Putting Humans in the Image Captioning Loop

Aliki Anagnostopoulou[1,2] **Mareike Hartmann**[1] **Daniel Sonntag**[1,2]
[1]German Research Center for Artificial Intelligence (DFKI), Germany
[2]Applied Artifical Intelligence (AAI), Oldenburg University, Germany
{aliki.anagnostopoulou, mareike.hartmann, daniel.sonntag}@dfki.de

Abstract

Image Captioning (IC) models can highly benefit from human feedback in the training process, especially in cases where data is limited. We present work-in-progress on adapting an IC system to integrate human feedback, with the goal to make it easily adaptable to user-specific data. Our approach builds on a base IC model pre-trained on the MS COCO dataset, which generates captions for unseen images. The user will then be able to offer feedback on the image and the generated/predicted caption, which will be augmented to create additional training instances for the adaptation of the model. The additional instances are integrated into the model using step-wise updates, and a sparse memory replay component is used to avoid catastrophic forgetting. We hope that this approach, while leading to improved results, will also result in customizable IC models.

1 Introduction

Image Captioning (IC) is the task of generating a natural language description for an image (Stefanini et al., 2021). State-of-the-art IC models are trained in the traditional offline setup, where large amounts of annotated training data are required (Zhou et al., 2020; Li et al., 2020; Wang et al., 2022). This requirement is impractical for models intended to caption user-specific images without large-scale annotations. Here, an *interactive* framework can be used to efficiently adapt a model to new data based on user feedback (Ling and Fidler, 2017; Shen et al., 2019). By exploiting user feedback, models can be trained with less annotated data. Furthermore, interactivity renders models more user-friendly, and the interaction with the user often leads to more trust in the AI/ML-based system (Bussone et al., 2015; Guo et al., 2022).

In the following, we present our work-in-progress on extending an IC model to an interactive setup. Our approach is shown in figure 1. We start with a pretrained IC model (subsection 2.1), which is used to caption new images. The user provides feedback for these captions (subsection 2.2), which is then used to generate more training instances via data augmentation (subsection 2.3). These augmented instances are used to update the model incrementally. In order to retain past knowledge, we employ sparse memory replay (subsection 2.4).

In this project, we plan to address four research questions:

1. What type of user feedback is most useful and how can it be collected?
2. Which data augmentation strategies are most useful to maximize the effect of the user feedback on model performance?
3. How helpful is user interaction in the data augmentation process?
4. How can the feedback best be integrated into the training process?

2 Experimental setup

In this section we describe our benchmark strategy, as well as the work-in-progress on data augmentation, model update and evaluation methods, including the human-in-the-loop intersections. The modules described in sections 2.1 and 2.3 are implemented, while the implementation of the ones described in sections 2.2 and 2.4 is ongoing.

2.1 Benchmark strategy

We experiment with a concrete implementation of the interactive approach outlined in Hartmann et al. (2022). As a starting point, we use a PyTorch implementation of the Show, Attend and Tell model (Xu et al., 2015). This architecture consists of a convolutional neural network (CNN) encoder, which is used to extract feature vectors, and a long-short term memory (LSTM) decoder, which generates a caption conditioned on these vectors with attention. The training strategy used is cross-entropy loss.

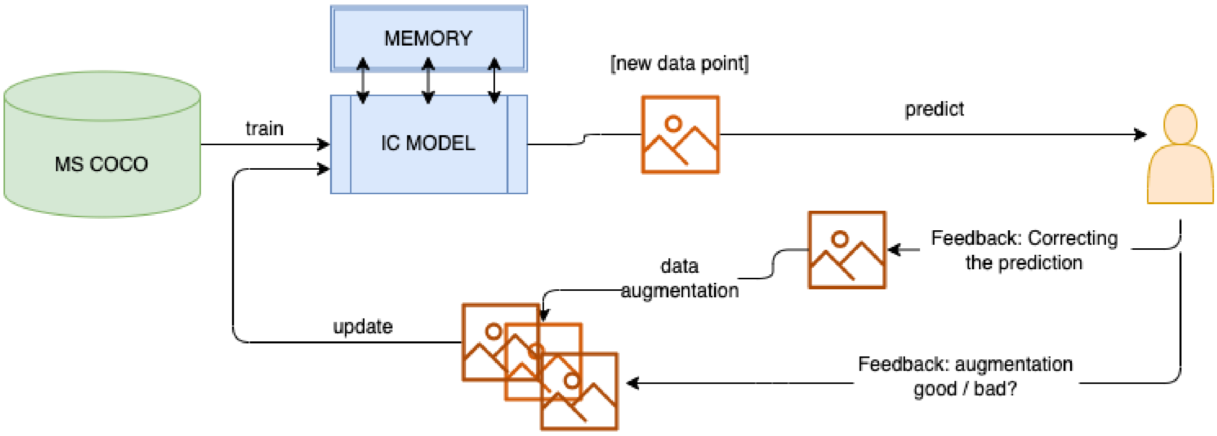

Figure 1: Our proposed pipeline. We pre-train our initial model on MS COCO. The model generates captions for new images, and the user gives feedback on the prediction (by correcting it or marking an area of interest on the image). This feedback is then augmented to create more training instances, and the user can evaluate the quality of the augmentations. The model is then updated accordingly, with a sparse memory replay in order to retain old knowledge.

In addition, we consider an architecture which requires more supervision, namely the Meshed-Memory (M2) Transformer (Cornia et al., 2020). This model is based on the Transformer architecture proposed by Vaswani et al. (2017). In the M2 Transformer, the encoder is extended with "slots" for additional, a priori information (*memory*). Additionally, *meshed* cross-attention is performed in the decoder, not only for the last encoding layer, but for all of them. This model requires the additional input of object detections. The model is trained on cross-entropy loss (for pre-training) and reinforcement learning (for fine-tuning).

To pre-train these base models, we use the MS COCO dataset (Lin et al., 2014). More specifically, we use the 2014 release, which contains 82,783 training and 40,504 validation images, with five captions per image. We make use of the Karpathy splits (Karpathy and Fei-Fei, 2017).

2.2 Feedback collection

Given the prediction of the base model on a new image, we collect useful feedback from users for both modalities. This could refer to the correction of a *caption*, but also the drawing of a bounding box around the object on the *image* that was described incorrectly. Additionally, we plan to experiment with feedback collection for the generated augmentations (see subsection 2.3).

Use case simulation: Ultimately, we want to apply this method in a real-life scenario where any user can provide feedback to adapt the model to their user-specific data. To simulate this feedback in this early stage, we currently work with an already available dataset, namely VizWiz (Gurari et al., 2020; Simons et al., 2020). VizWiz consists of 23,431 training images, 7,750 validation images and 8,000 test images (39,181 images in total). Each image is annotated with five captions. Since captions for the test set are not publicly available, we test on the validation set and use a small part of the training set as our validation set.

2.3 Augmenting the feedback

Data augmentation, or synthetic data generation, is a family of techniques that take an initial dataset (often limited in size) and automatically generate more examples (Atliha and Šešok, 2020). Our plan is to augment user feedback in both modalities. Furthermore, we intend to create a novel joint augmentation method, as described below. Some example augmentations from the VizWiz dataset are shown in figure 2.

Text For the captions, meaning-preserving operations will be used, since our goal is not to introduce noise to the data, but end up with captions that are similar to the feedback provided by the user. We choose three methods:

1. Lexical substitution. We follow the EDA (Wei and Zou, 2019) implementation, which leverages WordNet. 10% of the tokens of each caption are substituted by a synonym. We generate three augmentations with this method.

2. Back-translation. We use the ArgosTranslate[1]

[1]Models are downloaded from: https://www.

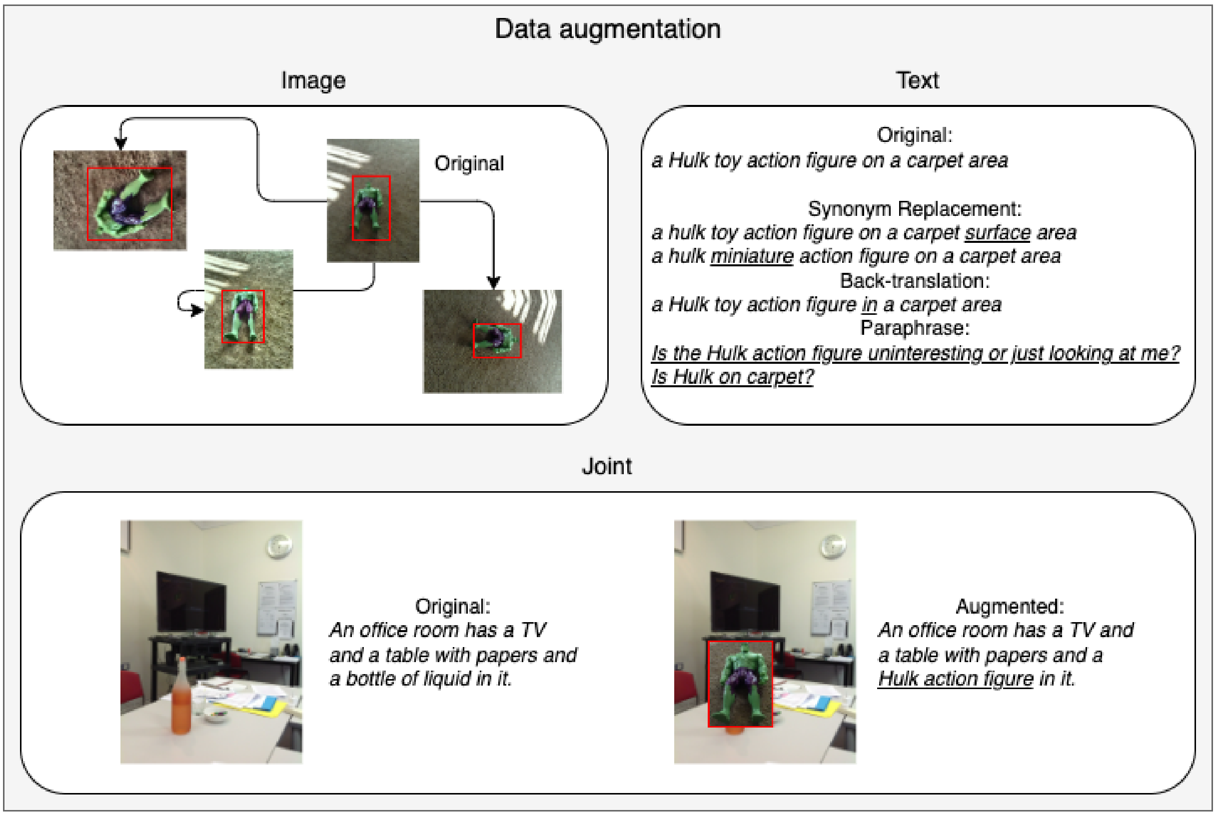

Figure 2: Augmentation examples for image, captions and the joint method. The original data points are from the VizWiz dataset.

library, translating our English captions to Arabic or Spanish and back to English. As a result, two augmented examples are created.

3. Paraphrasing with a T5 model (Raffel et al., 2020), which is specifically fine-tuned for this purpose. With this method, we create five more augmented captions.

Identical augmentations are discarded. The first and third method could provide more augmented examples; we notice, however, that with more examples, the augmentation quality drops significantly, and the number of identical augmentations increases. The whole procedure results in about 10 augmentations per caption.

Image We use the Albumentations (Buslaev et al., 2020) library for image augmentation. We use transformations like rotation, blur, optical distortion, grid distortion, and flip. An additional advantage of the Albumentations library is that it adjusts the bounding boxes with the augmentation. In this manner, feedback provided on the image can be retained.

Joint We also plan to implement a joint augmentation method, which is based on CutMix (Yun et al., 2019) and proposed by Feng et al. (2021). The idea is to cut objects from different images and insert them in other images. Since this would change the content of the image, the caption should be adjusted accordingly - that is, by addition of a description of the inserted object.

Interaction with users The examples resulting from the data augmentation step can be used as additional training examples right away. In addition, we consider user interaction with the augmented examples to assure their quality. More specifically, after user feedback for the prediction is processed and augmented, the user could choose to rank the augmentations (from best to worst), or evaluate them in terms of suitability (*good* / *bad*).

2.4 Model update and evaluation

In a real-life application scenario, the user will input images continuously, which means that the system has to be updated multiple times. In cases where a model is trained repeatedly on new data, *catastrophic forgetting* (Kirkpatrick et al., 2017) can be observed, namely the degradation of model

argosopentech.com/argospm/index/.

performance on older tasks when it is trained on new ones. We plan to tackle this problem with a continual or lifelong learning method, and more specifically with a sparse memory replay during training, adapting the idea of de Masson d'Autume et al. (2019): During training, some samples - experiences - are written into the memory. These past experiences are then sparsely replayed while the model is trained on new data.

In order to simulate the step-wise adaptation of the model to new data, we split the VizWiz dataset in parts of similar size, according to concepts contained in the captions. We follow a naive approach, collecting all noun phrases (NPs) from the captions, and grouping them according to their semantic similarity. We use k-means clustering (Hartigan and Wong, 1979) and pre-trained GloVe word embeddings (Pennington et al., 2014). All images with captions that contain NPs from the same cluster are then allocated to the same split.

We plan to train disjointly, namely treating each data split as a new task, which is one of the methods used both by Nguyen et al. (2019) (sequential class addition) and Del Chiaro et al. (2020) (disjoint procedure). Evaluation is carried out over individual classes or over all classes, each time the model is trained with a new class/task.

User evaluation Apart from evaluating the approach with respect to model performance using automated performance metrics, we plan to evaluate its usefulness and usability for end-users in a human study.

3 Possible extensions

Beyond the proposed architecture, we consider some extensions for our pipeline. As mentioned in subsection 2.4, evaluation from users can potentially point to improvements or the need for the addition of extensions. While VizWiz, which we use as substitute data in the initial implementation and experimentation stages, constitutes a use case by itself, we do not adapt out implementationto this specific dataset (for example, by employing optical character recognition/detection), as we aim to develop an approach that is applicable to a broad range of user-specific data. Implementing such specific adaptations might further improve performance and can be added on top of our approach depending on the use-case.

Further work can focus on the choice of experiences stored in the memory. This can be done either by integrating the user in the sampling process, or by employing active learning techniques to find the most suitable experiences for future replay.

Last but not least, we can leverage the advantage of interactive systems of learning from fewer labeled instances for cases which annotated data is limited. One such case is multilingual IC. For this reason, an extension of our architecture to support multiple languages looks promising.

Acknowledgments

We thank the reviewers for their insightful comments and suggestions. The research was funded by the XAINES project (BMBF, 01IW20005).

References

Viktar Atliha and Dmitrij Šešok. 2020. Text augmentation using bert for image captioning. Applied Sciences, 10(17):5978.

Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. 2020. Albumentations: Fast and flexible image augmentations. Information, 11(2).

Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In 2015 International Conference on Healthcare Informatics, pages 160–169.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-Memory Transformer for Image Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc.

Riccardo Del Chiaro, Bartłomiej Twardowski, Andrew D Bagdanov, and Joost Van de Weijer. 2020. Ratt: Recurrent attention to transient tasks for continual image captioning. arXiv preprint arXiv:2007.06271.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 968–988, Online. Association for Computational Linguistics.

Lijie Guo, Elizabeth M. Daly, Oznur Alkan, Massimiliano Mattetti, Owen Cornec, and Bart Knijnenburg. 2022. Building trust in interactive machine learning via user contributed interpretable rules. In 27th International Conference on Intelligent User Interfaces, IUI '22, page 537–548, New York, NY, USA. Association for Computing Machinery.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. CoRR, abs/2002.08565.

J. A. Hartigan and M. A. Wong. 1979. A k-means clustering algorithm. JSTOR: Applied Statistics, 28(1):100–108.

Mareike Hartmann, Aliki Anagnostopoulou, and Daniel Sonntag. 2022. Interactive machine learning for image captioning.

Andrej Karpathy and Li Fei-Fei. 2017. Deep visual-semantic alignments for generating image descriptions. IEEE Trans. Pattern Anal. Mach. Intell., 39(4):664–676.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114(13):3521–3526.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. ECCV 2020.

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.

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

Giang Nguyen, Tae Joon Jun, Trung Tran, Tolcha Yalew, and Daeyoung Kim. 2019. Contcap: A scalable framework for continual image captioning. arXiv preprint arXiv:1909.08745.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67.

Tingke Shen, Amlan Kar, and Sanja Fidler. 2019. Learning to caption images through a lifetime by asking questions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10393–10402.

Rachel N. Simons, Danna Gurari, and Kenneth R. Fleischmann. 2020. "i hope this is helpful": Understanding crowdworkers' challenges and motivations for an image description task. Proc. ACM Hum.-Comput. Interact., 4(CSCW2).

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2021. From show to tell: A survey on image captioning. CoRR, abs/2107.06912.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. arXiv preprint arXiv:2202.03052.

Jason Wei and Kai Zou. 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), pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, pages 2048–2057, Lille, France. PMLR.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6023–6032.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. Proceedings of the AAAI Conference on Artificial Intelligence, 34(07):13041–13049.