MSC Thesis Proposal: Combining multi-agent emergent communication, language models and realistic visual input. Tell me what you see.

By Polina Tsvilodub, Osnabrück university. **Supervisors:** Elia Bruni, Michael Franke
pstvilodub@uos.de

Idea: Combine the aspect of dealing with realistic visual input like MS Coco and the aspect of using natural language to deal with it like Lazaridou. Role of emergent communication: task-conditioning of language, with agent feedback instead of human feedback.

Task: Referential game. Might be rather simple. What is a realistic communication task? Maybe image guessing game as more advanced experiment (also number of distractors would be variable here)? Crazy idea: Live object detection captioning, see session 5 advancved DL topics seminar

Open questions: 1) Such input, 2) language drift, 3) Using a generic language model –> how?

Critical papers so far: Lazarisou 2020, Andreas and Klein 2016 for pragmatic visual captioning.

Lazaridou 2020 open questions:

- claim: langauge model pretrained on generic, non task-specific language data –> it is still an image captioning model, no? In what sense is it task-agnostic? –> In the sense that the captions are not reference game specific (i.e., not contrastive)
- what vs how distinction isn't fleshed out until the end of the experiment; in the end, these two are addressed by one combined learned system simultaneously
- "how" – langauge model, "what" – emergent, functional communication
- one distractor (?) –> what happens if there are more? Also distractors in this sense are unrealistic because we don't see many scenes next to each other - we usually have distractor objects within one scene (cf. image guessing game) –> a segmented images dataset would be necessary
- task essentially *discriminative image captioning* –> extend to other tasks
- !!! usage of artificial scenes –> naturalistic multi-modal datasets supposedly suffer from one-sided captions and a human reporting bias –> could task-conditioning improve on that?
- Maybe cast the task into what to say = what is worth mentioning as an explicit representation (cf Misra et al 2016)+ how to say = language model
- Human proof of concept solely addresses task sensibility, not task-conditioning capacity for the language use; a more sensible discriminative test here would have been to see how much the discriminative captions produced by humans would depart from the provided ones
- visual module: pretrained ResNet, language module: one-layer LSTM –> why not a pretrained LM like BART or whatever? (huggingface has these pretrained models)
- baseline oracle speakers sample from ground-truth data, and are not the grounded language models which were not task-fine-tuned (a sensible baseline for me)
- fixed listener: pre-trained on oracle speaker which is not a language model, i.e., it does not provide any signal for language adaptation, only re-ranking based on old captions
- structural only baseline: how is the caption chosen?

- loss functions: there are no pressures to keep the message short / to stop under the maximal length
- !!! reward fine-tuning: extend conditioning of the captioning model to distractor images, too –> how to combine two image representations?
- language drift (=catastrophic forgetting of language) addressed with KL regularization –> compare with approaches from Lu et al 2020; maybe add some utterance log likelihood pressure into loss
- multi-task learning: which λ_s was better? how can λ_f be 1, and the weights of the two components not sum up to 1?
- θ_{ResNet} is not optimized – the model learns to "better describe" static input, not to also attend to features that are more important
- the question between multi-task learning and reward fine-tuning is whether to use a pretrained LM or not, it seems
- especially with proper LMs, look into fine-tuning practices
- alternative: frozen language model, reranking loss on samples from the model –> difference to reward fine-tuning: no update of LM, sampling captions, not sequential words (effect of this is not discussed btw)
- –> essentially, the idea of learning a task-conditional language model would be to learn to use language in novel, contrastive / discriminative ways, by attending to discriminative features and applying (grounded) language to them correctly –> requires learning the discriminative visual features. –> discriminative learning = contrastive / max margin loss for the visual component; is there such a thing for language models?
- –> Huang et al 2018, Gunel et al 2021
- PoE: unclear how disctractor and target vectors are combined; reranker trainable as BoW embedding of samples, effectively adding a trainable language layer on top of the LSTM –> what would have been ranking alternatives? ordinal regression over the samples?
- !!! the similarity refers to listeral listener level RSA –> extend to pragmatic listener + pragmatic speaker level RSA !!! Furthermore, the learned semantcs are very similar to Xenia's work
- test time: what are sensible testing scenarios? learned listener as convergence evaluation; optimally, humans
- in human evaluation, they deliberately ignore / counteract humans' speaker-specific adaption –> it would have been interesting in order to see if the models develop consistent communication strategies that humans could understand –> consistency metrics
- ground truth captions speaker: access to labels; either picks one at random or uses world overlap heuristic for discriminative captioning
- results upshot: changing the language model decreases performance
- referential success of joint liseners: unclear, how co-adaptation of the two agents is being counteracted against

- why wasn't the noisy channel approach applied to the multi-task set up? (Part of) The losses are similar in that they use the conditional probabilities of the caption given the target
- PoE loss: why is the loss a proper distribution? (nachrechnen)
- they interpret their results as indicating the effectiveness of extending task conditioning of image captioning model to distractor image, i.e., doing discriminative captioning, but there is no indication whatsoever that the task conditioning can be used for fine-tuning a generic language model (becuase best results achieved on reranker models)
- (is there a reranker version withough ground truth captions?) Yes there is, the standard model samples from the pretrained language model which produced 20 captions.
- LM pretraining on image cpationing task objective = standard supervised learning of (caption | target image) using labels of dataset
- has anyone trained (caption | target, distractor) at all? Adreas and Klein? –> YES
- fixed grounded listener predictive of human performance (=pretrained with oracle discriminative speaker). This is quite an unsurprising result overall given that reranking the ground truth labels has best performance with humans overall.
- structural language drift: $logP(m)$ under pretrained unconditional language model –> which one? still the same captioning model? –> why not a grammar like PCFG, i.e., relation of neural models to syntactic proficiency rather unclear?
- semantic drift: $logP(m|i)$
- drift measures don't consider the functional component which is advantageous w r t adaptability of the measure, but looses information about the contrastivity which kind of overrides ungrammaticality (at least intuitively)
- pragmatic drift: divergenece between human interpretation and interpretation the speaker assumes; task specific; operationalized as difference between human and agent referential success –> the case where human performance is higher than agent performance is not discussed; includes co-adaptation and convention forming
- Lazaridou doesn't have any fluency measure for the generated captions(as Andreas&Klein 2016)

Given the overall goal to create a system that takes as input realistic images and generates discriminative natural language captions for them (ideally, by fine-tuning a general-purpose LM), the following major open questions remain:

1. Can we condition a language model on images, and specifically on a target + distractor?
    - Andreas & Klein 2016: L0 module $P(image\ ranking|target, distractor, message)$, So module $P(message|image)$ (equivalent to pretrained captioning model from Lazaridou)
    - L0 trained contrastively to maximize P to choose referent given a random sampled disctractor and description!!! –> add to Lazaridou's approach at listener level (it is unclear though if the loss is propagated to the description encoder)
    - S0 maximizes conditional likelihood of training data captions for a given embedded referent

- they pressure the speaker model ("S1" derived from L0 and S0) to produce contrastive descriptions indirectly, by using reasoning over L0's choises (there is no direct conditioning on two images)
- this is done by sampling candidate descriptions from S0 –> regularization towards fluent behavior –> point for counteracting potential language drift
- was the whole thing trained end-to-end? No, because the learning signal is from human raters (accuracy and fluency).
- (8) for scoring the samples is conceptually very similar to the PoE model of Lazaridou, except for computing sample caption score via considering a listener as in the noisy channel model version
- An alternative might be to pressure the model directly via an appropriate loss, or to learn to attend to contrastive image features and describe them with a "standard" language model. For instance, use visual attention and emphasize difference between attention scores.
- –> "Compiled" speaker from Andreas and Klein conditioning on concatenated image embeddings. It uses contrastive training data -> this could be produced by choosing most contrastive caption from non-contrastive training data via word overlap heuristic or similar as in Lazaridou
- Mao et al 2015

2. Can we finetune a language model for image captioning?
    - I would actually like to do something like the multi-task training situation from Lazaridou where the LM parameters are also trained (=finetuned)
    - Huggingface transformers library finetuning guide for pretrained LMs
    - Training a GRU on MS COCO: `https://www.tensorflow.org/tutorials/text/image_captioning`
    - VL-BERT: takes both visual and linguistic features as input; this would solve the issue of recasting the captioning model into a task conditional model
    - ViLBERT: combines single-mode networks combined via transformer module
    - ViLT: seems to be a simplified approach to generating such multimodal features
    - Zhou et al 2020 (AAAI): VLP model actually fine-tuned for image caption generation
3. Is there an appropriate image dataset and prior work?
    - **Major novelty** of my thesis is going to be using a natural image dataset like the MS Coco captions or Google's Conceptual Captions (only one caption per image though!)
    - previous work on both datasets hasn't been situated within the emergent communication framework (afaik). Although see the work on human reporting bias issues with captions (Misra et al 2016).
    - trained visual classifier with noisy human labels –> decouple into presense and relevance prediction modules;

- There is also the VizWiz dataset and challenge
- Visual Genome
- prior work from (https://arxiv.org/pdf/1701.02870.pdf):

4. Optional*: Is there a possibility to integrate a new loss (like contrastive loss) to (a) increase performance and (b) counteract language drift?
    - `https://arxiv.org/pdf/1701.02870.pdf` Context-aware Captions from Context-agnostic Supervision (2017): CUB 200-2011 dataset; claim: aviods training an L0 model (in contrast to Andreas & Klein)
    - introspective speaker model employs a listener function modelling discriminability via likelihood ration between speakers generative distributions conditioned on target over one conditioned on distractor –> might be a better adjustment of Lazaridou's multi-task objective combining language regularization and functional requirements –> would emphasize loss-function dependence
    - urlhttps://arxiv.org/pdf/1710.02534.pdf (Dai et al 2018): yet another mu;ti-component loss function. Details unclear at this moment –> important point is that there are different losses integrating the functional requirement of producing a caption conditioned on several images is fomralized in different ways
    - policy gradient methods cited in paper above
5. Optional*: In case we'd go for the crazy live caption idea, one might need to use live object detection which produces keywords, and construct descriptions from keywords.

Braindump area:

- the assumption is that i would pressure the model to attend to discriminative features and appropriately label them via applying the learning signal from the listener agent in the reference game task; appropriate loss will be critical here
- it is important to regularize the model to use fluent language and to not find loopholes
- problem with using ViLT or similar is the construction of the listener... we can assume the same text embedding module, but how do we retrieve the target image? One could use the same dot-product similarity as Lazaridou, or one could introduce some fancy attention, or one could try to do text2image. In case of dot-product, it is kind of ad hoc to use a different mechanism than used in the spaker module.
- one idea could be to compare three kinds of models on a captioning model for a natural image dataset: (1) a model like the one by Andreas&Klein 2016, but cast into a multi-agent setting (essentially adjusting Lazaridou's noisy-channel model's loss) (2) Lazaridou's multi-task or reranker model and (3) a multi-modal model like ViLT which embraces a better, or maybe even more explicit alignment between visual and textual features. This comparison would give as an intuition if pragmatic reasoning, reranking, or better feature alignment (vision-language grounding, one might say) are more beneficial for producing human-like captions. ViLT seems the most promising model due to computational efficiency here. the R@1 zero-shot image retrieval results seem to do exactly the type of reference game task we are considering here, but (?) even among all the candidates?

- This comparison aims at evaluating the best way to achieve the initial goal of best possible pragmatic scene description (pragmatic in the sense of being discriminative/contrastive). In a sense, a subgoal of this is turning the language model into a task-conditional language model which isn't really the true of the base ViLT. In the fine-tuned ViLT, the fine-tuning happens via (contrastive) self-training with cross-entropy, which is rather similar to a learning signal coming from a hypothetical listener. `https://github.com/dandelin/ViLT`
- it would be ideal to conduct a little human evaluation w.r.t accuracy and fluency on the model(s) outputs in the end (if they are sensible at all)
- is there a way of casting ViLT into a task-conditional model in Lazaridou's 2020 sense?
- diagnoze the different models with respect to different aspects like accuracy, fluency, inference / training time, how well each model performs in
- completely different approach: look into iange difference description models (Jhamtani, H., & Berg-Kirkpatrick, T. (2018). Learning to describe differences between pairs of similar images. arXiv preprint arXiv:1808.10584. or similar)