META LEARNING FOR MULTI-AGENT COMMUNICATION

Abhinav Gupta
MILA, DeepMind
abhinavg@nyu.edu

Angeliki Lazaridou
DeepMind

Marc Lanctot
DeepMind

ABSTRACT

Recent works have shown remarkable progress in training artificial agents to understand natural language but are focused on using large amounts of raw data involving huge compute requirements. An interesting hypothesis follows the idea of training artificial agents via multi-agent communication while using small amounts of task-specific human data to ground the emergent language into natural language. This allows agents to communicate with humans without needing enormous expensive human demonstrations. Evolutionary studies have showed that simpler and easily adaptable languages arise as a result of communicating with a diverse group of large population. We propose to model this supposition with artificial agents and propose an adaptive population-based meta-reinforcement learning approach that builds such a population in an iterative manner. We show empirical results on referential games involving natural language where our agents outperform all baselines on both the task performance and language score including human evaluation. We demonstrate that our method induces constructive diversity into a growing population of agents that is beneficial in training the meta-agent.

1 INTRODUCTION

Language is arguably the quintessential property of human intelligence. It allows us to communicate with others in order to coordinate tasks. Recent advances in deep learning has seen tremendous progress in training artificial learning agents that can perform tasks while coordinating with humans (Carroll et al., 2019; Hu et al., 2020). Multi-agent learning models this setup by incorporating other agents' behavior into an agent's own decision making. We are specifically interested in these systems with an added verbal communication channel that gives the agents a means to circumvent partial observability. Many recent works (Sukhbaatar et al., 2016; Lazaridou et al., 2017; Foerster et al., 2016; Kottur et al., 2017; Lazaridou & Baroni, 2020) have focused on different tasks and the languages that get evolved as a result (commonly referred to as 'emergent languages') studying its properties including compositionality Choi et al. (2018); Resnick et al. (2020); Chaabouni et al. (2020); Slowik et al. (2020) and correlation with a real human language (Lowe et al., 2020; Lazaridou et al., 2020; Lu et al., 2020). All these methods devise algorithms that use a single set of agents that jointly perform a task at any given instant. Therefore, these methods aren't able to take advantage of possibly beneficial diversity that might be present when interacting with a population of agents. Another axis that has recently been getting some interest is based on such a 'community' of agents where different agents in a group speak different languages while communicating with agents from another group (Tieleman et al., 2019; Lowe et al., 2019; Cogswell et al., 2019).

In this work, we propose to dynamically build this population based on agent's past behaviors. We consider fully cooperative games where multiple agents interact within the environment in an iterative manner trying to achieve high task reward while also keeping their language of communication closer to natural language. First, we pretrain the agents with a fixed dataset comprised of human demonstrations and put each of them in separate buffers. Now for each agent, a corresponding meta-agent is trained by interacting it with a population of past copies of other agents present in their respective buffers. Next, each agent interacts with the newly trained meta-agents corresponding to all other agents which is then fine-tuned with the human samples present in the pretraining dataset. Finally, these newly trained agents are added to their respective buffers of old agents which are then

used to train other meta-agents in the subsequent iteration. The process is repeated till adding more agents to the buffer does not change the meta-agents' performance

2 SETUP

Referential games are a type of Lewis Signaling games (Lewis, 1969) that have been used in human as well as artificial language experiments (Havrylov & Titov, 2017; Lee et al., 2018; Gupta et al., 2019). They consist of two players, a speaker $\mathcal{S}$ and a listener $\mathcal{L}$ who interact with each other to solve a common downstream task. Both the agents are parameterized using deep neural networks where the speaker's parameters are denoted by θ and the listener's parameters are denoted by ϕ. The speaker gets as input a target object t, encodes it using its own embedding and finally sends a discrete message $m = \mathcal{S}(t)$ which is obtained through a suitable decoding mechanism to convert the speaker's predictions (logits) to discrete tokens. We use greedy decoding in all our methods and the baselines to extract the discrete message unless explicitly specified. The message m is given as an input to the listener along with the set of distractor objects D and the target object t, shuffled uniformly at random, as a separate input. The listener embeds both of these into a shared vector representation to compute a similarity score between the message and the objects. Finally, it makes a prediction t' about the target object. The reward function r for both agents is the same and is given by $r = 1$ if $(t = t')$ or -0.1 otherwise We denote the number of distractor objects $|D|$ by K, the maximum length of the message m with l and the vocabulary set with V. We optimize the parameters of the two agents on the task performance using reinforcement learning similar to (Evtimova et al., 2018; Lazaridou et al., 2020). The listener is additionally optimized using a supervised learning loss since we know the ground truth label (the target object). Thus, the speaker can be trained using any policy gradient methods (we use REINFORCE (Williams, 1992)) while the listener is trained via policy gradients and a cross-entropy loss. Then the corresponding interactive loss functions for the speaker ($\mathcal{J}_{\mathcal{S}}^{\texttt{int}}$) and the listener ($\mathcal{J}_{\mathcal{L}}^{\texttt{int}}$) are given by:

$$\mathcal{J}_{\mathcal{S}}^{\texttt{int}}(t;\theta) = -\frac{r}{l}\sum_{j=1}^{l}\log p(m_j|m_{<j},t;\theta) + \lambda_{hs}H_{\mathcal{S}}(\theta) \tag{1}$$

$$\mathcal{J}_{\mathcal{L}}^{\texttt{int}}(m,t,D;\phi) = -r\log p(t'|m,t,D;\phi) + \lambda_s \log p(t'=t|m,t,D;\phi) + \lambda_{hl}H_{\mathcal{L}}(\phi) \tag{2}$$

where $H_{\mathcal{S}}$ and $H_{\mathcal{L}}$ denotes entropy regularization for the speaker and listener policies respectively. λ_{hs} and λ_{hl} are non-negative regularization coefficients and $\lambda_s \geq 0$ is a scalar quantity. On the other hand, we can also train these agents on their specific roles in the task independent of the other agent. The speaker's role is to describe the target object accurately and efficiently while the listener's role is to understand a message from a given language along with learning a feature rich representations of the objects and output predictions accordingly. We can collect a set of (object, description) pairs and form a training dataset for the speaker and the listener to separately train both agents via supervised learning without any interactive learning. Now, if we want these descriptions to be interpretable to humans or to allow agents to play with humans, it would be helpful if the agents could speak and understand natural language. So we let humans provide descriptions of the objects and collect such pairs to train our agents. Let us denote this dataset by $\mathcal{D}$. Then the corresponding cross-entropy losses for the speaker ($\mathcal{J}_{\mathcal{S}}^{\texttt{sup}}$) and the listener ($\mathcal{J}_{\mathcal{L}}^{\texttt{sup}}$) are:

3 META-LEARNING AND EMERGENT COMMUNICATION

$$\mathcal{J}_{\mathcal{S}}^{\texttt{sup}}(t;\theta) = -\frac{1}{l}\sum_{j=1}^{l}\sum_{c=1}^{|V|} m^*_{j,c}\log p(m_{j,c}|m_{<j},t;\theta) \tag{3}$$

$$\mathcal{J}_{\mathcal{L}}^{\texttt{sup}}(m^*,t,D;\phi) = -\sum_{j=1}^{K+1}\mathbb{1}_{(t_j=t)}p(t_j|m^*,t,D;\phi) \tag{4}$$

where m^* denotes the description for the target object t present in the dataset.

Let us assume we have a population of agents playing referential game and acting either as a speaker or as a listener. Then we can train a meta-speaker (meta-agent) by playing with a set of listeners (tasks). Similarly, we can obtain a meta-listener by playing with a group of speakers. The hypothesis is that by playing with a diverse group of agents, the meta-agent

should be able to learn new protocols faster. In other words, meta-training should help the meta-agent to generalize faster to a different set of agents (say humans) as compared to single and static pair of speaker and listener.

In this work, we use techniques that use gradient descent for optimizing the meta-agent. In particular, we use the popular MAML (Finn et al., 2017) algorithm to train our meta-agent. We also show some results using other algorithms that are derived from MAML in the Appendix. Similar to the individual agents, both the meta-agents are also parametrized using deep neural networks. We denote the parameters of meta-speaker $\mathcal{S}^m$ by ϑ and meta-listener $\mathcal{L}^m$ by φ. We assume a buffer of speakers denoted by $\mathcal{B}_\mathcal{S}$ and listeners $\mathcal{B}_\mathcal{L}$. We split the dataset $\mathcal{D}$ into a task-specific dataset $\mathcal{D}_t$ consisting of (object, description) pairs used for fine-tuning and a meta-dataset $\mathcal{D}_m$ containing only a set of objects used for computing the meta-objective. We further split $\mathcal{D}_m$ into two disjoint sets[1], $\mathcal{D}_m^0$ and $\mathcal{D}_m^1$ to compute the inner and outer loop losses in MAML respectively. Now, we can define the update rules of the meta-speaker and the meta-listener as follows:

$$\vartheta \leftarrow \vartheta - \beta \nabla_\vartheta \sum_{\mathcal{L} \in \mathcal{B}_\mathcal{L}} \mathcal{J}^{\texttt{int}}_{\mathcal{S}^m}\Big(t^1; \vartheta - \alpha \nabla_\vartheta \mathcal{J}^{\texttt{int}}_{\mathcal{S}^m}\left(t^0; \vartheta\right)\Big) \tag{5}$$

$$\varphi \leftarrow \varphi - \beta \nabla_\varphi \sum_{\mathcal{S} \in \mathcal{B}_\mathcal{S}} \mathcal{J}^{\texttt{int}}_{\mathcal{L}^m}\Big(m^1, t^1, D^1; \varphi - \alpha \nabla_\varphi \mathcal{J}^{\texttt{int}}_{\mathcal{L}^m}\left(m^0, t^0, D^0; \varphi\right)\Big) \tag{6}$$

where $t^0 \in \mathcal{D}_m^0$, $t^1 \in \mathcal{D}_m^1$, $m^0 = \mathcal{S}(t^0)$, $m^1 = \mathcal{S}(t^1)$, and D^0 and D^1 are sets of distractor objects sampled uniformly from $\mathcal{D}_m^0$ and $\mathcal{D}_m^1$ respectively. The fine-tuning losses are the same as the supervised losses Eq equation 3 equation 4 described in the previous section. Initially, both the speaker and the listener are pretrained on the dataset $\mathcal{D}$. Since the number of training iterations is unknown and could be potentially much larger than the size of the buffer the memory can hold, we use reservoir sampling to keep a uniform sample of past agents in the buffer. The detailed algorithm can be found in Algorithm 1.

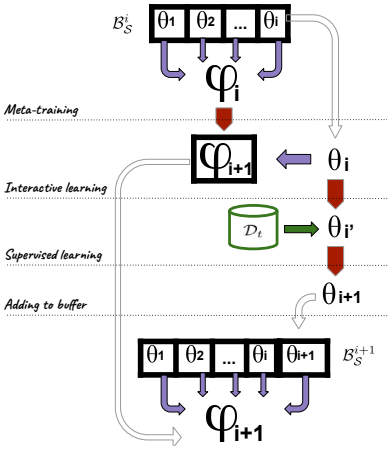


Figure 1: Algorithm outline. We show the different steps involved in training a meta-listener φ. The meta-speaker learning algorithm can be found in the Appendix. A buffer $\mathcal{B}_\mathcal{S}$ contains a list of speakers θ_j where $j \in \{1, 2, \ldots, i\}$. Purple arrows represent interaction between speaker and a listener. Red arrows denote update rule for the corresponding agent. Green arrow represent learning using samples present in the fine-tuning dataset $\mathcal{D}_t$. Hollow arrows are meant for copying the parameters of the specific agent. Black square boxes denote parameter freezing during backpropagation. For more details, refer to Sec 3 and Algorithm 1.

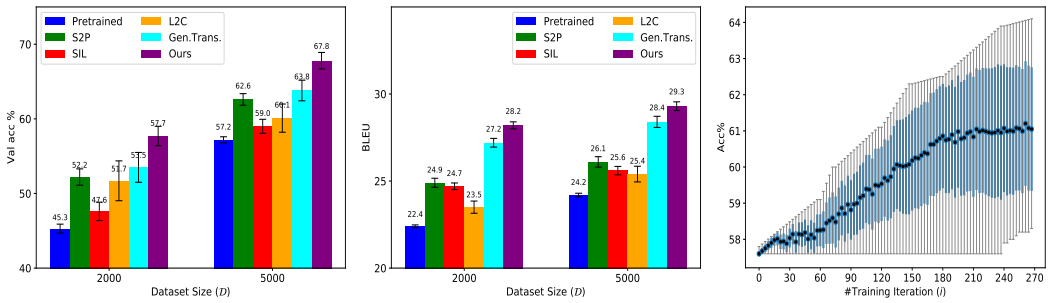
4 EXPERIMENTS

We use the image-based[2] referential game (Lee et al., 2018; Lowe et al., 2020; Lazaridou et al., 2020) which is a common environment used to analyze emergent protocols involving multimodal inputs. A set of images are uniformly sampled from a dataset consisting of diverse images. A target image is chosen uniformly from this set and the rest are set aside as distractors. The task for the listener is to correctly identify the target image among the set of distractors while the sender needs to give a suitable description containing discrete elements present in the target image so as to enable the listener to perform its task. Since we want these agents to understand and talk to humans, the sender is additionally tasked with uttering messages that are closer to natural language.

We use MSCOCO dataset (Lin et al., 2014) to obtain real world images and the corresponding ground truth English captions. Since MSCOCO has multiple gold captions for each image, we randomly

[1]Since we train the agents using minibatch gradient descent, a new split is done at each training iteration to allow more diversity in distractor objects.

[2]We also perform experiments on a novel text-based version of the game and show results in the Appendix.



(a) (b) (c)

Figure 2: (a) Final performance of the agents on the validation set as a function of number of samples in the dataset $\mathcal{D}$. (b) Final BLEU score of the (meta-)speaker agent on the validation set using English sentences. All runs are averaged over 3 random seeds and standard deviations shown. (c) Average referential accuracy on the validation set when a trained meta-listener plays with each speaker present in the buffer separately at the corresponding training iteration. The blue bars show the standard deviation across all agents present in the buffer. The gray bars show the variance between the minimum and maximum performing speaker in the buffer. All agents were trained in the image-game with $|\mathcal{D}| = 5000$ samples.

select one from the available set. Following Lee et al. (2019); Lowe et al. (2020), both speaker and listener are parameterized with recurrent neural networks (GRU (Cho et al., 2014)) of size 512 with an embedding layer of size 256. We use 9 distractors and a pretrained Resnet-50 model (He et al., 2016) (on ImageNet (Deng et al., 2009)) to extract the visual features from images. We set the vocabulary size to 100 and the maximum length of the sentences at 15. We use equal buffer sizes of 200 for reservoir sampling. Other implementation details are given in the Appendix.

5 Results and Analysis

We compute the task performance as well as the language score to evaluate the joint performance of both agents on the validation set. The task performance is measured by the referential accuracy i.e. how accurately the listener is able to predict the target. The language score is computed using BLEU score (Papineni et al., 2002) which is a common metric used in machine translation to compare a candidate translation with reference translations. We compute the BLEU score between the message generated by the speaker and the ground truth caption/sentence to determine the extent to which the speaker can understand and speak natural language. We show the final results for the meta-speaker and the meta-listener on both of these metrics. In Fig 2a we plot the referential accuracy on the validation set of 1000 images for the image game as a function of number of training samples. As we increase the number of samples from 2000 to 5000, the accuracy increases unsurprisingly. In both cases, our method outperforms all previous baselines along with the random chance of selecting the correct target at 10% (9 distractors) by a wide margin. The Pretrained baseline is computed by pretraining the agents using only the task-specific training dataset i.e. without any interactive learning. The emergent communication (*emecom*) baseline (i.e. agents trained via interaction and without any supervision) in this image game is 62.1%. A similar trend is observed in the game with textual inputs. In both cases, Gen.Trans. and our method outperform both single-agent methods, S2P and SIL, suggesting the role of using multiple diverse agents in a population. Moreover, our method surpassing Gen.Trans. and L2C indicates the importance of using the proposed meta-learning approach in conjunction with an adaptive population instead of using a static set of agents. In Fig 2b, we plot the BLEU score using English messages for the image game as a function of number of samples in the dataset. We show that our method again beats all baselines and is even able to perform better than the Pretrained agent improving by almost 6% in $|\mathcal{D}| = 2000$ and 5% in $|\mathcal{D}| = 5000$ game setting. The *emecom* baseline gave a BLEU score of 10.3. As expected, the score is higher when using more human samples in the dataset helping the agent to ground its language into natural language. In Fig 2c, we show the average performance of a trained meta-listener playing with multiple speakers at different stages of training. We show that as the training progresses, the standard deviation across all the speakers present in the buffer at that instant increases indicating that as the population

grows the diversity among the agents also improves. The large difference between the best and worst performing agents (denoted by gray bars) promotes this diversity helping in performing well on both RL and language tasks. We also show a similar plot for the language task (average BLEU score) in the Appendix along with further ablation studies and human evaluation results.

REFERENCES

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://www.aclweb.org/anthology/D15-1075.

Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca D. Dragan. On the utility of learning about humans for human-ai coordination. *CoRR*, abs/1910.05789, 2019. URL http://arxiv.org/abs/1910.05789.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. Compositionality and generalization in emergent languages. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 4427–4442. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.407. URL https://doi.org/10.18653/v1/2020.acl-main.407.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://www.aclweb.org/anthology/D14-1179.

Edward Choi, Angeliki Lazaridou, and Nando de Freitas. Multi-agent compositional communication learning from raw visual input. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rknt2Be0-.

Michael Cogswell, Jiasen Lu, Stefan Lee, Devi Parikh, and Dhruv Batra. Emergence of Compositional Language with Deep Generational Transmission. *arXiv:1904.09067 [cs, stat]*, April 2019. arXiv: 1904.09067.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Katrina Evtimova, Andrew Drozdov, Douwe Kiela, and Kyunghyun Cho. Emergent communication in a multi-modal, multi-step referential game. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJGZq6g0-.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *arXiv:1703.03400 [cs]*, March 2017. URL http://arxiv.org/abs/1703.03400. arXiv: 1703.03400.

Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. Learning to Communicate with Deep Multi-Agent Reinforcement Learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 2137–2145. Curran Associates, Inc., 2016.

Abhinav Gupta, Ryan Lowe, Jakob Foerster, Douwe Kiela, and Joelle Pineau. Seeded self-play for language learning. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pp. 62–66, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6409. URL https://www.aclweb.org/anthology/D19-6409.

Serhii Havrylov and Ivan Titov. Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 2149–2159. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6810-emergence-of-language-with-multi-agent-games-learning-to-communicate-with-sequences-of-symbols.pdf.

K. He, X. Zhang, S. Ren, and J. 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.

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob N. Foerster. "other-play" for zero-shot coordination. *CoRR*, abs/2003.02979, 2020. URL https://arxiv.org/abs/2003.02979.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. Natural language does not emerge 'naturally' in multi-agent dialog. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2962–2967, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1321. URL https://www.aclweb.org/anthology/D17-1321.

Angeliki Lazaridou and Marco Baroni. Emergent multi-agent communication in the deep learning era, 2020.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-Agent Cooperation and the Emergence of (Natural) Language. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Hk8N3Sclg.

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://www.aclweb.org/anthology/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. 4376–4386, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1447. URL https://www.aclweb.org/anthology/D19-1447.

David Lewis. *Convention: A philosophical study*. Harvard University Press, 1969.

Tsung-Yi Lin, M. Maire, Serge J. Belongie, James Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Ryan Lowe, Abhinav Gupta, Jakob Foerster, Douwe Kiela, and Joelle Pineau. Learning to learn to communicate. pp. 5, 2019.

Ryan Lowe, Abhinav Gupta, Jakob Foerster, Douwe Kiela, and Joelle Pineau. On the interaction between supervision and self-play in emergent communication. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rJxGLlBtwH.

Yuchen Lu, Soumye Singhal, Florian Strub, Olivier Pietquin, and Aaron C. Courville. Countering language drift with seeded iterated learning. *CoRR*, abs/2003.12694, 2020. URL https://arxiv.org/abs/2003.12694.

Luke Metz, Niru Maheswaranathan, Brian Cheung, and Jascha Sohl-Dickstein. Learning unsupervised learning rules. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HkNDsiC9KQ.

Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. In *AAAI Conference on Artificial Intelligence*, 2018. URL https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17007.

Alex Nichol, Joshua Achiam, and John Schulman. On First-Order Meta-Learning Algorithms. *arXiv:1803.02999 [cs]*, March 2018. URL http://arxiv.org/abs/1803.02999. arXiv: 1803.02999.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Adam Paszke, Sam Gross, Francisco Massa, et al. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://www.aclweb.org/anthology/D19-1410.

Cinjon Resnick, Abhinav Gupta, Jakob Foerster, Andrew Dai, and Kyunghyun Cho. Capacity, bandwidth, and compositionality in emergent language learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, Richland, SC, 2020. International Foundation for Autonomous Agents and Multiagent Systems. URL https://arxiv.org/abs/1910.11424.

Jonas Rothfuss, Dennis Lee, Ignasi Clavera, Tamim Asfour, and Pieter Abbeel. ProMP: Proximal meta-policy search. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SkxXCi0qFX.

Agnieszka Slowik, Abhinav Gupta, William L. Hamilton, Mateja Jamnik, Sean B. Holden, and Christopher J. Pal. Exploring structural inductive biases in emergent communication. *CoRR*, abs/2002.01335, 2020. URL https://arxiv.org/abs/2002.01335.

Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Learning multiagent communication with backpropagation. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, pp. 2244–2252. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6398-learning-multiagent-communication-with-backpropagation.pdf.

Olivier Tieleman, Angeliki Lazaridou, Shibl Mourad, Charles Blundell, and Doina Precup. Shaping representations through communication: community size effect in artificial learning systems. *NeurIPS Workshop on Visually Grounded Interaction and Language*, 2019. URL https://openreview.net/pdf?id=HkzL4hR9Ym.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. 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. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://www.aclweb.org/anthology/N18-1101.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Andreas Veit Xun Huang Yin Cui, Guandao Yang and Serge Belongie. Learning to evaluate image captioning. In *CVPR*, 2018.

A Related Work

Recent work has tackled similar questions of combining two objective functions of self-play and imitation learning with the goal of training agents that use natural language and perform well on a given emergent communication task. This transforms the problem into training a task conditional language model in a multi-agent setup. S2P (Lowe et al., 2020) investigates and proposes methods that devise a curriculum between the two training phases updating the speaker and the listener in an iterative manner. They use Gumbel-Softmax (GS) techniques to convert the logits to a discrete message allowing end-to-end backpropagation (Mordatch & Abbeel, 2018). Similarly, SIL (Lu et al., 2020) also uses GS to train the two agents in an iterated manner. They use a student-teacher paradigm that is trained sequentially, where the teacher agents, that are initially copies of student agents, are trained using interactive learning and then the student agents are trained to imitate the teacher agents. We note, this method is computationally expensive since it requires sampling data at every training iteration for distillation. Moreover, they use the human (pretraining) dataset to only initialize the student agents and is not used thereafter. Another recent work by (Lazaridou et al., 2020) also explores a similar research question by investigating different types of language (semantic and structural) and pragmatic drift. They propose a reranking module that first samples multiple messages from the speaker and then ranks them according to the task performance[3]. Crucially, this reranking module can be added on top of our method described here and thus is orthogonal to other approaches.

In all these works, the interaction between agents only happens with the current state of the agent and does not involve interaction with their old copies. (Tieleman et al., 2019) propose learning methods using a community of fixed number of agents where the diversity is obtained only via different seeds used for random initialization of parameters. L2C (Lowe et al., 2019) proposes a meta-training method on such a fixed population of agents. This method is closer to our setting but uses a static population of previously stored *trained* agents to train the meta-agent. Moreover, they only proposed a method for training the meta-listener and used GS to allow gradient flow to speakers. We adapt their method to use a REINFORCE loss to enable learning a meta-speaker as well.[4] Another similar method by (Cogswell et al., 2019) aims to learn a community of agents where groups of speakers and listeners are used to sample a pair uniformly at random who then play the game and jointly optimize for better task performance. During learning, few agents are reinitialized periodically/at random from a group of agents. The idea is to promote cultural transmission to induce compositionality over multiple generations of language transfer. For our experiments, we reinitialize agents to the pretrained agents using the human dataset. We denote this method as *Gen.Trans.* in the following sections.

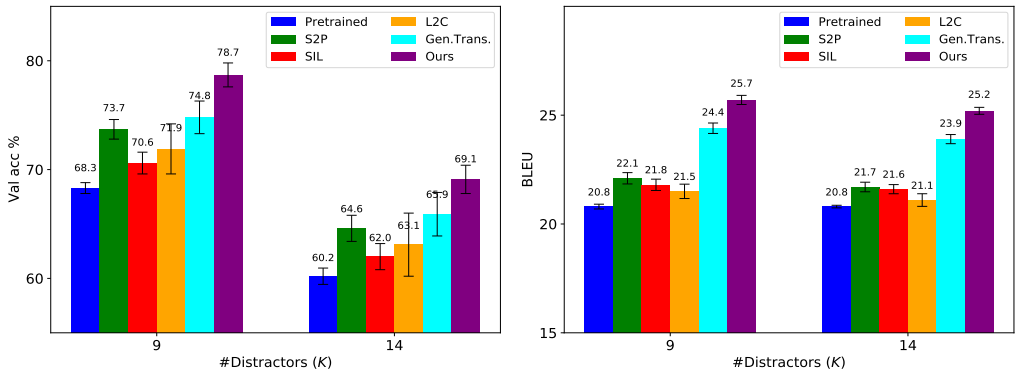
The diversity of agents obtained in such methods at a given iteration is limited and each agent is basically learning a remapping of the same solution resulting in agents achieving a similar performance on both objectives. In our method, we tackle this problem by building a dynamic population where each agent differs from the other agents on both metrics. Specifically, we add agents at various stages of training into a buffer that is then used to train a meta-agent. We hypothesize that using our proposed population training method, we can train agents that are able to perform better than single agent methods by countering the various drifts efficiently and utilizing the population diversity effectively while being faster to train and robust to environmental design.

B Experiments on Text Game

We propose to use text as an alternative input modality used in referential games (Lee et al., 2019). This allows us to use a different set of input representations that could be very different from the visual features that also encodes some form of spatial structures present in the images (Slowik et al., 2020). So we replace images in the aforementioned referential game with variable length sentences consisting of discrete words. Now, the task of the speaker is to translate a sentence from an input language to an intermediate language such that the listener is able to pick the correct sentence given other sentences in the input language.

[3]Their approach involves giving speaker access to the distractor objects which is not the case in our setup.

[4]Training via REINFORCE performed better than GS where we had to backpropagate gradients by freezing all listener's parameters.



(a) (b)

Figure 3: (a) Final validation accuracy as a function of number of distractor sentences. (b) Final BLEU score of the (meta-)speaker agent on the validation set using German sentences. All runs are averaged over 3 random seeds and standard deviations shown.

We use the publicly available IWSLT'14 English-German dataset for this purpose. We choose English as the input language and German as the intermediate language. Instead of randomly selecting the distractors from the set of English sentences, we use a pretrained Sentence-BERT model (Reimers & Gurevych, 2019) (on SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018)) to systematically choose them. Sentence-BERT outputs a dense vector representation (sentence embedding) given an input sentence. We use cosine similarity to compare these embeddings of two sentences and build a cluster consisting of sentences with minimum threshold similarity. The set of distractors are then chosen from this cluster. The higher the cosine similarity, the harder it is for the listener to identify the correct sentence. For our experiments, we keep the cosine similarity to 0.85 and the number of distractors to 14. The agents' architectures and other preprocessing details are the same as the game with images. We use the pretrained Sentence-BERT embeddings to get the textual features from the raw English sentences.

B.1 Results on Text Game

In Fig 3a, we show results for two sets of distractors, 9 (random choice 10%, *emecom* 72.6%) and 14 (random choice 6.7%, *emecom* 64.4%) for a fixed dataset size of 5000. As expected, the referential accuracy drops as we increase the number of distractors, making the game harder.

We also perform some ablation studies in Sec D.1. An interesting point to note is that given the same number of distractors (9), agents trained on the game with text perform better than the game with images. We think that this could be attributed to the pretrained textual features which are more aligned with the features suited for the task objective than the visual representations where the task objective is to learn maximally distinct features for all sentences in the dataset.

In Fig 3b, we show the BLEU score between the generated German sentences and the ground truth German translation in the text game by varying the number of distractor sentences from 9 (*emecom* 10.8) to 14 (*emecom* 9.9). We observe that the average score across all methods decreases as the game complexity (K) increases. We think that this is due to the potential misalignment of the task performance, that gets more challenging as K grows, with the linguistic loss of the speaker.

C Improved Captioning and Translation Model

We take the analysis of the meta-speaker in the game with images a step further and aim to evaluate the alignment of the generated captions with human judgement. We use the metric proposed by (Yin Cui & Belongie, 2018) to perform automatic evaluation of captions using a pretrained model on the MSCOCO dataset. Their method computes a similarity score between the candidate (generated) caption and the context (image and ground-truth caption). We evaluate 1000 generated captions from

Algorithm 1: Algorithm

Input : Datasets $\mathcal{D}_t$ and $\mathcal{D}_m$, pretrained speaker parameters θ and listener parameters ϕ, randomly initialized meta-speaker parameters ϑ and meta-listener parameters φ, empty buffers $\mathcal{B}^0_{\mathcal{S}}$ and $\mathcal{B}^0_{\mathcal{L}}$

$i \leftarrow 1$
$\theta_i \leftarrow \theta$
$\phi_i \leftarrow \phi$
repeat
  /* Add current agents to buffer */
  $\mathcal{B}^i_{\mathcal{S}} \leftarrow \mathcal{B}^{i-1}_{\mathcal{S}} \bigcup \theta_i$
  $\mathcal{B}^i_{\mathcal{L}} \leftarrow \mathcal{B}^{i-1}_{\mathcal{L}} \bigcup \phi_i$
  /* Meta-training loop */
  for $j \in \{1, 2, \ldots, n_{meta}\}$ **do**
    Train ϑ_i by playing with each $\phi \in \mathcal{B}^i_{\mathcal{L}}$ using Eq equation 5
    Train φ_i by playing with each $\theta \in \mathcal{B}^i_{\mathcal{S}}$ using Eq equation 6
  end
  /* Initialize i+1 parameters */
  $\theta_{i+1} \leftarrow \theta_i$
  $\phi_{i+1} \leftarrow \phi_i$
  /* Interactive learning loop */
  for $j \in \{1, 2, \ldots, n_{int}\}$ **do**
    Train θ_{i+1} by playing with φ_i using Eq equation 1
    Train ϕ_{i+1} by playing with ϑ_i using Eq equation 2
  end
  /* Supervised learning loop */
  for $j \in \{1, 2, \ldots, n_{sup}\}$ **do**
    Train θ_{i+1} with samples present in $\mathcal{D}_t$ using Eq equation 3
    Train ϕ_{i+1} with samples present in $\mathcal{D}_t$ using Eq equation 4
  end
  $i \leftarrow i + 1$
until *performance of ϑ and φ converge*

	$\lvert\mathcal{D}\rvert = 2000$	$\lvert\mathcal{D}\rvert = 5000$
S2P	5.5	6.7
SIL	5.5	6.6
L2C	5.4	6.4
Gen.Trans.	6.1	7.2
Ours	6.6	8.0
Oracle	*12.1*	

Table 1: Evaluating the (meta-)speaker on the captioning task using the metric proposed in (Yin Cui & Belongie, 2018). The Oracle in the table refers to the best score obtained by a trained captioning model using more than $100k$ samples as found in (Yin Cui & Belongie, 2018).

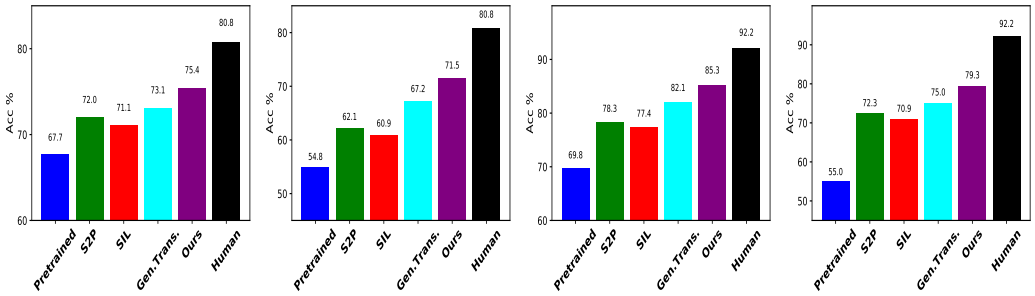
	BLEU
Pretrained + Greedy Decoding	20.8
Pretrained + Beam Search ($n = 2$)	21.1
Pretrained + Beam Search ($n = 4$)	22.4
Pretrained + Top-k Sampling ($k = 40$)	24.6
Ours (using greedy decoding)	25.7

Table 2: Comparative analysis of decoding strategies. We show the BLEU score on German sentences for (meta-)speaker in the text game trained with 9 distractor sentences and $|\mathcal{D}| = 5000$ human samples. Here n denotes number of beams in beam-search.

the validation set for each method and average the scores across all captions. The results are showed in Table 1. Our method beats all other baselines[5] using just a few human samples.

In Table 2, we perform comparative analysis using various decoding strategies with the Pretrained model in the text game and show the BLEU scores for English-German translation task. We show that our method that uses the simple greedy decoding is able to outperform all the variants of Pretrained model using sophisticated decoding strategies.

This implies that even though the task objective was not to train a better captioning or a translation model, we were able to obtain improved performances on both tasks given the limited set of examples. We note that we do not claim to propose the state-of-the-art captioning or translation model but instead show this analysis to reckon that our method can be combined with specialized models built for captioning or translation tasks. Furthermore, we show some qualitative samples by comparing the speaker generated messages across all baselines along with the ground truth captions.



(a) (b) (c) (d)

Figure 4: Human evaluation on the text game (a),(b) and the image game (c), (d). Here, (a),(c) refer to the referential accuracy of (Meta-speaker, Human-listener) pair and (b),(d) refer to the referential accuracy of (Human-speaker, Meta-listener) pair. The black bar in all plots represent the performance of the (Human-speaker, Human-listener) pair.

C.1 Human Evaluation

Although BLEU score is able to capture some form of syntactical and semantic drifts, it still fails to counter the phenomenon of pragmatic drift as introduced in (Lazaridou et al., 2020). For this reason, we evaluate the performances of our agents and the baselines by playing them with humans. We play both the (meta-)speaker with a human listener and the (meta-)listener with a human speaker separately. The (meta-)speaker is evaluated using 1000 human samples while the (meta-)listener is played 400 times. The final performance is computed using the referential accuracy. We also let humans play the game with other humans with similar game restrictions which defines our oracle. The participants weren't given the identity of the other agent they are playing with to make a fair comparison. In Fig 4c, we show the results for meta-speaker for the image game where our method outperforms other baselines by a significant margin. In Fig 4d we compare the performance of the meta-listener for the image game with other baselines and even here our method outclass them. This denotes that our method is able to understand human descriptions more accurately by learning diverse caption representations. The results for the game with text

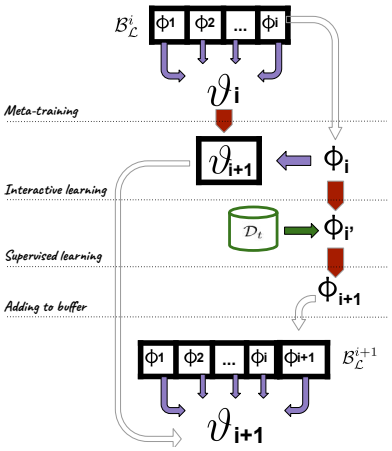
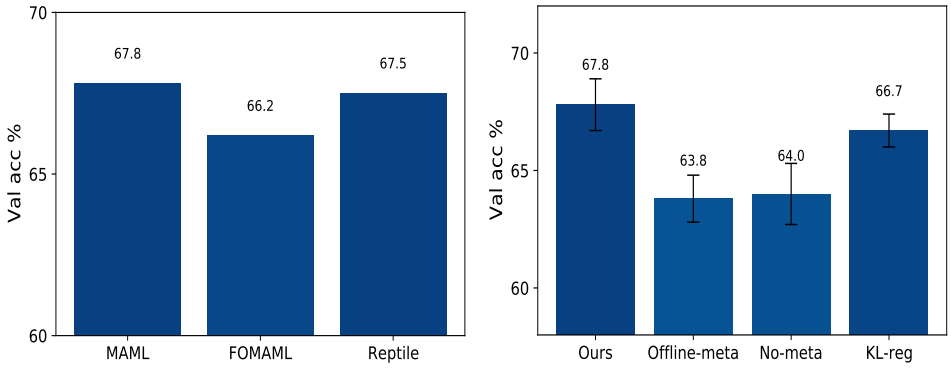


Figure 5: Algorithm outline for training a meta-speaker ϑ. A buffer $\mathcal{B}_{\mathcal{L}}$ contains a list of listeners ϕ_j where $j \in \{1, 2, \ldots, i\}$. See Sec 3 and Algorithm 1 for more details.

[5]The Pretrained performance is the same as the Oracle as it always outputs a description from the training set making it closer to a human-defined caption but suffers from the task accuracy.



(a) (b)

Figure 7: (a) Average referential accuracy on the validation set when a trained meta-listener plays with each speaker present in the buffer separately at the corresponding training iteration. The blue bars show the standard deviation across all agents present in the buffer. The gray bars show the variance between the minimum and maximum performing speaker in the buffer. All agents were trained in the image-game with $|\mathcal{D}| = 5000$. (b) Ablation study in the image game with $|\mathcal{D}| = 5000$.

follow similar pattern and can be found in Fig 4a 4b. We used 9 distractor objects and $|\mathcal{D}| = 5000$ for both games. The (meta-)speaker experiments were conducted with 1000 times (image game) and 350 times (text game) while the (meta-)listener was played 400 times (image game) and 100 times (text game) for each baseline. We present the results of the human evaluation experiment on the text game here. Similar to the image game, we show that agents trained using our method beat all prior baselines when paired both with both human listeners and human speakers. Furthermore, we observe that although the overall performance of the agents in the text game is lower (including the human pair), the gap between our agent and the human pair is smaller than the image game. We think that this could be attributed to the text game being harder (even for humans) than the image game due to systematically choosing the set of distractor objects and image game possibly having an easier way of discovering the 'differentiators' that help the agents to uniquely identify the target object. The experiment setup is the same as the image game. Overall we infer that our method suffers from the least amount of pragmatic drift as compared to other baselines measured by the performance gap with human-human gameplay.

D Further Results

We plot the average BLEU score of multiple speakers playing with a trained meta-listener in Fig 6 at different stages of training. Similar to Fig 2c, we show that the speakers learnt to speak wide variety of languages that are different from English as measured by the BLEU score.

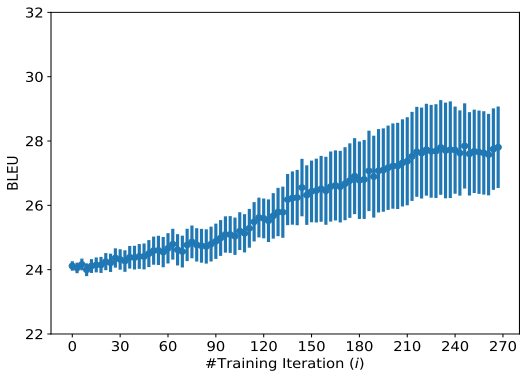


Figure 6: Average BLEU score on the validation set when a trained meta-listener plays with each speaker present in the buffer individually at the corresponding training iteration. All agents were trained in the image-game with $|\mathcal{D}| = 5000$.

D.1 Ablation Study

We further analyze the importance of each component of our proposed algorithm. Specifically, we compare three major baselines (a) Offline-meta: in the interaction phase, instead of playing with the other meta-agents, the agent plays with the current other agent itself (e.g., the speaker playing against the current listener instead of the meta-listener), in turn making the training of meta-agents interdependent of each other (b) No-meta: instead of learning a meta-agent, the

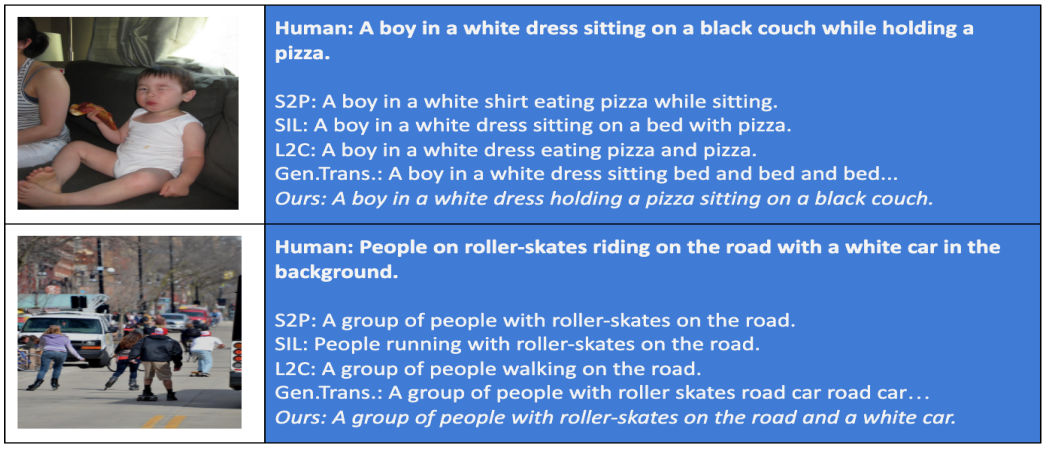


Figure 8: Qualitative samples generated by the (meta-)speaker in the game with images.

agent plays with past copies of other agents stored in a buffer demonstrating a knowledge gap similar to (Cogswell et al., 2019) but without resetting. (c) KL-reg: instead of having separate phases of interactive learning and supervised learning to avoid catastrophic forgetting, the agent is trained by minimizing a linear combination of (i) KL divergence of its distribution with the Pretrained model and (ii) interaction loss with the meta-agent, similar to (Lazaridou et al., 2020; Lu et al., 2020) In Fig 7b, we compare the (meta-)listener performance across all these baselines by measuring the referential accuracy on the validation set in the image game. We observe that both No-meta and Offline-meta achieve lower accuracy as compared to our method. This suggests that our meta-agent captures useful information from the past iterations of other agents while also helps in learning of the current other agent. KL-reg performance was close to our method indicating that one can use alternate ways to integrate the two loss functions in combination with our meta-learning approach.

We also performed an ablation study using different meta-learning algorithms (Finn et al., 2017; Nichol et al., 2018). FOMAML is the first-order approximation of MAML and Reptile is another first-order meta-learning algorithm that performs stochastic gradient descent for a few steps across all tasks and then updates the parameter towards the average of updated task-specific weights. We show the results (task performance) on the image game with $|\mathcal{D}| = 5000$ in Fig 7a. The performances of the all algorithms are competitive with each other indicating robustness across the three methods. Furthemore, we think that recent advancements in meta-learning algorithms (Rothfuss et al., 2019; Metz et al., 2019) could be combined with our algorithm to further analyze this effect and investigate biases resulting from a given meta-algorithm.

E HYPERPARAMETERS

We show here the range of parameter configurations we tried during training (bold indicates the ones used in the experiments):

- Batch Size: 512, **1024**
- Buffer Size: 50, 100, **200**
- Learning rate (outer loop): **1e-4**, 1e-5, 6e-5, 6e-4
- Learning rate (inner loop α): **1e-4**, 3e-4
- n_{meta}: 20, 40, **60**, 65, 70
- n_{sup}: 10, 20, **25**, 30
- n_{int}: 40, 60, 70, **80**, 100
- λ_{hs}: 0.1, **0.01**, 0.001
- λ_{hl}: 0.1, **0.03**, 0.007, 0.001

- λ_s: 0.1, 0.5, **0.8**, 1
- $\mathcal{D}_t$: 500, **1000**, 1500, 2500, **3500**, 4000
- $\mathcal{D}_m^{0,1}$: (1200,300), **(700,300)**, (400,100), (1700, 800), **(1000,500)**, (1200,300)

We use the Adam optimizer (Kingma & Ba, 2015) in PyTorch (Paszke et al., 2019) for training the agents. For the baselines (S2P, SIL, L2C, Gen.Trans.), we used the publicly available repositories attached with the respective publications. We adapt their codebase to train agents on the two referential games used in this work while tuning some hyperparameters to adapt to the task.