# Describing like Humans: on Diversity in Image Captioning

Qingzhong Wang and Antoni B. Chan
Department of Computer Science
City University of Hong Kong
qingzwang2-c@my.cityu.edu.hk, abchan@cityu.edu.hk

## Abstract

*Recently, the state-of-the-art models for image captioning have overtaken human performance based on the most popular metrics, such as BLEU, METEOR, ROUGE and CIDEr. Does this mean we have solved the task of image captioning? The above metrics only measure the similarity of the generated caption to the human annotations, which reflects its accuracy. However, an image contains many concepts and multiple levels of detail, and thus there is a variety of captions that express different concepts and details that might be interesting for different humans. Therefore only evaluating accuracy is not sufficient for measuring the performance of captioning models – the diversity of the generated captions should also be considered. In this paper, we proposed a new metric for measuring diversity of image captions, which is derived from latent semantic analysis and kernelized to use CIDEr similarity. We conduct extensive experiments to re-evaluate recent captioning models in the context of both diversity and accuracy. We find that there is still a large gap between the model and human performance in terms of both accuracy and diversity, and the models that have optimized accuracy (CIDEr) have low diversity. We also show that balancing the cross-entropy loss and CIDEr reward in reinforcement learning during training can effectively control the tradeoff between diversity and accuracy of the generated captions.*

## 1. Introduction

The task of image captioning is challenging and draws much attention from researchers in the fields of both computer vision and natural language processing. A large variety of models have been proposed to automatically generate image captions, and most of the models are engaged in improving the accuracy of the generated captions as measured by the current metrics, such as BLEU 1-4 [22], METEOR [7], ROUGE [16], CIDEr [28] and SPICE [1]. However, another important property, the **diversity** of captions generated for a given image, receives less attention. Generally, **diversity** refers to the differences among a set of captions generated by a method for a single image, and can be categorized into three levels: (1) *word diversity* refers to only changes of single words that do not change the caption's semantics, *e.g.*, using synonyms in different captions; (2) *syntactic diversity* refers to only differences in the word order, phrases, and sentence structures, such as pre-modification, post-modification, redundant and concise descriptions, which do not change the caption's concept. (3) *semantic diversity* refers to the differences of expressed concepts, including level of descriptive detail, changing of the sentence's subject, and addition/removal of sentence objects. For example, in Figure 1, the human captions 2 and 5 have syntactic diversity, as they both express the same concept of a zebra near a tree in a grass field using different word orderings. In contrast, captions 2 and 3 exhibit semantic diversity, as caption 3 describes the type of grass field ("lush dry") and omits "near a tree" as unimportant. Ideally, a caption system should be able to generate captions expressing different concepts in the image, and hence in this paper we focus on measuring *semantic diversity*.

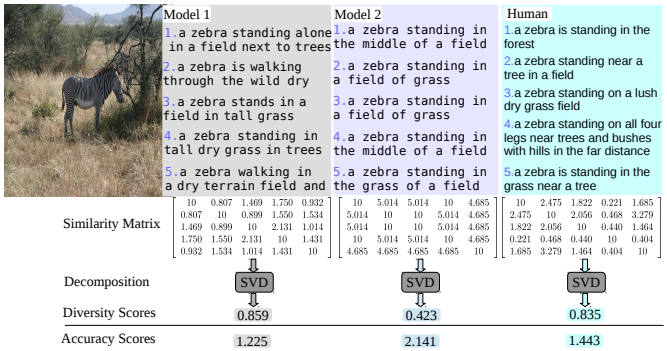


Figure 1: An overview of our diversity metric. Given a set of captions from a method, we first construct the self-similarity matrix $\mathbf{K}$, consisting of CIDEr [28] scores between all pairs of captions. The diversity score is computed from the singular values of $\mathbf{K}$. A higher diversity score indicates more variety in the set of generated captions, such as changes in the level of descriptive detail and inclusion of removal of objects. The accuracy (average CIDEr) of the captions with respect to the human ground-truth is on the bottom. For human annotations, this is the leave-one-out accuracy.

The motivations for considering diversity of image cap-

tions are as follows. First, an image may contain many concepts with multiple levels of detail — indeed, *an image is worth a thousand words* — and thus an image caption expresses a set of concepts that are interesting for a particular human. Hence, there is diversity among captions due to diversity among humans, and an automatic image captioning method should reflect this. Second, only focusing on increasing the caption accuracy will bias the captioning method to common phrases. For example, Figure 1 shows the set of captions generated by two models. Model 2 is the best when only considering accuracy. However, Model 2 just repeats the same common phrases, providing no particular additional details. In contrast, Model 1 recognizes that there are trees in the image and the the grass is dry, which also occurs in the human annotations. It even recognizes "walking", which does not appear in the human annotations, but is a plausible description. Thus, to mimic the ability of humans, the captioning models should also have the ability of generating diverse captions. Third, from the machine learning viewpoint, captioning models are typically trained on datasets where each image has at least 5 ground-truth captions (e.g., MSCOCO), and thus captioning models should also be evaluated on how well the learned conditional distribution of captions given an image approximates that of the ground-truth. In particular, while the caption accuracy measures the differences in the modes of the distributions, the caption diversity measures the variance of the distribution.

Recently, while a few works have focused on generating both diverse and accurate captions, such as conditional variational auto-encoders (CVAE) [30] and conditional generative adversarial network (CGAN) [5, 26], there is not a metric to well evaluate the diversity of captions. In [5], the diversity of captions is shown only qualitatively. [26, 30] evaluate the diversity of captions in three ways: 1) the percentage of novel sentences; 2) the percentage of unique uni-grams and bi-grams in the set of captions; 3) mBLEU, which is the average of the BLEU scores between each caption and the remaining captions. However, it is difficult to define a novel sentence, and only considering the percentage of unique uni-grams and bi-grams ignores the relationship between captions, e.g., the same n-gram could be used to construct sentences with different meanings. Because mBLEU uses the BLEU score, it aggregates n-grams over all the remaining captions, which obfuscates differences among the individual captions, thus under-representing the diversity. For example, the two caption sets, $\mathcal{C}_1 = \{$*"zebras grazing grass"*, *"grazing grass"*, *"zebras grazing"*$\}$ and $\mathcal{C}_2 = \{$*"zebras grazing"*, *"zebras grazing"*, *"zebras grazing"*$\}$, obtain the same mBLEU of 1.0. However, we may consider that $\mathcal{C}_1$ is more diverse, because each of its captions expresses different concepts or details. In contrast, captions in $\mathcal{C}_2$ describe exactly the same thing. Hence, considering all the pairwise relationships among captions will better reflect the structure of the set of captions. Moreover, BLEU is not a good metric for measuring semantic differences, since phrase-level changes and semantic changes may lead to the same BLEU score (e.g., see Table 1).

In this paper, we propose a diversity measure based on pairwise similarities between captions. In particular, we form a matrix of pairwise similarities (e.g., using CIDEr), and then use the singular values of the matrix to measure the diversity. We show that this is interpretable as applying latent semantic analysis (LSA) on the weighted n-gram feature representation of the captions to extract the topic-structure of the set of captions, where more topics indicates more diversity in the captions. The key contributions of this paper are three-fold: 1) we proposed a new metric for evaluating diversity of sets of captions, and we re-evaluate existing captioning models via considering both diversity and accuracy. Moreover, our proposed metric shows a stronger correlation to human evaluation than mBLEU; (2) we develop a framework that enables a tradeoff between diverse and accurate captions via balancing the rewards in reinforcement learning (RL) and the cross-entropy loss; (3) extensive experiments are conducted to demonstrate the effectiveness of the diversity metric and the effect of the loss function on diversity and accuracy – we find that RL and adversarial training are different approaches that provide equally satisfying results.

## 2. Related Work

**Image Captioning.** Early image captioning models normally contain 2 stages: 1) concept detection, 2) translation. In the first stage, object categories, attributes and activities are detected, then the translation stage uses the labels to generate sentences. The typical concept detection models are conditional random fields (CRFs) [9, 13], support vector machines (SVMs) [15] or convolutional neural networks (CNNs) [8], and the translation model is a sentence template [9] or n-gram model [15].

Recently, the encoder-decoder models, e.g., neural image captioning (NIC) [29], spatial attention [34] and adaptive attention [19], trained end-to-end have obtained much better results than the early models based on concept detection and translation. NIC [29] translates images into sentences via directly connecting the inception network to an LSTM. To improve NIC, [34] introduces a spatial attention module, which allows the model to "watch" different areas when it predicts different words. [10, 33, 35, 36] use image semantics detected using an additional network branch. In [19], a sentinel gate decides whether the visual feature or the semantic feature should be used for prediction. Instead of employing LSTM decoders for sentences, [2, 31, 32] apply convolutional decoders, which achieves faster training process and comparable results.

Both LSTM and convolutional models are trained using

| Modification | Caption | B1 | B2 | B3 | B4 | M | R | C/10 | S |
|---|---|---|---|---|---|---|---|---|---|
| Reference | a group of people are playing football on a grass covered field | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Word-level | a **couple** of **boys** are playing **soccer** on a grass covered field | 0.750 | 0.584 | 0.468 | 0.388 | 0.387 | 0.750 | 0.261 | 0.333 |
| Phrase-level | **some guys** are playing football on a **grassy ground** | 0.417 | 0.389 | 0.357 | 0.317 | 0.310 | 0.489 | 0.441 | 0.133 |
| Sentence-level | **on a grass covered field** a group of people are playing football | 1.000 | 0.953 | 0.899 | 0.834 | 0.581 | 0.583 | 0.676 | 0.941 |
| Redundancy | a group of people **in red soccer suits** are playing football on a grass covered field | 0.716 | 0.683 | 0.644 | 0.598 | 0.429 | 0.836 | 0.496 | 0.818 |
| Conciseness | a group of people are playing football | 0.583 | 0.564 | 0.542 | 0.516 | 0.526 | 0.774 | 0.482 | 0.714 |
| Average | | 0.693 | 0.635 | 0.582 | 0.531 | 0.447 | 0.693 | 0.471 | 0.588 |
| Semantic change | a group of people are **watching TV** | 0.417 | 0.389 | 0.357 | 0.317 | 0.270 | 0.553 | 0.072 | 0.429 |

Table 1: The similarity scores between a reference caption and a modified caption using different evaluation metrics. The caption in the first row is the reference caption, and the next five captions change different parts of the sentence (highlighted in bold) while keeping the same concepts. "Average" is the average metric value over these 5 modified captions. The bottom row shows an incorrect caption and the metric scores. B1-4, M, R, C/10, and S are BLEU1-4, METEOR, ROUGE, CIDEr divided by 10 (so that the maximum is 1), and SPICE.

cross-entropy. In contrast, [17, 23] directly improve the evaluation metric using reinforcement learning (RL). They also show that improving the CIDEr reward function also improves other evaluation metrics, but not vice versa. Instead of using metric rewards, [18, 20] employ the retrieval reward to generate more distinctive captions.

Generally, the above models are used to generate a single caption for one image, whereas [5, 26] use CGAN to generate a set of diverse captions for each image. The generator uses an LSTM to generate captions given an image, and the evaluator uses a retrieval model to evaluate the generated captions. The generator and evaluator are jointly trained in adversarial manner using policy gradients[1]. In the inference stage, latent noise vectors are sampled from a Gaussian distribution, generating different captions. CVAE [30] is another model that is able generate diverse caption by sampling the latent noise vector.

**Evaluation Metrics.** The most popular metrics are BLEU [22], METEOR [7], ROUGE [16], which are metrics from machine translation and document summarization, and CIDEr [28] and SPICE [1], which are metrics specific to image captioning. BLEU, METEOR, ROUGE and CIDEr are based on computing the overlap between the n-grams of a generated caption and those of the human annotations. BLEU considers the n-gram precision, ROUGE is related to n-gram recall, which benefits long texts, and METEOR takes both precision and recall of unigrams, while also applying synonym matching. CIDEr uses TF-IDF weighted n-grams to represent captions and calculates the cosine similarity.

Only considering n-gram overlap seems to ignore semantics of the captions. SPICE uses scene graphs [11, 24] to represent images – human annotations and one generated caption are first parsed into scene graphs, which are composed of object categories, attributes and relationships, and the F1-measure is computed between the two scene graphs. However, SPICE is highly dependent on the accuracy of the parsing results. [12] proposed a metric based on word2vec [21] and word mover distance (WMD) [14], which could leverage semantic information, but depends on the quality of word2vec. Recently, [4] proposed a learned metric that uses a CNN and an LSTM to extract features from images and captions, and then uses a classifier to assign a score that indicates whether the caption is generated by a human. While this metric is robust, it requires training and data augmentation, and the evaluation procedure takes more time.

Table 1 shows an example of similarity metrics between a reference caption and 5 modified captions that have the same semantic meaning, and an incorrect caption with different meaning. All metrics are less sensitive (have higher values) to sentence-level changes (due to the use of n-grams), in particular BLEU, ROUGE and SPICE. Furthermore, all metrics show sensitivity to word-level or phrase-level changes. Overall, CIDEr and METEOR have relatively low average metric value, which means that they are sensitive to sentence changes that keep the same semantics. On the other hand, CIDEr and METEOR also assign lower values to the incorrect caption that changes the semantic meaning, which indicates that they are better able to discriminate between semantic changes in the sentence. Hence, in this paper, we mainly consider CIDEr as the baseline metric to evaluate both the diversity and accuracy.

## 3. Measuring Diversity of Image Captions

Currently, the widely used metrics, such as BLEU, CIDEr, and SPICE are for a single caption prediction. To evaluate a set of captions $\mathcal{C} = \{c_1, c_2, \cdots, c_m\}$, two dimensions are required: accuracy and diversity. For accuracy, the standard approach is to average the similarity scores,

[1]This is similar to RL models, but RL models are trained by maximizing the rewards, which is not adversarial training.

$acc = \frac{1}{m}\sum_i s_i$, where $s_i = sim(c_i, \mathcal{C}_{GT})$ is the similarity measure (e.g., CIDEr) between caption $c_i$ and ground-truth caption set $\mathcal{C}_{GT}$. For diversity, we will consider the pairwise similarity between captions in $\mathcal{C}$, which is able to reflect the underlying structure of the set of captions.

### 3.1. Latent Semantic Analysis

Latent semantic analysis (LSA) [6] is a linear representation model, which is widely applied in information retrieval. LSA considers the co-occurrence information of words (or n-grams), and uses singular value decomposition (SVD) to obtain a low-dimensional representations of the documents in terms of topic vectors. Applying LSA to a caption set, more topics indicates a more diverse set of captions, whereas only one topic indicates a non-diverse set. To use LSA, we first represent each caption via a vector. In this subsection, we consider the simplest representation, bag-of-words (BoW), and kernelize it in the next subsection using CIDEr.

Given a set of captions $\mathcal{C} = \{c_1, \cdots, c_m\}$ that describe an image, and a dictionary $\mathcal{D} = \{w_1, w_2, \cdots, w_d\}$, we use the word-frequency vector to represent each caption $c_i$, $\mathbf{f}_i = [f_1^i, \cdots, f_d^i]^T$, where $f_j^i$ denotes the frequency of word $w_j$ occurring in caption $c_i$. The caption set $\mathcal{C}$ can be represented by a "word-caption" matrix, $\mathbf{M} = [\mathbf{f}_1 \cdots \mathbf{f}_m]$.

Applying SVD, we decompose $\mathbf{M}$ into three matrices, *i.e.*, $\mathbf{M} = \mathbf{U}\mathbf{S}\mathbf{V}^T$, where $\mathbf{U}$ is composed of the eigenvectors of $\mathbf{M}\mathbf{M}^T$ and $\mathbf{S} = \text{diag}(\sigma_1, \cdots, \sigma_m)$ is a diagonal matrix consisting of singular values $\sigma_1 > \sigma_2 > \cdots > 0$ , and $\mathbf{V}$ is composed of the eigenvectors of $\mathbf{M}^T\mathbf{M}$. Each column of $\mathbf{U}$ represents the words in a topic vector of the caption set, while the singular values in $\mathbf{S}$ represent the strength (frequency) of the topics. If all captions in $\mathcal{C}$ are the same, then only one singular value is non-zero, i.e., $\sigma_1 > 0$ and $\sigma_i = 0, \forall i > 1$, If all the captions are different, then all the singular values are the same, i.e., $\sigma_1 = \sigma_i, \forall i$. Hence, the ratio $r = \frac{\sigma_1}{\sum_{i=1}^m \sigma_i}$ represents how diverse the captions are, with larger $r$ meaning less diverse (i.e., the same caption), and smaller $r$ indicating more diversity (all different captions). The ratio $r$ is within $[\frac{1}{m}, 1]$. Thus we map the ratio to a value in $[0, 1]$, to obtain our diversity score $div = -\log_m(r)$, where larger $div$ means higher diversity.

Looking at the matrix $\mathbf{K} = \mathbf{M}^T\mathbf{M}$, each element $k_{ij} = \mathbf{f}_i^T\mathbf{f}_j$ is the dot-product similarity between the BoW vectors $\mathbf{f}_i$ and $\mathbf{f}_j$. As the dimension of $\mathbf{f}_i$ may be large, a more efficient approach to computing the singular values is to use the eigenvalue decomposition $\mathbf{K} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \cdots, \lambda_m)$ are the eigenvalues of $\mathbf{K}$, which are the squares of the singular values, $\sigma_i = \sqrt{\lambda_i}$. Note that $\mathbf{K}$ is a kernel matrix, and here LSA is using the linear kernel.

### 3.2. Kernelized Method via CIDEr

In Section 3.1, a caption is represented by BoW features $\mathbf{f}_i$. However, this only considers word frequency and ignores phrases and sentence structures. To address this problem, we use n-gram or p-spectrum kernels [25] with LSA. The mapping function from the caption space $\mathbb{C}$ to the feature space $\mathbb{F}$ associated with the n-gram kernel is

$$\phi^n(c) = [f_1^n(c) \cdots f_{|\mathcal{D}^n|}^n(c)]^T, \tag{1}$$

where $f_i^n(c)$ is the frequency of the $i$-th $n$-gram in caption $c$, and $\mathcal{D}^n$ is the $n$-gram dictionary.

CIDEr first projects the caption $c \in \mathbb{C}$ into a weighted feature space $\mathbb{F}$, $\Phi^n(c) = [\omega_i^n f_i^n(c)]_i$ where the weight $\omega_i^n$ for the $i$-th $n$-gram is its inverse document frequency (IDF). The CIDEr score is the average of the cosine similarities for each $n$,

$$CIDEr(c_i, c_j) = \frac{1}{4}\sum_{n=1}^{4} CIDEr_n(c_i, c_j), \tag{2}$$

where

$$CIDEr_n(c_i, c_j) = \frac{\Phi^n(c_i)^T\Phi^n(c_j)}{||\Phi^n(c_i)||\,||\Phi^n(c_j)||}. \tag{3}$$

In (3), $CIDEr_n$ is written as the cosine similarity kernel and the corresponding feature space is spanned by $\Phi^n(c)$. Since $CIDEr$ is the average of $CIDEr_n$ for different $n$, therefore, it is also a kernel function that accounts for uni-, bi-, tri- and quad-grams.

Since CIDEr can be interpreted as a kernel function, we reconsider the kernel matrix $\mathbf{K}$ in LSA, by using $k_{ij} = CIDEr(c_i, c_j)$. The diversity according to CIDEr can then be computed by finding the eigenvalues of the kernel matrix $\{\lambda_1, \cdots, \lambda_m\}$, computing the ratio $r = \frac{\sqrt{\lambda_1}}{\sum_{i=1}^m \sqrt{\lambda_i}}$, and applying the mapping function, $div = -\log_m(r)$. Here, we are computing the diversity by using LSA to find the caption topics in the weighted n-gram feature space, rather than the original BoW space. Other caption similarity measures could also be used in our framework to compute diversity if they can be written as positive definite kernel functions.

## 4. Experiment Setup

We next present our experiment setup re-evaluating current captioning methods using both diversity and accuracy.

### 4.1. Generating Diverse Captions

As most current models are trained to generate a single caption, we first must adapt them to generate a set of diverse captions. In this paper we propose 4 approaches to generate diverse captions from a baseline model. (1) **Random sampling (RS):** After training, a set of captions is generated by randomly sampling word-by-word from the learned conditional distribution $\hat{p}(c|I)$. (2) **Randomly cropped images (RCI):** The image is resized to $256 \times 256$, and then randomly cropped to $224 \times 224$ as input to generate the caption. (3) **Gaussian noise corruption (GNC):** Gaussian noise with different standard deviations is added to the input

image when predicting the caption. (4) **Synonym switch (SS):** The above 3 approached manipulate images to generate diverse captions, whereas the synonym switch approach directly manipulates a generated caption. First, a word2vec [21] model is trained on MSCOCO. Next, given a caption, the top-10 synonyms for each word are retrieved and given a weight based on the similarities of their word2vec representation. Finally, with probability $p$, each word is randomly switched with one of its 10 synonyms, where the synonyms are sampled according to their weights.

For the models that are able to generate diverse captions, such as CVAE and CGAN, **different random vectors (DRV)** are drawn from Gaussian distributions with different standard deviations to generate the captions.

### 4.2. Implementation Details

In this paper, we evaluate the following captioning models: (1) NIC [29] with VGG16 [27]; (2) SoftAtt [34] with VGG16; (3) AdapAtt [19] with VGG16; (4) Att2in [23] with cross-entropy (XE) and CIDEr reward, denoted as Att2in(XE) and Att2in(C); (5) FC [23] with cross-entropy and CIDEr reward, denoted as FC(XE) and FC(C); (6) Att2in and FC with retrieval reward[2] [20], denoted as Att2in(D5) and FC(D5), where the retrieval reward weight is 5 (the CIDEr reward weight is 1), and likewise for D10; (7) CVAE and GMMCVAE[3] [30], (8) CGAN [5].

Models (1)-(7) generate single caption for one image, and model (7) and (8) are able to generate diverse captions. The models are trained using Karpathy's training split of MSCOCO. We use each of the models to generate 10 captions for each image in the Karpathy's test split, which contains 5,000 images. The standard deviations of Gaussian noise for GNC and DRV are $\{1.0, 2.0, \cdots, 10.0\}$. For SS, we first generate a caption using beam search with beam-width 3, and then generate the other 9 captions by switching words with synonyms with probability $p \in \{0.1, 0.15, \cdots, 0.5\}$. Models and diversity generators are denoted as "model-generator", e.g., "NIC-RS".

The accuracy $acc$ of the generated captions $\mathcal{C}$ is the average CIDEr: $\frac{1}{m}\sum_{i=1}^{m} CIDEr(c_i, \mathcal{C}_{GT})$, where $c_i \in \mathcal{C}$ and $\mathcal{C}_{GT}$ is the set of human annotations. We also compute the *leave-one-out* accuracy of human annotations: $\frac{1}{N}\sum_{i=1}^{N} CIDEr(g_i, \mathcal{C}_{GT\setminus i})$, where $g_i \in \mathcal{C}_{GT}$ and $\mathcal{C}_{GT\setminus i}$ is the set of human annotations without the i-th annotation. The diversity of $\mathcal{C}$ is computed using the LSA-based method (denoted as LSA) and the kernel CIDEr method (denoted as Self-CIDEr), introduced in Sections 3.1 and 3.2.

## 5. Experiment Results

We next present our experiment results evaluating methods based on both diversity and accuracy.

[2]https://github.com/ruotianluo/DiscCaptioning

[3]https://github.com/yiyang92/vae_captioning

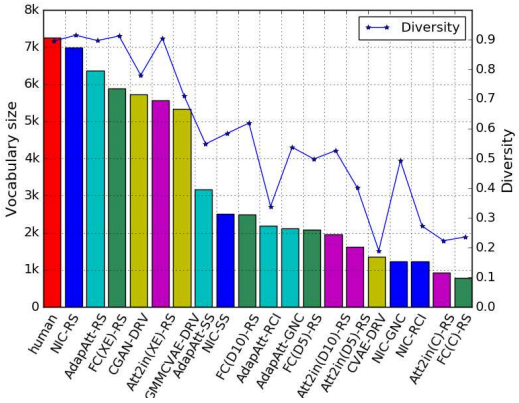


Figure 2: The vocabulary sizes and diversity scores (Self-CIDEr) of different caption models. The vocabulary of each trained model is collected from 50,000 captions (10 captions for each image), while the human annotations have 25,000 captions (5 captions for each image). For GNC, RCI, CVAE, GMMCVAE and CGAN, greedy search is used to generate captions.

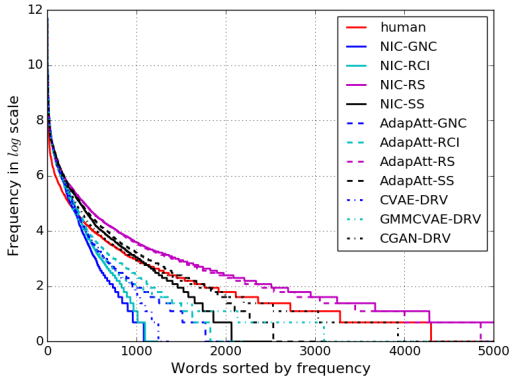


Figure 3: Word frequency plots for the top 5,000 most-frequent words for each captioning model.

### 5.1. Analysis of Caption Vocabulary

We first focus on the vocabulary of the generated captions from each model, including vocabulary size and word frequency. Generally, a large vocabulary size and long tail in the word frequency distribution indicates higher diversity.

Figure 2 shows the vocabulary sizes of different models (here we only show the most representative models), as well as the models diversity score (Self-CIDEr). Human annotations have the largest vocabulary, even though there are fewer human captions than model captions (25,000 for humans, and 50,000 for each model). For NIC and AdapAtt models, using RS results in larger vocabulary which also generates more diverse captions. Although AdapAtt is more advanced than NIC, the vocabulary size of AdapAtt-RS is smaller than that of NIC-RS. One possible reason is that models developed to obtain better accuracy metrics often learn to use more common words. Looking at reinforcement learning (e.g., Att2in(XE) vs. Att2in(C) vs. Att2in(D)),

| Corr Coef | Self-CIDEr | LSA | mBLEU-mix |
|---|---|---|---|
| overall Pearson $\rho$ | **0.616** | 0.601 | 0.585 |
| overall Spearman $\rho$ | **0.617** | 0.602 | 0.575 |
| avg. per image Spearman $\rho$ | 0.674 | **0.678** | 0.644 |

Table 2: Correlation between computed diversity metric and human diversity judgement: (top) overall correlation; (bottom) correlation of per-image rankings of methods.

using CIDEr reward to fine-tune the model will drastically decrease the vocabulary size so as to improve the accuracy metric (CIDEr) [23]. Interestingly, using a retrieval reward gives a larger vocabulary size compared to using the CIDEr reward. Improving retrieval reward encourages semantic similarity, while improving CIDEr reward encourages syntactic similarity, which leads to low diversity. Comparing the CGAN/CVAE methods, CVAE has a smaller vocabulary compared to CGAN and GMMCVAE, which indicates that the latter could generate more diverse captions. Note that the vocabulary sizes only roughly reflects the diversity – a small vocabulary could lead to diverse captions via using different combinations of words, Hence, it is important to look at the pairwise similarity between captions.

Figure 3 shows the frequency plots of each word used by the models. If a model employs diverse words, the plots in Figure 3 should have a long tail. However, most of the models have learned to use around 2,000 common words. In contrast, CGAN and GMMCVAE encourage a longer-tail distribution, and in particular the word frequency plot of CGAN is similar to the human annotations. RS tends to give the most words, but also fails to generate fluent sentences. Therefore, we suggest that both accuracy and diversity should be considered to evaluate a model. Interestingly, there is a very large gap between using cross-entropy and CIDEr rewards for reinforcement learning, which is bridged by the retrieval reward. In Section 5.3, we will show that balancing cross-entropy, CIDEr, and retrieval rewards can also provide good results in terms of diversity and accuracy.

## 5.2. Considering Diversity and Accuracy

Here we re-evaluate the models accounting for both diversity and accuracy. Figure 4 shows the diversity-accuracy (DA) plots for LSA-based diversity and CIDEr kernelized diversity (Self-CIDEr). The trends of LSA and Self-CIDEr are similar, although LSA yields overall lower values. Hence, we mainly discuss the results of Self-CIDEr.

After considering both diversity and accuracy, we may need to rethink what should be considered a good model. We suggest that a good model should be close to human performance in the DA space. Looking at the performance of humans, the diversity is much higher than Att2in(C), which is considered a state-of-the-art captioning model. On the other hand, the diversity using randomly sampling (RS) are closer to human annotations. However, the accuracy is poor, which indicates that the descriptions are not fluent or are off-topic. Therefore, a good model should well balance between diversity and accuracy. From this point of view, CGAN and GMMCVAE are among the best models, as they are closer to the human annotations in the DA space. Example caption results and their diversity/accuracy metrics can be found in the supplemental.

Most of the current state-of-the-art models are located in the bottom-right of the DA space, (high CIDEr score but poor diversity), as they aim to improve the accuracy. For example, directly improving CIDEr reward via RL is a popular approach to obtain higher CIDEr scores [17, 18, 20, 23], but it encourages using common words and phrases (also see Figure 2), which lowers the diversity. Using retrieval reward is able to improve diversity comparatively, e.g., Att2in(D5) vs Att2in(C), because it encourages distinctive words and semantic similarity, and suppresses common syntaxes that do not benefit retrieval. The drawback of using retrieval model is that the fluency of the captions could be poor [20], and using a very large weight for the retrieval reward will cause the model to repeat the distinctive words. Finally, note that there is a large gap between using the cross-entropy loss and the CIDEr reward for training, e.g., Att2in(XE) and Att2in(C). In the next subsection, we will consider building models to fill the performance gap by balancing between the losses.

Comparing the diversity generators, SS and GNC are more promising for generating diverse captions. Captions generated using RCI have higher accuracy, while those using RS have higher diversity. Interestingly, in the top-left of the DA plot, using RS, a more advanced model can generate more accurate captions without reducing the diversity, This shows that an advanced model is able to learn a better $\hat{p}(c|I)$, which is more similar to the ground-truth distribution $p(c|I)$. However, there is a long way to go to reach the accuracy of human annotations.

**Correlation to human evaluation.** We conduct human evaluation on Amazon Machine Turk (AMT). We use 100 images, and for each image we show the worker 9 sets of captions, which are generated in different ways: human annotations and 8 models, AdapAtt-SS, AdapAtt-GNC, AdapAtt-RCI, Att2in(XE)-RS, Att2in(C)-RS, Att2in(D5)-RS, Att2in(D10)-RS and CGAN-DRV. We require the workers to read all the sets of captions and then give scores (from 0 to 1) that reflects the diversity[4] of the set of captions. Each image is evaluated by 3 workers, and the diversity score for each image/model combination is the average score given by the 3 workers.

Fig. 5 (left, center) shows the correlation plots between our proposed metrics and human evaluation. The overall consistency between the proposed diversity metric and the human judgement is quantified using Pearson's (paramet-

[4]In our instructions, diversity refers to different words, phrases, sentence structures, semantics or other factors that impact diversity.

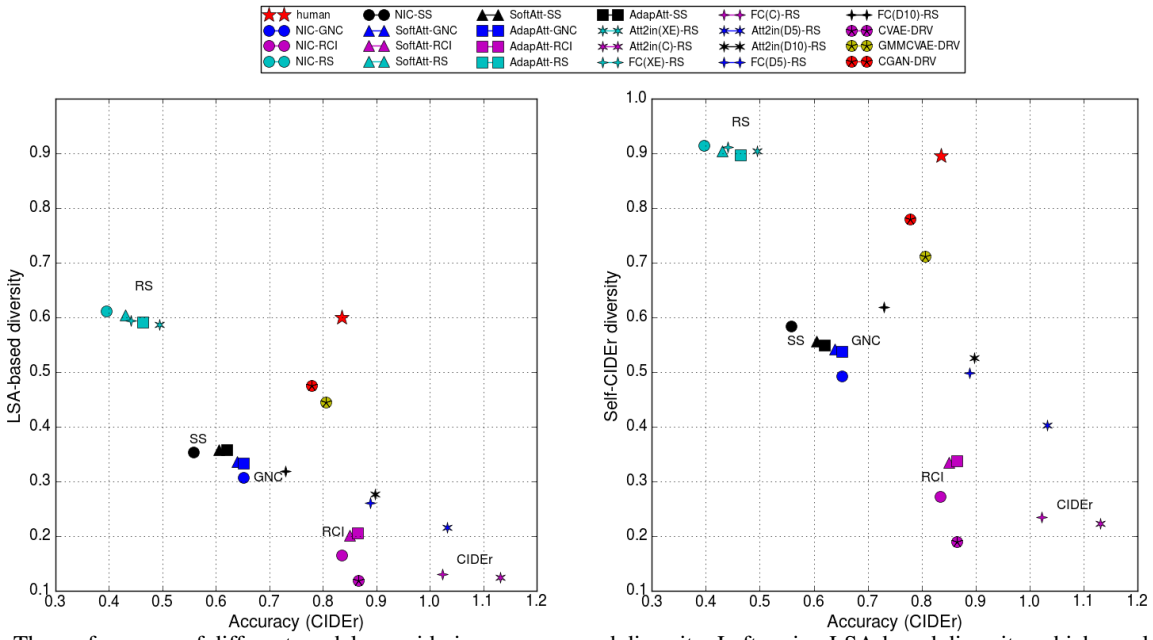


Figure 4: The performance of different models considering accuracy and diversity. Left: using LSA-based diversity, which employs BoW features. Right: using CIDEr kernelized diversity (Self-CIDEr). The marker shape indicates the caption model, while the marker color indicates the diversity generator or training method.

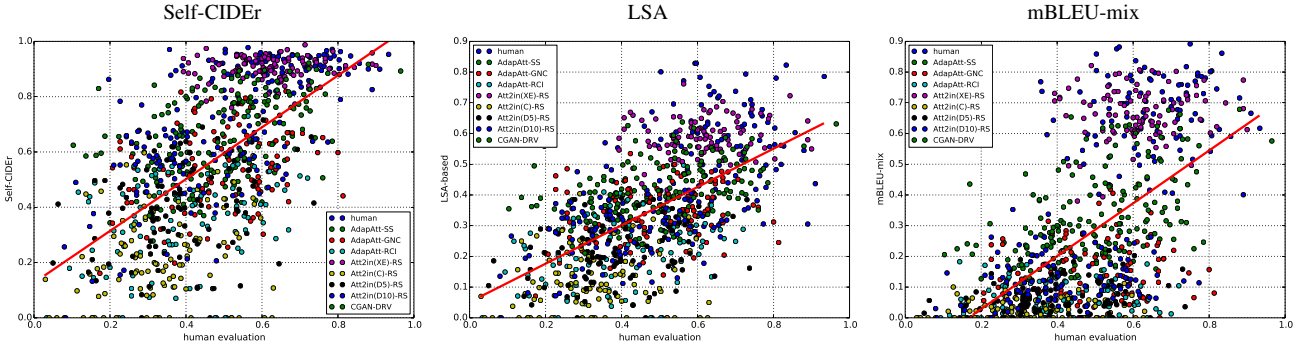


Figure 5: The correlation plots between the diversity scores computed by different metrics and human evaluation. The red lines are the best fit lines to the data.

ric) and Spearman's rank (non-parametric) correlation coefficients (see Table 2 top). Since the human annotator evaluated the diversity scores for all methods on each image, they were implicitly ranking the diversity of the methods. Hence, we also look at the consistency between the human rankings for an image and the rankings produced by the proposed metrics, as measured by the average per-image Spearman rank correlation (see Table 2 bottom). Both Self-CIDEr and LSA-based metrics are largely consistent with human evaluation of diveresity, with Self-CIDEr having higher overall correlation, while both have similar per-image ranking of methods.

We compare our metrics with $mBLEU_{mix} = 1 - \frac{1}{4}\sum_{n=1}^{4} mBLEU_n$, which accounts for mBLEU-{1,2,3,4}, and we invert the score so that it is consistent with our diversity metrics (higher values indicate more diversity). The correlation plot between mBLEU-mix and human judgement is shown in Fig. 5 (right). mBLEU-mix has lower correlation coefficient with human judgement, compared to LSA and Self-CIDEr (see Table 2). Similar results are obtained when looking at the mBLEU-n scores. Self-CIDEr has better overall correlation with human judgement, while the two methods are comparable in terms of per-image consistency of method ranking.

In addition the correlation plot shows the mBLEU scale is not uniformly varying, with more points falling at the lower and higher ends of the scale and less points in the middle. In contrast, LSA and Self-CIDEr have more uniform scales.

### 5.3. Re-thinking Reinforcement Learning

In this subsection we further investigate RL for image captioning, and how to bridge the gap between the perfor-

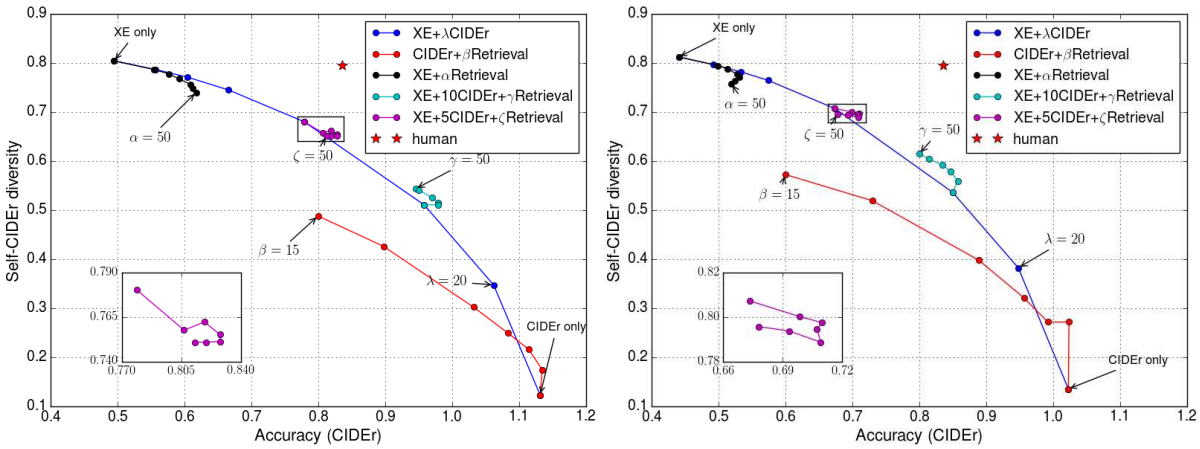


Figure 6: The diversity and accuracy performance of Att2in (left) and FC (right) with different loss functions. XE, CIDEr, and Retrieval denote the cross-entropy loss, CIDEr reward [23] and retrieval reward [20]. The weights are $\lambda \in \{0, 1, 2, 3, 5, 10, 20\}$, $\beta \in \{0, 1, 2, 3, 5, 10, 15\}$, $\alpha \in \{0, 5, 10, 20, 30, 40, 50\}$, $\gamma \in \{0, 10, 20, 30, 40, 50\}$ and $\zeta \in \{0, 5, 10, 20, 30, 40, 50\}$. The inset plot in the bottom-left is a zoom-in of rectangle in the main plot.

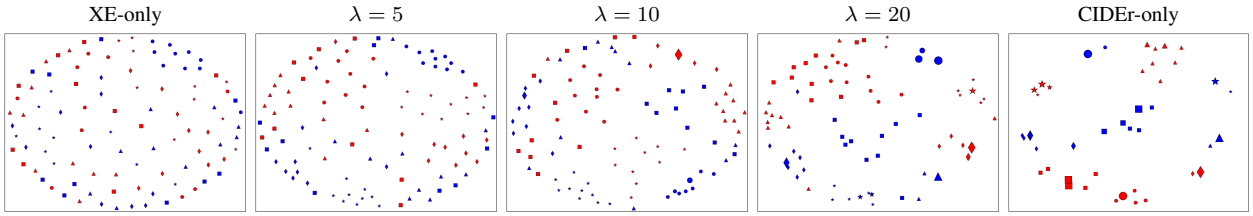


Figure 7: MDS visualization of the similarity between captions of 10 images. 10 captions are randomly sampled from Att2in for each image. Markers and colors indicate different images. Larger markers indicates multiple captions located at the same position.

mance using cross-entropy and using CIDEr reward. In particular, we train Att2in and FC using different loss functions that combine the cross-entropy, CIDEr reward, and retrieval reward, with varying weights.

The results are shown in Figure 6. Balancing the XE loss and CIDEr reward is the most effective way to bridge the gap. Using larger weight $\lambda$ results in higher accuracy but lower diversity. Based on our experiments, using $\lambda = 5$ well balances diversity and accuracy, resulting in performance that is closer to the human annotations, and similar to CGAN and GMMCVAE. Using XE loss only, the learned distribution $\hat{p}(c|I)$ has a large variance, which could be very flat and smooth, and thus incorrect words appear during sampling. In contrast, using CIDEr reward can suppress the probability of the words that cannot benefit CIDEr score, and encourage the words that improve CIDEr. Hence, combining the two losses suppresses the poor words and promotes good words (CIDEr), while also preventing the distribution from concentrating to a single point (XE). Figure 7 visualizes the similarity between captions using multi-dimensional scaling (MDS) [3], for different values of $\lambda$. As $\lambda$ increases, some captions are repeated, and points are merged in the MDS visualization.

Finally, using the retrieval reward in the combined loss function also slightly improves the diversity and accuracy, and generally results in a local move in the DA plot. However, a very large $\gamma$ or $\zeta$ could result in a repetition problem, *i.e.*, a model will repeat the distinctive words, since distinctive words are more crucial for the retrieval reward.

## 6. Conclusion

In this paper, we have developed a new metric for evaluating the diversity of a caption set generated for an image. Our diversity measure is based on computing singular values (or eigenvalues) of the kernel matrix composed of CIDEr values between all pairs of captions, which is interpretable as performing LSA on the weighted n-gram feature representation to extract the topic-structure of the captions. Using our diversity metric and CIDEr to re-evaluate recent captioning models, we found that: 1) models that have optimized accuracy tend to have very low diversity, and there is a large gap between model and human performances; 2) balancing the XE loss and other reward functions when using RL is a promising way to generate diverse and accurate captions, which can achieve performance that is on par with generative models (CGAN and GMMCVAE).

**Acknowledgments.** This work is supported by a Strategic Research Grant from City University of Hong Kong (Project NO. 7004682). We are grateful for the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

## References


[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016. 1, 3

[2] Jyoti Aneja, Aditya Deshpande, and Alexander Schwing. Convolutional image captioning. In *CVPR*, 2018. 2

[3] Andreas Buja, Deborah F Swayne, Michael L Littman, Nathaniel Dean, Heike Hofmann, and Lisha Chen. Data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics*, 17(2):444–472, 2008. 8

[4] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. Learning to evaluate image captioning. In *CVPR*, 2018. 3

[5] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *ICCV*, 2017. 2, 3, 5

[6] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990. 4

[7] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *EACL Workshop on Statistical Machine Translation*, 2014. 1, 3

[8] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. In *CVPR*, 2015. 2

[9] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Speaking the same language: Matching machine to human captions by adversarial training. In *ECCV*, 2010. 2

[10] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *CVPR*, 2017. 2

[11] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Ayman Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *CVPR*, 2015. 3

[12] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. Re-evaluating automatic metrics for image captioning. In *EACL*, 2017. 3

[13] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013. 2

[14] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *ICML*, 2015. 3

[15] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale n-grams. In *CoNLL*, 2011. 2

[16] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop*, 2004. 1, 3

[17] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *ICCV*, 2017. 3, 6

[18] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *ECCV*, 2018. 3, 6

[19] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017. 2, 5

[20] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In *CVPR*, 2018. 3, 5, 6, 8

[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 3, 5

[22] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 1, 3

[23] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017. 3, 5, 6, 8

[24] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *EMNLP-Vision and Language Workshop*, 2015. 3

[25] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004. 4

[26] Rakshith Shetty, Marcus Rohrbach, and Lisa Anne Hendricks. Speaking the same language: Matching machine to human captions by adversarial training. In *ICCV*, 2017. 2, 3

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[28] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 1, 3

[29] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 2, 5

[30] Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *NIPS*, 2017. 2, 3, 5

[31] Qingzhong Wang and Antoni B. Chan. Cnn+cnn: Convolutional decoders for image captioning. *arXiv preprint arXiv:1805.09019*, 2018. 2

[32] Qingzhong Wang and Antoni B. Chan. Gated hierarchical attention for image captioning. In *ACCV*, 2018. 2

[33] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel. What value do explicit high level concepts have in vision to language problems? In *CVPR*, 2016. 2

[34] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 2, 5

[35] T. Yao, Y. Pan, Y. Li, Z. Qiu, , and T. Mei. Boosting image captioning with attributes. In *ICCV*, 2017. 2

[36] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016. 2