Descriptive Image Captioning with Salient Retrieval Priors

Zhan Shi^{†,*}, Hui Liu[†], Xiaodan Zhu[†][†] Ingenuity Labs Research Institute & ECE, Queen's University, Canada

Abstract

Captions are often expected to carry detailed, essential information of images, but current image captioning models tend to play safe and generate generic captions that is less informative. Cross-modal retrieval is a promising solution as texts with more details has better performance in retrieval. In this work, we first explore two types of salient n-grams, i.e., Support N-grams (SN) and Deletion N-grams (DN), in captions which significantly affect the performance of typical cross-modal retrieval models. We further exploit these n-grams to enhance the original learning objectives for generating descriptive captions with more details. The experiments on two benchmark datasets show that our proposed model outperforms baselines significantly when evaluated with a wide range of metrics.

Keywords: Image Captioning, Natural Language Generation, Cross Modal Retrieval.

1. Introduction

Image captioning [1, 2] is a multimodal problem that lies in the intersection of the natural language processing and computer vision. Good captions are expected to have a high fidelity to images and are descriptive and fluent. The most recent years have seen a significant improvement of image captioning performance [1, 3–5]. However, existing models tend to play safe and generate generic captions, and generating informative captions needs further studies.

To assess how informative a caption is, we may design an explicitly discriminative task the success of which would depend on how accurately the caption describes the visual input. Leveraging cross-modal retrieval [6] is a promising approach to better solving this problem, due to the fact that more descriptive captions often results in better performance in finding relevant images in text-to-image retrieval or can be more easily retrieved when images are used as queries in image-to-text retrieval.

Inspired by the concept of saliency maps [7], we further investigate the cross-retrieval model by searching for fine-grained n-grams contributing to retrieval performance. To be more specific, given Figure 1, we try to find solution from answering the two questions: (1) What n-grams could cause a plunge of retrieval performance if removed from the caption? (Deletion N-grams), and (2) what n-grams could significantly support the retrieval performance alone? (Support N-grams) (see Section 3.2 for details). We leverage the salient n-grams into our training process by regularizing the original learning objective (see Section 3.3 for details) in the cross entropy (*CE*) and reinforcement learning (*RL*) optimization period to encourage the generated sentence covering them.

We perform extensive experiments with the proposed models on two benchmarks: MSCOCO [8] and FLICKR30K [9]. Our contributions are summarized as follows: (1) We investigate two categories of salient n-grams in captions and show how they are critical for cross-modal retrieval. (2) We design novel mechanisms to incorporate the salient n-grams to regularize the learning objective for generating more descriptive captions. (3) Our method achieves significantly better performance over a wide range metrics on two benchmark datasets over baselines. Human evaluation further confirms the effectiveness of the proposed method.

2. Related Work

Image captioning, aiming at generating visually grounded descriptions for images, often leverage an CNN or variants as the image encoder and an RNN as the decoder to generate sentences [2, 10, 11]. To improve the performance on reference-based automatic evaluation metrics, visual

*18zs11@queensu.ca

Figure 1. Salient SN and DN for a caption.

attention mechanism [1, 3, 4], explicit high-level attributes detection [12, 13], reinforcement learning methods [14], contrastive or adversarial learning [15, 16], multi-step decoding [17] and scene graph detection [18, 19] are proposed. The work of [20, 21] is most related to ours, which uses retrieval loss as a reward signal to produce descriptive captions. Unlike their approaches, our method utilizes salient n-grams, which are critical to retrieval performance, to guide caption generation. We build our method on top of these related methods to verify the effectiveness of our method.

3. Method

3.1. Cross Modal Retrieval

We leverage stacked cross attention network (SCAN) [22] as our cross-modal retrieval model. SCAN takes multi-modal inputs: an image with multiple object features $I = \{v_1, \dots, v_k\}$ and a sentence of contextualized word features $C = \{c_1, \dots, c_n\}$. The output is a similarity score $s(I, C)$ for the input image-sentence pair. SCAN calculates fine-grained latent alignment between each v_i and c_j to compute image-sentence similarity. We further train the score function $s(I, C)$ through the triplet loss:

$$\mathcal{L}_T = \sum_{\hat{C}} [\alpha - s(I, C) + s(I, \hat{C})]_+ + \sum_{\hat{I}} [\alpha - s(I, C) + s(\hat{I}, C)]_+ \quad (3.1)$$

where $[x]_+ \equiv \max(x, 0)$ and α is the margin between positive $s(I, C)$ and negatively sampled image-sentence pairs $s(I, \hat{C}), s(\hat{I}, C)$.

3.2. Salient N-grams Searching

To find n-grams critical for retrieval, we propose to search two types of salient n-grams (n from 1 to 5): *Deletion N-grams* and *Support N-grams*. The former focuses on finding n-grams that would significantly impair retrieval performance if removed, and the latter attempts to obtain n-grams that maximally support the retrieval.

Given any n-gram m in caption C , we can represent C as a union of two partitions $C = C_m \cup C_{\setminus m}$. We are interested in image-text similarity when, (1) C_m are unobserved, in which the similarity score can be expressed as $s(I, C_{\setminus m})$, and (2) $C_{\setminus m}$ are unobserved, in which the score can be written as $s(I, C_m)$. Our objective is to find all m that satisfy either of the two following conditions:

$$s(I, C) - s(I, C_{\setminus m}) > \alpha \quad (3.2)$$

$$s(I, C) - s(I, C_m) < \alpha \quad (3.3)$$

where α is the margin in cross-modal retrieval model. For simplicity, we denote the n-grams set satisfying Equation (3.2) as $\mathcal{M}_{DN}$ and Equation (3.3) as $\mathcal{M}_{SN}$.

3.3. Image Captioning Training

The goal of image captioning is to train a conditional text generation model $p_\theta(C | I)$ based on training dataset $(I, C) \in D$. We use the *CE* and then *RL* loss to optimize the model:

$$\mathcal{L}_{CE}(\theta) = -\sum_{I, C} \log p_\theta(C | I) \quad (3.4)$$

$$\mathcal{L}_{RL}(\theta) = -E_{C' \sim p_\theta(C|I)} [r(C, C')] \quad (3.5)$$

Regularized Learning Objectives. We leverage the salient n-grams set $\mathcal{M}$ to enhance the *CE* and *RL* objective, respectively:

$$\mathcal{L}'_C(\theta) = -\sum_{I,C} \sum_i w_i \cdot \log p_\theta(c_i | I, C_{<i}) \quad (3.6)$$

$$w_i = \begin{cases} 1 & c_i \notin m \\ \lambda_c & c_i \in m \end{cases} \quad (3.7)$$

where w_i is a weight for c_i and its value is 1 if it is not contained in any $m \in \mathcal{M}$, otherwise it would be a constant λ_c greater than 1.

$$r'(C, C') = r(C, C') + \lambda_r \cdot \frac{\#(C', \mathcal{M})}{|\mathcal{M}|} \quad (3.8)$$

where $\#(C', \mathcal{M})$ is the number of n-grams in C' belonging to $\mathcal{M}$, and λ_r is the balanced weight.

4. Experiment Set-Up

Data and Evaluation Metrics We perform experiments on the MSCOCO [8] and FLICKR30K [9] benchmarks. For MSCOCO, we use both the Karpathy split [2] (113K, 5K, 5K for training, validation and test) and the standard split (82K/5K/5K). And for FLICKR30K, there are 29K/1K/1K image caption pairs for training, validation and test respectively. We use a wide range of evaluation metrics, i.e., SPICE [23], CIDEr-D [24], METEOR [25], ROUGE-L [26], and BLEU [27]. In addition, we also evaluate descriptiveness with an extrinsic evaluation — we use the generated captions in a text-to-image retrieval model [22] and measure the retrieval performance with the R@K(K=1, 5) metric based on 1K images in test set [22]. Furthermore, we also perform human evaluation for descriptiveness, fluency, and fidelity.

Compared Models We use AoANet, ATTN and DISC as the baselines. ATTN [14] is a LSTM based visual attention network which encodes image into a set of features and decodes tokens using state from LSTM and weighted average of image features. AoANet [28] adopts attention on attention module, which achieves superior performance over automatic evaluation metrics. Moreover, we also leverage the self-retrieval enhanced captioning model DISC [20] which is built upon ATTN to verify if our method can further boost *descriptiveness*. We also list the performance when different salient n-grams set are used, namely, $\mathcal{M}_{SN}, \mathcal{M}_{DN}, \mathcal{M}_{SN} \cap \mathcal{M}_{DN}, \mathcal{M}_{SN} \cup \mathcal{M}_{DN}$.

Implementation Details To make fair comparison, we use the same experiment setup that the compared baselines used. The hyper-parameter λ_c is set to be 1.1. λ_r is set to be 0.1 for MSCOCO and 0.2 for FLICKR30K. Following Updown [4], we use the same bottom-up features which is extracted by Faster R-CNN [29] (ResNet-101 [30] as backbone) fine-tuned on Visual Genome [31] for AoANet and ATTN baselines over MSCOCO and FLICKR30K dataset. We set the initial learning rate to be 0.0002 (CE phase), 0.00002 (RL phase) for both baselines. The mini-batch size to be 128 for ATTN and 16 for AoANet. The maximum number of training epochs is 25 for CE and 20 for RL optimization respectively. For sequence generation in the inference stage, we adopt the beam search strategy with beam size to be 2.

4.1. Results and Analysis

Overall performance Table 1 and 2 show the results on the MSCOCO Karpathy/standard split and FLICKR standard split. The results shows that our proposed method outperforms the baselines on most metrics. Specifically on CIDEr-D, SPICE, METEOR, BLEU@4, our model reaches the score of 129.5, 22.8, 29.1, 39.2 on MSCOCO Karpathy split, and 71.1, 17.2, 23.2, 30.9 on FLICKR30K standard split, resulting in the superior single model performance. Furthermore, the performance can be further boosted when we apply our methods on DISC, with R@1 and CIDEr-D improved from 46.5 to 48.2, and 115.4 to 116.4, respectively.

	COCO Karpathy Split							COCO Standard Split						
	B@4	ME	RG	CD	SP	R@1	R@5	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	37.7	28.5	58.1	125.0	22.2	54.0	87.8
AoA+SN	39.2	29.1	58.9	129.5	22.8	55.1	89.3	38.2	28.6	58.2	126.0	22.4	55.3	88.7
AoA+DN	39.0	29.0	58.7	128.7	22.7	56.1	88.7	38.0	28.5	58.1	125.7	22.3	55.4	89.8
AoA+DN $\cap$ SN	38.6	28.8	58.5	128.1	22.6	55.5	89.9	38.0	28.7	58.3	126.1	22.4	53.6	87.9
AoA+DN $\cup$ SN	39.0	29.1	58.8	129.1	22.7	55.0	89.7	38.1	28.5	58.1	125.6	22.2	54.1	88.4
ATTN	35.6	27.2	57.0	118.0	20.6	40.8	77.4	34.7	26.9	56.5	115.7	20.2	39.2	75.5
ATTN+SN	35.6	27.2	57.0	118.4	20.6	42.2	78.9	35.2	27.0	56.7	116.7	20.5	40.2	76.6
ATTN+DN	35.6	27.1	56.9	118.2	20.4	41.8	78.4	35.0	26.8	56.4	115.7	20.3	39.7	75.8
ATTN+DN $\cap$ SN	35.5	27.1	56.8	118.1	20.5	40.9	77.7	35.1	26.7	56.4	116.0	20.3	39.8	75.9
ATTN+DN $\cup$ SN	35.7	27.2	56.9	118.0	20.5	41.8	77.2	34.9	26.8	56.5	115.8	20.3	38.4	75.2
DISC	35.6	27.1	56.7	115.4	21.0	46.5	83.6	35.1	26.9	56.5	114.0	20.6	42.2	79.3
DISC+SN	35.5	27.1	56.8	116.4	20.9	48.2	84.8	35.1	27.0	56.6	115.3	20.6	43.1	79.9

Table 1. Results on MSCOCO, B@4, ME, RG, CD, SP mean BLEU-4, METEOR, ROUGE, CIDEr-D, SPICE respectively. The AoA* result is the average score on three run of its official release.

	FLICKR Standard split						
	B@4	ME	RG	CD	SP	R@1	R@5
AoA	30.2	22.7	50.6	69.4	16.8	49.7	82.0
AoA+SN	30.9	23.1	51.1	70.7	17.2	54.0	86.0
AoA+DN	30.8	23.2	51.3	71.1	17.1	50.8	84.4
AoA+DN $\cap$ SN	30.7	23.1	51.3	69.9	17.1	52.7	85.5
AoA+DN $\cup$ SN	30.4	23.1	50.9	70.1	16.9	53.9	85.7

Table 2. Performances on the FLICKR30K dataset.

Performance with Fewer Training Data. We can see that the salient n-grams have a greater benefit when there are fewer training data. As the size of the training data of the MSCOCO standard split is three fourths of that of the Karpathy split, AoA+SN achieves a considerably higher relative improvement, improving CIDEr-D and R@1, by 0.9% and 2.4%, respectively, compared to 0.6% and 1.5% in the Karpathy Split. Moreover, the FLICKR30K standard split has only 30% training data compared to the MSCOCO Karpathy split. AoA+SN obtains a relative improvement of 8.7% and 1.8% on R@1 and SPICE, respectively, significantly higher than 1.5% and 0.9% on the MSCOCO Karpathy split.

Types of Salient N-grams. We additionally investigate the four salient n-grams sets. $\mathcal{M}_{SN}$ is much larger than $\mathcal{M}_{DN}$ in both datasets. In MSCOCO, the intersection $\mathcal{M}_{SN} \cap \mathcal{M}_{DN}$ covers more than half of the total n-grams in $\mathcal{M}_{DN}$, and in FLICKR30K, the coverage is more than 70%.

Hyper-parameters and Ablation Analysis There are two major hyper-parameters λ_c and λ_r : (1) λ_c is the weight for a token in the salient n-grams (Equation 3.7). We observe that λ_c has little effect on the final performance after the *RL* optimization, when its value is in the range of $[1, 1.15]$; otherwise it has a negative influence on the final performance. However, λ_c has a positive impact on the performance after the *CE* training (See ablation analysis below for details). (2) λ_r is the weight for salient n-grams items in regularized reward function. As shown in Figure 2, we find that the model achieves the best performance when λ_r is set to 0.1 for MSCOCO. In addition, we observe that the generated sentences would tend to repetitively produce salient n-grams in a sentence, e.g., “A wooden table with chairs on a wooden table”, when λ_r is too large.

Ablation Analysis. We conduct an ablation analysis on FLICKR30K with AoANet, presenting results for various combinations of different training stages. As shown in Figure 3, the regularized CE (RCE) outperforms the basic CE on CIDEr-D and SPICE. This improvement, however, is mitigated after

Figure 2. Effect of λ_r on MSCOCO standard split.

Figure 3. Ablation results on FLICKR standard split.

ATTN: two bowls of food on a table with a table.
DISC: a tray of food and a bowl of food.
ATTN+SN: two containers of food with broccoli and a table.
GT: two plastic containers sitting on top of a wooden table.

ATTN: a man riding a bike down a street.
DISC: a man riding a bike down a street with a bicycle.
ATTN+SN: a man riding a bike down a street next to a train.
GT: a man riding a bike past a train traveling along tracks.

Figure 4. Captions generated by ATTN+SN, in comparison to ATTN, DISC and the ground truth (GT).

the RL optimization. Additionally, regularized RL (RRL) achieves better results than basic RL on CIDEr-D and SPICE, showing that RRL has a positive effect on the final performance.

Human Evaluation and Case Study We further perform human evaluation on our methods (ATTN+SN, DISC+SN) and two baselines (ATTN, DISC) trained on the MSCOCO Karpathy split using 50 images randomly selected from the test set. Three human subjects rate captions with the 1-5 Likert scale (higher is better) with respect to three criteria: fluency, descriptiveness and fidelity.

	Fluency	Descriptiveness	Fidelity
ATTN	3.90	2.53	3.46
ATTN+SN	3.96	2.86	3.50
DISC	3.52	3.08	3.28
DISC+SN	3.58	3.20	3.31

Table 3. Human evaluation between different models.

Figure (3) shows that ATTN+SN performs better than ATTN on fluency, descriptiveness, and fidelity. Moreover, DISC+SN can further improve the performance on descriptiveness as well.

Figure (4) includes two examples in which we compare our model (ATTN+SN) with ATTN, DISC and the ground truth. Our model produces captions with more detailed and key content; e.g., we generate more details such as “*container*”, “*broccoli*” in the first case and “*next to a train*” in the second.

5. Conclusions

We propose an effective approach to produce more descriptive captions using salient n-grams from cross-modal retrieval models. Specifically, we incorporate two types of salient n-grams, i.e., deletion n-grams and support n-grams, into the *CE* and *RL* learning objectives. The proposed model is shown to outperform the compared models on two widely used benchmarks on a wide range of metrics and human evaluation.

6. Acknowledgement

We would like to thank the anonymous reviewers for their valuable comments. This research is supported by NSERC Discovery Grants and DND Supplement.

References

- [1] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. “Show, attend and tell: Neural image caption generation with visual attention”. In: *International conference on machine learning*. 2015, pp. 2048–2057.
- [2] A. Karpathy and L. Fei-Fei. “Deep visual-semantic alignments for generating image descriptions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3128–3137.
- [3] J. Lu, C. Xiong, D. Parikh, and R. Socher. “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*. 2017, pp. 375–383.
- [4] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. “Bottom-up and top-down attention for image captioning and visual question answering”. In: *CVPR*. 2018, pp. 6077–6086.
- [5] J. Lu, J. Yang, D. Batra, and D. Parikh. “Neural baby talk”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7219–7228.
- [6] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. “Vse++: Improved visual-semantic embeddings”. In: *arXiv 2.7* (2017), p. 8.
- [7] R. C. Fong and A. Vedaldi. “Interpretable explanations of black boxes by meaningful perturbation”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 3429–3437.
- [8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. “Microsoft coco: Common objects in context”. In: *European conference on computer vision*. Springer. 2014, pp. 740–755.
- [9] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions”. In: *TACL 2* (2014), pp. 67–78.
- [10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. “Show and tell: A neural image caption generator”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3156–3164.
- [11] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov. “Review networks for caption generation”. In: *Advances in neural information processing systems*. 2016, pp. 2361–2369.
- [12] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. “Boosting image captioning with attributes”. In: *ICCV*. 2017, pp. 4894–4902.
- [13] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. “Image captioning with semantic attention”. In: *CVPR*. 2016.
- [14] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. “Self-critical sequence training for image captioning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 7008–7024.
- [15] B. Dai and D. Lin. “Contrastive learning for image captioning”. In: *NIPS*. 2017, pp. 898–907.
- [16] B. Dai, S. Fidler, R. Urtasun, and D. Lin. “Towards diverse and natural image descriptions via a conditional gan”. In: *ICCV*. 2017.
- [17] L. Liu, J. Tang, X. Wan, and Z. Guo. “Generating Diverse and Descriptive Image Captions Using Visual Paraphrases”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 4240–4249.
- [18] T. Yao, Y. Pan, Y. Li, and T. Mei. “Exploring visual relationship for image captioning”. In: *ECCV*. 2018, pp. 684–699.
- [19] Z. Shi, X. Zhou, X. Qiu, and X. Zhu. “Improving Image Captioning with Better Use of Caption”. In: *ACL*. 2020, pp. 7454–7464.
- [20] R. Luo, B. Price, S. Cohen, and G. Shakhnarovich. “Discriminability objective for training descriptive captions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6964–6974.
- [21] X. Liu, H. Li, J. Shao, D. Chen, and X. Wang. “Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data”. In: *ECCV*. 2018, pp. 338–354.
- [22] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He. “Stacked cross attention for image-text matching”. In: *ECCV*. 2018, pp. 201–216.
- [23] P. Anderson, B. Fernando, M. Johnson, and S. Gould. “Spice: Semantic propositional image caption evaluation”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 382–398.
- [24] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. “Cider: Consensus-based image description evaluation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4566–4575.

- [25] M. Denkowski and A. Lavie. “Meteor universal: Language specific translation evaluation for any target language”. In: *Proceedings of the ninth workshop on statistical machine translation*. 2014.
- [26] C.-Y. Lin. “Rouge: A package for automatic evaluation of summaries”. In: *Text Summarization Branches Out* (2004).
- [27] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. “BLEU: a method for automatic evaluation of machine translation”. In: *Proceedings of the 40th annual meeting on association for computational linguistics*. ACL. 2002, pp. 311–318.
- [28] L. Huang, W. Wang, J. Chen, and X.-Y. Wei. “Attention on attention for image captioning”. In: *ICCV*. 2019.
- [29] S. Ren, K. He, R. Girshick, and J. Sun. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *NIPS*. 2015.
- [30] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [31] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. “Visual genome: Connecting language and vision using crowdsourced dense image annotations”. In: *International Journal of Computer Vision* 123.1 (2017).