ViLBERTScore: Evaluating Image Caption Using Vision-and-Language BERT

Hwanhee Lee[1], **Seunghyun Yoon**[1,2], **Franck Dernoncourt**[2]
Doo Soon Kim[2], **Trung Bui**[2] and **Kyomin Jung**[1]
[1]Dept. of Electrical and Computer Engineering, Seoul National University, Seoul, Korea
[2]Adobe Research, San Jose, CA, USA
{wanted1007, mysmilesh, kjung}@snu.ac.kr
{franck.dernoncourt, dkim, bui}@adobe.com

Abstract

In this paper, we propose an evaluation metric for image captioning systems using both image and text information. Unlike the previous methods that rely on textual representations in evaluating the caption, our approach uses visiolinguistic representations. The proposed method generates image-conditioned embeddings for each token using ViLBERT from both generated and reference texts. Then, these contextual embeddings from each of the two sentence-pair are compared to compute the similarity score. Experimental results on three benchmark datasets show that our method correlates significantly better with human judgments than all existing metrics.

1 Introduction

Image captioning is a task that aims to generate a text that describes a given image. While there have been many advances for caption generation algorithms (Vinyals et al., 2015; Anderson et al., 2018) and target datasets (Fang et al., 2015; Sharma et al., 2018), few studies have focused on assessing the quality of the generated captions with consideration to the image.

Most of the previous studies on evaluating image captioning tasks rely on n-gram similarity metrics such as BLEU (Papineni et al., 2002) or CIDEr (Vedantam et al., 2015). These approaches bear limitations in dealing with the text's diverse nature, similarly found in other text generation tasks (e.g., abstractive summarization and dialog) (Kryscinski et al., 2019; Liu et al., 2016). To alleviate the issues in the n-gram based approaches, researchers proposed word embedding-based techniques (Kusner et al., 2015; Zhang et al., 2019; Zhao et al., 2019; Lo, 2019; Clark et al., 2019). These techniques shows robust performance and achieve higher correlation with human judgment than that of other previous metrics in many text generation tasks, including image captioning. Especially, BERTScore (Zhang et al., 2019) shows that using contextualized embedding is effective for evaluating the text. As BERTScore does not utilize image content, it is still undiscovered how to effectively utilize the image content in the process of evaluating the captions.

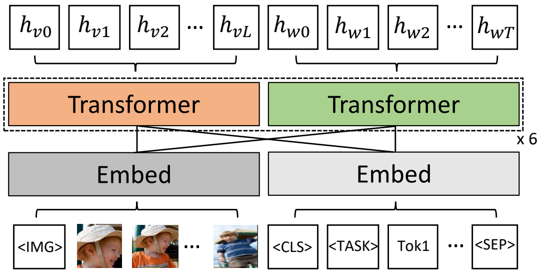

Figure 1: The overall architecture of ViLBERT. ViLBERT consists of a self-attention based embedding layer and co-attention layer for each image and text information.

To further reflect image context while utilizing the advantages of BERTScore, we propose ViLBERTScore[1] by employing the ViLBERT (Lu et al., 2019), which is a task-agnostic pre-trained visiolinguistic representation. ViLBERTScore computes cosine similarity between token embeddings for reference and candidate sentences similar to BERTScore. However, different from BERTScore, the token embedding is computed with the consideration of image contexts.

We evaluate our proposed method on three benchmark datasets (i.e., Composite, Flickr8k, and PASCAL-50S). Extensive experiments show that ViLBERTScore achieves a significantly higher correlation with human judgments than previous metrics. This result demonstrates that the use of contextualized embedding from vision and language is

[1] https://github.com/hwanheelee1993/ViLBERTScore

Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP), pages 34–39,
November 20, 2020. ©2020 Association for Computational Linguistics

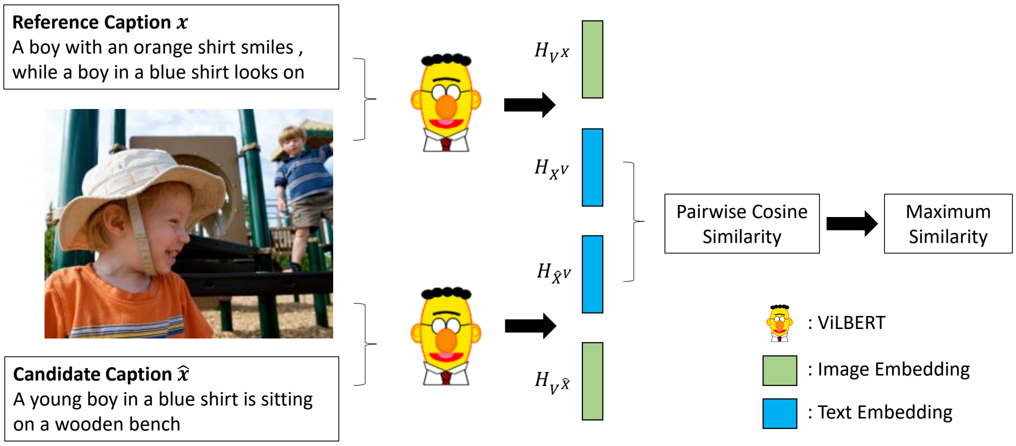
Reference Caption x
A boy with an orange shirt smiles , while a boy in a blue shirt looks on
Candidate Caption $\hat{x}$
A young boy in a blue shirt is sitting on a wooden bench
H_{V^x}
H_{X^V}
$H_{\hat{X}^V}$
$H_{V^{\hat{x}}}$
Pairwise Cosine Similarity
Maximum Similarity
: ViLBERT
🟩 : Image Embedding
🟦 : Text Embedding

Figure 2: Overall computation of ViLBERTScore. Given the image I, reference caption x and candidate caption $\hat{x}$, we compute contextual embeddings with ViLBERT for x and $\hat{x}$ respectively. Then, we extract the text embeddings H_{X^V} and $H_{\hat{X}^V}$ for each output embedding. Finally, we compute the pairwise cosine similarity between H_{X^V} and $H_{\hat{X}^V}$ as in (Zhang et al., 2019).

effective in evaluating image captioning tasks.

2 Related Work

2.1 Caption Evaluation

We provide a summary of the widely used metrics for evaluating image captions such as n-gram similarity metrics, embedding based metrics, and other task-specific metrics for captioning.

N-gram Similarity Metrics The most widely used metrics for evaluating the quality of text generation tasks are n-gram similarity metrics that compute the exact number of n-gram matches between reference and generated text. One example of these metrics is BLEU (Papineni et al., 2002) that computes the precision of overlap n-gram between reference and candidate. ROUGE (Lin, 2004) is a set of commonly used metrics for text summarization. In particular, ROUGE-N, the longest common subsequence based metric, is the most frequently used variants of ROUGE. CIDEr (Vedantam et al., 2015), which is proposed for evaluating image captions, computes the tf-idf weighted n-gram similarity between reference and candidate.

Embedding Based Metrics The n-gram similarity metrics possess critical limitations; they cannot count the synonym matches of the n-gram, even though the synonyms are widely found in the generated text. To overcome this weakness, embedding based metrics such as Word Mover Distance(WMD) (Kusner et al., 2015) and BERTScore (Zhang et al., 2019) are proposed. WMD computes minimum transportation distance among tokens using pre-trained word embeddings (i.e., GloVe (Pennington et al., 2014)). On the other hand, BERTScore computes cosine similarity among tokens using contextual embeddings from BERT (Devlin et al., 2019).

Captioning Specific Metrics After CIDEr is introduced, several metrics for image captioning are proposed. SPICE (Anderson et al., 2016) uses scene graph and LEIC (Cui et al., 2018) uses the trainable model to evaluate the captions. VIFIDEL (Madhyastha et al., 2019) is an extension of Wasserstein distance that utilizes the information from detected objects in the image. TIGEr (Jiang et al., 2019) uses the output of the visual grounding task. BERT-TBR (Yi et al., 2020) focuses on the variance of the captions and combine multiple reference captions to get improved BERTScore.

2.2 ViLBERT

To compute contextual representations from the visually-grounded text, researchers proposed a transformer-based model. One such example is ViLBERT (Lu et al., 2019), which is a task-agnostic pre-trained representation for vision and language. As shown in Fig. 1, ViLBERT employs two streams of transformer (Vaswani et al., 2017)-based architecture; one of each part processes visual and textual inputs, respectively. Specifically, the image and grounded-text inputs are fed into separate embedding layers; followed by two co-attentional transformer block that allows interaction between the two modalities. ViLBERT is pre-trained with two training objectives, masked multi-modal modeling, and multi-modal alignment. Lu et al. (2019) show that fine-tuning this pre-trained ViLBERT to vision-and-language related downstream tasks (e.g., visual question answering (Antol et al., 2015)) significantly outperforms previous approaches. Recently, Lu et al. (2020) investigate and reveal that training the ViLBERT with multi-task learning objectives provides further performance improvement for most of the vision and language tasks.

3 ViLBERTScore

We propose ViLBERTScore, a metric that utilizes visually-grounded representations for each token. The overall flow of our proposed ViLBERTScore is described in Fig. 2. Similar to BERTScore, we first compute contextual embeddings of both reference caption $X = (x_1, ..., x_n)$ and candidate caption $\hat{X}$ = $(\hat{x}_1, ..., \hat{x}_m)$. Since we use ViLBERT, we compute the embeddings for each caption conditioning with the target image I. For the target image, we extract N region-level features $V = (v_1, ..., v_N)$ using pre-trained object detection model (see 4.2 for detailed information). Then, we feed each pair of image and caption embeddings (X, V), $(\hat{X}, V)$ to pre-trained ViLBERT and compute the contextual embeddings (H_{VX}, H_{XV}) and $(H_{V\hat{X}}, H_{\hat{X}V})$. Note that H_V and H_X are image and text embeddings, respectively. Among these embeddings, we only utilize the text embeddings, $H_{XV} = (h_{w0}, ..., h_{wT})$ and $H_{\hat{X}V} = (\hat{h}_{w0}, ..., \hat{h}_{wT})$, and compute cosine similarity among the pair of tokens from the candidate and reference caption. Finally, the greedy matching process is exercised to the pair of tokens mentioned above for finding the most similar token-match between two sentences. We can formulate ViLBERTScore as follows.

$$\text{ViLBERTScore}_P = \frac{\Sigma_{i=1}^{m} \max_{\hat{h}_{wj} \in H_{\hat{X}V}} \mathbf{h}_{\mathbf{wi}}^{\mathbf{T}} \hat{\mathbf{h}}_{\mathbf{wj}}}{m} \quad (1)$$

$$\text{ViLBERTScore}_R = \frac{\Sigma_{i=1}^{n} \max_{h_{wj} \in H_{XV}} \hat{\mathbf{h}}_{\mathbf{wi}}^{\mathbf{T}} \mathbf{h}_{\mathbf{wj}}}{n} \quad (2)$$

$$\text{ViLBERTScore}_F = 2 \cdot \frac{\text{ViLBERTScore}_P \cdot \text{ViLBERTScore}_R}{\text{ViLBERTScore}_P + \text{ViLBERTScore}_R} \quad (3)$$

4 Experiments

4.1 Dataset

Composite Composite (Aditya et al., 2015) dataset consists of 11,985 human judgments for each candidate caption and image pair. The images in this dataset are from Flickr8k (Hodosh et al., 2013), Flickr30k (Plummer et al., 2017), and COCO captions (Lin et al., 2014). The human judgments scores range from 1 to 5, depending on the relevance between candidate caption and image.

Metric	Flickr8k	Composite
BLEU-1†	0.318	0.282
BLEU-4†	0.140	0.199
ROUGE-L†	0.323	0.313
METEOR†	0.436	0.381
CIDEr†	0.447	0.387
SPICE†	0.458	0.418
BERTScore†	0.393	0.399
BERT-TBR†	0.481	0.423
ViLBERTScore$_{\mathrm{P}}$	0.462	0.366
ViLBERTScore$_{\mathrm{R}}$	0.432	0.424
ViLBERTScore$_{\mathrm{F}}$	0.514	0.420
ViLBERTScore*$_{\mathrm{P}}$	0.541	0.499
ViLBERTScore*$_{\mathrm{R}}$	0.512	0.508
ViLBERTScore*$_{\mathrm{F}}$	**0.542**	**0.514**

Table 1: Kendall Correlation between human judgments and various metrics. Note that ViLBERTScore* uses the ViLBERT model from (Lu et al., 2020), which is fine-tuned on 12 downstream tasks. Scores with † are cited from (Yi et al., 2020).

Metric	HC	HI	HM	MM	All
BLEU-1	54.5	95.0	92.0	57.7	74.8
BLEU-4	51.8	92.3	86.9	59.3	72.6
ROUGE-L	53.4	94.3	93.8	57.2	74.7
METEOR	56.3	96.9	95.1	61.2	77.4
CIDEr	53.1	98.1	92.5	63.1	76.7
SPICE	59.7	95.1	87.2	61.6	75.9
ViLBERTScore$_{\mathrm{P}}$	43.4	95.3	75.4	67.7	70.4
ViLBERTScore$_{\mathrm{R}}$	**66.5**	99.2	**98.3**	61.1	81.3
ViLBERTScore$_{\mathrm{F}}$	50.3	98.1	91.4	69.6	77.4
ViLBERTScore*$_{\mathrm{P}}$	46.0	99.5	86.2	75.3	76.8
ViLBERTScore*$_{\mathrm{R}}$	61.4	**100.0**	97.1	75.0	**83.4**
ViLBERTScore*$_{\mathrm{F}}$	49.9	99.6	93.1	**75.8**	79.6

Table 2: Result for PASCAL-50S dataset. The paired ways HC, HI, HM and MM respectively mean human-correct, human-incorrect, human-model and model-model. We use five reference captions among 50 reference captions for each caption pair.

Flickr8k Flickr8k dataset is composed of 8,092 images with five corresponding human-generated captions. This dataset also provides three expert annotations for each image and candidate caption on 5,822 images. The score ranges from 1 to 4, depending on how well the caption and image match.

PASCAL-50S PASCAL-50S (Vedantam et al., 2015) dataset contains 1,000 images from UIUC PASCAL Sentence Dataset with 50 reference cap-

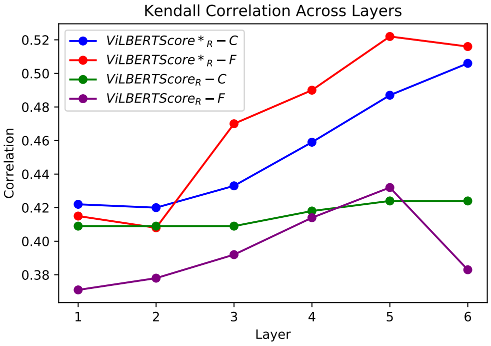

Figure 3: Kendall Correlation between human judgments across different layers. C and F are the results for Composite and Flickr8k datasets, respectively. Note that ViLBERTScore* uses the fine-tuned ViLBERT model from (Lu et al., 2020).

tions generated by humans for each image. Different from other datasets, this dataset provides 4,000 caption triplet <A, B, C> composed of 50 reference captions(A) and two candidate captions(B, C) for the given image. There are human annotated answers to which is more similar to "A", "B" or "C". Candidate captions are human-written or model-generated.

4.2 Implementation Details

We use two versions of ViLBERT, one from the pre-trained ViLBERT model from (Lu et al., 2019) and the other version from (Lu et al., 2020) that are fine-tuned on 12 downstream tasks. We set N = 100 boxes for each image using image detectron model (He et al., 2017) to compute contextual embedding as in (Lu et al., 2019). We use the textual representations in the 6-th layer, the last co-attention layer, of ViLBERT for the main results in Table 1 and Table 2. For the dataset containing multiple reference captions, we average the score over the pairs of candidate caption and reference captions.

4.3 Results

Evaluation Methods We compute Kendall's correlation coefficient with human judgments for the Composite dataset and Flickr8k dataset. For the PASCAL-50S dataset, we compute the number of matches between human judgments for each candidate caption pair.

Performance Comparison We present the correlation scores for the baseline metrics and our proposed ViLBERTScore for Composite dataset and Flickr8k dataset in Table 1. ViLBERTScore shows a higher correlation than all the existing metrics. For the PASCAL-50S dataset, Table 2 shows that ViLBERTScore$_\mathrm{R}$ is the best metric at comparing captions among all of the metrics. Interestingly, we observe that the performance of ViLBERTScore$_\mathrm{P}$ is lower than that of ViLBERTScore$_\mathrm{R}$ for the PASCAL-50S dataset. This is consistent behavior with the results of (Zhang et al., 2019). We speculate that the main objects in the image are the most critical words the human judgments as in (Zhang et al., 2019).

We further explore the performance of ViLBERTScore with different base model. We choose another ViLBERT model that is fine-tuned on 12 vision-and-language related tasks (see ViLBERTScore* in Table 1 and 2). This model shows better results than ViLBERTScore. We explain that some of the tasks such as image retrieval or visual entailment (Xie et al., 2019) are related to caption evaluation.

Correlation Across Layers The co-attentional block in ViLBERT is composed of six layers. To verify the effectiveness of each layer in computing the contextualized embedding of the data, we compute ViLBERTScore using the outputs of different layer. As shown in Fig. 3, the outputs of a higher layer show a better correlation with human judgments than the lower layer except for the last layer. This observation reveals that blending information among the modalities is essential in computing better contextual representations. We explain that the correlation drops in the last layer because the last layer has task specific property.

5 Conclusion

In this paper, we propose ViLBERTScore, a metric for image captioning task by using pre-trained visio-linguistic representations. Different from the BERTScore, ViLBERTScore utilizes image conditional embeddings for each token which is critical in evaluating vision-language combined task. Empirical results on Composite, Flickr8k, and PASCAL-50S datasets show that the proposed ViLBERTScore correlates better with human judgments than all of the previous metrics.

Acknowledgements

We also gratefully acknowledge support from Adobe Inc. in the form of a generous gift to Seoul

National University. K. Jung is with ASRI, Seoul National University, Korea. This work was supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Program (No.10073144).

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.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Elizabeth Clark, Asli Celikyilmaz, and Noah A Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

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.

Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 2019. Tiger: Text-to-image grounding for image caption evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2141–2152.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551.

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

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.

Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.

Chi-kiu Lo. 2019. Yisi-a unified semantic mt quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10437–10446.

Pranava Swaroop Madhyastha, Josiah Wang, and Lucia Specia. 2019. Vifidel: Evaluating the visual fidelity of image descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6539–6550.

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.

Jeffrey Pennington, Richard Socher, and Christopher D 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.

Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1):74–93.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

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.

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.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Yanzhi Yi, Hangyu Deng, and Jinglu Hu. 2020. Improving image captioning evaluation by considering inter references variance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 985–994.

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

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578.