Extricating from GroundTruth: An Unpaired Learning Based Evaluation Metric for Image Captioning

Zhong-Qiu Zhao[1,2]
[1]*Key Laboratory of Knowledge Engineering with Big Data*
Hefei University of Technology
[2]*School of Computer Science and Information Engineering*
Hefei University of Technology
Hefei, China
z.zhao@hfut.edu.cn

Yue-Lin Sun[1,2]
[1]*Key Laboratory of Knowledge Engineering with Big Data*
Hefei University of Technology
[2]*School of Computer Science and Information Engineering*
Hefei University of Technology
Hefei, China
sunyl@mail.hfut.edu.cn

Nan-Xun Wang[1,2]
[1]*Key Laboratory of Knowledge Engineering with Big Data*
Hefei University of Technology
[2]*School of Computer Science and Information Engineering*
Hefei University of Technology
Hefei, China
nanxun@mail.hfut.edu.cn

Wei-Dong Tian[1,2]
[1]*Key Laboratory of Knowledge Engineering with Big Data*
Hefei University of Technology
[2]*School of Computer Science and Information Engineering*
Hefei University of Technology
Hefei, China
wdtian@hfut.edu.cn

***Abstract*—Recently, instead of pursuing high performance on classical evaluation metrics, the research focus of image captioning has shifted to generating sentences which are more vivid and stylized than human-written ones. However, there are still no applicable metrics which can judge how close the generated captions are to the human-written ones. In this paper, we propose a novel learning-based evaluation metric, namely Unpaired Image Captioning Evaluation (UICE), which can be trained to distinguish between human-written and generated captions. Unlike existing metrics, our UICE consists of two parts: the semantic alignment module measuring the semantic distance between extracted image features and caption meanings, and the syntactic discriminating module syntactically judging how human-like the candidate caption is. The semantic alignment module is implemented by mapping the image features and the word embedding into a unified tensor space. And the syntactic discriminating module is designed to be learning-based, and thereby can be trained to be stylized by users' own, fed with additional personalized corpus during the training process. Extensive experiments indicate that our metric can correctly judge the grammatical correctness of generated captions and the semantic consistency between captions and corresponding images.**

***Index Terms*—image caption, unpaired learning, evaluation metric**

Fig. 1. Negative examples of traditional image caption metrics. These three examples respectively show the scenes of good sentence getting a low score, sentence that does not match the image getting a high score, and sentence that is neither grammatically nor semantically correct getting a fine score. Meanwhile, we also give out the corresponding scores given by our UICE metric and mark them red, which is obviously more reasonable.

I. Introduction

With the dramatic advancements of neural network [1] and the introduction of attention mechanism [2], automatic image captioning models have already achieved extremely high scores on classical metrics such as BiLingual Evaluation Understudy (BLEU) [3], Recall Oriented Understudy of Gisting Evaluation (ROUGE) [4], METEOR [5], Consensus-based Image Description Evaluation (CIDEr) [6] and Semantic Propositional Image Caption Evaluation (SPICE) [7], meaning that machine generated captions have been enough seman-

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

tically consistent with corresponding images. Recently, the research focus has shifted from further improving accuracy [2], [8] to generating captions with some specific styles [9]–[11]. However, all the metrics mentioned above depend on a finite number of ground-truth, which are stylistically fixed or monotonous so that cannot cover all interesting details in the image. So, it has become a common bottleneck that there are no applicable evaluation metrics to judge stylized captions.

The metrics, such as BLEU, ROUGE, METEOR and CIDEr, mainly measure the n-gram word overlap between generated and reference captions. However, n-gram overlap is neither necessary nor sufficient for two sentences to convey the same meaning [12]. SPICE estimates caption quality by transforming both candidate and reference captions into a graph-based semantic representation and shows better correlation with human judgments. But SPICE is sensitive to the semantic meaning of a caption and tends to ignore its syntactic quality [13]. And SPICE prefers to give high scores to long sentences with repeating clauses [14]. Some counter-examples are given in Figure 1. To remedy the aforementioned defects, Yin *et al.* [13] proposed a learning-based metric that directly discriminates between human-written and machine generated captions and is able to adapt to some pathological cases. For convenience of reference, we call the metric proposed by Yin *et al.* as YICE shorting for Yin's Image Captioning Evaluation. However, the YICE still depends on ground-truth captions, just like all previous metrics, which unavoidably results in little flexibility.

To break this limitation, we propose a metric that discriminates directly between wildly crawled human captions and machine generated captions, extricating from reference captions given in certain datasets, so that it can flexibly judge that how close the candidate caption is to be human-written. We crawl a large-scale image description corpus of more than 2 million natural sentences to facilitate the unpaired image caption evaluating scenario and artificially define several transformation to systematically generate some pathological sentences as negative training samples [15]. The crawled positive descriptions incorporating with the generated negative sentences are used to train a discriminator to distinguish machine-generated captions from human-written ones and to output a human-like score. Considering the correlation between caption and image, a semantic alignment block is designed. We use a CNN architecture [16] to capture high-level image representations and a RNN with LSTM cells to encode captions [17]. By projecting the image and sentence features into a common latent space, we can compute the distance between them as semantical consistent score of the candidate caption. Finally, we fuse these two scores as our Unpaired Image Captioning Evaluation (UICE) metric.

To sum up, our key contributions are as follows:

- We make the first attempt to construct an unpaired learning-based evaluation metric without relying on any image-sentence pairs for image captioning.
- We directly introduce image features into sentence matching to make it more reasonable to judge the semantic consistency between the generated statements and the images.
- For the first time, our UICE separates the evaluation into two modules which judge grammatical correctness and semantic consistency, respectively.
- We conduct comprehensive studies to demonstrate that our UICE metric is almost in agreement with human evaluation.

II. Related Works

A. Captioning Evaluation Metrics

Although human evaluation scores are more reliable, they are too costly to obtain. Thus, most image captioning models still adopt automatic metrics instead of human judgments. In the early times, there were no automatic metrics specifically for image captioning, so some were introduced from other tasks of Natural Language Processing (NLP). The most commonly used two metrics, BLEU [3] and METEOR [5], are originally for statistically evaluating machine translation task. Comparing with the precision-based BLEU, which has been repeatedly questioned [18]–[20], METEOR, which additionally considers the recall ratio, seems a little more promising. Another popular metric ROUGE [4] is originally used in the text summarization community task. As the name suggests, it is primarily recall-based.

With the interest in jointing visual and linguistic problems becoming increasingly considerable, there have been evaluation metrics specifically targeted at the task of image description. CIDEr [6] applies Term Frequency Inverse Document Frequency (TF-IDF) weights to n-grams in the candidate and reference sentences, and measures the sum of their cosine distance across n-grams. SPICE [7] provides a novel measuring method estimating caption quality by transforming both candidate and reference captions into a scene graph which explicitly encodes the main features found in the image captions and abstract most of the lexical and syntactic idiosyncrasies of natural language in the process. To correlate better with human judgments and avoid blind spot of those rule-based metrics, Yin *et al.* [13] propose YICE to adapt to pathological cases once identified, while correlating well with human judgments. Our work differs from all above methods as we separate image relevance from semantic correctness, which allow for machine generated captions extricating from reference captions

B. Unpaired training

As paired image-sentence data is expensive to collect, some researchers tried to build image captioning models which can learn from other available data [21], [22]. Inspired by unsupervised machine translation methods [23]–[25], Feng *et al.* [15] proposed an unsupervised image captioning method relying on a set of images, a set of sentences obtained from an external corpus and an existing visual concept detector. The train of thought that mapping unpaired images and sentences into a common latent space furtherly point us a way to

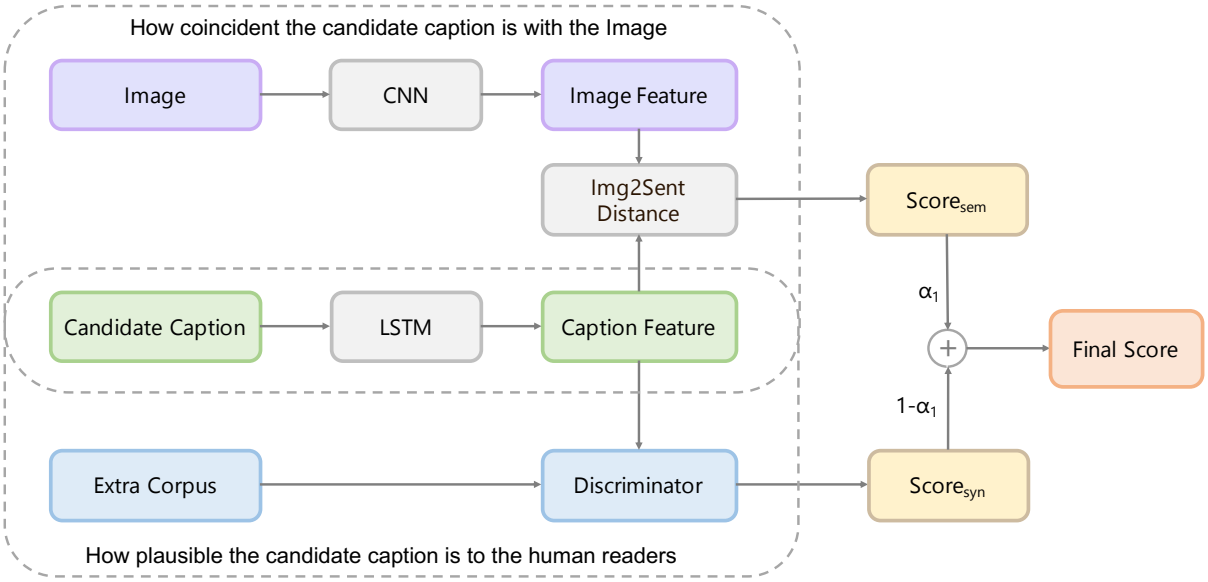

Fig. 2. The model architecture of the proposed learned metric. We use a convolution neural network to encode the reference image into context vector. An LSTM is applied to get the encoding of the candidate caption. We use the distance between the encoded vectors to score how well they match. Additionally, another LSTM is designed as a discriminator to distinguish whether a caption is machine-generated or human-written.

check whether the image and sentence of the same semantic meanings are well aligned.

C. Adversarial evaluation

Generative Adversarial Networks (GANs) [26] is demonstrated to be effective in generating more human-like captions [22], [27]–[29]. Composed of a generator and a discriminator, GANs can be regarded as a process of playing a min-max two-player game: the generator attempt to fool the discriminator, while the discriminator concentrates on minimizing this probability. Different from discriminators provided by GANs, our work focuses on evaluating instead of generating. [30] firstly introduced the evaluation strategy using adversarial classifier to evaluate sentence generation quality. This work impressed series of subsequent researches including ours. [31] preliminarily studied this idea in the context of dialogue generation, proposing to train a pair of GANs alternately. And on that basis, [32] discussed the necessity of assigning scores to partially decoded sequences in avoiding potential pitfalls of adversarial evaluations, while [33] aimed at training a single discriminator on large corpus of dialogue responses generated by different dialogue systems. [13] applied adversarial training to image caption evaluation and additionally explicitly defined several pathological transformations to enrich negative samples. The difference of our work is that we segment the image consistency evaluation into a separate module and judging the syntactic correctness of generated captions with extra corpus instead of ground-truth.

III. Proposed Model

As shown in Figure 2, we design two modules to evaluate semantic consistency and syntactic correctness respectively and an integration operation to fuse them into a final score.

A. Semantic Alignment Module

The semantic alignment score can be expressed as

$$\text{score}_{\text{sem}}(S, I) = P(S \text{ matches image } I \mid \Phi) \quad (1)$$

where Φ denotes the set of parameters.

As shown in Figure 3, there is an image encoder and a sentence encoder contained in this module. We simply choose the 152-layered ResNet convolutional neural network pretrained on ImageNet as the image feature extractor to extract the input image I into a feature representation x, and then embed it into a d-dimensional embedding image vector $f(x) \in \mathbb{R}^d$. Other common encoding networks, such as VGGNet and Inception, are also applicable here.

A long short-term memory (LSTM) is used to encode the candidate caption S into a fixed size embedded s. Before fed into the encoder, each word of the caption is represented as a word embedding vector, which has the same dimension with embedded image vector, $w \in \mathbb{R}^d$. The word embedding process is initialized with GloVe [34]. In this part,following the works in [25], [29], we introduce some object detection features in the COCO dataset into training process as visual concepts, and use the main words of candidate caption when computing the image-to-sentence distance.

The distance between $f(s)$ and $f(x)$ can be computed as

$$K = T_x \cdot f(s) \quad (2)$$

$$d_i(s) = exp(-\|K_{x,i} - K_{s,i}\|_{L_1}) \quad (3)$$

$$dist_x(s, x) = [d_1(s), \ldots, d_m(s)] \in \mathbb{R}^m \quad (4)$$

where T_x is a $d \times n \times m$ dimensional tensor and m is the number of different $d \times n$ distance kernels to use. This distance vector captures how well s matches the image x, and it is

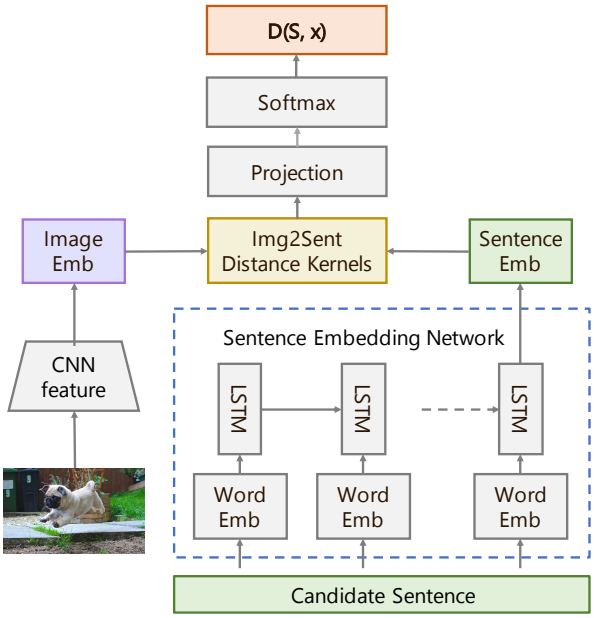

Fig. 3. Semantic Alignment Module. The candidate caption and the reference image are mapped into a common space and the distance between them is used to judge the caption.

multiplied with a output matrix followed by softmax to yield the discriminator output probability, $D(s, x)$. And the loss function of this part can be defined by

$$loss_{sem} = -log(D(s, x)) \tag{5}$$

B. Syntactic Discriminating Module

In this module, the syntactic discriminating score can be expressed as

$$\text{score}_{\text{syn}}(S) = P(S \text{ is human written} \mid \Theta) \tag{6}$$

where Θ represents the set of parameters.

As shown in Figure 4, we train an automatic metric by using extra corpus, which is also implemented as an LSTM, to distinguish generated captions from human-written ones.

$$[\ q_t, h_t^d\ ] = \text{LSTM}^d(x_t, h_{t-1}^d), \quad t \in \{-1, ..., n\} \tag{7}$$

where h_t^d is the hidden state of the LSTM. q_t indicates the probability that the generated partial sentence $S_t = [s_1, s_2, ..., s_t]$ is regarded as real by the discriminator. Similarly, given a real sentence $\hat{S}$ from the corpus, the discriminator outputs $\hat{q}_t, t \in \{1, ..., l\}$, where l is the length of $\hat{S}$. We take the q_n or q_l , the probability outputted at the last moment, as the global result. We set the score value as the logarithm of the probability

$$v_t = log(q_t) \tag{8}$$

So the loss function of this part can be defined as

$$loss_{syn} = -\frac{1}{n}\sum_{t=1}^{n} log(1 - q_t) \tag{9}$$

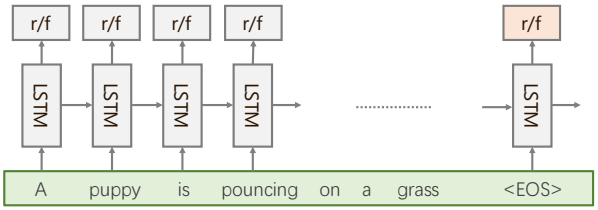

Fig. 4. The architecture of Syntactic Discriminating Module.

Referring to [13], we also set up several pathological cases as negative examples during the training process of syntactic discriminating module. We take two different transformations to generate pathological sentences: word permutation and random word. Word permutation means we will randomly choose several words in a certain sentence of our corpus and permute them:

$$T_{wp}(S, \tau) = \{S' | w \in S, w' \in S(\tau)\} \tag{10}$$

where τ is a hyper-parameter, and $S(\tau)$ represents permuting τ percent of words in the sentence S. Similarly, random word means replace τ percent of words in S with random words from the vocabulary:

$$T_{rw}(S, \tau) = \{S' | w \in S, w' \in V(\tau)\} \tag{11}$$

Note that τ percent of S should contain at least two words. For example, we take τ as 30:

original : *A beautiful lady is looking at the camera and smiling.*

word permute : *A beautiful and is looking at lady camera the smiling.*

random word : *A business lady is hands at the camera and garage.*

As analyzed in that paper, since the number of words machine generated captions contain is limited while those human-written sentences are more likely to use rare words, a discriminator tends to believe that a sequence of random words is written by human and tell those captions apart easily by simply looking at what words are used. To address this problem, in addition to augment the vocabulary, some captions generated using Monte Carlo Sampling, which contains a much higher variety of words, are also included in the corpus.

C. Integration

UICE (sem) evaluates the matching degree between image and candidate caption, but cannot judge whether the caption is grammatically correct, while UICE (syn) is opposite. They are both one-side. So we design a simply neural network with one hidden layer to assign a weight to each of above two scores and fuse them into the final score.

$$score(S) = \alpha_1 \boldsymbol{f}_1(score_{sem}) + 1 - \alpha_1 \boldsymbol{f}_2(score_{syn}) \tag{12}$$

We vector $score_{sem}$ and $score_{syn}$ into the input vector x. We take W as the vector of weights and b for bias. The activation function we choose is Sigmoid function. So, the output of the integration module is

$$y = sigmoid(Wx + b) \tag{13}$$

Here y represents the output score. The loss function of this part is

$$L(\hat{y}, y) = -(ylog\hat{y} + (1-y)log(1-\hat{y})) \tag{14}$$

So, the loss of the whole architecture is

$$\begin{aligned} loss = -log(D(s,x)) - \frac{1}{n}\sum_{t=1}^{n} log(1-q_t) \\ -(ylog\hat{y} + (1-y)log(1-\hat{y})) \end{aligned} \tag{15}$$

The performance of a model is then defined as the averaged performance on all the image-caption pairs in a test or validation set:

$$\text{score}(D_p) = \frac{1}{|D_p|}\sum_{(S,I)\in D_p} score(\Psi, (S,I)) \tag{16}$$

where D_p is the set of all image-caption pairs in a test or validation set, and Ψ represents the metric.

IV. Experiments

A. Experiment setup

a) Datasets: We use the images in the MSCOCO dataset [35] as the image set (excluding the captions). As it was done in previous works [36], we re-split the original COCO dataset into a train set with 113,287 images, a validation and a test set, each contains 5,000 images. For image feature extracting, the convolutional neural network (CNN) pretrained on ImageNet is used as the visual concept detector.

Regarding the syntactic discriminator, we randomly initialize the parameters and train it using two homemade corpus with human-made pathological samples. As for sentence corpuses, we imply two extra datasets to train the discriminator respectively to gain the ability of evaluating stylized captions: one for wild captions, where "wild" can be unserstood as "imperscriptible" or "unofficialand", and another one for romance novels [37]. The wild caption dataset is crawled from Shutterstock[1], an online stock photograph website providing massive royalty-free stock images. Image composers upload images with a description each. To ensure that the crawled sentences are relevant to the images, we directly use the name of eighty object categories in the MSCOCO dataset as the searching keywords, just like it was done in [38]. We finally got 2,322,635 distinct captions, which are used to train the discriminator evaluating normal captions. The romance novel dataset is crawled from SmashWords[2], where we can collect those free books written by yet unpublished authors. We choose books in the *Romance* genre, which contains 2,865 books and we automatically parse each book into sentences and struck down dialogues and sentences that are too short, too long or contain too much (over 20%) low-frequency words. The definition of "low-frequency words" will be shown in *Implementation Details*. Finally, we collect 1,123,200 sentences, which are used to train the discriminator for stylized captions. Finally, for training the integration module, we softmax the human scores in Flickr 8K datasets and use them as "y" in the process of training.

b) Image Captioning Models: For general image captioning task, we select three classic models, "NeuralTalk" [36], "Show and Tell" [1], "Show, Attend and Tell" [2] and Updown [39] to train and test our metric. Additionally, to prove the effectiveness of our model in the task of evaluating stylized image captions, we test our metric on SemStyle [10], style-factual LSTM(SF) [11] and StyleNet [40]. These three stylized models can generate captions of different style, but here we only use the romantic ones.

c) Implementation Details: Descriptions in the corpus crawled by us are tokenized by NLYK toolbox [41]. We count all tokenized words and build a vocabulary with top 15,000 words sorted by appearing frequency. Those abandoned words are replaced with one of the four special tokens, UNKNOW. Removing these low-frequency words will, on the one hand, greatly reduce our vocabulary capacity, speed up model training, and on the other hand, reduce noise. In some previous works, the sizes of vocabulary are commonly set into 10,000, but we find that the syntactic discriminator is more likely to judge sentences containing UNKNOW tokens as human-written, so we expand the vocabulary size to make the evaluation fairer. All words are embedded into 300 dimensional vectors initialized by GloVe [34].

For image feature extraction, we employ ResNet-152 model pretrained on ImageNet [16], after which we use a linear projection to reduce the dimension of image feature to match that of caption feature. For caption feature extraction, we fix the step size of LSTM to be 20, padding shorter sentences with the special token PAD while cutting longer ones to 20 words. The LSTM hidden dimension and the shared latent space dimension are both fixed to 512. Batch size is set to 100 with half positive samples and half negative samples in each. We train our model using Adam optimizer with a learning rate of 10^{-3} for 30 epochs. The decay factor is set to be 0.9. When fixing the architecture of the discriminator before using this learning-based metric, it is inescapable to exist deviation during different times, due to the randomness of parameters initialization in training.

B. Capability and Robustness

To measure the capability of our metric, we compare the evaluating result of five image captioning generating models and human-written captions under some existing automatic metrics and our UICE. Obviously, it is a matter of course for an evaluation metric to accurately tell machine generated captions apart from human-written ones and to give a fair score for the different models. In other words, the distinction between

[1]https://www.shutterstock.com

[2]https://www.smashwords.com

TABLE I
BLEU-1,2,3,4, METEOR, ROUGE, CIDER, SPICE, YICE, UICE SCORES OF NORMAL IMAGE CAPTION GENERATING MODELS. "UICE(SEM)" AND "UICE(SYN)" REPRESENT OUR SEMANTIC ALIGNMENT MODULE AND SYNTACTIC DISCRIMINATING MODULE.

	ShowAndTell	NeuralTalk	ShowAttendAndTell	Up-Down	Human
BLEU-1	60.3	66.3	66.9	79.8	-
BLEU-2	38.0	42.3	43.9	-	-
BLEU-3	25.4	27.7	29.6	-	-
BLEU-4	17.1	18.3	19.9	36.3	21.7
METEOR	16.9	19.5	18.5	27.7	25.2
CIDEr	-	66.0	-	120.1	85.4
SPICE	-	5.1	-	21.4	-
YICE	**7.2**	**6.1**	**9.7**	**-**	**73.6**
UICE(sem)	8.4	12.4	14.6	35.2	72.3
UICE(syn)	8.7	11.9	12.7	34.7	71.3
UICE	**7.4**	**7.7**	**9.8**	**34.3**	**72.8**

TABLE II
BLEU-1,2,3,4, METEOR, ROUGE, CIDER, SPICE, UICE SCORES OF STYLIZED IMAGE CAPTION GENERATING MODELS.

	SF-LSTM(Romantic)	SemStyle(Romantic)	StyleNet(Romantic)	Human
BLEU-1	27.8	38.9	46.1	71.3
BLEU-2	14.4	-	24.8	-
BLEU-3	8.2	-	15.2	-
BLEU-4	4.8	5.7	10.4	19.0
METEOR	11.2	15.6	15.4	22.4
ROUGE	25.5	-	38.0	-
CIDEr	37.5	29.7	31.2	57.3
UICE(sem)	8.9	13.1	19.9	73.6
UICE(syn)	6.1	17.8	25.1	63.3
UICE	**8.7**	**14.7**	**19.8**	**71.0**

the scores of different models and the scores of models and humans is an important criterion to evaluate the quality of an evaluation metric. Table 1 and Table 2 show that UICE can correctly distinguish machine generated captions and human-written ones consistent with the tendency of other metrics on not only normal caption sets but also stylized ones. For all these metrics, the higher the score, the better the model. Figures in these two table represent the percentage. We also display the score of semantic alignment module and syntactic discriminating module respectively in the tables, which are also in line with the trend. Since there are five ground-truth captions for each image in COCO dataset, we use one of them as a candidate caption and the other four as reference to judge it. Repeat this process in turn and pick the average score as the value of Human column. It can be seen that existing metrics are not reasonable enough in evaluating stylized captions, where UICE performs more flexible and fairer.

Due to those pathological cases we set up in Syntactic Discriminating Module, the Robustness of our model is better than some other metrics in theory, which is proven in [13].

C. Caption Level Correlation

At the caption level, following the procedure in SPICE [7], we compute the Kendall's τ correlation between evaluation metrics' scores and human annotations on Flickr 8K dataset [42]. The results are shown in Table III. Figures in Table 3 represents the proportions, which are also the higher the better. It is satisfying that UICE achieves a score almost equal to state-of-the-arts'.

TABLE III
CAPTION-LEVEL KANDALL'S τ CORRELATION BETWEEN EVALUATION METRICS AND HUMAN JUDGMENTS FOR THE 12 COMPETITION ENTRIES PLUS HUMAN CAPTIONS IN THE COCO VALIDATION SET.

	Flickr-8k
BLEU-1	0.32
BLEU-2	0.21
BLEU-3	0.20
BLEU-4	0.14
METEOR	0.42
ROUGE	0.32
CIDEr	0.44
SPICE	0.45
YICE	0.47
UICE	0.47
Inter-human	0.73

D. System Level Correlation

In this section, we compare UICE with some other existing evaluation metrics on the Pearson's ρ correlation with human judgments collected in the 2015 COCO Captioning Challenge and drew a visual image. Agreeing with the reason pointed out by [13] that M3, M4, M5 are not for ranking image caption models, the two items we use are M1: Percentage of captions that are evaluated as better or equal to human captions and M2: Percentage of captions that pass the Turing Test. The results are shown in Table IV. We perform our experiments on the COCO validation set, on which 12 of 15 teams submitted their results collected in the 2015 COCO Captioning Challenge and obtain a fine result. Although it seems a little weaker than YICE, we surpass it in the aspect of evaluating stylized

TABLE IV
SYSTEM-LEVEL PEARSON'S ρ CORRELATION BETWEEN EVALUATION METRICS AND HUMAN JUDGMENTS FOR THE 12 COMPETITION ENTRIES PLUS HUMAN CAPTIONS IN THE COCO VALIDATION SET.

	M1[a]		M2[b]	
	ρ	(p-value)	ρ	(p-value)
BLEU-1	0.124	(0.687)	0.135	(0.660)
BLEU-2	0.037	(0.903)	0.048	(0.877)
BLEU-3	0.004	(0.990)	0.016	(0.959)
BLEU-4	-0.019	(0.951)	-0.005	(0.987)
METEOR	0.606	(0.028)	0.594	(0.032)
CIDEr	0.438	(0.134)	0.440	(0.133)
SPICE	0.759	(0.003)	0.750	(0.003)
YICE	**0.939**	**(0.003)**	**0.949**	**(0.002)**
UICE	**0.937**	**(0.000)**	**0.940**	**(0.000)**

[a]M1: Percentage of captions that are evaluated as better or equal to human captions.

[b]M2: Percentage of captions that pass the Turing Test.

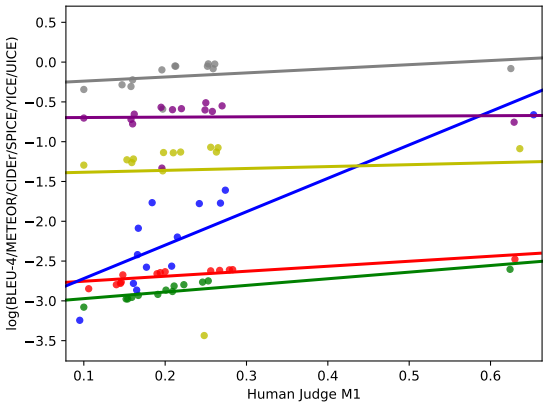

Fig. 5. BLEU-4(purple), METEOR(yellow), CIDEr(grey), SPICE(green), YICE(blue) and UICE(red),vs. human judgments M1 for 12 entries in the 2015 COCO Captioning Challenge on COCO validation set. Each of the data points in the lower left corner represents an entry and those in the higher right corner represent to human-written captions.

captions and we achieved extricating from ground-truth. In Figure 5, we compare the regression of above metrics with human judgments M1 on COCO validation set. It can be seen that UICE is more consistent with human judgments than the other five metrics.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel unpaired learning-based image captioning evaluation metric that can be trained to evaluate image captions from two aspects of semantics and syntactics. The improvement of our UICE over other existing metrics is that it can extricate from ground-truth. In addition, for both natural image captioning datasets and stylized ones, our UICE can give more human-like score than existing metrics over machine-generated captions. Excellent flexibility makes it reasonable for us to believe that UICE could be an effective complement to existing metrics, especially in the case of evaluating wild or personalized captions.

In the future work, one open issue could be training an adversarial image caption generating model using UICE score as the reward of discriminator. Another issue could be building a multi-round metric that can adaptively handle different styles of captions rather than separately training several discriminators for them. As for the improvement of the model itself, we consider using the idea of BERT to obtain a higher ability of grammatical correcting. Finally, it is also expected that UICE can be improved and extended to evaluate image storytelling or adapt to some other related tasks.

ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation of China (Nos.61672203 & 61976079) and Anhui Natural Science Funds for Distinguished Young Scholar (No.170808J08).

REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *IEEE Conference on Computer Vision Pattern Recognition*, 2015.

[2] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *Computer Science*, pp. 2048–2057, 2015.

[3] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Ibm research report bleu: a method for automatic evaluation of machine translation," 06 2019.

[4] C. Flick, "Rouge: A package for automatic evaluation of summaries," in *Workshop on Text Summarization Branches Out*, 2004.

[5] A. Lavie and A. Agarwal, "Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments," 2007.

[6] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Computer Vision Pattern Recognition*, 2015.

[7] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," *Adaptive Behavior*, vol. 11, no. 4, pp. 382–398, 2016.

[8] 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 visual question answering," 2017.

[9] W. Cheng, C. Wu, R. Song, J. Fu, X. Xie, and J. Nie, "Image inspired poetry generation in xiaoice," *CoRR*, vol. abs/1808.03090, 2018.

[10] A. Mathews, L. Xie, and X. He, "Semstyle: Learning to generate stylised image captions using unaligned text," 2018.

[11] T. Chen, Z. Zhang, Q. You, F. Chen, Z. Wang, H. Jin, and J. Luo, ""factual" or "emotional": Stylized image captioning with adaptive learning and attention," 2018.

[12] J. Giménez and L. Màrquez, "Linguistic features for automatic evaluation of heterogenous mt systems," in *Workshop on Statistical Machine Translation*, 2007.

[13] C. Yin, G. Yang, A. Veit, H. Xun, and S. Belongie, "Learning to evaluate image captioning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[14] S. Liu, Z. Zhu, Y. Ning, S. Guadarrama, and K. Murphy, "Improved image captioning via policy gradient optimization of spider," 2017.

[15] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," 11 2018.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[17] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," *Computer Science*, 2015.

[18] C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluating the Role of Bleu in Machine Translation Research," in *Proceedings the Eleventh Conference of the European Chapter of the Association for Computational Linguistics*, (Trento, Italia), pp. 249–256, 2006.

[19] D. Elliott and F. Keller, "Image description using visual dependency representations," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1292–1302, 10 2013.

[20] K. Girish, P. Visruth, O. Vicente, D. Sagnik, L. Siming, C. Yejin, A. C. Berg, and T. L. Berg, "Babytalk: understanding and generating simple image descriptions," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 35, no. 12, pp. 2891–2903, 2013.

[21] P. Anderson, S. Gould, and M. Johnson, "Partially-supervised image captioning," 2018.

[22] T. Chen, Y. Liao, C. Chuang, W. Hsu, J. Fu, and M. Sun, "Show, adapt and tell: Adversarial training of cross-domain image captioner," in *2017 IEEE International Conference on Computer Vision (ICCV)*, (Los Alamitos, CA, USA), pp. 521–530, IEEE Computer Society, oct 2017.

[23] M. Artetxe, G. Labaka, E. Agirre, and K. Cho, "Unsupervised neural machine translation," 2017.

[24] G. Lample, L. Denoyer, and M. Ranzato, "Unsupervised machine translation using monolingual corpora only," 2017.

[25] G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. Ranzato, "Phrase-based neural unsupervised machine translation," 2018.

[26] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, X. Bing, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *International Conference on Neural Information Processing Systems*, 2014.

[27] R. Shetty, M. Rohrbach, L. A. Hendricks, M. Fritz, and B. Schiele, "Speaking the same language: Matching machine to human captions by adversarial training," in *IEEE International Conference on Computer Vision*, 2017.

[28] X. Liang, Z. Hu, H. Zhang, C. Gan, and E. P. Xing, "Recurrent topic-transition gan for visual paragraph generation," 03 2017.

[29] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional gan," pp. 2989–2998, 10 2017.

[30] S. R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," pp. 10–21, 01 2016.

[31] A. Kannan and O. Vinyals, "Adversarial evaluation of dialogue models," 2017.

[32] J. Li, W. Monroe, T. Shi, A. Ritter, and D. Jurafsky, "Adversarial learning for neural dialogue generation," 2017.

[33] R. Lowe, M. Noseworthy, I. V. Serban, N. Angelard-Gontier, Y. Bengio, and J. Pineau, "Towards an automatic turing test: Learning to evaluate dialogue responses," 2017.

[34] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[35] X. Chen, H. Fang, T. Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *Computer Science*, 2015.

[36] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 39, no. 4, pp. 664–676, 2016.

[37] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *arXiv preprint arXiv:1506.06724*, 2015.

[38] C. K. Chen, Z. F. Pan, M. Sun, and M.-Y. Liu, "Unsupervised stylish image description generation via domain layer norm," 2018.

[39] 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 visual question answering," pp. 6077–6086, 06 2018.

[40] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, "Stylenet: Generating attractive visual captions with styles," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[41] W. Wagner, "Steven bird, ewan klein and edward loper: Natural language processing with python, analyzing text with the natural language toolkit," *Language Resources Evaluation*, vol. 44, no. 4, pp. 421–424, 2010.

[42] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, no. 1, pp. 853–899, 2013.