Language Structure in Auto-encoding Two-player Games

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

Abstract

Placeholder

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 shared information to solve specific tasks, communication is necessary. 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, 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. In this project, we consider an auto-encoding communication game, where agents need to process and deliver entangled content (e.g. image, audio) via combinations of language symbols. We show that this framework is similar to variational auto-encoder[]. In the experiment, we used ??? to demonstrate the language structure emerged in such communication game.

2 Related Work

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 percept the world in a structured manner. However, to our knowledge, no literature have addressed the relationship between emergent language structuredness and task complexity. Intuitively, a more complex task would require and encourage higher efficiency in communication protocol. We seek to demonstrate this relationship.

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 measuring the mutual information between an agent's message and

another agent's decision.

2.3 Unsupervised Learning

The significance of emergent language is due to the fact that researchers have been looking at it as a form of supervised learning. One interesting prospect is its potential as a building block for unsupervised language learning(Lazaridou et al., 2018). Unlike supervised learning where agents by 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, the output of the listener(e.g. images, audio) will be structured. Generative models such as GANs[], VAE(Kingma and Welling, 2013), and auto-regressive models[] have demonstrated their capability of generating realistic structured outputs. We can in fact look at most communication games as a variant of lossy autoencoder 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[?].

3 Approach

We explore emergent communication in the context of a two-player auto-encoding game. In the set up, two agents, speaker and listener, need to communicate to deliver a concrete structured message via a discrete symbol channel. Two agents share the same goal as to jointly optimize the reconstruction of the original message.

3.1 Game Setup

We used image auto-encoding game to demonstrate the effeteness of the communication channel. The setup of the game is as follows. Speaker is presented an image, it would need to describe to the listener the image using a fixed length sequence of tokens from a vocabulary. The listener, on the other hand, would need to reconstruct the image presented to the speaker. 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 speaker and listener is modeled by a convolutional neural network.

To bridge the gap between the discrete message from speaker 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 the speaker and listener $e \in R^{K \times D}$ where K is the number of the embedding vectors and D is the dimensionality of the embedding vectors. The speaker network takes an input image x, then output a encoding $z_s(x)R^{K \times D} \in$. Then the symbolic message z is generated by nearest neighbour lookup to find the K nearest embedding vectors, each symbolized as a one-hot vector of size K. Specifically The communication policy is formulated as follows, for the $k-th$ message symbol:

$$\pi(z_k|x) = argmin_j \left\| \left(z_s(x) - e_j \right) \right\|_2^2 \quad (1)$$

The listener is modeled by a decoding convolutional neural network. The listener will first obtain the latent embedding vectors corresponding to the message symbols, then use them as network input. One can notice now our set up is mathematically equivalent to VQ-VAE(van den Oord et al., 2017) as we have mentioned that two-player auto-encoding games are similar to auto-encoders. Note that this set up is different from traditional referential game not only in task complexity, but the speaker and listener share the same contextual information about the world via the latent embedding space. Then we can interpret this latent embedding the concrete language concept jointly learnt and grounded by two agents.

3.3 Training

TODO: Make this more RL-ish

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 policy gradient learning 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:

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

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.

$$\mathcal{L}_{VQ} = \left\| \left(sg[z_s(x)] - e \right) \right\|_2^2 \quad (3)$$

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 autoencoder.

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:

$$\mathcal{L}_commitment = \beta \left\| \left(z_s(x) - sg[e] \right) \right\|_2^2 \quad (4)$$

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

Other than the three main loss items, We adopt the following evaluation metrics for measuring the emerged language structure:

- Perplexity of message distribution
- Causal influence in Communication

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.

We also adopt Causal influence in Communication (CIC) (Lowe et al., 2019) to measure positive listening. Generally CIC look at the marginal effect of a certain symbol in the message on the trajectory. We will exhaustively enumerate the image gradient induced by each symbol, and look to find a generic pattern.

Measuring language structure and composibility has always been a challenging problem and different criteria have different emphasis. To demonstrate the effectiveness and potential power of the learnt language, we also show this practically via transfer learning. After a training session, we reinitialize the listener while freezing the speaker and embedding space. We show that the new listener will quickly catch up with the pre-trained knowledge in the latent embedding vectors and be able to perform the same task.

4 Experiment

4.1 Set up

We use image reconstruction as the task for the two agents. We conduct our experiment on two datasets: CIFAR10[?] and MNIST[?] to demonstrate the generalization capability of the framework by varying length of the message as well as the size of latent embedding space. For CIFAR10, we set the number of the embedding vectors K to be ? and the dimensionality of the embedding vectors D to be ?.

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[?] as a prior over the distribution of message z. We trained the model using Adam[?] optimizer for 100k epochs with a learning rate of 3e-4.

4.2 Result

Preliminary results show that our framework is able to reconstruct $32 \times 32 \times 3$ colored images with an embedding space of $K = 512$ and $D = 64$. Following are some initial reconstruction results.

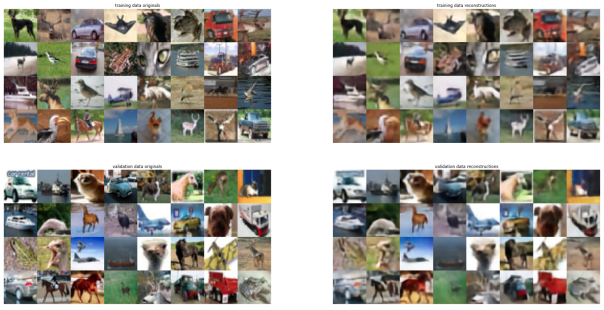

Figure 1: Examples of original image and reconstructed images from train and validation set

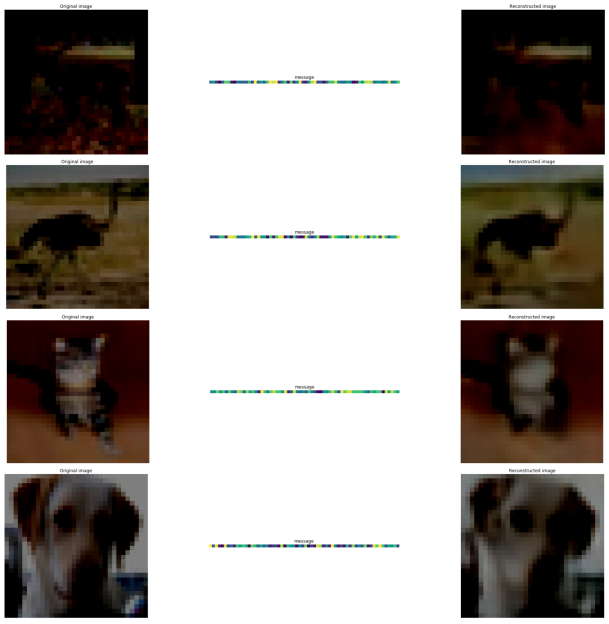

Figure 2: Examples of target images and corresponding messages

[TODO: Add systematic experiments and evaluation] - use image difference

5 Conclusion

Acknowledgments

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*.

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.

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