Variational Context Language Encoding as Multi-agent Communication Game

Steven Weikai Lu
University of Alberta
Edmonton, Alberta
T6G 2R3
weikai@ualberta.ca

Abstract

In traditional referential games based on multi-agent system, agents need to communicate via a discrete channel for some cooperative task. Typically, the speaker observe an object then describe to the listener. The listener demonstrate their grounding by pointing out the object that the speaker was referring to. In this project, we extend the granularity of the framework by adopting more complicated generative tasks. We show our perspective of looking at auto-encoding as a multi-player game, and demonstrate the validity of this theory with contextual image reconstruction.

1 Introduction

Emergent communication has been a long lasting topic in the field of natural language processing and linguistics. In the context of multi-agent systems where agents need to cooperate with partial observation and communicate information to solve specific tasks. Sign of natural language have been shown by many researches in this kind of communication protocol, indicating that it could be close to human natural language. Most of the literature in the area follow the setup of referential game, where in the simplest case of only two agents, one act as speaker and one act as listener. The game is characterized by a set of possible messages defined by a fixed-size vocabulary and possibly adaptive length of sentence, together with a set of candidate objects, and possibly some context. The speaker would choose from possible messages to describe the object of interest to the listener, the listener would need to point out the object of interest based on the message(s) from the speaker. If the listener can point out which object the speaker is referring to, we can argue there are positive signalling between the agents. In this project, we consider an variational context language encoding as communication game, where agents need to process and deliver a high-dimensional dense content via combinations of language symbols. Specifically, we present a framework in which we formulate a variation-encoding problem: image reconstruction as a multi-agent learning problem. In the experiment, we use muiti-agent framework to demonstrate the validity of our understanding. To our knowledge, no attempt in the field have been made from this perspective.

2 Related Work

The idea of emergent language stems from the functionalism and formalism in linguistics(MacWhinney, 1999). In the context of a communication based self-organized cooperative system, pressures from the general recognition task would lead to emergence of a potential formal structure of language. Majority of the current literatures base the study on the framework of referential game, where typically the speaker describes an object from a candidate list of objects given it's view(e.g. image), the listener receives messages from speaker via a discrete channel with symbols, then choose the object referred by speaker from possibly different view of an object candidate list(e.g. image or text description). Recent studies formulate this game as a reinforcement learning problem(Evtimova et al., 2017)(Li and Bowling, 2019) by situating the agents(e.g. speaker and listener) in a multi-agent environment. Variational context encoding models(Kingma and Welling, 2013)(Goodfellow et al., 2014) can be interpreted as a special case where only one speaker and one listener is presented with a cooperative or adversarial task. We aim to leverage the connection between language communication game in multi-agent system and discrete neural representa-

tion.

2.1 Referential Games

In the set up of referential games, (Evtimova et al., 2017) demonstrated that a multi-modal(speaker and listener have different perceptual information) and multi-step(adaptive length dialog conversation) setup is feasible. They demonstrate the properties of the emergent language by looking at the interaction pattern between the speaker and listener. (Li and Bowling, 2019) illustrated the concept of "ease-of-teaching" and showed that a more structured language with composibility is indeed easier to teach. (Lazaridou et al., 2018) scaled this setting up by adopting state-of-art deep learning model and reinforcement learning framework, showing that emergent language is more likely to be structured when the agent perceive the world in a structured manner. However, to our knowledge, no literature have addressed the relationship between emergent language structure and task complexity. Intuitively, a more complex task would require and encourage higher efficiency in communication protocol, which we seek to demonstrate.

2.2 Measuring Language Structure

Attempts have been made to demonstrate the existence of such evolutionary language: One naive approach is to simply conduct an ablation study to compare the rewards or degree of task completion with and without the communication channel, as many of language researches showed that cooperation may represent an important prerequisite for the evolution of language[]. Qualitative analysis of messages given states is another common approach, which look at the message generated given an input or task. Some papers quantify language structure by measuring alignment between an agent's message and its corresponding action. In some more recent studies, (Li and Bowling, 2019) exhaustively enumerate all referential target with corresponding deterministic messages, and measure the topological similarity between the two sets. (Lowe et al., 2019) considered causal influence of communication by also exhaustively measuring the mutual information between an agent's message and another agent's decision.

2.3 Unsupervised Learning

Due to the difficulties in data acquisition for many tasks, AI researchers have been emphasizing the importance of unsupervised learning. One interesting prospect of multi-agent communication game is its potential as a building block for unsupervised language learning(Lazaridou et al., 2018). Unlike supervised learning where agents passively learn language statistical associations via examples, agents in a communication game can learn or define its own symbolic language space via actively using this language to complete a certain task. As we want to extend the framework of communication games beyond referential games, more complicated task beyond this "multiple choice" setup is desired, which means the listener will now need to be somehow generative. Generative models such as GANs(Goodfellow et al., 2014), VAE(Kingma and Welling, 2013), and auto-regressive models(Van den Oord et al., 2016) have demonstrated their capability of generating realistic structured outputs. We can in fact look at most communication games as a variant of lossy auto-encoder with a discrete bottleneck. Indeed, speaker learned purely from communication game in (Lazaridou et al., 2018) have showed features somewhat similar to a trained ResNet(He et al., 2016).

3 Approach

We explore emergent communication in the context of a multi-player auto-encoding game. In the set up, multiple agents, speakers and the only one listeners, need to communicate to deliver a concrete structured message via a discrete symbol channel. The agents share the same goal as to jointly optimize the reconstruction of the original message. Each speaker has it's own corresponding code-book or embedding space shared with the listener. All speakers are considered to be independent with each other since they don't share the same embeddings. Hence, the novel component of our work compared to (van den Oord et al., 2017) is we consider multiple speaker while keeping the number of trainable parameters unchanged. Hence instead of having one powerful speaker to describe the image in the discrete channel, we utilize multiple speakers to describe the image via different language (i.e. embeddings).

3.1 Game Setup

We use image auto-encoding game to demonstrate the effeteness of the communication channel. The setup of the game is as follows. The framework is

constituted of one listener, two speakers and their own embedding spaces respectively. Both of the two speakers are presented the same image, they would need to describe to the listener the image using a fixed length sequence of tokens from their own vocabulary that only shared with the listener. The listener, on the other hand, would need to reconstruct the image presented to the speakers. Episode of the game is limited to one conversation. After each episode of the game, two agents get the same reward, which is the negative of the image reconstruction loss.

3.2 Agent Architecture

The policy of both the speakers and listener are modeled by convolutional neural networks. To leverage the bottleneck between the discrete message from speakers and the listener network, we take the idea from VQ-VAE(van den Oord et al., 2017) to define a trainable latent embedding space shared between each speaker and the global listener $e \in R^{K_n \times D_n}$ where K_n is the number of the embedding vectors of the $n-th$ speaker and D_n is the dimensionality of the embedding vectors(i.e. vocabulary size) of the $n-th$ speaker. The speaker network takes an input image x, then output a encoding $z_s, n(x) R^{K_n \times D_n} \in$. The final symbolic message z_n is generated by nearest neighbour lookup to find the K_n nearest embedding vectors, each symbolized as a one-hot vector of size D_n. Specifically The communication policy is formulated as follows, for the $k-th$ message symbol from the $n-th$ speaker:

$$\pi(z_{k,n}|x) = argmin_j \left\| \left(z_{s,n}(x) - e_j \right) \right\|_2^2 \quad (1)$$

The listener is modeled by a auto-regressive convolutional neural network. The listener will first obtain the latent embedding vectors corresponding to the messages from each speakers from corresponding embedding space shared between each speaker and the listener and use them as network input. One can notice now our set up is almost mathematically equivalent to VQ-VAE(van den Oord et al., 2017) while only differ in the number of speaker. Note that this set up is different from traditional referential game not only in task complexity, but the speakers and listener share the same contextual information about the world via the latent embedding spaces. We can interpret this latent embedding the concrete language concept jointly learnt and grounded by the two involved agents. And since all speakers need to independently ground the embedding with the same listener, though they are independent structure-wise, they still need to respect each other to leverage a common ground for the listener.

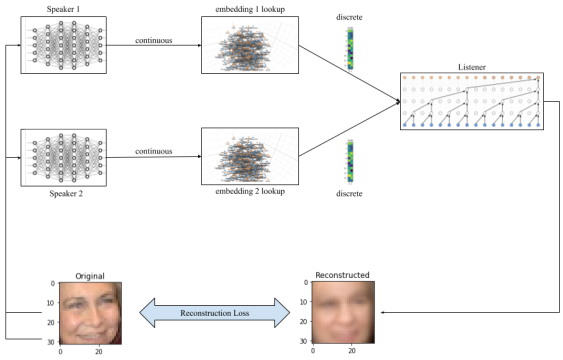

Figure 1: Model architecture.

3.3 Training

We train our model using batch gradient descent with an Adam optimizer. To accommodate for the $argmin$ operation in the speaker and listener, we use a gradient proxy to enable gradient flow through by passing the gradient to the chosen embedding similarly to straight-through estimator as done in VQ-VAE(van den Oord et al., 2017).

The overall loss function is constituted by three main components: reconstruction loss, VQ loss and commitment loss. The reconstruction loss optimizes both speaker and listener network, but leave all latent embedding untouched. This is the main objective of the framework, formulated as:

$$\mathcal{L}_{reconstruction} = log(x|z_1(x), z_2(x)) \quad (2)$$

where $z_n(x)$ is the output from $n-th$ speaker. As we not only want to train the two agent networks, we also want to train our latent embedding space to accommodate more contextual information regarding the task. Hence we include the second loss term embedding loss similar to the VQ quantisation(van den Oord et al., 2017). This loss is essentially a l2 loss between the latent embedding vectors and output from the speaker network. Intuitively, this loss push the embedding vectors

towards the output pattern of the speaker.

$$\mathcal{L}_{VQ} = \sum_{n}^{n \in speakers} \left\| \left(sg[z_{s,n}(x)] - e \right) \right\|_2^2 \quad (3)$$

where $sg[]$ is a stop gradient operator that pass the original argument value during forward pass but produce zero gradient during backpropagation. Hence this term optimize the latent embedding vectors only. In an ideal scenario, the scope of the speaker network output is spanned by the linear space defined by the latent embedding vectors. Then in this case the continuous information from speaker network would be delivered losslessly via the discrete message, and the framework becomes equivalent to an auto-encoder.

Finally, we want the embedding vectors to learn faster than the agent networks. As we mentioned above, embedding vectors represent the shared and grounded contextual knowledge for the two agents, agent networks represent agents' perception and decision making. Intuitively, we want the two agents to learn some common knowledge and contextual world concept before blindly using immature concept to solve tasks. This in fact can be one interpretation of the "posterior collapse" phenomenon. Hence we include a commitment loss as the third loss term:

$$\mathcal{L}_{commitment} = \sum_{n}^{n \in speakers} \beta \left\| \left(z_s, n(x) - sg[e] \right) \right\|_2^2 \quad (4)$$

where again $sg[]$ stands stop for the stopping gradient operator. Hence this term optimize only the speaker and listener networks but not the latent embedding vectors. The total loss function becomes:

$$\mathcal{L} = \mathcal{L}_{reconstruction} + \mathcal{L}_{VQ} + \mathcal{L}_{commitment} \quad (5)$$

3.4 Evaluation

We adopt the following evaluation metrics for measuring the emerged language structure:

- Task completion
- Perplexity of message distribution

For a quantitative task like image reconstruction, the most naive but practical approach is to study to what degree the communication protocol has enabled the task. We will evaluate the task capability of the framework by comparing mean square error between the original work(van den Oord et al., 2017) and our multi-branch model.

Perplexity of message distribution measures the perplexity of the message distribution conditional on target object. Generally speaking if for a fixed input the message distribution has low entropy, the speaker is using small set of message to describe that input. However, the drawback of this metric is that it does not look at how different the message will be for different input.

4 Experiment

4.1 Set up

In this experiment, we compare our model with VQ-VAE(van den Oord et al., 2017). We use image reconstruction as the task for the two models. We conduct our experiment on CIFAR10 to demonstrate the generalization capability of the framework. We set the number of the embedding vectors K to be 512 and the dimensionality of the embedding vectors D to be 64 as in (van den Oord et al., 2017). The speaker network has 2 convolutional layers with stride 2 and kernel size 4×4, followed by 2 3×3 residual blocks, implemented by ReLU. Total number of hidden units in the speaker network is 256. Similarly, the listener network has two 3×3 residual blocks, followed by two transposed convolutions with stride 2 and kernel size 4×4. For the purpose of image reconstruction, we used PixelCNN(Van den Oord et al., 2016) as a prior over the distribution of message z. Our model is similar to VQ-VAE but we adopted two speakers. During forward pass, output from each speakers will be concatenated into one tensor. We trained the model using Adam(Kingma and Ba, 2014) optimizer for 100k updates with a learning rate of 3e-4.

4.2 Result

Preliminary results show that both VQ-VAE and our framework are able to reconstruct $32 \times 32 \times 3$ colored images with an embedding space of $K = 512$ and $D = 64$. The training process averagely takes 38 minutes for the VE-VAE model and 48 minutes for our two branch model over 100k updates. The final MSE is 0.052 for VQ-VAE and 0.026 for our model. The MSE is close to half of that from VQ-VAE.

4.2 shows the parallel training dynamic comparison. The MSE error dynamic shows that our model consistently outperformed the original VQ-

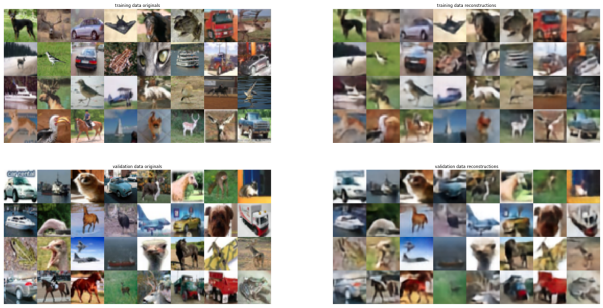

Figure 2: Examples of original image and reconstruction. Left: original batch. Right: reconstructed batch.

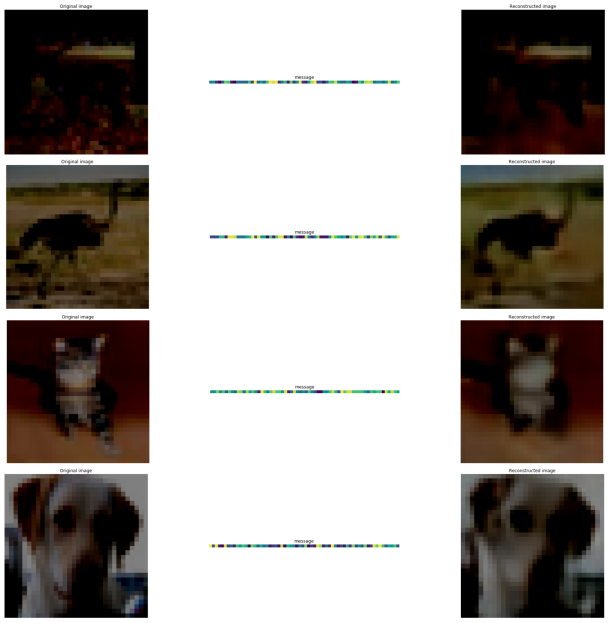

Figure 3: Examples from train and validation set by our model. Left: original image. Middle: visualized corresponding discrete message(z). Right: reconstructed image by our model.

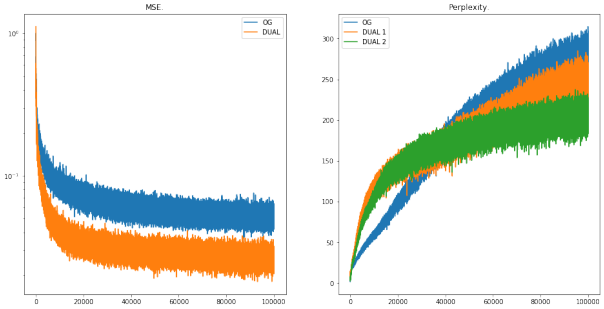

Figure 4: Comparison between original single branch model and our multi-agent model. Left: MSE, OG is the original VQ-VAE, DUAL is our model. Right: Perplexity measures, OG is the original VA-VAE, DUAL 1,2 is speaker 1,2 respectively.

VAE. Our model also demonstrate better asymptotic curve. The perplexity, on the other hand, shows more behind the scene. In the multi-branch model, the perplexity of the two speakers respectively with the listener increase relatively fast at the initial training phase, but soon was surpassed by the original model after around 20k updates. We can interpret this phenomenon by considering the level of grounding between listener and speakers. For our model after the initial training stage(20k updates), both speakers have grounded different view or features description of the images with the listener. As the two speakers learn with listener independently and are only softly constrained by the state of listener, they can potentially build up very different latent embedding spaces. These latent embedding spaces from two speakers is not very likely to be linear combinations of one another, which means there will be some features in the image captured by one speaker, but not the other. However, due to the multi-agent setup, each speaker do not need to encode the entire image. In fact, a desirable scenario would be each speaker describe a partition of features that are disjoint with other speakers, which is possible since image can be described by multiple independent sentence(e.g. "There is a cat on the table", "the background wall is green"). The less overlap of information among the speaker encoding, the more efficient our framework is.

5 Conclusion

In this project, we extend the setup of referential game in multi-agent system. We propose a framework by looking at auto-encoding models as muti-agent system. In this framework, we successfully reconstruct images via a discrete symbolized channel. Our experiment result shows that indeed multiple speaker can capture different features and demonstrate its advantage in terms of task completion.

Acknowledgments

We would like to thank Dr.Lili Mou for providing the project idea, and Fushan Li for high-level advice.

References

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

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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

Fushan Li and Michael Bowling. 2019. Ease-of-teaching and language structure from emergent communication. *arXiv preprint arXiv:1906.02403*.

Ryan Lowe, Jakob Foerster, Y-Lan Boureau, Joelle Pineau, and Yann Dauphin. 2019. On the pitfalls of measuring emergent communication. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 693–701. International Foundation for Autonomous Agents and Multiagent Systems.

Brian MacWhinney. 1999. Emergent language. *Functionalism and Formalism in Linguistics*, 1:361–386.

Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. 2016. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, pages 4790–4798.

Aaron van den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315.

A Appendices

B Supplemental Material