Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions

Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara
University of Modena and Reggio Emilia
Modena, Italy

ABSTRACT

Current captioning approaches can describe images using black-box architectures whose behavior is hardly controllable and explainable from the exterior. As an image can be described in infinite ways depending on the goal and the context at hand, a higher degree of controllability is needed to apply captioning algorithms in complex scenarios. In our paper, we introduced a novel framework for image captioning which can generate diverse descriptions by allowing both grounding and controllability. Given a control signal in the form of a sequence or set of image regions, we generate the corresponding caption through a recurrent architecture which predicts textual chunks explicitly grounded on regions, following the constraints of the given control. Experimental results demonstrate that our method achieves state of the art performances on controllable image captioning, in terms of caption quality and diversity. The source code is publicly available at: https://github.com/aimagelab/show-control-and-tell.

KEYWORDS

controllable captioning, image captioning, vision and language

ACM Reference Format:
Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara. 2019. *Show, Control and Tell*: A Framework for Generating Controllable and Grounded Captions. In *ACM Celebration of Women in Computing: womENcourage 2019, September 16–18, 2019, Rome, Italy*. ACM, New York, NY, USA, 1 page.

Image captioning brings vision and language together in a generative way. As a fundamental step towards machine intelligence, this task has been recently gaining much attention thanks to the spread of Deep Learning architectures which can effectively describe images in natural language [1, 3]. Image captioning approaches are usually capable of learning a correspondence between an input image and a probability distribution over time, from which captions can be sampled either using a greedy decoding strategy, or more sophisticated techniques like beam search and its variants.

As the two main components of captioning architectures are the image encoding stage and the language model, researchers have focused on improving both phases, which resulted in the emergence of attentive models on one side, and of more sophisticated interactions with the language model on the other. Recently, attentive models have been improved by replacing the attention over a grid of features with attention over image regions [1]. In these models, the generative process attends a set of regions which are softly selected while generating the caption.

Despite these advancements, captioning models still lack controllability and explainability – *i.e.*, their behavior can hardly be influenced and explained. As an example, in the case of attention-driven models, the architecture implicitly selects which regions to focus on at each timestep, but it cannot be supervised from the exterior. While an image can be described in multiple ways, such an architecture provides no way of controlling which regions are described and what importance is given to each region. This lack of controllability creates a distance between human and machine intelligence, as humans can manage the variety of ways in which an image can be described, and select the most appropriate one depending on the task and the context at hand. Most importantly, this also limits the applicability of captioning algorithms to complex scenarios in which some control over the generation process is needed. As an example, a captioning-based driver assistance system would need to focus on dangerous objects on the road to alert the driver, rather than describing the presence of trees and cars when a risky situation is detected. Eventually, such systems would also need to be explainable, so that their behavior could be easily interpreted in case of failures.

In our paper [2], we introduced *Show, Control and Tell*, that explicitly addresses these shortcomings. It can generate diverse natural language captions depending on a control signal which can be given either as a sequence or as a set of image regions which need to be described. As such, our method is capable of describing the same image by focusing on different regions and in a different order, following the given conditioning. Our model is built on a recurrent architecture which considers the decomposition of a sentence into noun chunks and models the relationship between image regions and textual chunks, so that the generation process can be explicitly grounded on image regions. To the best of our knowledge, this is the first captioning framework controllable from image regions. Experimental results demonstrate that our proposed method achieves state of the art results for controllable image captioning both in terms of diversity and caption quality, on different datasets commonly used for the image captioning task.

REFERENCES

[1] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[2] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[3] 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*.

womENcourage '19, September 16–18, 2019, Rome, Italy