FAIEr: Fidelity and Adequacy Ensured Image Caption Evaluation

Sijin Wang[1,2], Ziwei Yao[1,2], Ruiping Wang[1,2], Zhongqin Wu[3], Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]Tomorrow Advancing Life Education Group, Beijing, 100080, China
{sijin.wang, ziwei.yao}@vipl.ict.ac.cn, {wangruiping, xlchen}@ict.ac.cn, wuzhongqin@tal.com

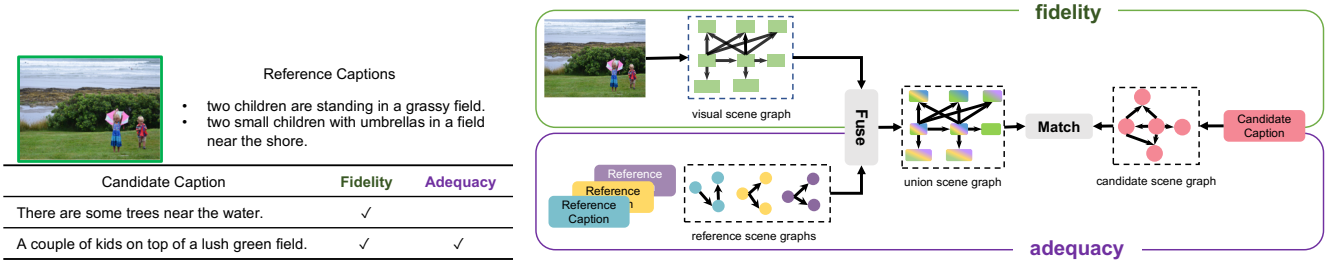
Abstract

*Image caption evaluation is a crucial task, which involves the semantic perception and matching of image and text. Good evaluation metrics aim to be fair, comprehensive, and consistent with human judge intentions. When humans evaluate a caption, they usually consider multiple aspects, such as whether it is related to the target image without distortion, how much image gist it conveys, as well as how fluent and beautiful the language and wording is. The above three different evaluation orientations can be summarized as **fidelity**, **adequacy**, and **fluency**. The former two rely on the image content, while fluency is purely related to linguistics and more subjective. Inspired by human judges, we propose a learning-based metric named **FAIEr** to ensure evaluating the fidelity and adequacy of the captions. Since image captioning involves two different modalities, we employ the scene graph as a bridge between them to represent both images and captions. FAIEr mainly regards the visual scene graph as the criterion to measure the fidelity. Then for evaluating the adequacy of the candidate caption, it highlights the image gist on the visual scene graph under the guidance of the reference captions. Comprehensive experimental results show that FAIEr has high consistency with human judgment as well as high stability, low reference dependency, and the capability of reference-free evaluation.*

1. Introduction

Good evaluations lead to continuous progress in many computer vision tasks. Different from other visual tasks, the evaluation of the image captioning [25, 39, 4, 40, 33, 38, 8, 7] is very difficult because the outputs of the image captioning are in the form of natural language and need to mirror the content of the given image, which involves multi-modals. Since image captioning can be regarded as translating the visual information into natural language, early image captioning methods follow the evaluation mode of the machine translation [30, 31, 12], which ignore the visual modal information. That is, the candidate caption is scored only based on the similarity with the human-labeled reference captions of the target image. In this formulation, early popular metrics [26, 9, 22, 32] measure the similarity of two sentences by the n-gram overlap, which results in low robustness to text ambiguity. To address this deficiency, SPICE [3] breaks the shackles of sentence structure by using the scene graph representations, which can measure the semantic similarity of sentences. However, "A picture paints a thousand words." Limited numbers of reference captions are hard to cover all contents in an image. Therefore, the reference-based metrics usually lead to biased evaluations. With recent breakthroughs, some studies [15, 6, 14] introduce the image information into the caption evaluation, which measures the similarity between the candidate caption and both the target image and reference captions simultaneously.

Developing automatic metrics aims to replace human judges, so the goal of a good metric is to reveal the human's evaluation intentions. From the human perspective, as shown in Fig.1(a), it is fundamental that the messages conveyed by a caption are related to the given image with no extra or distortion. Then, an adequate caption should describe the gist of the image concerned by humans [35]. Moreover, idiomatic wording and beautiful sentences will further get a higher score. We summarize them as three evaluating orientations: **fidelity**, **adequacy**, and **fluency**, which we think roughly form a multi-aspect criteria for image caption evaluation, to some extent similar to the machine translation evaluation systems [37, 27, 36]. With no consideration of image information, previous reference-based metrics cannot assess fidelity adequately, leading to biased evaluations. For example, if the candidate caption contains information not included in references but in the image, the reference-based metrics will fail to give a correct evaluation. It is also hard for those metrics based on n-gram overlap to ensure the evaluation of adequacy. Recent



(a) (b)

Figure 1. (a) An example image with human-labeled reference captions (from MS COCO [23]) and two candidates, showing the criterion of fidelity and adequacy. (b) The designing principle of FAIEr.

learning-based metrics [15, 6, 14] take fidelity into account by involving image information, yet fail to disassemble the complex human evaluation intentions. Beyond fidelity and adequacy, fluency is deemed to measure the quality of language expression more subjectively, which concerns little the image content and is purely related to linguistics.

In this paper, we focus on the first two objective orientations and propose a **F**idelity and **A**dequacy ensured **I**mage Caption **E**valuation metric named **FAIEr**. For fair evaluation, it gives the correct captions deserved scores, and ones containing more image gist will get more awards. FAIEr takes the image, reference captions, and candidate caption as input. The evaluation of fidelity mainly depends on the matching between the image and the candidate caption. To reward the adequacy, we need to compare the candidate and the reference captions (as the references convey the humans captured gist of the image). Therefore, the problem can be formulated as a multi-instance image-text matching task. To address such complex cross-modal matching task, FAIEr uses scene graphs as the intermediation to dissect and align the visual and textual information and then calculates the multi-modal similarity by scene graphs fusing and matching. As shown in Fig.1(b), FAIEr firstly represents the input image and captions as scene graphs. Taking the visual scene graph as the foundation, FAIEr further highlights the crucial contents that draw much human attention by fusing the reference and visual scene graphs into a union scene graph. The highlighted nodes will have larger weights in the evaluation process, which incorporates human evaluation intention intuitively. Finally, FAIEr scores the candidate caption by calculating the similarity between the candidate and union scene graphs.

Comprehensive experiments on Composite Dataset [1], Flickr8k [13], and PASCAL-50S [32] show the high consistency of FAIEr with human judgement, and verify the advantages of scene graph representations. In practical application scenarios, it is quite common that no human-labeled reference caption is available. Benefitting from the flexible scene graphs fusion module, FAIEr has the reference-free evaluation capability, which can readily address this issue.

2. Related work

According to whether involving the image information, popular evaluation metrics for image captioning can be divided into image-agnostic and image-based ones. In terms of the matching strategy, they can also be categorized into rule-based metrics, learning-based metrics, and a combination of them.

Image-agnostic metrics. Most of them calculate the similarity of the reference and candidate caption by word or n-gram overlap in a rule-based manner. BLEU [26] is a machine translation metric calculating the n-gram precision scores with a brevity penalty for short sentences. ROUGE-L [22] measures the similarity of a pair of sentences by the weighted harmonic average of the precision and recall of the longest common subsequence. METEOR [9] replaces the exact n-gram matching by WordNet-based synonym matching, and computes the similarity scores based on n-gram precision and recall. CIDEr [32] introduces the tf-idf weight to reduce the matching weight of the n-grams that are common in all image captions. In order to evaluate the similarity of two sentences from their semantic, SPICE [3] proposes to use semantic scene graphs to represent sentences, which breaks the constraint of the grammatical structure. To enhance the consistency with human judgment and robustness to textual ambiguity, recent studies propose learning-based metrics. [29] combines four different rule-based metrics through a learning-based framework and demonstrates that composite metrics can further impove caption evaluation. BERTScore [43] utilizes the BERT [10] to obtain the contextualized embeddings of text tokens for measuring the similarity of two sentences.

Image-based metrics. Recent studies [15, 6, 14] have found that only a limited number of reference captions are hard to cover all content in the given image, so they introduce the source image into the metrics as an additional "visual reference". [6] designs a neural network that learns to recognize whether the candidate caption is generated by humans. Though it augments pathological cases as negative examples to improve the model robustness, a simple

binary classification is not competent enough to supervise the model for learning the complex and subjective task. TIGEr [15] and REO apply the image-text matching model SCAN [21] to compute the caption-image and candidate-image grounding vectors, and then score candidate captions based on similarity of vectors. REO [14] additionally calculates mutual orthogonal projections between these vectors to assess the candidate quality from three perspectives - relevance, extraness, and omission. While these recent advances have made clear progress, their evaluation strategies either operate with a mixed measure score, or divide orientations while lacking adequate decomposition of evaluation progress. Our FAIEr tries to mimic humans' assessment and consider the multi-granularity semantic similarities from fidelity and adequacy. VIFIDEL [24] also proposes fidelity as a key point of image caption evaluation. It firstly takes the labels of objects detected in the image as visual representation, then gives them different weights according to reference captions, and finally calculates the WMD distance [20] between the weighted object labels and candidate caption. However, only using discrete textual object labels, VIFIDEL ignores other semantic image information such as relationships, attributes, and positions, so it is not able to reflect the target image completely. We utilize the structured scene graphs to represent images more comprehensively and encode all kinds of image features by learning process. What's more, scene graphs can represent both visual and textual semantic components, serving as a bridge between the two modalities to ease subsequent evaluation.

3. Method

An overall framework of our FAIEr metric is illustrated in Fig.2, which mainly consists of visual and textual embedding modules, attention fusion module, and matching module. Taking the image, N_f reference captions, and one candidate caption as input, FAIEr represents each of them as an instance of scene graph [34], and embeds the object-level and relationship-level representations respectively for each instance. Next, it fuses the visual and reference information at both object- and relationship-level by the attention fusion module to obtain the union reference representations. In the matching module, it computes the matching scores, S^o and S^r, between the candidate and union representations at two levels. Finally, the sum of the S^o and S^r serves as the evaluation score of the candidate caption w.r.t. the given image and references.

3.1. Scene graph embedding

Given an input image I, a pre-trained object detector is used to extract N_o object regions $O = \{o_i | i = 1, 2, ..., N_o\}$ in it. The initial visual feature of o_i is $\mathbf{u}_i \in R^{d_u}$, and the bounding box for o_i is b_i. The set of N_f reference captions is $F = \{F_j | 1 \leq j \leq N_f\}$, and the candidate caption is C. With such inputs, FAIEr firstly builds scene graphs and embeds the object- and relationship-level representations.

Visual scene graph embedding. A visual object encoder encodes the i-th object region as $\mathbf{v}_i^o = \mathbf{W}_2 \tanh(\mathbf{W}_1 \mathbf{u}_i)$, where $\mathbf{v}_i^o$ is the object-level representations, and $\mathbf{W}_1 \in R^{d \times d_u}$ and $\mathbf{W}_2 \in R^{d \times d}$ are trainable parameters. The visual scene graph is initialized as a complete graph, where each node represents an object.Then, a Graph Convolutional Network (GCN) acts as the visual graph encoder to embed the relationship-level representations between nodes. $\mathbf{v}_i^r$ is the relationship-level representation of o_i, which is computed as the attention weighted aggregate message from all its neighbor nodes. We use the offset between the bounding boxes of two objects, $\Delta \mathbf{b}_{ij} = b_i - b_j$, as the weight of the attention so that each node can selectively receive information from the connected nodes. The update process is computed as:

$$\gamma_{ij} = \frac{\exp(\tanh(\mathbf{W}_\Delta \Delta \mathbf{b}_{ij}))}{\sum_{k \in \mathcal{N}_i} \exp(\tanh(\mathbf{W}_\Delta \Delta \mathbf{b}_{ik}))}, \mathbf{v}_i^r = \sum_{j \in \mathcal{N}_i} \gamma_{ij} \mathbf{W}_r \mathbf{v}_j^o, \tag{1}$$

where γ_{ij} represents the message passing weight from node o_j to o_i. $\mathcal{N}_i$ denotes the neighbourhood of o_i, which also contains itself to retain its original characteristics. $\mathbf{W}_\Delta \in R^{1 \times 4}$ and $\mathbf{W}_r \in R^{d \times d}$ are trainable parameters.

Apart from the offset between the bounding boxes of two objects, we also tried other more complicated ways to calculate the weight of GCN attention, such as the fusion of region features, bounding boxes and textual label embeddings of each object pair, but they did not bring significant improvement.

Textual scene graph embedding. Suppose there are $L_{f_j}^o$ and L_c^o words in F_j and C, respectively. In case that no reference caption is available, we use "." as an empty reference sentence. For the textual object-level representations, each word w_k (k is the index of the word) is encoded by the embedding layer as $\mathbf{e}_{w_k}$ first. Then the textual word encoder, a bidirectional-GRU, will embed the words along with each sentence. We take the hidden state $\mathbf{h}_{w_k}^o$ of the word w_k as its object-level representation. The object-level representations of the reference caption F_j and the candidate caption C are represented by $\mathbf{h}_j^o = \{\mathbf{h}_{jk}^o | 1 \leq k \leq L_{f_j}^o\}$ and $\mathbf{h}_c^o = \{\mathbf{h}_{ck}^o | 1 \leq k \leq L_c^o\}$.

Next, FAIEr employs SPICE as the textual graph parser to parse each caption into a textual scene graph, in which the nodes are object words (*e.g.* "kids", "grass") and the edges are relationship phrases (*e.g.* "stand in"). After that, the textual relationship-level representations are encoded by the textual graph encoder. It takes the semantic triplets from the textual scene graph as input, such as "⟨kids-stand in-grass⟩", and then uses a bidirectional-GRU to encode each triplet. Finally, we take the last hidden state feature of each triplet as the relationship-level representations, $\mathbf{h}^r$. Suppose there are $L_{f_j}^r$ and L_c^r triplets in the reference F_j and

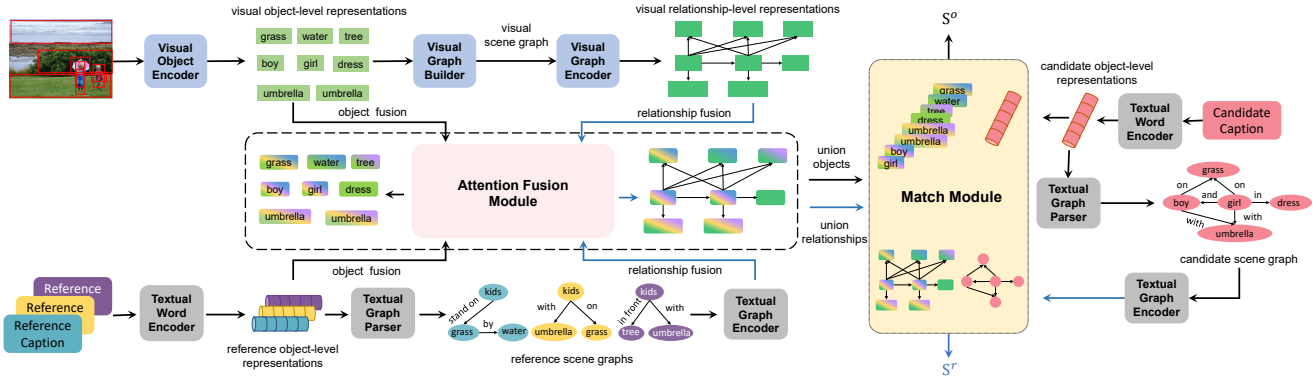


Figure 2. The framework of the image caption evaluation metric FAIEr (best viewed in a digital version).

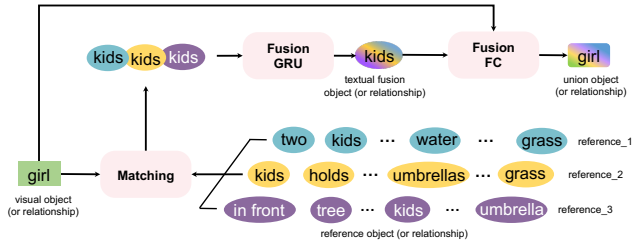


Figure 3. Illustration of the attention fusion layer.

the candidate C. Then the relationship-level representations of the reference F_j and candidate C are represented as $\mathbf{h}_j^r = \{\mathbf{h}_{jk}^r | 1 \leq k \leq L_{f_j}^r\}$ and $\mathbf{h}_c^r = \{\mathbf{h}_{ck}^r | 1 \leq k \leq L_c^r\}$, respectively.

3.2. Scene graph fusing

To obtain adequate reference information that reflects humans' major perception of the image, we fuse the visual and reference scene graphs via the attention mechanism. The framework of the attention fusion layer is illustrated in Fig.3. Given a visual object or relationship vector $\mathbf{v}_i^\chi$ (χ can be o or r), it first attends to the textual object or relationship vectors respectively in every reference as:

$$\alpha_{i,j,k} = softmax(\mathbf{W}_a(\mathbf{v}_i^\chi \odot \mathbf{h}_{jk}^\chi)), \mathbf{a}_{ij}^\chi = \sum_{k=1}^{L_{f_j}^\chi} (\alpha_{i,j,k} \mathbf{h}_{jk}^\chi), \tag{2}$$

where $\alpha_{i,j,k}$ is the attention weight of the k-th textual object or relationship node in F_j, $\mathbf{a}_{ij}^\chi$ is the attended information for $\mathbf{v}_i^\chi$ w.r.t F_j, $\mathbf{W}_a \in R^{1\times d}$ are trainable parameters. For all attended textual information $\mathbf{a}_i^\chi = \{\mathbf{a}_{ij}^\chi\}$ of $\mathbf{v}_i^\chi$, a fusion GRU (shown in Fig.3) fuses them into fusion textual information as $\mathbf{m}_i^\chi = GRU_m^\chi(\mathbf{a}_i^\chi)$. At last, a FC layer merges the visual and fusion textual information into the union information as $\mathbf{z}_i^\chi = \mathbf{W}_u([\mathbf{v}_i^\chi; \mathbf{m}_i^\chi])$, where $\mathbf{W}_u \in R^{d\times 2d}$ are trainable parameters. The union object- and relationship-level represetations are $\mathbf{z}^o = \{\mathbf{z}_i^o\}$ and $\mathbf{z}^r = \{\mathbf{z}_i^r\} (i \in [1, N_o])$, which highlights the content attended by humans.

3.3. Scene graph matching

The matching module aims to score the candidate caption by matching the candidate scene graph with the union scene graph. Inspired by [17], we define the similarity between two cross-modal vectors $\mathbf{z}_i^o$ and $\mathbf{h}_{ck}^o$ as their dot product $\mathbf{z}_i^{oT}\mathbf{h}_{ck}^o$. Then the matching scores between two scene graphs at the object-level and similarly at the relationship-level are computed as follows:

$$\begin{aligned} S^o &= \frac{\sum_{k=1}^{L_c^o} \max_{i\in[1,\,N_o]}(\mathbf{z}_i^{oT}\mathbf{h}_{ck}^o)}{L_c^o}, \\ S^r &= \frac{\sum_{k=1}^{L_c^r} \max_{i\in[1,\,N_o]}(\mathbf{z}_i^{rT}\mathbf{h}_{ck}^r)}{L_c^r}, \end{aligned} \tag{3}$$

which means for every candidate object or relationship, the most similar union object or relationship is picked up, and then the scores are averaged by the number of the objects or relationships. Finally, the score of the candidate caption w.r.t. the union reference information is $S = S^o + S^r$.

3.4. Loss function

We use the triplet loss Eq.(4) to train our FAIEr metric, which lets the candidate get higher score given its target image (along with its reference captions). In one mini-batch, the score of the k-th candidate with its target image I_k is set to S_{kk}. The k-th candidate given the l-th image ($k \neq l$) and the l-th candidate given the k-th image are unmatched pairs, whose scores are S_{kl} and S_{lk} respectively.

$$\begin{aligned} L = \sum_k (\sum_l &\max(0, m - S_{kk} + S_{kl}) \\ &+ \sum_l \max(0, m - S_{kk} + S_{lk})). \end{aligned} \tag{4}$$

After adding the hardest negative mining [11], our loss function is defined as

$$\begin{aligned} L_+ = \sum_k (&\max(0, m - S_{kk} + S_{k\hat{p}}) \\ &+ \max(0, m - S_{kk} + S_{\hat{q}k})), \end{aligned} \tag{5}$$

where $\hat{p} = \arg\max_{t \neq k} S_{kt}$ and $\hat{q} = \arg\max_{t \neq k} S_{tk}$ are hard negatives, and m is a margin parameter.

4. Experiments

4.1. Datasets and metrics

Human correlation and accuracy are commonly used to evaluate the image caption evaluation metrics. To calculate the caption-level correlation between the metrics and human judgments, each candidate caption in this kind of datasets is annotated with scores by human beings, such as Composite Dataset [1] and Flickr8k [13]. Following prior studies, we use Pearson's ρ, Kendall's τ, and Spearman's ρ correlations to calculate pairwise scores between human-beings and automatic metrics. The other kind of datasets, *e.g.* PASCAL-50S [32], ask annotators to choose the better one from candidate caption pairs, so we calculate the accuracy of metrics. Moreover, as a learning-based metric, we also conduct experiments on a cross-domain dataset Nocaps [2] to validate the generalization ability.

Composite Dataset. It has $3,995$ images from the testing splits of three different datasets, MS COCO [23], Flickr30k [41], and Flickr8k [13]. Each image has three candidate captions, of which one is written by human and the other two are machine captions generated by image captioning models [16, 1]. All candidate captions were scored by annotators on a graded correctness scale from 1 (not related to the image) to 5 (perfectly related to the image).

Flickr8k contains $8,092$ images, each of which has 5 human written captions. $5,822$ captions from the testing splits are scored by annotators from 1 (unrelated to the image) to 4 (describes the image correctly). To keep consistency with comparative work [3], we similarly excluded the 158 candidate captions that are scored according to their ground-truth image.

PASCAL-50S collects $1,000$ images from UIUC PASCAL Sentence Dataset [28], and 50 human written reference captions for each image. This dataset also contains $4,000$ candidate caption pairs with human judgments, which form 4 groups with 1,000 pairs in each group. 1) HC group includes correct human written pairs for the target images. 2) HI contains correct and incorrect human written caption pairs. 3) HM pairs include a human written caption and a machine-generated caption for the same image. 4) MM is machine-generated sentence pairs for each image.

Nocaps is a dataset for novel object captioning. It collects 166,100 captions generated by human for 15,100 images of validation and test set of Open Images Dataset V4 [19], which contains about 400 object classes associated little with training captions in MS COCO. A subset of its validation set, containing 1,000 images and 10 captions for each image, is used in our experiment.

4.2. Implementation details and experiment settings

Implementation details. We follow [16] to split $5,000$ images for validation and $5,000$ images for test from MS COCO [23], and trained all our models on the remaining $113,287$ images that not overlap with other datasets. We use the scene graph generator NeuralMotifs [42] to propose the bounding boxes for top 36 objects in each image, and then use the object detector in [4] to extract initial visual features $\mathbf{u}_i \in R^{d_u}$ ($d_u = 2048$) for each region. SPICE [3] is utilized as our textual scene graph parser to extract relation triplets from captions. Our method is implemented with the Pytorch platform[1]. The dimension of our embedding space $d = 512$. The margin m is set to 0.2. We use Adam [18] optimizer with a mini-batch size of 100 for training. The initial learning rate is 0.0005.

Due to the 512-dimension embedding space and two evaluation branches of object-level and relationship-level, the theoretical range of our metric is [-1024, 1024]. However, we find that scores are distributed over a much smaller range in practice. Through massive experiments on different datasets, we find that more than 99% of scores are in the range of -2 and 6, which can be seen as the experimental bound. In addition, the score distribution is also related to the margin in the loss function. The distribution range of results expands with the increasing of the chosen margin.

Experiment settings. Here we will clarify all our variant models. **FAIEr** is our full model. **FAIEr\rel** only has object-level cross-modal fusing and matching. **FIEr** evaluates candidates by only matching the candidate caption with the image information. **AIEr** denotes only reference captions are involved in evaluating the candidate caption. **X-n ref** means the model **X** uses n reference captions during training, *e.g.* **FAIEr-1 ref**. **X-r ref r**andomly picks 0 to 4 reference captions for each training sample, so it can work with no reference during test. We use MS COCO evaluation tool[2] to implement the rule-based metrics BLEU, METEOR, ROUGE-L, CIDEr, and SPICE. The learning-based metrics TIGEr [15] and LearnToEval [6] are implemented by their public codes. Note that we train LearnToEval on MS COCO and test on Composite and Flickr8k, which is different from their original usage. "*LearnToEval" in Table 1 and "*VIFIDEL" in Table 3 are original results copied from their paper [6] and [24]. Since the reference captions are randomly selected in the test, we test each model 5 times and take the average result.

4.3. Quantitative experiments

We conduct comprehensive quantitative experiments to compare our metric with the rule-based metrics [26, 22, 9, 32, 3] and the state-of-the-art learning-based metrics [15, 6, 24] that use image. Compared with the rule-

[1]Our source codes are available at http://vipl.ict.ac.cn/resources/codes.
[2]https://github.com/tylin/coco-caption

Table 1. Comparisons of state-of-the-art metrics on Composite Dataset and Flickr8k.

Method	Composite Dataset			Flickr8k		
	P-ρ	S-ρ	K-τ	P-ρ	S-ρ	K-τ
BLEU-1	0.392	0.386	0.287	0.324	0.294	0.218
BLEU-4	0.272	0.365	0.271	0.068	0.279	0.206
METEOR	0.376	0.449	0.338	0.493	0.464	0.350
ROUGE-L	0.397	0.392	0.296	0.365	0.319	0.239
CIDEr	0.321	0.429	0.324	0.422	0.415	0.314
SPICE	0.408	0.432	0.347	0.568	0.559	0.478
*LearnToEval [6]	-	-	-	-	-	0.466
LearnToEval [6]	0.318	0.368	0.273	0.415	0.475	0.355
TIGEr [15]	0.522	0.540	0.409	0.655	0.665	0.517
FAIEr-1 ref	**0.646**	**0.661**	**0.514**	0.702	0.713	0.563
FAIEr-4 ref	0.605	0.630	0.487	0.696	0.708	0.557
FAIEr-r ref	0.643	0.660	0.513	**0.709**	**0.718**	**0.568**
FAIEr-r ref (ref-free)	0.595	0.607	0.467	0.674	0.694	0.544

Table 2. Accuracy of different metrics and the average scores of human written and machine-generated captions in HM-MSCOCO. Testing with four reference captions.

Method	Accuracy(%)	Average Score	
		human	machine
BLEU-1	46.2	0.629	0.646
BLEU-4	44.2	0.195	0.241
METEOR	54.3	0.241	0.229
ROUGE-L	45.2	0.466	0.489
CIDEr	51.0	0.877	0.858
SPICE	55.8	0.210	0.179
TIGEr	62.7	0.732	0.706
FAIEr-4 ref	76.8	3.458	2.931

based metrics, FAIEr achieves higher consistency and stability and lower reference dependency, which benefits from the introduction of the image. Comparisons with other learning-based metrics validate that our evaluation strategy is more consistent with human intentions.

High human correlation. Table 1 displays the correlation between human judgement and different metrics. Except for **FAIEr-r ref (ref-free)** that uses no reference caption during testing, others use one reference. Compared with other metrics, our models achieve significantly higher correlation with human judgment, indicating that **FAIEr** can more accurately capture human evaluation intentions. More importantly, the promising performance of **FAIEr-r ref (ref-free)** shows that our metric is able to tackle the practical scenarios where no human-labeled reference caption is available. In addition, we conduct experiments under VIFIDEL's experimental settings on the Composite, calculating human correlation from two aspects: relevance and thoroughness, of which results also reflect our advantage. More details can be seen in supplementary materials.

High consistency. To further show the high consistency of our metric, we collect a new dataset named HM-MSCOCO (details in the supplementary material), which contains $15,000$ candidate caption pairs, each with a human written caption and a machine-generated one. Table 2 shows the results of different metrics. The accuracy is defined as the percentage of pairs whose human caption gets

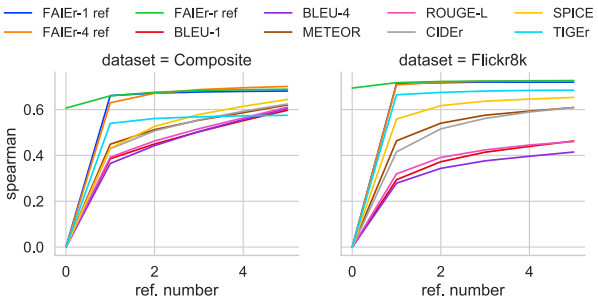


Figure 4. Testing metrics using different number of reference captions on Composite and Flickr8k.

a higher score than the machine-generated one. BLEU-1, BLEU-4, and ROUGE-L suggest that less than 50% of human captions are superior to machine ones. Their average scores of human captions are also smaller than machine captions. TIGEr expects 62.7% of human captions are better. **FAIEr** thinks human is better than machine in 76.8% pairs and gives the human a 18% (3.458 vs. 2.931) higher scores than machines. The possible reason here is if for the same image, the candidate with high-fidelity has different attentions from the reference captions, the rule-based metrics fail to give a fair score. Our **FAIEr** addresses this issue by considering both the image information and human evaluation intentions in the evaluation strategy.

Low reference dependency. We explore how the number of reference captions affects the metrics performance by testing from 0 to 5 references. Both candidate and reference captions are randomly chosen from human references. As shown in Fig.4, the Spearman correlation increases with the number of references growing, where the growth rate gradually slows down because the union of plenty of references tends to be a stable set. Our models generally outperform other metrics in almost all cases, especially with fewer references. Besides, the results also show the high stability and low reference dependency of our models, owing to the help of images. When testing with more reference captions, the rule-based metrics have large increase and even outperform TIGEr on Composite, because when randomly picking more references, the reference that is same as the candidate is more likely to be selected, which will favor the rule-based metrics. Comparing different invariants of **FAIEr**, we find **FAIEr-1 ref** performs better with fewer references, and then is surpassed by **FAIEr-4 ref** when testing with more references, because training the model using a fixed number of references will rigidify its evaluation strategy. **FAIEr-r ref** performs stably, even testing with zero reference. Overall, their performances have slight difference.

High stability. Table 3 displays the accuracy (%) and standard deviation of metrics on PASCAL-50S. The results indicate that the HC pairs are most difficult to judge while the HI pairs are easiest. We achieve the best performance on both of them. In the other two groups, HM and MM, **FAIEr**

Table 3. Comparisons of the accuracy of metrics on PASCAL-50S. Testing with five references.

Method	HC	HI	HM	MM
BLEU-1	50.9 ± 0.70	94.8 ± .72	91.8 ± .70	57.7 ± 1.50
BLEU-4	53.0 ± 0.90	92.3 ± .32	86.3 ± .82	60.6 ± 1.20
METEOR	58.1 ± 1.30	97.3 ± .31	**93.8** ± .55	63.0 ± 1.10
ROUGE-L	53.4 ± 1.70	95.3 ± .81	92.6 ± .45	58.3 ± 0.61
CIDEr	54.6 ± 1.50	98.1 ± .23	91.3 ± .77	64.0 ± 1.10
SPICE	55.2 ± 1.50	93.7 ± .63	85.8 ± .36	50.0 ± 1.00
TIGEr [15]	55.1 ± 0.42	99.7 ± .05	92.1 ± .31	**74.6** ± 0.79
*VIFIDEL [24]	**64.0**	97.0	75.0	72.0
FAIEr\rel-4 ref	58.4 ± 0.68	99.8 ± .08	92.6 ± .50	73.1 ± 0.70
FAIEr-4 ref	59.7 ± **0.39**	**99.9** ± **.00**	92.7 ± **.23**	73.4 ± **0.42**

Table 4. Evaluation of variants of our model on Composite Dataset and Flickr8k. Testing with one reference caption.

Method	Composite Dataset			Flickr8k		
	P-ρ	S-ρ	K-τ	P-ρ	S-ρ	K-τ
FIEr	0.602	0.612	0.471	0.683	0.697	0.548
AIEr-4 ref	0.202	0.169	0.125	0.037	0.049	0.036
FIEr+AIEr-4 ref	0.206	0.185	0.136	0.006	0.028	0.020
FAIEr\rel-4 ref	0.587	0.611	0.472	0.698	0.705	0.554
FAIEr-4 ref	0.605	0.630	0.487	0.696	0.708	0.557
FAIEr\rel-1 ref	0.643	0.656	0.509	0.693	0.710	0.561
FAIEr-1 ref	**0.646**	**0.661**	**0.514**	0.702	0.713	0.563
FAIEr\rel-r ref	0.645	0.661	0.514	0.694	0.708	0.559
FAIEr-r ref	0.643	0.660	0.513	**0.709**	**0.718**	**0.568**

shows comparable performance with the best metrics. Due to the introduction of image information, the standard deviations of **FAIEr** and TIGEr are significantly smaller than other rule-based metrics, which show high stability.

Effectiveness of modules. To verify the effectiveness of the proposed method, we evaluate many variants of our models in Table 4. The **FIEr** model only using image information for evaluation has high consistency with human judgment, which can reveal the importance of the image. Since our framework is designed for cross-modal matching and scoring, missing of image basis probably causes low and unstable performance of **AIEr**. **FIEr+AIEr-4 ref** merges the scores of **FIEr** and **AIEr** by a FC layer. However, generally merging two modules with different training difficulty by force has a negative effect, proving the effectiveness of our attention fusion layer that splits reference captions and fuses them into image. Due to **FAIEr\rel** fuses the image and reference for evaluation, it improves the evaluation ability. Comparing **FAIEr** to **FAIEr\rel** in Table 4 and Table 3, the scene graph representation shows its advantages, owing to the relationship information.

4.4. Qualitative analysis

Illustrative examples. Fig.5 shows an example image with three references from MS COCO test split and evaluation scores of 4 candidates by several metrics. The 1st candidate is a correct caption with similar words as references, so all metrics give it the highest score. The 2nd one is more detailed that mentions some objects not appearing in references. The 3rd one is also correct but has quite different expression.The 4th candidate is an incorrect caption just using similar words as references. The rule-based metrics sometimes cannot correctly evaluate captions with high-fidelity. BLEU-4, METEOR, and ROUGE-L give higher scores to the 4th incorrect candidate than the 2nd and 3rd correct ones, showing their low robustness against textual ambiguity. Without image information, SPICE cannot evaluate the second one correctly. Besides, the scores of three correct candidates have large gaps in some metrics, especially on BLEU-4, ROUGE-L and SPICE. Compared with them, our **FAIEr\rel** and **FAIEr** are much more reasonable. More cases can be seen in our supplementary materials.

Table 5. The average scores for matching the visual regions with its GT (ground-truth) words and non-GT words respectively in references and candidates.

ref. number	image-references pair		union-candidate pair	
	GT words	non-GT words	GT words	non-GT words
1	3.10	0.21	5.04	0.86
2	3.10	0.20	5.62	0.79
3	3.09	0.20	5.81	0.75
4	3.09	0.20	5.92	0.73

Grounding analysis. To analyze the evaluation process of our FAIEr in-depth, we utilize MSCOCO Entities [5] dataset (details in the supplementary materials) to match the visual object regions and fused object nodes with the reference and candidate words, respectively, and calculate the average scores of the GT (ground-truth) and non-GT words. As shown in Table 5, the GT words are scored much higher than non-GT words, which verify the effectiveness of our metric. As the number of references increases, the score of candidate GT words increases, while the score of candidate non-GT words decreases. It reveals our attention fusion module can give larger weight to important content. These results also reflect FAIEr can ensure to evaluate fidelity and adequacy. Fig.6 illustrates four major objects in an image and their heatmaps with regard to three references and four candidates. The three red heatmaps are matching between visual objects and three references. The visual objects can find their corresponding words in references, such as the obj1 **vs.** "children". After fusing the visual information and the references, we show the similarity scores between fusion objects and candidates in blue heatmaps, which is normalized in each caption. From these heatmaps, we find that the critical words in correct candidates can achieve high scores. Since the incorrect candidate contains no corresponding content with image and the references, it receives a low score. More detailed analysis and cases are shown in our supplementary materials.

4.5. Generalization on cross-domain datasets

Generalization ability is essential for learning-based evaluation metrics [6, 15]. Trained on a specific dataset,

Image	References	Candidates	w/o image				with image	
			BLEU-4	METEOR	ROUGE-L	SPICE	FAIEr/rel	FAIEr
	• two children are standing in a grassy field. • there are two children standing in the grass by water. • two small children with umbrellas in a field near the shore.	Two children with umbrellas in a grassy field.	0.84	0.389	0.797	0.636	4.282	4.670
		A girl with blue dress and a boy standing in front of trees.	3.91e-9	0.175	0.199	0.0	3.416	3.133
		A couple of kids on top of a lush green field.	7.08e-13	0.188	0.217	0.167	3.853	3.464
		Two dogs are standing in an oil field.	4.93e-5	0.258	0.625	0.1	1.894	1.580

Figure 5. Evaluation examples of different metrics.

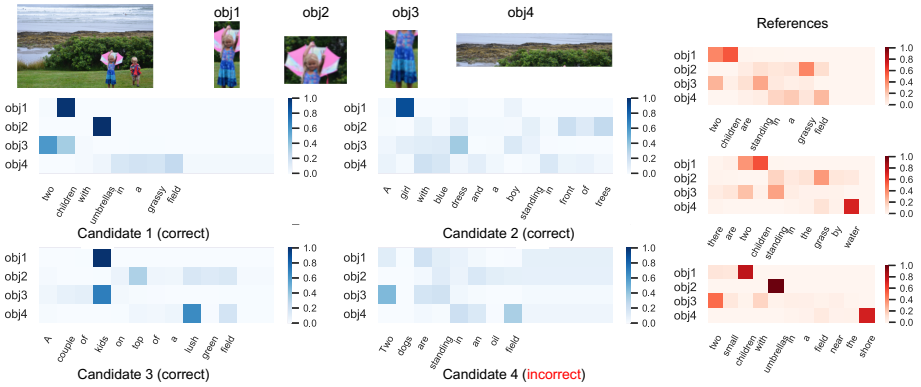


Figure 6. Visualizing the object-level evaluation of FAIEr\rel.

a good metric needs to be effective on other cross-domain datasets. In fact, with our model trained on MS COCO, all experiments above are conducted on other datasets without re-training or finetuning, but we additionally validate and compare its extensiveness on a dataset more different from MS COCO. Therefore, we test learning-based metrics on a subset of validation set of Nocaps Dataset, which contains about 400 object classes hardly seen during training. Since there are no human judgments for captions in Nocaps, we design and conduct cross-modal retrieval experiments. Image-text retrieval requires models to learn both image and text information and the correspondence between them precisely, which are also the key points of an image caption evaluation metric. More specifically, given an image or a caption as query, we score it on all provided captions or images and sort the results in descending order. An effective retrieval model should be able to identify the groundtruth answers among all the candidates and give them higher scores. As displayed in Table 6, owing to our comprehensive scene graph representations and delicate attention fusion mechanism, our model is better at understanding both modalities and shows higher extensiveness. More detailed examples can be seen in supplementary materials.

Table 6. Image-text retrieval results on Nocaps Dataset. Testing with four reference captions. R@k is the percentage of queries whose ground-truth is ranked within top K.

Method	image to text			text to image		
	R@1	R@5	R@10	R@1	R@5	R@10
TIGEr [15]	0.638	0.870	0.924	0.225	0.665	0.819
FAIEr-4 ref	**0.965**	**0.998**	**1.000**	**0.825**	**0.953**	**0.975**

5. Conclusion

In this paper, we tried to decompose the complex and subjective human's evaluation intentions as: fidelity, adequacy, and fluency for image captioning. To address the evaluation of fidelity and adequacy, we propose a learning-based metric **FAIEr**. Given the image, reference captions, and candidate caption as input, FAIEr uses the scene graphs to characterize the human semantic perception of visual and textual information. It measures the fidelity of the candidate mainly depending on the visual scene graph and fuses the reference information into the visual scene graph for further examining the adequacy. Comprehensive experiments show FAIEr can more accurately reveal the human evaluation intentions and ensure the assessment of fidelity and adequacy. Besides, FAIEr also shows high stability, low reference dependency, and the ability of reference-free evaluation. When applying FAIEr for practical evaluations, we believe that using larger training datasets with less bias or finetuning on the target dataset can generally boost its performance. For further comprehensive evaluation, it is indispensable to involve the factor of "fluency" which is expected to benefit from more advanced NLP techniques. It is also worth exploring the evaluation mechanism that fuses the learning-based and rule-based metrics.

Acknowledgements. This work is partially supported by Natural Science Foundation of China under contracts Nos. 61922080, U19B2036, 61772500, National Key R&D Program of China under Grant No. 2020AAA0104500, CAS Frontier Science Key Research Project No. QYZDJ-SSWJSC009, and Beijing Academy of Artificial Intelligence No. BAAI2020ZJ0201.

References

[1] Somak Aditya, Yezhou Yang, Chitta Baral, Yiannis Aloimonos, and Cornelia Fermuller. Image understanding using vision and reasoning through scene description graph. *Computer Vision and Image Understanding*, 173:33–45, 2017. 2, 5

[2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957, 2019. 5

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision*, pages 382–398. Springer, 2016. 1, 2, 5

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 1, 5

[5] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 7

[6] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. Learning to evaluate image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5804–5812, 2018. 1, 2, 5, 6, 7

[7] Bo Dai, Sanja Fidler, and Dahua Lin. A neural compositional paradigm for image captioning. In *Advances in Neural Information Processing Systems*, pages 658–668, 2018. 1

[8] Bo Dai and Dahua Lin. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems*, pages 898–907, 2017. 1

[9] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on Statistical Machine Translation*, pages 376–380, 2014. 1, 2, 5

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019. 2

[11] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference*, 2018. 4

[12] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828, 2016. 1

[13] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47(1):853–899, 2013. 2, 5

[14] Ming Jiang, Junjie Hu, Qiuyuan Huang, Lei Zhang, Jana Diesner, and Jianfeng Gao. Reo-relevance, extraness, omission: A fine-grained evaluation for image captioning. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1475–1480, 2019. 1, 2, 3

[15] Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. Tiger: Text-to-image grounding for image caption evaluation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2141–2152, 2019. 1, 2, 3, 5, 6, 7, 8

[16] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015. 5

[17] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems*, pages 1889–1897, 2014. 4

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

[19] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):2–3, 2017. 5

[20] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015. 3

[21] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision*, pages 201–216, 2018. 3

[22] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the 42th annual meeting on Association for Computational Linguistics*, pages 74–81, 2004. 1, 2, 5

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014. 2, 5

[24] Pranava Madhyastha, Josiah Wang, and Lucia Specia. VIFIDEL: Evaluating the visual fidelity of image descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6539–6550, Florence, Italy, July 2019. Association for Computational Linguistics. 3, 5, 7

[25] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *International Conference on Learning Representations*, 2015. 1

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine

translation. In *Proceedings of the 40th annual meeting on Association for Computational Linguistics*, pages 311–318, 2002. 1, 2, 5

[27] John R. Pierce and John B. Carroll. *Language and Machines: Computers in Translation and Linguistics*. National Academy of Sciences/National Research Council, 1966. 1

[28] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *North American Chapter of the Association for Computational Linguistics*, pages 139–147, 2010. 5

[29] Naeha Sharif, Lyndon White, Mohammed Bennamoun, and Syed Afaq Ali Shah. Learning-based composite metrics for improved caption evaluation. In *Proceedings of ACL 2018, student research workshop*, pages 14–20, 2018. 2

[30] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014. 1

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. 1

[32] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015. 1, 2, 5

[33] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015. 1

[34] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1508–1517, 2020. 3

[35] Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Sketching image gist: Human-mimetic hierarchical scene graph generation. In *Proceedings of the European Conference on Computer Vision*, pages 222–239, 2020. 1

[36] John S White, Theresa A O'Connell, and Francis E O'Mara. The arpa mt evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, 1994. 1

[37] Wikipedia. Evaluation of machine translation — Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Evaluation_of_machine_translation, 2004. [Online; accessed 26-September-2020]. 1

[38] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015. 1

[39] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019. 1

[40] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision*, pages 684–699, 2018. 1

[41] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. In *Transactions of the Association for Computational Linguistics*, volume 2, pages 67–78. MIT Press, 2014. 5

[42] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018. 5

[43] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. 2