Multi-lingual agents through multi-headed neural networks

Jonathan D. Thomas*[1], Raúl Santos-Rodríguez[1], Mihai Anca[1], and Robert Piechocki[1]

[1]University of Bristol

Abstract

This paper considers cooperative Multi-Agent Reinforcement Learning, focusing on emergent communication in settings where multiple pairs of independent learners interact at varying frequencies. In this context, multiple distinct and incompatible languages can emerge. When an agent encounters a speaker of an alternative language, there is a requirement for a period of adaptation before they can efficiently converse. This adaptation results in the emergence of a new language and the forgetting of the previous language. In principle, this is an example of the Catastrophic Forgetting problem which can be mitigated by enabling the agents to learn and maintain multiple languages. We take inspiration from the Continual Learning literature and equip our agents with multi-headed neural networks which enable our agents to be multi-lingual. Our method is empirically validated within a referential MNIST-based communication game and is shown to be able to maintain multiple languages where existing approaches cannot.

1 Introduction

Questions pertaining to communication naturally arise when considering Multi-Agent systems. It is natural as communication is such a vital part of our societies, enabling for the dissemination of ideas and large-scale coordination. By equipping agents with capacity to communicate they will likely be able to achieve greater levels of synergy with both artificial and biological entities.

This paper focuses on emergent communication within multi-agent reinforcement learning (MARL), specifically addressing settings where agents can be considered as independent learners (IL)[1]. This restriction removes common methodologies which are utilised to improve training speed and stability, such as centralised training decentralised execution (CDTE) [2], parameter sharing [1] and gradient propagation through other agents. This is justified by the motivation of creating algorithms that better approximate human learning, where, for example, models of other agents are unlikely to be available for gradient propagation. Recent work has attempted to improve training efficiency through a variety of methods. Jaques et al. [3] proposes an intrinsic reward based on social influence to encourage communication of useful information and Eccles et al. [4] proposes the introduction of biases to promote the emergence of communication.

In this more natural setting, experimentation has generally been restricted to two independent agents. However, realistic scenarios are likely to involve larger numbers of independent agents interacting at varying frequencies. As the agents do not use parameter sharing, it is conceivable that multiple unique languages may arise where these languages are unlikely to be compatible. As result of this, any interaction with a new agent mandates the learning of a shared language. Without specific modifications to the agent's architecture, this new language will overwrite the previous one as a consequence of a known phenomena within machine learning (ML) named catastrophic forgetting [5]. As the previous language has been lost, any interaction with the associated conversational partner will require re-training. Here, in order to address this issue, architectural modifications inspired by the Continual Learning literature are used to extend the algorithm proposed by Eccles et al. [4]. Namely, multi-headed neural networks are used where a different head is maintained for each language. This paper formalises this concept and demonstrates it within a novel environment called *Communication*

*Corresponding Author: jt17591@bristol.ac.uk

https://doi.org/10.7557/18.6818

© The author(s). Licensee Septentrio Academic Publishing, Tromsø, Norway. This is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

Carousel which extends a referential game to facilitate study of this adaptation problem.

2 Related Work

The general challenge of inter-agent communication has attracted much attention within the MARL community. A variety of approaches have been recently proposed. A few of the most relevant include RIAL [2], DIAL [2], CommNET [6], TarMAC [7] or DGN [8]. Works in the area of inter-agent communication can be loosely categorised into two main types, namely those that allow gradients to flow between agents and those that do not. Recently, there has been interest in the latter domain, whereby facilitation of centralised training and parameter sharing are removed and agents are only allowed to train via the environment reward. This is sometimes referred to as Independent Learners [1].

Independent Learners in emergent communication. Despite the additional difficulty, it is often argued that this is a more realistic setting as it is closer to the methods by which humans learn. State-of-the-art examples include Jaques et al. [3] and Eccles et al. [4]. In Jaques et al. [3], an intrinsic reward derived from causal influence is used to encourage the speaker to send messages that change the listeners policy. Differently, Eccles et al. [4] introduces biases into both the speaker and the listener. Here, the speaker is encouraged to maximise mutual information between its observation and its message while the listener is encouraged to modify its policy in response to the reception of a message. While both methods are related, following Lowe et al. [9], we can summarise Eccles et al. [4] as encouraging positive signalling and positive listening whereas Jaques et al. [3] only encourages positive signalling. In this paper we use Eccles et al. [4] as a baseline for our experimental work, as it can be shown to outperform Jaques et al. [3] in our setting.

Zero-shot coordination. A related area of growing interest is zero-shot coordination (ZSC) [10, 11, 12], where the objective is to derive policies for cooperative settings which allow for previously unseen partners. Hu et al. [10] consider issues posed by the standard self-play methodology where learnt policies are not compatible with novel partners due to agents not being able to exploit potential known symmetries in coordination tasks. They propose the *Other-play* algorithm (OP), which involves techniques based on domain randomisation. Treutlein et al. [11] build upon Hu et al. [10], formalising the setting as a label-free coordination problem (LPCB). Finally, Bullard et al. [12] explicitly consider communication within ZSC. The setting they study involves a costed communication channel with a non-uniform distribution over messaging intents. Based on OP, they introduce Quasi-Equivalence Discovery (QED).

Our work elaborates upon previous contributions within the emergent communication literature. We follow a deviation from the standard ZSC setting as in Bullard et al. [12]. In our setting multiple pairs of speakers and listeners are allowed to develop potentially unique languages. We then address how to both learn and maintain multiple languages and the mitigation of the issues introduced by catastrophic forgetting.

3 Setting

The MARL approach defined within this paper is applied to an N-player partially-observable Stochastic game [13], G. Where G is defined by the tuple $G = (S, A_1, ..., A_n, M_1, ..., M_n, T, O, r)$. The environment state is defined by $s \in S$. At each time-step each agent makes a local observation of the environment state according to the observation function $O : S \rightarrow o$. In addition to an agent's observation o, it also receives all messages from the previous time-step $\mathbf{m}$ (excluding it's own message). Using this information agents select an action $a_i \in A_i$ according to $\pi_{i,a}$ and a discrete message $m \in M_i$ according to the policy, $\pi_{i,m}$. All agents actions make up the joint action $\mathcal{A} = A_1 \times ... \times A_n$, which results in a state transition according to $T : S, \mathcal{A} \rightarrow S$ and all agents receive a reward $r : S, \mathcal{A} \rightarrow \mathcal{R}$. This work is constrained to fully-cooperative games where communication is provably advantageous. Agents are tasked with finding action policies $\pi_{i,a} : (o, m) \rightarrow A_i$ and a message policy $\pi_{i,m} : (o, m) \rightarrow M_i$ such that the cumulative discounted reward is maximised.

4 Method

4.1 Problem Statement

Let us consider the existence of two sets of agents, where these are referred to as speakers $T_x = \{\pi_{s,0}, ...\pi_{s,n}\}$ and listeners $R_x = \{\pi_{l,0}, ...\pi_{l,n}\}$, respectively[1]. All agents are parameterized by deep neural networks (DNN) according to the methodology described by Eccles et al. [4], where this includes introduction of inductive biases to promote the emergence of communication. For some pairing of T_x to R_x, the agents capacity to effectively convey information will be limited by their ability to understand one another. Overtime, the agents can adapt to each other and arrive at an emergent protocol which maximises task reward.

The first question this work intends to delve into is, what happens to their established emergent protocol when an agent (be that the speaker or the listener) interacts with a new partner? More formally, when the mapping from T_x to R_x is randomised and a period of training is allowed, how does this impact the agent's capacity for conversation with its previous partner? This problem exists within the continual learning setting, where Catastrophic Forgetting is known to be an issue [5]. It should be expected that as a pair of agents build up familiarity with one another, their previous languages will drift.

The fundamental issue with this mode of operation is that it always requires an agent to re-train upon interacting with a different partner even if they had previously arrived at an efficient protocol. Ideally, this should be avoided as this period of adaptation is costly. Naturally, the second question is simply, how can we mitigate this issue?

4.2 Multi-headed agents

As mentioned above, this primary issue in our scenario is Catastrophic Forgetting [5]. Following the naming convention from Delange et al. [14], our approach considers a simple parameter isolation method, where each speaker and listener maintains a separate output head for each possible partner. This idea is based on Donahue et al. [15]. It is assumed that the identity of each potential partner is observable and therefore the correct head can be chosen.

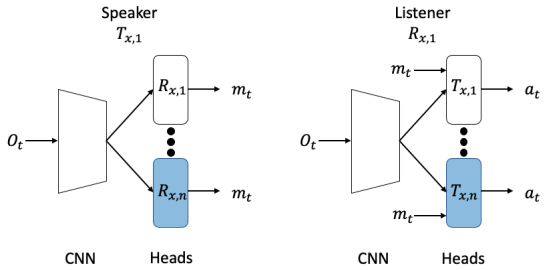


Figure 1: Speaker and listener DNN architecture shown on the left and right, respectively. The networks maintain a separate head for each partner, where the label indicates the conversational partner that it refers to. The white and blue colouring is representative of how gradients are allowed to propagate through the network. In both cases the CNN and first head are trained together, whereas the alternative heads are trained separately.

The architecture is presented in Fig. 1, where the CNN for both the speaker and listener are only trained with the first partner. This decision is justified by the assumption that, in most cases, languages consider mappings from a similar set of concepts to different words or phrases and, as such, the features learned by the CNN for one language should be transferable. An additional variant upon this model is proposed in which the weights of the non-primary heads are pre-initialised with those of the primary head upon establishment of the first language. This can be demonstrated to improve sample efficiency when compared to random initialisations.

5 Experiments

5.1 Implementation

All code is implemented in Pytorch [16] according to the methodology described in Section 4[2]. As previously introduced, the implementation of the

[1] To avoid clashes with standard RL notation, the speakers and listeners have symbols consistent with transmitter and receiver.

[2] Code available at `https://github.com/Jon17591/multi-lingual-agents`

speaker and listener follows the methodology described by Eccles et al. [4], where we train agents independently utilising REINFORCE and utilise the same hyperparameters. As we were unable to achieve convergence with the defined architecture we made one modification. We introduced an extra layer into the DNN which alleviated this issue. This minor modification to the method proposed by Eccles et al. [4] without the multi-headed output is utilised as a baseline within our experimentation.

5.2 Communication Carousel

This work intends to investigate the agents' capacity to maintain emergent languages after interacting with new partners. To achieve this, N-parallel referential games are instantiated and speakers and listeners are afforded E episodes with their initially assigned partner. After the initial E episodes, the agents are rotated and allowed the same number of episodes to interact with their new partner. We name this environment *Communication Carousel* and an illustration is provided in Figure 2. After a number of partner changes, ω, the speakers and listeners are returned to their initial partner and afforded a further E episodes to reconverge. All experimental parameters are introduced in Table 1. This environment formulation provides a simple and interpretable test-bed for studying agent adaptation where the complexity can be easily controlled through appropriate selection of the referential game.

The referential game maintains broadly the same structure as Eccles et al. [4] which is a simple MNIST based game. It comprises of two agents, a speaker and a listener who are both observe o_s and o_l which are images sampled from the MNIST dataset [17]. The speaker's input is an image from the dataset and it's output is a discrete discrete message m_t which gets passed to the listener. The listener observes it's own image and the speaker's message and is tasked with adding the two together, where it's answer is represented by it's action a_t. If the action is equal to the summation of the digits both agents receive a reward of 1, otherwise the reward is -1. By design, this game can only be successfully completed if an effective language is derived.

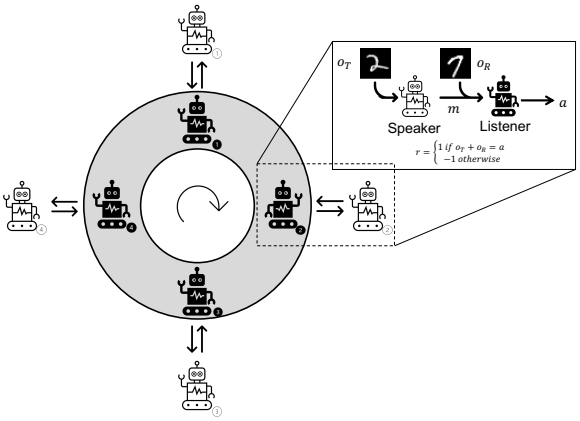


Figure 2: *Communication Carousel*, N-parallel referential games. After E episodes, the carousel rotates and all agents interact with a different partner. This continues for the desired number of rotations after which all agents are returned to their original partner for assessment of emergent language maintenance.

Table 1: Parameters used in carousel environment

Symbol	Meaning	Value
N	Number parallel environments	4
E	Episodes per interaction	75k
ω	Number of rotations	1

6 Results and Discussion

The results obtained support the hypothesis that the Multi-headed methods defined within Section 4.2 results in better maintenance of multiple emergent languages.

Figure 3 demonstrates the average reward which agents receive with their current conversational partner for the baseline, Multi-headed method and the Multi-headed method with pre-intialisatation of the non-primary heads. The most notable observation to draw from this Figure is that the reward for the baseline method reduces substantially when it returns to the initial conversational partner at $150k$ episodes, this reduction is not present in either of the Multi-headed method. This would suggest that Catastrophic Forgetting has been avoided. This claim is further supported by Figure 4a, 4b and 4c. These figures show the average reward obtained by all pairings of speakers and listeners in

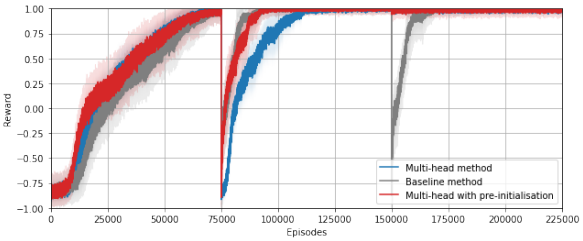
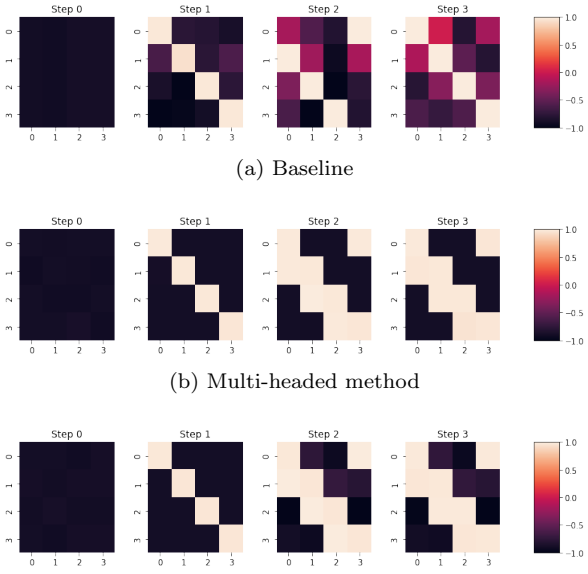


Figure 3: Average reward obtained by all 4-agents with their current partner. Partner is changed to a new partner at 75000 episodes and then to the original partner at 150000 episodes.



(a) Baseline

(b) Multi-headed method

(c) Multi-headed method with pre-initialisations

Figure 4: Heatmap for all method evaluated for all pairings at episodes=0, 75000, 150000 and 225000. Scale represents the average reward which is obtained over 100 episodes.

the form of a heatmap. The steps refer to the beginning of training, after every partner switch and at the end of training where this corresponds to episodes 0, 75k, 150k and 225k in Figure 3. Note that the baseline method experiences a significant reduction in reward acquisition once it has trained with a new partner whereas this is not present in either of the Multi-headed methods.

A drawback of the standard Multi-headed method appears to be the reduction in sample efficiency present when switching to the second partner (75k episodes) in Figure 3. The Multi-headed method seems to take longer to acquire the second language. This is as the additional heads are untrained and comprise of randomly initialised weights. The baseline method represent a policy that has converged to a solution. The entropy of both sets of speaker policies (shown in Figure 5) gives an indication as to why this occurs. It is clear that the Multi-headed method begins with significantly higher entropy. The introduction of this extra stochasticity may make the arrival at a common protocol more time intensive as there is less determinism to the respective messages and, as such, it is more difficult to achieve synchronisation between the agents. This can be overcome by pre-initialising the weights of each head with the solution of the primary head, thereby achieving comparable convergence speeds to the baseline.

7 Future Work

A current limitation of our method which we hope to address is that all derived languages are unique. The resulting multi-agent system has a quadratic relationship between the number of languages and the number of speakers/listeners. This is not the case in natural systems with the number of distinct languages being somewhat restricted. An interesting avenue to explore could consider methodologies which restrict the number of languages that may emerge thereby aiming to improve zero-shot performance.

Furthermore, although we focus on emergent communication in this work, we believe the results presented apply more generally to cooperative games. The idiosyncratic conventions which cooperative agents' can develop are equivalent to the languages which arise in referential games. In future work, we intend to expand our analysis to consider a broad set of cooperative problems where periodic interactions may arise. We believe this mode of operation and methodology may be applicable in a range of human-centric tasks where the personalisation of policies may be desirable.

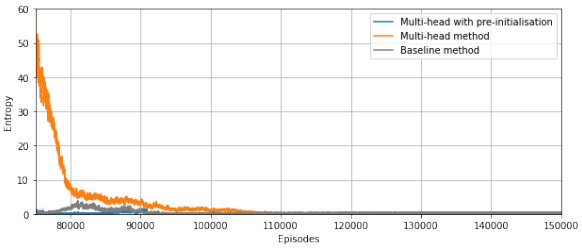


Figure 5: Entropy of speaker from 75k to 150k episodes.

8 Conclusion

We consider the development of agents which can maintain multiple languages without falling victim to catastrophic forgetting. This work builds upon that by Eccles et al. [4] and introduces a parameter isolation method into their neural network in order to mitigate the aforementioned issues. The modification involves the utilisation of a multi-headed output network, where each head is utilised for a specific language. This approach was validated empirically within a novel referential game formulation which facilitated evaluation of language maintenance through interactions with multiple unique agents and will serve as a simple test-bed for future work. The results demonstrate that the proposed method effectively avoids catastrophic forgetting when compared to the standard implementation of Eccles et al. [4]. Future work intends to consider this methodology within more complex domains and zero-shot scenarios.

Acknowledgement

This work is funded by the Next-Generation Converged Digital Infrastructure (NG-CDI) Project, supported by BT and Engineering and Physical Sciences Research Council (EPSRC), Grant ref. EP/R004935/1. RSR is partially funded by the UKRI Turing AI Fellowship EP/V024817/1.

References

[1] Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pages 330–337, 1993.

[2] Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2145–2153, 2016. ISBN 9781510838819. doi: 10.5555/3157096.3157336.

[3] 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 *International Conference on Machine Learning*, pages 3040–3049. PMLR, 2019.

[4] Tom Eccles, Yoram Bachrach, Guy Lever, Angeliki Lazaridou, and Thore Graepel. Biases for emergent communication in multi-agent reinforcement learning. *Advances in neural information processing systems*, 32, 2019. doi: 10.5555/3454287.3455463.

[5] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, (4):128–135, 1999. ISSN 1364-6613. doi: https://doi.org/10.1016/S1364-6613(99)01294-2.

[6] Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. Learning multiagent communication with backpropagation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2252–2260, 2016. ISBN 9781510838819. doi: 10.5555/3157096.3157348.

[7] Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. Tarmac: Targeted multi-agent communication. In *International Conference on Machine Learning*, pages 1538–1546. PMLR, 2019.

[8] Jiechuan Jiang, Chen Dun, Tiejun Huang, and Zongqing Lu. Graph convolutional reinforcement learning. In *International Conference on Machine Learning*, 2020.

[9] Ryan Lowe, Jakob Foerster, Y-Lan Boureau, Joelle Pineau, and Yann Dauphin. On the pitfalls of measuring emergent communication. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2019. doi: 10.5555/3306127.3331757.

[10] Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "other-play" for zero-shot coordination. In *International Conference on Machine Learning*, pages 4399–4410. PMLR, 2020. doi: 10.5555/3524938.3525347.

[11] Johannes Treutlein, Michael Dennis, Caspar Oesterheld, and Jakob Foerster. A new formalism, method and open issues for zero-shot coordination. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10413–10423. PMLR, Jul 2021.

[12] Kalesha Bullard, Douwe Kiela, Franziska Meier, Joelle Pineau, and Jakob Foerster. Quasi-equivalence discovery for zero-shot emergent communication. *arXiv preprint arXiv:2103.08067*, 2021.

[13] L. Shapley. Stochastic games*. *Proceedings of the National Academy of Sciences*, 39:1095 – 1100, 1953. doi: 10.1073/pnas.39.10.1095.

[14] Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. doi: 10.1109/TPAMI.2021.3057446.

[15] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 647–655. PMLR, Jun 2014. doi: 10.5555/3044805.3044879.

[16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. 2019. doi: 10.5555/3454287.3455008.

[17] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.