Adversarial Inference for Multi-Sentence Video Description

Jae Sung Park[1], Marcus Rohrbach[2], Trevor Darrell[1], Anna Rohrbach[1]
[1] University of California, Berkeley, [2] Facebook AI Research

Abstract

While significant progress has been made in the image captioning task, video description is still in its infancy due to the complex nature of video data. Among the main issues are the fluency and coherence of the generated descriptions, and their relevance to the video. Recently, reinforcement and adversarial learning based methods have been explored to improve the image captioning models; however, both types of methods suffer from a number of issues, e.g. *poor readability and high redundancy for RL and stability issues for GANs. In this work, we instead propose to apply adversarial techniques during inference, designing a discriminator which encourages better multi-sentence video description. In addition, we find that a multi-discriminator "hybrid" design, where each discriminator targets one aspect of a description, leads to the best results. Specifically, we decouple the discriminator to evaluate on three criteria: 1) visual relevance to the video, 2) language diversity and fluency, and 3) coherence across sentences. Our approach results in more accurate, diverse, and coherent multi-sentence video descriptions, as shown by automatic as well as human evaluation on the popular ActivityNet Captions dataset.* [1]

1. Introduction

Video captioning remains a highly challenging problem, despite high interest in the task and ongoing emergence of new datasets [4, 6, 19] and approaches [16, 17, 20]. Consider the outputs of the three recent video description methods on an example video from the ActivityNet Captions dataset [1, 6] in Figure 1. We notice that there are multiple issues with these descriptions, in addition to the errors with respect to the video content: there are semantic inconsistencies and lack of diversity within sentences, as well as redundancies across sentences.

To overcome these problems, we propose *Adversarial Inference* for video description, which relies on a discriminator to improve the description quality. Specifically, we are interested in the task of multi-sentence video description [13, 18], *i.e.* the output of our model is a paragraph that describes a video. We assume that the ground-truth temporal segments are given, *i.e.* we do not address the event detection task, but focus on obtaining a coherent multi-sentence description.

Figure 1: Comparison of the state-of-the-art video description approaches, Transformer [20], VideoStory [4], MoveForwardTell [16], and our proposed *Adversarial Inference*. Our approach generates more interesting and accurate descriptions with less redundancy. Red/bold indicates content errors, blue/italic indicates repetitive patterns, underscore highlights more interesting phrases.

2. Approach

Recent works in video captioning directly optimize for the sentence metrics using reinforcement learning based methods [8, 12]. We instead explicitly train a discriminator that scores the descriptions generated by generator for a given video. This includes, among others, to measure whether the multi-sentence descriptions are (1) correct with respect to the video, (2) fluent within individual sentences, and (3) form a coherent story across sentences. In a "single discriminator" design, the discriminator is given multiple tasks at once, *i.e.* to detect generated "fakes", which

[1] https://github.com/jamespark3922/adv-inf.

Method	Per video			Overall		Per act.	Per video		
	METEOR	BLEU@4	CIDEr-D	Vocab Size	Sent Length	RE-4 ↓	Div-1 ↑	Div-2 ↑	RE-4 ↓
MLE	16.70	9.95	20.32	1749	13.83	0.38	0.55	0.74	0.08
SCST	15.80	10.82	20.89	941	12.13	0.52	0.47	0.65	0.11
GAN	16.69	10.02	21.07	1930	13.60	0.36	0.56	0.74	0.07
MLE + SingleDis	16.29	9.25	18.17	2291	13.98	0.37	0.59	0.75	0.07
MLE + HybridDis (Ours)	16.48	9.91	20.60	2346	13.38	0.32	0.59	0.77	0.06
Human	-	-	-	8352	14.27	0.04	0.71	0.85	0.01

Table 1: Comparison to video description baselines Statistics over generated descriptions include N-gram Diversity (Div-1,2, higher better) and Repetition (RE-4, lower better) per video and per activity.

requires looking at linguistic characteristics, such as diversity or language structure, as well the visually mismatched "fakes", which requires looking at sentence semantics and relate it to the visual features. Moreover, for multi-sentence description, we would also like to detect cases where a sentence is inconsistent or redundant to a previous sentence. To obtain these properties, we find it to important to decouple the different tasks and allocate an individual discriminator for each one. In particular, we design visual, language, and pairwise discriminators, which jointly constitute our *hybrid discriminator*.

While adversarial learning methods [5] mostly rely on discriminators to train the generator for caption generation [3, 14], we argue that using them during inference is a more robust way of improving over the original generator. In our *Adversarial Inference*, the pre-trained generator presents discriminator with the sentence candidates by sampling from its probability distribution, and our *hybrid* discriminator selects the best sentence relying on the combination of its sub-discriminators.

3. Results

We benchmark our approach for multi-sentence video description on the ActivityNet Captions dataset [6]. We compare our Adversarial Inference (MLE+HybridDis) to: our baseline generator trained with maximum likelihood objective (MLE); Self Critical Sequence Tranining [12] which optimizes for CIDEr (SCST); GAN model built off [2, 9] with a single discriminator (GAN); inference with the single discriminator (MLE+SingleDis).

Automatic Evaluation.

Following [16], we conduct our evaluation at paragraph-level. We include standard metrics, *i.e*. METEOR [7], BLEU@4 [10] and CIDEr-D [15]. To see if our approach improves on content diversity and repetition, we report Div-1 and Div-2 scores [14], that measure a ratio of unique N-grams (N=1,2) to the total number of words, and RE-4 [16], that captures a degree of N-gram repetition (N=4) in a description[2]. Finally, to capture the degree of "discriminativeness" among the descriptions of videos with similar content, we also report RE-4 per activity label by combining all sentences associated with the same activity, and averaging the score over all activities.

Method	Better than MLE	Worse than MLE	Delta
SCST	22.0	62.0	-40.0
GAN	32.5	30.0	+2.5
MLE + SingleDis	29.0	30.0	-1.0
MLE + HybridDis (Ours)	38.0	31.5	**+6.5**

Table 2: Human evaluation of multi-sentence video descriptions, see text for details.

We compare our model to baselines in Table 1. In the standard metrics, we see that there is no significant difference across the models; however, our Adversarial Inference leads to more diverse descriptions with less repetition per video and activity than the baselines, including GANs. SCST, on the other hand, has the lowest diversity and highest repetition among all baselines. Our MLE+HybridDis model also outperforms the MLE+SingleDis in every metric, supporting our hybrid discriminator design.

Human Evaluation. We run our human evaluation on Amazon Mechanical Turk (AMT) with a set of 200 random videos, and compare each system to the MLE baseline. We ask 3 human judges to select that one description is better than another or that both as similar, and compute a majority vote (*i.e*. at least 2 out of 3 agree on a judgment). In Table 2, our proposed approach improves over all baselines. In particular, we see that the GAN is rather competitive, but still overall not scored as high as our approach. Notably, SCST is scored rather low, which we attribute to its grammatical issues and high redundancy in the descriptions.

4. Conclusion

We propose *Adversarial Inference*, where a discriminator selects the best from a set of sampled sentences, and introduce a *hybrid discriminator* which consists of three individual experts. For more details, qualitative results and comparison to the state-of-the-art models, please, see the full paper [11][3].

[2]For Div-1,2 higher is better, while for RE-4 lower is better.

[3]https://arxiv.org/abs/1812.05634

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[2] Chen Chen, Shuai Mu, Wanpeng Xiao, Zexiong Ye, Liesi Wu, Fuming Ma, and Qi Ju. Improving image captioning with conditional generative adversarial nets. *arXiv:1805.07112*, 2018.

[3] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[4] Spandana Gella, Mike Lewis, and Marcus Rohrbach. A dataset for telling the stories of social media videos. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 968–974, 2018.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014.

[6] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 706–715, 2017.

[7] Michael Denkowski Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, page 376, 2014.

[8] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, volume 3, page 3, 2017.

[9] Igor Melnyk, Tom Sercu, Pierre L Dognin, Jarret Ross, and Youssef Mroueh. Improved image captioning with adversarial semantic alignment. *arXiv:1805.00063*, 2018.

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.

[11] Jae Sung Park, Marcus Rohrbach, Trevor Darrell, and Anna Rohrbach. Adversarial inference for multi-sentence video description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[12] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *Proceedings of the German Confeence on Pattern Recognition (GCPR)*, 2014.

[14] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[15] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[16] Yilei Xiong, Bo Dai, and Dahua Lin. Move forward and tell: A progressive generator of video descriptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[17] Huanyu Yu, Shuo Cheng, Bingbing Ni, Minsi Wang, Jian Zhang, and Xiaokang Yang. Fine-grained video captioning for sports narrative. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6006–6015, 2018.

[18] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[19] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, 2018.

[20] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8739–8748, 2018.