# Towards Unsupervised Image Captioning with Shared Multimodal Embeddings

Iro Laina
Technische Universität München
iro.laina@tum.de

Christian Rupprecht
University of Oxford
chrisr@robots.ox.ac.uk

Nassir Navab
Technische Universität München
nassir.navab@tum.de

## Abstract

*Understanding images without explicit supervision has become an important problem in computer vision. In this paper, we address image captioning by generating language descriptions of scenes without learning from annotated pairs of images and their captions. The core component of our approach is a shared latent space that is structured by visual concepts. In this space, the two modalities should be indistinguishable. A language model is first trained to encode sentences into semantically structured embeddings. Image features that are translated into this embedding space can be decoded into descriptions through the same language model, similarly to sentence embeddings. This translation is learned from weakly paired images and text using a loss robust to noisy assignments and a conditional adversarial component. Our approach allows to exploit large text corpora outside the annotated distributions of image/caption data. Our experiments show that the proposed domain alignment learns a semantically meaningful representation which outperforms previous work.*

## 1. Introduction

Generating natural language descriptions for images has gained attention as it aims to teach machines how humans see, understand and talk about the world. Assisting visually impaired people [23, 62] and human-robot interaction [12, 39] are some examples of the importance of image captioning. Even though it is straightforward for humans to describe the contents of a scene, machine generation of image descriptions is a challenging problem that requires compositional perception of images translated into semantically and grammatically correct sentences.

Traditionally, image captioning has been carried out using full supervision in the form of image-caption pairs, given by human annotators. Crowd-sourcing captions is a cumbersome task that requires extensive quality control and further manual cleaning. Since annotators are often paid per image, the captions tend to be short and repetitive. In addition, current captioning benchmarks [38, 49] consist of a limited number of object categories and are focused on performance under imperfect evaluation metrics. Thus, methods developed on such datasets might not be easily adopted in the wild. Nevertheless, great efforts have been made to extend captioning to out-of-domain data [3, 9, 69] or different styles beyond mere factual descriptions [22, 55].

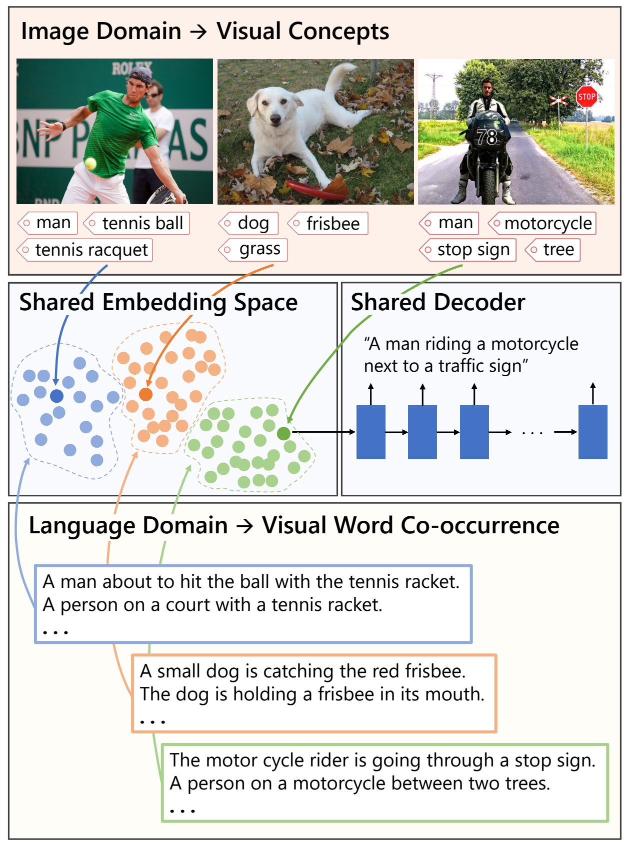



Figure 1. **Method overview.** Our model learns a joint embedding space of language and image features that is structured by visual concepts and their co-occurrence. Images and text come from disjoint sources. During inference, model embeds images into the shared space from which a caption can be decoded.

In this work we explore *unsupervised* captioning, where image and language sources are independent. The unsupervised setting can benefit from an almost unlimited amount of unlabeled or weakly labeled images as well as readily available large text corpora, without the need of bias-prone and costly human annotations. Although significant progress has been achieved in other unsupervised tasks [15, 34, 54, 70], unsupervised generation of image descriptions remains mostly unexplored.

The building blocks of our method are a language model and the translation of the image to the language domain. On the language side, we first learn a semantically structured embedding space, *i.e*. sentences describing similar visual concepts (*e.g*. *woman* and *person*) and similar context are encoded with similar embeddings. We then perform a weakly supervised domain alignment between image features and the learned text embeddings leveraging visual concepts in the image. This alignment allows to exploit co-occurrence statistics of visual concepts between sentences and images. For example, the words *boat* and *water* might often appear together in the language domain, similar to the fact that most images that contain a boat also contain water.

When language and images come from different sources, some *weak* supervisory signal is needed to align the manifold of visual concepts to the textual domain. Similar to previous work [18], we use a pre-trained object detector to generate an initial noisy alignment between the text source and visual entities that can be detected in the image.

We show that we can indeed learn to predict meaningful captions for images that extend beyond the limited capabilities of the object detector. Due to visual concept co-occurrence, the model learns to produce text descriptions including concepts that are not necessarily contained in the object detector's fixed set of labels (*e.g*. *beach*). This shows that the alignment is meaningful and the statistics of both domains help to discover more visual concepts. Quantitatively, our unsupervised approach nearly matches the performance of some early supervised methods and outperforms previous unsupervised methods. Finally, our approach makes it possible to leverage various language sources, for instance from a different language or with a particular style —poetic (Shakespeare), funny, story-telling— that cannot be easily obtained by crowdsourcing.

## 2. Related Work

**Fully supervised.** Pioneering work in neural-based image captioning [27, 60] established the commonly used framework of a Convolutional Neural Network (CNN) image encoder, followed by a Recurrent Neural Network (RNN) language decoder. There has been significant progress improving over the standard CNN-RNN approach. Xu *et al*. [63] introduced the concept of attention to image captioning and, subsequently, several methods focused on attention mechanisms to visualize the grounding of words on image context and effectively guide the generation process [4, 41, 64, 68]. Noteworthy efforts also include generating video descriptions [14] or dense captions on image regions [26], exploiting additional information such as attributes [67] or visual relationships [66] and optimizing evaluation metrics [40, 50]. Other methods focus on generating diverse and natural captions with adversarial models [11, 36, 53, 61], moving beyond just factual descriptions [19, 55] or addressing gender bias [5].

**Novel object captioning.** Recent approaches have also explored the task of novel object captioning to exploit large-scale visual object recognition from readily available datasets, such as ImageNet [51]. Their goal is to address the limitations of conventional models in integrating new entities into image descriptions without explicit training pairs. In [45] the problem is addressed by learning from few labeled pairs for novel categories. Copying mechanisms are employed in [6, 65] to transfer knowledge from the paired data to out-of-domain objects, while [59] jointly exploits semantic information from independent images and text sources. Another approach is to produce sentence templates and fill in the slots with detected concepts [42]. Instead of training the model to handle new concepts, [2] proposes to constrain beam search evaluation on target words.

**Partial supervision.** Recent work has further advanced the field towards generating image descriptions under more challenging settings, for example unpaired or unsupervised.

Chen *et al*. [9] address cross-domain captioning, where the source domain consists of image-caption pairs and the goal is to leverage unpaired data from a target domain through a critic. In [69], the cross-domain problem is addressed with a cycle objective. Similarly, unpaired data can be used to generate stylized descriptions [22, 46]. Anderson *et al*. [3] propose a method to complete partial sequence data, *e.g*. a sequence of detected visual concepts, without the need for paired image-caption datasets. Gu *et al*. [20] address unpaired image captioning from a different perspective, using an intermediary language where paired data is available, and then translating the captioner to the target language using parallel corpora. However, the goal of these methods is different to ours, as they typically align a target domain that contains limited paired or unpaired data with a source domain. A generic image captioner is first built from *full* supervision in the source domain and then adapted to a different language domain or novel object categories.

Most closely related to our work is [18] which does not require any image-sentence pairs. In this case, it is optimal to use a language domain which is rich in visual concepts. Therefore, their (and our) goal is to exploit image and language sources that are disjoint yet compatible, instead of aligning different language sources as in cross-domain approaches. Supervision comes in only through image recognition models, which are used to detect objects in the image.

**Multimodal embeddings.** A key component of our approach is the alignment of latent representations from two independent modalities. In unsupervised machine translation, although unimodal, [34, 35] create a shared latent space (interlingua) for both source and target languages. Kiros *et al*. [29] pose captioning as a translation problem and learn a multimodal embedding space that also allows them to perform vector arithmetics. Similarly, joint embedding spaces have been used in [16] for cross-modality retrieval and in [47] for video captioning. Finally, Fang *et al*. [17] predict visual words from images to produce caption candidates and use the similarity between images and sentences in a joint space to rank the captions.

## 3. Methods

An overview of our method is shown in Figure 2. The proposed approach consists of two components, a language model and a domain alignment model between images and text. The language model independently encodes samples from the language domain into a semantic-aware representation. The goal of the domain alignment is to translate image respresentations into the embedding space learned by the language model and decode these embeddings into meaningful image descriptions. In absence of paired image-caption data this is a challenging task.

We consider a visual domain $\mathcal{I}$ and an image $I_i \in \mathcal{I}$, represented by the set of visual entities that it encloses:

$$\mathcal{V}_i = \{v_k \mid k \in \mathbb{N},\ 1 \leq k \leq N_i\}, \tag{1}$$

where $i$ iterates over the total number of image samples and $N_i$ is the total number of visual concepts in image $i$.

Similarly, in the language domain $\mathcal{L}$, a text sequence $s_j \in \mathcal{L}$ can be described by a bag of words

$$\mathcal{W}_j = \{w_k \mid k \in \mathbb{N},\ 1 \leq k \leq M_j\}, \tag{2}$$

where $j$ enumerates sequences of length $M_j$.

For the purpose of this work, we assume that the image and language domains are not entirely disjoint. For example, it would seem unreasonable to attempt describing natural images based on text corpora of economics. Thus, we assume a universal set of concepts $\Omega = \mathcal{V} \cap \mathcal{W}$ that language and images have in common. We refer to joint concepts, such as *person*, as visual concepts.

### 3.1. Language Model

To create a basis for domain alignment, our first step is to create a meaningful textual domain. We learn an unsupervised sentence embedding by training a language model on the text corpus, following a standard sequence-to-sequence approach with maximum likelihood estimation [57]. The encoder $f$ embeds an input sentence $s$ into a $d$-dimensional latent representation which is reconstructed back into the same sentence by a decoder $g$:

$$f(s) = \phi, \quad g(\phi) = \tilde{s}, \quad \phi \in \Phi \subseteq \mathbb{R}^d. \tag{3}$$

RNNs are the most common choice for $f$ and $g$. Typically, language models of this structure are trained by minimizing the negative log-likelihood between $s$ and $\tilde{s}$ per word.

A model without any constraints on the latent space would learn a grammatical and syntactic embedding. Instead, we are primarily interested in creating a representation that encodes visual semantics. This means that we have to encourage the model to learn a manifold structured by visual concepts. As we show later, our representation encodes strong semantic properties in the sense that sentences with similar contents have a low distance in the embedding space. Since our goal is image captioning, our notion of *similar sentence contents* stems from visual concepts — words in a sentence that have visual grounding— and their co-occurrence statistics. We impose a visual concept-based structure on the manifold of $\phi$ with a triplet loss, defined as

$$L_t(\phi, \phi^+, \phi^-) = \max(0,\ \|\phi - \phi^+\|_2^2 - \|\phi - \phi^-\|_2^2 + m) \tag{4}$$

that operates on triplets of embeddings $\phi$. The loss is minimized when the distance from an anchor embedding $\phi$ to a *positive pair* $\phi^+$ is smaller than the distance to a *negative pair* $\phi^-$ by at least a margin $m \in \mathbb{R}^+$.

The positive and negative pairs can be defined based on the visual concepts that exist in the sentences. For a given sentence $s_j$ we define the set of negative pairs $\mathcal{S}_j^-$ as the set of sentences that do not have any concepts in common

$$\mathcal{S}_j^- = \{s_k \mid k \in \mathbb{N},\, \mathcal{W}_k \cap \mathcal{W}_j = \emptyset\}\,. \tag{5}$$

Analogously, we define the set of positive pairs $\mathcal{S}_j^+$ as the set of sentences that have at least *two* concepts in common

$$\mathcal{S}_j^+ = \{s_k \mid k \in \mathbb{N},\, k \neq j,\, |\mathcal{W}_k \cap \mathcal{W}_j| \geq 2\}\,. \tag{6}$$

We ignore sentence pairs that only have one overlapping concept to reduce bad alignments. For example, since many language datasets are human-centered, every sentence involving a person would be a positive pair to each other regardless of the context. The language model's total loss is

$$L_{\mathrm{LM}}(s_j) = L_{\mathrm{CE}}(\, g(\phi),\, s_j\,) \,+\, \lambda_t\, L_t(\phi_j, \phi_j^+, \phi_j^-). \tag{7}$$

During training, a positive sentence $s^+ \in \mathcal{S}_j^+$ is sampled from a multinomial distribution with probability proportional to the number of overlapping concepts. This favors positive pairs of sentences with many similar concepts. We sample a negative sentence $s^-$ uniformly from $\mathcal{S}_j^-$.

The triplet loss imposes a visually aware structure on the embedding space. Sentences with similar visual contents are encouraged to be close to each other, while sentences with different context will be pushed apart. This

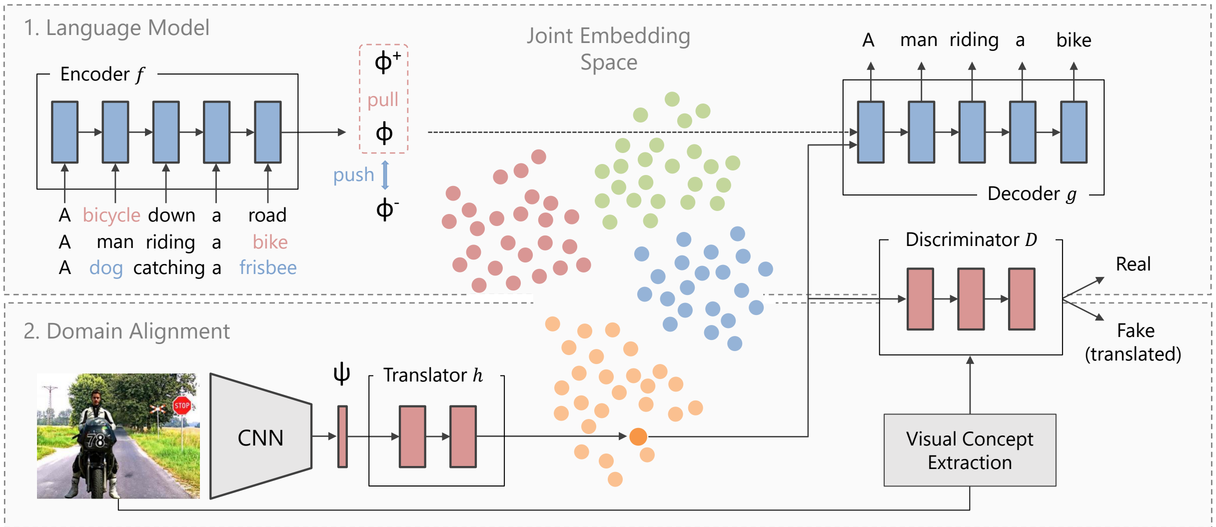



Figure 2. **Unsupervised image captioning architecture.** We first learn a language model with a triplet loss formulation that structures the embedding $\phi$ using visual concepts from the sentences. We then learn a mapping from images to the embedding space using a robust alignment scheme and adversarial training in feature space.

external emphasis on structure is important, since unconstrained language models are more likely to group sentences with similar words and grammar. Intuitively, generating image descriptions relies on visual content and thus the structured embedding space is presumably a more meaningful basis for the task at hand. A comparison between the visually constrained and unconstrained embedding space can be found in the supplementary material.

### 3.2. Joint Image and Language Domain

We have learned an encoder that projects text into a structured embedding. The next step is to project image features into the same embedding space so that they can be similarly decoded into sentences by the decoder. To do this, we need an initial alignment between the independent image and text sources for which we rely on the visual concepts they have in common. We build a bipartite graph $\mathcal{G}(\mathcal{L}, \mathcal{I}, P)$ with images $I_i$ and sentences $s_j$ as nodes. The edges $P_{i,j}$ represent weak assignments between $I_i$ and $s_j$, weighted by the number of overlapping concepts

$$P_{i,j} = |\mathcal{V}_i \cap \mathcal{W}_j|. \tag{8}$$

During training, for $I_i$ we sample $s_j$ with probability

$$p(s_j \mid I_i) = P_{i,j} \left( \sum_k P_{i,k} \right)^{-1}. \tag{9}$$

For sentence-image pairs without overlap $p(s_j \mid I_i) = 0$ and they are excluded from training. Highly visually correlated pairs will be sampled with higher probability. At this point, we have created a stochastic training set, which we could use to train a standard captioning model by sampling an image-caption pair at each iteration. Training this model with teacher forcing alone, collapses to certain caption-modes describing sets of images.

Visual concepts can be extracted from the images using any pretrained image recognition method. However, this would often result in only a limited number of categories. To lexically enrich the search space for matching sentences, we also query hyponyms of the predicted visual concepts $\mathcal{V}_i$, *i.e.* words among the text source concepts $\mathcal{W}_i$ that have a kind-of relationship with the predicted concepts (for example, man to person, puppy to dog).

### 3.3. Learning the Semantic Alignment

The initial alignment allows us to learn a mapping from images to text. We extract image features $\psi_i$ from $I_i$ using a standard pretrained CNN. The task is now to translate between the image feature domain $\psi_i \in \Psi$ to the visually structured text domain $\phi_j \in \Phi$. The stochastic alignment graph $\mathcal{G}$ is expected to be very noisy and full of imprecise correspondences. We thus propose a robust training scheme to exploit the underlying co-occurence information while ignoring problematic matches. We learn the translation function $h : \Psi \to \Phi$, where $h$ can be a simple multi-layer perceptron (MLP), using the correspondences $(s_j, I_i)$ and the following objectives.

**Robust Alignment.** If we train the alignment using a simple $L_2 = \sum_j \|h(\psi_i) - \phi_j\|_2^2$ loss the optimal mapping $h^*$ would be the conditional average $h^*(\psi_i) = \sum_j p(\phi_j \mid I_i)\, \phi_j$ which might not be an optimal or verbally rich sentence embedding as it could land between modes of the distribution.

Thus, we propose to learn the feature alignment using a robust formulation that encourages the mapping to be close to a real sentence embedding:

$$L_R(\psi_i) = \min_{\phi_j \sim p(s_j|I_i)} \| h(\psi_i) - \phi_j \|_2^2 . \qquad (10)$$

Since the set of matches is very large, we approximate the loss by sampling a fixed amount $K$ of $\phi_j$ for each image and by computing the minimum in this subset.

**Adversarial Training.** So far, the robust alignment encourages to learn a translation $h$ that adheres to the structure of the conceptual text embedding. However, we need to ensure that the mapping does not discard important concept information from the image feature vector. This is necessary so that the decoder can decode a caption that directly corresponds to the visual concepts in the image. To this end, we employ adversarial training using a conditional discriminator. Since adversarial training on discrete sequences is problematic [8, 56], we perform it in feature space $\Phi$ similar to [56]. The discriminator $D : \Phi \times \Omega \rightarrow \mathbb{R}$ is trained with a set of positive/real and a set of negative/fake examples. In our case a positive example is the concatenation of a translated feature $h(\psi_i)$ with the one-hot encoding of the image concepts $\mathcal{V}_i$. A negative example analogously is the concatenation of the sampled pair's text embedding $\phi_j$ and the image concepts $\mathcal{V}_i$. Thus, the discriminator learns the correlation of image concepts and text embeddings, which in turn encourages the mapping $h$ to encode image concepts correctly. Otherwise the discriminator can easily identify a real sentence feature from a translated image feature.

In practice, we use a WGAN-GP formulation [21] to train the discriminator $D$ to maximize its output for fake examples and minimize it for real. When training $h$ we thus maximize the discriminator for the translation.

$$L_{\text{adv}} = -D(h(\psi_i), \mathcal{V}_i) \qquad (11)$$

**Total loss.** Our final model is trained with all three aforementioned objectives:

$$L_{total} = \lambda_{CE} \, L_{\text{CE}} + \lambda_R \, L_R + \lambda_{\text{adv}} \, L_{\text{adv}}, \qquad (12)$$

where the weight factors $\lambda_{CE}, \lambda_R, \lambda_{\text{adv}} \in \mathbb{R}$ balance the contributions of the three losses.

## 4. Experiments and Results

The evaluation is structured as follows. First, we present ablation experiments in an *unpaired* setting on Microsoft COCO [38] to evaluate the effect of each component of our method. Second, we report the results in the *unsupervised* setting with independent image and language sources. We experiment with Flickr30k Images [49] paired with COCO captions and COCO images paired with Google's Conceptual Captions dataset (GCC) [52]. Finally, we show qualitative results for image descriptions with varying text sources.

**Implementation details.** We tokenize and process all natural language datasets, replacing the least frequently used words with `unk` tokens. The next step is to extract visual word synsets. We use the Visual Genome [31] object synsets as reference and look up nouns (or noun phrases) extracted by parsing each sentence with the Stanford CoreNLP toolkit [44]. This results in 1415 synsets for COCO and 3030 synsets for GCC which describe visual entities. During the semantic-aware training of the language model with Equation 4, positive and negative pairs of captions are defined using this synset vocabulary.

The encoder and decoder of the language model are implemented using Gated Recurrent Units (GRUs) [10] with 200 hidden units. The last hidden state of the encoder is projected through a linear layer into 256-d text features $\phi$. The decoder is followed by a linear layer that maps its output into a fixed-size vocabulary vector. We use 200-d GloVe embeddings [48] as inputs to the language model.

Similar to sentence pairs, we build weak image-sentence assignments based on (visual) synsets to train the image captioner. For richness in visual concepts, we use the OpenImages-v4 dataset [30, 33], which consists of 1.74 million images and 600 annotated object categories. Visual concepts are extracted using a Faster R-CNN detector [25] trained on OpenImages, which has been made publicly available[1]. Please note that we only make use of class labels and do not rely on image regions (bounding boxes) in order to keep the amount of supervision minimal. Thus, any multi-label classifier could be used instead.

The baseline for our image captioner is based on [60] and uses image features extracted by ResNet-101 [24] pretrained on ImageNet, without finetuning. The translator $h$ is implemented with a single-layer MLP of size 512 to map $\psi \in \mathbb{R}^{2048}$ into $\phi \in \mathbb{R}^{256}$.

**Training details.** We train the language model until convergence with a batch size of 64. The initial learning rates of the encoder and decoder are set to $10^{-4}$ and $10^{-3}$ respectively and $\lambda_t = 0.1$. When training the the alignment model, we further finetune the decoder so that it adapts to the joint embedding space. We optimize using Adam [28] with a learning rate of $10^{-3}$ and $\lambda_{CE} = \lambda_R = 1$, $\lambda_{\text{adv}} = 0.1$.

**Evaluation metrics.** We evaluate our method with the official COCO evaluation code and report performance under the commonly used metrics, BLEU 1-4 [10], ROUGE [37], METEOR [13], CIDEr [58], SPICE [1] and WMD [32].

### 4.1. Unpaired Captioning

The unpaired setting on COCO allows us to evaluate the effectiveness of the proposed method and to compare to previous work [18] using the same controlled setup. This is a

[1] https://github.com/tensorflow/models/tree/master/research/object_detection

| Component Evaluation | | | | | Metrics | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Abbreviation | $L_{\text{CE}}$ | $L_2$ | $L_R$ | $L_{\text{adv}}$ | B-1 | B-2 | B-3 | B-4 | METEOR | ROUGE | CIDER | SPICE | WMD |
| Supervised baseline | | | | | 67.4 | 50.0 | 35.4 | 24.8 | 22.6 | 50.1 | 80.2 | 15.9 | 17.9 |
| Oracle | | | | | 49.1 | 31.2 | 21.2 | 16.0 | 18.7 | 38.7 | 50.4 | 12.2 | 14.5 |
| Alignment only | | ✓ | | | 47.0 | 25.4 | 11.5 | 5.2 | 15.5 | 35.9 | 29.4 | 8.7 | 9.1 |
| MLE only | ✓ | | | | 59.9 | 40.2 | 26.0 | 17.1 | 19.1 | 43.7 | 57.9 | 11.6 | 13.0 |
| Joint, baseline | ✓ | ✓ | | | 59.7 | 40.2 | 25.8 | 16.6 | 18.3 | 43.1 | 53.8 | 10.8 | 12.6 |
| Joint, robust | ✓ | | ✓ | | 61.5 | 42.3 | 28.0 | 18.8 | 19.7 | 44.9 | 62.4 | 12.5 | 14.3 |
| Joint, robust ($\lambda_t = 0$) | ✓ | | ✓ | | 60.7 | 41.1 | 26.7 | 17.6 | 18.3 | 43.8 | 55.6 | 11.0 | 13.0 |
| Joint, adversarial | ✓ | | ✓ | ✓ | **61.7** | **42.8** | **28.6** | **19.3** | **20.1** | **45.4** | **63.6** | **12.8** | **14.4** |

Table 1. Ablation Experiments on COCO test set [27]. Image and language data are unpaired; COCO ground truth object categories are used for the initial alignment. Every component of our domain alignment model improves the performance on the captioning task.

simplification of the problem since the images and their descriptions come from the same distribution; however, we do not use the ground truth correspondences and treat images and text unpaired. We use the same data splits as in previous methods following [27], resulting in 113,287 training, 5,000 validation and 5,000 test images. Each image is originally annotated with 5 descriptions, resulting in over 560k training captions. After generating our initial image-caption assignments based on visual synsets, there are approximately 150k unique captions remaining in the graph $\mathcal{G}$.

**Ablation study.** We evaluate the proposed components through ablation experiments (Table 1). In these experiments, we use the 80 available COCO object categories as visual concepts. We compare the following models.

*Oracle:* We first evaluate the weak assignments using an oracle that selects the highest probability candidate among the ground truth captions assigned to an image. This candidate has the highest overlap of visual concepts with the image. Since there can be multiple captions with equally high probability, we randomly sample and report the best out of 100 runs. This baseline scores generally low as the initial assignments are very noisy.

*Alignment only:* The alignment is performed by training only the mapping $h$ of image features into the sentence manifold. We keep the decoder frozen, using the weights from the pretrained language model. The model shows understanding of the major visual concepts in the scene, meaning that relevant classes appear in the output sentence. However, the sentences are grammatically incoherent because the decoder cannot adapt to the latent space difference between the projected image features and the real sentence embeddings it was trained with. Thus, for subsequent experiments we also jointly finetune the decoder.

*MLE only:* The full model is trained using the weak pairs of image-captions and teacher forcing, in the standard supervised manner, but without any constraints to encourage a shared domain. The model is prone to the bias often seen in MLE models such as repeating sub-phrases.

*Joint, baseline:* In addition to MLE training, domain alignment is performed by minimizing the $L_2$-distance between $h(\psi)$ and $\phi$. This naïve alignment of the two domains does not improve over the MLE-only baseline.

*Joint, robust:* Instead of L2, the model is trained with the proposed robust alignment loss (10) which gives a significant boost in performance. We randomly sample $K = 10$ sentences as candidate pairs for each training image.

*Joint, robust ($\lambda_t = 0$):* To evaluate the importance of the embedding space, we also train the above model against sentence embeddings that come from a language model trained only with $L_{\text{LM}}{:=}L_{\text{CE}}$, *i.e.* without the triplet loss. It performs worse, suggesting that the semantic structure of the language model is indeed beneficial for captioning.

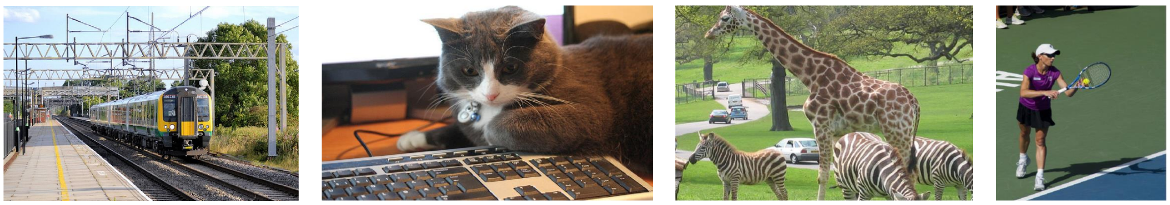
*Joint, adversarial:* The full model additionally includes adversarial training conditioned on visual concepts as categorical inputs. We observe that our unpaired model reaches performance close to its fully supervised counterpart [60] and is comparable to early work on image captioning.

The consistent improvement shows that our model is able to learn concepts beyond the initial weak assignments.

**Comparison to the State of the Art.** The field of image captioning without image-caption pairs has only been explored very recently. In Table 2, we compare our approach to previous methods. We follow the same *unpaired* setup on COCO as in [18]. We use the object detector trained on OpenImages (OID) to predict visual concepts for both creating the image-caption assignments and conditioning the discriminator during adversarial training. The reported results correspond to the predictions from our full model trained with $K = 10$ samples and evaluated using a beam size of 3. Our method sets a new state of the art on this problem.

**Qualitative Evaluation.** We show qualitative results of our full model in Figure 3, comparing captions predicted in the unpaired setting with two variants trained with different visual concept extractors (COCO and OID). We find

| | | | | |
|---|---|---|---|---|
| COCO | a teddy bear is sitting on a chair | a large passenger jet sitting on top of an airport tarmac | a building with a clock on the side of it | a traffic light on the side of the street |
| OID | a cat lying on a bed with a teddy bear | a large airplane flying through the sky | a large building with a clock tower on it | a traffic light and a street sign on a pole |
| GCC | a teddy bear with a cat | a passenger plane on the runway | a clock tower in the city | traffic sign on a street |
| VQA | is the cat in the tub? | is this plane in the air? | what time is on the clock? | what color is the traffic light on the right? |



| | | | | |
|---|---|---|---|---|
| COCO | a passenger train that is travelling down the tracks | a cat sitting on top of a laptop keyboard | a giraffe is standing in the grass near a tree | a man holding a tennis racket on a tennis court |
| OID | a train travelling down train tracks next to a tree | a cat laying on top of a laptop on a desk | a giraffe standing in the grass near a tree | a woman holding a tennis racket on a tennis court |
| GCC | train station | a cat with a laptop | a zebra in a giraffe | tennis player and a ball during a match |
| VQA | is the train on the track? | is the cat on the laptop? | is the giraffe standing straight? | is the tennis player wearing a headband? |

Figure 3. **Qualitative Results.** We show caption predictions on images from the COCO dataset. COCO and OID are results from our *unpaired* model trained with weak pairs coming from a detector trained on the respective dataset. GCC and VQA refer to the *unsupervised* model trained on COCO images using the Conceptual Captions and VQA-v2 datasets respectively.

that both the COCO model and the OID model capture the image contents well, whereas the OID model clearly benefits from the richer object detections. For example, in the last image the COCO model produces a description about a *man* —potentially due to bias. This is because only *person* is a category in COCO, but not *man* or *woman*, and therefore there can be no gender distinction in the captions that are weakly assigned to each image. The model trained with OID concepts has the capacity to resolve such ambiguities and correctly identifies *woman* in the last image. We note that the object detector is only used during training (for the weak assignments and the discriminator), but not during inference. The captioner learns to extrapolate from the labeled categories of the image domain; *e.g.* the generated words {tracks, airport, tower, passenger, grass} are unlabeled concepts that the model inferred due to co-occurrence with labeled concepts such as train, airplane, clock, etc.

## 4.2. Unsupervised Captioning

When training the image captioner in an unsupervised manner, the language model is pre-trained using an external text source and all other settings remain identical. We perform two cross-domain experiments: COCO images with GCC sentences and Flickr30k images with COCO captions. Quantitative results can be seen in Table 3 for the model variants with and without adversarial training. Adversarial training consistently improves our model. Naturally, we do not expect to match the performance of the unpaired setting since a different language domain implies vocabulary, context and style that differs from the *ground truth* captions in COCO.

Qualitatively, we show the predicted captions of the model trained on COCO images and GCC captions in Figure 3 (denoted as GCC). When using GCC as the language domain, we find that the initial image-caption assignments are even more noisy, which leads the model to produce short and simple descriptions. However, we also see that this model has learned some interesting concepts, not present in the unpaired setting, such as the difference between a plane being on the ground or in the air.

To produce descriptions with different styles that extend beyond captioning datasets, the choice of the language domain is not trivial, as it should be rich in visual descriptions. We thus experiment with VQA-v2 [7] as the language domain, using the questions provided by the dataset as the sentence source. Instead of captioning, the model learns to ask questions about the image content (Figure 3, VQA).

## 4.3. Joint Embedding Visualization

Finally, to verify that our training creates a meaningful *joint* latent space, we visualize the $t$-SNE embedding [43] of both the sentences (marked with [L]) and image-projected features ([I]) in Figure 4. The overall embedding is structured by visual categories due to the constraints we impose on the model during training. Within clusters, image and text features are well mixed. This means that the model has learned a joint embedding where it is not possible to separate text form images.

| Method | Metrics | | | | |
|---|---|---|---|---|---|
| | B-4 | M | R | C | S |
| Gu *et al.* [20] | 5.4 | 13.2 | - | 17.7 | - |
| Feng *et al.* [18] | 18.6 | 17.9 | 43.1 | 54.9 | 11.1 |
| Ours | **19.3** | **20.2** | **45.0** | **61.8** | **12.9** |

Table 2. Comparison with the state of the art on COCO test set [27] under the *unpaired* setting of [18]. OpenImages [30] categories are used for concept extraction.

| Method | Metrics | | | | |
|---|---|---|---|---|---|
| | B-4 | M | R | C | W |
| **Flickr Images ↔ COCO Captions** | | | | | |
| Ours (w/o adv) | 5.9 | 10.9 | 31.1 | 8.2 | 7.0 |
| Ours | 7.9 | 13.0 | 32.8 | 9.9 | 7.5 |
| **COCO Images ↔ Conceptual Captions** | | | | | |
| Ours (w/o adv) | 5.5 | 11.1 | 30.1 | 20.8 | 6.7 |
| Ours | 6.5 | 12.9 | 35.1 | 22.7 | 7.4 |

Table 3. Evaluation under the unsupervised setting using image and captions from independent sources.

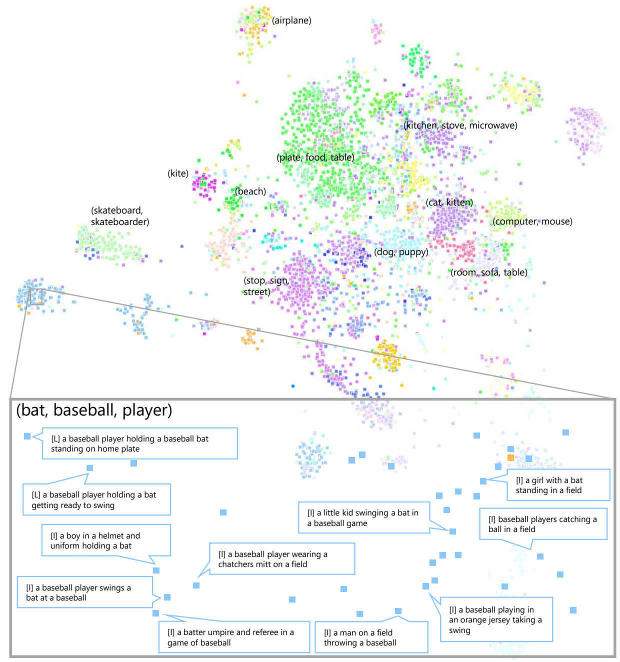



Figure 4. ***t*-SNE Embedding.** We show a projection of the learned joint embedding of our model and zoom into a cluster to visualize that sentences from the text corpus (denoted by [L]) lie in visual-semantic groups together with image embeddings [I]. Colors are generated by groups of visual concepts. A large-scale version of this figure can be found in the supplementary material.

## 5. Limitations and Discussion

Although our approach sets the state of the art in unsupervised image captioning, there are still several limitations. As mentioned before, to generate the initial assignments, the language source needs to contain sufficient visual concepts overlapping with the image domain. We believe it is possible to alleviate this problem by learning from a combination of text sources with varying contents and styles.

Another limitation is the capability of the model to extend to novel compositions and atypical scene descriptions. We observe two factors that decide the model's behavior in this respect. First, the capabilities of the base captioner itself, *i.e.* unsupervised training will not solve limitations that are present even for the supervised model [60]. In our experiments, the output often collapses into caption modes that are generic enough to describe a set of images; this results in approximately 20% of the generated captions actually being unique and 16% novel captions, not found in the training set. This is on par with the findings of [60].

The second factor is the amount of *discoverable* visual concepts. For example, it is not possible to discover the difference between a whole pizza and a slice of pizza, when only the concept *pizza* is known, unless *slice* also appears in other context. Naturally, learning from more concepts holds the potential for more diversity. One could enrich the search space of weak assignments by including predicates in the set of known visual concepts, thus relying on relationship detection. This could greatly help in resolving ambiguities such as *a person riding a bike* or *carrying a bike*, however it goes against the idea of weak or no supervision.

## 6. Conclusion

We have presented a novel method to align images and text in a shared latent repesentation that is structured through visual concepts. Our method is minimally supervised in the sense that it requires a standard, pre-trained image recognition model to obtain initial noisy correspondences between the image and the text domain. Our robust training scheme and the adversarial learning of the translation from image features to text allows the model to successfully learn the captioning task. In our experiments we show different combinations of image and text sources and improve the state of the art in the unpaired COCO setting.

For the future we are interested in investigating several directions. One could improve the decoder architecture with typical components, such as attention, or follow a template approach to encourage novel compositions of objects. Overall, unsupervised image captioning is an upcoming research direction that is gaining traction in the community.

**Acknowledgements.** Christian Rupprecht is supported by ERC Stg Grant IDIU-638009.

## References


[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016. 5

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. In *EMNLP*, 2017. 2

[3] Peter Anderson, Stephen Gould, and Mark Johnson. Partially-supervised image captioning. In *Advances in Neural Information Processing Systems*, pages 1879–1890, 2018. 1, 2

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 2

[5] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787, 2018. 2

[6] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2016. 2

[7] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015. 7

[8] Liqun Chen, Shuyang Dai, Chenyang Tao, Haichao Zhang, Zhe Gan, Dinghan Shen, Yizhe Zhang, Guoyin Wang, Ruiyi Zhang, and Lawrence Carin. Adversarial text generation via feature-mover's distance. In *Advances in Neural Information Processing Systems*, pages 4671–4682, 2018. 5

[9] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 521–530, 2017. 1, 2

[10] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 5

[11] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional GAN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2970–2979, 2017. 2

[12] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017. 1

[13] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 5

[14] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 2

[15] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016. 2

[16] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 3

[17] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015. 3

[18] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. *arXiv preprint arXiv:1811.10787*, 2018. 2, 5, 6, 8

[19] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. StyleNet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146, 2017. 2

[20] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. Unpaired image captioning by language pivoting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 503–519, 2018. 2, 8

[21] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017. 5

[22] Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. MSCap: Multi-style image captioning with unpaired stylized text. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2

[23] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018. 1

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[25] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7310–7311, 2017. 5

[26] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. DenseCap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016. 2

[27] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2, 6, 8

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[29] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 3

[30] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. OpenImages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017. 5, 8

[31] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 5

[32] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966, 2015. 5

[33] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982*, 2018. 5

[34] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations (ICLR)*, 2018. 2, 3

[35] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 3

[36] Dianqi Li, Xiaodong He, Qiuyuan Huang, Ming-Ting Sun, and Lei Zhang. Generating diverse and accurate visual captions by comparative adversarial learning. *arXiv preprint arXiv:1804.00861*, 2018. 2

[37] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004. 5

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 5

[39] Huan Ling and Sanja Fidler. Teaching machines to describe images via natural language feedback. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5075–5085. Curran Associates Inc., 2017. 1

[40] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881, 2017. 2

[41] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017. 2

[42] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7219–7228, 2018. 2

[43] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 7

[44] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014. 5

[45] Junhua Mao, Xu Wei, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan L Yuille. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *Proceedings of the IEEE international conference on computer vision*, pages 2533–2541, 2015. 2

[46] Alexander Mathews, Lexing Xie, and Xuming He. SemStyle: Learning to generate stylised image captions using unaligned text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8591–8600, 2018. 2

[47] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4594–4602, 2016. 3

[48] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 5

[49] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1, 5

[50] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017. 2

[51] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 2

[52] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2556–2565, 2018. 5

[53] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4135–4144, 2017. 2

[54] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[55] Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. Engaging image captioning via personality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12516–12526, 2019. 1, 2

[56] Sandeep Subramanian, Sai Rajeswar Mudumba, Alessandro Sordoni, Adam Trischler, Aaron C Courville, and Chris Pal. Towards text generation with adversarially learned neural outlines. In *Advances in Neural Information Processing Systems*, pages 7562–7574, 2018. 5

[57] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014. 3

[58] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 5

[59] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5753–5761, 2017. 2

[60] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 2, 5, 6, 8

[61] Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Advances in Neural Information Processing Systems*, pages 5756–5766, 2017. 2

[62] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1180–1192. ACM, 2017. 1

[63] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 2

[64] Zhilin Yang, Ye Yuan, Yuexin Wu, William W Cohen, and Ruslan R Salakhutdinov. Review networks for caption generation. In *Advances in Neural Information Processing Systems*, pages 2361–2369, 2016. 2

[65] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Incorporating copying mechanism in image captioning for learning novel objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6580–6588, 2017. 2

[66] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 684–699, 2018. 2

[67] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4894–4902, 2017. 2

[68] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 2

[69] Wei Zhao, Wei Xu, Min Yang, Jianbo Ye, Zhou Zhao, Yabing Feng, and Yu Qiao. Dual learning for cross-domain image captioning. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 29–38. ACM, 2017. 1, 2

[70] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 2