Interactive Machine Learning for Image Captioning

Mareike Hartmann,[1] **Aliki Anagnostopoulou,**[1] **Daniel Sonntag** [1]

[1] German Research Center for Artificial Intelligence (DFKI), Germany
mareike.hartmann@dfki.de, aliki.anagnostopoulou@dfki.de, daniel.sonntag@dfki.de

Abstract

We propose an approach for interactive learning for an image captioning model. As human feedback is expensive and modern neural network based approaches often require large amounts of supervised data to be trained, we envision a system that exploits human feedback as good as possible by multiplying the feedback using data augmentation methods, and integrating the resulting training examples into the model in a smart way. This approach has three key components, for which we need to find suitable practical implementations: feedback collection, data augmentation, and model update. We outline our idea and review different possibilities to address these tasks.

Introduction

Our goal is to improve an image captioning system (Biswas, Barz, and Sonntag 2020) by learning from human feedback. The envisioned use case is an image captioning system initially trained on publicly available image captioning data, e.g., the MS COCO dataset (Lin et al. 2014), that can be adjusted to user-specific data based on corrective feedback provided by the end-user, ideally after deployment. To make use of user-specific feedback in an efficient way, we suggest to use data augmentation techniques to build additional training examples based on the user feedback, and then find the most suitable way to update model parameters based on the additional training data. This approach raises three research questions that we plan to address:

1. What type of feedback is most useful and how can it be collected?
2. Which augmentation strategies are most useful to maximize the impact of the feedback?
3. How can the feedback best be integrated into the training process?

We expect the answers to these questions to be interdependent, as, e.g., some augmentation strategies might fit well with specific types of feedback. Also, best solutions might differ depending on the specific type of captioning model used. In the following, we briefly review two image captioning architectures that we plan to work with, and then discuss several ideas for implementing our approach.

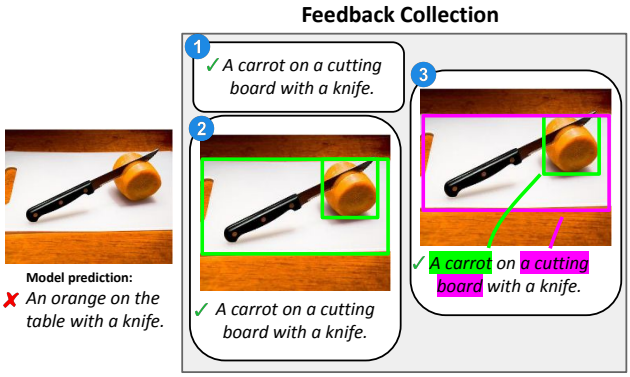

Figure 1: Examples for three different types of feedback: (1) corrected caption, (2) corrected caption with image annotations, (3) additional explicit alignment between objects and corrected caption words.

Approach

Our interactive learning approach for image captioning aims at improving a model based on user feedback, focusing on feedback collection, data augmentation, and feedback integration, rather than developing a new model architecture for image captioning. We consider using two different types of captioning models for our study: standard encoder-decoder models, such as Show-Attend-and-Tell (Xu et al. 2015) or the Top-Down Bottom-Up Attention model (Anderson et al. 2018) generate captions by feeding an attention-weighted representation of visual features, e.g., the output of a pre-trained object detection model, to the decoder at each time-step. The second family of models explicitly forces the decoder to generate words that are grounded in the image, which means that they refer to concepts detected in the image in a first step (Lu et al. 2018; Cornia, Baraldi, and Cucchiara 2019; Chen et al. 2020, 2021). The former models can be trained on pairs of images and corresponding captions. The latter models require finer-grained supervision, such as alignments between image regions and (sub)phrases, or scene graphs, but we expect them to be more suitable for our purpose, as they seem to offer a more direct way of integrating user feedback.

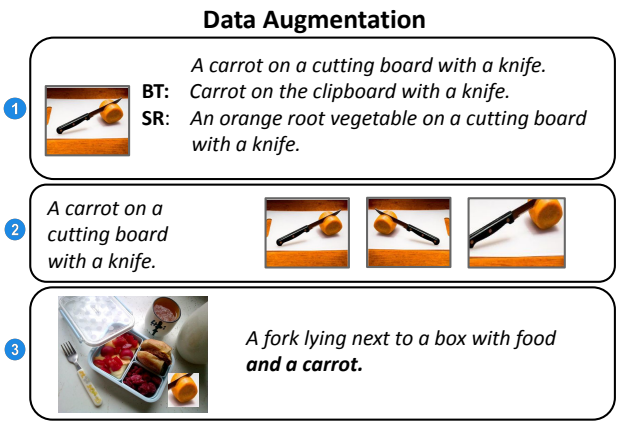

Figure 2: Examples of three different types of data augmentation: (1) caption-based including back-translation (BT) and synonym replacement (SR), (2) image-based, and (3) multi-modal by inserting patches cropped from a different image and adjusting the caption accordingly (added tokens are marked in bold).

Feedback collection The first step is the collection of user feedback for model predictions on user-specific data. The user could provide different types of feedback of varying complexity (see Figure 1), that might require more or less (cognitive) user efforts. Our goal should be to find a good balance between collecting rich feedback while not straining the user too much and keeping them engaged in the task, which makes it obligatory to implement a well-suited user interface. The least complex feedback type (in terms of user effort and requirements to the collection interface) would be a corrected image caption. Richer feedback could be collected by asking the user to additionally mark or select objects or image regions that the model incorrectly omitted from the generated caption, or even explicitly mark the alignment between corrected caption words and corresponding images. This type of feedback might be most useful for the controlled generation models.

Another question to address is whether we can increase the impact of user feedback on model performance by requesting feedback for specific examples selected by a suitable deep active learning acquisition function (Asghar et al. 2017; Siddhant and Lipton 2018; Lowell, Lipton, and Wallace 2019). Shen, Kar, and Fidler (2019) propose to collect feedback via an agent that learns to ask questions about specific visual concepts that it is uncertain about, which are answered by a human and used to construct new training examples. Their findings suggest that answering questions reduces human cost compared to providing a new caption and might be a reasonable alternative way of feedback collection to be explored.

Data augmentation An image-caption example paired with user feedback initially constitutes one additional training example for the captioning model. In the data augmentation step, we want to exploit the feedback to generate a larger amount of additional training examples. To this end, different augmentation strategies are applicable, either caption-based, image-based, or a combination of both modalities (see Figure 2).

For *image-based* augmentation, several image transformations such as cropping, warping, and flipping have previously been applied for image captioning (Wang, Yang, and Meinel 2018; Takahashi, Matsubara, and Uehara 2019; Katiyar and Borgohain 2021). Such image transformations are possible, but might introduce noise in the form of mismatch between image and caption, for example if spatial relations are referred to in the caption (... *is to the right of* ...).

For *caption-based* augmentation, Atliha and Šešok (2020) explored synonym replacement (McCarthy and Navigli 2007; Zhang, Zhao, and LeCun 2015) and paraphrasing using a pre-trained language model (Devlin et al. 2019). Other text augmentation methods might be applicable, including random insertion/deletion/swapping of words (Wei and Zou 2019), and backtranslation (Sennrich, Haddow, and Birch 2016). For both modalities, retrieval-based augmentation from additional resources is possible as well (Li et al. 2021).

Another research direction is *multi-modal* augmentation, where instead of modifying either caption or image, both modalities are modified at the same time. Feng et al. (2021) bring up the idea of combining caption editing with image manipulation based on the CutMix method (Yun et al. 2019), which augments data for image classification and object localization by inserting image patches (showing parts of objects) cut out from a different image, and respectively adjusting the label distribution. An adaptation for image captioning would require a re-write of the caption such that it correctly describes the modified image, which could either be done by replacing single words or by combining parts of different captions, similar to instance crossover augmentation proposed by Luque (2019).

Model update Once we have generated a reasonable amount of training data based on the user feedback, we need to update the model based on these new examples. Instead of re-training the model from scratch on the augmented training dataset (Shen, Kar, and Fidler 2019), we are more interested in the continual learning paradigm of batch-wise model updates, that allows us to adjust the model to new information more efficiently (see Riccardo Del Chiaro (2020) for continual learning approaches applicable for image captioning). Ling and Fidler (2017) propose an approach for interactive image captioning that integrates phrase-level human feedback using reinforcement learning methods (Rennie et al. 2017), which we plan to compare to gradient-based methods. The main challenges we need to address in the continual learning setting include avoiding catastrophic forgetting of previously learned knowledge (Goodfellow et al. 2013), expanding the decoder to account for user-specific vocabulary, and integrating information about novel objects that were not observed previously (Hendricks et al. 2016). For the latter problem, we plan to include experiments with inference-time strategies for novel object captioning (Anderson et al. 2017; Zheng, Li, and Wang 2019).

Acknowledgments

The research was funded by the XAINES project (BMBF, 01IW20005).

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2017. Guided Open Vocabulary Image Captioning with Constrained Beam Search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 936–945.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 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*, 6077–6086.

Asghar, N.; Poupart, P.; Jiang, X.; and Li, H. 2017. Deep Active Learning for Dialogue Generation. *c2017 The Association for Computational Linguistics*, 78.

Atliha, V.; and Šešok, D. 2020. Text augmentation using BERT for image captioning. *Applied Sciences*, 10(17): 5978.

Biswas, R.; Barz, M.; and Sonntag, D. 2020. Towards Explanatory Interactive Image Captioning Using Top-Down and Bottom-Up Features, Beam Search and Re-ranking. *Künstliche Intell.*, 34(4): 571–584.

Chen, L.; Jiang, Z.; Xiao, J.; and Liu, W. 2021. Human-like Controllable Image Captioning with Verb-specific Semantic Roles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16846–16856.

Chen, S.; Jin, Q.; Wang, P.; and Wu, Q. 2020. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9962–9971.

Cornia, M.; Baraldi, L.; and Cucchiara, R. 2019. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8307–8316.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Feng, S. Y.; Gangal, V.; Wei, J.; Chandar, S.; Vosoughi, S.; Mitamura, T.; and Hovy, E. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.

Goodfellow, I. J.; Mirza, M.; Xiao, D.; Courville, A.; and Bengio, Y. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

Hendricks, L. A.; Venugopalan, S.; Rohrbach, M.; Mooney, R.; Saenko, K.; and Darrell, T. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–10.

Katiyar, S.; and Borgohain, S. K. 2021. Image Captioning using Deep Stacked LSTMs, Contextual Word Embeddings and Data Augmentation. *arXiv preprint arXiv:2102.11237*.

Li, G.; Zhai, Y.; Lin, Z.; and Zhang, Y. 2021. Similar Scenes arouse Similar Emotions: Parallel Data Augmentation for Stylized Image Captioning. *arXiv preprint arXiv:2108.11912*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Ling, H.; and Fidler, S. 2017. Teaching machines to describe images via natural language feedback. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 5075–5085.

Lowell, D.; Lipton, Z. C.; and Wallace, B. C. 2019. Practical Obstacles to Deploying Active Learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 21–30.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2018. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7219–7228.

Luque, F. M. 2019. Atalaya at TASS 2019: Data Augmentation and Robust Embeddings for Sentiment Analysis. *CoRR*, abs/1909.11241.

McCarthy, D.; and Navigli, R. 2007. SemEval-2007 Task 10: English Lexical Substitution Task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 48–53. Prague, Czech Republic: Association for Computational Linguistics.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7008–7024.

Riccardo Del Chiaro, A. D. B. J. v. d. W., Bartłomiej Twardowski. 2020. RATT: Recurrent Attention to Transient Tasks for Continual Image Captioning. In *Proceedings of the 37 the International Conference on Machine Learning*.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 86–96. Berlin, Germany: Association for Computational Linguistics.

Shen, T.; Kar, A.; and Fidler, S. 2019. Learning to caption images through a lifetime by asking questions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10393–10402.

Siddhant, A.; and Lipton, Z. C. 2018. Deep Bayesian Active Learning for Natural Language Processing: Results of a Large-Scale Empirical Study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2904–2909.

Takahashi, R.; Matsubara, T.; and Uehara, K. 2019. Data augmentation using random image cropping and patching for deep CNNs. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(9): 2917–2931.

Wang, C.; Yang, H.; and Meinel, C. 2018. Image captioning with deep bidirectional LSTMs and multi-task learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(2s): 1–20.

Wei, J.; and Zou, K. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6382–6388. Hong Kong, China: Association for Computational Linguistics.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057. PMLR.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6023–6032.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-Level Convolutional Networks for Text Classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, 649–657. Cambridge, MA, USA: MIT Press.

Zheng, Y.; Li, Y.; and Wang, S. 2019. Intention oriented image captions with guiding objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8395–8404.