EMERGENT COMMUNICATION FINE-TUNING (EC-FT) FOR PRE-TRAINED LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

It has recently been argued that the currently dominant paradigm in NLP of pretraining on text-only corpora will not yield robust natural language understanding systems. One strain of this argumentation highlights the need for grounded, goal-oriented, and interactive language learning. In this position paper, we articulate how **E**mergent **C**ommunication (EC) can be used in conjunction with large pretrained language models as a '**F**ine-**T**uning' (FT) step (hence, EC-FT) in order to provide them with supervision from such learning scenarios. We discuss methodological issues and difficulties with making this work, and then illustrate the overall idea with a case study in unsupervised machine translation, before concluding with a discussion on the relation to multimodal pretraining.

1 EC-FT: MOTIVATION

Recent breakthroughs in myriad NLP tasks and datasets have been driven by large-scale pretraining from raw text corpora using semi-supervised learning signals (Howard & Ruder, 2018; Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019; Conneau et al., 2020; Liu et al., 2020; Brown et al., 2020, i.a.).Despite these breakthroughs, several recent results suggest limitations on their generalization performance (McCoy et al., 2019; Niven & Kao, 2019; Ettinger, 2020; Rogers et al., 2020, i.a.). More fundamentally, a number of recent papers have argued that pretraining on text alone will not deliver fully general and robust NLP systems. For example, through several detailed thought experiments, Bender & Koller (2020) argue that models trained on text alone, in principle, will not be able to recover either the conventional/standing meaning of expressions or the communicative intent of an expression in context. Many of their arguments highlight the importance of the interaction between linguistic expressions and extra-linguistic context communicative intents (e.g. acting in the world, executing programs, etc.).[1] Similarly, Bisk et al. (2020) articulate progressively broader *world scopes* in which language use is embedded, and argue that present pretraining methods work at a relatively limited scope. They too emphasize the importance of embodied interaction with the environment and with the social world for future NLP systems.

At the same time, the growing field of emergent communication (Lazaridou & Baroni) studies artificial agents communicating with each other in order to accomplish particular goals in an environment. This can be viewed as a subfield of reinforcement learning, wherein language (i.e. the communcation protocol) is shaped by rewards determined by interacting with an external environment and with other agents. Typical work in this area starts, however, from a *tabula rasa* and studies under what conditions—e.g. environments, tasks/goals, social settings—the resulting communication protocols among agents resembles human language, along axes like word length economy (Chaabouni et al., 2019a), word-order biases (Chaabouni et al., 2019b), and compositionality (Chaabouni et al., 2020; Steinert-Threlkeld, 2020; Geffen Lan et al., 2020), among others (Mu & Goodman, 2021).

In this short position paper, we articulate a vision for using the goal-directed, interactive setting of emergent communication as a method for fine-tuning pretrained language models. This EC-FT approach has the potential to combine the strengths of large-scale pretraining with the benefits of grounded, goal-directed communication to build better NLP systems. We first (§ 2) articulate some implementation choices for executing an EC-FT project before illustrating them with a case study on unsupervised machine translation (§ 3) and finally discussing related (§ 4) and future (§ 5) work.

[1] See Merrill et al. (2021) for a formalization of Bender & Koller (2020)'s argument about learning a programming language from form alone.

2 IMPLEMENTATION CHOICE POINTS

Executing EC-FT is non-trivial: many issues arise when integrating text-only models into a larger, interactive scenario. We here outline some of the choice points, focusing on pieces that are unique to embedding a pretrained model in an EC setting as opposed to EC in general.

Type of pretrained model: while this choice will partially be determined by downstream task desires, we note that a primary choice for transformers comes from whether to use an encoder-only (e.g. BERT/RoBERTa), decoder-only (e.g. the GPT series) or an encoder-decoder model (e.g. (m)BART (Lewis et al., 2020; Liu et al., 2020) and MASS (Song et al., 2019)). Given that EC effectively 'inverts' sequence-to-sequence tasks, with a sender (a decoder) generating a message that's consumed by a receiver (an encoder), it is natural to use pretrained models with both components.[2]

Environment-to-sender adapter: Depending on the choice of EC game, some extra-linguistic environment will need to be presented to a sender pretrained as a text-only decoder. For example , in a traditional image reference game (Lazaridou et al., 2017; Havrylov & Titov; Lee et al., 2018, i.a.), this will involve presenting one or several images to the sender. Two special considerations arise in the EC-FT setting. First, images are often represented by extraction from a pretrained vision model, and then applying a small learned transformation to the sender's embedding space. In the EC-FT setting, we are primarily interested in training the language-modeling components and exporting them to other language-only tasks. This provides some *prima facie* reason to prefer fixed, non-learnable transformations of image representations into the sender's embedding space. Candidates include various strided pooling operations or other dimensionality reduction techniques like PCA.

Second, an issue arises when the sender is initialized from a pretrained transformer decoder (as will be common in the current era), which heavily relies on cross-attention to a set of representations (typically from an encoder). When using recurrent networks, a fixed image representation can be "presented" to the sender by supplying it as the initial hidden or cell state. Here, however, the environment needs to be represented not just as one representation, but as a sequence for the purposes of cross-attention.[3] In the image reference game scenario, this can be accomplished with a small LSTM that takes an image representation and generates a sequence, or via more sophisticated region-of-interest extraction methods as used in vision transformers (Li et al., 2019b; Tan & Bansal, 2019; Du et al., 2022). Intuitively, inputs to the cross-attention of the sender/decoder should occupy a similar space as the encodings that the model was pre-trained with, so as to mitigate catastrophic forgetting. For this reason, a learned transformation may be more appropriate than a fixed one.

Receiver-to-environment adapter: On the other end of the pipeline, a transformer encoder receiver takes the sender's natural language generation as input and produces a set of representations rather than a single one. Thus, a pooling or aggregation step is necessary. This could involve using a special position token (e.g. `[CLS]`) as the sentence representation, mean/max pooling, or learning a light aggregator over the final representations (e.g. a small recurrent network).

Drift / forgetting: it has been observed in other contexts that EC can cause significant language drift (Lee et al., 2018; 2019; Lu et al., 2020; Lazaridou et al., 2020). This is likely to be particularly acute in the EC-FT setting, where the pretraining tasks are very different from typical EC tasks. Methods such as KL regularization (Havrylov & Titov; Baziotis et al., 2020) are likely to be helpful here, as are other techniques from the catastrophic forgetting literature (Parisi et al., 2019).

3 CASE STUDY: UNSUPERVISED TRANSLATION

We here describe an ongoing project which aims to build better systems for *unsupervised* neural machine translation (UNMT) by taking large pretrained multilingual models and embedding them in an EC game, i.e. having them partake in goal-directed communication. At a high level, communication should promote translation in the following way. The task of translation can be viewed as 'aligning'

[2]This is, of course, not necessary. For instance, XLM(-R) (Conneau et al., 2020) is an encoder-only model that was fine-tuned for sequence-to-sequence tasks by initializing a decoder with the same weights.

[3]There are at least two reasons that an environment (e.g. image) embedding cannot simply be the first token to the sender. (1) Many pretrained models use the first token as a special control code (e.g. a language ID in multingual models (Liu et al., 2020)). (2) It is not obvious that a transformer decoder's cross-attention can be 'turned off' without effecting overall performance, since the model has always been trained with it.

a model's representations for sentences in several languages. In the supervised case, the parallel text data instructs the model how to do this alignment. In the unsupervised case, there's no such direct supervision. But through communication, a model can align its representations for different languages *with the same shared environment*, which will thereby promote alignment between the languages themselves. A schematic overview of this paradigm can be seen in Figure 1.

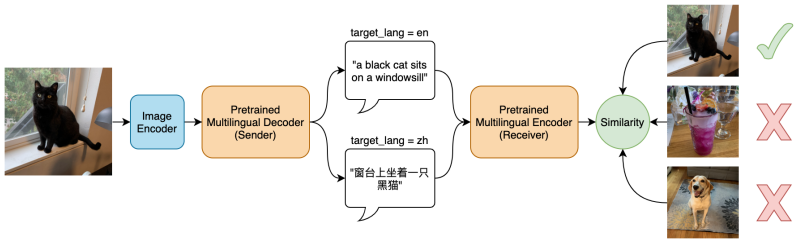

Figure 1: Learning to translate by learning to communicate: a standard image reference game, but with the sender and receiver initialized from a pretrained multilingual decoder and encoder. The communication language alternates between several of those on which the model was pre-trained.

We have operationalized some of the above choice points as follows. (Appendix A has more details.) We use mBART-25 (Liu et al., 2020), a multilingual seq2seq model that has shown strong results on unsupervised translation via the use of backtranslation (Sennrich et al., 2016; Lample et al., 2018), in a standard image reference game. For the environment-to-sender adapter, we extract the final hidden layer of a pretrained ResNet, use a single linear layer to transform its dimensionality to that of mBART,[4] and finally use a small LSTM to generate a bag of representations for cross-attention. This adapter is trained via a small amount of (monolingual) supervised image captioning before EC. For the receiver-to-image adapter, we take the receiver's representation of a CLS token prepended to its input sequence; we also train this token via light supervision on choosing the correct image from a gold caption at the same time as the sender's adapter. Our main loss is the same as Lee et al. (2018); Li et al. (2020), where logits for image choice are inversely proportional to the MSE between image representations and the receiver's encoding of the sender's text. Pilot results exhibited significant drift, with EC causing significant repetition of content words (as observed by Lazaridou et al. (2020)); adding KL regularization (Havrylov & Titov; Baziotis et al., 2020) with an mBART decoder lightly finetuned for multilingual causal language modeling mitigated the problem.

Table 1 reports promising pilot results (see Appendix A for details): adding EC to a baseline model trained with backtranslation produces a substantial increase on average BLEU in zh↔en unsupervised translation (with largest effect on the en→zh direction). This suggests that the EC-FT paradigm can produce results on a core NLP task: unsupervised translation. We are in the process of extending these proof-of-concept results, and will write a full paper detailing them in the near future.

Model	Mean (en ↔ zh)	en→zh	zh→en
mBART + BT	10.02	12.45	**7.58**
mBART + BT + EC	**11.01**	**14.75**	7.27

Table 1: BLEU scores on WMT '19 (Barrault et al., 2019).

4 Related Work

EC and NLP Several other works have investigated the use of emergent communication for NLP tasks (and two for translation in particular). These provide much inspiration for the EC-FT approach, but do not leverage large-scale pretrained language models in the EC context.

Lee et al. (2018) use EC together with image captioning to build unsupervised translation models, showing that EC promotes better translation than the multimodal alignment technique of Nakayama & Nishida (2017) (though the unsupervised setting still lags behind supervised translation). The EC-FT approach to UNMT above differs in several important respects: we initialize our EC game with

[4]We share these parameters to present the images to the receiver as well.

pretrained language models; we integrate EC with backtranslation; and we do not simultaneously train on the EC objective and image captioning objective. Moreover, because we use one multilingual model, our caption grounding only uses one language, instead of all languages. Our pilot EC-FT results suggest that the advantages of using EC for UNMT found by Lee et al. (2018) can be had even in the context of large-scale pretrained multilingual models together with backtranslation.

Li et al. (2020) use emergent communication as a pretraining step for NMT systems. That is: they have agents play an EC game, and then use those parameters to initialize an NMT system. They find that (together with adapters and weight-distance regularization) EC pretraining improves in BLEU over a standard NMT baseline, with especially large gains coming in the few-shot (500 or 1000 examples) setting. While this work shows that EC can provide a good initialization for recurrent NMT system, our pilot results suggest that EC can provide a good fine-tuning signal for a transformer-based pretrained multilingual language model in the unsupervised NMT setting.

Our approach further differs from these two in using one multilingual model, controlled by language ID codes, instead of separate models for each language.

Finally, Lazaridou et al. (2020) study various ways of combining EC with a standard NLP task, namely image captioning. As discussed in Section 2, they identify several forms of language drift and explore various ways of combining the two forms of loss. This work heavily inspired our own, since many of their settings correspond to using a pretrained image-caption system. Our focus, however, has been on using EC to fine-tune large-scale pretrained text-only models, which introduces its own challenges and has its own benefits.

Multimodal Pretraining Recently, efforts in multimodal pretraining are surging, especially in vision-language (V-L) pretraining (Du et al., 2022). Most of the works create joint V-L representation through fusion encoder (Li et al., 2019b;a; Tan & Bansal, 2019), where the fused representation is the joint representation of image and text with one encoder. Other recent works also attempt using different encoders for images and texts to make the framework more efficient, such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021). While this previous V-L pretraining work models image and text data jointly (Du et al., 2022; Wang et al., 2021), we first use the existing pretrained language model and further pretrain it through the communication process in the referential game. Although we expect the alignment between image and text to arise through this process, we view the visual modality as an additional signal to ground the multilingual communication process.

We also note that most previous works on V-L pretraining evaluate solely on vision or V-L tasks (Li et al., 2019b; Radford et al., 2021; Jia et al., 2021), while the advantage of this joint pretraining for language-only tasks remains unclear (Yun et al., 2021; Pezzelle et al., 2021). We therefore focus on language tasks, specifically machine translation, as a way to examine the advantage of emergent communication for using the visual modality as an additional signal after pre-training.

Finally, we note that EC-FT is in a sense strictly more general than typical approaches to multimodal pre-training. While the image-based referential game used in the case study works by promoting multimodal alignment, the range of possible communication tasks that can be used in EC-FT is huge, from directing other agents in an environment (Mordatch & Abbeel, 2018) to controlling a robot (Das et al., 2019) to playing various types of games and reasoning about social dilemmas (Jaques et al., 2019). This wide range of tasks can incorporate many dimensions of communication that should be beneficial for NLP systems—e.g. other agents with their own private goals, social context, embodied control—that are not easily captured by multimodal pretraining (Bender & Koller, 2020; Bisk et al., 2020). In terms of Bisk et al. (2020)'s *world scopes* mentioned in the introduction, multimodal pretraining corresponds to world scope 3 (perception); EC-FT has the ability to move us much closer towards the final scopes 4 (embodiment and action) and 5 (the social world).

5 Conclusion

We have articulated the EC-FT approach of using emergent communication to finetune pretrained language models for use in text-only NLP tasks and illustrated this approach via a proof-of-concept on unsupervised machine translation. This general approach has the possibility of leveraging the benefits of large-scale pretraining while overcoming some of that approach's limitations by embedding pretrained models in goal-directed, interactive communicative scenarios. Much future work remains to be done: we hope that others will be encouraged pursue it.

ACKNOWLEDGMENTS

We gratefully acknowledge [removed for anonymity] for helpful discussion. This work was partially supported by funding from [removed for anonymity].

REFERENCES

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 1–61, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5301. URL https://aclanthology.org/W19-5301.

Christos Baziotis, Barry Haddow, and Alexandra Birch. Language model prior for low-resource neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7622–7634, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.615. URL https://aclanthology.org/2020.emnlp-main.615.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL https://aclanthology.org/2020.acl-main.463.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8718–8735, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.703. URL https://aclanthology.org/2020.emnlp-main.703.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. Anti-efficient encoding in emergent communication. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019a.

Rahma Chaabouni, Eugene Kharitonov, Alessandro Lazaric, Emmanuel Dupoux, and Marco Baroni. Word-order Biases in Deep-agent Emergent Communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5166–5175, Stroudsburg, PA, USA, 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1509. URL https://www.aclweb.org/anthology/P19-1509.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. Compositionality and Generalization In Emergent Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4427–4442, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.407. URL http://arxiv.org/abs/2004.09124https://www.aclweb.org/anthology/2020.acl-main.407.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. TarMAC: Targeted multi-agent communication. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1538–1546. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/das19a.html.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of Vision-Language Pre-Trained models. February 2022.

Allyson Ettinger. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48, 2020. doi: 10.1162/tacl_a_00298. URL https://aclanthology.org/2020.tacl-1.3.

Nur Geffen Lan, Emmanuel Chemla, and Shane Steinert-Threlkeld. On the Spontaneous Emergence of Discrete and Compositional Signals. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4794–4800, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.433. URL https://www.aclweb.org/anthology/2020.acl-main.433.

Serhii Havrylov and Ivan Titov. Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*. URL http://arxiv.org/abs/1705.11192.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1031. URL https://aclanthology.org/P18-1031.

Eric Jang, Shixiang Shane Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *ArXiv*, abs/1611.01144, 2017.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, Dj Strouse, Joel Z. Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3040–3049. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/jaques19a.html.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and Vision-Language representation learning with noisy text supervision. February 2021.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5039–5049, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1549. URL https://aclanthology.org/D18-1549.

Angeliki Lazaridou and Marco Baroni. Emergent Multi-Agent Communication in the Deep Learning Era. URL http://arxiv.org/abs/2006.02419.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-Agent Cooperation and the Emergence of (Natural) Language. In *International Conference of Learning Representations (ICLR2017)*, 2017. URL http://arxiv.org/abs/1612.07182.

Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7663–7674, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.685. URL https://aclanthology.org/2020.acl-main.685.

Jason Lee, Kyunghyun Cho, Jason Weston, and Douwe Kiela. Emergent translation in multi-agent communication. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1vEXaxA-.

Jason Lee, Kyunghyun Cho, and Douwe Kiela. Countering language drift via visual grounding. 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)*, pp. 4385–4395, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1447. URL https://aclanthology.org/D19-1447.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. August 2019a.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. August 2019b.

Yaoyiran Li, Edoardo Maria Ponti, Ivan Vulić, and Anna Korhonen. Emergent communication pre-training for few-shot machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4716–4731, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.416. URL https://aclanthology.org/2020.coling-main.416.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL http://arxiv.org/abs/1405.0312.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020. doi: 10.1162/tacl_a_00343. URL https://aclanthology.org/2020.tacl-1.47.

Yuchen Lu, Soumye Singhal, Florian Strub, Aaron Courville, and Olivier Pietquin. Countering language drift with seeded iterated learning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6437–6447. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/lu20c.html.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *ArXiv*, abs/1611.00712, 2017.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060, 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00412.

Igor Mordatch and P. Abbeel. Emergence of grounded compositional language in multi-agent populations. In *AAAI*, 2018.

Jesse Mu and Noah D. Goodman. Emergent communication of generalizations. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2021. URL https://arxiv.org/abs/2106.02668.

Hideki Nakayama and Noriki Nishida. Zero-resource machine translation by multimodal encoder–decoder network with multimedia pivot. *Machine Translation*, 31(1-2):49–64, jun 2017. ISSN 0922-6567. doi: 10.1007/s10590-017-9197-z. URL http://link.springer.com/10.1007/s10590-017-9197-z.

Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4658–4664, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1459. URL https://aclanthology.org/P19-1459.

German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019. ISSN 0893-6080. doi: 10.1016/j.neunet.2019.01.012.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

Sandro Pezzelle, Ece Takmaz, and Raquel Fernández. Word Representation Learning in Multimodal Pre-Trained Transformers: An Intrinsic Evaluation. *Transactions of the Association for Computational Linguistics*, 9:1563–1579, 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00443.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349. URL https://aclanthology.org/2020.tacl-1.54.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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)*, pp. 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: Masked sequence to sequence pre-training for language generation. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5926–5936. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/song19d.html.

Shane Steinert-Threlkeld. Toward the Emergence of Nontrivial Compositionality. *Philosophy of Science*, 87(5):897–909, 2020. ISSN 0031-8248. doi: 10.1086/710628.

Hao Tan and Mohit Bansal. LXMERT: Learning Cross-Modality encoder representations from transformers. August 2019.

Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. UFO: A UniFied TransfOrmer for Vision-Language representation learning. November 2021.

Tian Yun, Chen Sun, and Ellie Pavlick. Does vision-and-language pretraining improve lexical grounding? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4357–4366, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.370. URL https://aclanthology.org/2021.findings-emnlp.370.

A Training Details for Pilot Study

We here include more details about the training protocol for the pilot results reported in Section 3. As mentioned there, a more detailed and exhaustive set of experiments will appear in a future paper.

Pretrained model We start from mBART 25 (Liu et al., 2020).

BT data Following the mBART paper, we use the zh/en portions of the Common Crawl 25 dataset, which is the same data used for its pretraining, to generate synthetic data in backtranslation.

Evaluation data We evaluate using BLEU on the first 4096 examples of the WMT'19 news translation shared task data (Barrault et al., 2019). We generate using beam search with a width of 5.

mBART + BT training Our baseline UNMT model is trained using on-the-fly iterative backtranslation. We do 1536 steps for each direction, using a batch size of 32 and a maximum generated sequence length of 64. We use a learning rate of 1×10^{-5} with a linear warmup of 192 steps. Inspired by Liu et al. (2020), we constrain generation: for the first 384 steps, the model can only generate tokens in the desired language among the 90% most frequent; afterwards, this is lifted to the 99% most frequent tokens in the language. (This helps prevent the model from copying the source text.) The best model is chosen by cross entropy loss on the WMT '19 validation set, evaluated every 64 steps.

mBART + BT + EC training We aim to show that EC training can improve UNMT performance in the setting of using a large pretrained model together with BT. To that end, we use the same amount of BT, but additionally train on an image reference communication game in the middle. In particular, we have the following pipeline:

1. 512 steps of backtranslation: following the protocol as above, but with no linear warmup of the learning rate.
2. Supervised image captioning. In order to provide initial training signal to the various small adapters, we do a small amount of supervised training.

   We train the sender (mBART decoder + image adapters) on 10,000 English image captions from MS-COCO (Lin et al., 2014). At the same time, we train the receiver to choose the

original image from $k = 15$ distractors on the basis of the gold caption, prepended with a special CLS token. The sender uses standard cross-entropy loss; the receiver uses the same loss as in (Lee et al., 2018; Li et al., 2020).

We do both of these for 2048 steps, using the same learning rate schedule as in BT, with batch size of 16.

Following Lee et al. (2018); Li et al. (2020), image features are extracted from ResNet 50 He et al. (2016). These 2048 feature-vectors are mapped to the 1024-dimensional mBART hidden size using a single linear layer, which is shared between sender and receiver.

On the sender side, we take this 1024-dimensional output, and unroll it into a sequence of length 32 using a single-layer LSTM. The sequence of representations generated by this LSTM is used for the cross-attention in the sender in place of the encoder representations it saw during pretraining.

3. EC training. We do 1024 steps of EC training on the image reference task. The images also come from MS-COCO, but are ensured not to overlap with the ones used in the previous supervision step. We used the same number of distractors and image selection loss as described above, using the straight-through Gumbel-Softmax trick to backpropagate through the sender's generation (Jang et al., 2017; Maddison et al., 2017).

   In addition to the image-selection loss, we add KL regularization (Havrylov & Titov; Baziotis et al., 2020) with a language model, using $lambda = 0.125$ as a weight for this term. In particular, we fine-tuned mBART's decoder only, with cross-attention entirely turned off, as a multilingual causal language model on 500k examples from the CC25 corpus.

4. Final BT. We finish with a remaining 1024 steps of BT. This has a constant learning rate, and the 99% mask.

Thus, steps (1) and (4) together are the same BT training protocol as our baseline; but we have added steps (2) and (3).