Contextual knowledge is important: utilizing top-down information in generating structured and ordered image paragraphs

Anonymous ACL submission

Abstract

abstract

1 Introduction

Humans are typically able to effortlessly describe real-world images when required: we easily identify objects, attributes and relations between them. Diversity, richness and complexity of such human-produced image descriptions have been observed in several benchmark image description datasets, including MSCOCO (Lin et al., 2014; Chen et al., 2015), Flickr8k (Hodosh et al., 2013), Flickr30k (Young et al., 2014), Visual Genome (Krishna et al., 2016). These datasets were collected to address the task of automatic image description (Bernardi et al., 2016), a long-standing and active field of research, placed in intersection between computer vision and natural language processing (generation, in particular). This problem of mapping visual data to text can be viewed as the specific example of one of the core goals of NLG: 'translating' source data into a natural language representation.

In natural language generation community, the task of text generation has been typically divided into multiple sub-tasks, including *content determination (selection)*, the task of deciding which parts of the source information should be included in the output description, and *text structuring*, the task of ordering selected information (Gatt and Krahmer, 2017). However, with the rise of neural networks in many NLP areas, the generation tasks are now seen as a continuous, non-modular process of automatically learning relations between input and expected output. Specifically, neural models of image captioning (Kiros et al., 2014; Vinyals et al., 2014) are trying to implicitly learn what is important about an image (content selection) and how this information should be structured in the generated caption (text structuring). Such mechanisms as attention (Xu et al., 2015; Anderson et al., 2017) further improve ability of the models to locate important parts in an image and utilize them for caption generation. Some recent advances in image captioning include application of transformer architecture (Vaswani et al., 2017; Herdade et al., 2019).

While it is clear that neural networks demonstrate good performance in generating *well-structured* single-sentence image captions with *relevant knowledge*, the problem of selecting and ordering information becomes significantly harder when generating multiple sentences for a single image. The corresponding task of *image paragraph generation* has been initially introduced in (Krause et al., 2017), proposing the challenge of generating a text, consisting of several sentences that would form a coherent whole. Most of the following work (Liang et al., 2017; Chatterjee and Schwing, 2018; Wang et al., 2019) has focused on *generating* good, diverse and human-like paragraphs as measured by various automatic evaluation metrics like BLEU (Papineni et al., 2002) or CIDEr (Vedantam et al., 2014).

In this paper we look at the different aspect of image paragraph generation and address the problem of **information order** in the multi-sentence image captioning setting. We argue that utilizing top-down knowledge (information about context available to the captioner) is beneficial for the task of image paragraph generation. We show that the model conditioned on both low-level visual features and high-level top-down information is able to learn human-like distributions of attended objects, attributes and relations generated in the paragraph. We introduce several image paragraph models based on the hierarchical paragraph generator Krause et al. (2017) and also demonstrate that using bottom-up information exclusively is not enough to learn good paragraph structure. We employ transfer learning and use model pre-trained for the dense image captioning task (Johnson et al., 2016) to obtain representations of background information, which we treat as our top-down features. We evaluate how close our models are compared to the human performance in terms of attending to objects in visual scenes in a particular order.

2 Related Work

[Nikolai: This whole section needs to be shrinked. Or not?]

Human visual attention Humans are quite efficient in detecting objects and separating them from the rest of the visual scene (Ullman, 1987). We are also fluent in using visual attention, which allows us to single out particular objects in the visual scene, significantly reducing our perceptual load when needed (Lavie et al., 2004), therefore, preventing us from being overwhelmed by typically complex real-world visual scenes. Our ability to attend to particular parts of the environment is based on both bottom-up information (low-level visual stimuli) and top-down information (high-level goal-related stimuli, discourse knowledge) (Zarcone et al., 2016). Stimuli that attracts our attention is said to be *salient* (relevant). Salience of objects affects our *surprisal* towards particular visual input: discourse-salient entities cause less surprisal (e.g. 'bed' in bedroom) unlike the visually salient objects (e.g. 'large pink elephant' in bedroom).

Attention has also been employed in formal theories of interaction. One of the approaches has been proposed by Dobnik and Kelleher (2016), who link attention with judgements as defined in Type Theory of Records (Cooper, 2008). They introduce a Bayesian-based framework in which attention controls the extent to which context induced judgements ($\sim$ task-based top-down information) are utilized by an agent. This allows for topic modelling at each timestamp in interaction. Thus, it follows along the lines of our proposal about attention using both contextual and low-level visual information in selecting relevant information for each individual sentence in the image paragraph.

Neural image paragraph captioning The task of generating more complex descriptions of images such as paragraphs has been introduced in Krause et al. (2017) along with the dataset of image-paragraph pairs. The paper adopts a hierarchical structure for the model of paragraph generation: sentence RNN is conditioned on visual features and unrolled for each sentence in the paragraph, giving a sentence topic as its output. Then, each of these topics is used by another RNN to generate actual sentences. We start by implementing this hierarchical image paragraph model, since it inherits the modular nature of human image paragraph production (given an image, plan structure of your paragraph and identify its sentence topics, then incrementally generate sentences). Liang et al. (2017) use similar hierarchical network in addition with adversarial discriminator, that forces model to generate realistic paragraphs with smooth transitions between sentences. Chatterjee and Schwing (2018) also address cross-sentence topic consistency by modelling global coherence vector, conditioned on all sentence topics. Different from these approaches, Melas-Kyriazi et al. (2019) employ self-critical training technique (Rennie et al., 2016) to directly optimize a target evaluation metric for image paragraph generation. Lastly, Wang et al. (2019) use convolutional auto-encoder for topic modelling based on region-level image features. They demonstrate that extracted topics are more representative and contain information relevant for sentence generation. In this paper we similarly model better topic representations. However, we use additional language representations as part of the input to our topic generator, which is an LSTM.

Language attention in language and vision models [Nikolai: wordy section, needs to be shorter, keep all information here for now]

Only a limited number of models for image captioning has been supplied with both visual and background information for caption generation. You et al. (2016) detect visual concepts found in the scene (objects, attributes, etc.) and extract top-down visual features. Both of these modalities are then fed to the RNN-based caption generator. Attention is applied on detected concepts to inform generator about how relevant a particular concept is at each timestamp. Different to their model, we do not use any attribute detectors to identify objects in the scene, instead relying on the output of the model pre-trained for the task of dense captioning. Lu et al. (2016a) emphasize that image is not always useful in generating some function words ('of', 'the', etc.). They introduce adaptive attention, which determines when to look at the image and when it is more important to use the language model to generate the next word. In their work, the attention vector is the mixture of visual features and visual sentinel, a vector obtained through the additional gate function on decoder memory state. Our model is focused on a similar task: we are interested in deciding which type information is more important at a particular timestamp, but we also look at how *merging* two modalities into a single representation performs and how it affects attention of the model. Closest to our work is the work by (Liang et al., 2017), who apply language attention on region captions and use it to assist recurrent word generation in producing sentences in a paragraph. They embed region descriptions into the same embedding space that their word RNN is operating on. While we also believe that feeding information about semantic concepts found in an image is beneficial for the model, we propose to employ transfer learning. We use hidden states of the RNN trained for the task of dense captioning (Johnson et al., 2016) as our background information representation. Outside of image paragraph captioning, Lu et al. (2016b) have proposed a joint image and question attention model for the task of visual question answering. [Nikolai: Any work on language attention in visual dialog? I think one sentence with some citations would be nice to have.]

3 Approach

[Nikolai: The stuff below needs to be updated! Nothing important to read there (yet)!]

In our experiments we largely adopt architecture of the hierarchical paragraph generation model described in (Krause et al., 2017), applying numerous changes. For all our models, we change the stopping probability threshold parameter T_{STOP} from 0.5 to 0.4. At each timestamp, both sentence LSTM and word LSTM use hidden states from previous timestamps respectively. Hyperparameters of all our models are identical to the ones reported in the original paper. All models are implemented in PyTorch.

Baseline As our baseline, we implement hierarchical model described in the original paper. The only difference is that we use a single fully-connected layer to obtain input to the word LSTM, while the original paper uses two.

No-FC Our second variant of the model does not use any fully-connected layers between sentence LSTM and word LSTM, directly passing current hidden state h_t of sentence LSTM as input to the word LSTM. Such change is supposed to reduce complexity of the model.

DC-wordLSTM Our third model has two layers in word LSTM, where the first layer is initialised with the DenseCap RNN weights and embeddings. We follow similar transfer learning strategy described in the original paper.

All training is done via teacher forcing. During inference stage, we use predicted word as an input at the next timestamp. To predict the word, we test multiple decoding strategies and observe that nucleus sampling (p = 0.9) (Holtzman et al., 2019) with temperature over softmax (0.5) gives us the most interesting and coherent descriptions compared to the other decoding algorithms.

4 Conclusion

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 visual question answering.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures.

Moitreya Chatterjee and Alexander G. Schwing. 2018. Diverse and coherent paragraph generation from images. In *ECCV*.

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

Robin Cooper. 2008. Type theory with records and unification-based grammar. In *Logics for Linguistic Structures, pages 9 – 34. Mouton de Gruyter*.

Simon Dobnik and John D. Kelleher. 2016. A model for attention-driven judgements in type theory with records.

Albert Gatt and Emiel Krahmer. 2017. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res.*, 47(1):853–899.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 595–603, Bejing, China. PMLR.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Computer Vision and Patterm Recognition (CVPR)*.

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.

Nilli Lavie, Aleksandra Hirst, Jan W de Fockert, and Essi Viding. 2004. Load theory of selective attention and cognitive control. *Journal of experimental psychology. General*, 133(3):339—354.

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, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft coco: Common objects in context.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2016a. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016b. Hierarchical Question-Image Co-Attention for Visual Question Answering.

Luke Melas-Kyriazi, Alexander Rush, and George Han. 2019. Training for Diversity in Image Paragraph Captioning.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2016. Self-critical Sequence Training for Image Captioning.

S. Ullman. 1987. Visual routines. In M. A. Fischler and O. Firschein, editors, *Readings in Computer Vision: Issues, Problems, Principles, and Paradigms*, pages 298–328. Kaufmann, Los Altos, CA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator.

Jing Wang, Yingwei Pan, Ting Yao, Jinhui Tang, and Tao Mei. 2019. Convolutional auto-encoding of sentence topics for image paragraph generation.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image Captioning with Semantic Attention.

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.

Alessandra Zarcone, Marten van Schijndel, Jorrig Vogels, and Vera Demberg. 2016. Salience and attention in surprisal-based accounts of language processing.

A Appendices

Appendices are material that can be read, and include lemmas, formulas, proofs, and tables that are not critical to the reading and understanding of the paper. Appendices should be **uploaded as supplementary material** when submitting the paper for review. Upon acceptance, the appendices come after the references, as shown here.

LaTeX-specific details: Use `\appendix` before any appendix section to switch the section numbering over to letters.

B Supplemental Material

Submissions may include non-readable supplementary material used in the work and described in the paper. Any accompanying software and/or data should include licenses and documentation of research review as appropriate. Supplementary material may report preprocessing decisions, model parameters, and other details necessary for the replication of the experiments reported in the paper. Seemingly small preprocessing decisions can sometimes make a large difference in performance, so it is crucial to record such decisions to precisely characterize state-of-the-art methods.

Nonetheless, supplementary material should be supplementary (rather than central) to the paper. **Submissions that misuse the supplementary material may be rejected without review.** Supplementary material may include explanations or details of proofs or derivations that do not fit into the paper, lists of features or feature templates, sample inputs and outputs for a system, pseudo-code or source code, and data. (Source code and data should be separate uploads, rather than part of the paper).

The paper should not rely on the supplementary material: while the paper may refer to and cite the supplementary material and the supplementary material will be available to the reviewers, they will not be asked to review the supplementary material.