

Training for Diversity in Image Paragraph Captioning

Luke Melas-Kyriazi George Han Alexander M. Rush

School of Engineering and Applied Sciences
Harvard University

{lmelaskyriazi@college, hanz@college, srush@seas}.harvard.edu

Abstract

Image paragraph captioning models aim to produce detailed descriptions of a source image. These models use similar techniques as standard image captioning models, but they have encountered issues in text generation, notably a lack of diversity between sentences, that have limited their effectiveness. In this work, we consider applying sequence-level training for this task. We find that standard self-critical training produces poor results, but when combined with an integrated penalty on trigram repetition produces much more diverse paragraphs. This simple training approach improves on the best result on the Visual Genome paragraph captioning dataset from 16.9 to 30.6 CIDEr, with gains on METEOR and BLEU as well, without requiring any architectural changes.

1 Introduction

Image captioning aims to describe the objects, actions, and details present in an image using natural language. Most image captioning research has focused on single-sentence captions, but the descriptive capacity of this form is limited; a single sentence can only describe in detail a small aspect of an image. Recent work has argued instead for image paragraph captioning with the aim of generating a (usually 5-8 sentence) paragraph describing an image.

Compared with single-sentence captioning, paragraph captioning is a relatively new task. The main paragraph captioning dataset is the Visual Genome corpus, introduced by Krause et al. (2016). When strong single-sentence captioning models are trained on this dataset, they produce repetitive paragraphs that are unable to describe diverse aspects of images. The generated paragraphs repeat a slight variant of the same sentence multiple times, even when beam search is used. Prior work, discussed in the following section,

tried to address this repetition with architectural changes, such as hierarchical LSTMs, which separate the generation of sentence topics and words.

In this work, we consider an approach for training paragraph captioning models that focuses on increasing the diversity of the output paragraph. In particular, we note that self-critical sequence training (SCST) (Ranzato et al., 2015; Rennie et al., 2016), a technique which uses policy gradient methods to directly optimize a target metric, has been successfully employed in standard captioning, but not in paragraph captioning. We observe that during SCST training the intermediate results of the system lack diversity, which makes it difficult for the model to improve. We address this issue with a simple repetition penalty which down-weights trigram overlap.

Experiments show that this technique greatly improves the baseline model. A simple baseline, non-hierarchical model trained with repetition-penalized SCST outperforms complex hierarchical models trained with both cross-entropy and customized adversarial losses. We demonstrate that this strong performance gain comes from the combination of repetition-penalized search and SCST, rather than from either individually, and discuss how this impacts the output paragraphs.

2 Background and Related Work

Nearly all modern image captioning models employ variants of an encoder-decoder architecture. As introduced by Vinyals et al. (2014), the encoder is a CNN pre-trained for classification and the decoder is a LSTM or GRU. Following work in machine translation, Xu et al. (2015) added an attention mechanism over the encoder features. Recently, Anderson et al. (2017) further improved single-sentence captioning performance by incorporating object detection in the encoder (bottom-up attention) and adding an LSTM layer before attending to spatial features in the decoder (top-

down attention).

Single-sentence and paragraph captioning models are evaluated with a number of metrics, including some designed specifically for captioning (CIDEr) and some adopted from machine translation (BLEU, METEOR). CIDEr and BLEU measure accuracy with n-gram overlaps, with CIDEr weighting n-grams by TF-IDF (term-frequency inverse-document-frequency), and METEOR uses unigram overlap, incorporating synonym and paraphrase matches. We discuss these metrics in greater detail when analyzing our experiments.

Related Models Krause et al. (2016) introduced the first large-scale paragraph captioning dataset, a subset of the Visual Genome dataset, along with a number of models for paragraph captioning. Empirically, they showed that paragraphs contain significantly more pronouns, verbs, coreferences, and greater overall "diversity" than single-sentence captions. Whereas most single-sentence captions in the MSCOCO dataset describe only the most important object or action in an image, paragraph captions usually touch on multiple objects and actions.

The paragraph captioning models proposed by Krause et al. (2016) included template-based (non-neural) approaches and two encoder-decoder models. In both neural models, the encoder is an object detector pre-trained for dense captioning. In the first model, called the flat model, the decoder is a single LSTM which outputs an entire paragraph word-by-word. In the second model, called the hierarchical model, the decoder is composed of two LSTMs, where the output of one sentence-level LSTM is used as input to the other word-level LSTM.

Recently, Liang et al. (2017) extended this model with a third (paragraph-level) LSTM and added adversarial training. In total, their model (RTT-GAN) incorporates three LSTMs, two attention mechanisms, a phrase copy mechanism, and two adversarial discriminators. To the best of our knowledge, this model achieves state-of-the-art performance of 16.9 CIDEr on the Visual Genome dataset (without external data).

For our experiments, we use the top-down single-sentence captioning model in Anderson et al. (2017). This model is similar to the "flat" model in Krause et al. (2016), except that it incorporates attention with a top-down mechanism.

3 Approach

The primary issue in current paragraph captioning models, especially non-hierarchical ones, is lack of diversity of topic in the output paragraph. For example, for the image of a skateboarder in Figure 1, the flat model outputs "The man is wearing a black shirt and black pants" seven times. This example is not anomalous: it is a typical failure case of the model. Empirically, in validation, ground truth paragraphs contain 0.62 repeated trigrams on average, whereas paragraphs produced by the flat cross-entropy model contain 25.9 repeated trigrams on average.

3.1 Self-Critical Sequence Training

Self-critical sequence training (SCST) is a sequence-level optimization procedure proposed by Rennie et al. (2016), which has been widely adopted in single-sentence captioning but has not yet been applied to paragraph captioning. This method provides an alternative approach to word-level cross-entropy which can incorporate a task specific metric.

Sequence-level training employs a policy gradient method to optimize directly for a non-differentiable metric, such as CIDEr or BLEU. This idea was first applied to machine translation by Ranzato et al. (2015) in a procedure called MIXER, which incrementally transitions from cross-entropy to policy gradient training. To normalize the policy gradient reward and reduce variance during training, MIXER subtracts a baseline estimate of the reward as calculated by a linear regressor.

SCST replaces this baseline reward estimate with the reward obtained by the test-time inference algorithm, namely the CIDEr score of greedy search. This weights the gradient by the difference in reward given to a sampled paragraph compared to the current greedy output (see Eq. 3-9 in (Rennie et al., 2016)). Additionally, SCST uses a hard transition from cross-entropy to policy gradient training. The final gradient is:

$$-\mathbb{E}_{w^s \sim p_\theta} [(r(w^s) - r(w^g)) \nabla_\theta \log p_\theta(w^s | x)]$$

Where w^s is a sampled paragraph, w^g is a greedy decoded paragraph, r is the reward (e.g CIDEr), p_θ is the captioning model.




Input Image	Paragraph Caption		
	Trained with cross-entropy Tested without repetition penalty	Trained with cross-entropy Tested with repetition penalty	Trained with repetition penalty
	Two people are sitting on a bench. The elephant is a light brown color. The man is wearing a white shirt and a blue shirt. The man is sitting on a dirt ground. There is a large tree behind the man.	Two people are sitting on a bench. The elephant is a light brown color. The man is wearing a white shirt and a blue shirt. The woman is wearing a black shirt and black pants. There is a large tree behind the elephant.	Two people are sitting on a bench. The elephant is sitting on the dirt. The man is sitting on top of the elephant. The woman is wearing a white shirt. The man is wearing a black shirt. There is a tree behind the elephant. There are trees on the ground. There are trees in the background.
	A man is standing on a snow covered mountain. He is wearing a black jacket and black pants. The man is holding a ski poles in his hands. The man is wearing a black jacket and black pants. The man is holding a ski poles in his hands. The snow is covered in snow. The snow is covered in snow. The snow is covered in snow.	A man is standing on a snow covered mountain. He is wearing a black jacket and black pants. The man is holding a ski poles in his hands. The snow is covered in snow. There are trees in the background.	A man is standing on a snow covered mountain. The man is wearing a black jacket. The skier is wearing a black jacket. The man is holding a ski poles. The person is wearing black pants. The snow is white. The trees are covered in snow. The sky is covered in snow. There are trees on the ground.
	A man is skateboarding on a skateboard. He is wearing a black shirt and black pants. He is wearing a black cap and a black hat. The man is wearing a black shirt and black pants. The man is wearing a black shirt and black pants. The man is wearing a black shirt and black pants. The man is wearing a black shirt and black pants. The man is wearing a black shirt and black pants. The man is wearing a black shirt and black pants. The man is wearing a black shirt and black pants.	A man is skateboarding on a skateboard. He is wearing a black shirt and black pants. He has a black hat on his head. The man is wearing a white cap and a black cap. The ramp is made of concrete. The ramp has a white and white ramp on it.	A man is skateboarding on a skateboard. He is wearing a black shirt and black pants. The man is wearing a white hat. The man is jumping on a skateboard. The skateboard is black. The boy is wearing a hat. The ramp is black. There is a man on the ramp. There is a skater on the skateboard.

Figure 1: Example paragraph outputs of our model. The final example is a failure case of both our model and the non-penalized model. Our model does not suffer from the repetition problem of the non-penalized, but it does not produce a great caption because it does not understand that the image is black-and-white.

3.2 Repetition Penalty

In preliminary experiments, we find that directly applying SCST is not effective for paragraph captioning models. Table 1 shows that when training with SCST, the model performs only marginally better than cross-entropy. In further analysis, we see that the greedy baseline in SCST training has very non-diverse output, which leads to poor policy gradients. Unlike in standard image captioning, the cross-entropy model is too weak for SCST to be effective.

To address this problem, we take inspiration from recent work in abstractive text summarization, which encounters the same challenge when producing paragraph-length summaries of documents (Paulus et al., 2017). These models target the repetition problem by simply preventing the model from producing the same trigram more than once during inference. We therefore introduce an inference constraint that penalizes the log-probabilities of words that would result in repeated trigrams. The penalty is proportional to the number of times the trigram has already been gener-

ated.

Formally, denote the (pre-softmax) output of the LSTM by o , where the length of o is the size of the target vocabulary and o_w is the log-probability of word w . We modify o_w by $o_w \rightarrow o_w - k_w \cdot \alpha$, where k_w is the number of times the trigram completed by word w has previously been generated in the paragraph, and α is a hyperparameter which controls the degree of blocking. When $\alpha = 0$, there is no penalty, so we have standard greedy search. When $\alpha \rightarrow \infty$, or in practice when α exceeds about 5, we have full trigram blocking.

We incorporate this penalty into the greedy baseline used to compute policy gradients in SCST. During inference, we employ the same repetition-penalized greedy search.

4 Methods and Results

For our paragraph captioning model we use the top-down model from Anderson et al. (2017). Our encoder is a convolutional network pretrained for object detection (as opposed to dense captioning, as in Krause et al. (2016) and Liang et al. (2017)).

	METEOR	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Krause et al. (Template)	14.31	12.15	37.47	21.02	12.30	7.38
Krause et al. (Flat w/o object detector)	12.82	11.06	34.04	19.95	12.20	7.71
Krause et al. (Flat)	13.54	11.14	37.30	21.70	13.07	8.07
Krause et al. (Hierarchical)	15.95	13.52	41.90	24.11	14.23	8.69
Liang et al. (w/o discriminator)	16.57	15.07	41.86	24.33	14.56	8.99
Liang et al.	17.12	16.87	41.99	24.86	14.89	9.03
Ours (XE training, w/o rep. penalty)	13.66	12.89	32.78	19.00	11.40	6.89
Ours (XE training, w/ rep. penalty)	15.17	22.68	35.68	22.40	14.04	8.70
Ours (SCST training, w/o rep. penalty)	13.63	13.77	29.67	16.45	9.74	5.88
Ours (SCST training, w/ rep. penalty)	17.86	30.63	43.54	27.44	17.33	10.58

Table 1: Results of our model compared with prior published results. Note that Liang et al. (2017) also trains a model on additional data, but here we only compare models trained on Visual Genome. Also note that our models employ greedy search, whereas other models employ beam search.

The encoder extracts between 10 and 100 objects per image and applies spatial max-pooling to yield a single feature vector of dimension 2048 per object. The decoder is a 1-layer LSTM with hidden dimension 512 and top-down attention.

Evaluation is done on the Visual Genome dataset with the splits provided by Krause et al. (2016). We first train for 25 epochs with cross-entropy (XE) loss, using Adam with learning rate $5 \cdot 10^{-4}$. We then train an additional 25 epochs with repetition-penalized SCST targeting a CIDEr-based reward, using Adam with learning rate $5 \cdot 10^{-5}$.

Our PyTorch-based implementation is available at <https://github.com/lukemelas/image-paragraph-captioning>.

Results Table 1 shows the main experimental results. Our baseline cross-entropy captioning model gets similar scores to the original flat model. When the repetition penalty is applied to a model trained with cross-entropy, we see a large improvement on CIDEr and a minor improvement on other metrics.¹ When combining the repetition penalty with SCST, we see a dramatic improvement across all metrics, and particularly on CIDEr. Interestingly, SCST only works when its baseline reward model is strong; for this reason the combination of the repetition penalty and SCST is particularly effective.

Table 2 compares the effect of training with

¹We believe the improvement may be largest on CIDEr because, in the calculation of CIDEr, n-gram counts are clipped to the number of times each n-gram appears in the reference sentence. However, a similar procedure is applied in calculating BLEU, on which we show lesser improvements.

different values of the penalty hyperparameter α , demonstrating that intermediate values of α (≈ 2.0) perform slightly better than large values. An intermediate value of α discourages the model from producing repeat trigrams, but still permits the model to output them when there are no likely alternative phrases.

α	METEOR	CIDEr	BLEU-4
0.0	13.8	13.6	5.9
1.0	17.4	28.9	10.2
2.0	17.7	31.4	10.8
4.0	17.6	30.1	10.4
10.0	17.5	30.6	9.9

Table 2: Varying the repetition penalty α (on the validation set). $\alpha = 10$ is equivalent to trigram blocking.

	XE w/o penalty	XE w/ penalty	SCST w/ penalty
Avg. # of trigram repeats in output	25.9	0.67	3.70
Avg. # unique trigram overlaps btw. output and ground truth	2.23	2.97	3.49

Table 3: Analysis of different model outputs ($\alpha = 2.0$ for models w/ penalty)

Finally, Table 3 shows quantitative changes in trigram repetition and ground truth matches. The cross-entropy model fails to generate enough unique phrases. Blocking these entirely gives some benefit, but the SCST model is able to raise the total number of matched trigrams while reintroducing few repeats.

5 Conclusion

This work targets increased diversity in image paragraph captioning. We show that training with SCST combined with a repetition penalty leads to a substantial improvement in the state-of-the-art for this task, without requiring architectural changes or adversarial training. In future work, we hope to further address the language issues of paragraph generation as well as extend this simple approach to other tasks requiring long-form text or paragraph generation.

Acknowledgements

We would like to thank Ruotian Luo, Sang Phan, and all contributors to OpenNMT-py, for their work on open-source implementations of different image captioning, video captioning, and translation models. AMR is supported by NSF-CCF 1704834, Google, Facebook, Bloomberg, and Amazon research awards.

References

- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998*.
- Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2016. Re-evaluating automatic metrics for image captioning. *arXiv preprint arXiv:1612.07600*.
- Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2016. A hierarchical approach for generating descriptive image paragraphs. *arXiv preprint arXiv:1611.06607*.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.
- Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. 2017. Recurrent topic-transition gan for visual paragraph generation.
- 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.
- Ruotian Luo, Brian L. Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. *CoRR*, abs/1803.04376.
- Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.
- 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, Jarret Ross, and Vaibhava Goel. 2016. Self-critical sequence training for image captioning.
- 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.
- Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044.