On Diversity in Image Captioning: Metrics and Methods

Qingzhong Wang, Jia Wan and Antoni B. Chan, *Senior Member, IEEE*

Abstract—Diversity is one of the most important properties in image captioning, as it reflects various expressions of important concepts presented in an image. However, the most popular metrics cannot well evaluate the diversity of multiple captions. In this paper, we first propose a metric to measure the diversity of a set of captions, which is derived from latent semantic analysis (LSA), and then kernelize LSA using CIDEr [1] similarity. Compared with mBLEU [2], our proposed diversity metrics show a relatively strong correlation to human evaluation. We conduct extensive experiments, finding there is a large gap between the performance of the current state-of-the-art models and human annotations considering both diversity and accuracy; the models that aim to generate captions with higher CIDEr scores normally obtain lower diversity scores, which generally learn to describe images using common words. To bridge this "diversity" gap, we consider several methods for training caption models to generate diverse captions. First, we 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. Second, we develop approaches that directly optimize our diversity metric and CIDEr score using reinforcement learning. These proposed approaches using reinforcement learning (RL) can be unified into a self-critical [3] framework with new RL baselines. Third, we combine accuracy and diversity into a single measure using an ensemble matrix, and then maximize the determinant of the ensemble matrix via reinforcement learning to boost diversity and accuracy, which outperforms its counterparts on the oracle test. Finally, inspired by Determinantal Point Processes (DPP), we develop a DPP selection algorithm to select a subset of captions from a large number of candidate captions. The experimental results show that maximizing the determinant of the ensemble matrix outperforms other methods considerably improving diversity and accuracy.

Index Terms—Image captioning, diverse captions, reinforcement learning, policy gradient, adversarial training, diversity metric.

1 Introduction

Recently, the task of image captioning has drawn much attention from researchers in the fields of computer vision and natural language processing, and a wide range of captioning models have been developed [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], the performance of which has even overtaken human performance based on the most popular metrics, such as BLEU [26], METEOR [27], ROUGEL [28], CIDEr [1], and SPICE [29]. However, the above metrics only account for the similarity between human annotations and the generated captions, which reflects the accuracy of the generated captions. Another property, the *diversity* of multiple generated captions, receives less attention from the community of image captioning. Generally, *diversity* refers to the differences among a set of captions from an image, such as expressing different concepts and different sentence structures. Hence, the *diversity* of a set of captions is categorized as follows:

- *Word diversity.* the same concepts can be expressed using different words. Thus, single words are modified (e.g., using synonyms to describe an object) to obtain a different caption.
- *Syntactic diversity.* Generally, the syntax of a sentence represents the organization of words and phrases, which indicates different grammars. For example, applying pre-, post-modifications, clauses, redundant and concise descriptions change the organization of the caption, while not changing the semantics.
- *Semantic diversity.* An image is worth a thousand words. Hence, an image could contain multiple concepts, and different captions could express different concepts that are relatively more important for the particular annotator. For example, in Fig. 1, Caption 3 of the Human annotations uses "lush dry grass" but treats "trees" and "bushes" as unimportant objects. In contrast, Captions 4, 5 describe "trees" and "bushes". Also, the captions generated by Model 1 show multiple concepts, and therefore, the diversity score of the set of captions is relatively high.

Our motivations for accounting for the diversity of multiple captions are threefold. First, an image contains many concepts and different people could be interested in different concepts, which results in diverse descriptions for one image. Therefore, there is diversity among captions due to the diversity among humans. To describe like humans, the automatic captioning methods should reflect this important property—*diversity* of captions. However, diversity receives less attention than accuracy.

Second, focusing only on accuracy could bias a captioning model towards common words and phrases. Generally, a captioning model learns a projection from image space to caption space (see Fig. 2). For a dataset that provides multiple human annotation for an image, the model will converge to the "average" captions that employs common words and phrases. For example, the captions generated

- *Qingzhong Wang, Jia Wan and Antoni B. Chan are with Department of Computer Science, City University of Hong, Kowloon, Hong Kong. E-mail: qingzwang2-c@my.cityu.edu.hk, jiawan1998@gmail.com, abchan@cityu.edu.hk*

Manuscript received xxx, xxx; revised xxx

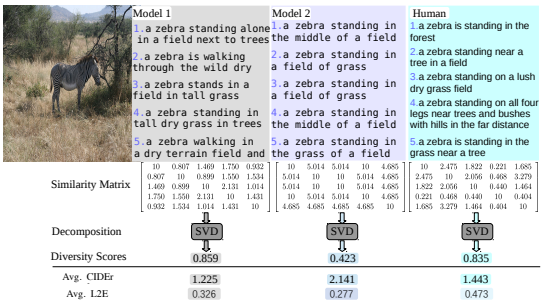

Fig. 1. An overview of our diversity metric. Given a set of captions from a method, we first construct the self-similarity matrix $\mathbf{K}$, consisting of CIDEr [1] scores between all pairs of captions. The diversity score is computed from the singular values of $\mathbf{K}$. A higher diversity score indicates more variety in the set of generated captions, such as changes in the level of descriptive detail and inclusion or removal of objects. The accuracy (average CIDEr and average L2E [32]) of the captions with respect to the human ground-truth is on the bottom. For human annotations, this is the leave-one-out accuracy.

by Model 2 in Fig. 1 obtain a relatively high accuracy score (average CIDEr), but all captions roughly employ the same words and same concepts. In contrast, for captions generated by Model 1, the action of the zebra is described differently as "standing" and "walking". Since this is a static image, "walking" should be a plausible description. Moreover, Model 1 also recognizes the concept of "dry", which also occurs in human annotations, while Model 2 cannot achieve this. Thus, to imitate the ability of humans, a captioning model is supposed to generate diverse captions.

Third, from the viewpoint of machine learning, captioning models should generate multiple captions, since they are typically trained on datasets that provide at least 5 ground-truth captions for one image, such as MSCOCO [30] and Visual Genome [31]. Hence, a trained captioning model should be evaluated on how well the learned conditional distribution of captions given an image approximates the ground-truth distribution. Furthermore, the captioning models that obtain high accuracy score typically employ beam search to generate a caption for an image, which reflects the mode of the learned distribution. Thus, in this case, the accuracy metrics evaluate the differences in the modes of the distributions. However, this ignores another property, the variance of the distribution, which is also a reflection of caption diversity.

Previous works that endeavor to generate diverse captions [2], [33] have measured diversity by analyzing the statistics of words, n-grams and the number of novel captions, which just roughly reflects the diversity of captions, since they treat all captions of the test images as a whole, but ignore the relationship between captions. In addition, it is difficult to define the novelty of a caption. In this paper, we proposed a novel diversity metric that considers the pairwise similarities (e.g. CIDEr similarity) between captions. We first construct a similarity matrix and then use singular value decomposition (SVD) to calculate the diversity score (see Fig. 1). Our proposed diversity metric can be interpreted as latent semantic analysis (LSA), and we further kernelize LSA to enable using any similarity function. Thus we can extract the latent semantics to show the semantic diversity of multiple captions. In addition, examining the topic vectors show the common concept structures between captions, similar to principle component analysis (PCA) [34], [35]. The contributions of this paper are summarized as follows:

1) We proposed a novel diversity metric to evaluate the diversity of a set of captions, and we re-evaluate a wide range of existing image captioning models accounting for both diversity and accuracy. We show that there is a large gap between the performance of the existing models and that of humans. We conduct human evaluation on the diversity of multiple captions, and our proposed diversity metric shows relatively strong correlation to human evaluation.
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. To further improve the performance of a captioning model considering diversity and accuracy, we directly maximize the proposed diversity metric and CIDEr score via reinforcement learning, and we show that our approach employs new baselines to calculate policy gradients during training. The experimental results show that our approaches are superior to CGAN and CVAE.
3) We combine accuracy and diversity into an ensemble matrix, and show that maximizing the determinant of the ensemble matrix via reinforcement learning leads to generating both diverse and accurate captions. Moreover, inspired by determinantal point processes [36], we propose an algorithm to select a subset of captions that have relatively high quality and diversity from a large number of generated captions.
4) We conduct extensive experiments to demonstrate the effectiveness of the diversity metric and the effect of the loss function on diversity and accuracy. In terms of the oracle performance, our proposed method outperforms its counterparts, which obtains 1.696 on CIDEr and 0.309 on SPICE when we sample 20 captions from the learned model.

The organization of the rest of this paper is as follows. In Section 2, we present related works, including metrics and image captioning methods. We propose our diversity metrics in Section 3, and our RL-based methods to generate diverse captions in Section 4. The experimental settings and results are presented in Sections 5 and 6. Finally, we conclude and discuss future directions in Section 7.

2 Related Work

In this section we review image captioning methods and their evaluation metrics. Compared with our preliminary conference version [37], this paper has the following additions: 1) we propose approaches that directly maximize accuracy and diversity rewards using reinforcement learning, which performs relatively well on both diversity and accuracy; 2) we propose a DPP selection method to select a subset of captions that obtains high quality and diversity from a large number of candidate captions; 3) we conduct more experiments to analyze the effects of the hyperparameters and the combinations of different loss functions; 4) we report more evaluation metric scores, such as L2E [32] and word mover distance (WMD) [38], [39].

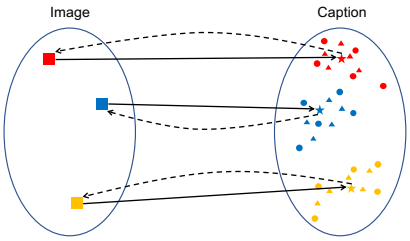

Fig. 2. An illustration of captioning models. A captioning model learns a projection from image space to caption space (solid arrows). Methods focusing only on accuracy will predict "average" captions (represented by ★) that contain common words and concepts among the human annotations (represented by •). In contrast, a diverse captioning model predicts a set of captions (represented by ▲) that span all the concepts present in the human annotations. The dashed arrows represent image retrieval, which can improve the distinctiveness of the generated caption. The color indicates the image-caption correspondence.

2.1 Image Captioning Methods

Image captioning combines computer vision and natural language processing [40]. Generally, there are two stages in a captioning model: (1) concept detection—object recognition, detection, localization, attribute and relationship detection, (2) language generation—translating visual information into sentences. Earlier works typically train the two stages separately. First, concepts are detected using support vector machines [41], conditional random fields [4], [5] or convolutional neural networks (CNNs) [13]. Next, the detected concepts are used to construct sentences using templates [4] or sequence models, such as n-gram models [41]. The quality of the captions generated by these models is highly related to the quality of concept detection, and detecting object relationships and object attributes is difficult.

Recently, encoder-decoder models have become more popular in the field of image captioning, and most of them are trained in an end-to-end manner. The encoder in m-RNN [42] is a CNN [43], [44], [45] and the decoder is a vanilla recurrent neural network (RNN) for modeling sentences – in each time-step, language features and image features are fused to predict the current word. Neural image captioning (NIC) [6] employs a more powerful CNN [46] to extract image features, which are then fed into an LSTM [47] to predict a sequence of words. However, different words should correspond to different image areas, and NIC and m-RNN cannot learn this correspondence. To address this problem, [7] presents a captioning model that uses visual attention, which is able to learn the correspondence between words and image regions, and thus, image captioning models become more interpretable. [17] shows that employing semantics is able to improve the performance, and presents a semantic attention-based captioning model. However, it requires an additional branch to predict semantics and the generated captions could be highly related to the predicted semantics. Also, [10], [11] apply semantics to generate high quality captions. There is an issue in visual attention-based models—some words correspond to image regions while some depend on the context. To address this issue, [48] introduces a sentinel gate to decide whether the image feature or the context feature should be used to predict the current word. Although LSTMs are popular decoders in the encoder-decoder captioning models, language CNNs [49], [50] can alternatively be employed as decoders, and the advantages are (1) faster training [19], and (2) multi-level representations for sentences [18], [22]. [16] applies LSTMs and language CNNs, where language CNNs enhance the long-term dependency of LSTMs.

The above models are typically trained by minimizing cross-entropy loss and using beam search [6] for inference. However, cross-entropy is not directly related to the evaluation metrics used for captioning, such as BLEU [26] and CIDEr [1]. Reinforcement learning (RL) can be used to directly maximize the metric scores [3], [20]. However, RL-based methods could lead to the problem of readability, such as bad endings [51]. To mitigate the issue, [51] employ an n-gram prior to constrain sentence generation, which also reduces the action space of RL, accelerating training. Another drawback of RL-based methods is that the generated captions tend to use common words and phrases, resulting in many images having the same caption, which reduces distinctiveness. Visual-semantic embedding reward encourages the distinctive words given an image, and [52] develops a decision-making framework that uses embedding reward and reinforcement learning. Similarly, [23], [24], [53] present the self-retrieval reward to improve distinctiveness. Alternatively, [54] combines SCST [3] and GAN [55] to train captioning models that are able to describe images that contain objects that do not usually co-occur together, which alleviates the problems of using common words. However, the models trained by minimizing cross-entropy or maximizing the metric scores via RL just learn projections from the image space to caption space, which does not preserve the image-caption structure and the inter-caption structure for one image (see Fig. 2). Although employing self-retrieval reward is able to preserve the image-caption structure, the inter-caption structure is still ignored because the goal of these captioning models is to generate only one caption given an image, which cannot reflect the diversity of human annotations. In contrast, generating diverse captions reflects both the inter-caption structure and the distribution of captions (see Fig. 2).

Some researchers have considered generating diverse captions, developing a large variety of methods to generate both accurate and diverse captions given one image. We categorize these methods into 4 classes: (1) *distribution approximation methods*, such as conditional generative adversarial nets (CGAN) [21], which applies adversarial training [55] to better approximate the conditional distribution; (2) *extra information guidance*, such as conditional variational auto-encoder with Gaussian mixture model prior (GMM-CVAE) [33], where the tags of an image are used to control the diversity. Similarly, [56] also employ detected concepts in the image to guide image captioning and using different concepts could lead to different captions, whereas part-of-speech (POS) guidance [57] employs POS to guide image captioning, which first generates a sequence of POS tags and then uses different words for each POS tag; (3) *inter-caption structure preserving methods*, such as multi-model captioning (GroupTalk) [58], where a weak classifier is applied to discriminate the captions, and captioning with structure relevance and diversity constrains (GroupCap) [59], where similarity matrices are used to represent relevance and diversity; (4) *multi-approach fusion*, such as CGAN with diversity objective [2] and comparative adversarial learning (CAL) [60], which fuse GAN and inter-caption structure

preserving methods.

2.2 Evaluation Metrics

Accuracy evaluation. To evaluate the accuracy of a generated caption, the overlap between one generated caption and human annotations is normally considered in metrics, such as BLEU [26], METEOR [27], ROUGEL [28] and CIDEr [1]. BLEU accounts for the n-gram precision, METEOR considers both precision and recall of uni-grams and applies synonym matching, ROUGEL also considers precision and recall of n-grams, which benefits long texts, and CIDEr applies TF-IDF weighted n-grams, which reduces the score of common n-grams and assigns higher scores to the distinctive words. SPICE [29] is another metric specific to image captioning, which first parses one generated caption and human annotations into scene graphs that are composed of object categories, attributes and relationships [61], [62], and then computes F1-score between the two scene graphs. Although SPICE shows higher correlation to human judgment, it depends on the quality of the scene graph parsing results. Since measuring the overlap between captions could ignore semantic similarity, to better evaluate semantic similarity WMD [39] projects words into a semantic space via word2vec [63], and computes the distance between captions. Inspired by [21], L2E [32] assigns a score to an image-caption pair which is similar to the evaluator in [21]. However, these learning-based metrics could be highly related to the training dataset and corpus.

Diversity evaluation. Vocabulary size is able to roughly reflect diversity [37] and a large vocabulary normally indicates more diverse captions. [2], [33] also employ the percentage of novel sentences that have not been seen in the training set to evaluate diversity, but it is difficult to define the novelty. Another diversity metric is mBLEU [2], [57], [58], which is the average of the BLEU scores between each caption and the remaining captions. However, mBLEU cannot well reflect the semantic diversity and the inter-caption structure, since mBLUE is not able to extract latent semantics and it treats the remaining captions as a whole instead of computing pairwise similarity. Our previous work [37] shows that mBLEU is less correlated to human judgment.

3 Measuring Diversity of Image Captions

In this section, we present our proposed diversity metrics—LSA-based metric and the kernelized metric Self-CIDEr. To evaluate a set of captions $\mathcal{C} = \{c_1, c_2, \cdots, c_m\}$ requires two dimensions: accuracy and diversity. For accuracy, the standard approach is to use the average similarity scores, $acc = \frac{1}{m}\sum_i s_i$, where $s_i = sim(c_i, \mathcal{C}_{GT})$ is the similarity measurement (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 reflects the inter-caption structure, i.e., the structure between captions.

3.1 Latent Semantic Analysis (LSA)

Latent semantic analysis (LSA) [64] is widely applied in information retrieval and topic analysis. As a linear model, LSA first constructs a term-document matrix, each element of which represents the frequency of a term that occurs in a document and then *singular value decomposition* (SVD) is applied to obtain low-dimensional representations of documents in terms of topic vectors. To analyze the diversity of a set of captions via LSA, each caption is represented by a bag-of-word (BoW) vector, which is composed of word frequency, and using SVD we can obtain the latent topics. More latent topics indicate a more diverse set of captions, while only one topic indicates a non-diverse set.

Formally, 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, e.g., $\mathbf{M} = \mathbf{U}\mathbf{S}\mathbf{V}^T$, 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 $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$, 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 and each column of $\mathbf{V}$ represents a caption and its correlation to all latent 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 (Self-CIDEr)

In LSA-based metric, 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 [65] with LSA. The mapping function from the caption space to the feature space 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 into a weighted feature space, $\Phi^n(c) = [\omega_i^n f_i^n(c)]_i$ where the weight ω_i^n is the inverse document frequency for the i-th n-gram. The CIDEr score is the average of the cosine similarities for each n,

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

where

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

In (3), **CIDEr**$_n$ is written as the cosine similarity kernel and the corresponding feature space is spanned by $\Phi^n(c)$. As **CIDEr** is the average over n of **CIDEr**$_n$, 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 $k_{ij} = \mathbf{CIDEr}(c_i, c_j)$. 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.

3.3 Considering Both Accuracy and Diversity

To measure both diversity and accuracy of a set of captions, we can compute an F-measure [66],

$$F_\beta = \frac{(\beta^2+1)\cdot acc \cdot div}{\beta^2 \cdot acc + div}, \tag{4}$$

where β is a weight to balance the accuracy and diversity scores, acc and div respectively. However, the scales of acc and div are different. The accuracy score is the average of the CIDEr scores between each generated caption and the human annotations, ranging from 0 to 10 and the diversity score is $-\log_m(r)$ (see 3.1), ranging from 0 to 1, and it is difficult to select a β.

Inspired by *determinantal point processes (DPPs)* [36], which accounts for the quality and diversity of samples, we present a metric that is able to reflect both accuracy and diversity. Generally, the goal of DPPs is to find a subset $\mathcal{Y} = \{y_1, \cdots, y_m\}$ of items that maximizes the determinant of the matrix $L = [l_{ij}]$ with entries $l_{ij} = q_i \tilde{s}_{ij} q_j$, where q_i represents the "quality" of the ith element in $\mathcal{Y}$ and $\tilde{s}_{ij}$ denotes the similarity between the ith and jth elements. The determinant of L is

$$\det(L) = \det(\tilde{\mathbf{s}}) \prod_{i=1}^{m} q_i^2 \tag{5}$$

where $\tilde{\mathbf{s}} = [\tilde{s}_{ij}]$ is a positive semi-definite matrix. Thus, a set of items that has both high quality and high diversity leads to large value of $\det(L)$.

In this paper, we define the caption quality as the CIDEr score to the ground-truth, $q_i = \mathbf{CIDEr}(c_i, \mathcal{C}_{GT})$, and the similarity as the CIDEr score between two captions, $\tilde{s}_{ij} = \mathbf{CIDEr}(c_i, c_j)$. Thus, L is computed as

$$L = \mathbf{q}\mathbf{q}^T \odot \mathbf{K}, \tag{6}$$

where $\mathbf{q} = [\mathbf{CIDEr}(c_1, \mathcal{C}_{GT}), \cdots, \mathbf{CIDEr}(c_m, \mathcal{C}_{GT})]^T$, $\odot$ represents element-wise multiplication and $\mathbf{K}$ denotes the kernel matrix in Section 3.2.

4 Generating Diverse Captions via Reinforcement Learning

In this section, we present our proposed frameworks for generating diverse captions via reinforcement learning. Our experiments (see Section 6) show that the captions randomly drawn from the model trained by cross-entropy loss obtain a relatively high diversity score but low accuracy score, while the captions randomly drawn from the self-critical model [3] obtain a much higher accuracy score but lower diversity score – there is a large gap between the models trained by cross-entropy loss and CIDEr reward. Thus, it is promising to generate diverse captions via balancing the reward and the cross entropy loss in reinforcement learning. Similar to SCST [3] that directly maximizes CIDEr score, we propose RL-based methods that maximize CIDEr and our proposed Self-CIDEr scores to further improve accuracy and diversity.

4.1 Combining Cross-entropy Loss and CIDEr Reward (XE-CIDEr)

The cross-entropy loss is defined

$$\mathcal{L}_{XE}(c_{gt}, I; \theta) = -\sum_{t=1}^{T} \log p_\theta(w_t^{gt}), \tag{7}$$

where $c_{gt} = (w_1^{gt}, \cdots, w_T^{gt})$ represents the ground-truth caption composed of words $w_{1:T}^{gt}$ for the image I, w_t^{gt} denotes the t-th ground-truth token, and T denotes the caption length. The caption model is represented by p_θ, which is the conditional probability distribution of caption $c = (w_1, \cdots, w_T)$ given the input image I[1], where θ represents its learnable parameters. The gradient of $\mathcal{L}_{XE}$ is

$$\nabla_\theta \mathcal{L}_{XE} = -\sum_{t=1}^{T} \nabla_\theta \log p_\theta(w_t^{gt}), \tag{8}$$

which encourages predicting words that occur in human annotations. Note that as there are multiple captions for an image, the ground-truth conditional distribution could have multiple modes and applying cross-entropy loss to train a model would force the learned distribution to cover all modes[2]. Therefore, if we randomly draw samples from the learned distribution, the variance could be very large.

The loss function in reinforcement learning is the negative expectation of the reward, which is computed as

$$\mathcal{L}_{RL}(c; \theta) = -\mathbb{E}_{c\sim p_\theta}\left[R(c) - b\right], \tag{9}$$

where $R(c)$ is the reward for the caption c and b is a baseline that can be an arbitrary function. If b does not depend on c, e.g., in [3], $b = \mathbf{CIDEr}(c_g, \mathcal{C}_{GT})$, where c_g represents the caption generated via greedy search, it will not change the expected gradient but reduce the variance. We assume that b is not a function of θ.[3] The gradients of $\mathcal{L}_{RL}$ are

$$\nabla_\theta \mathcal{L}_{RL} = -\mathbb{E}_{c\sim p_\theta}\left[(R(c) - b)\nabla_\theta \log p_\theta(c)\right], \tag{10}$$

which encourages captions that obtain rewards higher than b, and suppresses captions that have lower rewards. Using b can reduce the variance of the gradient, making the training process more stable. However, it also significantly reduces the diversity of the captions. For example, self-critical method [3] employs the CIDEr score of c_g as the baseline, which is the elitist strategy, and after training, the samples will converge to the elitist, yielding low diversity.

1. We do not explicitly write the conditional dependence on I to reduce clutter.
2. The quality of the learned distribution is related to the model.
3. A baseline b as a function of θ is still valid.

Our approach uses a weight α to balance L_{XE} and L_{RL}, giving a new loss function

$$\mathcal{L}_{RXE} = \mathcal{L}_{XE} + \alpha \mathcal{L}_{RL}, \tag{11}$$

which has gradient (suppose we draw one sample):

$$\nabla_\theta \mathcal{L}_{RXE} = -\sum_{t=1}^{T} \left(\nabla_\theta \log p_\theta(w_t^{gt}) + \tilde{\alpha} \nabla_\theta \log p_\theta(w_t) \right), \tag{12}$$

where $\tilde{\alpha} = \alpha \cdot (R(c) - b)$. Eq. 12 shows that the probabilities of the words that both occur in human annotations and obtain high reward are improved.[4]

4.2 Maximizing Accuracy and Diversity Rewards

In Section 3, we develop two metrics that are able to measure the diversity of captions. Similar to [3], we can directly maximize the diversity metrics via reinforcement learning. Here we consider 3 methods: 1) we treat the set of captions as a whole and use the joint probability in reinforcement leaning, 2) since our diversity reward function is differentiable w.r.t. the pairwise similarity, we compute the derivatives of the diversity reward w.r.t. the pairwise similarity, which indicates whether the similarity should be enlarged or reduced, resulting in a new method that considers each pair of captions, 3) instead of considering diversity and accuracy separately, we unify them using an ensemble matrix and maximize its determinant, improving both diversity and accuracy [36].

4.2.1 Using the Joint Probability of Captions (JPC)

To measure the diversity, we draw m samples $\mathcal{C} = \{c_1, \cdots, c_m\}$ from $p_\theta(c)$. We use two reward functions, accuracy reward $R_a = \textbf{CIDEr}(c, \mathcal{C}_{GT})$ and diversity reward $R_d = -\frac{\lambda_1}{\sum_{i=1}^{m} \lambda_i}$, to generate both accurate and diverse captions, where λ_i denotes the i-th eigenvalue of the kernel matrix $\mathbf{K}$. The accuracy loss $\mathcal{L}_a$ is computed using (9). Likewise, the diversity loss is

$$\mathcal{L}_d = -\sum_{\mathcal{C}} R_d \cdot p_\theta(\mathcal{C}). \tag{13}$$

Considering that $\{c_1, \cdots, c_m\}$ are *independent and identically distributed (i.i.d)*, the joint distribution $p_\theta(\mathcal{C}) = \prod_{i=1}^{m} p_\theta(c_i)$. Therefore, the holistic loss is computed as (suppose we only sample one set of captions):

$$\mathcal{L}_J = \underbrace{-\sum_{i=1}^{m} (R_a - b) p_\theta(c_i)}_{\mathcal{L}_a} \underbrace{-\eta R_d \prod_{i=1}^{m} p_\theta(c_i)}_{\mathcal{L}_d}, \tag{14}$$

where η is the weight to balance accuracy and diversity.

The gradient of $\mathcal{L}_a$ is calculated using (10), and the gradient of $\mathcal{L}_d$ is computed as

$$\nabla_\theta \mathcal{L}_d = -\eta R_d \sum_{i=1}^{m} \left(\nabla_\theta \log p_\theta(c_i) \cdot p_\theta(\mathcal{C}_{\setminus i}) \cdot p_\theta(c_i) \right), \tag{15}$$

where $\mathcal{C}_{\setminus i}$ represents the set of captions without the i-th caption. Finally,

$$\nabla_\theta \mathcal{L}_J = -\sum_{i=1}^{m} \left(R_a - \underbrace{(b - \eta R_d \cdot p_\theta(\mathcal{C}_{\setminus i}))}_{\text{baseline}} \right) \nabla_\theta \log p_\theta(c_i) p_\theta(c_i), \tag{16}$$

4. The generated and ground-truth captions could have different lengths, and we use padding and masks to make the lengths equal.

where we employ a new baseline, which pushes the sampled captions far away from the caption generated by greedy search. Although $p_\theta(\mathcal{C}_{\setminus i})$ is a function of θ, the value is not that important, because we can adjust the value of η adaptively to be $\eta / p_\theta(\mathcal{C}_{\setminus i})$. Hence, we can simplify $\eta R_d \cdot p_\theta(\mathcal{C}_{\setminus i})$ as ηR_d and the expected gradient becomes

$$\nabla_\theta \mathcal{L}_J = -\mathbb{E}\left[(R_a - (b - \eta R_d)) \nabla_\theta \log p_\theta(c) \right]. \tag{17}$$

4.2.2 Considering Each Pair of Captions (EPC)

Recall the diversity reward $R_d = -\frac{\lambda_1}{\sum_{i=1}^{m} \lambda_i}$ where λ_i denotes the eigenvalue of $\mathbf{K} = [k_{ij}]$, which is differentiable w.r.t. k_{ij}. Generally, we can apply gradient ascent to maximize R_d and the derivatives w.r.t. θ can be computed using the chain rule $\frac{\partial R_d}{\partial k_{ij}} \cdot \frac{\partial k_{ij}}{\partial \theta}$. Note that k_{ij} is *not* differentiable, but we can still apply reinforcement learning to train θ. Thus we introduce a new reward function to maximize R_d.

Reconsidering the derivative $\frac{\partial R_d}{\partial k_{ij}}$, if $\frac{\partial R_d}{\partial k_{ij}} > 0$, we should enlarge k_{ij} and if $\frac{\partial R_d}{\partial k_{ij}} < 0$, we should reduce k_{ij} to maximize R_d. The sign of the derivative of R_d w.r.t. k_{ij} indicates whether we should enlarge or reduce k_{ij} to maximize R_d. Thus, we define the new reward as $\textbf{sign}\left(\frac{\partial R_d}{\partial k_{ij}}\right) k_{ij}$, where $\textbf{sign}(x) = 1$, if $x > 0$, and $\textbf{sign}(x) = -1$, if $x < 0$. Given a set of captions $\mathcal{C} = \{c_1, \cdots, c_m\}$ the loss function is

$$\mathcal{L}_d = -\sum_{i=1}^{m} \sum_{j=1}^{m} \textbf{sign}\left(\frac{\partial R_d}{\partial k_{ij}}\right) k_{ij} \cdot p_\theta(c_i, c_j), \tag{18}$$

where k_{ij} is the similarity between c_i and c_j, such as $\textbf{CIDEr}(c_i, c_j)$ and $p_\theta(c_i, c_j) = p_\theta(c_i) p_\theta(c_j)$. Eq. 18 shows that our new loss function considers each pair of captions in $\mathcal{C}$, which contains more details of the inter-caption structure than (13), since (13) treats the captions as a whole. The policy gradient can be computed as

$$\begin{aligned} \nabla_\theta \mathcal{L}_d = &-\sum_{i=1}^{m} \sum_{j=1}^{m} \textbf{sign}\left(\frac{\partial R_d}{\partial k_{ij}}\right) k_{ij} \nabla_\theta \log p_\theta(c_i) p_\theta(c_i, c_j) \\ &-\sum_{i=1}^{m} \sum_{j=1}^{m} \textbf{sign}\left(\frac{\partial R_d}{\partial k_{ij}}\right) k_{ij} \nabla_\theta \log p_\theta(c_j) p_\theta(c_i, c_j), \end{aligned} \tag{19}$$

since $k_{ij} = k_{ji}$ and $\frac{\partial R_d}{\partial k_{ij}} = \frac{\partial R_d}{\partial k_{ji}}$, (19) can be rewritten as

$$\nabla_\theta \mathcal{L}_d = -2 \sum_{i=1}^{m} \nabla_\theta \log p_\theta(c_i) p_\theta(c_i) \underbrace{\sum_{j=1}^{m} \textbf{sign}\left(\frac{\partial R_d}{\partial k_{ij}}\right) k_{ij} p_\theta(c_j)}_{\mathbb{E}\left[\textbf{sign}\left(\frac{\partial R_d}{\partial k_{ij}}\right) k_{ij}\right]}. \tag{20}$$

Similar to (16), which unifies the policy gradients of accuracy loss and diversity loss, we can fuse the policy gradient of our new diversity loss function with that of the accuracy loss function,

$$\begin{aligned} \nabla_\theta \mathcal{L}_J &= -\sum_{i=1}^{m} (R_a - b) \nabla_\theta \log p(c_i) \cdot p_\theta(c_i) \\ &\quad - 2\gamma \sum_{i=1}^{m} \nabla_\theta \log p_\theta(c_i) p_\theta(c_i) \mathbb{E}\left[\textbf{sign}\left(\frac{\partial R_d}{\partial k_{ij}}\right) k_{ij}\right] \\ &= -\sum_{i=1}^{m} (R_a - B) \nabla_\theta \log p(c_i) \cdot p_\theta(c_i), \end{aligned} \tag{21}$$

where γ is a hyperparameter to balance the diversity and accuracy rewards, and $B = b - 2\gamma \cdot \mathbb{E}\left[\textbf{sign}\left(\frac{\partial R_d}{\partial k_{ij}}\right) k_{ij}\right]$ is

a new baseline. Compared with the baseline in (17), B considers the similarity between one caption and the remaining captions in $\mathcal{C}$ instead of treating $\mathcal{C}$ as a whole. Thus, it encourages a caption that is similar to human annotations and different with the other sampled captions.

Alternatively, to consider each pair of captions in $\mathcal{C}$, we first define mCIDEr as the average of the CIDEr scores between each caption and the remaining captions $\mathbf{mCIDEr}(c_i) = \frac{1}{m}\sum_{j=1}^{m}\mathbf{CIDEr}(c_i, c_j)$, and then the diversity loss $\mathcal{L}_d$ becomes

$$\mathcal{L}_d = \sum_{i=1}^{m} \mathbf{mCIDEr}(c_i) \cdot p_\theta(c_i). \tag{22}$$

Using this new diversity loss, the gradient of $\mathcal{L}_J$ is written as (21), where the baseline is now $B = b + 2\gamma \cdot \frac{1}{m}\sum_{j=1}^{m}\mathbf{CIDEr}(c_i, c_j)$, which is similar to using the loss in (18)[5], except that the weight in mCIDEr is always 1. In contrast, in (18), the weight could be $\{1, 0, -1\}$, depending on $\frac{\partial R_d}{\partial k_{ij}}$, which is computed as [67]

$$\begin{aligned}\frac{\partial R_d}{\partial k_{ij}} &= -\frac{1}{(\sum_{l=1}^{m}\lambda_l)^2}\left(\frac{\partial \lambda_1}{\partial k_{ij}}\sum_{l=1}^{m}\lambda_l - \lambda_1\sum_{l=1}^{m}\frac{\partial \lambda_l}{\partial k_{ij}}\right) \\ &= -\frac{1}{(\sum_{l=1}^{m}\lambda_l)^2}\left(u_{i1}u_{j1}\sum_{l=1}^{m}\lambda_l - \lambda_1\sum_{l=1}^{m}u_{il}u_{jl}\right),\end{aligned} \tag{23}$$

where $\mathbf{K} = \mathbf{U}\Lambda\mathbf{U}^T$, $\mathbf{U} = [u_{ij}]$, $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_m)$ and $\mathbf{U}\mathbf{U}^T = \mathbf{I}$. Hence, $\sum_{l=1}^{m} u_{il}u_{jl} = 1$, if $i = j$, otherwise, $\sum_{l=1}^{m} u_{il}u_{jl} = 0$. We conduct experiments to compare these two approaches (see Section 6), and they obtain similar performance, but note that the reward functions are different.

4.2.3 Unifying Diversity and Accuracy (UDA)

In Sections 4.2.1 and 4.2.2, the accuracy and diversity rewards are considered separately, resulting in new baselines for policy gradient. In this section, we unify diversity and accuracy via maximizing the determinant of $L = [l_{ij}]$ in (6).

Similar to Section 4.2.2, we first compute the derivative of $\det(L)$ w.r.t. l_{ij}, thus

$$\frac{\partial \det(L)}{\partial l_{ij}} = \hat{l}_{ij}, \tag{24}$$

where $l_{ij} = \underbrace{\mathbf{CIDEr}(c_i, \mathcal{C}_{GT})\mathbf{CIDEr}(c_j, \mathcal{C}_{GT})}_{\text{quality}} \cdot \underbrace{\mathbf{CIDEr}(c_i, c_j)}_{\text{similarity}}$ and $\hat{L} = [\hat{l}_{ij}] = L^{-1}$.[6] Then we define a new loss function as follows:

$$\mathcal{L}_J = -\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{sign}(\hat{l}_{ij})l_{ij}p_\theta(c_i)p_\theta(c_j). \tag{25}$$

The gradient can be computed similar to (20),

$$\begin{aligned}&\nabla_\theta \mathcal{L}_J = \\ &-2\sum_{i=1}^{m}\nabla_\theta \log(p_\theta(c_i))p_\theta(c_i)\underbrace{\sum_{j=1}^{m}\mathbf{sign}(\hat{l}_{ij})l_{ij}p_\theta(c_j)}_{\mathbb{E}[\mathbf{sign}(\hat{l}_{ij})l_{ij}]}.\end{aligned} \tag{26}$$

5. Using (18), the second term of B is $\mathbb{E}\left[\mathbf{sign}\left(\frac{\partial R_d}{\partial k_{ij}}\right)k_{ij}\right] = \frac{1}{m}\sum_{j=1}^{m}\mathbf{sign}\left(\frac{\partial R_d}{\partial k_{ij}}\right)k_{ij}$ and k_{ij} could be $\mathbf{CIDEr}(c_i, c_j)$.

6. Adding a small constant ϵI to L ensures invertability.

Algorithm 1 DPP Selection

Input: $\mathbf{C} = \{c_1, \cdots, c_N\}$, and m, where $N \gg m$
Output: $\mathcal{C}$ {subset composed of m captions}
$\mathcal{C} = \emptyset$ {initialization}
for each $i \in 1 : m$ **do**
  Compute L
  {if $i == 1$, c^* is the the caption with the highest quality score}
  $c^* = \arg\max_{c \in \mathbf{C}} \log\det(L(\mathcal{C} \cup c))$
  $\mathcal{C} = \mathcal{C} \cup c^*$
  $\mathbf{C} = \mathbf{C}_{\setminus c^*}$ {remove c^* from $\mathbf{C}$}
end for

To further improve the quality of the generated captions, we employed the linear combination of CIDEr and retrieval rewards [23] and the quality function is as follows:

$$q_i = \mathbf{CIDEr}(c_i, \mathcal{C}_{GT}) + \zeta \cdot \mathbf{RETr}(c_i, I), \tag{27}$$

where ζ is a hyperparameter and $\mathbf{RETr}(c_i, I)$ represents the retrieval score. Hence, $l_{ij} = q_i q_j \cdot \mathbf{CIDEr}(c_i, c_j)$. Although there is no baseline in the proposed UDA model, the variance of the loss is relatively small during training, since the reward in (26) is estimated using multiple samples, and thus the training is stable.

4.3 DPP Selection

In Section 4.2, we presented our RL approaches to generate diverse captions. Typically, the captions are randomly sampled from the learned conditional distribution $\hat{p}(c|I)$ and to further improve diversity-accuracy performance, we employ DPP selection [36], [68] (see Alg. 1) to select a subset of captions from a large number of random samples. First, given an image I, N captions are randomly drawn from the learned distribution $\hat{p}(c|I)$, then we use Alg. 1 to select one caption in each iteration to enlarge the determinant of L, which is a greedy search algorithm.

Note that to compute L, the quality of each caption in $\mathbf{C}$ should be given. If we apply $\mathbf{CIDEr}(c_i, \mathcal{C}_{GT})$ as the quality score, then the ground-truth captions $\mathcal{C}_{GT}$ are required, however, in most cases $\mathcal{C}_{GT}$ is inaccessible. Hence, we use L2E metric [32], which is similar to retrieval models but trained using adversarial samples. In this paper, **DPP-CIDEr** and **DPP-L2E** denote DPP selection methods that use CIDEr and L2E as quality functions. As **DPP-CIDEr** uses the ground-truth captions for the quality metric at test time, it can be considered as an upper-bound performance of DPP.

5 Experiment Setup

In this section, we present the settings for the experiments, including the dataset, captioning models and metrics.

5.1 Dataset

We conduct our experiments on MSCOCO [30], which contains 122,218 training and validation images, with at least 5 human annotations for each image. We split the dataset as [69]—5,000 images for validation, 5,000 images for testing and the remaining images for training. The words that occur less than 6 times in the training split are ignored, resulting in a vocabulary composed of 9,489 words. During inference, we set the maximum length of the generated caption to 16.

5.2 Diverse Captioning

Since most of the existing models are used to generate one caption for an image, we first extend these models to generating multiple captions for one image. Four approaches are adopted in this paper. (1) **Random Sampling (RS):** we randomly draw a sample from the learned distribution $\hat{p}(c|I)$. (2) **Randomly Cropped Images (RCI):** first the given images are resized to 256×256 and then we randomly crop 224×224 patches to generate different captions using beam search algorithm. (3) **Gaussian Noise Corruption (GNC):** a given image is resized to 224×224, after that Gaussian noise with different standard deviations is added to the image to predict captions. (4) **Synonym Switch (SS):** RCI and GNC try to generate different captions by changing the input image, while SS directly manipulates the generated caption. We first train a word2vec model[7] [63] using the texts in the training split. For each word in a given caption, we retrieve its top-10 synonyms using the trained word2vec model and assign a weight to each synonym based on the similarity scores. After that, with probability p, each word is randomly replaced by one of its 10 synonyms, where the synonyms are sampled according to their weights.

We also employ conditional variational auto-encoders (CVAEs), conditional generative adversarial nets (CGNs) and their variants to generate diverse captions. For these models that use Gaussian noise to control the diversity among the generated captions, we draw **Different Random Vectors (DRV)** from Gaussian distributions with different standard deviations to generate captions.

In terms of our proposed approaches and other reinforcement learning based models, we use **Random Sampling (RS)** to generate multiple captions. For DPP selection, we first generate 100 captions using the diverse-captioning models and then we respectively employ **DPP-CIDEr** and **DPP-L2E** to select 10 captions for evaluation.

5.3 Caption Models

We re-evaluate the following existing captioning models: (1) NIC [6] with VGG16 [44]; (2) SoftAtt [7] with VGG16; (3) AdapAtt [48] with VGG16; (4) Att2in [3] with cross-entropy loss (XE) loss and CIDEr reward, denoted as Att2in(XE) and Att2in(C); (5) FC [3] with cross-entropy loss (XE) and CIDEr reward, denoted as FC(XE) and FC(C); (6) Att2in and FC with retrieval reward [23], demoted 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 [33]; (8) CGAN [21].

The models are trained using the training split mentioned in Section 5.1. For models (1)-(7), we randomly sample 10 captions from the trained models, and for (7) and (8), the random noise vectors are drawn from Gaussian distributions with standard deviations $\{1.0, 2.0, \cdots, 10.0\}$ and beam search is used to generate captions. The standard deviations of Gaussian noise for **GNC** are also $\{1.0, 2.0, \cdots, 10.0\}$. For **SS**, we first generate one caption using beam search with beam-width 3 and then generated the other 9 captions by switching words with probabilities $p \in \{0.1, 0.15, \cdots, 0.5\}$. Models and diversity generators are denoted as "model-generator", e.g., "NIC-RS".

7. In this paper, we use continuous bag-of-word model.

Our proposed approaches use reinforcement learning with different reward functions: (1) cross entropy and CIDER, denoted as $\mathbf{XE} + \alpha \cdot \mathbf{CIDEr}$ (Section 4.1); (2) CIDEr and joint probability of captions, denoted as $\mathbf{CIDEr} + \eta \cdot \mathbf{JPC}$ (Section 4.2.1); (3) CIDEr and caption pairs, denoted as $\mathbf{CIDEr} + \gamma \cdot \mathbf{EPC}(m)$ and $\mathbf{CIDEr} + \gamma_1 \cdot \mathbf{mCIDEr}(m)$ (Section 4.2.2). During training we sample $m \in \{2, 5, 8\}$ captions to calculate diversity; (4) unified diversity and accuracy, denoted as **UDA**-m-ζ (Section 4.2.3), where $m \in \{2, 5, 8\}$ and $\zeta \in \{1, 3, 5, 10\}$. Note that the proposed UDA model can only use CIDEr as the quality function, denoted as **UDA**-m and $m \in \{2, 3, \cdots, 8\}$. We consider different values of the hyperparameters to balance the loss functions, $\alpha \in \{5, 10, 20\}$, $\eta \in \{1, 2, 3, 5\}$, $\gamma \in \{0.02, 0.03, \cdots, 0.07\}$ and $\gamma_1 \in \{0.04, 0.06, \cdots, 0.14\}$.

We also investigate different combinations of the loss functions, such as cross-entropy loss and retrieval reward, denoted as $\mathbf{XE} + \zeta_1 \cdot \mathbf{RETr}$, where $\zeta_1 \in \{10, 20, \cdots, 50\}$, CIDEr reward and retrieval reward, denoted as $\mathbf{CIDEr} + \zeta_2 \cdot \mathbf{RETr}$, where $\zeta_2 \in \{1, 3, 5, 10\}$, and cross-entropy loss, CIDEr and retrieval rewards, denoted as $\mathbf{XE} + \alpha_1 \cdot \mathbf{CIDEr} + \zeta_3 \cdot \mathbf{RETr}$, where $\alpha_1 \in \{5, 10\}$ and $\zeta_3 \in \{10, 20, \cdots, 50\}$. The caption model that we use is ATTN proposed by [23].

The models are first trained using cross-entropy loss for 100 epochs with the batch size of 128, and then we apply reinforcement learning to train the model for another 100 epochs. We employ Adam optimizer [70] to update the learnable parameters and the initial learning rate is 0.0004, which decays every 15 epochs with the rate of 0.8. To accelerate the training process, we use Bottom-up features [14], where each image contains 10-100 objects.

5.4 Evaluating Diversity

To evaluate the diversity, we generate 10 captions using the above trained models. For DPP selection, we first randomly sample $N = 100$ captions, and then select 10 captions. We also show the results that use CIDEr as the quality function to select captions, which can be treated as the oracle performance in the diversity-accuracy space. The accuracy of the generated captions $\mathcal{C}$ is the average CIDEr score, thus $acc = \frac{1}{m}\sum_{i=1}^{m} \mathbf{CIDEr}(c_i, \mathcal{C}_{GT})$, where c_i denotes the i-th caption in $\mathcal{C}$. For human annotations, we compute the leave-one-out accuracy score: $acc = \frac{1}{m}\sum_{i=1}^{m} \mathbf{CIDEr}(g_i, \mathcal{C}_{GT\setminus i})$, where $g_i \in \mathcal{C}_{GT}$ and $\mathcal{C}_{GT\setminus i}$ is the set of human annotation without the ith annotation. To calculate the diversity score, we use our proposed LSA-based and the kernelized (Self-CIDEr) metrics (see Section 3).

For DPP selection (see Section 4.3), L2E [32] is trained using the training data in MSCOCO and the adversarial samples are generated by Softatt [7] using beam search. During training, the model takes an image-caption pair as input[8] and the output is a probability indicates whether the caption is a human annotation of the input image.

For the WMD metric [39], we first train a word2vec [63] model using the human annotations in MSCOCO, where each word is represented by a 300-D vector $\mathbf{v}$. Given a set of captions $\{c_1, \cdots, c_{m_1}\}$ and human annotations $\{\hat{c}_1, \cdots, \hat{c}_{m_2}\}$ for an image, we obtain two dic-

8. Note that in [32] the model also takes a human annotation as a reference input, hence, the input is a triplet—(I, c_{GT}, c), where I, c_{GT} and c represent an image, a human annotation and a generated caption.

TABLE 1
Correlation between diversity metrics and human judgement: (top) overall correlation; (bottom) correlation of per-image rankings of methods. SC and mB denote Self-CIDEr and mBLEU, respectively.

Corr Coef	SC	LSA	mB-mix	mB-1	mB-2	mB-3	mB-4
overall Pearson ρ	**0.616**	0.601	0.585	0.567	0.576	0.581	0.585
overall Spearman ρ	**0.617**	0.602	0.575	0.564	0.575	0.574	0.572
avg. per image Spearman ρ	0.674	**0.678**	0.644	0.633	0.643	0.646	0.639

tionaries $\mathcal{D} = \{(w_1, T_1), \cdots, (w_{n_1}, T_{n_1})\}$ and $\mathcal{D}_{GT} = \{(\hat{w}_1, \hat{T}_1), \cdots, (\hat{w}_{n_2}, \hat{T}_{n_2})\}$, where $T_i = \frac{t_i}{\sum_{j=1}^{n} t_j}$ if a word w_i occurs t_i times in the captions. Then the word mover distance between the two sets of captions is $d = \mathbf{min}\{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \mathbf{T}_{ij}\mathbf{dist}(w_i, \hat{w}_j)\}$, where $\mathbf{T}_{ij}$ is the flow between w_i and w_j and $\mathbf{dist}(w_i, \hat{w}_j) = 1 - \mathbf{cos}(\mathbf{v}_i, \hat{\mathbf{v}}_j)$ is the distance between w_i and w_j. The minimization is over the flow matrix $\mathbf{T}_{ij}$, which can be solved using the simplex algorithm. Finally, the WMD score is e^{-d}, which reflects how the generated captions cover the concepts that occur in human annotations but ignores the order of words. In our implementation, we ignore the words contained in the stop words list[9] in the captions.

6 Results

In this section, we present the experiment results. We first show the correlation between our proposed diversity metrics and human judgment. Then we re-evaluate the existing models using both diversity and accuracy metrics. Finally, we show the performance of our proposed approaches, including diverse-caption generation (randomly sample multiple captions) and single-caption generation (using beam search to generate one caption for an image).

6.1 Correlation to Human Judgment

In human evaluation, we use 100 images, each of which has 9 sets of captions generated by human and 8 models—AdapAtt-SS, AdapAtt-GNC, AdapAtt-RCI, Att2in(XE)-RS, Att2in(C)-RS, Att2in(D5)-RS, Att2in(D10)-RS and CGAN-DRV. At each time we first show a worker our instructions—"diversity refers to different words, phrases, sentence structures, semantics or other factors that impact diversity and the score ranges from 0 to 1" — and then show an image and its 9 sets of captions. A worker is required to read all sets of captions and assign a diversity score to each set. For each image, we employ 3 workers to evaluate the diversity and we use the average score as the final human evaluation score. We conduct human evaluation on Amazon Mechanical Turk (AMT).

Fig. 3 (left, center) shows the correlation plots between our proposed metrics and human evaluation. The overall consistency between the proposed diversity metric and the human judgment is quantified using Pearson's (parametric) and Spearman's rank (non-parametric) correlation coefficients (see Table 1 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 1 bottom). Both Self-CIDEr and LSA-based metrics are largely consistent with human evaluation of diversity, with Self-CIDEr having higher 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 judgment is shown in Fig. 3 (right). mBLEU-mix has lower correlation coefficient with human judgment, compared to LSA and Self-CIDEr (see Table 1). Similar results are obtained when looking at the individual mBLEU-n scores. Self-CIDEr has better overall correlation with human judgment, while the two methods are comparable in terms of per-image consistency of method ranking. In terms of mBLEU-{1,2,3,4,mix}, mBLEU-mix obtain equivalent or marginally higher correlation coefficients, hence, considering n-grams could benefit diversity evaluation.

Finally, 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.

6.2 Re-evaluating the Existing Captioning Models

We next re-evaluate the models accounting for both diversity and accuracy. Fig. 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. Since the experiment in Section 6.1 shows that Self-CIDEr metric has higher overall correlation coefficients (Table 1), 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.

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 [3], [20], [23], [24], but it encourages using common words and phrases, 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 [23], 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

9. https://www.nltk.org/book/ch02.html

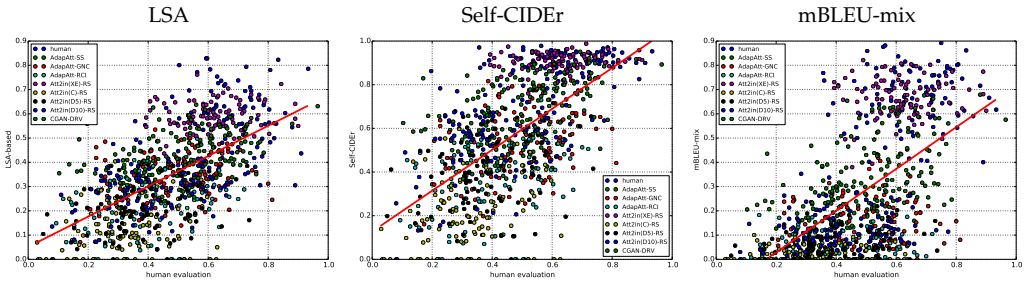

Fig. 3. Correlation plots between the diversity scores of computed metrics and human evaluation. Red lines are the best fit lines to the data.

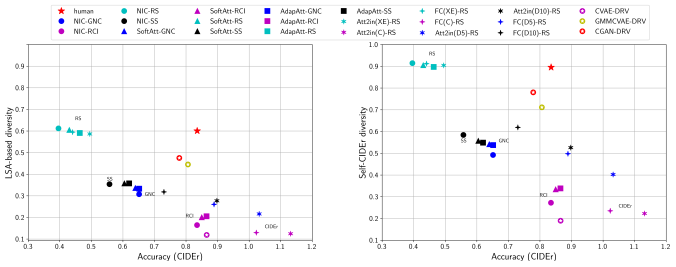

Fig. 4. The performance of different models considering accuracy and diversity. Left: using LSA-based diversity, which employs BoW features. Right: using CIDEr kernelized diversity (Self-CIDEr). The marker shape indicates the caption model, while the marker color indicates the diversity generator or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

Att2in(C). This motivates our method to fill the performance gap by balancing between the two 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.

6.3 The Performance of Our Proposed Methods

We use Self-CIDEr metric to evaluate the diversity of the proposed approaches (see Fig. 5). In addition, we show the oracle performance (upper bound) based on each accuracy metric (see Table 2) and compared with the existing models, our proposed approaches perform much better, obtaining CIDEr(best@20) of 1.696 and CIDEr(best@100) of 1.924. Instead of generating diverse captions for one image, the proposed models are able to generate single caption for one image using beam search, which performs relatively well on MSCOCO test split (see Table 3). Tables 2 and 3 only show the best results based on CIDEr and we show full experimental results in Tables A.2 and A.3 in our supplemental.

6.3.1 *Diversity by Random Sampling*

Fig. 5 (top) shows the performance of our proposed methods on accuracy and Self-CIDEr diversity. Human annotations obtain relatively high diversity and accuracy scores and compared to CGAN, CVAE models, our proposed approaches are relatively promising and efficient to balance diversity and accuracy. The proposed UDA dominates other methods, obtaining higher accuracy scores without

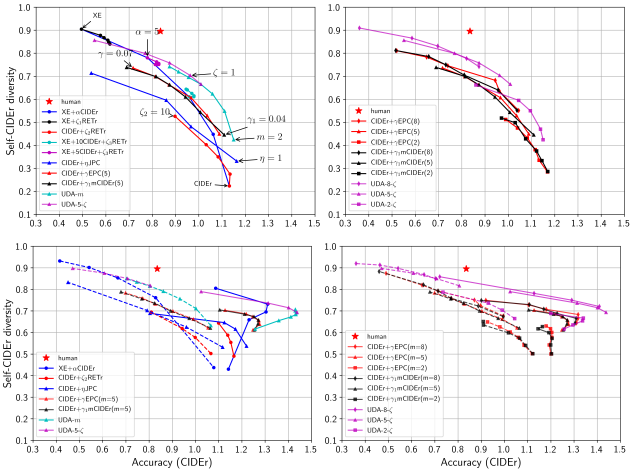

Fig. 5. The performance of our proposed models. Top: using random sampling. Bottom: using DPP selection, where the dashed lines denote using L2E [32] as the quality function (DPP-L2E) and the solid lines represent using CIDEr as the quality function (DPP-CIDEr).

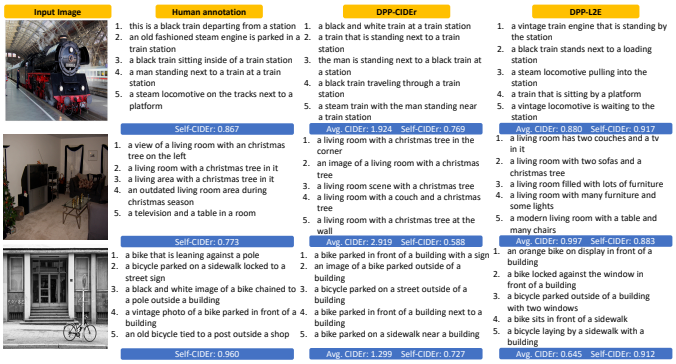

Fig. 6. Qualitative results. For DPP selection, we randomly sample 100 captions from the trained UDA-m-ζ, where $m = 8$ and $\zeta = 1$, then apply DPP to select 5 captions that have both high quality and diversity.

reduction of diversity scores, e.g., UDA-m obtains (accuracy, diversity) of (1.111, 0.548) using $m = 3$, whereas CIDEr+γ_1mCIDEr($m = 5$) obtains (1.007, 0.543) using $\gamma_1 = 0.06$ and CIDEr+γEPC($m = 5$) obtains (1.036, 0.543). The reason is that CIDEr+γEPC only considers the largest eigenvalue λ_1 of $\mathbf{K}$, but ignores the other eigenvalues. Although reducing $\frac{\lambda_1}{\sum_i \lambda_i}$ is able to encourage diversity, it could introduce randomness, since the model reduces the largest eigenvalue, but does not know which eigenvalues should be enlarged. Thus CIDEr+γEPC cannot well preserve the inter-caption structure of human annotations. In contrast, the UDA maximizes the determinant of L, which unifies accuracy and diversity, thus accounting for all eigenvalues of $\mathbf{K}$. Hence, UDA is more effective in balancing diversity and accuracy.

TABLE 2
The oracle (upper bound) and average performance based on each metric. **#samples** denotes the number of samples, **best** denotes the highest score and **avg** denotes the average score. **B** stands for BLEU [26], **M** for METEOR [27], **R** for ROUGEL [28], **C** for CIDEr [1] and **S** for SPICE [29].

Model		#samples	B-4		M		R		C		S		L2E		WMD	Self-CIDEr
			best	avg	best	avg	best	avg	best	avg	best	avg	best	avg		
human		5	-	-	-	-	-	-	-	-	-	-	0.884	0.761	1.0	0.895
CVAE [33]		20	0.312	-	0.244	-	0.541	-	0.910	-	0.176	-	-	-	-	0.193
AG-CVAE [33]		20	0.471	-	0.309	-	0.638	-	1.308	-	0.244	-	-	-	-	-
		100	0.557	-	0.345	-	0.690	-	1.517	-	0.277	-	-	-	-	
GMM-CVAE [33]		20	0.449	-	0.299	-	0.624	-	1.251	-	0.232	-	-	-	-	0.710
		100	0.527	-	0.329	-	0.670	-	1.430	-	0.263	-	-	-	-	
POS [57]		20	0.449	-	0.357	-	0.678	-	1.468	-	0.277	-	-	-	-	-
		100	0.578	-	0.423	-	0.739	-	1.710	-	0.322	-	-	-	-	
SCT [71]		20	0.448	-	0.366	-	0.689	-	1.565	-	0.309	-	-	-	-	-
XE loss		20	0.329	0.045	0.325	0.192	0.621	0.399	1.208	0.509	0.266	0.128	0.895	0.579	0.651	0.904
CIDEr reward		20	0.335	**0.238**	0.323	0.283	0.631	**0.575**	1.385	1.161	0.245	0.206	0.623	0.428	0.618	0.223
XE+αCIDEr	$\alpha = 10$	20	0.464	0.192	0.371	0.264	0.689	0.538	1.568	0.984	0.290	0.187	0.865	0.520	0.650	0.611
CIDEr+ζ_2RETr	$\zeta_2 = 1$	20	0.347	0.240	0.332	0.288	0.639	**0.575**	1.420	1.164	0.260	0.213	0.653	0.426	0.626	0.278
CIDEr+ηJPC	$\eta = 1$	20	0.360	0.225	0.345	0.292	0.640	0.563	1.473	**1.190**	0.272	**0.218**	0.554	0.277	0.634	0.331
CIDEr+γEPC	$m = 5, \gamma = 0.02$	20	0.404	0.218	0.359	0.289	0.659	0.560	1.528	1.134	0.284	0.214	0.645	0.278	0.642	0.449
CIDEr+γ_1mCIDEr	$m = 5, \gamma_1 = 0.04$	20	0.406	0.220	0.360	**0.290**	0.661	0.560	1.531	1.149	0.285	0.215	0.657	0.296	0.642	0.445
UDA-m	$m = 5$	20	0.524	0.174	**0.403**	0.269	0.716	0.536	**1.696**	1.034	0.309	0.196	0.868	0.552	0.667	0.665
UDA-m-ζ	$m = 5, \zeta = 1$	20	0.521	0.158	0.400	0.264	0.714	0.527	1.681	0.990	**0.311**	0.194	0.876	0.591	0.672	0.703
	$m = 5, \zeta = 5$	20	0.430	0.097	0.363	0.238	0.670	0.474	1.479	0.792	0.298	0.176	0.896	0.711	0.670	0.799

Looking at CIDEr+γEPC with CIDEr+ηJPC, the performance of CIDEr+γEPC is much better. As we have mentioned in Section 4.2.2, CIDEr+ηJPC treats a set of captions as a whole, ignoring the inter-caption structure, while CIDEr+γEPC considers the pairwise similarity, which is capable of reflecting the inter-caption structure. Interestingly, CIDEr+γ EPC and CIDEr+γ_1mCIDEr show similar performance. As we showed in Section 4.2.2, the gradient of CIDEr+γEPC has the same form as that of CIDEr+γ_1mCIDEr. Normally, the model is first trained using cross-entropy loss, and the captions drawn from this learned distribution are very different. Hence, $\mathbf{K}$ could be a diagonal matrix and $\frac{\partial R_d}{\partial k_{ij}} = 0$ if $i \neq j$, $\frac{\partial R_d}{\partial k_{ij}} < 0$ if $i = j$ (see Eq. 23). In this case the gradient of CIDEr+γEPC could be the same as that of CIDEr+γ_1mCIDEr, and thus, the two approaches have similar performance.

Interestingly, diversity can be improved by increasing m (see Fig. 5). UDA-2 obtains diversity of 0.424, while UDA-8 boosts diversity up to 0.741. Also, CIDEr+γEPC and CIDEr+γ_1mCIDEr show the same trend with the increase of m. However, a large m results in low accuracy, e.g., the average CIDEr score reduces from 1.150 to 0.874 with m increasing from 2 to 8 for UDA. Another drawback of using a large m is the computational complexity. Compared to SCST [3] that only samples one caption during training, the proposed models require m captions to compute diversity, which is around m times slower than SCST in the training phase (e.g., UDA-5 takes 5.3s per batch on a M40 GPU with batch size 128, while SCST takes $\sim$1s per batch). Note that SCST cannot generate diverse captions and the inference times of the proposed models are the same as SCST.

In terms of different combinations of XE, RETr and CIDEr, XE+αCIDEr is more effective at balancing diversity and accuracy, obtaining wider ranges of diversity and accuracy scores, e.g., the diversity score ranges from 0.223 to 0.904 and accuracy score ranges from 0.495 to 1.131. In contrast, RETr plays the role of local search (see the curves of XE+5CIDEr+ζ_3RETr and XE+10CIDEr+ζ_3RETr). The reason is that RETr reward is relatively smaller than XE and CIDEr reward, and thus, XE and CIDEr could dominate the trend of the curve. It is believed that RETr is important for improving the distinctiveness of the generated captions [53], and introducing RETr reward to captioning models is able to improve diversity (see the curves of CIDEr+ζ_2RETr and UDA-m-ζ). However, using a large weight of RETr could result in repetition problems—a caption repeats the distinctive words for several times, yielding less fluency.

6.3.2 Diversity by DPP Selection

To further improve diversity and accuracy, we apply DPP selection (Alg. 1) to select 10 captions from 100 candidates. Looking at Fig. 5 (bottom), both diversity and accuracy could be significantly improved using CIDEr as the quality function in DPP selection (solid lines). Compared to random sampling (left), the (accuracy, diversity) score of the model that only employs CIDEr reward increases from (1.131, 0.223) to (1.142, 0.430) using DPP-CIDEr, which is comparable to UDA-m using $m = 2$. For the model trained by XE loss, the accuracy score of which surges from 0.495 to 1.087, while the diversity score decreases by 0.1.

However, DPP-CIDEr requires human annotations at test time, which is difficult and expensive. Instead of using CIDEr as the quality function, we employ L2E [32] as the quality function to select a subset of caption (dashed lines), yielding a rise in diversity scores and drop in accuracy scores. In particular for the models that have low diversity scores, DPP-L2E is more effective at improving the diversity scores, e.g., the diversity score of the model that train by CIDEr reward soars from 0.223 to 0.437, which nearly doubles. Interestingly, for the models that obtain low diversity scores using random search (see Fig. 5 (top, left)), DPP-CIDEr and DPP-L2E could obtain similar diversity score, e.g., CIDEr+ζ_2 ($\zeta_2 = 1$), CIDEr+ηJPC ($\eta = 1$) and UDA-m ($m = 2$) have the diversity scores of 0.491, 0.537 and 0.616 by using DPP-CIDEr, while using DPP-L2E, the three models obtain the diversity scores of 0.503, 0.532 and 0.632, respectively. The possible reason is that both DPP-CIDEr and DPP-L2E employ Self-CIDEr matrix and most of the captions in a less diverse set could be the same, thus, the quality of captions is less important and DPP selection algorithm pays more attention to the difference among captions. Whereas, for the models that have high diversity scores, DPP-CIDEr and DPP-L2E lead to quite different diversity scores, the reason for which is that in this case, quality plays a more important role during selection. Note that the accuracy score that we use is the average CIDEr, thus DPP-L2E generally results in less accuracy than DPP-CIDEr.

Fig. 6 shows qualitative results. Our DPP-L2E is able to recognize "vintage train engine" (first row), "sofas",

"chairs", "table" (second row), and "windows" (third row), which are correct. However, these concepts do not occur in human annotations, thus the captions obtain low CIDEr (compared to DPP-CIDEr). In contrast, DPP-CIDEr obtains relatively high CIDEr, since it directly uses CIDEr as the quality function. However, DPP-CIDEr prefers common words, e.g., "a black and white train at a train station", achieving the highest CIDEr score (first row), but the important information "old fashioned steam engine" is missing, which is more distinctive for the image. More examples can be found in the appendix.

6.3.3 Comparison with state-of-the-art

Table 2 shows the oracle (best) and average performance based on each metric and higher scores indicate that a model is able to sample more accurate captions. Compared to the most models that focus on generating diverse captions, our proposed methods are capable of obtaining higher scores using 20 and 100 samples. In particular, our proposed UDA-m ($m = 8$) obtains BLEU-4(best@20) of 0.528, 17.9% higher than SCT [71], BLEU-4(best@100) of 0.655, 13.3% higher than POS [57], CIDEr(best@100) of 1.942, 12.5% higher than POS [57], SPICE(best@20) of 0.311, 0.6% higher than SCT [71] and SPICE(best@100) of 0.351, 9.0% higher than POS [57]. Note that POS [57] and SCT [71] use other information, such as part-of-speech tags and the order of image regions to guide the captioning procedure, while UDA-m does not require any external information. Moreover, POS/SCT require beam search during inference, while our methods employ random sampling, which considerably reduces the inference time.

In this experiment we also observe similar trend that is shown in Fig 5(top)—increasing diversity could reduce the average accuracy scores based on BLEU-4, METEOR, ROUGEL, CIDEr and SPICE. For example, for UDA-m, when m ranges from 2 to 8, the diversity score surges from 0.424 to 0.741, while BLEU-4(avg@20), METEOR(avg@20) , ROUGEL(avg@20), CIDEr(avg@20) and SPICE(avg@20) gradually drop by 38.3%, 11.2%, 9.9%, 23.5% and 12.9%, respectively. In contrast, improving diversity could encourage the highest accuracy scores – for UDA-m, BLEU-4(best@20), METEOR(best@20), ROUGEL(best@20), CIDEr(best@20) and SPICE(best@20) increase by 19.5%, 9.8%, 5.6%, 6.9% and 11.1%, respectively, which means that improving diversity could lead to a model that has strong exploration ability, and thus, it is able to find the "best" caption. See Supplemental Table A.2 for more details.

We also report the L2E and WMD scores of human annotations, since both the candidate and reference captions are human annotations – the WMD score is 1.0. Looking at the average L2E scores, human annotations obtain the highest score (0.761) and much better than that obtained by using CIDEr reward (0.428). In terms of our proposed methods, UDA model obtains the highest L2E (0.711) and WMD (0.674) scores. Also, the model trained by cross-entropy loss obtains a relatively high score, L2E(avg@20) of 0.579, while the proposed models CIDEr+ηJPC, CIDEr+γEPC and CIDEr+γ_1mCIDEr obtain much lower L2E scores. The reason is that to train L2E, we treat human annotations as positive samples and the captions generated by Softatt [7] with beam search, using random words and word permutation [32] as negative samples. Thus using CIDEr reward to train a model could bias it to using common words, which is similar to using beam search, and results in lower L2E scores for CIDEr reward models. CIDEr+ηJPC, CIDEr+γEPC and CIDEr+γ_1mCIDEr force the generated captions far away from the greedy search caption by introducing a new baseline into SCST [3] model (see section 4), however, they could use random words, resulting in non-fluency, hence, the captions generated by these models can be easily recognized as non-human annotations. Interestingly, CIDEr+ηJPC, CIDEr+γEPC and CIDEr+γ_1mCIDEr obtain relatively high WMD scores, which means that these models are capable of capturing the concepts occur in human annotations.

TABLE 3
The performance on generating single caption for one image. **bw** represents beam width and **length** denotes the average length of the generated captions. Full results are shown in supplemental Table A.3.

Model		bw	length	B-4	M	R	C	S	L2E	WMD
Adaptive-XE [48]		3	-	0.322	0.266	-	1.085	-	-	-
Updown-XE [14]		5	-	0.362	0.270	0.564	1.135	0.203	-	-
Updown-RL [14]		5	-	0.363	0.277	0.569	1.201	0.214	-	-
DISC-RL [23]		2	-	0.363	0.273	0.571	1.141	0.211	-	-
SCST [3]		-	-	0.333	0.263	0.553	1.114	-	-	-
Att2in-XE [3]		-	-	0.313	0.260	0.543	1.013	-	-	-
Hieratt-XE [72]		3	-	0.362	0.275	0.566	1.148	0.206	-	-
Hieratt-RL [72]		3	-	0.376	0.278	**0.581**	1.217	0.215	-	-
baseline	XE loss	3	9.0	0.364	0.274	0.569	1.117	0.203	0.446	0.602
	CIDEr reward	3	9.0	0.367	0.273	0.577	1.177	0.208	0.428	0.604
XE+αCIDEr	$\alpha = 10$	3	8.9	0.378	0.276	0.580	1.174	0.207	0.447	0.604
CIDEr+ζ_2RETr	$\zeta_2 = 1$	3	9.4	0.368	0.278	0.578	1.185	0.216	0.421	0.610
CIDEr+ηJPC	$\eta = 1$	3	10.0	0.332	**0.286**	0.569	1.219	**0.222**	0.271	**0.616**
CIDEr+γEPC	$m = 5, \gamma = 0.02$	3	10.8	0.335	0.285	0.570	1.182	0.219	0.266	**0.616**
CIDEr+γ_1mCIDEr	$m = 5, \gamma_1 = 0.06$	3	11.4	0.317	0.284	0.561	1.120	0.216	0.240	**0.616**
UDA-m	$m = 2$	3	9.3	0.371	0.279	0.578	**1.223**	0.213	0.428	0.610
	$m = 5$	3	9.5	0.357	0.278	0.568	1.179	0.212	0.494	0.610
UDA-m-ζ	$m = 5, \zeta = 1$	3	9.3	0.358	0.278	0.570	1.169	0.212	0.518	0.609
	$m = 5, \zeta = 10$	3	10.6	0.272	0.263	0.526	0.950	0.199	**0.589**	0.608

Finally, retrieval reward significantly benefits L2E and WMD scores, e.g., UDA-m ($m = 5$) obtains L2E(avg@20) of 0.552, whereas UDA-m-ζ ($m = 5, \zeta = 5$) has L2E(avg@20) of 0.711, 28.8% higher. Although using retrieval reward could improve the relevance of the generated captions, it leads to repetition problems and reduces fluency.

6.3.4 Single Caption Generation

We evaluate the proposed models in the typical way focusing only on accuracy—given one image we generate a caption using beam search, and the results are shown in Table 3. Using beam search to obtain the top-K captions based on the probability is able to reflect how well a trained model can capture the modes of the ground-truth distribution.

Compared to the state-of-the-art models, our proposed approaches obtain comparable or better results in terms of the most popular metrics, such as BLEU and CIDEr, e.g., UDA-m ($m = 2$) has CIDEr of 1.223, while Hieratt-RL [72] obtains CIDEr of 1.217, which is slight worse.

Looking at the models that are able to generate diverse captions, as diversity increases, BLEU, METEOR, ROUGEL, CIDEr and SPICE scores typically decrease. For example, CIDEr+ζ_2RETr (ζ_2=1) obtains CIDEr of 1.185, while CIDEr+ζ_2RETr (ζ_2=10) has CIDEr of 0.989. In contrast, L2E and WMD scores show different trends, and they are lower than the corresponding average L2E scores and WMD scores shown in Table 2. The reason is that we regard the captions generated by beam search as negative samples to train L2E. Moreover, encouraging diversity could benefit the length of captions and a caption contains more words could provide more information. Employing retrieval reward significantly improves L2E score, e.g., UDA-m ($m = 5$) has L2E of 0.494, whereas UDA-m-ζ ($m = 5, \zeta = 10$) obtains L2E of 0.589.

7 Conclusions

In this paper, we modelled the diversity of human annotations via considering the similarity between each pair of annotations, and presented a diversity metric derived from latent semantic analysis and then kernelized it using CIDEr, which are correlated to human judgment of diversity. We re-evaluated the existing captioning models and found that the models that focus on accuracy generally use common words and obtain low diversity. To improve the diversity of generated captions, we proposed a variety of methods based on reinforcement learning with different reward functions. Extensive experiments were conducted, showing that the proposed methods are effective at balancing diversity and accuracy. In particular, our proposed UDA significantly improves the state-of-the-art oracle performance, and also outperforms the other proposed methods in this paper. UDA maximizes the determinant of the ensemble matrix, which accounts for all eigenvalues of $\mathbf{K}$, whereas the other proposed methods only consider the largest eigenvalue.

Although we have proposed metrics and methods for diverse image captioning, the following directions could be considered in the future. First, the existing metrics can be improved, as the overlap-based metrics, such as BLEU and CIDEr, cannot reflect semantic relevance, while WMD that employs word2vec cannot reflect fluency, and L2E is highly related to the dataset and data augmentations. To well evaluate a captioning model, relevance, fluency, diversity and descriptiveness should be considered. Second, note that the proposed UDA employs unlearnable quality and similarity functions – improvements could be obtained by extending UDA by parameterizing the quality and similarity functions. Third, the proposed methods can be extended to other text generation tasks, such as dialogue and machine translation, providing more choices to the users. Fourth, the existing dataset could be limited on diversity, since the annotations are normally composed of common words. Generating detailed captions that contain more interesting concepts could be an interesting direction for future work.

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] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015.
[2] R. Shetty, M. Rohrbach, and L. A. Hendricks, "Speaking the same language: Matching machine to human captions by adversarial training," in *ICCV*, 2017.
[3] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *CVPR*, 2017.
[4] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *ECCV*, 2010.
[5] 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," *TPAMI*, vol. 35, no. 12, pp. 2891–2903, 2013.
[6] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015.
[7] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.
[8] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov, "Review networks for caption generation," in *Proc. Neural Information Processing Systems*, 2016.
[9] Y. H. Tan and C. S. Chan, "phi-lstm: a phrase-based hierarchical lstm model for image captioning," in *ACCV*, 2016.
[10] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng, "Semantic compositional networks for visual captioning," in *CVPR*, 2017.
[11] T. Yao, Y. Pan, Y. Li, Z. Qiu, , and T. Mei, "Boosting image captioning with attributes," in *ICCV*, 2017.
[12] Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *CVPR*, 2016.
[13] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, C. Lawrence Zitnick, and G. Zweig, "From captions to visual concepts and back," in *CVPR*, 2015.
[14] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and vqa," in *CVPR*, 2018.
[15] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of attention for image captioning," in *ICCV*, 2017.
[16] J. Gu, G. Wang, J. Cai, and T. Chen, "An empirical study of language cnn for image captioning," in *ICCV*, 2017.
[17] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *CVPR*, 2016.
[18] J. Aneja, A. Deshpande, and A. Schwing, "Convolutional image captioning," in *CVPR*, 2018.
[19] Q. Wang and A. B. Chan, "Cnn+cnn: Convolutional decoders for image captioning," *arXiv*, 2018.
[20] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," in *ICCV*, 2017.
[21] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional gan," in *ICCV*, 2017.
[22] Q. Wang and A. B. Chan, "Gated hierarchical attention for image captioning," in *ACCV*, 2018.
[23] R. Luo, B. Price, S. Cohen, and G. Shakhnarovich, "Discriminability objective for training descriptive captions," in *CVPR*, 2018.
[24] X. Liu, H. Li, J. Shao, D. Chen, and X. Wang, "Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data," in *ECCV*, 2018.
[25] C. C. Park, B. Kim, and G. Kim, "Towards personalized image captioning via multimodal memory networks," *TPAMI*, vol. 41, no. 4, pp. 999–1012, 2019.
[26] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002.
[27] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *EACL Workshop*, 2014.
[28] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *ACL Workshop*, 2004.
[29] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *ECCV*, 2016.
[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.
[31] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, vol. 123, no. 1, pp. 32–73, 2017.
[32] Y. Cui, G. Yang, A. Veit, X. Huang, and S. Belongie, "Learning to evaluate image captioning," in *CVPR*, 2018.
[33] L. Wang, A. Schwing, and S. Lazebnik, "Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space," in *NIPS*, 2017.
[34] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 7, pp. 498–520, 1933.
[35] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation*, vol. 10, no. 5, pp. 1299–1319, 1998.
[36] A. Kulesza and B. Taskar, "Learning determinantal point processes," in *UAI*, 2011.

[37] Q. Wang and A. B. Chan, "Describing like humans: on diversity in image captioning," in *CVPR*, June 2019.
[38] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *ICML*, 2015.
[39] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem, "Re-evaluating automatic metrics for image captioning," in *EACL*, 2017.
[40] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys*, vol. 51, no. 6, pp. 1–36, 2019.
[41] 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.
[42] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," in *ICLR*, 2015.
[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.
[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[46] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.
[47] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[48] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *CVPR*, 2017.
[49] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *ICML*, 2017.
[50] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *ICML*, 2017.
[51] T. Guo, S. Chang, M. Yu, and K. Bai, "Improving reinforcement learning based image captioning with natural language prior," in *EMNLP*, 2018.
[52] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in *CVPR*, 2017.
[53] B. Dai and D. Lin, "Contrastive learning for image captioning," in *NIPS*, 2017.
[54] P. Dognin, I. Melnyk, Y. Mroueh, J. Ross, and T. Sercu, "Adversarial semantic alignment for improved image captions," in *CVPR*, June 2019.
[55] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.
[56] Y. Zheng, Y. Li, and S. Wang, "Intention oriented image captions with guiding objects," in *CVPR*, June 2019.
[57] A. Deshpande, J. Aneja, L. Wang, A. G. Schwing, and D. Forsyth, "Fast, diverse and accurate image captioning guided by part-of-speech," in *CVPR*, June 2019.
[58] Z. Wang, F. Wu, W. Lu, J. Xiao, X. Li, Z. Zhang, and Y. Zhuang, "Diverse image captioning via grouptalk," in *AAAI*. AAAI Press, 2016.
[59] F. Chen, R. Ji, X. Sun, Y. Wu, and J. Su, "Groupcap: Group-based image captioning with structured relevance and diversity constraints," in *CVPR*, June 2018.
[60] D. Li, Q. Huang, X. He, L. Zhang, and M.-T. Sun, "Generating diverse and accurate visual captions by comparative adversarial learning," *arXiv*, 2018.
[61] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *CVPR*, 2015.
[62] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning, "Generating semantically precise scene graphs from textual descriptions for improved image retrieval," in *EMNLP Workshop*, 2015.
[63] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013.
[64] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, pp. 391–407, 1990.
[65] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
[66] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *ECIR*. Springer, 2005.
[67] T. Papadopoulo and M. I. Lourakis, "Estimating the jacobian of the singular value decomposition: Theory and applications," in *ECCV*, 2000.
[68] L. Chen, G. Zhang, and E. Zhou, "Fast greedy map inference for determinantal point process to improve recommendation diversity," in *NeurIPS*, 2018.
[69] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015.
[70] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
[71] M. Cornia, L. Baraldi, and R. Cucchiara, "Show, control and tell: A framework for generating controllable and grounded captions," in *CVPR*, June 2019.
[72] W. Wang, Z. Chen, and H. Hu, "Hierarchical attention network for image captioning," in *AAAI*, 2019.

Qinzhong Wang received the B.Eng. and M.Eng. degrees in control science and engineering from Harbin Engineering University, Harbin, China, in 2013 and 2016. Now he is a Ph.D candidate in the Department of Computer Science, City University of Hong Kong. His research interests include computer vision, natural language processing and generative models.

Jia Wan received the B.Eng. degree in software engineering from Northwestern Polytechnical University, Xi'an, China, and M.Phil. degree from School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China, in 2015 and 2018, respectively. He is currently working towards the Ph.D. degree in Computer Science at the City University of Hong Kong. His research interests include congestion analysis and crowd counting.

Antoni B. Chan received the B.S. and M.Eng. degrees in electrical engineering from Cornell University, Ithaca, NY, in 2000 and 2001, and the Ph.D. degree in electrical and computer engineering from the University of California, San Diego (UCSD), San Diego, in 2008. He is currently an Associate Professor in the Department of Computer Science, City University of Hong Kong. His research interests include computer vision, machine learning, pattern recognition, and music analysis.