Combining natural language, realistic images, and multi-agent communication: MSc Thesis Proposal

January 11th 2022

By Polina Tsvilodub (ptsvilodub@uos.de), Osnabrück University

The following proposal draft outlines some ideas for my MSc thesis in the area of using natural language for describing realistic images in a multi-agent communication setting. The goal of this proposal is to summarize my ideas regarding the topic to facilitate the initial thesis planning and constraining the scope of the work. However, these are just rough ideas with remaining open questions, and I am completely open to improving these during the kick-off meeting. If possible, I would like to focus on acquiring methodological skills in using different deep learning visual and language modeling modules within this work. This focus has guided the ideas presented below.

The proposal is structured as follows:

(1) presents the ideas regarding the topic, a possible research question and methods in an abstract-style fashion. More specifically, the ideas so far are largely based on **comparing and combining aspects of work from these papers** (ordered by decreasing relevance): [Lazaridou et al., 2020, Andreas and Klein, 2016, Kim et al., 2021].

(2) presents a generic *suggested thesis structure*, which will be adjusted and mapped to specific completion milestones and dates once the topic has been finalized. Nevertheless, I suggest three most important and already plannable dates below. Finally, if possible, I would also like to *schedule monthly review meetings* to discuss the aforementioned milestones, once agreed upon (tbd).

Most important dates:

- Finalize thesis topic, structure and timeline by February 13th 2022
- Finish thesis draft by August 1st 2022
- Submit thesis by August 31st 2022

(3) summarizes two ideas that were not planned through at all yet, but which I would like to nevertheless share with you.

(4) contains references.

(1) Proposal: The area of multi-agent communication has gained increased popularity both as a field for studying the mechanisms of language evolution, as well as as a field for developing potential human-machine communication architectures (e.g., [Lazaridou and Baroni, 2020]). In particular, multi-agent communication experiments employing communication in natural language bear great potential for developing easy to use and scalable human-machine interaction (e.g., [Andreas and Klein, 2016, Mao et al., 2016, Lazaridou et al., 2020, Gupta et al., 2021]).

However, to my knowledge, multi-agent communication experiments employing natural language so far have been conducted on synthetic images with curated labels (e.g., the Abstract Scenes dataset, [Zitnick and Parikh, 2013]). Yet it is critical for potential applications to scale multi-agent communication to natural realistic visual input. Realistic visual input typically refers to natural photographic images, as opposed to synthetic ones, compiled and labelled in datasets like MS

COCO (up to ∼five captions / image), Visual Genome, Google Conceptual Captions (one caption / image) etc [Krishna et al., 2017, Sharma et al., 2018, Lin et al., 2014].

This kind of input has received increased attention in the area of image captioning from the computer vision perspective (e.g., [Kim et al., 2021, Su et al., 2019, Lu et al., 2019] etc, among others). Moreover, some work has focused on training models to produce *discriminative* image captions, which arguably are more cognitively plausible ([Andreas and Klein, 2016, Dai and Lin, 2017, Vedantam et al., 2017, Nie et al., 2020]). That is, they focus on producing captions that would maximally aid a listener in retrieving a target image among distractors, for instance in a *reference game* task. At the same time, this task has been a dominant traning paradigm in multi-agent communication experiments. Approaching the intersection of multi-agent communication and image captioning in natural language, in particular, [Lazaridou et al., 2020] train a model to generate discriminative image captions by training it in a multi-agent communication setting on synthetic images. Other multi-agent communication experiments have focused on other aspects like properties of the emergent communication protocols, the effects of different losses or experimental set-ups (e.g., [Lazaridou et al., 2018, van der Wal et al., 2020]).

The **goal of this thesis would be to extend existing work on multi-agent communication to agents generating pragmatic captions of realistic visual input** like the MS COCO Captions or the Flickr30k dataset [Lin et al., 2014, Young et al., 2014], to move towards the goal of realistic human-machine communication about naturalistic environments, while employing a framework which potentially allows for listener-specific adaptation. Depending on the chosen models (see below), the final dataset choice might depend on available pretrained models and on the required number of labels per image. Transitioning to such a dataset might already pose some issues for the langauge model due to the nature of the labels which might suffer from the human reporting bias [Misra et al., 2016], but this depends on the nature of the chosen dataset, as well as the objective to generate rather human-like discriminative captions (and therefore, possibly even benefitting from such a bias).

Since this goal of natural image captioning has been approached with different methods like vision-language models (e.g., [Kim et al., 2021]), the *modus operandi* of the thesis could be to compare three different methods of generating pragmatic image captions.[1]

The three methods might include:

1. training a speaker model with a pragmatic caption generation objective in a multi-agent communication setting akin to the set-up by [Lazaridou et al., 2020]. While one could just replicate one of their architectures with other training data (e.g., the multi-task learning architecture), one might also try to use a different loss function actually conditioning on both images, e.g., as proposed by [Vedantam et al., 2017] (the so-called introspective speaker), as differences in loss functions might result in relatively large impact on performance. However, specifics of choosing a loss / policy function are definitely subject to further investigation. Other architectures proposed by [Lazaridou et al., 2020] could also be replicated, but this one seemed to be well-motivated to me in the sense that the model is supposed to simultaneously learn to extract important discriminative features from images and to refer to them using language, as opposed to learning the two skills separately.
2. replicating the set-up proposed by [Andreas and Klein, 2016] on the new training dataset. Their architecture seems to be transferable to a different dataset without any major adjust-

[1] "pragmatic" in the sense of [Andreas and Klein, 2016]

ments, but follow-up related work like [Nie et al., 2020] could also be considered.

3. using a vision-language-pretraining based model like the ViLT [Kim et al., 2021] which combines visual and linguistic feature representations early on in the training process and can be fine-tuned to produce image captions or perform caption-based image retrieval. While several such vision-language model variants exit, the ViLT seems to be the most computationally lightweight one.

This comparison would allow to access whether (1) fine-tuning the language model on a listener's downstream task feedback, (2) performing pragmatic reasoning, or (3) early text-vision attention-based feature integration for multimodal datasets produces best pragmatic image captions, e.g., for a referential game. Ideally, this assessment could involve a small human study where participants could rate sample captions with respect to their fluency, as well as play the reference game. Alternatively, a listener agent could be used for evaluation as suggested by [Lazaridou et al., 2020].

Furthermore, such a comparison would allow to compare different state-of-the-art pragmatic caption generation methods with respect to scalability to realistic visual input data, training and inference time / resources, and susceptibility to language drift. Especially the latter aspect has been addressed separately for different model architecures, but no clear comparison between different ways to integrate visual and textual modules with respect to their impact on language drift has been conducted [Lu et al., 2020, Lazaridou et al., 2020, Lee et al., 2019]. Choosing appropriate evaluation criteria for the latter point would be a critical part of finalizing the thesis planning.

(2) Structure: The thesis would consist of roughly the following chapters, assuming the rough topic suggested above:

1. Introduction and motivation: outline of the gap in prior research and motivation for the thesis, framing the topic within the research field
2. Previous work and theoretical background: review of previous work addressing related questions and using related methods, discussed in subchapters per model type. Possibly review of theoretical background for cognitive plausibility aspects and model comparison metrics
3. Experiments: a dataset subchapter, a subchapter for each experiment (split into architecture, training, results)
4. Model comparisons: a subchapter per comparison aspect, discussing methods of evaluation and results
5. Discussion and conclusion of the results

(3) Different ideas: While reference games have been an influential training paradigm for multi-agent communication, intuitively, such a situation contrasting entirely different scenes against each other might seem rather unrealistic. A more realistic set-up might be considering gradual changes within the same scene and detecting meaningful contrasts within a scene. Some work exists in this direction (e.g., [Jhamtani and Berg-Kirkpatrick, 2018]). Maybe this task might be interesting for multi-agent communication.

A different more realistic task might be visual question answering, which has also been addressed within vision-language-pretraining models (e.g., [Lu et al., 2019, Nie et al., 2020]).

Finally, another different idea would be to build a system producing live grammatical natural language captions for live image input (e.g., via webcam) which is segmented, e.g., via an

object detection model. The latter part exists, e.g., based on the YOLOv4 object detector [Bochkovskiy et al., 2020], but I don't know whether the former part has been addressed.

References

[Andreas and Klein, 2016] Andreas, J. and Klein, D. (2016). Reasoning about pragmatics with neural listeners and speakers. *arXiv preprint arXiv:1604.00562*.

[Bochkovskiy et al., 2020] Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

[Dai and Lin, 2017] Dai, B. and Lin, D. (2017). Contrastive learning for image captioning. *arXiv preprint arXiv:1710.02534*.

[Gupta et al., 2021] Gupta, A., Lanctot, M., and Lazaridou, A. (2021). Dynamic population-based meta-learning for multi-agent communication with natural language. *Advances in Neural Information Processing Systems*, 34.

[Jhamtani and Berg-Kirkpatrick, 2018] Jhamtani, H. and Berg-Kirkpatrick, T. (2018). Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584*.

[Kim et al., 2021] Kim, W., Son, B., and Kim, I. (2021). Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*.

[Krishna et al., 2017] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

[Lazaridou and Baroni, 2020] Lazaridou, A. and Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*.

[Lazaridou et al., 2018] Lazaridou, A., Hermann, K. M., Tuyls, K., and Clark, S. (2018). Emergence of linguistic communication from referential games with symbolic and pixel input. *arXiv preprint arXiv:1804.03984*.

[Lazaridou et al., 2020] Lazaridou, A., Potapenko, A., and Tieleman, O. (2020). Multi-agent communication meets natural language: Synergies between functional and structural language learning. *arXiv preprint arXiv:2005.07064*.

[Lee et al., 2019] Lee, J., Cho, K., and Kiela, D. (2019). Countering language drift via visual grounding. *arXiv preprint arXiv:1909.04499*.

[Lin et al., 2014] 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*, pages 740–755. Springer.

[Lu et al., 2019] Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

[Lu et al., 2020] Lu, Y., Singhal, S., Strub, F., Courville, A., and Pietquin, O. (2020). Countering language drift with seeded iterated learning. In *International Conference on Machine Learning*, pages 6437–6447. PMLR.

[Mao et al., 2016] Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., and Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

[Misra et al., 2016] Misra, I., Lawrence Zitnick, C., Mitchell, M., and Girshick, R. (2016). Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2939.

[Nie et al., 2020] Nie, A., Cohn-Gordon, R., and Potts, C. (2020). Pragmatic issue-sensitive image captioning. *arXiv preprint arXiv:2004.14451*.

[Sharma et al., 2018] Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

[Su et al., 2019] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. (2019). Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

[van der Wal et al., 2020] van der Wal, O., de Boer, S., Bruni, E., and Hupkes, D. (2020). The grammar of emergent languages. *arXiv preprint arXiv:2010.02069*.

[Vedantam et al., 2017] Vedantam, R., Bengio, S., Murphy, K., Parikh, D., and Chechik, G. (2017). Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260.

[Young et al., 2014] Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (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.

[Zitnick and Parikh, 2013] Zitnick, C. L. and Parikh, D. (2013). Bringing semantics into focus using visual abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3009–3016.