Improving Image Captioning Evaluation by Considering Inter References Variance

Yanzhi Yi and **Hangyu Deng** and **Jinglu Hu**
Graduate School of Information, Production and Systems, Waseda University,
2-7 Hibikino, Wakamatsu, Kitakyushu-shi, Fukuoka, Japan, 808-0135
yiyanzhi@akane.waseda.jp, deng.hangyu@fuji.waseda.jp, jinglu@waseda.jp

Abstract

Evaluating image captions is very challenging partially due to the fact that there are multiple correct captions for every single image. Most of the existing one-to-one metrics operate by penalizing mismatches between reference and generative caption without considering the intrinsic variance between ground truth captions. It usually leads to over-penalization and thus a bad correlation to human judgment. Recently, the latest one-to-one metric BERTScore can achieve high human correlation in system-level tasks while some issues can be fixed for better performance. In this paper, we propose a novel metric based on BERTScore that could handle such a challenge and extend BERTScore with a few new features appropriately for image captioning evaluation. The experimental results show that our metric achieves state-of-the-art human judgment correlation.

1 Introduction

Image captioning is one of the key visual-linguistic tasks that asks for generated captions with specific images. Researchers look forward to inexpensive evaluation metrics that closely resemble human judgment, which remains a challenging task since most of the metrics can hardly get close to human judgment.

Image captioning is a one-to-many task since each image can correspond to many possible captions. Different captions may focus on different parts of the image; this not only creates a challenge for generating the captions (Dai et al., 2017; Venugopalan et al., 2017), but also for evaluating them. Most of the existing one-to-one evaluation metrics, however, overlook such a challenge. These one-to-one metrics (Lin, 2004; Vedantam et al., 2015; Zhang et al., 2019) ignore other reference captions since the score is computed by comparing the candidate capture with one single reference caption.

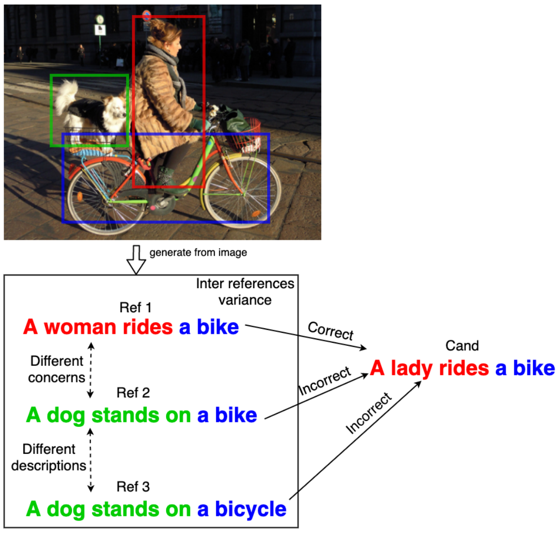

Figure 1: Intrinsic variance exists in a set of ground truth captions for an image. Differences between two references are commonly caused by two reasons: different concerns or different descriptions. Different concerns mean different expressions between references are caused by different regions of interest in an image, while different descriptions mean references focus on the same part but use different ways to explain it. One-to-one metrics can hardly deal with the cases caused by different concerns. For example, they may regard **Cand** as a good caption compared with **Ref1**; while regard **Cand** as a bad caption compared with **Ref2** or **Ref3**.

When there are multiple reference captions, prior works compute individual scores for each reference caption and pool these scores together afterward. Intrinsic variance exists in a set of ground truth captions for an image, since different captions may have different concerns or descriptions. It's challenging to find a remedy for such over-penalization if the metric looks at only one single reference caption.

BERTScore (Zhang et al., 2019) is the latest one-to-one metric that computes token-level cosine

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 985–994
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

similarity between two sentences by contextual embeddings of pre-trained models, and greedily picks and adds up cosine values as a score. It reaches high performance in machine translation tasks and a system-level image captioning evaluation task.

In one-to-one evaluation, although it is hard to consider all references directly, it is possible to combine references into a single one using contextual embedding from the pre-trained language model. In this work, we propose a metric where all of the references are combined as a new comprehensive embedding by detecting the mismatches between two contextual embeddings. To achieve this goal, we add the concept of mismatch into cosine similarity by a threshold for mismatch detection and proper penalization. Also, our metric considers the importance of different words, and our research shows that adding a stop word list is an efficient way.

Using various image captioning evaluation datasets with human annotations like Microsoft COCO (Lin et al., 2014), Flickr8k (Hodosh et al., 2013), COMPOSITE (Aditya et al., 2015) and PASCAL-50S (Vedantam et al., 2015), the experimental results show that our metric achieves state-of-the-art correlation in several tasks, especially in caption-level tasks. Our main contribution is a novel metric that can detect mismatches among captions, build a combined caption with multi-references, and achieve high human correlation in image captioning evaluation tasks. The code for our metric is released at here[1].

2 Related work

2.1 Automated caption evaluation

For captions evaluation, a traditional method is scoring by human experts, which is a precise but expensive way. Current image captioning models are evaluated by automatic metrics, which compute the similarity between generated captions and ground truth captions.

Currently, most widely used caption metrics are n-gram matching metrics such as BLEU, METEOR, ROUGE, CIDEr. BLEU (Papineni et al., 2002) is a precision-based n-gram overlap matching metric that counts the number of overlap n-grams among all of references and the candidate. Several modifications can be applied to improve BELU, such as different n-gram (e.g. n=1,2,3,4), brevity penalty for a short candidate, and geometrical average. BLEU is a fast, low-cost metric but has a low correlation with human judgment. METEOR (Denkowski and Lavie, 2014) computes both precision and recall in unigram, and consider more factors such as word stems, synonyms, and paraphrases. ROUGE (Lin, 2004) is a package of measures for automatic text summaries evaluation: ROUGE-N uses n-gram co-occurrence statistics; ROUGE-L uses the longest common sub-sequence; ROUGE-W uses weighted longest common sub-sequence; ROUGE-S uses skip-bigram co-occurrence statistics. CIDEr (Vedantam et al., 2015) represents a sentence as an n-grams vector with tf-idf (term frequency-inverse document frequency), and compute the cosine similarity between reference and candidate.

LEIC (Cui et al., 2018) uses a trained neural model to predict whether a caption is generated by humans. LEIC is trained with COCO images data and uses data augmentation, which helps to achieve a high human correlation. However, LEIC suffers from high computational cost to train in the COCO data. SPICE (Anderson et al., 2016) computes F1 score according to the scene graph created by captions. SPICE reaches a high correlation with human judgment while suffers from long repetitive sentence evaluation (Liu et al., 2017).

2.2 Pre-trained language models and BERTScore

Thanks to the development of a pre-trained language model, better sentence representation can be used in diverse kinds of NLP tasks. Previous works mainly focus on linguistic representation such as word embedding (Mikolov et al., 2013; Pennington et al., 2014; Goldberg and Levy, 2014), which are only word-level embedding without positional information. After the success of Transformer (Vaswani et al., 2017) , a series of language model approaches are proposed such as GPT (Radford et al., 2018), BERT (Devlin et al., 2018), GPT-2 (Radford et al., 2019), XLNET (Yang et al., 2019), XLM (Lample and Conneau, 2019), RoBERTa (Liu et al., 2019). These approaches learn from a huge number of unlabeled text data as a pre-trained process and can fine-tune in downstream tasks with a few epochs.

BERTScore is the latest one-to-one matching metric for text similarity. Benefiting from the contextual embedding of the pretrained language

[1] https://github.com/ck0123/improved-bertscore-for-image-captioning-evaluation

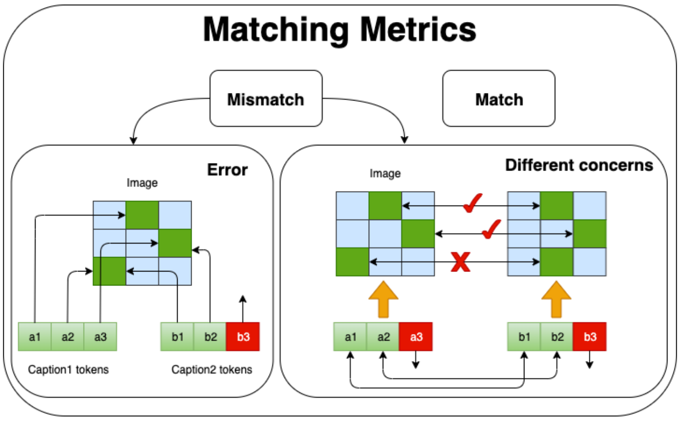

Figure 2: This figure explains the differences between an error and different concerns when mismatches occur. We show an image as nine parts grid chart (3×3). In error case, $caption_1$ focuses on three parts while b_3 cannot represent the part that matches a_1. In different-concerns cases, two captions can represent well in different parts of the same image. However, a mismatch occurs since a_3 and b_3 attend to a different part of the image.

models, BERTScore measures two texts similarity by token-level cosine similarity computing and greedily pick strategy: (1) feed reference text and candidate text into pre-trained model, and extract two contextual embeddings $r = [r_1, .., r_n]$, $c = [c_1, , .., c_m]$; (2) compute the cosine similarity matrix between r and c by $\frac{\mathbf{r} \times \mathbf{c}}{\|\mathbf{r}\| \|\mathbf{c}\|}$; (3) greedily pick the maximum value from cosine similarity matrix for each reference token as a matching value; (4) collect all the matching values with optional inverse document frequency weights.

Inverse document frequency (idf) computes a score for word frequency in the whole corpus. Given N documents $[s_1, .., s_N]$ and each word w, idf score is :

$$\text{idf}(w) = -lg \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[w \in s]$$

where $\mathbb{I}[\cdot]$ is an indicator function and lg is the base 10 logarithm.

The recall of BERTScore (BS for short) is :

$$\text{BS} = \frac{\sum_{r_i \in r} \text{idf}(r_i) \max_{c_j \in c} \mathbf{r}_i^\top \mathbf{c}_j}{\sum_{r_i \in r} \text{idf}(r_i)}$$

BERTScore can adequately deal with different descriptions by knowledge from a pre-trained model and achieves high performance in both machine translation, image captioning evaluation tasks. However, as a one-to-one metric approach, it still suffers from different-concerns problems. Another pitfall in BERTScore comes from the strategy "greedy pick": when no candidate word attends to a specific reference word, this reference word still gets value by picking a maximum cosine value greedily, which causes under-penalization.

Inspired by BERTScore, our metric treats the mismatches between captions carefully, and try to give a proper score for the similarity between captions.

3 Method

Proper scoring for generated captions should consider the information about multi-references and avoid the wrong penalization. In this section, we provide the idea about references combination and fix some under or over penalization issues for cosine similarity-based metrics.

3.1 Preliminary concept of references combination

Token-level mismatches lead to two kinds of problems: different descriptions and different concerns. We introduce these two concepts in Figure 2. Some methods are available for description problems like thesaurus or similarity with contextual embedding, while few of methods handle the different-concerns problem in multi-references cases.

The common ways for one-to-one text metrics to deal with multi-references cases are pooling the results by some strategies like average or maximum. Maximum picks the maximum of results, which can get a higher score than average meanwhile ignores other references directly. Average merges all the results with each reference, which can consider all references. Although average slightly reduce the impact of different concerns, both of the two over-penalize the generated caption since they already

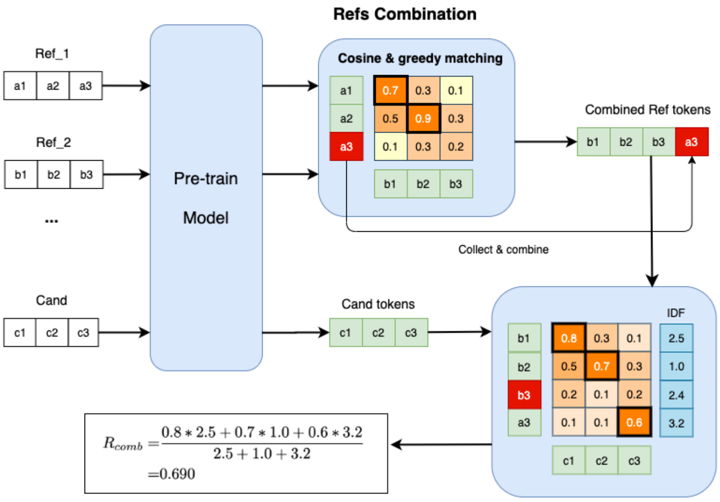

Figure 3: Combination of references comes from a phenomenon that: mismatches between two ground truth captions can't be errors but different concerns. After the greedy matching process, we collect all the mismatch tokens and create a combined contextual embedding as a combined caption. For example, the threshold value β is 0.4, and all the tokens in Ref_2 can't match a_3 with a value bigger than 0.4. After the combination of all references, our metric provides a better recall score between the combined caption and the candidate caption with idf weighted.

regard those mismatches from different concerns as errors during the one-to-one text evaluation process.

Different from average and maximum strategies, the strategy of our metric is to combine reference captions. The combination works based on a fact that: all of the reference captions are ground truth captions so that the mismatches between references should not be errors, but considering different concerns (cosine similarity with contextual embedding also ensures that mismatches are not from errors).

Once we choose a base reference caption and pick up all the mismatches among base and others, the combination among the base and mismatches contains all the information in references without duplicate.

After that, the evaluation between the candidate caption and the combined caption does not suffer from the problems from inter references variance any more.

3.2 Mismatch detection with overlap and cosine similarity

It is hard to define the "differences" between captions clearly. To simplify the problem, we regard mismatches in token-level between two embeddings as differences between two captions.

Mismatch is a concept from n-gram overlap matching metrics like BLEU. We find a mismatch when a word from one sentence cannot be found in the other sentence. Although mismatch is a clear concept to word-level comparison, overlap-based mismatch results in some problems like synonyms. Meanwhile, cosine similarity-based metrics like BERTScore can address this problem quite well. BERTScore uses a pre-trained language model's contextual embedding and regard cosine similarity between two tokens as their similarity. Therefore, the match values change from overlap's discrete value (0 or 1) to cosine's continuous value (0 to 1) with semantic and positional information, which make similarity values more precise.

However, a weakness of cosine similarity is that we cannot distinguish match and mismatch directly since the concept of mismatch does not exist in cosine similarity. To achieve references combination, we simply set a threshold function φ for distinguish the mismatch: when the cosine value is bigger than the threshold, we keep it; otherwise, we set it to 0, which is shown as follows.

$$\varphi(x,\beta) = \begin{cases} x & x > \beta \\ 0.0 & x \leq \beta \end{cases} \tag{1}$$

where x is the cosine value and β is the threshold value.

$\mathcal{S}$ is the improved "greedy pick" function for each r_i reference token with threshold:

$$\mathcal{S}(r_i, c, \beta) = \varphi(\max_{c_j \in c} \mathbf{r}_i^\top \mathbf{c}_j, \beta) \tag{2}$$

where $r = [r_1, .., r_n]$ and $c = [c_1, .., c_m]$ are contextual embedding. We call this process "cut" for removing low cosine values. The standard "cosine & greedy pick" process is a case when the threshold value β equals to 0. Then we can get the greedy recall similarity with threshold indicator:

$$R = \frac{\sum_{r_i \in r} \mathrm{idf}(r_i) \mathcal{S}(r_i, c, \beta)}{\sum_{r_i \in r} \mathrm{idf}(r_i) \mathrm{sgn}(\mathcal{S}(r_i, c, \beta))} \tag{3}$$

where sgn is the sign function.

With a threshold indicator, our metric acquires the ability to detect mismatches. Furthermore, since we cut all the low cosine value, the bad impact of greedy pick (mentioned in Section 2) will be eliminated, which means our metric provides a more reasonable similarity for each token pair.

Empirically for a pre-trained language model, the threshold value in different tasks is similar due to the same architecture and the same widely pretraining process in an ample amount of text data. In this work, we use the threshold value 0.4 for BERT (base) and 0.83 for RoBERTa (large) as the recommended settings.

3.3 The combination of references

Contextual embeddings are extracted from the pretrained model. Since the inputs of the model contain both token embedding and position embedding, contextual embedding for each token also contains its semantic and positional information. Therefore, the change of tokens' position does not change the inner positional information for each token. For example, [**embed A**, **embed B**] is the contextual embedding sentence generated from [**word A**, **word B**]. Both [**embed A**, **embed B**] and [**embed B**, **embed A**] (only switch tokens' position) still provide same positional information.

Using this characteristic, we can now easily combine all of the references with the following steps: (1) choose a random reference caption embedding as a base, A; (2) compute the similarity between A and another reference B with a threshold; (3) collect those tokens from B that mismatch comparing with A, B'; (4) concatenate A and B' as a new base caption A; (5) repeat steps above until used all the references;

R_{comb} computes the recall score for combined reference and candidate. Figure 3 shows references combination $Comb$ and the computation of R_{comb}.

$$R_{comb} = R(Comb([r_1, .., r_M]), c) \tag{4}$$

where M is the number of references.

3.4 Importance of different words

For proper scoring, our metric also focuses on a problem that token-level matching sometimes does not mean similarity between captions. *A bird standing on the blue handrail* and *A bird flying on the blue sky* are describing different images with only two words different but five words the same. The meaning of a caption is sensitive to the replacement of essential components like subject, predicate, object, while some replacement (like **a** → **the**) are not.

The problem is: in matching metric, we only focus on the match and mismatch while ignoring the importance of each word in the sentence. It is hard to provide optimal importance with each word and pick the important ones; in contrast, the removal of unimportant words is more comfortable to achieve.

In this work, our metric removes all the stop words and computes an another greedy cosine score as an additional score without idf weight, R_{rm}:

$$R_{rm} = \frac{\sum_{r_i \in r'} \mathcal{S}(r_i, c', \beta)}{|r'|} \tag{5}$$

where r' and c' are embeddings without stop words and $|r'|$ means the length of sentence r' .

Although taking idf weight into consideration is convenient, using the stop word removal additionally is still necessary. The definition of idf points out that idf is an indicator of frequency, while frequency does not equate to importance. Take COCO caption corpus as an example: all the idf weights of common subjects are low such as **man**, **dog**, **girl**, etc; while those of **playfully**, **sleepy** are high. However, there is no doubt that mismatches occur in these common subjects will change the meaning dramatically.

3.5 Summary and metric formula

In this section, we discussed the mismatches between references, under-penalization of "greedy pick", and the importance of words. Moreover, we showed our idea about captions combination, greedy recall similarity with threshold indicator, and stop word removal. Including all of formulas above, the final expression of our metric is the product of R_{comb} and R_{rm}:

$$Score = R_{comb} \times R_{rm} \tag{6}$$

Type	Metric	M1	M2
Task-agnostic	ROUGE-L	0.062 (0.846)	0.215 (0.503)
	BLEU-1	0.029 (0.927)	0.165 (0.607)
	BLEU-4	0.236 (0.459)	0.380 (0.222)
	CIDEr	0.440 (0.151)	0.539 (0.071)
	METEOR	0.703 (0.011)	0.701 (0.011)
	BS (BERT-base)	0.807 (0.001)	0.735 (0.006)
	BS (RoBERTa-large)	**0.873** (0.000)	**0.841** (0.000)
	Ours (BERT)	0.875 (0.000)	0.797 (0.002)
	Ours (RoBERTa)	**0.932** (0.000)	**0.869** (0.000)
Task-specific	SPICE	0.715 (0.009)	0.688 (0.013)
	LEIC	**0.939*** (0.000)	**0.949*** (0.000)

Table 1: Pearson correlation of system level metrics scores with human judgment in 2015 COCO Captioning Challenge. We use 12 teams results on validation set with "Karpathy split". M1: the percentage of captions that are evaluated as better or equal to human captions; M2: the percentage of captions that are indistinguishable from human caption. BS means BERTScore and score with * are cited from (Cui et al., 2018).

4 Experiments

The most convincing way for metric evaluation is the human correlation in caption-level and system-level tasks. In this section, we evaluate our metric in four typical image captioning evaluation datasets with standard metrics. We also consider the impact of each part in our metric by ablation experiment and key part replacements.

4.1 Dataset

Microsoft COCO 2014 COCO dataset contains 123,293 images with 82,783 images in training set, 40,504 images in the validation set and 40,775 images in the test set. Each image has five human-annotated captions as ground truth captions.

In 2015 COCO Captioning Challenge (Chen et al., 2015), submissions of the challenge are evaluated by human judgments with five kinds of metrics: M1, percentage of captions that are evaluated as better or equal to human caption; M2, percentage of captions that pass the Turing Test; M3, average correctness of the captions on a scale 1-5 (incorrect - correct); M4, the average amount of detail of the captions on a scale 1-5 (lack of details - very detailed); M5, percentage of captions that are similar to human description.

Flickr 8K Flickr 8K dataset contains 8,092 images with five human-generated captions for each image. Flickr 8K provides an annotation called Expert Annotation, and each row contains one image, one candidate caption from Flickr 8K dataset (it may matches this image or not), and three expert scores for the image-caption pair. Scores range from 1: indicating that the caption does not describe the image at all to 4: indicating that the caption describes the image.

COMPOSITE The COMPOSITE dataset contains 11985 human judgments from Flickr 8K, Flickr 30K, and COCO captions re-coined. Candidate captions come from human and two caption models scoring by Amazon Mechanical Turk (AMT) workers. All the captions score a 5-point scale from 1 (The description has no relevance to the image) to 5 (The description relates perfectly to the image).

PASCAL-50S PASCAL-50S dataset has 1000 images from UIUC PASCAL Sentence Dataset, and each image has 50 reference captions annotated by AMT worker. PASCAL-50S includes over 4000 candidate captions pair with human judgments. Different from COCO and Flickr format, PASCAL-50S consists of the triplet: $\langle A, B, C\rangle$. A is the reference sentence from an image, and B, C are two candidate sentences. AMT workers are asked *Which of the two sentences, B or C, is more similar to A?*. This kind of question is more accessible for workers to judge than provide correct scores. Candidate sentences come from human-written, or model generated, and four kinds of paired ways: human-correct (HC), human-incorrect (HI), human-model (HM), and model-model (MM).

Metric	M1	M2
BLEU-1	0.307	0.338
ROUGE-L	0.062	0.215
ROUGE-L (cat)	0.096	0.180
BLEU-1 (cat)	0.351	0.175
METEOR (cat)	0.662	0.571
METEOR	**0.703**	**0.701**
BS(BERT-base)	0.807	0.735
Ours (Unigram)	0.809	0.714
Ours (BERT+T)	0.822	0.741
Ours (BERT+T+cat+R)	0.857	0.783
Ours (BERT+TB)	0.867	0.755
Ours (BERT+TBR)	**0.875**	**0.797**

Table 2: We provide the ablation study and the replacement study in 2015 COCO Captioning dataset. As an additional experiment, we also compare concatenation with average in some standard metrics like: BLEU-1, ROUGE-L and METEOR. Abbreviation: **BS** means BERTScore, **T** means cut, **B** means combine, **R** means remove, **cat** means concatenation of references.

	Flickr 8K	COMPOSITE
BLEU-1	0.318	0.282
BLEU-4	0.140	0.199
ROUGE-L	0.323	0.313
BS (RoBERTa)	0.367	0.392
BS (BERT)	0.393	0.399
METEOR	0.436	0.381
CIDEr	0.447	0.387
SPICE	0.458	0.418
LEIC	0.466*	-
Ours (RoBERTa)	0.451	**0.449**
Ours (Unigram)	0.471	0.420
Ours (BERT)	**0.481**	0.423
Inter-human	0.736*	-

Table 3: In caption-level experiments, we compute the Kendall correlation between human judgments and scores of metrics. Two dataset results are given: Flickr 8K and COMPOSITE. Both our unigram metric and BERT based metric outperform other metrics. Scores with * are cited from (Cui et al., 2018)

4.2 Compared Metrics

For comparison, we use common standard metrics in our scoring tasks, such as BLEU-1, ROUGE-L, METEOR, CIDEr, and SPICE. All these metrics are implemented in MS COCO evaluation tool.[2] We also use the original BERTscore to check the improvement of our metrics. To be more convincing, we compare with the current SOTA training-based approach LEIC in COCO captioning 2015 and Flickr 8K.

4.3 Baselines

Two metrics are implemented as baselines: (1) unigram overlap matching metric and (2) references concatenation metric with BERT. Unigram overlap matching metric is implemented for verifying the importance of contextual embedding from the pre-trained language model. References concatenation metric with BERT is implemented for verifying the importance of references combination.

Unigram overlap matching metric. In our unigram overlap matching metric, we remove contextual embedding from the pre-trained language model and only use unigram overlap matching. Different from continuous value methods like BERTScore, it is easy for overlap matching to distinguish the match and mismatch (1 or 0). In combination part, we collect all the mismatch words and combine them with the base reference caption. To reduce the impact of unimportant words, we remove stop words from the combined caption directly.

References concatenation. We also regard the concatenation of references as another baseline comparing with our combination method. The concatenation of references combines all the information from references as well. The difference between concatenation and our combination is the duplicate tokens in majority references. In this metric, we follow all the steps of our metric with BERT, except the combination.

4.4 System-Level Correlation

In system-level evaluation, we use twelve teams of human judgment results from COCO 2015 Captioning Challenge. We use data from "Karpathy splits" (Karpathy and Fei-Fei, 2015), which contains 113,287 train images, 5000 test images, and 5000 validation images. Each image has 5 references human captions. Following prior works (Anderson et al., 2016; Cui et al., 2018), we compute the Pearson correlation with human judgment. In the pre-trained model selection for BERTScore, we choose BERT (base), which is the most common model in the set of transformer language models, and RoBERTa (large), which is an optimized ver-

[2] https://github.com/tylin/coco-caption

	HC	HI	HM	MM	All
BLEU-1	53.1	94.7	90.9	56.9	73.9
BLEU-4	53.3	92.8	85.2	60.5	73.0
ROUGE-L	55.6	95.1	93.3	57.7	75.4
METEOR	61.4	97.2	94.9	64.5	79.5
CIDEr	55.0	98.0	91.0	64.6	77.2
SPICE	57.7	96.1	88.3	**65.3**	76.9
Ours (RBT)	62.5	97.7	95.0	59.4	78.7
BS (BERT)	64.4	97.9	96.6	59.0	79.5
Ours (BERT)	**65.4**	**98.1**	**96.4**	60.3	**80.1**

Table 4: In PASCAL-50S, candidate sentences come from human written or model generated. There are 4 kinds of paired ways: human-correct (HC), human-incorrect (HI), human-model (HM), and model-model (MM). Ours (BERT) outperforms in HC, HI and HM. Abbreviation: **RBT** means RoBERTa.

sion of BERT.

The experimental results in Table 1 show that our metrics with both BERT and with RoBERTa perform better than BERTScore and other standard metrics. What is more, our metric with RoBERTa can reach a high correlation of 0.932 with human judgment, which is even close to the training-based task-specific metric LEIC with image features.

4.5 Ablation and replacement

To check the influence of each part, we provide both ablation study and replacement study in 2015 COCO Captioning dataset. The results are showed in Table 2.

In ablation study, we use our metric with BERT and remove **remove**, **combine** and **cut** one by one. The result shows that each part of our metric is useful, and **combine** is the most influential part.

In the replacement study, we compare our metric with the unigram metric and concatenation metric to check the influence of contextual embedding and combination. The comparison between Ours (Unigram) and Ours (BERT+TBR) shows that contextual embedding is better than unigram matching in the system-level correlation task. The comparison between Ours (BERT+T+cat+R) and Ours (BERT+TBR) shows that the combination process is better than concatenation directly. Furthermore, we also show the comparison between concatenation and average in some standard metrics.

Model	Ours (BERT)	CIDEr-D
AoAnet	0.3529	1.296
M2-Transformer	0.3481	1.321
SAT	0.3296	0.893
CNN+LSTM	0.3055	0.946
NeuralTalk	0.2845	0.692

Table 5: We present some results on current state-of-the-art models (M2-Transformer and AoAnet) for image captioning models with respect to CIDEr-D. The experimental results show that: on both our metric and CIDEr-D, current models perform better. Abbreviation: **SAT** means Show, Attend and Tell.

4.6 Caption-Level Correlation

In caption-level evaluation tasks, we compute Kendall's correlation (Kendall, 1938) between metrics results and expert judgments.

In Flickr 8K, we use Expert Annotation with 5822 samples, including 158 correct image-caption pairs where the candidate caption equals one of captions in references set. Following the prior work (Anderson et al., 2016), we use 5664 samples and exclude those correct image-caption pairs. In COMPOSITE, captions are estimated by two kinds of standards: correctness and throughness, and we only focus on correctnesss in this work.

The experimental results in Table 3 show that our metric is quite suitable for caption-level evaluation in image captioning. Our metric outperforms other metrics (including training-based metric LEIC in Flickr 8K). Another interesting fact is that the unigram metric also has high performance in caption-level correlation tasks. In COMPOSITE, our unigram metric is comparable to our metric with BERT.

In PASCAL-50S, we use five references for metrics computation, which is comparable with previous experiments. The results in Table 4 show that in four kinds of caption pairs, our metric performs better than others in human-correct (HC), human-incorrect (HI), human-model (HM) classification.

5 More model results on our metric

In table 5, We evaluate some current state-of-the-art image captioning models reported from Codalab competition: Meshed-Memory-Transformer (Cornia et al., 2020), AoAnet (Huang et al., 2019).[3] Some of models in 2015 COCO Captioning Chal-

[3] https://competitions.codalab.org/competitions/3221

lenge are listed for comparison: (1) Show, Attend and Tell (Xu et al., 2015); (2) CNN+LSTM (Vinyals et al., 2015); (3) NeuralTalk (Karpathy and Fei-Fei, 2015). The result shows that: on our metric, current models perform better than previous models. It is worth noting that different judgments exist between AoAnet and M2-Transformer on our metric and CIDEr-D. According to our observation, several captions (1558/5000) generated by M2-Transformer are incomplete, like *a bedroom with a bed and a tv in a* or *a wooden door with a skateboard on a*. It may explain why M2-Transformer is a little worse than AoAnet on our metric.

6 Conclusion

In this work, we study the intrinsic variance among ground truth captions in image captioning evaluation. We propose an improved matching metrics based on BERTScore, which can combine all of the references for taking full advantage of multi-references. Our metric also benefits from stop word removal by reducing the impact of stop words. The experimental results show that our metric can reach state-of-the-art human correlation in several evaluation tasks.

References

Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. 2015. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *arXiv preprint arXiv:1511.03292*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer.

X. Chen, H. Fang, TY Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv:1504.00325*.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-Memory Transformer for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2970–2979.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yoav Goldberg and Omer Levy. 2014. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *International Conference on Computer Vision*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2017. Captioning images with diverse objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5753–5761.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.