ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2.
Available from http://www.i6doc.com/en/.

Influence of image encoders and image features transformations in emergent communication

Bastien Vanderplaetse[1], Stéphane Dupont[1] and Xavier Siebert[2] *

1- University of Mons - MAIA Artificial Intelligence Lab - Belgium

2- University of Mons - Mathematics and Operational Research - Belgium

Abstract. Emergent communication in multi-agent systems is a research field exploring how autonomous agents can develop unique communication protocols without human programming, showing adaptability in various contexts. This study investigates the influence of image encoders and spatial information within image features on agent performance and the compositionality of emergent languages in multi-agent systems. By exploring various image encoding strategies, including the application of different processing methods to image features, we assess their impact on agents' abilities in a structured communication task. Our findings indicate that while certain encoding processes enhance overall task performance, they do not necessarily improve language compositionality.

1 Introduction

The domain of emergent communication within multi-agent systems lies at the intersection of computational linguistics, artificial intelligence, and robotics, targeting the challenge of enabling autonomous agents to develop and utilize complex communication protocols. Traditional approaches oscillate between rigid handcrafted protocols, adapted for specific tasks but limited by their complexity, and learned continuous protocols, which are more flexible but suffer from issues of interpretability. These issues lead to the emphasis on using discrete symbols and hierarchical syntactic arrangements inspired by the inherent properties of human language, renowned for its robustness and capacity to convey complex concepts across novel contexts. This is why emergent communication proposes a paradigm shift by advocating the development of communication protocols not only scalable and adaptable to a broad spectrum of tasks, but also bear a resemblance to human linguistic structures, thereby enhancing interpretability.

Our goal is to study experimentally how agents' image encoders and spatial information in image features affect agents' performance and the compositionality of emergent language in a given task.

2 Related Works

The field of emergent communication in multi-agent systems is a research area seeking to understand and develop methods by which autonomous agents can establish and evolve their own protocols for interaction. This research is driven

*This research was funded by the TRAIL/ARIAC (trail.ac).

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

by the premise that agents, when placed in a shared environment with common goals, can develop unique ways of communicating that are not pre-programmed by humans. Recent studies reveal the adaptability of emergent communication strategies across different contexts, showing their potential to enhance collaborative performance in multi-agent systems [1, 2, 3, 4, 5].

However, the literature reveals a gap in establishing a definitive set of tasks or environments where the advantages of emergent communication protocols decisively outperform traditional methods, such as handcrafted or learned continuous communication protocols [1, 5, 6, 7]. The comparisons place these traditional methods as benchmarks, highlighting a need for more nuanced evaluations that demonstrate the unique benefits of emergent language-based communication in real-world scenarios. These investigations tend to center on improving multi-agent performance, with additional considerations given to the robustness of emergent communication systems [3, 4, 5].

This suggests a critical juncture in the study of emergent communication: the need for systematic and comparative research that identifies the specific conditions and mechanisms facilitating the superior performance of emergent languages. Such an inquiry would also advance our theoretical understanding of how autonomous agents can develop complex, adaptive, and efficient communication systems from scratch.

In this paper, we present a study of the influence of image encoders and spatial information contained in image features on the performance of agents in the context of emergent communication within multi-agent systems. Through experimentation, we investigate how variations in image representation affect both agent accuracy in the Lewis Game [8] and the compositionality of the emergent language. Our experiments show that the choice of image encoder and the processing applied to image features affect the accuracy of the agents and the compositionality of the language. We also observe that the accuracy and the compositionality are not necessarily correlated.

3 Experimental Setup

Our setup explores how agents' performance and language compositionality are affected by agents' image encoders and spatial information in image features.

3.1 Task: Lewis Game

The Lewis Game [8] is a referential game involving two players: the *Speaker* and the *Listener*. It unfolds through three stages: Firstly, the *Speaker* is presented with a target image x and generates a message m. Secondly, the *Listener* receives m and selects an image $\hat{x}$ from a set of candidates $\mathcal{C}$ including x. Finally, both players receive a reward based on whether $\hat{x}$ matches x. The parameterization of the *Speaker* and *Listener* is denoted by sets of parameters θ and ϕ, respectively. The message m is a sequence of T words from a vocabulary $\mathcal{V}$. For clarity, we simplify notation by omitting certain dependencies and variables when no ambiguity arises.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

3.2 Experiments

Our agents use the *Speaker* and *Listener* architectures from [9], as well as their loss functions. The *Speaker* is a neural network, composed of several modules depicted in Figure 1, tasked with generating a message m for an input image x. The *Listener*, depicted in Figure 2, receives the *Speaker*'s message m and a set of image candidates $\mathcal{C}$, including the target image x. It computes the probability of each candidate being the target image. The training process involves a population of agents [9] since it has been shown that considering a simple pair of agents may lead to overfitting and extreme co-adaptation [10]. However, these negative results can be countered by sampling agents within a population [11].

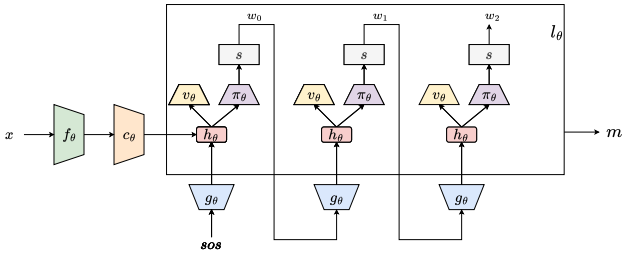

Fig. 1: *Speaker*'s architecture to generate message m for target image x [9].

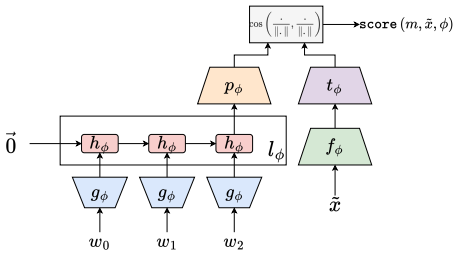

Fig. 2: *Listener*'s architecture evaluating if message m describes image $\tilde{x}$ [9].

To enhance coherence in message generation, at each step, five *Speakers* are uniformly sampled to be trained with imitation learning [9], one of them playing the role of teacher and the others the role of students. Then, students learn from the teacher's policy via cross-entropy loss.

In our experiments, we use 20 candidates per round of the Lewis Game and a population size of 10 *Speakers* and 10 *Listeners*. The messages have a length of $T = 10$ tokens sampled from a set of vocabulary containing 20 tokens. The images are from a Multi-Object Positional Relationships Dataset (MOPRD), tailored for training agents in communicating object relationships within the Lewis Game [12]. Each image is characterized by a tuple (*shape*,

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

shape, *relationship*) for a total of 100 combinations. 20 of them are reserved for the test set, which comprises 2000 images (100 for each combination).

To validate our hypotheses, we employ two distinct ResNets [13], pretrained on ImageNet [14], as image encoders: ResNet-18 and ResNet-50. Specifically, we extract image features from their 4th layer. For each one, we experiment with three methods on the image features: 1) $\mathcal{F}_1(x) = f_\theta(x) = f_\phi(x)$, applying average pooling, 2) $\mathcal{F}_2(x)$, flattening the features, and 3) $\mathcal{F}_3(x)$, similar to $\mathcal{F}_2$ with a custom initialization of the weights from the layer following f_θ and f_ϕ in the agents' architecture. This initialization involves connecting each feature value f_i for an image to the ith neuron of the next layer (with weight $w_{i,i} = 1, \forall i$) while effectively disconnecting this value from other neurons ($w_{i,j} = 0$ if $i \neq j$).

4 Results

In order to evaluate our models, we use two metrics: topographic similarity (TopSim) and test accuracy.

TopSim [15] is a frequently used metric for assessing the compositionality of emergent languages [16, 17] and evaluating generalization in multi-agent communication systems [18]. It measures the closeness between images and their associated messages using Spearman correlation. However, recent research raises doubts about its adequacy, particularly in complex scenarios with natural images, due to potential factors like agents encoding features not captured by human-labeled attributes or the emergence of synonyms, which may lead to low TopSim values despite effective communication [9]. Following prior works [9, 12], we use the edit-distance in the message space and the cosine distance for the image space. Additionally, we use the number of different attributes in the (*shape*, *shape*, *relationship*) tuple to compute TopSim with another distance metric.

Test accuracy represents how well a *Listener* is able to guess the correct image over several rounds of the Lewis Game. In this paper, each possible (`Speaker`, `Listener`) pair is evaluated over 2000 rounds of the Lewis Game. We ensure that each tuple is not represented by more than one image.

Table 1 shows the results for the accuracy. We observe that using $\mathcal{F}_3(x)$ outperforms the other methods in terms of accuracy (61.88 for ResNet-18 and 60.88 for ResNet-50). For TopSim (Table 2), we observe that, when we use $\mathcal{F}_1(x)$, the TopSim using the edit distance is higher, which means a higher compositionality of the emergent language, while with $\mathcal{F}_2(x)$ and $\mathcal{F}_3(x)$, the cosine distance gives a higher compositionality. This can be explained by the fact that the relative position of the figures in the images is important in determining the target image. However, while flattening ($\mathcal{F}_2(x)$ and $\mathcal{F}_3(x)$) retains the spatial information of the figures, average pooling ($\mathcal{F}_1(x)$) loses some of this information. Since these features contain more information with flattening, agents can make better use of it to enhance compositionality in the emerging language.

Based on these observations, we also conclude that a higher accuracy does not imply a better compositionality of the emergent language. Indeed, $\mathcal{F}_3(x)$ gives better accuracy for both ResNets. However, compositionality is better with

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

$\mathcal{F}_3(x)$ for ResNet-18, and with $\mathcal{F}_1(x)$ for ResNet-50.

	$\mathcal{F}_1(x)$	$\mathcal{F}_2(x)$	$\mathcal{F}_3(x)$
ResNet-18	19.49 ± 0.09	33.91 ± 0.51	$\mathbf{61.88 \pm 1.53}$
ResNet-50	53.57 ± 0.52	30.53 ± 0.52	$\mathbf{60.88 \pm 2.09}$

Table 1: Accuracies of the experiments. Bold values show the preprocessing method providing the best accuracy for each ResNet.

		$\mathcal{F}_1(x)$	$\mathcal{F}_2(x)$	$\mathcal{F}_3(x)$
ResNet-18	Cos. TopSim	14.78 ± 0.10	$\mathbf{23.51 \pm 0.48}$	$\mathbf{33.63 \pm 0.93}$
	Edit TopSim	$\mathbf{22.74 \pm 0.18}$	13.68 ± 0.28	25.79 ± 1.83
ResNet-50	Cos. TopSim	27.71 ± 0.33	$\mathbf{27.62 \pm 0.47}$	$\mathbf{25.65 \pm 1.13}$
	Edit TopSim	$\mathbf{29.99 \pm 0.35}$	4.97 ± 0.48	9.99 ± 1.32

Table 2: TopSim values for the experiments. Bold values show which considered distance gets the best compositionality for each ResNet.

5 Discussion and Conclusion

In this paper, we propose an experimental approach to study the influence of image encoders and the spatial information contained in image features on 1) the performance of agents in the Lewis Game and 2) the compositionality of the emergent language. To this end, we have used a learning phase and architectures from the literature, as well as a dataset allowing easy measurement of topographic similarity, which is generally used as an indicator of language compositionality. The agents were trained with two different image encoders: ResNet-18 and ResNet-50. We used three different processing on the image features: 1) average pooling, 2) flattening, and 3) flattening with a custom initialization of the next linear layer.

Our results show that the choice of image encoder and the processing applied to image features have an impact on agent performance (up to tripling accuracy on ResNet-18). However, higher accuracy does not necessarily imply better language compositionality (custom initialization gives better accuracy on both image encoders, but does not provide the best TopSim for ResNet-50). Experiments also show that the choice of processing applied to the image features influences the distance measure to be used on the image space to calculate the TopSim. This observation can be explained by the way image information is stored in the features after these processing steps. However, those results contradict what has been observed in terms of performance in [19] which uses a different setup. This observation could imply that the choice of the image encoding process is influenced by the setup we use.

In conclusion, our study shows an intricate relationship between initialization methods, network architectures, and emergent communication compo-

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

sitionality. While our experiments highlight the impact of pooling methods on language emergence, future research should explore additional architectural enhancements, such as attention mechanisms or graph-based representations, which may offer alternative approaches to preserving spatial information and enhancing language expressivity. Furthermore, examining the interpretability and generalization capabilities of emergent communication remains an important area for future exploration.

References

[1] Sheng Li et al. Learning Emergent Discrete Message Communication for Cooperative Reinforcement Learning. In *2022 International Conference on Robotics and Automation*.

[2] Elías Masquil et al. *Intrinsically-Motivated Goal-Conditioned Reinforcement Learning in Multi-Agent Environments*. 2022.

[3] Kalesha Bullard et al. Quasi-Equivalence Discovery for Zero-Shot Emergent Communication, 2021.

[4] Dylan Cope and Nandi Schoots. Learning to Communicate with Strangers via Channel Randomisation Methods, 2021.

[5] Yuqi Wang et al. *Emergence of Machine Language: Towards Symbolic Intelligence with Neural Networks*. 2022.

[6] Shubham Gupta et al. Networked Multi-Agent Reinforcement Learning with Emergent Communication. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 2020.

[7] Siqi Chen et al. Deep reinforcement learning with emergent communication for coalitional negotiation games. *Mathematical Biosciences and Engineering*, 2022.

[8] David Kellogg Lewis. *Convention: A Philosophical Study*. Cambridge, MA, USA: Wiley-Blackwell, 1969.

[9] Rahma Chaabouni et al. Emergent Communication at Scale. 2022.

[10] Marc Lanctot et al. A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2017.

[11] Max Jaderberg et al. *Human-level performance in first-person multiplayer games with population-based deep reinforcement learning*. 2018.

[12] Yicheng Feng et al. Learning Multi-Object Positional Relationships via Emergent Communication, 2023.

[13] Kaiming He et al. Deep Residual Learning for Image Recognition, 2015.

[14] Jia Deng et al. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[15] Aaron van den Oord et al. Representation Learning with Contrastive Predictive Coding, 2019.

[16] Angeliki Lazaridou et al. Emergence of Linguistic Communication from Referential Games with Symbolic and Pixel Input. 2018.

[17] Fushan Li and Michael Bowling. Ease-of-Teaching and Language Structure from Emergent Communication. In *Advances in Neural Information Processing Systems*, 2019.

[18] Jerry Fodor and Ernest Lepore. *The Compositionality Papers*. Oxford University Press, 2002.

[19] Daniela Mihai and Jonathon Hare. Avoiding hashing and encouraging visual semantics in referential emergent language games, 2019.