PKU-IAI Technical Report: TR-PKU-IAI-2023-0006

A preliminary study of interpretable emergent language in visual referential games

Yiheng Du
Yuanpei College
Peking University
duyiheng@stu.pku.edu.cn

Yuran Xiang
Yuanpei College
Peking University
2000017477@stu.pku.edu.cn

Haoran Sun
Yuanpei College
Peking University
2000017768@stu.pku.edu.cn

Abstract

Deep neural networks have achieved high accuracy in many computer vision and natural language processing tasks through their complicated structures and operations. But their formulation is numerical and the computation process cannot be interpreted by humans directly. Therefore they are often criticized for their inability to align with human language. In this project, we aim at discovering emergent language and how it evolves in multi-agent communication games. We design a visual referential game combining the ideas of the referential game (Lazaridou et al. [11]) and the sketch drawing game (Qiu et al. [16]). We consider both continuous and discrete communication types and propose a model which can be efficiently optimized by back-propagation. Experiments conducted on MNIST and CIFAR-10 have shown the efficiency and interpretability of this communication, with t-SNE of visual embedding as visualization. Our code is available at `https://github.com/ran1812/pkucore-nonverbal_communication`. Our video is available at `https://disk.pku.edu.cn:443/link/5C38F27C4323A6AE8B28389A5C851212`

1 Introduction

General-purpose AI requires various kinds of communication abilities. For cooperating with other agents, they need to develop communication under possibly strict constraints. Also, the ultimate goal of AI is to assist and cooperate with humans, which induces the issue to properly communicate in its own way, not necessarily in natural language. All these tasks require agents to align with human values. Thus, handling natural-language-based communication is a key step toward the development of AI that can thrive in a world populated by other agents. For this purpose, we need to first model the communication process between agents.

A recent successful model example of language emergence in simple environments is the *multi-agent coordination communication games*, in which agents start as blank states and need to develop a language to communicate and earn payoffs. An important type of communication game is the

referential game (i.e. Lewis signaling game) defined in Lazaridou et al. [11]. In a referential game, the sender is given an image and the receiver is given a bunch of images. And the receiver needs to find the particular image in the same class as the sender's image with the sender's signal.

But as the neural network is a black-box method, the symbols agents developed during communication is hard to align with human and hard to have global semantic meaning. In this project, we try to investigate whether we can make the symbols more interpretable and try to align them with humans.

Our main insight is to make use of different constraints. We consider two different scenarios: continuous communication and discrete communication. Continuous communication occurs when constraints are weak and the agent is able to transmit relatively unlimited information. Specifically, in visual reference games, the agent is able to control every pixel on the canvas. On the other hand, discrete communication simulates a harsh environment where the agent is allowed only to provide a rough description of the full information to be transmitted.

This report is structured as follows: First, we introduce the related paper about the emergent model in section 2. Then we describe our model with higher interpretability in section 3. The experiment results are included in section 4. We conclude our paper and discuss future works in section 5.

2 Related Work

Observation Study Fay et al. [4] explore the role of iconicity in spoken language and argue that the simplification of iconic signs is driven by repeated interactions. Fay et al. [3] also argues that interaction is critical to the creation of shared sign systems through studying the evolution of human communication.

Multi-agent Communication Lazaridou et al. [11] consider a two-agent communication and design a referential game where both sender and receiver are feed-forward networks. They ground agents' communication in human language by integrating supervised classification tasks with referential games. Mordatch and Abbeel [15] consider multi-agent scenario and formalize it as a cooperative partially observable Markov game (Littman [13]). The emergence of an abstract compositional language can be observed during the process of achieving goals.

Communication Type Communication in multi-agent games can be categorized into two types: continuous and discrete [10]. In the continuous scenario, agents send and receive continuous vectors. This allows the whole learning process to be efficiently optimized through back-propagation. However, in the discrete scenario, agents' communication is composed of sequences of symbols or actions. This means back-propagation can't be directly applied. One natural idea is to solve this by reinforcement learning [5, 12, 16], where agents receive different rewards and update their policies based on that. Another possible solution is to approximate discrete representations by continuous one [14, 7]. Besides, some literature focuses on a certain form of communication and uses parameterization tricks to make them continuous. For example, [8, 6] studied the problem of continuously parameterizing strokes. In this project, we consider both continuous and discrete scenarios. In continuous one, we explicitly add reconstruction loss to make the information interpretable. In discrete one, we use a VAE structure to parameterize the strokes, which can be seen as a simple form of [8, 6].

Evaluation Metric Accuracy (success rate) is certainly an important metric but it's not the whole story. A high accuracy does not imply high-quality communication (Lazaridou et al. [12]). Therefore various kinds of metrics have been proposed to evaluate the quality of communication. Lazaridou et al. [11] form clusters by extracting the representations in CNN fully-connected layer. They assess the quality of the clusters by measuring their *purity* which is first proposed by [18]. [7] define *omission score* of a sentence to quantify the change in the target image probability after removing the most important word. To be specific, the omission score of a word is equal to the difference between the target image probability given the original message and the probability given a message with the removed word. Lazaridou et al. [12] use *topographic similarity*, first proposed by Brighton and Kirby [1], to measure the extent of topographic relation between meanings and signals. More recent works Qiu et al. [16] take inspiration from linguistic works and propose evaluation metrics *iconicity*, *symbolicity*, and *semanticity*. Aside from direct evaluation, there is also some work from other perspectives. Evtimova et al. [2] put their focus on the effect of conversation length, attention mechanism, and message dimensionality on communication accuracy.

3 Framework

We formalize the *multi-agent coordination communication games* as a referential game (Lazaridou et al. [11]). The object of this game is an image set $\mathcal{I}$ with label set $\mathcal{L}$ and the message set is $\mathcal{M}$. We only consider two agents: a sender and a receiver. In this game, an image $I \in \mathcal{I}$ is shown to the sender. The sender generates a message $M \in \mathcal{M}$ after seeing I. Then the message is sent to the receiver and she chooses the most possible label $L \in \mathcal{L}$ of I. It's worth mentioning that we increase the difficulty by not only referencing one of the two images but possibly a large bag of candidates. In a mathematical form, the sender and the receiver are two functions as follows.

$$\begin{aligned} \text{sender} \quad & s : \mathcal{I} \longrightarrow \mathcal{M} \\ \text{receiver} \quad & r : \mathcal{M} \longrightarrow \mathcal{L} \end{aligned}$$

3.1 Choice of $\mathcal{I}$, $\mathcal{M}$ and $\mathcal{L}$

Currently, we conduct experiments on both MNIST and CIFAR-10 dataset; $\mathcal{I}$ is any image in the dataset and $\mathcal{L} = \{0, 1, \cdots, 9\}$ represents its category. We'll describe the reason we choose these two datasets in the Experiments section 4.2. We restrict the message set $\mathcal{M}$ to be the image space, representing the canvas that agents can draw on in the visual reference game. It simulates the *Draw Something* game between sender and receiver. We consider two different approaches to generate M, corresponding to the continuous and discrete scenarios.

3.2 Shared Eyes

In our framework design, we impose the assumption that both the sender and receiver share a common structure of "eyes". This is natural since humans also share the construction of their primitive vision system. Here this assumption is represented as a shared convolution network that preprocesses image data: Selected input image for the sender, transmitted visual message, and candidate images for the receiver.

3.3 Continuous Communication

In the continuous scenario, we directly use a fully-connected layer to transform the sender's output to a 2D canvas of raw pixels. As it will be shown in section Sec. 4, applying only *cross-entropy loss* for misprediction is not enough for generating an interpretable message, since the agents are not designed to treat its output as a canvas. Therefore we explicitly add *reconstruction loss* between message M and original image I so that the sender is encouraged to generate a message which is similar to what she sees. Intuitively, by doing so the sender will be penalized for generating images not resembling the training set and the message will become more interpretable. Besides, we can also adjust the relative weight of these two losses to let the sender generate a message with different styles: from more *iconicity* in the early stages to more *efficiency* in the later stages. The whole learning process is efficiently optimized through back-propagation.

3.4 Discrete Communication

In the discrete scenario, inspired by Qiu et al. [16], we assume the process of generating M is stroke by stroke. And the strokes we consider are straight or approximate straight lines. The reason for this assumption is that this process is closer to humans. But meantime the drawing process itself also brings about the problem of indifferentiability.

To solve this problem, we use a VAE structure to parameterize these straight lines, which can be seen as a simple form of Huang et al. [8], Ha and Eck [6]. Specifically, the *Drawer* network is the composition of encoder $\mathcal{E} : \mathcal{I} \mapsto \mathcal{H}$ and decoder $\mathcal{D} : \mathcal{H} \mapsto \mathcal{I}$. We generate the training set $\mathcal{I}$ on the fly as uniformly random straight lines and optimize its $L2$ difference with the generated image. After training, the *Drawer* network is fixed as $\mathcal{D}$ and plugged into the sender network. Reconstruction loss is also added to the sender network.

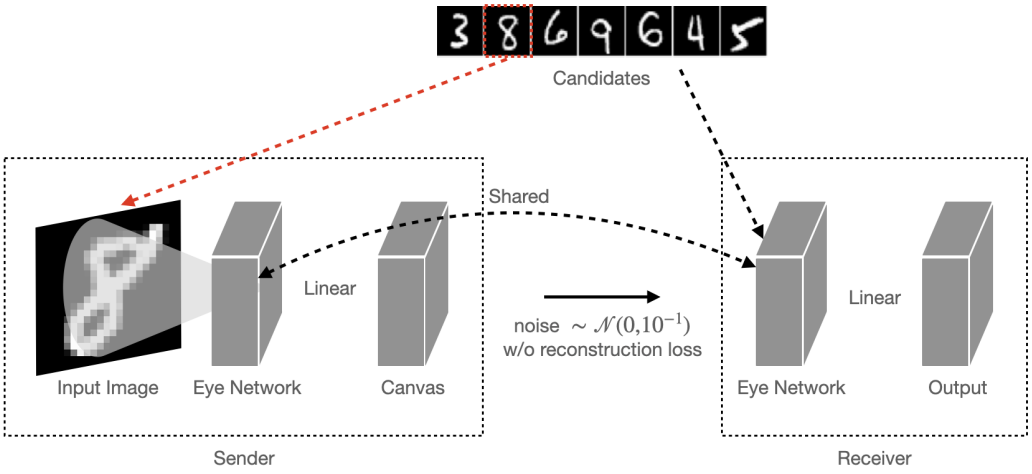

Figure 1: Continuous Communication Framework

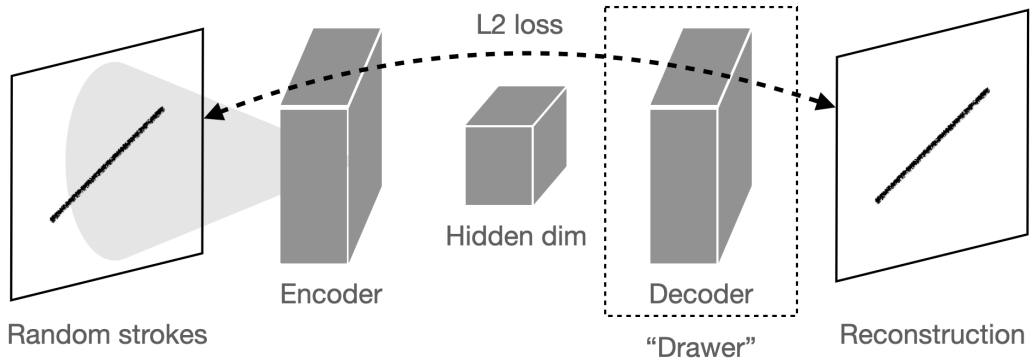

Figure 2: Reparametrized Drawer Network in Discrete Communication

3.5 Evaluation Metric

Inspired by [18], we introduce random noise in the channel between the sender and the receiver. As mentioned in [12], a highly accurate communication sender-receiver system could communicate with simple statistics of the input image. In [18] the authors avoided this problem by randomly dropping words, which can be viewed as a certain degree of regularization. Here we use uniform noise in both the training and evaluation periods to ensure that the sender and receiver are not taking such shortcuts.

In addition, we randomly select different categories of data. We exact their original images and messages M generated by the sender in a different model. We then adopt t-SNE to map them into 2D space and present the visualization. More details and results are shown in Sec. 4.3.A

4 Experiments

4.1 Experiment Setup

EGG Toolkit EGG Toolkit is a toolkit that greatly simplifies the implementation of emergent-language communication games. [9] It includes primitives for implementing single-symbol or variable-length communication and training with optimization of the communication channel through REINFORCE or Gumbel-Softmax relaxation via a common interface.

Specifically, we use the SymbolGameGS class as the game module for our aggregation of the sender and receiver structure, the build_optimizer class for our optimizer module during training, and the Trainer module which implements the training loop for training our model conveniently. This toolkit is effective as it simplifies our code and reduces our workload.

Parameters Details All the experiments are finished on Boya NO.1 with NVIDIA RTX 3090. For both $\mathcal{M}$, we trained it for 50 epochs using Adam optimizer with learning rate 1e-4 and batch size 4096.

Game Details We choose the referential game for the test. Specifically, in each example, we randomly choose a picture as the sender's input, and the receiver need to choose the image with the same label from 10 examples in different classes. The receiver is penalized by cross-entropy loss between its prediction and the true label.

Model Details The decoding module for the sender and receiver is defined as a 2-layer convolutional network with 64 hidden dimensions in the end. The encoding module for continuous $\mathcal{M}$ is a 2-layer transposed convolutional network. And the encoding module for discrete $\mathcal{M}$ is the VAE we discussed above with a 4-dimension hidden space, as we can determine a line on 2D-images with 4 parameters. (start x, start y, end x, end y) To get more general results in reconstruction, we add noise $n \sim \mathcal{N}(0, 0.1)$ on it.

4.2 Results

We choose the MNIST and CIFAR-10 for experiments and get some results. As a dataset of a handwritten numeral, MNIST has more texton information with fixed simple structure units in images, which may have more similarity to the strokes we used. As a dataset of universal objects with color image data, CIFAR-10 has more texture information with noise in the real world, which is important for us to detect the generalization of strokes.

The experimental results are shown below. We use accuracy to evaluate the referential game, and we show the reconstructed image to evaluate the quality of the message. The results are shown as follows:

Datasets	Continuous (w/o reconstruction)	Continuous (w/ reconstruction)	Discrete (strokes)
MNIST	0.9966	0.9989	0.9809
CIFAR-10	0.9980	0.9990	0.8200

Table 1: Test accuracy for two types of communication on MNIST

For both of the games in two datasets, the accuracy shows that even if we constrain the type of information, we can still get success in referential games with both stroke communication and image communication.

Continuous Communication We trained our model both with and without reconstruction loss. When the transmitted message is not constrained, the sender converged to output non-interpretable messages. These images, however, still exhibit distinct shapes on different input categories, which have no resemblance to images in the dataset as expected.

When reconstruction loss is explicitly introduced, the sender actually outputs images that are visually similar to those in the dataset. But as the sender is only asked to transmit category information, it converged to output the common visual features of images of the same category. This type of result in two datasets with its comparison to no reconstruction error is shown below.

Discrete Communication For unconstrained discrete communication, we found that the generated strokes have more distinct features compared to continuous communication. We attribute this phenomenon to the relatively constrained message space that the sender can make use of since in the continuous scenario it can fully control every pixel of the message and here it is restricted to a fixed stroke drawer network. The sender has to find a way to better separate different categories. Both the results of the stroke in two different datasets are shown below.

Constrained with reconstruction loss, the sender transmits images that resemble those in the dataset, just as in the continuous scenario.

4.3 Visualizing the Message

In the previous section, we see that our model achieves a very high communication accuracy. We present the visualization of the message M generated by our model. We use the same 10×10 data as in Sec. 4.2. Since all messages are in a high dimensional space, we first map them to a 2D space

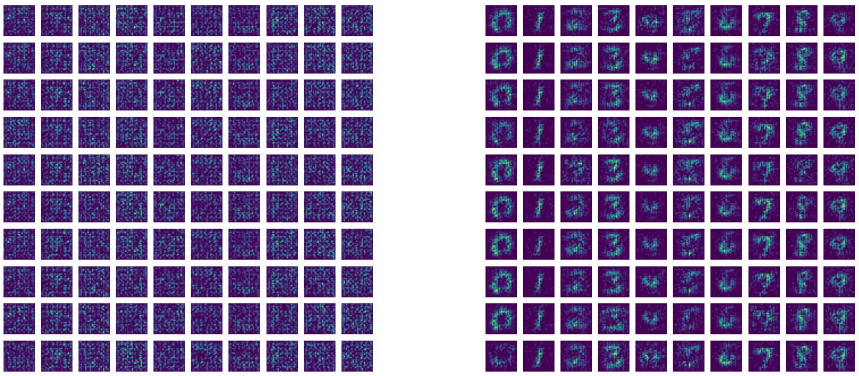

Figure 3: Generated images without reconstruction(left) and with reconstruction(right) for different classes of data in MNIST, each column is a class for numbers from 0 to 9 (from left to right), and each class has 10 examples.

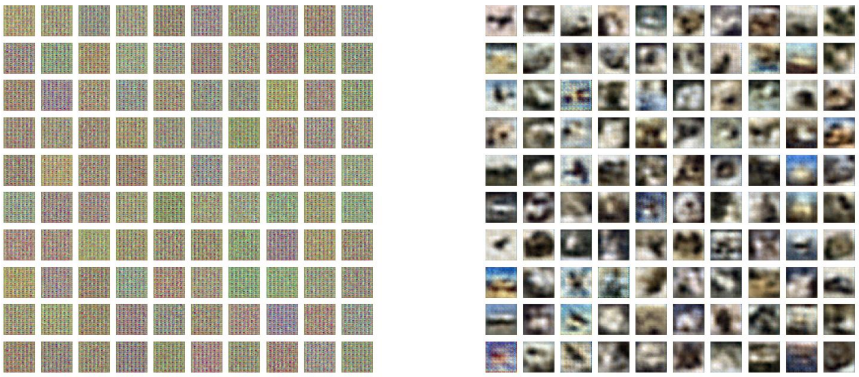

Figure 4: Generated images without reconstruction(left) and with reconstruction(right) for different classes of data in CIFAR-10, each class has 10 examples.

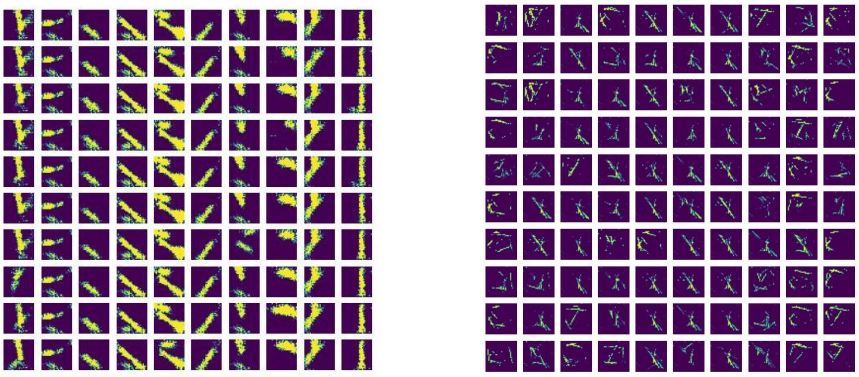

Figure 5: Generated strokes for MNIST(left) and CIFAR-10(right) for different classes of data, each class has 10 examples.

via t-SNE (Van der Maaten and Hinton [17]). The results are shown in Fig. 6 and Fig. 7 for MNIST and CIFAR-10, respectively. Each figure contains four visualization results, original images, the message from continuous communication trained without reconstruction error, the message from continuous communication trained with reconstruction error, and the message (strokes) from discrete communication.

The results of MNIST show that different categories form natural clusters, indicating our three different models can all generate messages with good semantic information. This is also evident from the generated images in Sec. 4.2. However, in CIFAR-10, there is no significant difference between the original images and messages generated by our three models. We think the reason is that we directly reconstruct the original image in our process. But the pictures in CIFAR-10 are not necessarily similar, even if they are in the same category. So this kind of reconstruction may be influenced by the background of the pictures and not capture the most important features of each category. One possible effective way is to first extract the sketch of image (Qiu et al. [16]), and reconstruct the sketch.

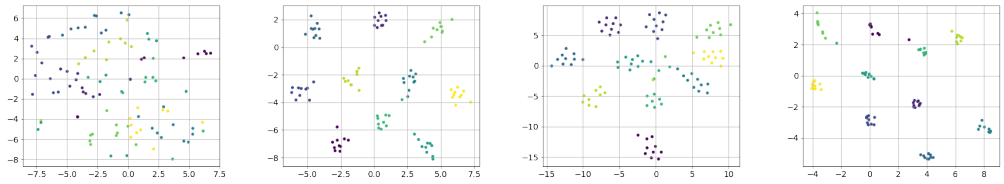

Figure 6: Clustering results for MNIST. From left to right, t-SNE of visual embedding of the original images, the message from continuous communication trained without reconstruction error, the message from continuous communication trained with reconstruction error, and message (strokes) from discrete communication.

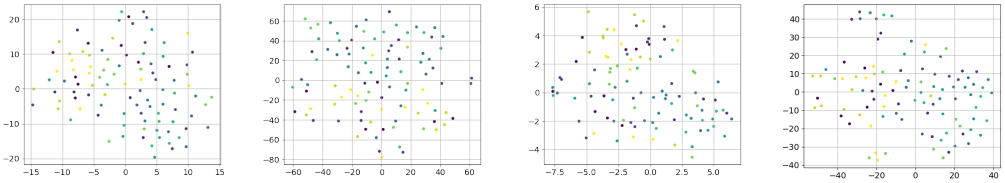

Figure 7: Clustering results for CIFAR-10. From left to right, t-SNE of visual embedding of the original images, the message from continuous communication trained without reconstruction error, the message from continuous communication trained with reconstruction error, and message (strokes) from discrete communication.

5 Conclusion

In this project, we explore the emergent language in multi-agent communication. We formalize it as a referential game and consider both continuous and discrete communications. Experiments on MNIST and CIFAR-10 datasets show that high communication accuracy is achieved. When reconstruction loss is introduced, the sender's outcome (message) exhibits visual interpretability.

5.1 Future Work

We made several simple but crucial assumptions that greatly simplify our experiment but in the meantime introduce a gap between our framework and the real model for humans.

The most basic assumption we made is to represent the alignment between AI and human value as visual similarity, specifically the reconstruction loss. This oversimplified setting allows us to investigate the significance of value alignment in a simple manner, but in the real-world humans have sophisticated value systems that cannot be expressed as a visual similarity.

Another direction that could be further investigated is evaluation metrics. Here we evaluate the emergence of language as the accuracy of communication in the noisy reference game. One could introduce more techniques that investigate the robustness of the communication, say random masking and transformations. Our visualization results via t-SNE on CIFAR-10 haven't shown the clustering effect although it still reached a high communication accuracy. One possible improvement is that

the sender, after getting the picture, first extracts the sketch in the picture and then reconstructs it based on the sketch. This allows the message to capture more key information that can be used to distinguish between different categories.

5.2 Contribution

Haoran Sun contributed to the initial writing of our project report, visualization results, and analysis. Yuran Xiang contributed to improving the drawer network, adapting our model to the CIFAR-10 dataset, and experimenting with various hyperparameters to finetune our model. Yiheng Du contributed to the initial codebase implementation on the MNIST dataset and the revision of our project report.

References

[1] Henry Brighton and Simon Kirby. Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial life*, 12(2):229–242, 2006. 2

[2] Katrina Evtimova, Andrew Drozdov, Douwe Kiela, and Kyunghyun Cho. Emergent communication in a multi-modal, multi-step referential game. *arXiv preprint arXiv:1705.10369*, 2017. 2

[3] Nicolas Fay, Simon Garrod, Leo Roberts, and Nik Swoboda. The interactive evolution of human communication systems. *Cognitive science*, 34(3):351–386, 2010. 2

[4] Nicolas Fay, Mark Ellison, and Simon Garrod. Iconicity: From sign to system in human communication and language. *Pragmatics & Cognition*, 22(2):244–263, 2014. 2

[5] Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/c7635bfd99248a2cdef8249ef7bfbef4-Paper.pdf. 2

[6] David Ha and Douglas Eck. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017. 2, 3

[7] Serhii Havrylov and Ivan Titov. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *Advances in neural information processing systems*, 30, 2017. 2

[8] Zhewei Huang, Wen Heng, and Shuchang Zhou. Learning to paint with model-based deep reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8709–8718, 2019. 2, 3

[9] Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. Egg: a toolkit for research on emergence of language in games. *arXiv preprint arXiv:1907.00852*, 2019. 4

[10] Angeliki Lazaridou and Marco Baroni. Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*, 2020. 2

[11] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language. *arXiv preprint arXiv:1612.07182*, 2016. 1, 2, 3

[12] Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. Emergence of linguistic communication from referential games with symbolic and pixel input. *arXiv preprint arXiv:1804.03984*, 2018. 2, 4

[13] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994. 2

[14] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016. 2

[15] Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[16] Shuwen Qiu, Sirui Xie, Lifeng Fan, Tao Gao, Song-Chun Zhu, and Yixin Zhu. Emergent graphical conventions in a visual communication game. *arXiv preprint arXiv:2111.14210*, 2021. 1, 2, 3, 7

[17] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7

[18] Ying Zhao and George Karypis. Criterion functions for document clustering: Experiments and analysis. 2001. 2, 4