words (*e.g.*, “Clock.”), noun phrases (*e.g.*, “White clock.”), relational phrases (*e.g.*, “Clocks on wall”) and full sentences. For all baselines, the texts of different types are treated as full sentences. The result is presented in Table 3.

We generate positive image-text pairs by randomly choosing an image and a semantic component from 5 matched captions with the chosen image. It is worth mentioning that the semantic components extracted from captions may not cover all visual concepts in the corresponding image, which makes the annotation noisy. To address this, we also leverage the MS-COCO detection annotations to facilitate the evaluation (see *obj(det)* column). We treat the labels for detection bounding boxes as the annotation of objects in the scene.

Ablation study: contrastive learning of components. We evaluate the effectiveness of using contrastive samples for different semantic components. Shown in Table 3, UniVSE_{obj} denotes the model trained with only contrastive samples of noun components. The same notation applies to other models. The UniVSE trained with a certain type of contrastive examples (*e.g.*, UniVSE_{obj} with contrastive nouns) consistently improves the retrieval performance of the same type of queries (*e.g.*, retrieving images from a single noun). UniVSE trained with all kinds of contrastive samples performs best in overall and shows a significant gap w.r.t. other baselines.

Visualization of the semantic alignment. We visualize the semantic-relevance map on an image w.r.t. a given query \mathbf{u}_q for a qualitative evaluation of the alignment performance of various semantic components. The map M_i is computed as the similarity between each image region \mathbf{v}_i and \mathbf{u}_q , in a similar way as Eq. (2). Shown as Fig. 6, this visualization helps to verify that our model successfully aligns different semantic components with the corresponding image regions.



Figure 6. The relevance maps and grounded areas obtained from the retrieved images w.r.t. three queries. The temperature of the softmax for visualizing the relevance map is $\tau = 0.1$. Pixels in white indicates a higher matching score. Note that the third image of the query “black dog” contains two dogs, while our model successfully locates the black one (on the left). It also succeeded in finding the white dog in the first image of “white dog”. Moreover, for the query “player swing bat”, although there are many players in the image, our model only attend to the man swinging the bat.

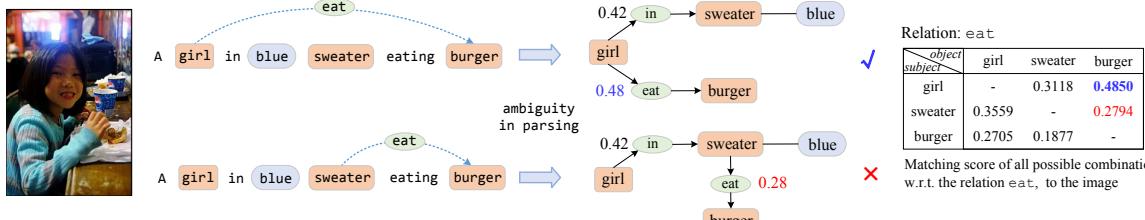


Figure 7. Example showing that Unified VSE can leverage image to parse sentences with ambiguity. The matching score of “girl eat burger” is much higher than “sweater eat burger”, which resolves the ambiguity. Other components are also correctly inferred.

Task	attributed object	relational phrase
Random	37.41	31.90
VSE++	41.12	43.31
VSE-C	43.44	41.08
UniVSE	64.82	62.69

Table 4. The accuracy of different models on recovering word dependencies with visual cues. In the “Random” baseline, we randomly assign the word dependencies.

4.4. Semantic Parsing with Visual Cues

As a side application, we show how the learned unified VSE space can provide the visual cues to help the semantic parsing of sentences. Fig. 7 shows the general idea. When parsing a sentence, ambiguity may occur, *e.g.*, the subject of the relational word eat may be sweater or burger. It is not easy for a textual parser to decide which one is correct because of the innate syntactic ambiguity. However, we can use the image which is depicted by this sentence to assist the parsing by. This is related to previous works on using image segmentation models to facilitate the sentence parsing [6].

This motivates us to design two tasks, 1) recovering the dependency between attributes and entities, and 2) recovering the relational triples. In detail, we first extract the entities, attributes and relational words from the raw sentence without knowing their dependencies. For each possible combination of certain semantic component, our model computes its embedding in the unified joint space. *E.g.*, in Fig. 7, there are in total $3 \times (3 - 1) = 6$ possible dependencies for eat. We choose the combination with the highest matching score with the image to decide the subject/object dependencies of

the relation eat. We use parsed semantic components as the ground-truth and report the accuracy, defined as the fraction of the number of correct dependency resolution and the total number of attributes/relations. Table 4 reports the results on assisting semantic parsing with visual cues, compared with other baselines. Fig. 7 shows a real case in which we successfully resolve the textual ambiguity.

5. Conclusion

We present a *unified visual-semantic embedding* approach that learns a joint representation space of vision and language in a factorized manner: Different levels of textual semantic components such as objects and relations get aligned with regions of images. A contrastive learning approach for semantic components is proposed for the efficient learning of the fine-grained alignment. We also introduce the enforcement of semantic coverage: each caption embedding should have a coverage of all semantic components in the sentence. Unified VSE shows superiority on multiple cross-modal retrieval tasks and can effectively defend text-domain adversarial attacks. We hope the proposed approach can empower machines that learn vision and language jointly, efficiently and robustly.

6. Acknowledgements

We thank Haoyue Shi for helpful discussions and suggestions. This research is supported in part by the National Key Research and Development Program of China under grant 2018YFB0505000 and the National Natural Science Foundation of China under grant 61772138.

References

- [1] VSE-C open-sourced code. <https://github.com/ExplorerFreda/VSE-C>. 6
- [2] VSE++ open-sourced code. <https://github.com/fartashf/vsepp>. 6
- [3] O. Abend, T. Kwiatkowski, N. J. Smith, S. Goldwater, and M. Steedman. Bootstrapping Language Acquisition. *Cognition*, 164:116–143, 2017. 1, 2
- [4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015. 2
- [5] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. Abstract Meaning Representation for Sembanking. In *Linguistic Annotation Workshop and Interoperability with Discourse*, 2013. 3
- [6] G. Christie, A. Laddha, A. Agrawal, S. Antol, Y. Goyal, K. Kochersberger, and D. Batra. Resolving Language and Vision Ambiguities Together: Joint Segmentation & Prepositional Attachment Resolution in Captioned Scenes. *arXiv:1604.02125*, 2016. 8
- [7] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. In *NIPS 2014 Workshop on Deep Learning*, 2014. 4
- [8] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2
- [9] A. Eisenschitz and L. Wolf. Linking Image and Text with 2-way Nets. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 6
- [10] M. Engilberge, L. Chevallier, P. Pérez, and M. Cord. Finding Beans in Burgers: Deep Semantic-Visual Embedding with Localization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 6
- [11] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *Proceedings of British Machine Vision Conference (BMVC)*, 2018. 2, 3, 5, 6
- [12] A. Fazly, A. Alishahi, and S. Stevenson. A Probabilistic Computational Model of Cross-Situational Word Learning. *Cognitive Science*, 34(6):1017–1063, 2010. 2
- [13] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A Deep Visual-Semantic Embedding Model. In *Advances in Neural Information Processing Systems (NIPS)*, 2013. 2, 5
- [14] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4
- [15] Y. Huang, W. Wang, and L. Wang. Instance-Aware Image and Sentence Matching with Selective Multimodal LSTM. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3, 6
- [16] Y. Huang, Q. Wu, and L. Wang. Learning Semantic Concepts and Order for Image and Sentence Matching. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6
- [17] J. Johnson, A. Gupta, and L. Fei-Fei. Image Generation from Scene Graphs. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2
- [18] J. Johnson, R. Krishna, M. Stark, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Image Retrieval using Scene Graphs. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2
- [19] A. Karpathy and L. Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 5
- [20] A. Karpathy and L. Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6
- [21] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. 2
- [22] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980*, 2017. 5
- [23] R. Kiros, R. Salakhutdinov, and R. Zemel. Multimodal Neural Language Models. In *Proceedings of International Conference on Machine Learning (ICML)*, 2014. 2
- [24] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *arXiv:1411.2539*, 2014. 2, 6
- [25] O. Levy, K. Lee, N. Fitzgerald, and L. Zettlemoyer. Long Short-Term Memory as a Dynamically Computed Element-wise Weighted Sum. *arXiv:1805.03716*, 2018. 4
- [26] P. Liang, M. I. Jordan, and D. Klein. Learning Dependency-Based Compositional Semantics. *Computational Linguistics*, 39(2):389–446, 2013. 3
- [27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2014. 3, 5
- [28] Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew. Learning a Recurrent Residual Fusion Network for Multimodal Matching. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 6
- [29] C. Lu, R. Krishna, M. Bernstein, and L. Fei-Fei. Visual Relationship Detection with Language Priors. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 2
- [30] L. Ma, Z. Lu, L. Shang, and H. Li. Multimodal Convolutional Neural Networks for Matching Image and Sentence. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2015. 2, 6
- [31] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep Captioning with Multimodal Recurrent Neural Net-

- works (m-RNN). In *Proceedings of International Conference on Learning Representations (ICLR)*, 2015. 6
- [32] R. Montague. Universal Grammar. *Theoria*, 36(3):373–398, 1970. 3
- [33] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Hierarchical Multimodal LSTM for Dense Visual-Semantic Embedding. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 6
- [34] J. Pennington, R. Socher, and C. Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 4
- [35] Z. Ren, H. Jin, Z. Lin, C. Fang, and A. Yuille. Joint Image-Text Representation by Gaussian Visual-Semantic Embedding. In *Proceedings of ACM Multimedia (ACM-MM)*, 2016. 2
- [36] Z. Ren, H. Jin, Z. Lin, C. Fang, and A. Yuille. Multiple Instance Visual-Semantic Embedding. In *Proceedings of British Machine Vision Conference (BMVC)*, 2017. 2
- [37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 4
- [38] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning. Generating Semantically Precise Scene Graphs from Textual Descriptions for Improved Image Retrieval. In *Workshop on Vision and Language (VL15)*, Lisbon, Portugal, 2015. 3
- [39] R. Shekhar, S. Pezzelle, Y. Klimovich, A. Herbelot, M. Nabi, E. Sangineto, and R. Bernardi. FOIL it! Find One Mismatch between Image and Language Caption. *arXiv:1705.01359*, 2017. 2, 6
- [40] H. Shi, J. Mao, T. Xiao, Y. Jiang, and J. Sun. Learning Visually-Grounded Semantics from Contrastive Adversarial Samples. In *Proceedings of International Conference on Computational Linguistics (COLING)*, 2018. 2, 3, 4, 5, 6, 7
- [41] A. Shrivastava, A. Gupta, and R. Girshick. Training Region-Based Object Detectors with Online Hard Example Mining. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3
- [42] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. Order-Embeddings of Images and Language. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2016. 5, 6
- [43] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and Tell: A Neural Image Caption Generator. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2
- [44] L. Wang, Y. Li, and S. Lazebnik. Learning Deep Structure-Preserving Image-Text Embeddings. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 6
- [45] J. Wu, Y. Yu, C. Huang, and K. Yu. Deep Multiple Instance Learning for Image Classification and Auto-Annotation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 5
- [46] F. Xiao, L. Sigal, and Y. J. Lee. Weakly-Supervised Visual Grounding of Phrases with Linguistic Structures. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2
- [47] H. Xu and K. Saenko. Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016. 2
- [48] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of International Conference on Machine Learning (ICML)*, 2015. 2
- [49] Q. You, Z. Zhang, and J. Luo. End-to-End Convolutional Semantic Embeddings. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 5, 6
- [50] L. S. Zettlemoyer and M. Collins. Learning to Map Sentences to Logical Form: Structured Classification with Probabilistic Categorial Grammars. In *Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI)*, 2005. 3