Enhancing Descriptive Image Captioning with Natural Language Inference

Zhan Shi, Hui Liu, Xiaodan Zhu
Ingenuity Labs Research Institute, Queen's University
Department of Electrical and Computer Engineering, Queen's University
{z.shi, hui.liu, xiaodan.zhu}@queensu.ca

Abstract

Generating *descriptive* sentences that convey non-trivial, detailed, and salient information about images is an important goal of image captioning. In this paper we propose a novel approach to encourage captioning models to produce more detailed captions using natural language inference, based on the motivation that, among different captions of an image, descriptive captions are more likely to entail less descriptive captions. Specifically, we construct directed inference graphs for reference captions based on natural language inference. A PageRank algorithm is then employed to estimate the descriptiveness score of each node. Built on that, we use reference sampling and weighted designated rewards to guide captioning to generate descriptive captions. The results on MSCOCO show that the proposed method outperforms the baselines significantly on a wide range of conventional and descriptiveness-related evaluation metrics[1].

1 Introduction

Automatically generating visually grounded descriptions for given images, a problem known as image captioning (Chen et al., 2015), has drawn extensive attention recently. In spite of the significant improvement of image captioning performance (Lu et al., 2017; Anderson et al., 2018; Xu et al., 2015; Lu et al., 2018), existing models tend to *play safe* and generate generic captions. However, generating *descriptive* captions that carry detailed and salient information is an important goal of image captioning. For example, recent work (Luo et al., 2018; Liu et al., 2018b, 2019a) leveraged cross-modal retrieval (Faghri et al., 2017; Feng et al., 2014) to solve this problem, based on the observation that more *descriptive* captions often result in better discriminativity in retrieval.

In the paper, we explore to develop better descriptive image captioning models from a novel perspective— considering that among different captions of an image, descriptive captions are more likely to entail less descriptive ones, we develop descriptive image captioning models that leverage natural language inference (NLI, or also known as recognizing textual entailment) (Dagan et al., 2005; MacCartney and Manning, 2009; Bowman et al., 2015), which can utilize multiple references of captions (Young et al., 2014; Lin et al., 2014) to guide the models to produce more descriptive captions.

Specifically, the proposed model first predicts NLI relations for all pairs of references, i.e., *entailment* or *neutral*[2]. Built on that, we construct inference graphs and employ a PageRank algorithm to estimate descriptiveness scores for individual captions. We use reference sampling and weighted designated rewards to incorporate the descriptiveness signal into the Maximum Likelihood Estimation and Reinforcement Learning phase, respectively, to guide captioning models to produce *descriptive* captions. Extensive experiments were conducted on the MSCOCO dataset using different benchmark baseline methods (Huang et al., 2019; Luo et al., 2018; Rennie et al., 2017).

We demonstrate that the proposed method outperforms the baselines, achieving better performances on various evaluation metrics. In summary, the major contributions of the paper are three-fold: (1) To the best of our knowledge, this is the first attempt to connect natural language inference to image captioning, which helps generate more descriptive captions; (2) we propose a reference sampling distribution and weighted designated rewards to guide captioning model to produce more descriptive captions; (3) the proposed method attains better performance on various evaluation metrics over the

[1] https://github.com/Gitsamshi/Nli-image-caption

[2] As reference captions are unlikely to contradict to each other, we ignore the *contradiction* relation in our study.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 269–277
August 1–6, 2021.

state-of-the-art baselines.

2 Related Work

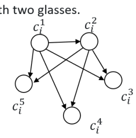
Image captioning Image captioning aims at generating visually grounded descriptions for images. It often leverages a CNN or variants as the image encoder and an RNN as the decoder to generate sentences (Vinyals et al., 2015; Karpathy and Fei-Fei, 2015; Donahue et al., 2015; Yang et al., 2016). To improve the performance on reference-based automatic evaluation metrics, previous work has used visual attention mechanism (Anderson et al., 2018; Lu et al., 2017; Pedersoli et al., 2017; Xu et al., 2015; Pan et al., 2020), explicit high-level attributes detection (Yao et al., 2017; Wu et al., 2016; You et al., 2016), reinforcement learning methods (Rennie et al., 2017; Ranzato et al., 2015; Liu et al., 2018a), contrastive or adversarial learning (Dai and Lin, 2017; Dai et al., 2017), multi-step decoding (Liu et al., 2019a; Gu et al., 2018), weighted training by word-image correlation (Ding et al., 2019) and scene graph detection (Yao et al., 2018; Yang et al., 2019; Shi et al., 2020).

The work of (Luo et al., 2018; Liu et al., 2018b) is most related to ours, which uses retrieval loss as a rewarding signal to encourage descriptive captioning. Different from the above approaches, our method explicitly explore the different *descriptiveness* in references using NLI models and incorporate the information into the training objectives to guide the model to generate more informative sentences. We build our method on top of the existing methods to verify the effectiveness.

Applications of NLI There are basically three major application types for NLI, (1) Direct application of trained NLI models. Trained NLI models are directly used in Fact Extraction and Verification (Thorne et al., 2018) to decide whether a piece of evidence supports a claim (Nie et al., 2019) and generation of longer sentences as a discriminator (Holtzman et al., 2018) to prevent a text decoder from contradicting itself; (2) NLI as a research and evaluation task for new methods. It is widely used as a major evaluation when developing novel language model pretraining (Devlin et al., 2018; Peters et al., 2018; Liu et al., 2019c); (3) NLI as a pre-training task in transfer learning. Training neural network models on NLI corpora and then fine-tuning them on target tasks often yields substantial improvements in performance (Liu et al., 2019b; Phang et al., 2018).

c_i^1: A glass for wine sitting next to a bottle.
c_i^2 : A wine bottle and some wine glasses in the hay.
c_i^3 : Two wine glasses lie beside a bottle of wine in straw.
c_i^4 : Two wine glasses lie beside a wine bottle on straw.
c_i^5 : A bottle of wine on it's side with two glasses.

$$\begin{array}{cc} & \begin{array}{ccccc} c_i^1 & c_i^2 & c_i^3 & c_i^4 & c_i^5 \end{array} \\ \begin{array}{c} c_i^1 \\ c_i^2 \\ c_i^3 \\ c_i^4 \\ c_i^5 \end{array} & \begin{pmatrix} e & n & n & n & n \\ e & e & n & n & n \\ e & e & e & e & e \\ e & e & e & e & e \\ e & e & e & e & e \end{pmatrix} \end{array}$$



Label	Description	Example
Paraphrase	Two way entailment, X entails Y and vice versa	X: Two wine glasses lie beside a bottle of wine in straw Y: Two wine glasses lie beside a wine bottle on straw
Forward Entailment	One way entailment, X entails Y and Y is neutral to X	X: Two wine glasses lie beside a bottle of wine in straw Y: A wine bottle and some wine glasses in the hay
Reverse Entailment	One way entailment, Y entails X and X is neutral to Y	X: A wine bottle and some wine glasses in the hay Y: Two wine glasses lie beside a bottle of wine in straw
Mutual Neutral	Two way neutral, X is neutral to Y and vice versa	X: a tall house with a large clock mounted on its face Y: a building and outdoor seating for a café

Figure 1: A NLI matrix and inference graph.

3 Our Method

The goal of image captioning is to train conditional generation model $p_\theta(c \mid x)$ based on training instances $(x_i, C_i)_{i=1}^m$ in a training dataset and $C_i = \{c_i^1, \cdots, c_i^n\}$, where m is the number of training instances and n is the number of reference captions for an image.

The typical models leverage a two-phase learning process to estimate $p_\theta(c \mid x)$: the first uses MLE objective, which minimizes a cross-entropy loss with regard to the ground truth captions:

$$\mathcal{L}_{\mathrm{ML}}(\theta) = -\sum_{i=1}^{m}\sum_{j=1}^{n} \log p_\theta(c_i^j \mid x_i) \tag{1}$$

RL is then used to optimize models by maximizing the expected reward for generating captions.

$$\mathcal{L}_{\mathrm{RL}}(\theta) = -\sum_{i=1}^{m} E_{\hat{c} \sim p_\theta(c \mid x_i)}[r(\hat{c}, x_i)] \tag{2}$$

where $r(\hat{c}, x_i)$ could be CIDEr reward (r_{cd}) (Rennie et al., 2017) or a combination of CIDEr (r_{cd}) and discriminative loss (l_{dis}) (Luo et al., 2018).

In this work, we enhance these two basic learning objectives by considering the descriptiveness of references $\{c_i^1, \cdots, c_i^n\}$.

3.1 Constructing Inference Graphs

NLI Matrix The SNLI corpus (Bowman et al., 2015) is widely used for training natural language inference models. To leverage the data for our task, we extract a subset of SNLI to fit our needs, e.g., removing *contradiction* sentence pairs (see Appendix B for details). Our NLI model is built upon BERT (Devlin et al., 2018), which achieves near state-of-the-art performance and is sufficient for our purpose. Given reference captions $C_i =$

$\{c_i^1, \cdots, c_i^n\}$ of an image, we obtain a NLI label for each ordered pair $\langle c_i^j, c_i^k \rangle$, forming a NLI relation matrix, as shown in Figure 1. Note that a NLI relation matrix is not necessary to be a symmetric matrix. For example, it is possible that $\langle c_i^j, c_i^k \rangle$ has an entailment relation (i.e., c_i^j entails c_i^k) and $\langle c_i^k, c_i^j \rangle$ is neutral, by the definition in NLI (Bowman et al., 2015).

Inference Graphs Built on the NLI matrix, we construct the inference graphs. For c_i^j and c_i^k, if the ordered pair $\langle c_i^j, c_i^k \rangle$ and $\langle c_i^k, c_i^j \rangle$ are both *entailment* in the NLI matrix, c_i^j and c_i^k are *paraphrases*. If $\langle c_i^j, c_i^k \rangle$ is entailment and $\langle c_i^k, c_i^j \rangle$ is neutral, then $\langle c_i^j, c_i^k \rangle$ is said to be a *forward entailment* (FwdEntail). On the contrary, if $\langle c_i^j, c_i^k \rangle$ is neural and $\langle c_i^k, c_i^j \rangle$ is entailment, then $\langle c_i^j, c_i^k \rangle$ is said to be a *reverse entailment* (RevEntail). If both directions are neutral, we call it mutual neutral (muNeutral).

To construct a directed inference graph, captions in a given image are added as vertexes. We add a directed edge from c_i^j to c_i^k if $\langle c_i^j, c_i^k \rangle$ is revEntail; i.e., the edge's head c_i^k is expected to be more descriptive than the tail c_i^j, and the edge points towards c_i^k. If $\langle c_i^j, c_i^k \rangle$ is fwdEntail, we add an edge from c_i^k to c_i^j. We do not add edges for paraphrase and muNeural pairs.

Descriptiveness Scorer PageRank (Page et al., 1999) is a link analysis model applied to collections of nodes with quotations or references. We perform PageRank on a inference graph to compute the *descriptiveness* score for each node/caption, which measures at which node a random walk is more likely to stop. Nodes with a higher score assigned by PageRank can be viewed as more *descriptive*. We then normalize the score to obtain distribution $q(c \mid x_i), c \in C_i$.

3.2 Descriptiveness Regularized Learning

Reference sampling (Rs) for MLE We can verify that $\mathcal{L}_{\text{ML}}$ in Equation (1) is equivalent to the KL divergence between a uniform target reference distribution $U(c \mid x_i)$ and model distribution $p_\theta(c \mid x_i)$:

$$\mathcal{L}_{\text{ML}}(\theta) = \sum_{i=1}^{m} \text{KL}(U(c \mid x_i) || p_\theta(c \mid x_i)) \quad (3)$$

Note that Equation (3) indicates that any c that belongs to reference set of C_i will be equally learned without considering their *descriptiveness*. To resolve the issue, for an image x_i, we use the probability distribution q obtained from graph nodes. We obtain an enhanced MLE loss $\mathcal{L}'_{\text{ML}}$, which is equivalent to minimizing the KL divergence between the target reference sampling distribution q and p_θ:

$$\mathcal{L}'_{\text{ML}}(\theta) = \sum_{i=1}^{m} \text{KL}(q(c \mid x_i) || p_\theta(c \mid x_i)) \quad (4)$$

Weighted reward (Wr) for RL We modify the reward function in RL to integrate the *descriptiveness* score to encourage more contribution from descriptive references in designated reward. Specifically, we change the CIDEr reward item r_{cd} in $r(\hat{c}, x_i)$ as shown in equation (2) by replacing $U(c \mid x_i)$ with $q(c \mid x_i)$:

$$r'_{cd}(\hat{c}, x_i) = \sum_{j=1}^{n} q(c_i^j \mid x_i) \cdot \text{CD}(\hat{c}, c_i^j) \quad (5)$$

where CD denotes the CIDEr similarity score.

4 Experiment

4.1 Setup

Dataset and Evaluation Metrics We perform experiments on the Karpathy split of the MSCOCO dataset (Lin et al., 2014; Karpathy and Fei-Fei, 2015). We employ a wide range of conventional image caption evaluation metrics, i.e., SPICE(SP) (Anderson et al., 2016), CIDEr(CD) (Vedantam et al., 2015), METEOR(ME) (Denkowski and Lavie, 2014), ROUGE-L(RG) (Lin, 2004), and BLEU (Papineni et al., 2002) to evaluate the generated captions. Following (Liu et al., 2019a), we also use the caption generated $\hat{c}$ to retrieve image x using a separately trained image-matching model (Lee et al., 2018). The retrieval evaluation is based on 1K images (Lee et al., 2018) from the Karpathy test set. Retrieval performances are measured by R@K ($K = 1, 5$), i.e., whether x is retrieved within the top K retrieved images. We also perform human evaluation on *descriptiveness, fluency*, and *fidelity*.

Implementation Details To make a fair comparison, we use the same experiment setup that the compared baselines used. See more implementation details for NLI model, retrieval model in evaluation, and descriptiveness score normalization in appendix B.

Compared Models We use AoANet, ATTN, and DISC(λ set to 1) as the baselines. ATTN (Rennie et al., 2017) is a LSTM based decoder with visual attention mechanism. AoANet (Huang et al., 2019) adopts the attention on attention module. We also leverage the discriminativity enhanced model DISC (Luo et al., 2018) which is built upon ATTN.

4.2 Results and Analyses

Overall Performance Table 1 shows the overall performance of different models.

Results on conventional metrics. Our method consistently outperforms the baseline models on most conventional metrics, especially SPICE and CIDEr; e.g., the proposed model improves the AoANet baseline from 118.4 to 119.1 on CIDEr, 21.5 to 21.7 on SPICE in the MLE phase, and improves the ATTN baseline on CIDEr from 117.4 to 120.1, SPICE from 20.5 to 21.0 in the RL phase. As CIDEr is based on tf-idf weighting, it helps to differentiate methods that generate more image-specific details that are less commonly occur across the dataset. As our method is designed to encourage models to generate sentences with more objects, attributes, or relations, the effect was also suggested by the improvement on SPICE.

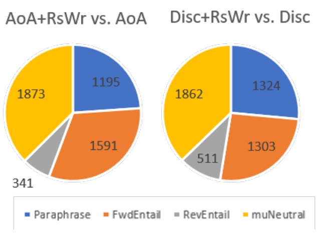


Figure 2: Inference labels in different models

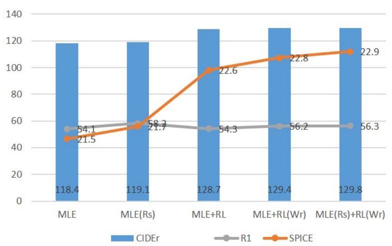


Figure 3: Ablation Analysis based on AoANet

Performance on descriptiveness related metrics. Our methods achieve consistently better results on R@1 and R@5 in both the MLE and RL optimization phases. Note that the proposed model can further boost the retrieval performance on the discriminativity enhanced baseline (DISC), improving R@1 from 46.5 to 48.1 and R@5 from 83.6 to 87.9. Our weighted CIDEr reward is complementary to the discriminative loss item in DISC and further boost the retrieval performance.

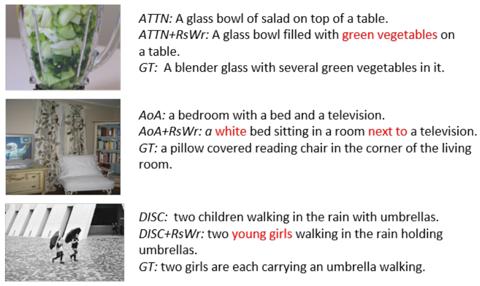


Figure 4: Examples generated by different models.

Labels between generated sentences. We use the externally trained NLI model (Section 3.1) to further investigate the NLI relationships between the captions generated by our method and by the baselines (AoA and DISC) on the testset. Figure 2 shows that our model generates more descriptive sentences. For example, comparing the generation results of AoA+RsWr and AoA on 5,000 testing images, captions generated by AoA+RsWr *forward-entails* those generated by AoA on 1,591 images, and *reverse-entails* on 341 images.

Ablation analysis. As shown in Figure 3, both reference sampling (Rs) and weighted reward (Wr) can improve performance in their respective optimization period, i.e., MLE to MLE(Rs), MLE+RL to MLE+RL(Wr). There is also a marginal improvement when using MLE(Rs) instead of MLE before the RL(Wr) optimization period, i.e., MLE+RL(Wr) to MLE(Rs)+RL(Wr), showing that MLE(Rs) has a positive impact even after RL(Wr) optimization.

Human Evaluation We further perform human evaluation on our method and two baselines (here, ATTN and DISC) using 100 images randomly sampled from the test set. Three human subjects rate captions with 1-5 Likert scales (higher is better) with respect to three criteria: *fluency*, *descriptiveness*, and *fidelity*. See more details in appendix A for rating details. Table 2 shows that ATTN+RsWr performs better than ATTN on descriptiveness. Moreover, DISC+RsWr can further improve the *descriptiveness* performance over the baseline discriminativity enhanced captioning model.

Case Study. Figure 4 includes three examples, in which our model produces captions with more attributes, objects, or relations.

5 Discussion

5.1 Descriptiveness and Entailment

We perform human analysis between descriptiveness and entailment. Specifically we randomly

	Maximum Likelihood Estimation							Reinforcement Learning						
	BLEU4	ME	RG	CD	SP	R@1	R@5	BLEU4	ME	RG	CD	SP	R@1	R@5
AoA	36.8	28.3	57.3	118.4	21.5	54.1	**87.6**	39.0	29.0	**58.9**	128.7	22.6	54.3	88.6
AoA+RsWr	**36.9**	**28.5**	**57.5**	**119.1**	**21.7**	**58.2**	87.4	39.0	**29.1**	58.7	**129.8**	**22.9**	**56.3**	**90.2**
ATTN	35.5	27.0	56.0	108.9	19.8	42.8	79.7	35.8	27.1	56.7	117.4	20.5	40.8	77.3
ATTN+RsWr	**35.8**	**27.3**	**56.3**	**112.1**	**20.5**	**48.2**	**84.4**	**36.2**	**27.3**	56.7	**120.1**	**21.0**	**44.9**	**84.8**
DISC	-	-	-	-	-	-	-	35.6	27.2	**57.0**	115.4	21.0	46.5	83.6
DISC+RsWr	-	-	-	-	-	-	-	**35.9**	27.2	56.8	**118.3**	**21.4**	**48.1**	**87.9**

Table 1: Results on MSCOCO karpathy split. RsWr detnotes Reference sampling and Weighted reward.

	Fluency	Descriptiveness	Fidelity
ATTN	3.90	2.53	3.46
ATTN+RsWr	**3.91**	**2.86**	**3.50**
DISC	**3.52**	3.08	3.28
DISC+RsWr	3.49	**3.30**	**3.31**

Table 2: Human evaluation on different models.

	Pairwise Similarity Comparison						
	B@4	ME	RG	CD	SP	R@1	R@5
AoA	39.0	29.0	**58.9**	128.7	22.6	54.3	88.6
AoA+Sim	38.8	28.8	58.6	128.3	22.5	54.0	87.4
AoA+RsWr	39.0	**29.1**	58.7	**129.8**	**22.9**	**56.3**	**90.2**
	Re-ranking Comparison						
ATTN	35.8	27.1	56.7	117.4	20.5	40.8	77.3
ATTN+re-rank	35.7	27.2	**56.8**	117.0	20.6	41.5	78.8
ATTN+RsWr	**36.2**	**27.3**	56.7	**120.1**	**21.0**	**44.9**	**84.8**

Table 3: Comparison with pairwise similarity and re-ranking.

sample 50 images from the MSCOCO training set. For one image, there are five references, constituting ten reference pairs. So we have 500 reference pairs. For each reference pair, we ask three subjects to annotate whether one sentence conveys more non-trivial, important and detailed information than the other in terms of the described image. If the majority of the three subjects annotate yes, they further annotate the NLI relation—entailment or neutral, with the more informative caption as premise and the other as the hypothesis. As a result, out of the 500 reference pairs, we obtained 208 pairs that have differences in descriptiveness. The annotated NLI relations show that 164 of the 208 collected pairs have the entailment relation; i.e., for around 80% of the 208 pairs, "descriptive captions entail less descriptive captions" holds in the randomly sampled MSCOCO subset, where MSCOCO is a widely used multi-reference image caption benchmark.

5.2 Pairwise similarity and Re-ranking

We apply a pairwise similarity approach to AoA, in which we use Jaccard similarity between a pair of sentences to build the graph and run PageRank to get scores. Table 3 shows that pairwise similarity baseline approach (AoA+Sim) did not further improve performance over the corresponding baselines, showing pairwise similarity does not suggest descriptiveness, unlike entailment.

We perform re-ranking on the ATTN baseline; we use beam search with a beam size of 3, and then rank the captions in the beam by descriptiveness scores, which is calculated by BERT based NLI model. As shown in Table 3, the re-ranked sentences in the beam do not have much improvement in terms of baseline. Sentences generated by beam search (c.f. appendix C) do not vary significantly in terms of descriptiveness; these sentences are usually neutral to each other and sentences ranked low in the beam may have the fidelity/fluency issues.

6 Conclusions

We explore a novel approach to encourage image captioning models to produce more descriptive sentences using natural language inference. We construct inference graphs and descriptiveness scores are assigned to nodes using the PageRank algorithm. Built on that, we use reference sampling and weighted designated rewards to guide captioning to generate descriptive captions. We demonstrate the effectiveness of the model on various evaluation metrics and perform detailed analyses.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments. This research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

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 *CVPR*, pages 6077–6086.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards diverse and natural image descriptions via a conditional gan. In *ICCV*.

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

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*.

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*.

Guiguang Ding, Minghai Chen, Sicheng Zhao, Hui Chen, Jungong Han, and Qiang Liu. 2019. Neural image caption generation with weighted training and reference. *Cognitive Computation*, 11(6):763–777.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improved visual-semantic embeddings. *arXiv*, 2(7):8.

Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16.

Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. 2018. Stack-captioning: Coarse-to-fine learning for image captioning. In *AAAI*.

Aric Hagberg, Pieter Swart, and Daniel S Chult. 2008. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. *arXiv preprint arXiv:1805.06087*.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE 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.

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 (ECCV)*, pages 201–216.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

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.

Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. 2018a. Context-aware visual policy network for sequence-level image captioning. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1416–1424.

Lixin Liu, Jiajun Tang, Xiaojun Wan, and Zongming Guo. 2019a. Generating diverse and descriptive image captions using visual paraphrases. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4240–4249.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019b. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. 2018b. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *ECCV*, pages 338–354.

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

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228.

Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6964–6974.

Bill MacCartney and Christopher D Manning. 2009. *Natural language inference*. Citeseer.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6859–6866.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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.

Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. 2017. Areas of attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1242–1250.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024.

Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. 2020. Improving image captioning with better use of caption. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7454–7464.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and verification (fever) shared task. *arXiv preprint arXiv:1811.10971*.

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.

Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 203–212.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

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

Zhilin Yang, Ye Yuan, Yuexin Wu, William W Cohen, and Russ R Salakhutdinov. 2016. Review networks for caption generation. In *Advances in neural information processing systems*, pages 2361–2369.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699.

Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4894–4902.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *CVPR*.

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

A Human Evaluation Details

The human evaluation is performed with three non-author human subjects. We ask the subjects to rate on three 1-5 Likert scales, corresponding to *fidelity* (the sentences' fidelity to the corresponding images), *fluency* (the quality of captions in terms of grammatical correctness and fluency), and *descriptiveness* (how much the sentences convey more detailed and faithful information about the images).

B More Implementation Details

NLI We exclude the training instances labeled with *contradiction*, since our task does not need to consider contradiction—reference captions for the same image are unlikely to contradict each other. We also sample training instances in the SNLI dataset to make the subset's length distribution similar to the caption references. We obtained a filtered dataset with around 250K sentence pairs as our training set, 4K and 4K as validation and test set, respectively. We leverage BERT (Devlin et al., 2018) as the framework for training which is a basis for many state-of-the-art models and achieve near state-of-the-art performance, which is sufficient for our purpose. The training gets stabled after 3 epochs, reaching an accuracy around 88% on the test set.

Retrieval Model in Evaluation The model is trained with the published package of SCAN (Lee et al., 2018). For the specific parameters, we followed the "SCAN t-i LSE" setting in their published report.

Descriptiveness Score We use the entailment probability as the weights on the edges and then we perform PageRank using the toolkit from (Hagberg et al., 2008). We set the damping parameter of 0.95 for descriptiveness score at MLE training stage and 0.1 for descriptiveness score at RL training stage, as we find that a smooth score distribution on reward (c.f. Equation 5) and a peaked score distribution on MLE(c.f. Equation 4) lead to improved performance in the RL and MLE training stage respectively.

C Beam Search Generation

Example 1. {*"image˙id": 247625, "caption": a man holding a snowboard in the snow, a man standing on a snowboard in the snow, a man is standing on a snowboard in the snow*}

{*"image˙id": 131019, "caption": a group of zebras are standing in a field, a group of zebras are standing in a field with a zebra, a group of zebras are walking in a field*}

These are sentences generated by beam search by ATTN model after RL stage (before re-ranking).