

Word to Sentence Visual Semantic Similarity for Caption Generation: Lessons Learned

Ahmed Sabir

Universitat Politècnica de Catalunya, TALP Research Center, Barcelona, Spain

Abstract

This paper focuses on enhancing the captions generated by image captioning systems. We propose an approach for improving caption generation systems by choosing the most closely related output to the image rather than the most likely output produced by the model. Our model revises the language generation output beam search from a visual context perspective. We employ a visual semantic measure in a word and sentence level manner to match the proper caption to the related information in the image. This approach can be applied to any caption system as a post-processing method.

1 Introduction

Automatic caption generation is a fundamental task that incorporates vision and language. The task can be tackled in two stages: first, image-visual information extraction and then linguistic description generation. Most models couple the relations between visual and linguistic information via a Convolutional Neural Network (CNN) to encode the input image and Long Short Term Memory for language generation (LSTM) [32, 1, 22]. Recently, self-attention has been used to learn these relations via Transformers [14, 8, 25] or Transformer encoder based models like Vision and Language BERT [20, 21, 16]. These systems show promising results on benchmark datasets such as Flickr [13] and COCO [18]. However, the lexical diversity¹ of the generated caption remains a relatively unexplored research problem. Lexical diversity refers to how accurate the generated description is for a given image. An accurate caption should provide details regarding specific and relevant aspects of the image [23]. Caption lexical diversity can be divided into three levels: word level (different words), syntactic level (word order), and semantic level (relevant concepts) [33]. In this work, we approach word level diversity at the semantic level by learning the semantic correlation between the caption and its visual context, as shown in Figure 1, where the visual information from the image is used to learn the semantic relation from the caption in a word and sentence manner.

Modern image captioning systems heavily rely on visual grounding, using objects or regions from the image to guide the caption generation [10, 7, 35]. The informativeness of visual information will help the model

¹Lexical diversity is a measure that counts the different words (unique words) used in a sentence.

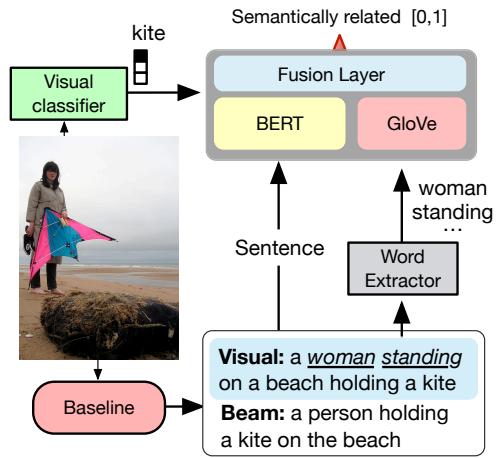


Figure 1. An overview of our visual semantic re-ranker. We employ the visual context in a word and sentence level manner from the image to re-rank the most closely related caption to its visual context. An example from the caption Transformer [8] shows how the **Visual** re-ranker uses the semantic relation to re-rank a more diverse caption.

to narrow the search (*i.e.* beam search) to determine the most related candidate caption to the object or scene in the image. Inspired by these works, we propose an object-based re-ranker to re-rank the most closely related caption with both static and contextualized semantic similarity.

Our main contributions in this paper are: (1) we propose a post-processing method for any caption generation system via visual semantic related measures; (2) as an addendum to the main analysis of this work, we note that the visual re-ranker does not apply to less diverse short beam search, and suffer from the fluctuating of independent stand-alone word similarity score.

2 Beam Search Caption Extraction

We employ the three most common architectures for caption generation to extract the top beam search. The first baseline is based on the standard CNN-LSTM model [32]. The second, VilBERT [21], is fine-tuned on a total of 12 different vision and language datasets such as caption image retrieval. Finally, the third baseline is a specialized Transformer caption generator [8].

Model	B-1	B-2	B-3	B-4	M	R	C	BERTscore
Show and Tell CNN-LSTM [32] ♠								
TellBeamS	0.331	0.159	0.071	0.035	0.093	0.270	0.035	0.8871
Tell+VR_V1BERT-GloVe	0.330	0.158	0.069	0.035	0.095	0.273	0.036	0.8855
Tell+VR_V2BERT-GloVe	0.320	0.154	0.073	0.037	0.099	0.277	0.041	0.8850
Tell+VR_V1RoBERTa-GloVe (sts)	0.313	0.153	0.072	0.037	0.101	0.273	0.036	0.8839
Tell+VR_V2RoBERTa-GloVe (sts)	0.330	0.158	0.069	0.035	0.095	0.273	0.036	0.8869
VilBERT [21] ♣								
VilBeamS	0.739	0.577	0.440	0.336	0.271	0.543	1.027	0.9363
Vil+VR_V1BERT-GloVe	0.739	0.576	0.438	0.334	0.273	0.544	1.034	0.9365
Vil+VR_V2BERT-GloVe	0.740	0.578	0.439	0.334	0.273	0.545	1.034	0.9365
Vil+VR_V1RoBERTa-GloVe (sts)	0.738	0.576	0.440	0.335	0.273	0.544	1.036	0.9365
Vil+VR_V2RoBERTa-GloVe (sts)	0.740	0.579	0.442	0.338	0.272	0.545	1.040	0.9366
Transformer based Caption Generator [8] ♣								
TransBeamS	0.780	0.631	0.491	0.374	0.278	0.569	1.153	0.9399
Trans+VR_V1BERT-GloVe	0.780	0.629	0.487	0.371	0.278	0.567	1.149	0.9398
Trans+VR_V2BERT-GloVe	0.780	0.630	0.488	0.371	0.278	0.568	1.150	0.9399
Trans+VR_V1RoBERTa-GloVe (sts)	0.779	0.629	0.487	0.370	0.277	0.567	1.145	0.9395
Trans+VR_V2RoBERTa-GloVe (sts)	0.779	0.629	0.487	0.370	0.277	0.567	1.145	0.9395

Table 1. Performance of compared baselines on the 5k Karpathy test split ♣ (for VilBERT and Transformer baselines) and Flickr8k 1730 test set ♠ (for Show and Tell CNN-LSTM baseline) with/without Visual semantic Re-ranking. At inference, we use only Top- k -2 (Visual 1 or Visual 2) object visual context once at a time.

3 Visual Re-ranking for Image Captioning

Problem Formulation. Beam search is the dominant method for approximate decoding in structured prediction tasks such as machine translation and image captioning. The larger beam size allows the model to perform a better exploration of the search space compared to greedy decoding. Our goal is to leverage the visual context information of the image to re-rank the candidate sequences obtained through the beam search, thereby moving the most visually relevant candidate up in the list, while moving incorrect candidates down.

Beam Search Visual Re-ranking. We introduce a word and sentence level semantic relation with the visual context in the image. Inspired by [27], who propose a joint BERT [9] with topic modelling for semantic similarity, we propose a joint BERT with GloVe for visual semantic similarity.

Word Level Similarity. To learn the semantic relation between a caption and its visual context in a word level manner, we first employ a bidirectional LSTM based CopyRNN keyphrase extractor [24] to extract keyphrases from the sentence as context. The model is trained on two combined pre-processed datasets: (1) wikidump (*i.e.* keyword, short sentence) and (2) SemEval 2017 Task 10 (Keyphrases from scientific publications) [2]. Secondly, GloVe [28] is used to compute the cosine similarity between the visual context and its related context. Lastly, following [30] we use the object confidence score (*i.e.* visual context) in the image to convert the similarity to probability [4]. For example, from *a woman in a red dress and a black skirt walks down a sidewalk* the model will extract *dress* and *walks*, which are the highlights keywords of the caption.

Sentence Level Similarity. We fine-tune the BERT base model to learn the visual context information. The model learns a dictionary-like relation word-to-sentence

paradigm. We use the visual data (*i.e.* object as context for the caption) to compute the relatedness score.

- **BERT.** BERT achieves remarkable results on many sentence level tasks and especially in the Semantic Textual Similarity task (STS-B) [6]. Therefore, we fine-tuned BERT_{base} on the training dataset, (textual information, 460k captions: 373k for training and 87k for validation) *i.e.* visual, caption, label ([semantically related/not related]), with a binary classification cross-entropy loss function [0,1] where the target is the semantic similarity between the visual and the candidate caption, with a batch size of 16 for 2 epochs.
- **RoBERTa [19].** RoBERTa is an improved version of BERT, and since RoBERTa_{Large} is more robust, we rely on pre-trained Sentence RoBERTa-sts [29] that are fine-tuned on general STS-B dataset [6].

Fusion Similarity Expert. Inspired by Product of Experts PoE [12], we combined the two experts at the word and sentence level as a Late Fusion layer as shown in Figure 1. The combined probability of a given candidate caption \mathbf{w} , with the semantic relation with the visual context, can be written as:

$$P(\mathbf{w}|\theta_1 \dots \theta_n) = \frac{\prod_m p_m(\mathbf{w}|\theta_m)}{\sum_c \prod_m p_m(\mathbf{c}|\theta_m)} \quad (1)$$

where θ_m are the parameters of each model m , $p_m(\mathbf{w}|\theta_m)$ is the probability of \mathbf{w} under the model m , and c is the indexes of all possible vectors in the data space. Since we are just interested in retrieving the candidate caption with higher probability after re-ranking, we do not need to normalize. Therefore, we compute:

$$\arg \max_{\mathbf{w}} P(\mathbf{w}|\theta_1 \dots \theta_n) = \arg \max_{\mathbf{w}} \prod_m p_m(\mathbf{w}|\theta_m) \quad (2)$$

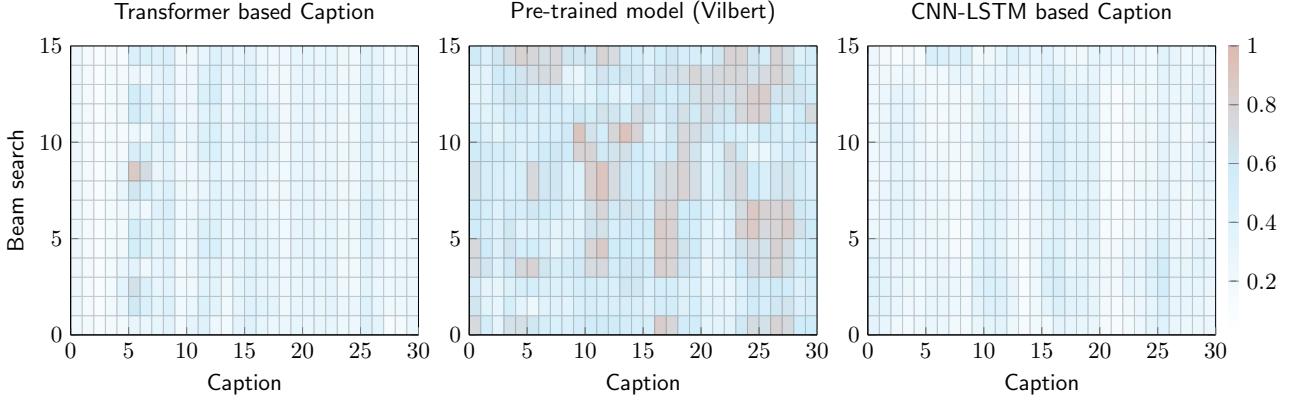


Figure 2. Visualization of the top-15 beam search after visual re-ranking. The color $\square \leq 0$, $\square \leq 0.4$ and $\square \leq 0.8$ represents the degree of change in probability after visual re-ranking, respectively. Also, we can observe that a less diverse beam negatively impacted the score, as in the case of Transformer and CNN-LSTM baselines.

where, in our case, $p_m(\mathbf{w}|\theta_m)$ are the probabilities (*i.e.* semantic relatedness score) assigned by each expert to the candidate caption \mathbf{w} with the semantic relation with the visual context of the image.

4 Datasets

We evaluate the proposed approach on two different size datasets. The idea is to evaluate our approach with (1) a shallow model CNN-LSTM (*i.e.* less data scenario), and on a system that is trained on a huge amount of data (*i.e.* Transformer).

Flickr8k [13]. The dataset contains 8k images, and each image has five human label annotated captions. We use this data to train the shallow model CNN-LSTM baseline (6270 train/1730 test).

COCO [18]. It contains around 120k images, and each image is annotated with five different human label captions. We use the most commonly used split as provided by Karpathy *et al.* [15], where 5k images are used for testing and 5k for validation, and the rest for model training for the Transformer baseline.

Visual Context Dataset. Since there are many public datasets for image captioning, they contain no visual information like objects in the image. We enrich the two datasets, as mentioned above, with textual visual context information. In particular, to automate visual context generation and dispense with the need for human labeling, we use ResNet-152 [11] to extract top- k 3 visual context from each image in the caption dataset.

Evaluation Metric. We use the official COCO offline evaluation suite, producing several widely used caption quality metrics: BLEU [26] METEOR [3], ROUGE [17], CIDEr [31] and BERTscore or (B-S) [34].

5 Experiments

We use visual semantic information to re-rank the candidate captions produced by out-of-the-box state-of-the-art caption generators. We extracted the top-20

Model	Voc	TTR	Uniq	WPC
Show and Tell CNN-LSTM [32] ♠				
TellBeamS	304	0.79	10.4	12.7
Tell+VR _{RoBERTa-GloVe}	310	0.82	9.42	13.5
VilBERT [21] ♣				
VilBeamS	894	0.87	8.05	10.5
Vil+VR _{RoBERTa-GloVe}	953	0.85	8.86	10.8
Transformer based Caption Generator [8] ♣				
TransBeamS	935	0.86	7.44	9.62
Trans+VR _{BERT-GloVe}	936	0.86	7.48	8.68

Table 2. Measuring the lexical diversity of caption before and after re-ranking. Uniq and WPC columns indicate the average of unique/total Words Per Caption, respectively. (The ♠ refers to the Flickr8k 1730 test set, and ♣ refers to the Karpathy 5k test split on COCO Caption dataset.

beam search candidate captions from three different architectures (1) standard CNN-LSTM model [32], (2) a pre-trained vision and language model VilBERT [21], fine-tuned on a total of 12 different vision and language datasets such as caption image retrieval, and (3) a specialized caption-based Transformer [8].

Experiments applying different re-rankers to each base system are shown in Table 1. The tested re-rankers are: (1) VR_{BERT-GloVe}, which uses BERT and GloVe similarity between the candidate caption and the visual context (top- k V1 or V2 during the inference) to obtain the re-ranked score. (2) VR_{RoBERTa-GloVe}, which carries out the same procedure using S_{RoBERTa}.

Our re-ranker produced mixed results as the model struggles when the beam search is less diverse. The model is therefore not able to select the most closely related caption to its environmental context as shown in Figure 2, which is a visualization of the final visual beam re-ranking. However, our models improve the lexical



BL_{BeamS}: a plate of food on a table
VR_{BERT-GloVe}: a plate of food and a drink on a table
Human: a white plate with some food on it.

Visual: food



BL_{greedy}: a green bus parked in front of a building
VR_{BERT-GloVe}: a green double decker bus parked in front of a building X
Human: a passenger bus that is parked in front of a library.

Visual: trolleybus

Figure 3. Examples of the re-ranked captions by our visual re-ranker (VR) against the original caption (**greedy** and **Beam Search**) by the baseline (BL). The (Top) example shows that our model re-ranked a more diverse caption than the baseline. (Bottom) the model struggles when there is a high similarity word with the visual context and without direct relation to the image *i.e.* $\text{sim}(\text{trolleybus}, \text{decker})$ which influences the final score negatively.

diversity, as shown in Table 2 and we can conclude that we have (1) more Vocabulary and (2) the Unique words/total Words Per Caption are also improved, even with a lower Type-Token Ratio (TTR [5]²).

Figure 3 shows examples of the re-ranked captions by our visual re-ranker (VR) against greedy/beam search. (Top) the re-ranked caption has a precise description of the image *food and drink*. In the (Bottom) example, the independent word level (*i.e.* single word without surrounding context) high similarity score $\text{sim}(\text{trolleybus}, \text{decker})$ influences the expert decision negatively, which results in an incorrect re-ranked caption as the object is not present in the image.

Ablation Study. We performed an ablation study, with our worst model in Table 1 *i.e.* caption Transformer, to investigate the effectiveness of each model (*i.e.* word and sentence experts). In this experiment, we trained each model separately, as shown in Table 3. The GloVe performed better as a stand-alone than the combined model (and thus, the combined model breaks the accuracy). To investigate this even further we visualized each expert before the fusion layers as shown in Figure 4 (Bottom), BERT struggles with the Transformer less diverse short caption, and therefore, the word level *i.e.* GloVe dominates as the main expert.

Limitation. In contrast to CNN-LSTM Figure 4 (Top), where each expert is contributing to the final decisions, we observed that having a shorter caption (with less context) can influence the BERT similarity score negatively. Another limitation is the fluctuating of indepen-

²TTR is the number of unique words or types divided by the total number of tokens in a text fragment.

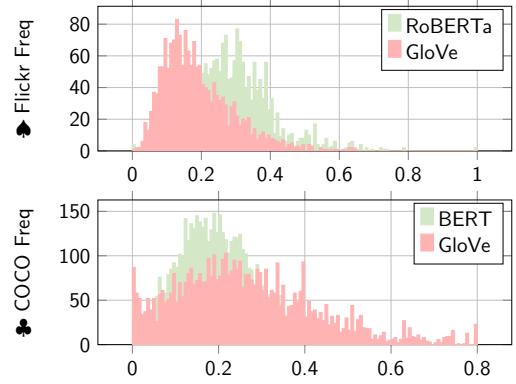


Figure 4. (♠ Top) 1k random sample from Flickr test set with CNN-LSTM baseline. Each Expert is contributing different probability confidence and therefore the model is learning the semantic relation in word and sentence level. (♣ Bottom) 5k Karpathy test split from COCO Caption with caption Transformer baseline. The GloVe score is dominating the distribution to become the expert.

Model	B-4	M	R	C	B-S
Transformer based Caption Generator [8]					
TransBeamS	0.374	0.278	0.569	1.153	0.9399
+VR _{RoBERT-GloVe}	0.370	0.277	0.567	1.145	0.9395
+VR _{BERT-GloVe}	0.371	0.278	0.567	1.149	0.9398
+VR _{RoBERT-BERT}	0.369	0.278	0.567	1.144	0.9395
+VR_V1GloVe	0.371	0.278	0.568	1.148	0.9398
+VR_V2GloVe	0.371	0.278	0.568	1.149	0.9398

Table 3. Ablation study using different model compared to GloVe alone visual re-ranker on the Transformer baseline. ♣ (Bottom) Figure 4 shows that **BERT** is not contributing, as **GloVe**, to the final score for two reasons: (1) less diverse short beam, and (2) the fluctuating of independent word level similarity score as shown in Figure 3 (Bottom).

dent stand-alone word similarity score *i.e.* keyphrases from the caption and the visual $\text{sim}(\text{keyphrase}, \text{visual})$ as shown in the Figure 3 (Bottom) with $\text{sim}(\text{trolleybus}, \text{decker})$ example. Also, the visual classifier struggles with complex backgrounds (*i.e.* misclassified/hallucinated objects), which results in an inaccurate semantic score.

Conclusion

In this work, we have introduced an approach that overcomes the limitation of beam search and avoids re-training for better accuracy. We proposed a combined word and sentence visual beam search re-ranker. However, we discovered that word and sentence similarity disagree with each other when the beam search is less diverse. Our experiments also highlight the usefulness of the proposed model by showing successful cases.

References

- [1] 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 *CVPR*, 2018.
- [2] Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. *arXiv preprint arXiv:1704.02853*, 2017.
- [3] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACLW*, 2005.
- [4] Sergey Blok, Douglas Medin, and Daniel Osherson. Probability from similarity. In *AAAI*, 2003.
- [5] Keith Brown. *Encyclopedia of language and linguistics*. Elsevier, 2005.
- [6] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.
- [7] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *CVPR*, 2019.
- [8] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, 2020.
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.
- [10] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, John C, et al. From captions to visual concepts and back. In *CVPR*, 2015.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [12] Geoffrey E Hinton. Products of experts. 1999.
- [13] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR*, 2013.
- [14] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *CVPR*, 2019.
- [15] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.
- [16] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.
- [17] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 2004.
- [18] 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 *ECCV*, 2014.
- [19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, et al. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [20] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 2019.
- [21] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020.
- [22] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *CVPR*, 2018.
- [23] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In *CVPR*, 2018.
- [24] Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, and Yu Chi. Deep keyphrase generation. *arXiv preprint arXiv:1704.06879*, 2017.
- [25] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *CVPR*, 2020.
- [26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.
- [27] Nicole Peinelt, Dong Nguyen, and Maria Liakata. tbert: Topic models and bert joining forces for semantic similarity detection. In *ACL*, 2020.
- [28] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.
- [29] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
- [30] Ahmed Sabir, Francesc Moreno-Noguer, and Lluís Padró. Visual re-ranking with natural language understanding for text spotting. In *ACCV*, 2019.
- [31] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.
- [32] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.
- [33] Qingzhong Wang and Antoni B Chan. Describing like humans: on diversity in image captioning. In *CVPR*, 2019.
- [34] Tianyi Zhang, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *ICLR*, 2020.
- [35] Yuanan Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. More grounded image captioning by distilling image-text matching model. In *CVPR*, 2020.