Removing Word-Level Spurious Alignment between Images and Pseudo-Captions in Unsupervised Image Captioning

Ukyo Honda[1,3] **Yoshitaka Ushiku**[2] **Atsushi Hashimoto**[2]
Taro Watanabe[1] **Yuji Matsumoto**[3]
[1] Nara Institute of Science and Technology [2] OMRON SINIC X Corp.
[3] RIKEN Center for Advanced Intelligence Project
[1] {honda.ukyo.hn6, taro}@is.naist.jp
[2] {yoshitaka.ushiku, atsushi.hashimoto}@sinicx.com
[3] yuji.matsumoto@riken.jp

Abstract

Unsupervised image captioning is a challenging task that aims at generating captions without the supervision of image–sentence pairs, but only with images and sentences drawn from different sources and object labels detected from the images. In previous work, *pseudo-captions*, *i.e.*, sentences that contain the detected object labels, were assigned to a given image. The focus of the previous work was on the alignment of input images and pseudo-captions at the sentence level. However, pseudo-captions contain many words that are irrelevant to a given image. In this work, we investigate the effect of removing mismatched words from image–sentence alignment to determine how they make this task difficult. We propose a simple gating mechanism that is trained to align image features with only the most reliable words in pseudo-captions: the detected object labels. The experimental results show that our proposed method outperforms the previous methods without introducing complex sentence-level learning objectives. Combined with the sentence-level alignment method of previous work, our method further improves its performance. These results confirm the importance of careful alignment in word-level details.[1]

1 Introduction

Image captioning is a task to describe images in natural languages. This is a fundamental challenge with regard to automatically retrieving and summarizing the visual information in a human-readable form. Recently, considerable progress has been made (Vinyals et al., 2015; Xu et al., 2015; Anderson et al., 2018b) owing to the development of neural networks and a large number of annotated image–sentence pairs (Young et al., 2014; Lin et al., 2014; Krishna et al., 2017). However, these pairs are limited in their coverage of scenes[2], and scaling them is difficult owing to the cost of manual annotation.

Unsupervised image captioning (Feng et al., 2019) aims to describe scenes that have no corresponding image–sentence pairs, without requiring additional annotation of the pairs. The only available resources are images and sentences drawn from different sources and object labels detected from the images. Although it is highly challenging, unsupervised image captioning has the potential to cover a broad range of scenes by exploiting a large number of images and sentences that are not paired by expensive manual annotation.

To train a captioning model in this setting, previous work (Feng et al., 2019; Laina et al., 2019) employed sentences that contained the object labels detected from given images. We refer to these sentences as *pseudo-captions*. However, pseudo-captions are problematic in that they are likely to contain words that are irrelevant to the given images. Assume that an image contains two objects *cat* and *girl* (Figure 1). This situation could give rise to various possible pseudo-captions, *e.g.*, "a girl is holding a cat," "a cat is sleeping with a girl," "a girl is running with a cat." When the first sentence is the correct caption of the image, the words *sleeping* and *running* of the other sentences are irrelevant to the image. As the detected object labels provide insufficient information to judge which sentence corresponds to the image, many pseudo-captions containing such mismatched words can be produced.

[1] Code will be available at https://github.com/ukyh/RemovingSpuriousAlignment

[2] For example, the standard captioning dataset MS COCO (Lin et al., 2014) covers only approximately 100 object categories out of 500 object categories defined in an object detection dataset (Agrawal et al., 2019). In addition to objects, attributes and relations are also not covered well owing to the small vocabulary size, 8791 (Karpathy and Fei-Fei, 2015).

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3692–3702
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

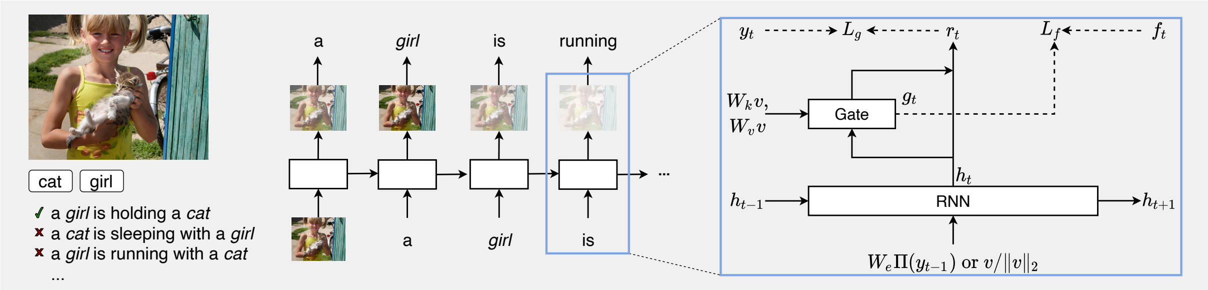

cat
girl
✓ a *girl* is holding a *cat*
✗ a *cat* is sleeping with a *girl*
✗ a *girl* is running with a *cat*
...
a
girl
is
running
a
girl
is
...
y_t
L_g
r_t
L_f
f_t
$W_k v$,
$W_v v$
Gate
g_t
h_t
h_{t-1}
RNN
h_{t+1}
$W_e \Pi(y_{t-1})$ or $v/\|v\|_2$

Figure 1: Overview of our model. The input is listed on the left-hand side: an image, its detected object labels, and its pseudo-captions. The model learns to generate the pseudo-captions while considering the correspondence between the image and each word being generated. The detailed process is shown in the blue box on the right-hand side. The base encoder–decoder model output $\boldsymbol{h}_t$, a gate value g_t, and a pseudo-label f_t on the gate are described in Sections 2.1, 2.2, and 2.3, respectively. The dashed arrows indicate the processes conducted only during training.

Regardless of the problem in pseudo-captions, previous work (Feng et al., 2019; Laina et al., 2019) did not explicitly remove word-level mismatches. They tried aligning the features of images and their pseudo-captions at the sentence level. Although this line of approach can potentially align the images and sentences correctly if there are sentences that exactly describe each image, it is not likely to hold for the images and sentences retrieved from different sources.

To shed light on the problem of word-level spurious alignment in the previous work, we focus on removing mismatched words from image–sentence alignment. To this end, we introduce a simple gating mechanism that is trained to exclude image features when generating words other than the most reliable words in pseudo-captions: the detected objects. The experimental results show that the proposed method outperforms previous methods without introducing complex sentence-level learning objectives. Combined with the sentence-level alignment method of previous work, our method further improves its performance. These results confirm the importance of careful alignment in word-level details.

2 Method

Our model comprises a sequential encoder–decoder model, a gating mechanism on the encoder–decoder model, a pseudo-label on the gating mechanism, and a decoding rule to avoid the repetition of object labels, as presented in Figure 1.

2.1 Base Encoder–Decoder Model

Typical supervised, encoder–decoder captioning models maximize the following objective function during training:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_{(\boldsymbol{I},\boldsymbol{y})} \log p(\boldsymbol{y}|\boldsymbol{I};\boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\theta}$ are the parameters of the models, $\boldsymbol{I}$ is a given image, and $\boldsymbol{y} = y_1, ..., y_T$ is its corresponding caption, the last token y_T is a special end-of-sentence token.

However, in unsupervised image captioning, the corresponding caption $\boldsymbol{y}$ is not available. Instead, object labels in given images are provided by pre-trained object detectors. Previous work utilized the detected object labels to assign a roughly corresponding caption $\hat{\boldsymbol{y}}$, *i.e.*, a pseudo-caption, to the given image. Following the previous work, we define pseudo-captions of an image as sentences containing the object labels detected from the image. Given the pseudo-caption $\hat{\boldsymbol{y}}$, our base encoder–decoder model maximizes the following objective function:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_{(\boldsymbol{I},\hat{\boldsymbol{y}})} \log p(\hat{\boldsymbol{y}}|\boldsymbol{I};\boldsymbol{\theta}). \tag{2}$$

In encoder–decoder captioning models, the probability $p(\boldsymbol{y}|\boldsymbol{I})$[3] is auto-regressively factorized as $p(\boldsymbol{y}|\boldsymbol{I}) = \prod_{t=1}^{T} p(y_t|y_{<t}, \boldsymbol{I})$ and each $p(y_t|y_{<t}, \boldsymbol{I})$ is computed by a single step of recurrent neural networks (RNNs). The encoder encodes the given image $\boldsymbol{I}$ to an image representation $\boldsymbol{v} \in \mathbb{R}^{d'}$ that is fed to the decoder as an initial input to generate a sequence of words auto-regressively. The detailed

[3] Hereafter, we omit the model parameter $\boldsymbol{\theta}$ for brevity.

computation of $p(\hat{y}_t|\hat{y}_{<t}, \boldsymbol{I})$ is as follows:

$$p(\hat{y}_t|\hat{y}_{<t}, \boldsymbol{I}) = \frac{\exp(\boldsymbol{h}_t^\top \boldsymbol{W}_o \Pi(\hat{y}_t))}{\sum_{y' \in \mathcal{Y}} \exp(\boldsymbol{h}_t^\top \boldsymbol{W}_o \Pi(y'))}, \quad (3)$$

$$\boldsymbol{h}_t = \begin{cases} \text{Dec}\left(\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}, \boldsymbol{h}_0\right), & \text{if } t = 1; \\ \text{Dec}(\boldsymbol{e}_t, \boldsymbol{h}_{t-1}), & \text{otherwise}, \end{cases} \quad (4)$$

$$\boldsymbol{v} = \boldsymbol{W}_a \text{Enc}(\boldsymbol{I}), \quad (5)$$

$$\boldsymbol{e}_t = \boldsymbol{W}_e \Pi(\hat{y}_{t-1}), \quad (6)$$

where Enc$(\cdot)$ is a pre-trained image encoder with a linear transformation matrix $\boldsymbol{W}_a \in \mathbb{R}^{d \times d'}$ on top of it, Dec$(\cdot)$ is an RNN decoder, $\Pi(\cdot)$ is the one-hot encoding function, $\boldsymbol{h}_0 \in \mathbb{R}^d$ is a zero vector, $\mathcal{Y}$ is the whole vocabulary to use, and $\boldsymbol{W}_e, \boldsymbol{W}_o \in \mathbb{R}^{d \times |\mathcal{Y}|}$ are the word embedding matrices. Details of the encoder and decoder are provided in Section 3.2.

2.2 Gating Mechanism to Consider Word-Level Correspondence

As indicated in Eq. 2, our base encoder–decoder model decodes all of the words in pseudo-captions from the images. However, pseudo-captions are highly likely to contain words that are irrelevant to the given images. Thus, forcing a model to decode the pseudo-captions in their entirety from the images might be more disadvantageous than beneficial for training precise captioning models.

To enable our model to handle word-level mismatches, we introduce a simple gating mechanism. Our model, which is equipped with this gating mechanism, takes an image representation at each t-th time step. The gate is designed to control the amount of image representation used to generate the t-th word. In other words, the gate is expected to determine the extent to which the given image corresponds to the t-th word. With a slight modification to Eq. 3, our model with the gating mechanism is defined as follows:

$$p(\hat{y}_t|\hat{y}_{<t}, \boldsymbol{I}) = \frac{\exp(\boldsymbol{r}_t^\top \boldsymbol{W}_o \Pi(\hat{y}_t))}{\sum_{y' \in \mathcal{Y}} \exp(\boldsymbol{r}_t^\top \boldsymbol{W}_o \Pi(y'))}, \quad (7)$$

$$\boldsymbol{r}_t = g_t \frac{\boldsymbol{W}_v \boldsymbol{v}}{\|\boldsymbol{W}_v \boldsymbol{v}\|_2} + (1 - g_t)\boldsymbol{h}_t, \quad (8)$$

$$g_t = \text{sigmoid}(\tanh(\boldsymbol{W}_k \boldsymbol{v})^\top \boldsymbol{h}_t), \quad (9)$$

where $\boldsymbol{W}_k, \boldsymbol{W}_v \in \mathbb{R}^{d \times d}$ are the linear transformation matrices for computing the gate value $g_t \in [0, 1]$ and the output of the gate $\boldsymbol{r}_t \in \mathbb{R}^d$. When g_t is close to one, it forces the model to use more information from the image ($\boldsymbol{v}$) to generate the t-th word; when g_t is close to zero, it forces the model to do the opposite.

The fed image representation $\boldsymbol{W}_v \boldsymbol{v}$ is kept constant at every time step t. Thus, even when the t-th word is correctly pictured in the image $\boldsymbol{I}$, $\boldsymbol{W}_v \boldsymbol{v}$ itself cannot determine which specific object in the image should be generated according to the current context in the output caption. Therefore, we apply L2 normalization to the image representation in Eq. 8 to ensure that a relatively greater amount of the contextual information ($\boldsymbol{h}_t$) is used.

To train our model with the gating mechanism, we minimize the following cross-entropy loss for each pair of images and their pseudo-captions:

$$\mathcal{L}_g = -\frac{1}{T} \sum_{t=1}^{T} \log p(\hat{y}_t|\hat{y}_{<t}, \boldsymbol{I}). \quad (10)$$

2.3 Pseudo-Labels on Gate to Remove Word-Level Spurious Alignment

The above gate is expected to reflect the correspondence between images and words in pseudo-captions. However, learning to reflect the correspondence correctly is difficult for the gate under the noisy and weak supervision of pseudo-captions.

In this work, our focus is to remove the spurious alignment between images and words in pseudo-captions. Consequently, we apply the following rule to the gate that largely suppresses image representations to use: g_t should be close to one if the t-th word to generate is a detected object label; otherwise, it should be close to zero. This is based on the assumption that, given an image and its pseudo-caption, the reliable words in the pseudo-caption are only the detected object labels, and the others are likely to be irrelevant to the image.

We assign a pseudo-label $f \in \{0, 1\}$ on the gate: $f_t = 1$ if the word $\hat{y}_t$ corresponds to any of the object labels detected from a given image; otherwise, $f_t = 0$. The gate then learns the correspondence by minimizing the following loss function:

$$\mathcal{L}_f = -\frac{1}{T} \sum_{t=1}^{T} \Big[\alpha f_t \log g_t + (1 - f_t) \log(1 - g_t)\Big], \quad (11)$$

where α is the weight to emphasize the loss when $f_t = 1$. A relatively large value is recommended for α to prevent g_t from always being zero because the number of detected object labels (where $f_t = 1$)

	Training Text	Object Detector	Image Encoder	Text Decoder
A (Feng et al., 2019)	SS	Faster-RCNN trained on OpneImages-v2	Inception-v4	1-layer LSTM of 512 dimensions
B (Laina et al., 2019)	GCC	Faster-RCNN trained on OpneImages-v4	ResNet-101	1-layer GRU of 200 dimensions

Table 1: Summary of the difference in the experimental settings.

in pseudo-captions is generally smaller than the number of the other words (where $f_t = 0$).

Combined with the loss function of Eq. 10, the final loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_g + \mathcal{L}_f. \tag{12}$$

2.4 Unique-Object Decoding

An evaluation of our model revealed that it tends to repeat words in object categories. Although repetition is common in encoder–decoder models, this repetition was generated owing to a different cause. As mentioned in Section 2.2, the image representation $\boldsymbol{v}$ cannot correctly predict the word $\hat{y}_t$ without the context representation $\boldsymbol{h}_t$; if the gate value g_t is exactly one, the model always outputs the most salient object label in the given image.

To avoid ignoring contextual information, we applied a simple decoding rule during the evaluation. Given that the model generates a word y_t at t-th time step, our decoding rule checks whether y_t is in predefined object categories, *i.e.*, object categories defined for object detectors. If y_t is found in the object categories, the rule forces the probability of generating y_t to be zero in the subsequent time steps.

3 Experiments

We ran the experiments under two different settings, Feng et al. (2019) and Laina et al. (2019), for a fair comparison with each. For brevity, we refer to the settings in Feng et al. (2019) and Laina et al. (2019) as setting A and B, respectively. The difference of the settings is clarified in Table 1.

3.1 Datasets

Evaluation Set. To evaluate our proposed method, we used the MS COCO dataset (Lin et al., 2014) with the validation/test split defined by Karpathy and Fei-Fei (2015). Each split has 5,000 images and five reference captions for each image.

Training Images. We used the images (without their captions) in the remaining training split of MS COCO (113,286 images), and a pre-traind object detector (Huang et al., 2017) to retrieve the object labels found in the images[4]. The detector is a publicly available Faster-RCNN model[5] (Ren et al., 2015). The training data of the object detector differs depending on the previous work; thus, we used the object detector trained on OpenImages-v2 (Krasin et al., 2017) to compare with Feng et al. (2019) and that trained on OpenImages-v4 (Kuznetsova et al., 2020) to compare with Laina et al. (2019). Note that these object detectors were not trained on MS COCO images. Following the previous work, we refrained from using the detected bounding boxes and their features.

Training Text. Following the previous work, we used the Shutterstock image description corpus (SS) (Feng et al., 2019) and the training split captions (without images) of Google's Conceptual Captions (GCC) (Sharma et al., 2018) for comparison with Feng et al. (2019) and Laina et al. (2019), respectively. SS consists of 2.3M image descriptions crawled from Shutterstock, an online stock photography website; GCC consists of 3.3M image descriptions crawled from the web. Note that these sentences are not the descriptions of the images in MS COCO.

3.2 Implementation Details

Image Encoder. For a fair comparison with the previous work, we employed different image encoders depending on the compared method: Inception-v4 (Szegedy et al., 2017) in the settings of Feng et al. (2019) and ResNet-101 (He et al., 2016a,b) in the settings of Laina et al. (2019). Both image encoders were pre-trained on ImageNet (Russakovsky et al., 2015) and are publicly available[6]. The parameters of the image encoder were fixed during training and prediction.

Text Decoder. Similar to the image encoder, we

[4] Although this pre-trained object detector requires bounding box and semantic label annotations, it can be replaced with any multi-label image classifier, which can be trained on image-tag pairs that are largely and freely available on the web. To ensure this compatibility, bounding box features are not used in unsupervised image captioning.

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

[6] https://github.com/tensorflow/models/tree/master/research/slim

		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
A	Feng et al. (2019)	41.0	22.5	11.2	5.6	12.4	28.7	28.6	8.1
	Ours	**49.5 ± 0.7**	**27.3 ± 1.2**	**13.1 ± 0.8**	**6.3 ± 0.5**	**14.0 ± 0.1**	**34.5 ± 0.3**	**31.9 ± 1.0**	**8.6 ± 0.2**
B	Laina et al. (2019)				6.5	12.9	35.1	22.7	
	Ours	50.4 ± 1.5	29.5 ± 0.8	14.4 ± 0.5	**7.6 ± 0.4**	**13.5 ± 0.3**	**37.3 ± 0.2**	**31.8 ± 0.7**	8.4 ± 0.1

Table 2: Comparison with the state-of-the-art results on the experimental settings A and B. The scores of our model are the *mean* ± *standard deviation* of five runs. The scores obtained for BLEU-1 to 3 and SPICE are not provided in the original paper of Laina et al. (2019).

used a different RNN as our decoder: LSTM (Hochreiter and Schmidhuber, 1997) and GRU (Cho et al., 2014) to enable us to compare our results with those of Feng et al. (2019) and Laina et al. (2019), respectively. Following the previous work, the number of hidden layers' dimensions was set to 512 for LSTM and 200 for GRU. The number of the RNN layer was set to one. Word embeddings were randomly initialized and had the same dimensions as the RNN hidden layer.

Pseudo-Captions. Captions tend to describe salient objects, not all detected objects. For example, the frequent object *person* often co-occurs with *face* and *clothing* in images, but these three are not always the salient objects to be described in a caption. To avoid collecting the pseudo-captions that only contain these frequent objects, we picked up each detected object and their pairs to retrieve pseudo-captions, rather than using all detected objects. In this retrieval, we converted object labels to their plural forms using a dictionary used in Feng et al. (2019) so that the pseudo-captions could also cover the plural forms of the objects.

Pseudo-Caption Preprocessing. For each pair of objects, we selected sentences where fewer than four words existed between the objects. This is to pick up the sentences likely to describe the relations of the target objects. We then removed the sentences wherein the target objects were adjacent to avoid collecting the objects' compound words. For each object, we selected sentences wherein fewer than two words were in between the object and its dependent adjective to pick up the sentences likely to describe the object in detail. We used spaCy[7] en_core_web_lg model for parsing.

Value of α. As described above, each pseudo-caption contains only one or two detected objects, which is very few compared with the average sentence lengths of the text corpora (12.0 in SS and 10.7 in GCC). To balance the label imbalance of f_t, we searched the value for α (Eq. 11) at a power of 2 and found that $\alpha = 16$, which roughly equals the quotient of $\frac{\text{Sentence Length}}{\text{Detected Objects}}$, worked well across the settings.

Training Iteration. After collecting the pseudo-captions, we created a set of the objects and pairs that were used to collect the pseudo-captions. The training is iterated over the pairs in this set, rather than over each image, to avoid overfitting for the most frequent object labels. On each iteration of the pairs of objects, we randomly sampled the image and pseudo-caption, wherein both of the objects were contained. Likewise, we did the same sampling on each object in the pairs. The number of the object pairs was 11,607 and 10,612 in the settings A and B, respectively. We set the batch size to eight and terminated the training when the best validation score (specifically, the CIDEr score) did not exceed for 20 epochs. For the optimizer, we used Adam with the recommended hyperparameters (Kingma and Ba, 2015).

Evaluation. In the evaluation, we set the maximum decoding length to 20. Our model decoded captions by using greedy search and unique-object decoding, described in Section 2.4. The evaluation metrics we used were BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Denkowski and Lavie, 2014), CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016).

3.3 Comparison with the State-of-the-Art Results

Table 2 lists the results of our model compared with the previous state-of-the-art results. To avoid evaluating cherry-picked scores, we computed the mean and standard deviation of five results obtained with different seeds[8]. Our method outperforms the previous approaches in terms of all evaluation metrics. These results confirm the effectiveness of our simple method.

[7] `https://spacy.io`

[8] In all the experiments, we specified a seed of 0, 1, 2, 3, 4 for each run.

		gate	*pseudoL*	*unique*	*image*	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
A	Ours (full)	✓	✓	✓	✓	**49.5**	**27.3**	**13.1**	6.3	14.0	34.5	**31.9**	**8.6**
	w/o *pseudoL*	✓		✓	✓	0.0	0.0	0.0	0.0	0.0	0.5	0.9	0.3
	w/o *gate*			✓	✓	40.9	21.5	10.1	4.8	12.7	32.1	17.6	6.0
	w/o *unique*	✓	✓		✓	47.2	26.2	13.0	**6.4**	**14.1**	**34.9**	28.3	8.5
	w/o *image*	✓	✓	✓		43.3	23.3	10.8	5.1	13.1	31.7	25.5	7.8
B	Ours (full)	✓	✓	✓	✓	**50.4**	**29.5**	**14.4**	**7.6**	**13.5**	**37.3**	**31.8**	**8.4**
	w/o *pseudoL*	✓		✓	✓	44.5	25.4	12.2	6.2	12.4	36.7	29.2	7.5
	w/o *gate*			✓	✓	44.5	24.2	12.0	6.2	11.6	34.2	19.4	5.8
	w/o *unique*	✓	✓		✓	47.9	27.1	13.0	6.4	12.6	36.3	26.9	7.4
	w/o *image*	✓	✓	✓		47.1	26.0	12.8	6.6	13.1	34.7	29.7	8.0

Table 3: Ablation studies on the experimental settings A and B. The scores of Ours (full) are the mean of five runs; those of the other ablated models are the results of a single run.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
Feng et al. (2019)	41.0	22.5	11.2	5.6	12.4	28.7	28.6	8.1
Ours	49.5 ± 0.7	**27.3 ± 1.2**	**13.1 ± 0.8**	6.3 ± 0.5	**14.0 ± 0.1**	34.5 ± 0.3	31.9 ± 1.0	8.6 ± 0.2
Ours + Feng et al. (2019)	**50.9 ± 0.1**	**28.0 ± 0.1**	**14.0 ± 0.1**	**7.1 ± 0.0**	**14.1 ± 0.0**	**35.2 ± 0.1**	**35.7 ± 0.1**	**9.2 ± 0.0**

Table 4: Results of combining our method with previous methods (Feng et al., 2019). Scores of our model and the combined model are the *mean* ± *standard deviation* of five runs. We marked in bold the scores within the standard deviation of the best scores.

3.4 Ablation Study

Table 3 lists the results of our model obtained in the ablation studies. We tested the ablation of the gating mechanism (*gate*), pseudo-labels on the gating mechanism (*pseudoL*), unique-object decoding (*unique*), and image features (*image*). The pseudo-labels cannot be implemented without the base gating mechanism. Thus, the model "w/o *gate* w/ *pseudoL*" is not applicable. The model w/o *image* is the same as Ours (full) except that it only uses the word embeddings of detected object labels, rather than image features. It encodes detected object labels into word embeddings and then takes their mean[9] and replaces the image feature $\boldsymbol{v}$ with it. All models here were trained in the same manner as described in Section 3.2.

The results show that the pseudo-labels on the gating mechanism contribute a lot to the performance; the score degrades significantly from Ours (full) to w/o *pseudoL* in all metrics. On the other hand, the base gating mechanism does not function well by itself; not all scores of w/o *gate* are lower than those of w/o *pseudoL*. These results demonstrate that explicitly removing the word-level spurious alignment contributes the most to the relatively high performance of our model. Although it is a relatively low contribution compared with the pseudo-labels, unique-object decoding also enhances performance.

The degraded performance of w/o *image* suggests that object labels themselves are insufficient to describe images correctly. We observed that this model was vulnerable to errors propagated through object detectors. See Section 3.8 for the examples.

3.5 Combining with Previous Methods

Our method focuses on removing word-level spurious alignment between images and pseudo-captions, whereas the previous methods focus on aligning images and pseudo-captions at the sentence level. To utilize the strength of each, we combined our method with the previous methods as a model initialization method.

We first trained our model on the setting A and generated captions for the images in training data. We then paired the captions with the images as their pseudo-captions[10]. With the pairs, the caption generator of Feng et al. (2019) was initialized by learning to generate the pseudo-captions from the images. After the initialization, we trained the previous model using their publicly available code[11]. We used the same hyperparameters as Feng et al. (2019) except for the learning rate of 10^{-5} for the

[9] The number of detected objects was 3.0 in setting A and 4.0 in setting B on average. Thus, taking the mean does not break the detected information significantly.

[10] To avoid assigning obviously incorrect pseudo-captions, we omitted the captions that contained fewer than one detected object for the images with more than two detected objects. For the images with fewer than one detected object, we omitted the captions that contained no detected objects.

[11] `https://github.com/fengyang0317/unsupervised_captioning`

		Precision	Recall	F1
Detected	Feng et al. (2019)	**56.6**	57.4	**55.4**
	Ours	51.0	56.7	51.6
	Ours + Feng et al. (2019)	54.0	**61.8**	**55.4**
Others	Feng et al. (2019)	22.3	17.0	18.8
	Ours	27.8	**21.9**	23.4
	Ours + Feng et al. (2019)	**29.9**	**21.9**	**24.2**

Table 5: Bag-of-words matching scores with respect to detected object labels and the other words.

generator and 10^{-8} for the discriminator.

Table 4 shows the results. The combined model further improves the performance of our model and Feng et al. (2019). In particular, the improvement from Feng et al. (2019) is much larger than that from our model. These results suggest that removing the word-level spurious alignment is critical for the subsequent sentence-level alignment.

3.6 Negative Effect of Spurious Alignment

To further investigate the effect of removing the spurious alignment, we evaluated our model on *noisier words*: words other than the detected object labels. Our method discourages from aligning them with images because they are likely to be irrelevant to given images, while previous methods force the alignment. We tested the following bag-of-words matching on the MS COCO test set.

Let S be the bag of words of a caption generated from an image $\boldsymbol{I}$ and T^m be the m-th reference caption of $\boldsymbol{I}$. Given a set of detected object labels $\mathcal{O}$ of $\boldsymbol{I}$, we took the intersections $S_{det} = S \cap \mathcal{O}$ and $S_{other} = S \cap \overline{\mathcal{O}}$ for S, as well as for T^m. We define the precision (P), recall (R) and F1 score (F) of S against T^m as follows: $P = \frac{|S \cap T^m|}{|S|}$, $R = \frac{|S \cap T^m|}{|T^m|}$, $F = 2 \cdot \frac{P \cdot R}{P+R}$. Based on this, we define the precision, recall, and F1 score of S_{det} against T^m_{det} by replacing S with S_{det} and T^m with T^m_{det}, and likewise for those of S_{other} against T^m_{other}. We calculated the above scores for each pair of a generated captions and their reference captions and subsequently averaged it across the pairs. The pairs with empty T^m_* were excluded from the calculation.

Table 5 shows the results. Overall, the scores on detected object labels (Detected) are about two times higher than those on the other words (Others), indicating the difficulty of learning the alignment of the latter, noisier words. Our model performs better in predicting the noisier words, outperforming Feng et al. (2019) in all metrics. These results indicate that refraining from the alignment works better than forcing it for the noisier words.

		Word Type	Frequency
Object	Feng et al. (2019)	205	20013
	Ours	306	15052
	Ours + Feng et al. (2019)	239	18226
Others	Feng et al. (2019)	827	24865
	Ours	169	83693
	Ours + Feng et al. (2019)	121	110358

Table 6: Analysis of generated captions with respect to object labels and the other words. Word Type is the number of unique words, and Frequency is the mean of the frequency of the words in the training text corpus.

On the other hand, our model performs worse in predicting detected object labels. This is because our method trusts all detected object labels and aligns them with images without any constraints used in previous work. Combined with the previous method (Ours + Feng et al. (2019)), our model improves the prediction on detected object labels.

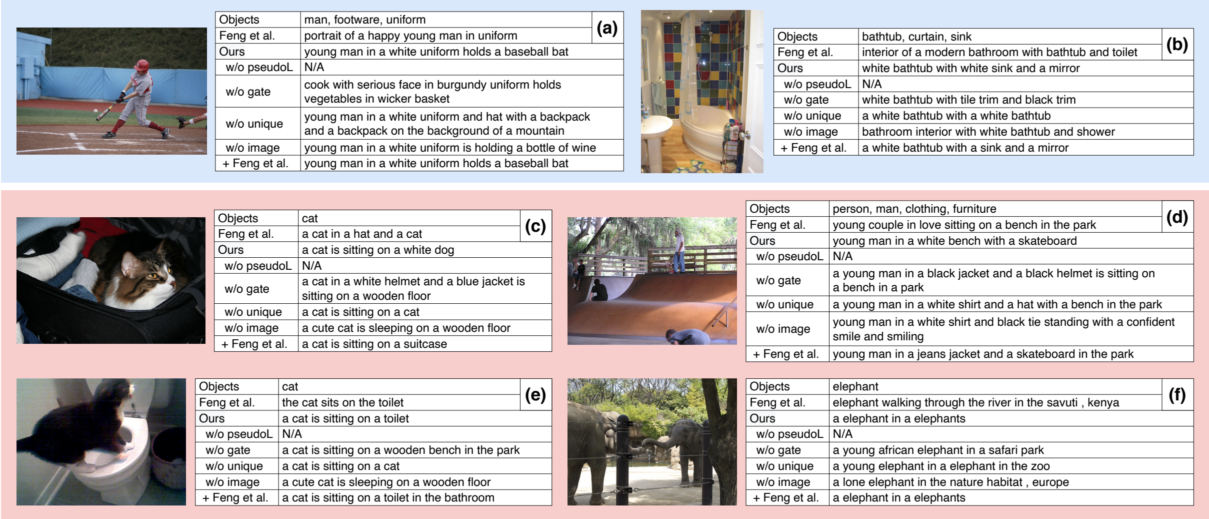
3.7 Positive Effect of Frequency

By assigning the pseudo-label f, our method encourages to align detected object labels with the image representation $\boldsymbol{v}$ and the other words with the contextual representation $\boldsymbol{h}$. Thus, our model is likely to predict the latter words mostly based on the previous output sequences, as language models do. If this is the case, then the latter words predicted tend to be the frequent words in the training text corpus.

To verify this tendency, we analyzed the frequency of output words in the training text corpus for object labels and the other words[12]. Table 6 presents the results. In contrast to object labels, our outputs' vocabulary is about five times smaller than that of Feng et al. (2019), and the words tend to be highly frequent in the training text corpus.

The results also show that a model performs better if it has the smaller and more frequent vocabulary of the words other than object labels. This correlation is convincing considering the coverage of frequent words. For example, a general caption such as "a man *with* a bike" can correctly cover various scenes in which a man is riding/sitting on/leaning on/standing near/... a bike. This positive effect of frequency suggests that firstly aligning the frequent words and gradually extending them can

[12] As we analyzed each unique word across all output captions in the MS COCO test set, we roughly divided the words into object labels and the others, not into detected object labels and the others.



Objects	man, footware, uniform	(a)
Feng et al.	portrait of a happy young man in uniform	
Ours	young man in a white uniform holds a baseball bat	
w/o pseudoL	N/A	
w/o gate	cook with serious face in burgundy uniform holds vegetables in wicker basket	
w/o unique	young man in a white uniform and hat with a backpack and a backpack on the background of a mountain	
w/o image	young man in a white uniform is holding a bottle of wine	
+ Feng et al.	young man in a white uniform holds a baseball bat	

Objects	bathtub, curtain, sink	(b)
Feng et al.	interior of a modern bathroom with bathtub and toilet	
Ours	white bathtub with white sink and a mirror	
w/o pseudoL	N/A	
w/o gate	white bathtub with tile trim and black trim	
w/o unique	a white bathtub with a white bathtub	
w/o image	bathroom interior with white bathtub and shower	
+ Feng et al.	a white bathtub with a sink and a mirror	

Objects	cat	(c)
Feng et al.	a cat in a hat and a cat	
Ours	a cat is sitting on a white dog	
w/o pseudoL	N/A	
w/o gate	a cat in a white helmet and a blue jacket is sitting on a wooden floor	
w/o unique	a cat is sitting on a cat	
w/o image	a cute cat is sleeping on a wooden floor	
+ Feng et al.	a cat is sitting on a suitcase	

Objects	person, man, clothing, furniture	(d)
Feng et al.	young couple in love sitting on a bench in the park	
Ours	young man in a white bench with a skateboard	
w/o pseudoL	N/A	
w/o gate	a young man in a black jacket and a black helmet is sitting on a bench in a park	
w/o unique	a young man in a white shirt and a hat with a bench in the park	
w/o image	young man in a white shirt and black tie standing with a confident smile and smiling	
+ Feng et al.	young man in a jeans jacket and a skateboard in the park	

Objects	cat	(e)
Feng et al.	the cat sits on the toilet	
Ours	a cat is sitting on a toilet	
w/o pseudoL	N/A	
w/o gate	a cat is sitting on a wooden bench in the park	
w/o unique	a cat is sitting on a cat	
w/o image	a cute cat is sleeping on a wooden floor	
+ Feng et al.	a cat is sitting on a toilet in the bathroom	

Objects	elephant	(f)
Feng et al.	elephant walking through the river in the savuti , kenya	
Ours	a elephant in a elephants	
w/o pseudoL	N/A	
w/o gate	a young african elephant in a safari park	
w/o unique	a young elephant in a elephant in the zoo	
w/o image	a lone elephant in the nature habitat , europe	
+ Feng et al.	a elephant in a elephants	

Figure 2: Sample captions for six input images taken from the MS COCO validation set. Our model generated correct captions for the images in the top row and wrong captions for the rest. Best viewed by zooming in.

be a promising approach.

3.8 Qualitative Analysis on Outputs

Figure 2 shows the captions generated by our model, its ablated models, Feng et al. (2019), and the combined model trained on the setting A. Our model generated correct captions for images (a) and (b). It successfully generated object labels that were not even detected by the object detector: *bat* in (a) and *mirror* in (b). On the other hand, errors of the object detector directly propagated to the output captions of w/o *image* model: the model generated an incorrect object *a bottle of wine*, owing to the missing object *bat* in (a).

Captions of the other images are negative results of our model. We observed that our model tended to repeat similar objects: *cat* and *dog* in (c), and *elephant* and *elephants* in (f). Without unique-object decoding, this tendency got worse: w/o *unique* model repeated *cat* in (c) and (e), and *elephant* in (f). Ours + Feng et al. (2019) model did not change much of the prediction of our model, as we set the learning rate low (see Section 3.5). However, it allowed the partial correction seen in (c): the combined model modified *dog* to *suitcase*.

In our outputs, words other than object labels tended to be frequent words and composed short phrases. On the contrary, Feng et al. (2019) tended to generate less frequent words (*savuti* and *kenya* in (f)) and longer phrases (*portrait of a happy young* in (a) and *young couple in love* in (d)), which were incorrect predictions in these examples.

Figure 3 shows output captions of our model and the gate values for each word. Overall, the gate values were high for object labels and low for the other words. Although our model was correct on the words other than object labels in these examples, these words were generated mostly by contextual features, thus heavily relied on contextual frequency. This heavy reliance on contexts resulted in generating the same word after an object label without considering images: *is sitting on* followed *cat* in both (c) and (e).

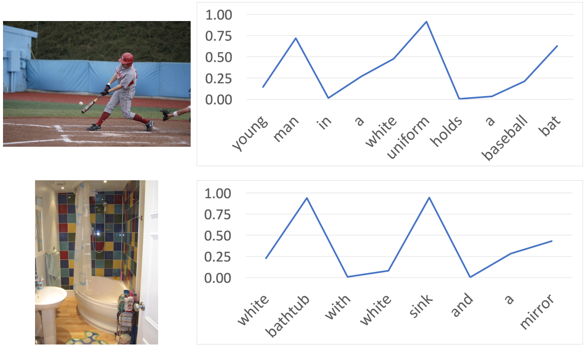


Figure 3: Sample captions with gate values. The plot represents the values of g_t for each predicted word. The value of g_t becomes high when the word is predicted using mainly image representation.

4 Related Work

There has been considerable research with different settings and approaches to describe scenes that have no image–sentence pairs. Novel object captioning (Hendricks et al., 2016; Venugopalan et al., 2017; Anderson et al., 2018a; Agrawal et al., 2019) attempted describing unseen objects in captions.

They incorporated an image classifier or object detector trained on objects not included in image–sentence pairs. Lu et al. (2018) tested captioning models on the generation of unseen combinations of objects, and Nikolaus et al. (2019) extended this to the unseen combinations of objects, attributes, and relations. In both settings, only the combinations were unseen, but each word in the combinations appeared in the training data. Semi-supervised approaches utilized caption retrieval models to automatically collect the corresponding captions for unannotated images to augment image–sentence pairs (Liu et al., 2018; Kim et al., 2019).

The above work was evaluated on the scenes where correct descriptions partially overlapped with those in the training image–sentence pairs. However, there can be scenes with no such overlap due to the limited coverage of the currently available image–sentence pairs. Taking a step further, unsupervised image captioning (Feng et al., 2019; Laina et al., 2019) aims to describe scenes that have no overlap with the image–sentence pairs, without the annotation of the pairs. To test in that situation, the task does not allow to use any image–sentence pairs. The only available resources are images and sentences drawn from different sources and object labels detected from the images.

Feng et al. (2019) first trained an encoder–decoder model that takes object labels in a sentence as its input and outputs the original sentence. After training, this model took the object labels detected from each image and outputted a sentence to pair with the image as its pseudo-caption. These pairs were then used to initialize a caption generator for the subsequent image–sentence alignment: bi-directional (image-to-sentence and sentence-to-image) feature reconstruction and GAN training (Goodfellow et al., 2014) to ensure fluency in generated captions. In the work of Laina et al. (2019), pseudo-captions were sentences that contained object labels detected from a given image. They employed metric learning and GAN training to minimize the difference between images and pseudo-captions in their latent space, as well as to maximize the difference between images and sentences wherein no detected object label was included.

Our approach is different from them in that it focuses on removing the mismatched words of pseudo-captions to take reliable supervision only, rather than forcing the use of the entire pseudo-captions for image–sentence alignment. Although the previous work additionally ensured to align detected object labels to images, they did not prevent the spurious alignment between images and words.

As an eased setting of unsupervised image captioning, unpaired image captioning has also been explored (Feng et al., 2019; Laina et al., 2019; Gu et al., 2019; Liu et al., 2019). The major difference from unsupervised image captioning is that images and sentences are drawn from image–sentence pairs, rather than from different sources. That is, every image has completely matched captions in pseudo-captions, which is not the case in unsupervised image captioning. As correct captions exist for each image, previous approaches focused on matching images and sentences at the sentence level. Contrary to these approaches, we focus on employing unsupervised image captioning and devising a method to remove word-level spurious alignment in the much noisier pseudo-captions.

Another variation of unpaired image captioning is the generation of captions in one language that has no image–sentence pairs, using paired images and captions in another language (Gu et al., 2018; Song et al., 2019). However, this line of research is beyond the scope of our work, as it requires image–sentence pairs to be at least in one language.

Our gating mechanism borrowed the idea of adaptive attention (Lu et al., 2017, 2018). Adaptive attention serves to control when generating words from image representations. Although these methods assume that the control is automatically learned from image–sentence pairs, this is not the case in an unsupervised setting. Our method is different from theirs in that we add heuristic pseudo-labels to train the gate when using image representations.

5 Conclusion

We investigated the importance of removing word-level spurious alignment between images and pseudo-captions in the task of unsupervised image captioning. For this purpose, we introduced a simple gating mechanism trained to align image features with only the most reliable words in pseudo-captions. The experimental results showed that our proposed method outperformed the previous methods without the sentence-level learning objectives used in the previous methods. Moreover, our method improved the performance further by combining with the previous methods. These results confirmed the importance of careful alignment in word-level details.

References

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *ICCV*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*.

Peter Anderson, Stephen Gould, and Mark Johnson. 2018a. Partially-supervised image captioning. In *NeurIPS*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018b. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *EMNLP*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *the ninth workshop on statistical machine translation*.

Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Unsupervised image captioning. In *CVPR*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NeurIPS*.

Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. 2018. Unpaired image captioning by language pivoting. In *ECCV*.

Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. 2019. Unpaired image captioning via scene graph alignments. In *ICCV*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016a. Deep residual learning for image recognition. In *CVPR*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016b. Identity mappings in deep residual networks. In *ECCV*.

Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. 2017. Speed/accuracy trade-offs for modern convolutional object detectors. In *CVPR*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. 2019. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. In *EMNLP-IJCNLP*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. 2017. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*.

Iro Laina, Christian Rupprecht, and Nassir Navab. 2019. Towards unsupervised image captioning with shared multimodal embeddings. In *ICCV*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

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 *ECCV*.

Fenglin Liu, Meng Gao, Tianhao Zhang, and Yuexian Zou. 2019. Exploring semantic relationships for image captioning without parallel data. In *ICDM*.

Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. 2018. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *ECCV*.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *CVPR*.

Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. 2019. Compositional generalization in image captioning. In *CoNLL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Yuqing Song, Shizhe Chen, Yida Zhao, and Qin Jin. 2019. Unpaired cross-lingual image caption generation with self-supervised rewards. In *ACMMM*.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2017. Captioning images with diverse objects. In *CVPR*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

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*.