

# Learning Visually-Grounded Semantics from Contrastive Adversarial Samples

Haoyue Shi<sup>1\*</sup> Jiayuan Mao<sup>2\*</sup> Tete Xiao<sup>1\*</sup> Yuning Jiang<sup>3</sup> Jian Sun<sup>3</sup>

<sup>1</sup>: School of Electronics Engineering and Computer Science, Peking University, China

<sup>2</sup>: ITCS, Institute for Interdisciplinary Information Sciences, Tsinghua University, China

<sup>3</sup>: Megvii, Inc.

{hyshi, jasonhsiao97}@pku.edu.cn

mjy14@mails.tsinghua.edu.cn, {jyn, sunjian}@megvii.com

## Abstract

We study the problem of grounding distributional representations of texts on the visual domain, namely visual-semantic embeddings (VSE for short). Begin with an insightful adversarial attack on VSE embeddings, we show the limitation of current frameworks and image-text datasets (*e.g.*, MS-COCO) both quantitatively and qualitatively. The large gap between the number of possible constitutions of real-world semantics and the size of parallel data, to a large extent, restricts the model to establish a strong link between textual semantics and visual concepts. We alleviate this problem by augmenting the MS-COCO image captioning datasets with textual contrastive adversarial samples. These samples are synthesized using language priors of human and the WordNet knowledge base, and enforce the model to ground learned embeddings to concrete concepts within the image. This simple but powerful technique brings a noticeable improvement over the baselines on a diverse set of downstream tasks, in addition to defending known-type adversarial attacks. Codes are available at <https://github.com/ExplorerFreda/VSE-C>.

## 1 Introduction

The visual grounding of language plays an indispensable role in our daily lives. We use language to name, refer, and describe objects, their properties and generally, visual concepts. Distributional semantics (*e.g.*, global word embedding (Pennington et al., 2014)) based on large-scale corpora have shown great success in modeling the functionality and correlation of words in the natural language domain. This further contributes to the success in numerous natural language processing (NLP) tasks such as language modeling (Cheng et al., 2016; Inan et al., 2017), sentiment analysis (Cheng et al., 2016; Kumar et al., 2016), and reading comprehension (Cheng et al., 2016; Chen et al., 2016; Shen et al., 2017). However, effective and efficient grounding of distributional embeddings remains challenging. Being ignorant of the corresponding visual concepts, pure textual embeddings demonstrate inferior performances when incorporating with visual inputs. Typical tasks include image/video captioning, multi-modal retrieval/understanding, and visual reasoning, some of which are further extensively studied in the paper.

Visual concept and its link with textual semantics, as a cognitive alignment, provide rich supervision to learning systems. Visual-Semantic Embedding (VSE) aims at building the bridge between natural language and the underlying visual world. Introduced by Kiros et al. (2014), the embedding spaces of both images and descriptive texts (captions) are jointly optimized and aligned. Nevertheless, even for large-scale datasets such as MS-COCO (Lin et al., 2014), the number of image-caption pairs are far less than the number of possible constitutions of real-world semantics, making the dataset inevitably sparse and biased.

To reveal this, we begin with constructing textual adversarial samples to attack the state-of-the-art system VSE++ (Faghri et al., 2017). Specifically, we study the composition of sentences in two aspects: (1) content words including nouns and numerals and (2) prepositions indicating spatial relations (*e.g.*, in,

---

\* Work was done when HS, JM and TX were intern researchers at Megvii Inc. HS, JM and TX contribute equally to this paper.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: <http://creativecommons.org/licenses/by/4.0/>

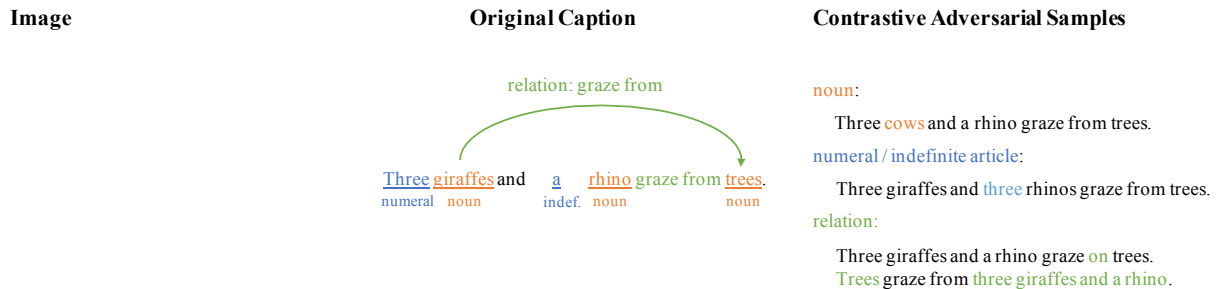


Figure 1: An overview of our textual contrastive adversarial samples. For each caption, we have three paradigms to generate contrastive adversarial samples, *i.e.*, noun, numeral and relation. For an given image, we expect the model to distinguish the real captions against the generated adversarial ones.

on, above, below). As shown in Figure 1, we manipulate the original caption to construct hard negative captions with similar structure but completely contradictory semantics. We found that the models easily get confused, suffering a noticeable drop in confidence or even wrong predictions in the caption retrieval task.

We propose VSE-C, which enforces the learning of correlation and correspondence between textual semantics and visual concepts by providing contrastive adversarial samples during the training procedure, incorporating with an intra-pair hard negative sample mining. Instead of defending adversarial attacks, we focus on the study of limitations of current visual-semantic datasets and the transferability of learned embeddings. To fulfill the large gap between the number of parallel image-caption pairs and the expressiveness of natural languages, we augment the data by employing a set of heuristic rules to generate large sets of contrastive negative captions, as demonstrated in Figure 1. The candidates are selectively used for training by an intra-pair hard-example mining technique. VSE-C alleviate the bias of dataset and provide rich and effective samples on par with original image captions. This strengthen the link between text and visual concepts by requiring models to detect a mismatch on the level of some precise concepts.

VSE-C learns discriminative and visually-grounded word embeddings on the MS-COCO dataset (Lin et al., 2014). It is extensively compared with existing works with rich experiments and analyses. Most importantly, we explore the transferability of the learned embeddings on several real-world applications both qualitatively and quantitatively, including image-to-text retrieval and bidirectional word-to-concept retrieval. Furthermore, VSE-C demonstrates a general framework for augmenting textual inputs considering semantical consistency. The introduction human priors and knowledge bases alleviates the sparsity and non-contiguity of languages. We hope the framework and the released data are beneficial for building more robust and data-efficient models.

## 2 Related works

**Joint embeddings** Joint embedding is a common technique for a wide range of tasks incorporating multiple domains, including audio-video embeddings for unsupervised representation learning (Ngiam et al., 2011), shape-image embeddings (Li et al., 2015) for shape inference, bilingual word embeddings for machine translation (Zou et al., 2013), human pose-image embeddings for pose inference (Li, 2011), image-text embeddings for visual description (Reed et al., 2016), and global representation learning from multiple domains (Castrejon et al., 2016). These embeddings map multiple domains into a joint vector space which describes the semantical relations between inputs (*e.g.*, distance, correlation).

We focus on the visual-semantic embedding (Mao et al., 2016; Kiros et al., 2014; Faghri et al., 2017), learning word embeddings with visually-grounded semantics. Examples of related applications include image caption retrieval and generation (Kiros et al., 2014; Karpathy and Fei-Fei, 2015), and visual question-answering (Malinowski et al., 2015).

**Image-to-text translation** Canonical Correlation Analysis (CCA) (Hotelling, 1936) is a statistical method that projects two views linearly into a common space to maximize their correlation. (Andrew et al., 2013) proposes a deep learning framework to extend CCA so that it is able to learn nonlinear projections

and has better scalability on relatively large datasets.

In the state-of-the-art frameworks, the pairwise ranking is often adopted to learn a distance metric (Socher et al., 2014; Niu et al., 2017; Nam et al., 2017). (Frome et al., 2013) proposes a cross-modal feature embedding framework that uses CNN and Skip-Gram (Mikolov et al., 2013) to extract representations for images and texts respectively, then an objective is applied to ensure that the distance between the matched image-text pair is smaller than that between the mismatched pair. A similar framework proposed by (Kiros et al., 2014) uses a Gated Recurrent Unit (GRU) as the sentence encoder. (Wang et al., 2016) uses a bidirectional loss function with structure-preserving constraints. An attention mechanism on both image and caption is used by (Nam et al., 2017) where the model estimates the similarity between images and texts by sequentially focusing on a subset of image regions and words that have shared semantics. (Huang et al., 2017) utilizes a multi-modal context-modulated attention mechanism to compute the similarity between an image and a caption. (Faghri et al., 2017) proposes a novel loss to penalize the hard negatives, *i.e.*, the closest mismatched pairs, instead of averaging the individual violations across all negatives in (Kiros et al., 2014).

**Adversarial attack in text domain** Adversarial attacks have recently drawn significant attention in the deep learning community. The adversarial attacks spread over multiple domains including image classification (Nguyen et al., 2015), image segmentation and object detection (Xie et al., 2017), and textual reading comprehension (Jia and Liang, 2017), and deep reinforcement learning (Kos and Song, 2017).

In this paper, we present textual adversarial attacks in image-to-text translation systems such as image caption frameworks. While focusing on the problem of learning visually-grounded semantics, the adversarial attack brings new solutions to fulfill the gap between limited training data and numerous constitutions of natural languages. With extensive experiments on the effects of the adversarial samples, we reach the conclusion that current visual-semantic embeddings are “insensitive” to the underlying semantics. The proposed VSE-C shows advance across multiple visual-semantic tasks.

### 3 Method

#### 3.1 Preliminaries

**Word embeddings** We manually split the embeddings of each word into two parts: distributional embeddings, and visually-grounded embeddings. We use GloVe (Pennington et al., 2014) as the distributional embeddings, pre-trained unsupervisedly on large-scale corpora. We focus on the visually-grounded embeddings of words. The embeddings are optimized using the visual-semantic embedding (VSE) technique.

**Visual-semantic embeddings** VSE optimizes and aligns the latent space of both visual and textual domains. Parallel data are typically obtained from image captioning datasets such as Flickr30K (Young et al., 2014) or MS-COCO (Lin et al., 2014). The training set  $S = \{(i_n, c_n)\}_N$  contains  $N$  image-caption pairs. Typically all  $(i_n, c_m), n \neq m$  and  $(i_m, c_n), n \neq m$  form the negative samples for a specific pair  $(i_n, c_n)$ .

Following the notations used by (Kiros et al., 2014), domain-specific encoders are first employed to extract latent features of both images and captions, denoted as  $\phi(i)$  and  $\psi(c)$ , respectively. We use ResNet-152 (He et al., 2016) as visual domain encoder and GRU as text domain sentence encoder, which are both effective for VSE. They are projected into a joint latent space with a linear transformation. A hinge loss with margin  $\alpha$  is employed to optimize the alignment:

$$\ell^{VSE}(i, c) = \sum_{c'} [\alpha + s(i, c') - s(i, c)]_+ + \sum_{i'} [\alpha + s(i', c) - s(i, c)]_+, \quad (1)$$

where  $[\cdot]_+ = \max(0, \cdot)$ , and  $s(i, c) = W_i^T f(i; \theta_i) \cdot W_c^T g(c; \theta_c)$  measuring the distance between projected image embedding  $W_i^T f(i; \theta_i)$  and caption embedding  $W_c^T g(c; \theta_c)$ . The summations are taken over all image-caption pairs within a sampled batch.

| Class      | Original Caption                                    | Contrastive Adversarial Example                     |
|------------|-----------------------------------------------------|-----------------------------------------------------|
| Noun       | A person feeding a <b>cat</b> with a banana.        | A person feeding a <b>dog</b> with a banana.        |
| Numeral    | A person feeding <b>a cat</b> with a banana.        | A person feeding <b>five cats</b> with a banana.    |
| Relation-1 | <b>A person</b> feeding <b>a cat</b> with a banana. | <b>A cat</b> feeding <b>a person</b> with a banana. |
| Relation-2 | A person feeding a cat <b>with</b> a banana.        | A person feeding a cat <b>in</b> a banana.          |

Table 1: Examples of contrastive adversarial samples generated with our heuristic rules and knowledge from WordNet. The samples can be classified into four types: noun replacement, numeral replacement, relation shuffling, and relation replacement.

### 3.2 Generating contrastive adversarial samples

Our contrastive adversarial samples can be split into three classes: *noun*, *numeral* and *relation*. Each class of samples is generated separately.

**Noun.** We extract a list of heads (Zwicky, 1985) of noun phrases in MS-COCO dataset and label those with frequency larger than 200 be frequent heads. In addition, since images usually reflect concrete concepts better than abstract ones, we compute the concreteness of words following Turney et al. (2011), and only consider those heads with concreteness larger than  $\theta = 0.6$ . Only frequent concrete heads can be replaced by other frequent concrete heads with different meaning to form contrastive adversarial samples.

While replacing, we utilize the hypernymy/hyponymy relations in WordNet (Miller, 1995) to confirm the original noun and the corresponding contrastive adversarial sample are semantically different. Only words without hypernymy or hyponymy relations can be used as the replacement for adversarial sample generation. For example, “animal” is a hypernym of “cat”. Therefore, “A person feeding an animal with a banana” cannot be a valid generated contrastive adversarial caption for the image with the caption of “A person feeding a cat with a banana.”

**Numeral.** For each caption, we detect numerals and replace them with other numerals indicating different quantities to form contrastive adversarial samples. Note that “a” and “an” are treated as “one” here, though they are (indefinite) articles instead of numerals. Meanwhile, we singularize or pluralize the corresponding nouns when necessary.

**Relation.** The relation class includes two different paradigms.

The first one can be viewed as *shuffle of noninterchangeable noun phrases*. After extracting noun phrases of a caption, we shuffle them and put them back to the original positions. Although the bag of words features of the two sentences (caption) remain the same, the semantic meaning alters through this process.

The second one is *replacement of prepositions*. We extract the prepositions with frequency higher than 200 in MS-COCO dataset. Then we manually annotate a semantic overlap table, which can be found in Appendix A. In this table, words in the same set may have semantic overlap with each other, *e.g.*, by and with, in and among.

The noun phrase detection, preposition detection and numeral detection mentioned above are performed with SpaCy (Honnibal and Johnson, 2015). Examples of different classes of contrastive adversarial sample generation are shown in Table 1.

### 3.3 Intra-pair hard negative mining

We extend the online hard example mining (OHEM) technique used by VSE++(Faghri et al., 2017). The original hinge loss is computed by choosing the hardest sample within an batch (inter-pair). Mathematically,

$$\ell^{\text{VSE++}}(i, c) = \max_{c' \neq c} [\alpha + s(i, c') - s(i, c)] + \max_{i' \neq i} [\alpha + s(i', c) - s(i, c)]. \quad (2)$$

There are two major concerns regarding the in-batch hard negative mining. On one hand, mining negatives from a single batch is inefficient when batch size is not comparable with the size of the dataset.

On the other hand, for real-world datasets, taking the max in loss function tends to be very sensitive to label noise, resulting in fake negative samples.

In contrast, given an image-caption pair  $(i, c)$ , we employ human heuristics and WordNet knowledge base to generate contrastive negative samples  $\mathcal{C}'(c)$ . To utilize these candidate caption sets, we employ an intra-pair hard negative mining strategy. Specifically, during the optimization, we add an extra loss term:

$$\ell^{\text{VSE-C}}(i, c) = \ell^{\text{VSE++}}(i, c) + \max_{c'' \in \mathcal{C}'(c)} [\alpha + s(i, c'') - s(i, c)]_+. \quad (3)$$

In our implementation, the candidate set  $\mathcal{C}'$  has approximately 1,000 samples. In each iteration, we randomly sample  $N = 8$  negatives from it. This simple sample technique are effective and computation-friendly based on our empirical studies.

## 4 Experiments

We begin our experiments with an extensive study on the effect of adversarial samples on the baseline models. Even trained with hard negative mining techniques, VSE++ fails to discriminate words with completely contradictory visually-grounded semantics. Furthermore, we study the improvement brought by the introduction of contrastive adversarial samples on a diverse set of tasks. We release our code at <https://github.com/ExplorerFreda/VSE-C>.

### 4.1 Adversarial attacks

We select 1,000 images for test in MS-COCO 5k test split following (Karpathy and Fei-Fei, 2015). Each image is associated with five captions. Each caption in the selected test set can be manipulated to generate at least 20 contrastive adversarial samples by all manners (noun, numeral, and relation adversary). The image-to-caption retrieval task is defined as ranking the candidate captions based on the distance between their semantics and the given image.

We follow the metric used in Faghri et al. (2017) computing R@1, R@10, median rank and mean rank w.r.t. the top-ranked correct caption for each image. For each image, the database of retrieval contains the full set of  $1000 \times 5$  captions, in which only 5 captions are labeled as positive. The R@k metric essentially measures the percentage of images where the set of top-k ranked captions contains at least one positive caption.

We attack the existing models by adversarial samples. We extend each caption with 60 adversarial samples (20 noun-typed, 20 numeral-typed and 20 relation-typed). Therefore, each image has  $60 \times 5$  contrastive adversarial samples in total. The candidate retrieval set for each image now becomes  $5000 + 300$ . We discuss the experimental results as follows:

**VSE-C are more robust to known-typed adversarial attacks than VSE and VSE++.** We compare the performance of VSE-C with VSE (Kiros et al., 2014) and VSE++ (Faghri et al., 2017) in Table 2. Both VSE and VSE++ have a significant drop in performance after adding adversarial samples, while VSE-C training with contrastive adversarial samples is less vulnerable to the attacks. This phenomenon reflects that the text encoders of VSE and VSE++ do not actually make a good use of the image encodings, as the image encodings are fixed in all experiments.

Detailed attacking results are shown in Table 3. The three hyper-columns show the ability of the models to defend the adversarial attack of noun, numeral, and relation-typed respectively. Among three types of attacks, VSE and VSE++ suffer least from the noun attack. As the constitution of the dataset ensures the frequency in the entire dataset of the words used for replacement, the visual grounding of these frequent nouns is easy to obtain. However, the semantics of relations (including prepositions and entity relations) or numbers are not diverse enough in the dataset, leading to the poor performance of VSE against these attacks.

**Numeral-typed VSE-C improves the counting ability of models.** As shown in Table 3, numeral-typed contrastive adversarial samples improve the counting ability of models. However, it is still not clear about where the gain comes from, as the creation of numeral-typed samples may change the form (*i.e.*,

| Model         | MS-COCO Test |             |        |            | MS-COCO Test (w/. adversarial) |             |        |            |
|---------------|--------------|-------------|--------|------------|--------------------------------|-------------|--------|------------|
|               | R@1          | R@10        | Med r. | Mean r.    | R@1                            | R@10        | Med r. | Mean r.    |
| VSE           | 47.7         | 87.8        | 2.0    | 5.8        | 28.0                           | 71.6        | 4.0    | 11.7       |
| VSE++         | <b>55.7</b>  | <b>92.4</b> | 1.0    | <b>4.3</b> | 35.6                           | 72.5        | 3.0    | 11.8       |
| VSE-C (+n.)   | 50.7         | 90.7        | 1.0    | 5.2        | 40.3                           | 80.2        | 2.0    | 9.2        |
| VSE-C (+num.) | 53.3         | 90.2        | 1.0    | 5.8        | 46.9                           | 86.3        | 2.0    | 6.9        |
| VSE-C (+rel.) | 52.4         | 89.0        | 1.0    | 5.7        | 42.3                           | 82.5        | 2.0    | 7.2        |
| VSE-C (+all)  | 50.2         | 89.8        | 1.0    | 5.2        | <b>47.4</b>                    | <b>88.8</b> | 2.0    | <b>5.5</b> |

Table 2: Evaluation on image-to-caption retrieval. Although VSE++ (Faghri et al., 2017) obtains the best performance on original MS-COCO test set, it is more vulnerable to the caption-specific adversarial attack compared with the proposed VSE-C, and so does VSE (Kiros et al., 2014).

| Model         | MS-COCO Test (+n.) |             |            | MS-COCO Test (+num.) |             |            | MS-COCO Test (+rel.) |             |            |
|---------------|--------------------|-------------|------------|----------------------|-------------|------------|----------------------|-------------|------------|
|               | R@1                | R@10        | Mean r.    | R@1                  | R@10        | Mean r.    | R@1                  | R@10        | Mean r.    |
| VSE           | 37.6               | 85.8        | 6.9        | 38.5                 | 82.3        | 7.7        | 30.7                 | 76.7        | 8.8        |
| VSE++         | 45.7               | 89.1        | 5.5        | 45.9                 | 82.3        | 7.2        | 42.3                 | 80.0        | 7.6        |
| VSE-C (+n.)   | 49.2               | 88.4        | 5.7        | 42.1                 | 80.3        | 9.1        | 40.4                 | 83.3        | 7.1        |
| VSE-C (+num.) | <b>51.0</b>        | <b>89.5</b> | 6.1        | <b>53.3</b>          | <b>90.2</b> | 5.8        | 49.0                 | 87.0        | 6.6        |
| VSE-C (+rel.) | 48.0               | 88.8        | <b>5.3</b> | 45.4                 | 83.9        | 6.7        | <b>50.1</b>          | <b>90.2</b> | <b>4.9</b> |
| VSE-C (+all.) | 49.4               | 89.3        | <b>5.3</b> | 49.9                 | 89.6        | <b>5.2</b> | 47.9                 | 89.4        | 5.3        |

Table 3: Detailed results on each type of adversarial attack. Training VSE-C on one class gains the best performance on robustness against the adversarial attack of the class itself. In addition, training with numeral-typed adversarial samples helps improve the robustness against noun-typed and relation-typed attack. We hypothesis that this is attributed to the singularization or pluralization of the corresponding nouns in the process of numeral-typed adversarial sample generation.

| Model         | MS-COCO Test (plural split, +plurals) |             |            |
|---------------|---------------------------------------|-------------|------------|
|               | R@1                                   | R@10        | Mean r.    |
| VSE++         | 43.7                                  | 78.3        | 9.1        |
| VSE-C (+num.) | <b>50.6</b>                           | <b>84.4</b> | <b>7.8</b> |

Table 4: Results on plural-typed adversarial attack to the plural split of MS-COCO test set. This split consists of 205 images, together with the 1,025 original captions. VSE-C outperforms VSE++ by a large margin on all the three considered metrics.

singular or plural) of nouns to make a sentence plausible. Does the gain comes from the improved ability to distinguish singulars and plurals?

We conduct the following evaluation to study the counting ability. We extract all the images associated with captions including plurals from our test split of MS-COCO, forming a plural-split of the dataset, and generate only plural (numeral)-typed contrastive adversarial samples (changing the numerals) w.r.t. the plurals in the captions. We report the performance of VSE++ and numeral-typed VSE-C on this plural split in Table 4. It clearly shows that what VSE-C does not only distinguish singulars against plurals, but also, at least, distinguish plurals against other plurals (*e.g.*, 3 vs. 5).

It is worth noting that such counting ability is still not evaluated completely due to the limitation of the current MS-COCO test split. We find that 99.8% of the plurals in MS-COCO test set comes from one of “two”, “three”, “four” and “five”. This may reduce the counting problem to a much simpler classification one.

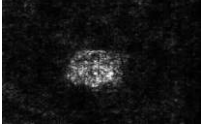
|                        |                                                                                   |       |                                                                                                                            |
|------------------------|-----------------------------------------------------------------------------------|-------|----------------------------------------------------------------------------------------------------------------------------|
| Original Image         |  |       | an elephant walking through the weeds in the forest                                                                        |
| Image Saliency (VSE-C) |  | VSE++ | an elephant walking through the weeds in the <u>bird</u><br>0.021 0.134 0.061 0.067 0.029 0.248 0.063 0.065 0.313          |
|                        |                                                                                   | VSE-C | an elephant walking through the weeds in the <u>bird</u><br>0.031 0.136 0.056 0.083 0.050 0.192 0.080 0.069 0.299          |
| Image Saliency (VSE-C) |  | VSE++ | <u>eighteen</u> elephants walking through the weeds in the forest<br>0.192 0.203 0.069 0.094 0.051 0.198 0.036 0.037 0.121 |
|                        |                                                                                   | VSE-C | <u>eighteen</u> elephants walking through the weeds in the forest<br>0.306 0.112 0.087 0.087 0.040 0.101 0.079 0.041 0.136 |
| Image Saliency (VSE-C) |  | VSE++ | an elephant walking <u>against</u> the weeds in the forest<br>0.039 0.176 0.101 0.087 0.051 0.248 0.060 0.057 0.181        |
|                        |                                                                                   | VSE-C | an elephant walking <u>against</u> the weeds in the forest<br>0.030 0.108 0.125 0.258 0.108 0.176 0.077 0.027 0.090        |

Figure 2: Saliency analysis on adversarial samples. The left column shows the saliency of VSE-C on the image (what is the difference between the image and the image you imagine from caption  $c'$ ), while the right column shows the saliency of both VSE++ and VSE-C on the caption (what is the difference between the caption and the caption you summarize from image  $i$ ). The magnitude of values indicates the level of saliency. For better visualization, the image saliency is  $L_{inf}$ -normalized and the caption saliency is  $L_1$ -normalized. For all the three classes of textual adversarial samples, the image encoding model (ResNet-152) almost only focuses on the main part of the image, *i.e.*, elephant. For numeral-typed and relation-typed adversarial samples, VSE-C pays much more attention to the manipulated segments of the sentence than VSE++.

## 4.2 Saliency visualization

Given an image-caption pair and its corresponding textual adversarial samples, we are interested in the following question: what is the semantic distance between the image and an adversarial caption? In other words, *which part in the image or caption, in particular, makes them semantically different?*

We visualize the saliency on input images and captions w.r.t. changes in sentence semantics. Specifically, given an image-caption pair  $(i, c)$ , we manually modify the semantics of the caption  $c$  with the techniques introduced in Section 3.2, and obtain  $c' \neq c$ . We compute the saliency of  $i$  or  $c'$  w.r.t. this change by visualizing the Jacobian:

$$\mathbf{J} = \nabla_i s(i, c') = \nabla_i W_i^T f(i; \theta_i) \cdot W_c^T g(c'; \theta_c), \quad (4)$$

where  $s(i, c')$  is the similarity metric for image-caption pairs.

Shown in Figure 2, as for captions, VSE-C captures the change in sentence semantics and thus possesses large saliency on the manipulated words. In contrast, although trained with hard-negative mining, it is difficult for VSE++ to capture differences other than nouns.

Interestingly, the saliency of images shows less correlated response to semantics changes while the replaced word is not the major component in the image. We attribute this to the image embedding extractor, ResNet, because it is pre-trained on the ImageNet classification task. As the ResNet learns to produce shift-invariant features focusing on the major components (or concepts) of images, it inevitably learns less about secondary (and other) concepts.

## 4.3 Correlate words and objects

As only textual adversarial samples are provided during the training, the model may overfit the training samples by memorizing incorrect co-occurrence of words or concepts. To quantitatively evaluate the

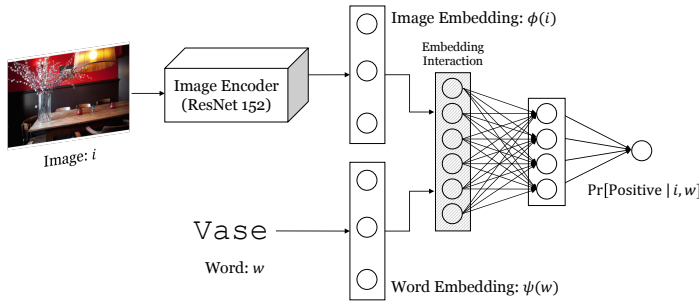
| Image                                                                         | Captions                                                                                                       |
|-------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|
|                                                                               | A <u>table</u> with a huge glass <u>vase</u> and fake <u>flowers</u> come out of it.                           |
|                                                                               | A <u>plant</u> in a <u>vase</u> sits at the end of a <u>table</u> .                                            |
|                                                                               | A <u>vase</u> with <u>flowers</u> in it with long <u>stems</u> sitting on a <u>table</u> with <u>candles</u> . |
|                                                                               | A large <u>centerpiece</u> that is sitting on the <u>edge</u> of a dining <u>table</u> .                       |
|                                                                               | <u>Flowers</u> in a clear <u>vase</u> sitting on a <u>table</u> .                                              |
| <b>Positive Objects:</b> table, plant, vase.                                  |                                                                                                                |
| <b>Negative Objects:</b> screen, pickle, sandwich, toy, hill, coat, cat, etc. |                                                                                                                |

Table 5: An example of the image-to-word retrieval dataset. We extract objects by detecting heads of noun phrases in captions. We only collect the “object” words with frequency higher than 200 in MS-COCO full dataset as available positive/negative objects for each image.

learned word embeddings, we conduct experiments on word-level image-to-word retrieval. Specifically, we first examine how each noun is linked with a visual object. This task shows the concrete link between words and image concepts, which supports the effectiveness of adversarial samples in enforcing the learning of visually-grounded semantics beyond co-occurrence memorizing.

**Dataset** Based on captions, we extract *positive objects* for each image in MS-COCO dataset by detecting heads of noun phrases using SpaCy. As mentioned in Section 3.2, we only let those objects without direct hypernymy/hyponymy relation to positive objects of the image be *negative objects* to avoid ambiguity. Table 5 shows an example of the preparation of image-object dataset.

**Training** Inspired by Gong et al. (2018), we train an image-word alignment network through the interaction space, since this structure reflects the property of “alignment” better than just concatenating the feature vectors of word and image. In the training stage, the network is fed by batches of samples in the form of (image, word, label), where the label is 0 or 1, indicating whether the word is a negative or positive object of the image. Let  $\mathbf{v}_W(w)$  denote the embedding of word  $w$  and  $\mathbf{v}_I(img)$  denote the feature vector of image  $img$  extracted by ResNet-152 (He et al., 2016). As shown in Figure 3, we use the full interaction matrix  $\mathbf{v}_W(w)\mathbf{v}_I(img)^T$  as the feature for object retrieval. While fixing both the image and word features, we only tune the parameters of multi-layer perceptron (MLP).



| Model         | MAP          |
|---------------|--------------|
| GloVe         | 58.69        |
| VSE           | 61.65        |
| VSE++         | 61.08        |
| VSE-C (+all)  | 62.16        |
| VSE-C (+n.)   | <b>62.80</b> |
| VSE-C (+rel.) | 62.29        |
| VSE-C (+num.) | 61.96        |

Figure 3: Model structure for image-to-word retrieval. The network is trained through the interaction space. Note that only parameters in MLP are tuned during training.

Table 6: Evaluation result (MAP in percentage) on image-to-word retrieval.

**Testing** We use mean average precision (MAP), which is a widely-applied metric in information retrieval, to evaluate the performance of the word embeddings. For each image, we treat it as a query. The average precision (AP) is defined by

$$AP = \frac{\sum_{k=1}^n P(k) \times \text{positive}(k)}{\text{number of positive objects}} \quad (5)$$

where  $n$  is the quantity of objects in data base, *i.e.*, both positive and negative objects,  $P(k)$  is the precision at cut-off  $k$  in the list,  $\text{positive}(k)$  is an indicator function equaling 1 if the object at rank  $k$  is a positive one, 0 otherwise (Turpin and Scholer, 2006).



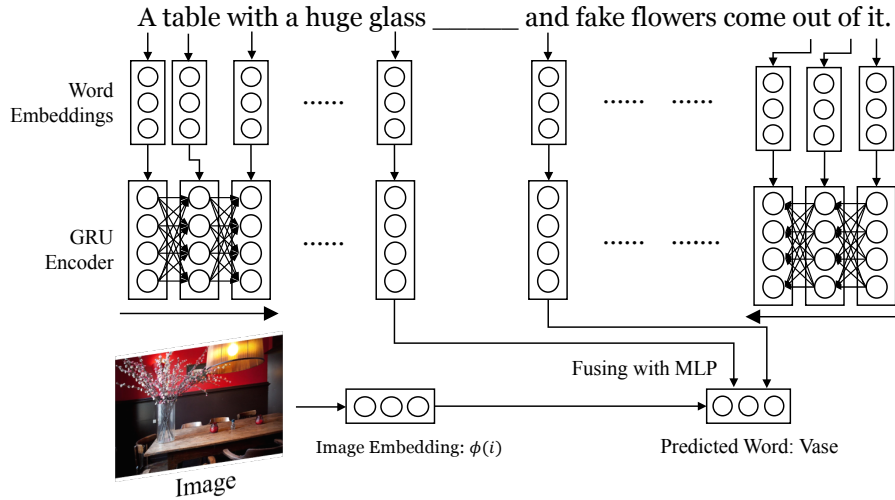


Figure 4: Model structure for fill-in-the-blank.

Based on the definition of AP, MAP can be computed by  $MAP = \frac{\sum_{i=1}^{|Q|} AP(i)}{|Q|}$ , where  $Q$  is the query set, *i.e.*, image set. It is worth noting that the database for retrieval of each query may be different from others, which is similar to Section 4.1.

**Results** We show the evaluation results in Table 6. It is as expected that VSE-C (+n.) achieves the best performance in the image-object retrieval task. All of the VSE-C models outperform the baselines produced by VSE (Kiros et al., 2014), VSE++ (Faghri et al., 2017) and GloVe (Pennington et al., 2014), showing the concrete link between learned word semantics and visual concepts. With surprise, VSE-C with only relation adversarial samples shows comparable performance as VSE-C with noun adversarial samples. This further supports the effectiveness of sentence-level manipulation (relation-shuffle in Figure 1) on strengthening the link.

#### 4.4 Concept to word

We quantitatively evaluate the performance of concept-to-word retrieval performance by introducing a sentence completion task. Given an image-caption pair  $(i, c)$ , we manually replace concept words (nouns and relational words) with blanks. A separate model is trained to fill in the blanks.

**Dataset and implementation details** Based on captions, we extract nouns and relational words from captions for each image in MS-COCO dataset using SpaCy. These selected words are marked as “concept” representatives. During training, we randomly sample a word from the representative set, and the word is masked as a blank to be filled. Given the image and the rest of the words, the model is trained to predict the embedding of the word.

**Model** The sentences with blank are encoded by two mono-directional GRU layer. The words before the blank and after the blank are separately encoded using  $GRU_f$  and  $GRU_b$  respectively. The image feature extracted from a pre-trained ResNet152 is then concatenated with the last output of both GRUs. The prediction of embedding is made by a two-layer MLP taking in the concatenated feature. We use cosine similarity as the loss function. Figure 4 shows the demonstration of our fill-in-the-blank model.

**Results** We present in Table 7 the performance of the proposed VSE-C on filling in both nouns and prepositions. VSE-based models outperform GloVe without visual grounding and concretely correlate word semantics with image embeddings. We found that only small gaps exist between VSE++ and VSE-C on preposition filling, which again shows the limited diversity on visual relations within the dataset.

| Model        | Noun Filling |              | Prep. Filling |              | All (n. + prep.) |              |
|--------------|--------------|--------------|---------------|--------------|------------------|--------------|
|              | R@1          | R@10         | R@1           | R@10         | R@1              | R@10         |
| GloVe        | 23.22        | 58.75        | 23.34         | 79.92        | 23.27            | 66.55        |
| VSE++        | 24.98        | 61.74        | 34.88         | 84.88        | 28.44            | 68.12        |
| VSE-C (ours) | <b>27.30</b> | <b>62.87</b> | <b>35.17</b>  | <b>85.24</b> | <b>30.02</b>     | <b>70.98</b> |

Table 7: Evaluation result on the fill-in-the-blank task (in percentage). The word embeddings learned by VSE-C with all classes of contrastive adversarial samples help reach a better performance than those learned by VSE++ (Faghri et al., 2017).

## 5 Discussion and conclusion

In this paper, we focus on the problem of learning visually-grounded semantics using parallel image-text data. With extensive experiments on adversarial attacks against existing frameworks (Kiros et al., 2014; Faghri et al., 2017), we obtain new insights on the limitation of datasets as well as frameworks. (1) Even for large-scale datasets such as MS-COCO captioning, the large gap between the number of possible constitutions of real-world visual semantics and the size of dataset still exists. (2) Existing models are not powerful enough to fully capture or extract the information contained in visual embeddings.

We propose VSE-C, introducing contrastive adversarial samples in the text domain and an intra-pair hard-example mining technique. To delve deeper into the embedding space and its transferability, we study a set of multi-modal tasks both qualitatively and quantitatively. Beyond being robust to adversarial attacks on image-to-caption retrieval tasks, experimental results on image-to-word retrieval and fill-in-the-blank reveal the correlation between the learned word embeddings and visual concepts.

VSE-C also demonstrates a general framework for augmenting textual inputs considering semantical consistency. The introduction human priors and knowledge bases alleviates the sparsity and non-contiguity of languages. We hope the framework and the released data are beneficial for building more robust and data-efficient models.

## References

- Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep Canonical Correlation Analysis. In *Proc. of ICML*.
- Lluís Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Learning Aligned Cross-Modal Representations from Weakly Aligned Data. In *Proc. of CVPR*.
- Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. In *Proc. of ACL*.
- Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long Short-Term Memory-Networks for Machine Reading. In *Proc. of EMNLP*.
- Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. VSE++: Improved Visual-Semantic Embeddings. *arXiv preprint arXiv:1707.05612*.
- Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A Deep Visual-Semantic Embedding Model. In *Proc. of NIPS*.
- Yichen Gong, Heng Luo, and Jian Zhang. 2018. Natural Language Inference over Interaction Space. In *Proc. of ICLR*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proc. of CVPR*.
- Matthew Honnibal and Mark Johnson. 2015. An Improved Non-Monotonic Transition System for Dependency Parsing. In *Proc. of EMNLP*.
- Harold Hotelling. 1936. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.
- Yan Huang, Wei Wang, and Liang Wang. 2017. Instance-Aware Image and Sentence Matching with Selective Multimodal LSTM. In *Proc. of CVPR*.
- Hakan Inan, Khashayar Khosravi, and Richard Socher. 2017. Tying Word Vectors and Word Classifiers: A Loss Framework for Language Modeling. In *Proc. of ICLR*.
- Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proc. of EMNLP*.
- Andrej Karpathy and Li Fei-Fei. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *Proc. of CVPR*.
- Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proc. of ICLR*.
- Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *arXiv preprint arXiv:1411.2539*.
- Jernej Kos and Dawn Song. 2017. Delving into Adversarial Attacks on Deep Policies. *arXiv preprint arXiv:1705.06452*.
- Ankit Kumar, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, and Richard Socher. 2016. Ask Me Anything: Dynamic Memory Networks for Natural Language Processing. In *Proc. of ICML*.
- Yangyan Li, Hao Su, Charles Ruizhongtai Qi, Noa Fish, Daniel Cohen-Or, and Leonidas J Guibas. 2015. Joint embeddings of shapes and images via cnn image purification. *ACM Trans. Graph.*
- Hang Li. 2011. Learning to rank for information retrieval and natural language processing. *Synthesis Lectures on Human Language Technologies*, 4(1):1–113.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proc. of ECCV*.
- Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask Your Neurons: A Neural-Based Approach to Answering Questions about Images. In *Proc. of ICCV*.
- Junhua Mao, Jiajing Xu, Kevin Jing, and Alan L Yuille. 2016. Training and Evaluating Multimodal Word Embeddings with Large-Scale Web Annotated Images. In *Proc. of NIPS*.

- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.
- George A Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*.
- Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual Attention Networks for Multimodal Reasoning and Matching. In *Proc. of CVPR*.
- Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal Deep Learning. In *Proc. of ICML*.
- Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In *Proc. of CVPR*.
- Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. 2017. Hierarchical Multimodal LSTM for Dense Visual-Semantic Embedding. In *Proc. of CVPR*.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proc. of EMNLP*.
- Scott Reed, Zeynep Akata, Bernt Schiele, and Honglak Lee. 2016. Learning Deep Representations of Fine-Grained Visual Descriptions. In *Proc. of CVPR*.
- Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. 2017. Reasonet: Learning to stop reading in machine comprehension. In *Proc. of SIGKDD*.
- Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded Compositional Semantics for Finding and Describing Images with Sentences. *TACL*.
- Peter D Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and Metaphorical Sense Identification through Concrete and Abstract Context. In *Proc. of EMNLP*.
- Andrew Turpin and Falk Scholer. 2006. User Performance versus Precision Measures for Simple Search Tasks. In *Proc. of SIGIR*.
- Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning Deep Structure-Preserving Image-Text Embeddings. In *Proc. of CVPR*.
- Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. 2017. Adversarial examples for semantic segmentation and object detection. In *Proc. of ICCV*.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *TACL*.
- Will Y Zou, Richard Socher, Daniel Cer, and Christopher D Manning. 2013. Bilingual Word Embeddings for Phrase-Based Machine Translation. In *Proc. of EMNLP*.
- Arnold M Zwicky. 1985. Heads. *Journal of Linguistics*.

## A Semantic Overlap Table of Frequent Prepositions

| Set # | Words in a Semantical Set                                  |
|-------|------------------------------------------------------------|
| 1     | towards, toward, beyond, to                                |
| 2     | behind, after, past                                        |
| 3     | outside, out                                               |
| 4     | underneath, under, beneath, down, below                    |
| 5     | on, upon, up, un, atop, onto, over, above, beyond          |
| 6     | in, within, among, at, during, into, inside, from, between |
| 7     | if, while                                                  |
| 8     | with, by, beside                                           |
| 9     | around, like                                               |
| 10    | to, for, of                                                |
| 11    | about, within                                              |
| 12    | because, as, for                                           |
| 13    | as, like                                                   |
| 14    | near, next, beside                                         |
| 15    | though                                                     |
| 16    | thru, through                                              |
| 17    | besides, along                                             |
| 18    | against, next, to                                          |
| 19    | along, during, across, while                               |
| 20    | off, out                                                   |
| 21    | without                                                    |
| 22    | than                                                       |
| 23    | before                                                     |

Table 8: Manually annotated semantic overlap table.

Table 8 shows our manually annotated semantic overlap sets. Prepositions in each row has overlap in semantics, *i.e.*, can be replaced by each other in some level. A preposition can appear in several sets.

## B Training Details of VSE-C

In all experiments, we use Adam (Kingma and Ba, 2015) as the optimizer of which the learning rate is set to  $1e-3$ , with the batch size of 128. The learning rate is updated by multiplying 0.1 after every 15 epochs. We do not apply any regularization or dropout term. Word embeddings are initialized with the 300-dimensional GloVe (Pennington et al., 2014)<sup>1</sup>. The text encoder is a bidirectional 512-dimensional (in total 1024D) 1-layer GRU. The dimensionality of joint (multimodal) embedding is also 1,024. Empirically, with training data and hyper-parameters fixed, there is no significant variance in performance caused by different random seeds for the sampling.

<sup>1</sup><http://nlp.stanford.edu/data/glove.840B.300d.zip>