

REO-Relevance, Extraness, Omission: A Fine-grained Evaluation for Image Captioning

Ming Jiang¹, Junjie Hu², Qiuyuan Huang³, Lei Zhang³, Jana Diesner¹, Jianfeng Gao³

¹University of Illinois at Urbana-Champaign, ²Carnegie Mellon University

³Microsoft Research, Redmond

{mjiang17, jdiesner}@illinois.edu, junjieh@cs.cmu.edu
{qihua, leizhang, jfgao}@microsoft.com

Abstract

Popular metrics used for evaluating image captioning systems, such as BLEU and CIDEr, provide a single score to gauge the system’s overall effectiveness. This score is often not informative enough to indicate what specific errors are made by a given system. In this study, we present a fine-grained evaluation method REO for automatically measuring the performance of image captioning systems. REO assesses the quality of captions from three perspectives: 1) **R**elevance to the ground truth, 2) **E**xtraness of the content that is irrelevant to the ground truth, and 3) **O**mission of the elements in the images and human references. Experiments on three benchmark datasets demonstrate that our method achieves a higher consistency with human judgments and provides more intuitive evaluation results than alternative metrics.¹

1 Introduction

Image captioning is an interdisciplinary task that aims to automatically generate a text description for a given image. The task is fundamental to a wide range of applications, including image retrieval (Rui et al., 1999) and vision language navigation (Wang et al., 2019). Though remarkable progress has been made (Gan et al., 2017; Karpathy and Li, 2015), the automatic evaluation of image captioning systems remains a challenge, particularly with respect to quantifying the generation errors made by these systems (Bernardi et al., 2016).

Existing metrics for caption evaluation can be grouped into two categories: 1) rule-based metrics (Papineni et al., 2002; Vedantam et al., 2015) that are based on exact string matching, and 2) learning-based metrics (Cui et al., 2018; Sharif

¹Code is released at <https://github.com/SeleenaJM/CapEval>.



References

- A man instructing a group of kids on a soccer field.
- A soccer coach is instructing the children on the field.
- Pair of adult males with group of small children with soccer balls.

Candidate A: Two men are teaching kids football <i>in a shopping mall with many stores.</i>			Candidate B: Two men are teaching football. (<i>missing "kids"</i>)		
BLEU4: 0.00	CIDEr: 0.05	SPICE: 0.05	BLEU4: 0.00	CIDEr: 0.00	SPICE: 0.06
REO_R: 0.28	REO_E: 9.39	REO_O: 10.64	REO_R: 0.26	REO_E: 10.40	REO_O: 9.32

Figure 1: An example of caption evaluation. Given two caption candidates, even though Caption A covers more image information than Caption B (e.g., missing “kids”), Caption A contains extra irrelevant description (highlighted in red). Prior metrics (e.g., BLEU4) only provide an overall quality score, which is difficult to infer specific description mistakes in a caption. In contrast, REO provides three indicators (i.e., relevance, extraness, and omission) that can properly achieve a fine-grained assessment for each caption.

et al., 2018) that predict the probability of a testing caption as a human-generated caption by using a learning model. In general, prior work has shown that description adequacy with respect to the ground truth data is a main concern for evaluating text generation systems (Gatt and Krahmer, 2018). Though this aspect has been emphasized by prior work for assessing image captions (Papineni et al., 2002; Banerjee and Lavie, 2005; Gao et al., 2019), one common limitation of existing metrics is the lack of interpretability to the description errors because existing metrics only provide a composite score for the caption quality. Without fine-grained analysis, the developers may not be able to understand the specific description errors made by their developed captioning systems.

To fill this gap, we propose an evaluation method called **REO** that considers three specific pieces of information for measuring each caption with respect to: 1) **R**elevance: relevant information of a candidate caption with respect to the ground truth, 2) **E**xtraness: extra information of a

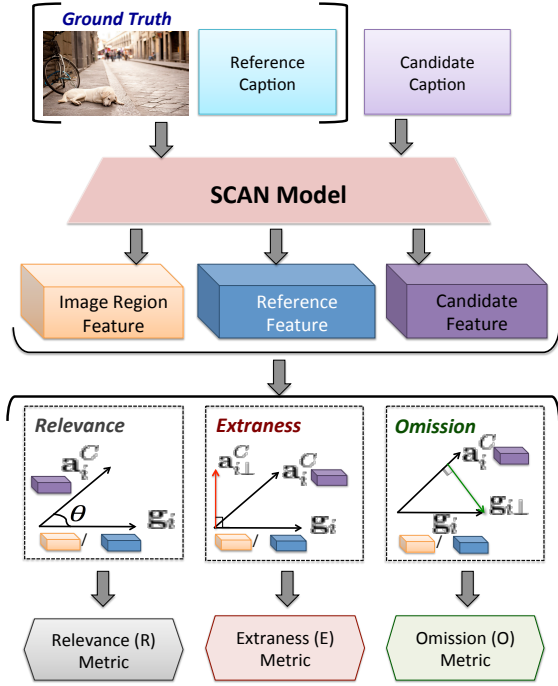


Figure 2: Overview of REO evaluation for image captioning: (1) Feature Extraction; (2) Metric Measurement.

candidate caption beyond ground truth data, and 3) **Omission**: missing information that a candidate fails to describe from an image and human-generated reference captions. Figure 1 shows a comparison between existing metrics and our proposed metrics that measure caption quality at a fine-grained level. If we view caption generation as a process of decoding the information embedded in an image, we can evaluate an image captioning system by measuring the effectiveness of the decoding process in terms of the relevance of the decoded information regarding the image content, and the amount of missing or extra information. Using both the images and reference captions as ground truth information for evaluation, our approach is built based on a shared image-text embedding space defined by a grounding model that has been pre-trained on a large benchmark dataset. Given a pair of vectors representing a candidate caption and its ground truth (i.e., the target image and associated reference captions), respectively, we compute the relevance of the candidate caption and ground truth based on vector similarities. By applying vector orthogonal projection, we identify the extra and missing information carried by the candidate caption. Each aspect that we consider here (i.e., relevance, extraneous, omission) is measured by an independent score.

We test our method on three datasets. The experimental results show that our proposed metrics are more consistent with human evaluations than alternative metrics. Interestingly, our study finds that human annotators pay more attention to extra or missing information in a caption (i.e., false positive and false negatives) than the caption’s relevance for the given image (true positives). We also find that considering both image and references as ground truth information is more helpful for caption evaluation than considering image or references individually.

2 Methods

Figure 2 provides an overview for the calculation of REO, which happens in two stages. The first stage is *feature extraction*, where we aim to obtain feature vectors to encode the candidate caption C and corresponding ground truth G for further comparisons. The second stage is to measure *three metric scores*. Specifically, we measure relevance using standard cosine similarity. To measure irrelevance (i.e., extraneous and omissions), we compare the information carried by C and G , respectively. We will give a detailed description of our method in the following two subsections.

2.1 Feature Extraction

Following Lee et al., we leverage a pre-trained Stacked Cross Attention Neural Network (SCAN) to build a multi-modal semantic space. Specifically, we obtain word features $\mathbf{H}^\tau = [\mathbf{h}_1^\tau; \dots; \mathbf{h}_M^\tau] \in \mathbb{R}^{M \times D}$ by averaging the forward and backward hidden states per word from a bidirectional GRU (Bahdanau et al., 2014), where $\tau = \{C, R\}$ denotes either a candidate C or a reference R sentence of M words. Based on Anderson et al., we achieve image features $\mathbf{U} \in \mathbb{R}^{N \times D'}$ by detecting N salient regions per image ($N = 36$ in this paper). A linear layer is applied to transform image features to D -dimensional features $\mathbf{V} = [\mathbf{v}_1; \dots; \mathbf{v}_N] \in \mathbb{R}^{N \times D}$.

Based on the SCAN model, we further extract the context information from the caption words for each detected region. To this end, we compute a context feature \mathbf{a}_i^τ for the i^{th} region by a weighted sum of caption word features in Eq. (1). Notice that $\mathbf{A}^\tau = [\mathbf{a}_1^\tau; \dots; \mathbf{a}_N^\tau] \in \mathbb{R}^{N \times D}$ extracts the context information of the caption τ with respect to all regions in the image.

$$\mathbf{a}_i^\tau = \sum_{j=1}^m \alpha_{ij} \mathbf{h}_j^\tau \quad (1)$$

$$\alpha_{ij} = \frac{\exp(\lambda \text{sim}(\mathbf{v}_i, \mathbf{h}_j^\tau))}{\sum_{k=1}^m \exp(\lambda \text{sim}(\mathbf{v}_i, \mathbf{h}_k^\tau))} \quad (2)$$

where λ is a smoothing factor, and $\text{sim}(\mathbf{v}, \mathbf{h}^\tau)$ is a normalized similarity function defined as

$$\text{sim}(\mathbf{v}_i, \mathbf{h}_j^\tau) = \frac{\max(0, \text{cosine}(\mathbf{v}_i, \mathbf{h}_j^\tau))}{\sqrt{\sum_{k=1}^n \max(0, \text{cosine}(\mathbf{v}_k, \mathbf{h}_j^\tau))^2}} \quad (3)$$

2.2 Metric Scores

In order to explore the impact of image data on evaluation, we focus on comparing context features \mathbf{A}^C of the candidate caption C to ground-truth references $\mathbf{G} \equiv [\mathbf{g}_1; \dots; \mathbf{g}_N] = \{\mathbf{V}, \mathbf{A}^R\}$, where \mathbf{G} denotes either image features \mathbf{V} or the context features of R (i.e., \mathbf{A}^R).

Relevance : The relevance between a candidate caption and a ground-truth reference based on the i -th region is computed by the cosine similarity of \mathbf{a}_i^C and \mathbf{g}_i . We average similarity over all regions to get the relevance score of a candidate caption with respect to an image.

$$\mathcal{R} = \frac{1}{N} \sum_{i=1}^N \text{sim}(\mathbf{a}_i^C, \mathbf{g}_i) \quad (4)$$

Extraneous : The extraneous of C is captured by performing an orthogonal projection of \mathbf{a}_i^C to \mathbf{g}_i , which returns the vertical context vector $\mathbf{a}_{i\perp}^C$ to represent the irrelevant content of C to the ground truth at the i^{th} region.

$$\mathbf{a}_{i\perp}^C = \mathbf{a}_i^C - \frac{\mathbf{a}_i^C \cdot \mathbf{g}_i}{\|\mathbf{g}_i\|^2} \mathbf{g}_i. \quad (5)$$

To avoid potential disturbance due to correlated feature vectors, we measure the Mahalanobis distance between the vertical context vector $\mathbf{a}_{i\perp}^C$ and its original context vector \mathbf{a}_i^C (see Eq. (7)). Notice that a small distance value indicates that the irrelevant context vector $\mathbf{a}_{i\perp}^C$ is closed to the original context vector \mathbf{a}_i^C . In other words, the original context contains a large amount of extra information. Therefore, the higher this metric is, the less extra information the caption contains.

$$\mathcal{E} = \frac{1}{N} \sum_{i=1}^N d(\mathbf{a}_i^C, \mathbf{a}_{i\perp}^C) \quad (6)$$

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(\mathbf{p} - \mathbf{q})^T \mathbf{S}^{-1} (\mathbf{p} - \mathbf{q})} \quad (7)$$

Omission : The measurement of omission is similar to that of extraneous, where we capture the missing information of C by the vertical context features $\mathbf{g}_{i\perp}$ based on the orthogonal projection of \mathbf{g}_i to \mathbf{a}_i^C . The omission score is denoted as \mathcal{O} . Similarly, the higher the omission score is, the less missing information the caption contains.

$$\mathbf{g}_{i\perp} = \mathbf{g}_i - \frac{\mathbf{g}_i \cdot \mathbf{a}_i^C}{\|\mathbf{a}_i^C\|^2} \mathbf{a}_i^C \quad (8)$$

$$\mathcal{O} = \frac{1}{N} \sum_{i=1}^N d(\mathbf{g}_i, \mathbf{g}_{i\perp}) \quad (9)$$

Considering that each image may have multiple reference captions, we further average the score of the aforementioned three aspects over all reference captions while considering \mathbf{A}^R as ground truth.

3 Experiments

3.1 Experimental Setup

We perform experiments on three human-evaluated caption sets. **Composite Dataset** (Aditya et al., 2015) contains the candidate captions of images from MS-COCO, Flickr8k, and Flickr30k. Captions were generated by humans and two caption models (11,985 instances in total). Human judgments for these candidate captions was provided on a 5-point scale rating that represents description correctness. **PublicSys Dataset** has 2,500 captions collected by Rohrbach et al., which were generated by five state-of-the-art captioning systems on 500 MS-COCO images, respectively. Human grading was done on a 5-point scale based on annotators' preferences to descriptions. **Pascal-50S Dataset** (Vedantam et al., 2015) includes 4000 caption pairs that describe images from the UIUC PASCAL Sentence dataset. Each annotator was asked to select one sentence per pair that is closer to the expression of the given reference sentence. Candidate pairs were grouped into four categories: 1) human-human correct (HC, i.e., a pair of captions are written by humans for the same image), 2) human-human incorrect (HI, i.e., Two human-written captions of which one describes another image instead of the target image), 3) human-machine (HM, i.e., two captions are generated by a human and a machine, respectively.), and 4) machine-machine (MM, i.e., both machine-generated caption).

Following standard practice (Anderson et al., 2016; Elliott and Keller, 2014), we compared with

Ground Truth	Metric	Composite (τ)	PublicSys (τ)	Pascal-50S (<i>accuracy%</i>)				
				HC	HI	HM	MM	ALL
Reference	BLEU-1	0.280	0.267	50.50	94.50	92.30	56.00	73.33
	BLEU-4	0.205	0.223	50.60	91.90	85.60	60.90	72.25
	ROUGE_L	0.307	0.232	53.30	94.60	93.50	58.20	74.90
	METEOR	0.379	0.254	58.00	97.60	94.90	63.40	78.48
	CIDEr	0.378	0.278	54.80	97.90	91.50	63.80	77.00
	SPICE	0.419	0.258	56.60	94.70	85.00	49.00	71.33
Image (ours)	Relevance	0.423	0.148	58.40	99.40	93.10	73.40	81.08
	Extraneous	0.430	0.149	56.80	99.70	92.80	74.80	81.03
	Omission	0.445	0.165	61.00	99.40	93.80	69.10	80.83
Image + Reference (ours)	Relevance	0.502	0.313	56.40	99.70	93.50	77.10	81.68
	Extraneous	0.507	0.320	54.30	99.60	92.60	77.20	80.93
	Omission	0.533	0.291	60.00	99.60	95.40	72.50	81.88

Table 1: Caption-level correlation between metrics and human grading scores in Composite and PublicSys dataset by using Kendall tau (τ). All p-values < 0.01 . For PASCAL-50S, we display the accuracy of metrics at matching human judgments with 5 reference captions per image. The highest value per column is in bold font. Column titles are explained in Section 3.1. Ground truth refers to two points of information: human-written references and images.

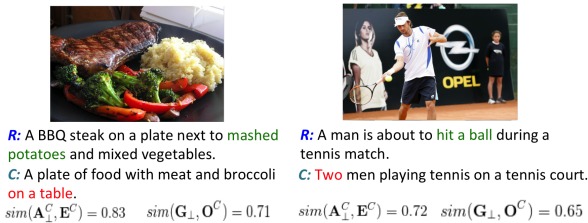


Figure 3: Examples of validating error identification. Text in red is extra information, while text in green is missing information. $sim(x, y)$ is the average similarity between machine-identified and true error vectors over image regions.

rule-based metrics (see Table 1). All existing metrics were implemented with the MS-COCO evaluation tool². The performance of metrics was assessed via Kendall’s tau (τ) rank correlation for the scoring-based datasets (i.e., Composite & PublicSys) and accuracy of the pairwise comparison for the Pascal-50S dataset.

3.2 Experimental Results

Can extra & missing information be captured?

In order to measure the effectiveness of error identification (i.e., extraneous and omission), we randomly sampled a subset of data, and manually identify the actual extraneous (i.e., \mathbf{E}^C) and true omission (i.e., \mathbf{O}^C) of each candidate caption. We conduct validation based on the average co-

sine similarity between the machine-identified error (i.e., $\mathbf{a}_{i\perp}^C$ & $\mathbf{g}_{i\perp}$) and true error description.

Figure 3 provides two illustrative examples of the validation process. Phrases highlighted in red (e.g., ”on a table”) are extra information (more text in the candidate caption than in the ground truth). Meanwhile, phrases in green (e.g., ”mashed potatoes”) are missing from the candidate description, but occur in the image and the reference caption. We observe that machine-identified errors are highly similar to the true error information in both cases (≥ 0.65). This result suggests that our method can capture extraneous and omission from an image caption.

Do error-aware evaluation metrics help?

The results of metric performance in Table 1 show that overall, using the three metrics proposed in REO, especially extraneous and omission, led to a noticeable improvement in Kendall tau’s correlation compared to the best reported results based on prior metrics. Our results suggest that human evaluation tends to be more sensitive to the irrelevance than the relevance of a candidate caption regarding ground truth. We also find that jointly considering both images and human-written references contributes more to caption evaluation than each of the two data sources alone - except for the case of HC pair comparison. This exception can be explained by the phenomenon that human-written descriptions are flexible in terms of word

²<https://github.com/tylin/coco-caption>



References:

- Some baseball players are playing baseball on a field.
- A professional baseball game with a player getting ready to swing a bat.
- A batter ready for a pitch at a baseball game.
- A baseball player is ready to hit a ball at a game.
- A professional baseball game shot of the batter waiting for a pitch.

Candidate:

- Sys1: A batter catcher and umpire during a baseball game.
- Sys2: A crowd watches as a crowd watches as the crowd watches.
- Sys3: A baseball game in progress with a crowd watching.
- Sys4: A group of people watching a baseball game.
- Sys5: A baseball player swinging a bat at a ball.

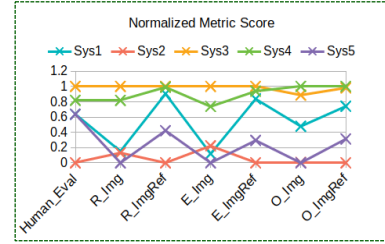


Figure 4: Case study of REO metric scores. Candidates are highlighted in: 1) green: a detailed but incomplected caption, 2) red: repetition, 3) yellow: high-level description, and 4) blue: extra information not shown in the image.

choice and sentence structure, and such diversity may lead to the challenge of comparing a reference to a candidate in cases where both captions were provided by humans. By further looking into each considered aspect in REO, we find that extraneous metric is more appropriate to evaluate machine-generated captions (e.g., the PublicSys dataset and the MM pairs of Pascal-50S dataset), while the omission metric can be a better choice to access caption quality when the testing data consists of human-written descriptions.

What can we learn from the metric outputs?

To analyze our metric scores more in depth, we compare the outputs of five captioning systems on a set of images in the PublicSys dataset. Figure 4 shows an illustrative example. To make the scale of human grading and REO metrics comparable, we normalize scores per metric by using max-min normalization.

We find that metrics calculated in cases where the ground truth contained both the target image and human references are more likely to identify expression errors. For example, though the phrase “a crowd watches” in the caption of system 2 is relevant to the image, this phrase is repeated by three times in a sentence. As a result, the scores for relevance and extraneous are decreasing when the ground truth involves references. Also, metrics focusing only on image content return higher values when the testing captions provide a high-level description of the whole image (e.g., captions of system 3 and 4) compared to the detailed captions for a specific image part (e.g., captions of system 1 and 5 focus on the baseball player). By comparing the herein considered three aspects of each caption, we observe that a caption that mainly focuses on describing a part of image in detail boosts relevance, but the sentence achieves a reduced metric score in terms of omission.

4 Conclusion

This paper presents a fine-grained, error-aware evaluation method REO to measure the quality of machine-generated image captions according to three aspects of descriptions: relevance regarding ground truth, extra description beyond image content, and omitted ground truth information. Comparing these metrics to alternative metrics, we find that our proposed solution produces evaluations that are more consistent with the assessment of human judges. Moreover, we find that human judgment tends to penalize extra and missing information (false positives and false negatives) more than it appreciates relevant content. Finally, and to no surprise, we conclude that using a combination of image content and human-written references as ground truth data allows for a more comprehensive evaluation than using either type of information separately. Our method can be extended to evaluate other text generation tasks.

Acknowledgments

We appreciate anonymous reviewers for their constructive comments and insightful suggestions. This work was partly performed when Ming Jiang was interning at Microsoft Research. The authors would like to thank Pengchuan Zhang for his help with pre-training the grounding model.

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.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.
- Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikinler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.
- 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.
- Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 452–457.
- Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2.
- Jianfeng Gao, Michel Galley, Lihong Li, et al. 2019. Neural approaches to conversational ai. *Foundations and Trends® in Information Retrieval*, 13(2-3):127–298.
- Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.
- Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.
- Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision*, pages 201–216.
- 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 on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. **Object hallucination in image captioning**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.
- Yong Rui, Thomas S Huang, and Shih-Fu Chang. 1999. Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation*, 10(1):39–62.
- Naeha Sharif, Lyndon White, Mohammed Bennamoun, and Syed Afaq Ali Shah. 2018. Learning-based composite metrics for improved caption evaluation. In *Proceedings of ACL 2018, Student Research Workshop*, pages 14–20.
- 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.
- Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. 2019. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6629–6638.