Generating Diverse and Descriptive Image Captions Using Visual Paraphrases

Lixin Liu[1,2,3] Jiajun Tang[1] Xiaojun Wan[1,2,3] Zongming Guo[1,2]
[1]Institute of Computer Science and Technology, Peking University
[2]Center for Data Science, Peking University
[3]The MOE Key Laboratory of Computational Linguistics, Peking University
{liulixin, jiajun.tang, wanxiaojun, guozongming}@pku.edu.cn

Abstract

Recently there has been significant progress in image captioning with the help of deep learning. However, captions generated by current state-of-the-art models are still far from satisfactory, despite high scores in terms of conventional metrics such as BLEU and CIDEr. Human-written captions are diverse, informative and precise, but machine-generated captions seem to be simple, vague and dull. In this paper, aimed at improving diversity and descriptiveness characteristics of generated image captions, we propose a model utilizing visual paraphrases (different sentences describing the same image) in captioning datasets. We explore different strategies to select useful visual paraphrase pairs for training by designing a variety of scoring functions. Our model consists of two decoding stages, where a preliminary caption is generated in the first stage and then paraphrased into a more diverse and descriptive caption in the second stage. Extensive experiments are conducted on the benchmark MS COCO dataset, with automatic evaluation and human evaluation results verifying the effectiveness of our model.

1. Introduction

Image captioning is a task aiming to describe images with natural languages. There have been remarkable developments in rent years with the emergence of deep learning methods [18, 38, 47]. However, captions generated by current methods still require improvement.

Figure 1 shows a case where the imperfection of machine-generated captions can be easily identified. We insist that a good caption which resembles a human-written caption should have several properties. (1) **Fluency**: The caption should be a fluent sentence. (2) **Relevance**: The caption should correctly describe the visual content and be closely relevant to the image. (3) **Diversity**: Language is a rich, colorful and varied system. Good captions contain diverse wordings and rich expressions. (4) **Descriptiveness**:

Machine-generated Caption:
a man standing next to a white car
Human-written Captions:
Caption 1: a white compact car parked on a sandy dirt road
Caption 2: a car being driven onto two white flat things
Caption 3: a man getting in a sport utility vehicle with surf boards on the roof
Caption 4: man standing in open door of car on a desert road
Caption 5: a man gets back into his car in the desert

Figure 1. A machine-generated caption by a state-of-the-art attention-based image captioning model [34] and five human-written captions from MS COCO dataset [24].

A good caption describes an image by referring to the important, specific, and detailed aspects of the image, which is precise, informative and descriptive [28]. As is shown in Figure 1, the machine-generated caption is a fluent and correct description of the image. However, it is very simple and vague. Computers prefer "safe" output sentences with very high-frequency expressions [22], and they tend to describe only the obvious facts, ignoring key details. On the contrary, humans prefer writing captions with more diversity by using more varied wordings (like *sandy dirt road* and *standing in open door*, etc.) and with more descriptiveness by describing more important details (like *in the desert* and *with surfboards*, etc.).

Paraphrases are sentences or phrases that convey approximately the same meaning in different expressions [4]. In the task of image captioning, different people may describe the same image from different perspectives. Even they focus on the same scene in an image, their expressions can

hardly be identical. For example, five human-written captions in Figure 1 differ from each other significantly. Different sentences *describing the same image* can be considered as a set of paraphrases, which is called **visual paraphrases**.

In this paper, we would like to generate diverse and descriptive image captions by taking advantages of visual paraphrases from captioning datasets. An image is usually annotated with a set of visual paraphrases consists of d different captions. Typical methods simply ignore the relationship between these paraphrases and regard them as d independent samples. We explore the relationship between them and select several visual paraphrase pairs (C_i, C_j) with a specific scoring function (see Section 3.1 and 3.3) for training. Concerning that writing a diverse, descriptive caption directly is challenging, we propose a captioning model with two-stage decoding which first generates a preliminary caption (less diverse and descriptive) given the visual input, and then paraphrases it into a more diverse and descriptive caption using these visual paraphrase pairs. Our model not only learns from visual-semantic information but also utilizes textual relationships from different wordings of visual paraphrases.

Our major contributions are summarized as follows:

- We explore the role of visual paraphrases for image caption generation. And we investigate different scoring functions for selecting useful visual paraphrase pairs from captioning training data.
- We propose a captioning model which fuses visual and textual information with two-step decoding by firstly generating a preliminary caption and then paraphrasing it into a more diverse and descriptive caption.
- Results in terms of a variety of automatic metrics and human evaluation demonstrate that our model can generate more diverse and descriptive captions while maintaining fluency and relevance.

2. Related Work

Image Captioning Text generation from images [38, 17, 26, 20] is a problem at the intersection of computer vision and natural language processing. Image captioning, aimed at generating natural language descriptions for images, usually consists of a CNN as an image encoder and an RNN as a decoder to generate sentences [38, 10, 18, 48]. Attention mechanism [47, 27, 30, 2], explicit attributes detection [12, 52, 44, 50], reinforcement learning (RL) methods [32, 34], and visual relations detection [49] are proposed for improvement.

Diverse and Discriminative Captioning Some work pays attention to the diversity or distinctness of image captions, with goals similar to ours. Dai et al. [7] adopt conditional generative adversarial networks (GAN) to produce diverse and natural captions. Some other work addresses distinctiveness or discriminability, which is closely related to the descriptiveness we refer to, by emphasizing the distinctive aspects of an image that distinguishes it from other images. Introspective speaker (IS) model [35], as a modification of beam search, generates discriminative image captions using a distractor image. Dai et al. [8] adopts a contrastive loss to push the probabilities of captions to be higher for matched images and lower for mismatched images than the reference model. Luo et al. [28] add an extra discriminability reward to a CIDEr reward for policy gradient for generating discriminative captions. Some prior work [7, 9, 39, 42, 37, 40] focus on improving the diversity of captions. However, when it comes to diversity, they refer to generating *multiple* mutually diverse captions for each image with methods like beam search, while we refer to producing a *single* caption with diverse and rich expressions rather than simple and common wordings. The interpretation of diversity in some other works for text generation [22, 46, 53] are similar to ours.

Paraphrases Paraphrases are alternative ways of expressing the same meaning using different wordings [4]. Our work is inspired by some work addressing paraphrases associated with other modalities and paraphrase generation task. Chu et al. [6] propose a clustering method to extract phrasal expressions describing the same visual concept (called visually grounded paraphrases) from image captions. Chen et al. [5] build an image captioning dataset with visually-situated paraphrase pairs by crowd-sourcing and retrieval-based methods. Lin et al. [25] address the task called visual paraphrasing as verifying if the two textual descriptions describe the same image by visual imagination. As for paraphrase generation, the mainstream approach is attention-based sequence-to-sequence model [3, 31]. Some improvements such as the use of reinforcement learning [23] and variational autoencoders [14] are proposed. Caption pairs in COCO dataset are utilized to constitute a paraphrase corpora in their experiments. However, they randomly choose caption pairs without addressing different characteristics of captions and utilizing visual information.

Two-stage Text Generation A problem of the current encoder-decoder framework for text generation is when generating words, only the previously generated words can be utilized, ignoring future words [45]. So methods with two-stage decoding are proposed. In deliberation network [45] for machine translation, two decoders are utilized, with the first decoder generating a sequence and the second decoder for refining. Stack-Cap [13] consists of one coarse decoder and a sequence of fine decoders for image captioning. Their intermediate outputs from the first decoder are randomly sampled but not well-defined during training. Without clear targets for training the first RNN decoder, they are prone to accumulate errors thus very hard to train. On the contrary, our model uses two different sentences from the

training set to train the two-step decoders, which is easy to optimize. Preview network [54] uses a pipeline with two stages of decoding using two visual encoders and two language decoders. Sentences in their two-stage decoding are identical during training, which is different from ours. Skeleton Key [41] first generates skeleton sentences and attributes, and then rewrites them to full sentences. POS [9] is a VAE-based network using part-of-speech as a language prior. These methods define an intermediate sequence for caption generation. In this paper, sentences generated in two steps are both complete and correct captions with different properties.

3. Our Method

In this section, we discuss our method in details. Our model relies on selecting visual paraphrase pairs (Section 3.1) from image captioning datasets using a variety of scoring functions (Section 3.3). Then these visual paraphrase pairs are utilized to train our captioning model with two stages of decoding (Section 3.2), as is shown in Figure 2.

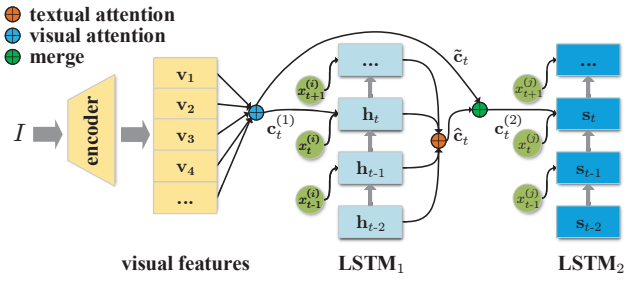

Figure 2. Framework of our model.

3.1. Visual Paraphrase Pairs Selection

An image captioning dataset $\mathcal{D}$ is composed of N images and M captions. Each image I has d annotated captions $\mathcal{C} = \{C_1, ..., C_d\}$. These d captions constitute a set of **visual paraphrases** corresponding to image I. Each pair of captions is a **visual paraphrase pair** corresponding to I. We select a portion of visual paraphrase pairs $\mathcal{P}$:

$$\mathcal{P} = \{(C_i, C_j) | S(C_i, C_j) > \epsilon, \\ \forall C_i, C_j \in \mathcal{C}, C_i \neq C_j\} \quad (1)$$

where $S(C_i, C_j)$ is a scoring function measuring the difference within visual paraphrase pairs on a specific characteristic (e.g. diversity), and ϵ is a threshold.

In this way, d captions for an image are reorganized into a series of selected visual paraphrase pairs $(C_i, C_j) \in \mathcal{P}$. In our experiments, C_j is more "complex" than C_i from the point of a scoring function, so the number of different selected paraphrase pairs is at most $\binom{d}{2}$. These selected visual paraphrase pairs are utilized for training. The scoring functions used for selection are elaborated in Section 3.3.

3.2. Caption Generation

During training, our image captioning model learns to first produce a "simpler" (less diverse and descriptive) caption C_i and then rewrites the caption considering the image content I to get a more "complex" (more diverse and descriptive) caption C_j that may better describe the image. As is shown in Figure 2, our model adopts a two-stage decoding process and it is composed of two parts: a standard attention-based captioning module for generating a preliminary caption and a visual paraphrase generation module with multimodal fusion for paraphrasing the preliminary caption into a final caption.

The image I is first encoded by an image encoder to get a set of spatial visual features $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_{|\mathcal{V}|}\}$ where each feature represents a sub-region of the image. Then we use a first LSTM decoder $\mathbf{LSTM}_1$ to generate the preliminary caption $C_i = \{x_t^{(i)}\}_{t=1}^{|C_i|}$ by paying attention to the visual input. We adopt **Att2in** [34] model in our experiments. It is an effective modification of the vanilla LSTM [16] with attention mechanism [3, 47].

$$\mathbf{h}_t = \mathbf{LSTM}_1\left(x_t^{(i)}, \mathbf{h}_{t-1}, \mathbf{c}_t^{(1)}\right) \quad (2)$$

$$\mathbf{c}_t^{(1)} = \sum_{n=1}^{|\mathcal{V}|} \alpha_t^n \mathbf{v}_n \quad (3)$$

$$\alpha_t^n = \frac{\exp(a(\mathbf{h}_{t-1}, \mathbf{v}_n))}{\sum_{m=1}^{|\mathcal{V}|} \exp(a(\mathbf{h}_{t-1}, \mathbf{v}_m))} \quad (4)$$

$$a(\mathbf{h}_{t-1}, \mathbf{v}_n) = \mathbf{u}_a^\intercal \tanh(\mathbf{W}_h \mathbf{h}_{t-1} + \mathbf{W}_v \mathbf{v}_n) \quad (5)$$

where $\mathbf{c}_t^{(1)}$ is the context vector of the input of $\mathbf{LSTM}_1$ representing the weighted sum of spatial image features $\mathbf{v}_n \in \mathcal{V}$, with hidden state $\mathbf{h}_t$ as query. $\mathbf{u}_a$, $\mathbf{W}_h$ and $\mathbf{W}_v$ are model parameters. For brevity, we use $\mathbf{c}_t^{(1)} = \mathbf{Attn}(\mathbf{h}_{t-1}, \mathcal{V})$ to denote the the context vector obtained from the attention on $\mathcal{V}$ using $\mathbf{h}_{t-1}$ as the query via Equations 3, 4 and 5.

After the preliminary caption C_i has been generated, the hidden states of $\mathbf{LSTM}_1$ and original image features $\mathcal{V}$ are fed to another decoder $\mathbf{LSTM}_2$. The hidden states of $\mathbf{LSTM}_1$ $\mathcal{H} = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{|C_i|}\}$ consist of the information of the preliminary sentence C_i. We aggregate the textual information of sentence C_i from these hidden states and visual information of image regions from $\mathcal{V}_k$ for rewriting to C_j. Different from textual paraphrase generation, visual paraphrase generation requires information from two modalities to rephrase the sentence. The textual information from the first sentence C_i and the attentive visual information are merged when generating each word.

We adopt attention mechanism to get contextual information from visual content for the rephrased caption C_j:

$$\tilde{\mathbf{c}}_t = \mathbf{U}_v \mathbf{Attn}\left(\mathbf{s}_{t-1}, \mathcal{V}\right) \quad (6)$$

where $\mathbf{s}_t$ is the hidden state of the second decoder $\mathbf{LSTM}_2$.

And the contextual information from textual content is calculated by the attention on $\mathcal{H}$ with $\mathbf{s}_t$ as a query:

$$\hat{\mathbf{c}}_t = \mathbf{U}_c \mathbf{Attn}\left(\mathbf{s}_{t-1}, \mathcal{H}\right) \tag{7}$$

Note that the two original context vectors $\mathbf{Attn}\left(\mathbf{s}_t, \mathcal{V}\right)$ and $\mathbf{Attn}\left(\mathbf{s}_t, \mathcal{H}\right)$ obtained from attention mechanisms are mapped by linear layers $\mathbf{U}_v$ and $\mathbf{U}_c$ respectively to get the transformed context vectors $\tilde{\mathbf{c}}_t$ and $\hat{\mathbf{c}}_t$ in a shared multimodal embedding space, because features in $\mathcal{H}$ and $\mathcal{V}$ represent different modalities thus are in different feature spaces.

To fuse the information of two modalities, we use a merging gate $\mathbf{g}_t$ for the fusion of multimodal features.

$$\mathbf{g}_t = \sigma\left(\mathbf{W}_{\tilde{c}}\tilde{\mathbf{c}}_t + \mathbf{W}_{\hat{c}}\hat{\mathbf{c}}_t + \mathbf{W}_s\mathbf{s}_{t-1}\right) \tag{8}$$

$$\mathbf{c}_t^{(2)} = \mathbf{g}_t \odot \tilde{\mathbf{c}}_t + (1 - \mathbf{g}_t) \odot \hat{\mathbf{c}}_t \tag{9}$$

where σ denotes sigmoid function, and $\odot$ indicates element-wise vector multiplication. $\mathbf{W}_{\tilde{c}}$, $\mathbf{W}_{\hat{c}}$ and $\mathbf{W}_s$ are model parameters. When generating the paraphrased sentence $C_j = \{x_t^{(j)}\}_{t=1}^{|C_j|}$, $\mathbf{LSTM}_2$ pays attention to different sub-regions of images by visual attention and different words of sentence C_i by textual attention for generating the next word. The information fusion of two modalities is controlled by the merging state for adaptively determining different weights for the two-sides. The probability distribution of the next generated word $\hat{x}_{t+1}^{(j)}$ is calculated through a softmax activation function on the output of $\mathbf{LSTM}_2$.

$$\mathbf{s}_t = \mathbf{LSTM}_2\left(x_t^{(j)}, \mathbf{s}_{t-1}, \mathbf{c}_t^{(2)}\right) \tag{10}$$

$$\hat{x}_{t+1}^{(j)} \sim \text{Softmax}\left(\mathbf{W}_o\mathbf{s}_t\right) \tag{11}$$

3.3. Scoring Functions

We employ a variety of scoring strategies to select a series of paraphrase pairs in $\mathcal{P}$ (Equation 1). Scoring functions define the extent of specific characteristics (e.g. diversity) of paraphrase pairs. Sentences in visual paraphrase pairs differ from each other.

3.3.1 Sentence Length

If a sentence is longer, it is more likely to provide informative descriptions or use diverse expressions. We define the scoring function $S_{\text{len}}(C_i, C_j) = \textit{length}(C_j) - \textit{length}(C_i)$. We set $\epsilon \geq 0$ in Equation 1, so caption C_j is longer than C_i in the selected visual paraphrase pair.

3.3.2 Syntactic Complexity

If a sentence contains more complex syntactic structure, it may contain more modifiers to give a detailed description. Diverse and rich wording may also increase the syntactic complexity of sentences. Yngve score [51] which measures the number of branches in a sentence's syntactic tree is adopted as the measure of syntactic complexity. The Yngve scoring function is defined as $S_{\text{Yngve}}(C_i, C_j) = \textit{Yngve}(C_j) - \textit{Yngve}(C_i)$. We set $\epsilon \geq 0$ in Equation 1, so C_j is generally more syntactically complex than C_i.

3.3.3 TF-IDF Diversity

We design to score a sentence C using the sum of TF-IDF scores of n-grams. TF-IDF reflects the importance of an n-gram. N-grams with lower frequency in the training corpus have higher IDFs. TF-IDF Diversity (**Tdiv**) metric and scoring function are defined as:

$$\text{Tdiv}(C) = \sum_{n=1}^{\mathcal{N}} \frac{\sum_{\text{n-gram}\in C} \text{TF-IDF(n-gram)}}{V_n} \tag{12}$$

$$S_{\text{Tdiv}}(C_i, C_j) = \text{Tdiv}(C_j) - \text{Tdiv}(C_i) \tag{13}$$

where $V_n = \frac{1}{M}\sum_{C\in\mathcal{D}}\sum_{\text{n-gram}\in C}\text{TF-IDF(n-gram)}$ is the normalization term for the n-th gram and M is the number of captions in $\mathcal{D}$. When calculating the document frequency of TF-IDF scores, each caption is considered as a document. We refer equation 12 to *Tdiv* metric in our experiments. We choose $\mathcal{N} = 3$ so unigrams, bigrams, and trigrams of captions are used for calculation. Tdiv metric denotes the extent of diverse expressions of a sentence. A sentence has a low Tdiv score when it contains mostly commonly-used words and phrases, and vice versa. In a visual paraphrase pair (C_i, C_j) chosen by Tdiv scoring function ($\epsilon \geq 0$), C_j generally has richer expressions than C_i.

3.3.4 Image Retrieval Rank

Previous works [8, 28] focus on the distinctness or discriminability aspect of caption by image retrieval, aiming to retrieve the original image given its corresponding caption. It is based on a visual-semantic retrieval system. We adopt a similar network architecture as Luo et al. [28]. An image I and its corresponding caption $C \in \mathcal{C}$ are encoded by a CNN and an LSTM respectively to get corresponding feature vectors. The feature vectors are mapped into the same embedding space to get the image embedding $f(I)$ and caption embedding $g(C)$. The similarity of I and C is computed by the cosine similarity of the embeddings:

$$\mathbf{sim}(I, C) = \frac{f(I)^\intercal g(C)}{\|f(I)\|\|g(C)\|} \tag{14}$$

A bi-directional ranking loss is defined as follows:

$$\begin{aligned} L_e = &\sum_I \sum_{C^-} \max(0,\ \beta - \mathbf{sim}(I, C) + \mathbf{sim}(I, C^-)) \\ &+ \sum_C \sum_{I^-} \max(0, \beta - \mathbf{sim}(I, C) + \mathbf{sim}(I^-, C)) \end{aligned} \tag{15}$$

where $\beta \in \mathbb{R}$ serves as a margin parameter. Every (I, C) is a ground truth image-caption pair in training data, C^- denotes an unpaired caption for image I, and vice-versa for I^-. So the retrieval system can retrieve relevant images given captions as queries, and vice versa.

Given the caption C as the query, we use the pretrained retrieval system on COCO training set to rank a large number of images in the training set including the original image I. Ideally, a correct, detailed caption with high distinctness can be used to retrieve the corresponding image as the best matching. So retrieval performances can be used to reflect how correct and descriptive a caption is. $\mathit{rank}(I|C)$ is the rank of the corresponding image I using C as query. A larger value of $\mathit{rank}(I|C)$ denotes that the description is either incorrect or not detailed enough so the retrieval system cannot find the correct image I given the query C. We define the image retrieval (**IR**) scoring function as

$$S_{\text{IR}}(C_i, C_j) = \frac{\mathit{rank}(I|C_i)}{\mathit{rank}(I|C_j)} \tag{16}$$

Captions in a caption pair (C_i, C_j) selected by S_{IR} and Equation 1 ($\epsilon \geq 1$) are both human-annotated captions so they are correct captions. But caption C_j may provide more detailed and informative descriptions of the image than C_i.

The image retrieval ranks reflect *relevance* and *descriptiveness* aspects of captions. Human-written captions are usually relevant and informative. Moreover, the retrieval system is trained on the same training set, so captions will have small retrieval ranks. But some human-written captions which do not describe the highly distinct parts may also be correct captions for describing another image in the training set. Therefore, they will have a larger rank compared to other captions when we use the huge numbers of images in the training set as retrieval candidates.

3.4. Training and Inference

Our model can be trained jointly by minimizing the negative log-likelihood of generating C_i and C_j:

$$\mathcal{L} = \sum_{I \in \mathcal{D}} \sum_{(C_i, C_i) \in \mathcal{P}} -\log p(C_i|I) - \log p(C_j|C_i, I) \tag{17}$$

As captions for the first and the second decoding stages are available during training, we use standard teacher forcing [43] strategy for training RNNs by feeding the words of ground-truth captions as the inputs.

During testing, we adopt the beam search strategy. When the candidate captions in the first stage are completely decoded by beam search, the candidate caption with the highest probability is selected as the preliminary caption. The hidden states corresponding to it are collected for the textual attention of the second-stage decoding. Then another beam search is applied to decode the polished final caption.

4. Experiments

4.1. Dataset

We conduct experiments on the Microsoft COCO dataset [24]. It has 123,287 images with five different human-annotated captions per image[1]. We adopt the standard "Karpathy split" [18], with 5,000 images for validation, 5,000 images for testing and the rest for training.

4.2. Experimental Details

For our image captioning model, we utilize bottom-up spatial features [2] extracted by Faster R-CNN [33] in conjunction with Resnet-101 [15] trained by object and attribute annotations from Visual Genome Dataset [21]. We set the word embedding size and LSTM hidden size to 512. The vocabulary size is 9,488. During training, Adam [19] with a learning rate of 5×10^{-4} is utilized for optimization. The batch size is set to be 16 and the beam size is 3. The maximum sentence length is 16. Retrieval ranks for Equation 16 are calculated using 20,000 candidate images in the COCO training set.

4.3. Evaluation Metrics

We consider a variety of evaluation metrics aiming for better evaluation of caption quality from different perspectives. We employ widely-adopted conventional metrics, including BLEU-4 [29], CIDEr [36], and SPICE [1]. These metrics compare generated sentences with references so they mainly focus on the *relevance* aspect of captions. BLEU and CIDEr are n-gram based metrics while SPICE measures how effectively captions recover objects, attributes and the relations between them, which is proved to have more correlation with human judgments [1].

However, these conventional metrics are not perfect especially insufficient to evaluate *diversity* and *descriptiveness*, which are also crucial aspects for high-quality image descriptions. Particularly, metrics like BLEU and CIDEr are primary to n-gram overlap so a sentence with very common n-grams but lacking diversity and descriptiveness, which is shown to have negative correlation with human judgments on the detailedness of captions [1].

Therefore, we also report some other statistics and metrics reflecting some important aspects of captions. These metrics are mostly ignored in prior works. In addition to average length (*length*), Yngve scores (*Yngve*) and *Tdiv* (Equation 12) defined in Section 3.3, we report *Dist-2*, *Dist-3*, and *Dist-S* [46] results. They are respectively the number of distinct bigrams, trigrams, and sentences in the generated captions. Higher *Dist* scores indicate more diverse expressions of captions so *Dist* metrics are measurements of sentence diversity. The descriptiveness of captions is vague

[1]There are very few (327) images with 6 or 7 annotated captions.

and difficult to evaluate. We follow prior work [28, 8] with a self-retrieval strategy. We try to retrieve the original image given the generated captions as queries. Retrieval performances are measured by $R@K$ (K=1, 5, 10), i.e., recall at K [2]. To prevent overfitting to the retrieval model during training with image retrieval scoring function, we use pretrained VSE++ [11] for self-retrieval evaluation, which is a strong visual-semantic retrieval model utilizing different network architectures and image features (finetuned ResNet-152 [15]). Better self-retrieval performance indicates that the model generates relevant, informative, and descriptive captions.

We also conduct human evaluations on Amazon Mechanical Turk with 20 volunteers. We compared methods on 100 images randomly sampled from test set. Each caption is rated by 4 different people. Volunteers rate captions from the 1-5 scale (higher is better) with respect to four criteria: fluency, relevance, diversity, and descriptiveness. Definitions are shown in the Supplementary Material.

4.4. Baseline Approaches

We adopt a variety of baseline methods, including:

Attention [34]: Attention based captioning model (Att2in) using bottom-up image features from Faster R-CNN (ResNet-101). Our model is based on it.

GAN [7]: Conditional generative adversarial network for diverse and natural image captioning.

IS [35]: Introspective Speaker method for discriminative image captioning.

CL [8]: Contrastive learning method focusing on the distinctness aspect of captions.

We use baseline methods of Attention, GAN and IS with the same image features and attention-based architecture as ours for fair comparisons. While CL [8] also utilizes a strong adaptive attention [27] as base model with a fine-tuned ResNet-152 [15] encoder.

We also report results using reinforcement learning with CIDEr as an optimization objective, including:

CIDEr-RL [34]: Self-critical sequence training with CIDEr rewards.

DiscCap [28]: Self-critical training with a mixed objective of CIDEr and discriminative objective rewards (model ATTN+CIDER+DISC-1 in their paper).

Stack-Cap [13]: A coarse-to-fine strategy using two-step decoders with CIDEr optimization.

4.5. Automatic Evaluation Results

Table 1 shows the automatic evaluation results. To further see the properties of human-written captions, we also report metrics' scores of captions in MS COCO test set in the table. BLEU/CIDEr/SPICE scores of human-written captions are calculated following previous work [3] [7]. Compared to **Attention** baseline, humans have higher metrics' scores except for BLEU and CIDEr. It further demonstrates that humans write more diverse and informative captions than machines. But there are not many overlapping n-grams among these human-written visual paraphrases.

We first compare our models with **Attention** baseline to evaluate different influences of the choice of scoring functions. **Ours (len)** notably improves diversity and descriptiveness of captions, with increases in *Tdiv*, *Dist* and retrieval performances ($R@K$). But it generates excessively long sentences which heavily damages overall quality, with huge drops in BLEU and CIDEr. **Ours (Yngve)** only slightly improves diversity and descriptiveness of captions. Compared to these two models with simple scoring functions, **Ours (IR)** and **Ours (Tdiv)** show much better performances. **Ours (IR)** significantly improves retrieval performances, with slight gains in conventional metrics SPICE, BLEU, and CIDEr. **Ours (Tdiv)** shows better results in diversity and retrieval performances than other scoring functions. With no surprise, scores of n-gram based metrics like BLEU and CIDEr drop while diversity increases. Furthermore, we observe an obvious improvement in SPICE metric. Higher SPICE scores demonstrate that **Ours (Tdiv)** correctly describes objects, attributes, and their relations. Comparisons suggest that sentence length and syntactic complexity (Yngve scores) may not be accurate indicators of diversity and descriptiveness of captions. We observe that retrieval performances of **Ours (Tdiv)** are even better than that of **Ours (IR)** which utilizes a retrieval model explicitly. It may be due to the difference of retrieval models utilized for training and evaluation [4].

Then we compare **Ours (Tdiv)** and **Ours (IR)** with other MLE baselines which focus on similar goals to ours. **GAN** is prominent in generating highly diverse captions, but it has negative effects on the correctness of captions, with much lower scores in SPICE, BLEU and CIDEr, and not much boost in retrieval performances. Human evaluation results in Section 4.6 further demonstrate that. By comparison, **Ours (Tdiv)** achieves comparable results in diversity (*Tdiv* and *Dist* scores) and outperforms **GAN** in retrieval performances and conventional metrics such as SPICE. **IS** and **CL** focus on generating discriminative captions which are mainly evaluated by retrieval performances. They achieve high results in $R@K$, but have lower conventional metrics'

[2] The fraction of generated captions where the correct image is retrieved in the closest K points to the query caption in the shared embedding space.

[3] For each image, one sentence are randomly sampled from the annotations as the candidate and the others as the references. We notice that in this method we only have 4 references, so we further calculate BLEU/CIDEr/SPICE of models by randomly selecting 4 annotated captions out of 5 as references. BLEU and CIDEr results show a similar trend. Some results: Attention 30.7/108.5/21.1; Ours (Tdiv, 0.1) 27.8/104.8/22.2; Ours (Tdiv, 0.3) 24.1/86.7/22.3; Ours (IR, 2) 30.6/108.7/21.5.

[4] We find Ours (IR, 2) has higher R@K than Ours (Tdiv, 0.1) and Ours (Tdiv, 0.3) evaluated with the retrieval model of the same architecture as that used for training.

	length	Yngve	Tdiv	Dist-2	Dist-3	Dist-S	R@1	R@5	R@10	BLEU-4	CIDEr	SPICE
Attention (Base)	9.1	12.5	1.78	2511	4972	3228	19.2	47.8	61.5	35.0	109.8	19.9
GAN	10.6	15.1	2.38	4418	9321	4365	21.4	48.8	62.4	23.1	86.0	18.7
IS	9.4	13.4	2.04	4772	9016	4248	24.8	54.9	68.8	31.9	101.8	19.7
CL	9.3	12.7	1.86	3103	6130	3499	24.1	52.5	67.5	33.6	106.5	19.7
CIDEr-RL	9.4	12.8	1.79	1843	3694	2909	19.4	47.0	61.3	36.0	115.5	20.9
DiscCap	9.3	12.3	1.74	1743	3512	3093	21.6	50.3	65.4	**36.1**	114.2	21.0
Stack-Cap	9.4	12.7	1.74	1930	3999	3268	21.9	49.7	63.7	**36.1**	**120.4**	20.9
Ours (len, 0)	**15.4**	**28.7**	**3.10**	3916	8683	4565	26.2	56.5	**70.8**	24.6	69.7	21.0
Ours (Yngve, 0)	10.6	17.7	2.12	2904	5895	3634	22.6	50.8	65.6	32.3	106.4	20.8
Ours (IR, 2)	9.3	13.0	1.88	3439	6726	3884	25.3	55.5	69.1	35.0	109.8	20.3
Ours (Tdiv, 0.1)	10.8	16.3	2.26	3873	7810	4196	25.0	55.5	69.7	31.5	105.4	21.0
Ours (Tdiv, 0.3)	12.9	21.7	2.78	**4790**	**10053**	**4576**	**26.3**	**57.2**	**70.8**	27.1	86.9	**21.1**
Human	10.5	15.9	2.90	15381	25309	4992	30.3	59.4	72.4	19.4	85.8	21.3

Table 1. Automatic evaluation results on COCO test set. The words and numbers in parenthesis denote the choices of scoring function and ϵ, respectively. (e.g. (IR, 2) denotes using image retrieval (IR) scoring function and ϵ=2). R@K, BLEU-4, CIDEr and SPICE values are reported as the percentage.

scores. **Ours (IR, 2)** not only outperforms **IS** and **CL** in *R@K*, but also shows advantages in conventional metrics.

With no surprise, captions generated by reinforcement learning (RL) based methods (**CIDEr-RL**, **DiscCap** and **Stack-Cap**) achieve high BLEU and CIDEr scores, as they directly use CIDEr as an optimization objective. However, they have even lower *Dist* scores than **Attention** baseline, indicating CIDEr optimization hurts the diversity of captions. Furthermore, retrieval performances of these methods only improve a bit, which are still much lower than ours. The results are consistent with prior work finding n-gram-based metrics like BLEU and CIDEr correlate negatively with detailedness [1] and diversity [40] of image captions.

In summary, automatic evaluation results demonstrate that our models with **IR** and **Tdiv** scoring functions effectively generate diverse and descriptive captions. **Ours (IR)** has advantages in retrieval performances, and comparable BLEU and CIDEr scores with MLE baselines. **Ours (Tdiv)** is prominent in producing diverse and descriptive captions according to *Tdiv*, *Dist*, and *R@K* scores, while maintaining correctness reflected by higher SPICE and *R@K* scores, despite fewer overlapping n-grams with references.

	Relevance	Fluency	Diversity	Descriptiveness
Attention	3.48	3.88	2.90	2.88
Stack-Cap	3.47	**3.97**	2.91	2.89
IS	3.41	3.65	3.16*	3.11*
GAN	3.32	3.58	3.51*	3.19*
CL	3.43	3.92	3.06*	2.98
DiscCap	3.67*	3.95	2.87	2.97
Ours (IR, 2)	3.64*	3.93	3.15*	3.14*
Ours (Tdiv, 0.1)	**3.79***	3.87	3.54*	3.44**
Ours (Tdiv, 0.3)	3.69*	3.81	**4.06****	**3.94****

Table 2. Human evaluation results. * and ** indicate that a model has a statistically significantly higher score than Attention baseline and all baselines, respectively (t-test with $p \leq 0.05$).

	Tdiv	Dist-3	R@1	BLEU	CIDEr	SPICE
w/o first	2.15	6957	24.1	31.0	102.6	20.0
pretrain w/o first	2.17	6985	23.6	30.9	102.9	20.3
Ours (Tdiv, 0.1)	**2.26**	**7810**	**25.0**	**31.5**	**105.4**	**21.0**

Table 3. Results validating the effectiveness of the two-step process.

4.6. Human Evaluation Results

Table 2 shows the results of human evaluation of our models and several baseline methods. **Attention** and **Stack-Cap** are baseline methods with MLE and RL training respectively. Other baselines focus on similar objectives with ours in diversity or descriptiveness of captions. Compared to **Attention** and **Stack-Cap**, our three models perform significantly better in *Diversity* and *Descriptiveness*, while basically maintaining *Fluency*. **IS** and **GAN** damage *Fluency* and *Relevance* while obtaining gains in *Diversity* and *Descriptiveness*. **CL** slightly increases scores of *Diversity* and *Descriptiveness*. **DiscCap** improves the *Relevance* score which is comparable to ours, but only has little gain in *Descriptiveness*. Compared to these related baselines, our models achieve satisfactory performances in all four metrics. In particular, **Ours (Tdiv)** achieves better results than all baselines in *Relevance*, *Diversity* and *Descriptiveness*. The results are also in consistency with their higher scores in automatic metrics like SPICE, *Tdiv*, *Dist*, and *R@K*. *Fluency* scores are slightly lower as sentences are longer and much more diverse.

	Tdiv	Dist-3	R@1	BLEU	CIDEr	SPICE
Attention	1.78	4972	19.2	35.0	109.8	19.9
First Decoding	1.72	4468	19.6	**35.4**	**110.7**	19.9
Second Decoding	**2.26**	**7810**	**25.0**	31.5	105.4	**21.0**

Table 4. Comparison of first-step output (First Decoding) and final output (Second Decoding (Tdiv, $\epsilon = 0.1$).

4.7. Model Analysis

Effects of parameter ϵ: Hyperparameter ϵ controls how much difference C_i and C_j in a selected paraphrase pair (C_i, C_j) have in terms of the scoring function. It also influences the number of paraphrase pairs in $\mathcal{P}$. Automatic results in Table 1 show that increasing ϵ with Tdiv scoring function will lead to higher diversity and descriptiveness but lower BLEU and CIDEr scores, as the gap between C_i and C_j in diversity is increasing. There is a trade-off between diversity and n-gram accuracy. Changing hyper-parameter ϵ is a way to controls the trade-off. Higher ϵ encourages diversity (with higher *Dist/R@K*), resulting in less n-gram overlaps (lower BLEU/CIDEr). Apart from high scores in R@K and Dist, Ours(Tdiv) is consistent in SPICE with the change of ϵ. Table 2 indicates that **Ours (Tdiv, 0.3)** achieves much higher scores than **Ours (Tdiv, 0.1)** in *Diversity* and *Descriptiveness*, with a bit lower scores in *Relevance* and *Fluency*.

Effectiveness of Two-Step Decoding Process: Concerning that directly generating a more diverse and descriptive caption while maintaining the correctness is difficult, our model adopts a two-step decoding strategy. We evaluate the effectiveness of the two-step process by experiments shown in Table 3. We take **Ours (Tdiv, 0.1)** as an example and compare it to two baselines which only utilize one-step decoding. The baseline (w/o first) only uses the selected second-stage caption C_j in selected pair (C_i, C_j) with higher Tdiv scores for training, so it utilizes the same data distribution as the training of $\mathbf{LSTM}_2$ in our model. Another baseline (pretrain, w/o first) first pretrains the attention-based captioning models on the full training set and then only uses the second-stage captions with higher Tdiv scores for fine-tuning. Experimental results in Table 3 show that baseline models generate less diverse and descriptive sentences with lower *Tdiv*, *Dist-3* and *R@1* scores than our model. Furthermore, without the two-step process, the correctness of captions is also affected, resulting in all lower scores in conventional metrics and *R@1*.

Comparison of Outputs of Two Decoding Steps: Table 4 compares the results of the first-step output (First Decoding) and the final output (Second Decoding) from our two-stage decoding model (**Tdiv,** $\epsilon = 0.1$). Preliminary captions generated by the first stage is "simpler" than the outputs of attention baseline and final outputs, with lower *Tdiv* and *Dist-3* scores. But BLEU and CIDEr scores of the first output are higher. Meanwhile, our final output sentences are much more diversified and achieve a much higher SPICE score.

Attention: a person riding a motorcycle on a road
Ours (IR): a person on a motorcycle in a race
Ours (Tdiv): a man riding a motorcycle down a road next to a pile of hay
Human: a person on a motorcycle riding beside hay bales

Attention: a black and white photo of an airplane
Ours (IR): an airplane sitting on a runway with people on the ground
Ours (Tdiv): a black and white photo of an airplane parked at an airport with people standing around
Human: an airplane with people under the wings at a field

Figure 3. Examples of captions generated by different models and a human-written caption from COCO test set.

4.8. Examples

Figure 3 shows example images with captions generated by **Attention** baseline model, **Ours (IR, 2)** and **Ours (Tdiv, 0.3)**. Caption elements in different colors indicate different detailed descriptions of images. Our model describes images with more diverse expressions using phrases like *parked at an airport* and more detailed descriptions referring to important parts like *a pile of hay*. Captions generated by attention baseline are correct but lacking diversity and descriptiveness. More examples can be found in the supplementary material.

5. Conclusions

In this study, we focus on improving the diversity and descriptiveness of image captions by proposing a captioning model exploring the role of visual paraphrases, together with a variety of scoring functions for selecting useful paraphrase pairs. Our model firstly generates a preliminary caption and then paraphrases it into a polished caption. Our model can generate better captions with diversity and descriptiveness compared to some state-of-the-art models while maintaining correctness. We will explore better scoring functions and network architectures in the future.

Acknowledgment

This work was supported by National Natural Science Foundation of China (61772036) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We appreciate Pengcheng Yang and the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*. Springer, 2016. 5, 7

[2] 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 *IEEE International Conference on Computer Vision*, 2018. 2, 5

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2014. 2, 3

[4] Rahul Bhagat and Eduard Hovy. What is a paraphrase? *Computational Linguistics*, 39(3):463–472, 2013. 1, 2

[5] Jianfu Chen, Polina Kuznetsova, David Warren, and Yejin Choi. Déja image-captions: A corpus of expressive descriptions in repetition. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 504–514, 2015. 2

[6] Chenhui Chu, Mayu Otani, and Yuta Nakashima. iParaphrasing: Extracting visually grounded paraphrases via an image. In *International Conference on Computational Linguistics*, 2018. 2

[7] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional GAN. In *IEEE International Conference on Computer Vision*, 2017. 2, 6

[8] Bo Dai and Dahua Lin. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems*, pages 898–907, 2017. 2, 4, 6

[9] Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G. Schwing, and David A. Forsyth. Diverse and controllable image captioning with part-of-speech guidance. *arXiv preprint arXiv:1805.12589*, abs/1805.12589, 2018. 2, 3

[10] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[11] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 6

[12] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[13] Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. Stack-captioning: Coarse-to-fine learning for image captioning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2, 6

[14] Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. A deep generative framework for paraphrase generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE International Conference on Computer Vision*, 2016. 5, 6

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8), 1997. 3

[17] Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, 2016. 2

[18] Andrej Karpathy and Li Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *IEEE International Conference on Computer Vision*, 2015. 1, 2, 5

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015. 5

[20] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 317–325, 2017. 2

[21] 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

[22] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*, 2015. 1, 2

[23] Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. Paraphrase generation with deep reinforcement learning. *arXiv preprint arXiv:1711.00279*, 2017. 2

[24] 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*. Springer, 2014. 1, 5

[25] Xiao Lin and Devi Parikh. Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 2

[26] Lixin Liu, Xiaojun Wan, and Zongming Guo. Images2poem: Generating chinese poetry from image streams. In *2018 ACM Multimedia Conference*, pages 1967–1975. ACM, 2018. 2

[27] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 6

[28] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 4, 6

[29] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics, 2002. 5

[30] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. Areas of attention for image captioning. In *IEEE International Conference on Computer Vision*, 2017. 2

[31] Aaditya Prakash, Sadid A Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. Neural paraphrase generation with stacked residual LSTM networks. *arXiv preprint arXiv:1610.03098*, 2016. 2

[32] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*, 2016. 2

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 5

[34] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 3, 6

[35] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2, 6

[36] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 5

[37] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *The Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[38] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1, 2

[39] 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

[40] Qingzhong Wang and Antoni B Chan. Describing like humans: on diversity in image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 7

[41] Yufei Wang, Zhe Lin, Xiaohui Shen, Scott Cohen, and Garrison W Cottrell. Skeleton key: Image captioning by skeleton-attribute decomposition. *IEEE International Conference on Computer Vision*, 2017. 3

[42] Zhuhao Wang, Fei Wu, Weiming Lu, Jun Xiao, Xi Li, Zitong Zhang, and Yueting Zhuang. Diverse image captioning via grouptalk. In *International Joint Conferences on Artificial Intelligence*, 2016. 2

[43] Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, June 1989. 5

[44] Qi Wu, Chunhua Shen, Anton van den Hengel, Peng Wang, and Anthony Dick. Image captioning and visual question answering based on attributes and external knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1367–1381, 2018. 2

[45] Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Deliberation networks: Sequence generation beyond one-pass decoding. In *Advances in Neural Information Processing Systems*, pages 1784–1794, 2017. 2

[46] Jingjing Xu, Xuancheng Ren, Junyang Lin, and Xu Sun. Diversity-promoting GAN: A cross-entropy based generative adversarial network for diversified text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3940–3949, 2018. 2, 5

[47] Kelvin Xu, Aaron Courville, Richard S Zemel, and Yoshua Bengio. Show, attend and tell : Neural image caption generation with visual attention. In *International Conference on Machine Learning*, 2015. 1, 2, 3

[48] 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

[49] 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

[50] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *IEEE International Conference on Computer Vision*, 2017. 2

[51] Victor H Yngve. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104(5):444–466, 1960. 4

[52] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[53] Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems*, pages 1810–1820, 2018. 2

[54] Zhihao Zhu, Zhan Xue, and Zejian Yuan. Think and tell: Preview network for image captioning. In *British Machine Vision Conference*, 2018. 3