Emergent Communication with Attention

Ryokan Ri (ryou0634@email.com)
Ryo Ueda (ryoryoueda@is.s.u-tokyo.ac.jp)
Jason Naradowsky (jason.narad@gmail.com)
The University of Tokyo, Japan

Abstract

To develop computational agents that better communicate using their own emergent language, we endow the agents with an ability to focus their attention on particular concepts in the environment. Humans often understand an object or scene as a composite of concepts and those concepts are further mapped onto words. We implement this intuition as cross-modal attention mechanisms in Speaker and Listener agents in a referential game and show attention leads to more compositional and interpretable emergent language. We also demonstrate how attention aids in understanding the learned communication protocol by investigating the attention weights associated with each message symbol and the alignment of attention weights between Speaker and Listener agents. Overall, our results suggest that attention is a promising mechanism for developing more human-like emergent language.

Keywords: emergent communication; attention; language compositionality

Introduction

Language is a defining characteristic of human beings, allowing us to efficiently convey a wide range of ideas. One key aspect of language is compositionality, the ability to represent complex concepts through the combination of atomic units such as morphemes or words. To understand the development of compositionality in human language, a field of research called *emergent communication* (Lazaridou & Baroni, 2020) studies the communication protocols developed by computational agents. These agents are often constructed using artificial neural networks and optimized to solve a task requiring inter-agent communication. During the optimization process, their "language" emerges.

Existing studies have shown that the compositionality of emergent language can be enhanced by various factors such as learning across generations (Li & Bowling, 2019; Ren, Guo, Labeau, Cohen, & Kirby, 2020) or applying noise to the communication channel (Łukasz Kuciński, Korbak, Kołodziej, & Miłoś, 2021). However, the model architecture of the agents has generally been kept with minimal assumptions about the inductive bias. Typical speaker agents encode information into a single fixed-length vector to initialize the hidden state of a RNN decoder and generate symbols (Lazaridou, Peysakhovich, & Baroni, 2017; Mordatch & Abbeel, 2018; Ren et al., 2020). Only a few studies have explored architectural variations (Słowik et al., 2020; Evtimova, Drozdov, Kiela, & Cho, 2018), and there remains much to be discussed about the effects of the inductive bias from different architectures, especially ones that reflect the human cognitive process.

In this study, we explore *the attention mechanism*. The conceptual core of attention is the ability to adaptively focus on relevant information, and attention has been shown to play an important role in human cognition (Rensink, 2000), language development (de Diego-Balaguer, Martinez-Alvarez, & Pons, 2016), and intentional communication (Brinck, 2000). Introducing the notion of attention into emergent communication can enhance its resemblance to human communication and expand the scope of the research field. In this paper, we hypothesize that attention can help agents learn clear associations between subparts of input stimuli and language symbols, resulting in more compositional language.

Another reason to explore the attention mechanism is its interpretability. Emergent language is usually optimized to maximize task rewards and the learned communication protocol often results in counter-intuitive and opaque encoding (Bouchacourt & Baroni, 2018). Several metrics have been proposed to measure specific characteristics of emergent language (Brighton & Kirby, 2006; Lowe, Foerster, Boureau, Pineau, & Dauphin, 2019) but these metrics provide rather a holistic view of emergent language and do not tell us a fine-grained view of what each symbol is meant for or understood as. Attention weights, on the other hand, have been shown to provide insights into the basis of the network's prediction (Bahdanau et al., 2015; Xu et al., 2015; Yang et al., 2016). Incorporating attention in the process of symbol production/comprehension will allow us to inspect the meaning of each symbol in the messages.

In this paper, we test attention agents with the visual referential game (Lewis, 1969; Lazaridou et al., 2017), which involves two agents: *Speaker* and *Listener*. The goal of the game is to convey the types of items in an image that Speaker sees to Listener. To offer extensive empirical results, we experiment with two types of popular network architectures, LSTM (Hochreiter & Schmidhuber, 1997) and Transformer (Vaswani et al., 2017), to implement the agents. We compare the attention agents against their non-attention counterparts to show that adding attention mechanisms to either/both Speaker or/and Listener helps develop a more compositional language. We also examine the attention weights and investigate how they shed light on the learned language.

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

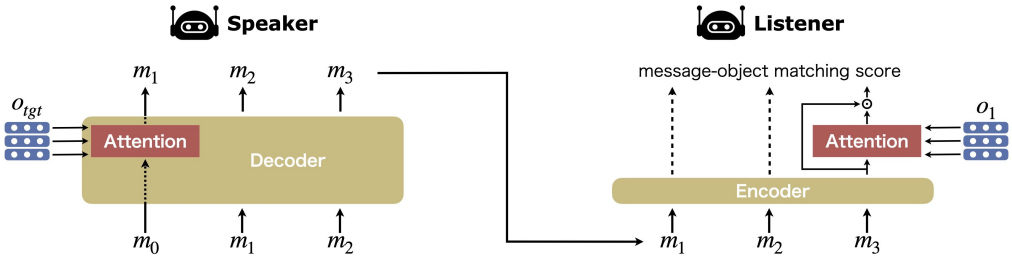


Figure 1: Illustration of the attention agents in the referential game.

Experimental Framework

Referential Game

We study emergent language in the referential game (Lewis, 1969; Lazaridou et al., 2017). The game focuses on a basic feature of language, referring to things. The version of the referential game used in this paper is structured as follows:

1. Speaker is presented with a target object $o_{tgt} \in O$ and generates a message m that consists of a sequence of symbols.
2. Listener receives the message m and a candidate set $C = \{o_1, o_2, ..., o_{|C|}\}$ including the target object o_{tgt} and distractor objects sampled randomly without replacement from O.
3. Listener chooses one object from the candidate set and if it is the target object, the game is considered successful.

The objects can be represented as a set of attributes. The focus here is whether agents can represent the objects in a compositional message based on the attributes.

Agent Architectures

Our goal in this paper is to test the effect of the attention mechanism on emergent language. The attention mechanism in machine learning takes a query vector $\boldsymbol{x}$ and key-value vectors $\{\boldsymbol{y}_1, ..., \boldsymbol{y}_L\}$ as input. The key-value vectors are optionally transformed into separate keys and values. Attention scores $\{s_1, ..., s_L\}$ are calculated as the similarity between the query and keys to produce the attention weights via the softmax function. Finally, the attention weights are used to produce a weighted sum of the value vectors.

A key feature of attention is that it allows the agents to selectively attend to a part of multi-vector inputs. Our hypothesis is that by modeling direct associations between symbols as queries and multi-vector inputs as key-value, agents will tend to assign symbols to meaningful subparts of the inputs instead of opaque and non-compositional information. To test this hypothesis, we design non-attention and attention agents for both Speaker and Listener (Figure 1).

Object Encoder The objects are presented to the agents in the form of a set of real-valued vectors, in our case, patch-wise pretrained CNN feature vectors, as $\{\boldsymbol{o}^1, ..., \boldsymbol{o}^A\}$.

Speaker and Listener agents have their individual object encoders. The input vectors are independently linear-transformed into the size of the agent's hidden size and then successively go through the `gelu` activation (Hendrycks & Gimpel, 2016), which is commonly adapted in Transformer-based neural networks. The non-attention agents average the transformed vectors into a single vector $\hat{\boldsymbol{o}}$ for the subsequent computations, whereas the attention agents leave the vectors intact and will attend to the set of vectors $\{\hat{\boldsymbol{o}}^1, ..., \hat{\boldsymbol{o}}^A\}$.

Speaker Agents Speaker has a message decoder that takes the encoded vector(s) of the target object as input and generates a multi-symbol message $m = (m_1, ..., m_T)$. To provide extensive empirical evidence on the effect of the attention mechanism, we experiment with two common decoder architectures: the **LSTM** decoder from Luong, Pham, and Manning (2015) and the **Transformer** decoder from Vaswani et al. (2017).

At each time step t, the decoders embed a previously generated symbol into a vector $\boldsymbol{m}_{t-1}$ and produce an output hidden vector through three steps: (1) contextualization; (2) attention; (3) post-processing. As the decoders basically follow the original architecture, we only briefly describe each step in the LSTM and Transformer decoder with emphasis on how attention is incorporated. The contextualization step updates the input vector with the information of previous inputs. The LSTM decoder uses a LSTM cell (Hochreiter & Schmidhuber, 1997) and the Transformer decoder uses the self-attention mechanism (Vaswani et al., 2017).

Then, with the contextualized input vector as the query $\boldsymbol{x}_t$ and the object vectors as the key-value vectors $\{\hat{\boldsymbol{o}}^1, ..., \hat{\boldsymbol{o}}^A\}$, the decoders perform attention. The LSTM decoder uses the bilinear attention, where the attention score s_i is computed as $s_t^i = \boldsymbol{x}_t^\top \boldsymbol{W}_b \hat{\boldsymbol{o}}^i$, where $\boldsymbol{W}_b$ is a learnable matrix. The attention vector is calculated as the weighted sum of the original key-value vectors. The Transformer attention first linear-transforms the input vector as $\boldsymbol{q}_t = \boldsymbol{W}_q \boldsymbol{x}_t, \boldsymbol{k}^i = \boldsymbol{W}_k \hat{\boldsymbol{o}}^i, \boldsymbol{v}^i = \boldsymbol{W}_v \hat{\boldsymbol{o}}^i$, where $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v$ are learnable matrices. Then the attention score is calculated using the scaled dot attention: $s_t^i = (\boldsymbol{q}_t^\top \boldsymbol{k}^i)/\sqrt{d}$, where d is the dimension of the query and key vectors. Finally, the attention vector is calculated as the weighted sum of the value vectors $\boldsymbol{v}^i$. The original Transformer also has the multi-head attention mechanism, but in the main experiments, we set the number of the attention heads to one for interpretability and ease of analysis.

In the post-processing step, the original query vector $\boldsymbol{x}_t$ and attended vector $\hat{\boldsymbol{x}}_t$ combine to generate the hidden vector to

predict the next symbol. This integration often includes vector concatenation, addition, and transformation, as detailed in the original papers (Luong et al., 2015; Vaswani et al., 2017).

Non-attention (NoAT) Speaker is a baseline agent that encodes the target object as a single vector $\hat{\boldsymbol{o}}_{tgt}$. In the decoder's source-target attention, this vector is always attended to, and the focus remains fixed throughout message generation.

Attention (AT) Speaker, in contrast, encodes the target object into a set of vectors $\{\hat{\boldsymbol{o}}^1_{tgt}, ..., \hat{\boldsymbol{o}}^A_{tgt}\}$ and the source-target attention dynamically changes its focus at each time step.

Our agent design aims to provide a fair comparison between non-attention and attention agents, by ensuring they possess the same modules and number of parameters. The only distinction is the latter's ability to adjust its focus dynamically when generating each symbol.

Listener Agents Listener tries to predict the target object from a set of candidate objects $C = \{o_1, o_2, ..., o_{|C|}\}$ given the speaker message m by computing message-object matching scores $\{s_1, ..., s_{|C|}\}$ and choosing the object with the maximum score. Listener first encodes the objects using the object encoder and also encodes each symbol in the message into vectors $\{\boldsymbol{m}^1, ..., \boldsymbol{m}^T\}$ using a message encoder, for which the LSTM-based agent uses the bidirectional LSTM and the Transformer-based agent uses the Transformer encoder.

Non-attention (NoAT) Listener encodes each candidate object into a single vector $\hat{\boldsymbol{o}}_i$. The agent also averages the encoded symbol vectors into a single vector $\boldsymbol{m} = \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{m}^i$. The message-object matching score is computed by taking the dot product of the object and message vector $s_i = \hat{\boldsymbol{o}}_i^\top \boldsymbol{m}$.

Attention (AT) Listener encodes each object into a set of attribute vectors $\{\hat{\boldsymbol{o}}^1_i, ..., \hat{\boldsymbol{o}}^A_i\}$ and use the encoded symbol vectors as it is. With each encoded symbol vector $\boldsymbol{m}^t$ as query, the model produces an attention vector $\hat{\boldsymbol{m}}^t_i$ with the object attribute vectors as key-value using the dot-product attention. Intuitively, the attention vector $\hat{\boldsymbol{m}}^t_i$ is supposed to represent the attributes of the object o_i relevant to the symbol m^t. Then, the symbol-object matching scores are computed by taking the dot product between the attention vector and each symbol vector: $s^t_i = \hat{\boldsymbol{m}}^{t\top}_i \boldsymbol{m}^t$. Finally, the symbol-object matching scores are averaged to produce the message-object matching score: $s^i = \frac{1}{T}\sum_t s^t_i$.

Optimization

The parameters of Speaker θ_S and Listener θ_L are both optimized toward the task success.

Speaker is trained with the REINFORCE algorithm (Williams, 1992). The message decoder produces the probability distribution of which symbol to generate $\pi_{\theta_S}(\cdot|t)$ at each time step t. At training time, message symbols are randomly sampled according to the predicted probabilities and the loss function for the Speaker message policy is $\mathcal{L}_\pi(\theta_S) = \sum_t r \log(\pi_{\theta_S}(m_t|t))$ where m_t denotes the t-th symbol in the message. The reward r is set to 1 if Listener selects the correct target object from the candidate set and 0 otherwise.

As an auxiliary loss function, we employ an entropy regularization loss $L_H(\theta_S) = -\sum_t H\left(\pi_{\theta_S}(\cdot|t)\right)$, where H is the entropy of a probability distribution, to encourage exploration. We also add a KL loss $L_{\mathrm{KL}}(\theta_S) = \sum_t D_{\mathrm{KL}}\left(\pi_{\theta_S}(\cdot|t) \| \pi_{\bar{\theta}_S}(\cdot|t)\right)$, where the policy $\pi_{\bar{\theta}_S}$ is obtained by taking an exponential moving average of the weights of θ_S over training, to stabilize the training (Chaabouni et al., 2022). In summary, the final speaker loss is $\mathcal{L}(\theta_S) = \mathcal{L}_\pi(\theta_S) + \alpha\mathcal{L}_H(\theta_S) + \beta\mathcal{L}_{KL}(\theta_S)$, where α and β are hyperparameters.

Listener is trained with a multi-class classification loss. The message-object matching scores are converted through the softmax operation to $p_{\theta_L}(o_i|C)$, the probability of choosing the object o_i as the target from the candidate set C. Then Listener is trained to maximize the probability of predicting the target object by minimizing the loss function $\mathcal{L}(\theta_L) = -\log p_{\theta_L}(o_{tgt}|C)$.

Evaluation Metrics

We quantitatively evaluate emergent languages from how well the language can be used to solve the task and how well the language exhibits compositionality.

Training accuracy (TrainAcc) measures the task performance with objects seen during training. This indicates how the agent architectures are simply effective to solve the referential game.

Generalization accuracy (GenAcc) measures the task performance with objects unseen during training. We split the distinct object types in the game into train and evaluation sets and the generalization accuracy is computed with the evaluation set. As each object can be represented as a combination of attribute values, what we expect for the agents is to learn to combine symbols denoting each attribute value in a systematic way so that the language can express unseen combinations of known attribute values.

Topographic similarity (TopSim), also known as Representational Similarity Analysis (Kriegeskorte, Mur, & Bandettini, 2008), is one of the most commonly used metrics to assess the compositionality of emergent language (Brighton & Kirby, 2006; Lazaridou, Hermann, Tuyls, & Clark, 2018; Ren et al., 2020; Chaabouni, Kharitonov, Bouchacourt, Dupoux, & Baroni, 2020). Intuitively, TopSim checks if similar objects have similar messages assigned. To compute TopSim, we enumerate all the object-message pairs $\{(o^1, m^1), ..., (o^{|O|}, m^{|O|})\}$ with a trained Speaker and define a distance function for objects $d_O(o^i, o^j)$ and messages $d_{\mathcal{M}}(m^i, m^j)$. Then we compute Spearman's correlation between pairwise distances in the object and message space. For the distance function for objects $d_O(o^i, o^j)$, we use the cosine distance of the binary attribute value vectors and for the distance function of messages $d_{\mathcal{M}}(m^i, m^j)$ the edit distance of message symbols.

(a) TrainAcc (b) GenAcc (c) TopSim

Figure 2: The results of the referential game. The color of the boxes indicates the base architecture of the agents (LSTM or Transformer) and the x-axis labels indicates whether Speaker and Listener use attention.

Experimental Setup

Fashion-MNIST Game

Attention has been shown to be able to associate a symbol and a relevant region of an image to solve the task (Xu et al., 2015; Yang et al., 2016). To develop human-like emergent languages, one important question is whether the attention agents develop an interpretable language such that we can understand the meaning of each symbol by inspecting the attended region. For this purpose, we design a multi-item image referential game using the Fashion-MNIST dataset[1] (Xiao, Rasul, & Vollgraf, 2017).

The game objective is to communicate a type of object between Speaker and Listener. Each object is defined as a combination of two classes from the Fashion-MNIST dataset (e.g., *T-shirt* and *Sneaker*). As the dataset has 10 classes, there are a total of ${}_{10}C_2 = 45$ object types. These 45 types are randomly split into training and evaluation sets at a ratio of 30/15, and the number of candidates is set to 15.

The FashionMNIST items on the images are spatially disentangled and ideal for evaluating attention mechanisms. While learning to attend to abstract attributes such as color and shape in images can be more challenging, we anticipate that it is achievable with the use of high-quality feature extractors and learning configurations.

Input Representation

Each object is presented to the agents as feature vectors extracted from a pixel image. The image is created by placing on a 224×224 black canvas two item images, each of which is rescaled to the size of 48×48 (Figure 3). The places are randomly sampled so that the items never overlap.

The specific item images and their positions are randomly sampled every time the agents process the objects both during training and evaluation time to avoid degenerated solutions that exploit spurious features of an image (Lazaridou et al., 2018; Bouchacourt & Baroni, 2018). Also, the Speaker and Listener are presented as the target object with two images depicting the same item types, but with different instances and locations, to facilitate learning a robust communication protocol (Rodríguez Luna, Ponti, Hupkes, & Bruni, 2020).

[1]https://github.com/zalandoresearch/fashion-mnist

Each image is encoded into $7 \times 7 \times 768$-dim feature vectors with a pretrained ConvNet[2] (Z. Liu et al., 2022). For non-attention models, the feature vectors are averaged across spatial axes into a single 768-dim feature vector.

Agent Configurations

The vocabulary size of the agents is set to 20 and the message length is 2. A perfectly compositional language would refer to each item in the image with different symbols with a consistent one-to-one mapping. The sizes of embeddings and hidden sizes of the Speaker and Listener are set to 256.

The hyperparameters (Table 1) are tuned for the entropy loss weight α. The reported scores are obtained from the top 10 agents in terms of the generalization score for each setting.

Training Batch size	480
Max Training steps	50K
Evaluation Rounds	15000
Entropy loss weight α	[0.1, 0.01, 0.001]
KL loss weight β	0.1
Learning rate	1e-4

Table 1: Hyperparameters of the experiments.

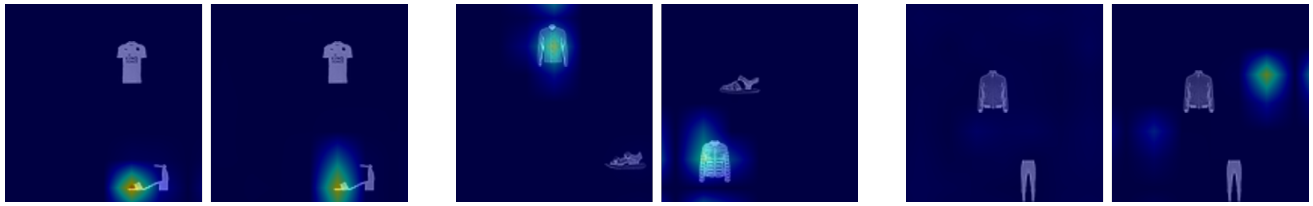
Results

Attention agents find more compositional solutions

We evaluate non-attention agents and attention agents where either/both Speaker and Listener have dynamic attention (Figure 2). We will not discuss the distinctions between LSTM and Transformer as they vary in multiple aspects of their architecture. We focus on the difference between the non-attention and attention agents within each architecture.

We observe a general trend that the attention agents perform better than the non-attention baseline (NoAT-NoAT), which is indicated by the better average scores in task generalization (GenAcc) and compositionality metrics (TopSim). This observation provides evidence in support of the hypothesis that the attention mechanism creates pressure for learning more compositional emergent languages.

[2]The pretrained model is registered as `convnext_tiny` in the `torchvision` library

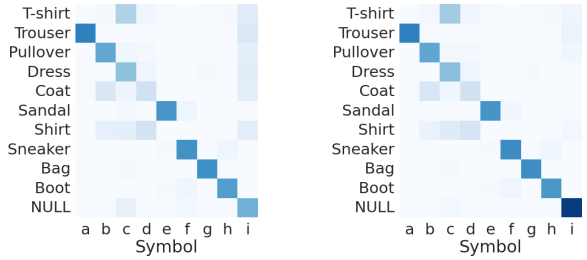


(a) Successful coordination with the same image.

(b) Successful coordination with the different images with the same type.

(c) Failure case. The attention maps are not focused.

Figure 3: Attention maps produced by a AT-AT agent pair (Left: Speaker, Right: Listener).



(a) Speaker (b) Listener

Figure 4: A frequency heatmap of symbol-concept association derived from the attention weights of Transformer AT-AT agent pairs.

One possible interpretation is that the attention mechanism adds flexibility to the model and it simply leads to better learning of the task. We observe the effect in TrainAcc, where the attention models consistently outperform their non-attention baselines. However, we can still see the contribution of the attention mechanism besides the flexibility. We can see some NoAT-NoAT agents and AT-AT agents exhibit comparable TrainAcc around 75%, which means they are successful at optimization to a similar degree. However, we observe all the AT-AT agents significantly outperform any of the NoAT-NoAT agents, which indicates the degree of optimization alone cannot explain the better GenAcc and TopSim scores of the attention model. Therefore, the results supports our initial hypothesis that the attention mechanism facilitates developing compositional languages.

In some settings, we observe that adding the attention mechanism to only one of Speaker and Listener does not lead to better TopSim scores compared the NoAT-NoAT agents to as in the AT-NoAT and NoAT-AT LSTM agents. However, when both Speaker and Listener agents have the attention mechanism, they all exhibit more generalizable and compositional languages, which indicates the effect of attention in Speaker and Listener is multiplicative.

Attention agents learn to associate input attributes and symbols

Having confirmed that the attention agents give rise to more compositional languages, we proceed to examine if they use attention in an expected way, i.e., producing/understanding each symbol by associating them with a single input concept. We focus on analyzing the Transformer agents below.

We inspect the attention weights and confirm that attention agents generally learn to focus on a single object when generating a symbol as in Figure 3(a) and (b), although there are some failure cases as shown in Figure 3(c).

Given the observations above, we can associate each symbol in each message with the concepts defined in the game via attention weights. We visualize the association from a pair of AT-AT agents in Figure 4 to inspect the mapping patterns developed by the agents. A symbol is considered to be associated with a concept when the center of gravity of the attention weights is within the bounding box of the item in an image. The results show that the mappings learned by Speaker and Listener have a strong tendency to agree in almost all cases. We identify three types of symbol-to-concept mapping patterns.

Monosemy. A single symbol always refers to a single concept, e.g., `a`, `e`, and `g`. This is a desired mapping pattern that allows an unambiguous interpretation of symbols.

Polysemy. A single symbol refers to multiple concepts, e.g., `b`, `c`, and `d`. These symbols are somewhat ambiguous, but they seem to be affected by the visual similarity of the fashion items, e.g., `b` refers to tops (Pullover, Coat, and Shirt) and `f` refers to shoes (Sandal, Sneaker, and Boot). This pattern demonstrates that the semantics of emergent language can be heavily influenced by the property of the input objects.

Gibberish. We observe a few cases where the attention weights do not consistently focus on any particular regions, e.g., the symbol `i`. These gibberish symbols could have conveyed something informative but uninterpretable to humans, but we confirmed that this indicates the results of optimization failure. We run 10 rounds of referential games with 45 candidates with a single agent pair, and the communication success rate is much lower when the Speaker's message contained gibberish symbols, compared to the overall score ($27.7 < 45.3$).

The coordination of Speaker and Listener attentions predicts the task performance

An important prerequisite of successful communication is the participants engaging in joint attention and establishing a shared understanding of each word (Garrod & Pickering, 2004). Here we show that the degree of the alignment between Speaker's and Listener's attention weights can be regarded as a proxy of their mutual understanding and predictive of communication success.

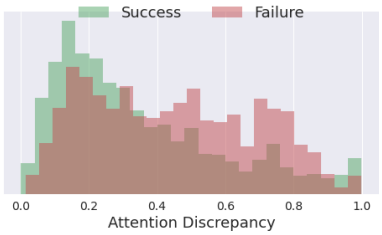


Figure 5: The distribution of attention discrepancy scores from the trials of the AT-AT agents. The frequency is normalized within successful and failed trials respectively.

We define a metric called *attention discrepancy*, which measures the difference between the attention weights of Speaker and Listener given the same inputs. For each symbol m_t in a message, Speaker and Listener have attention weights $\boldsymbol{a}_t^{(S)}, \boldsymbol{a}_t^{(L)}$. The metric is calculated by computing the average of the Jensen–Shannon Divergence (Lin, 1991) between the attention maps: $\frac{1}{T}\sum_{t=1}^{T} D_{\mathrm{JS}}\left(\boldsymbol{a}_t^{(S)} \| \boldsymbol{a}_t^{(L)}\right)$.

Figure 5 shows the distributions of attention discrepancy scores within successful and failed communications. These distributions are significantly dissimilar ($p < 0.001$ in the Kolmogorov-Smirnov test), with successful communications exhibiting lower scores than failed communications The results indicate a correlation between attention weight alignment and communication success to some degree. This suggests that attention agents develop intuitive communication protocols that rely on a shared understanding of symbols.

Related Work

Emergent Language

Emergent language should exhibit *compositionality*, allowing interpretability, generalizability, and ease of learning (Li & Bowling, 2019; Ren et al., 2020). Existing studies have shown that the compositionality of language can be improved by learning across generations (Li & Bowling, 2019; Ren et al., 2020), learning with a population (Rita, Strub, et al., 2022), applying noise to the communication channel (Łukasz Kuciński et al., 2021), and balancing the learning speed of the agents (Rita, Tallec, et al., 2022).

While recent studies have primarily used simple RNN-based Speaker and Listener agents, other agent architectures have been employed for specific purposes. For example, Chaabouni et al. (2019) and Ryo et al. (2022) used the LSTM sequence-to-sequence (with attention) architecture to study transduction from a grammar-generated input to an emergent language. Evtimova et al. (2018) compared attentional and non-attentional agents in a multi-modal and multi-step referential game, observing improvements in the out-of-domain test but not the in-domain test. However, it is unclear whether these trends generalize to other settings.

Our study aims to investigate the impact of the model architecture's inductive biases on the compositionality of the emergent language. Similarly, Słowik et al. (2020) compared Speaker agents that process inputs as a graph, sequence, or bag-of-words to show that the graph architecture results in more compositional emergent languages. Our study contributes further empirical evidence in this direction, specifically exploring the role of the attention mechanism.

Attention Mechanism in Machine Learning

The attention mechanism has proven effective in supervised learning (Xu et al., 2015; Vaswani et al., 2017), becoming an integral part of modern neural networks. Conceptually, this mechanism models pairwise associations between a query and a subset of key-values. This enables the model to focus on a subpart of compositional representation and has been shown to enforce compositional solutions in visual reasoning (Hudson & Manning, 2018), symbolic reasoning (Korrel, Hupkes, Dankers, & Bruni, 2019), image generation (Hudson & Zitnick, 2021). Our study shows that by utilizing attention to model the association between a symbol and an object attribute, the agents can discover more compositional languages in a referential game without additional supervision.

Attention may also offer interpretability. Attention has been shown to provide plausible alignment patterns between inputs and outputs, such as source and target words in machine translation (Bahdanau et al., 2015) and image regions and words in image captioning (Xu et al., 2015; Yang et al., 2016). However, the alignment patterns may not always align with human intuitions (Alkhouli, Bretschner, & Ney, 2018; F. Liu et al., 2020) and there is an ongoing debate about the extent to which the attention weights can be used as explanations for model predictions (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Bibal et al., 2022). In this paper, our agents exhibit relatively straightforward attention patterns that allow for unambiguous interpretations. However, the interpretability of attention weights with more complex models and environments requires further investigation.

Discussion and Conclusion

When we seek to interpret a complex object, we are likely to dynamically change our focus on its subpart to describe the whole (Rensink, 2000). Hearing a word aids in recognizing its referent (Boutonnet & Lupyan, 2015) and affects where to focus in an image (Estes, Verges, & Barsalou, 2008). Motivated by these observations, we implemented agents with the dynamic interaction between symbols and inputs in the form of the attention mechanism. We showed that the attention agents develop more compositional languages than their non-attention counterparts. This implies that the human capacity for dynamic focus may have contributed to developing compositional language.

To better understand how cognitive properties shape language, future research should explore additional architectural variations inspired by human cognitive processing. For instance, incorporating joint attention (Kwisthout et al., 2008) into language training may enhance communication success, as alignment between speaker and listener attention appears to be a factor.

Acknowledgement

We thank the anonymous reviewers for their insightful helpful comments to improve the manuscript.

References

Alkhouli, T., Bretschner, G., & Ney, H. (2018). On the alignment problem in multi-head attention-based neural machine translation. In *Proceedings of the third conference on machine translation: Research papers.*

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International conference on learning representations.*

Bibal, A., Cardon, R., Alfter, D., Wilkens, R., Wang, X., François, T., & Watrin, P. (2022). Is attention explanation? an introduction to the debate. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers).*

Bouchacourt, D., & Baroni, M. (2018). How agents see things: On visual representations in an emergent language game. In *Proceedings of the 2018 conference on empirical methods in natural language processing.*

Boutonnet, B., & Lupyan, G. (2015). Words jump-start vision: a label advantage in object recognition. *Journal of vision*, *15 12*, 11.

Brighton, H., & Kirby, S. (2006). Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial Life*, *12*, 229-242.

Brinck, I. (2000). Attention and the evolution of intentional communication. *Pragmatics & Cognition*, *9*, 255-272.

Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., & Baroni, M. (2020). Compositionality and generalization in emergent languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics.*

Chaabouni, R., Kharitonov, E., Lazaric, A., Dupoux, E., & Baroni, M. (2019). Word-order biases in deep-agent emergent communication. In *Proceedings of the 57th annual meeting of the association for computational linguistics.*

Chaabouni, R., Strub, F., Altché, F., Tarassov, E., Tallec, C., Davoodi, E., ... Piot, B. (2022). Emergent communication at scale. In *International conference on learning representations.*

de Diego-Balaguer, R., Martinez-Alvarez, A., & Pons, F. (2016). Temporal attention as a scaffold for language development. *Frontiers in Psychology*, *7*.

Estes, Z., Verges, M., & Barsalou, L. W. (2008). Head up, foot down. *Psychological Science*, *19*, 93 - 97.

Evtimova, K., Drozdov, A., Kiela, D., & Cho, K. (2018). Emergent communication in a multi-modal, multi-step referential game. In *International conference on learning representations.*

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences*, *8*(1), 8 - 11. doi: DOI: 10.1016/j.tics.2003.10.016

Hendrycks, D., & Gimpel, K. (2016). Gaussian error linear units (gelus). *ArXiv*, *abs/1606.08415*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hudson, D. A., & Manning, C. D. (2018). Compositional attention networks for machine reasoning. In *International conference on learning representations.*

Hudson, D. A., & Zitnick, C. L. (2021). Compositional transformers for scene generation. In *Advances in neural information processing systems.*

Jain, S., & Wallace, B. C. (2019). Attention is not Explanation. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers).*

Korrel, K., Hupkes, D., Dankers, V., & Bruni, E. (2019). Transcoding compositionally: Using attention to find more generalizable solutions. In *Proceedings of the 2019 acl workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP.*

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*.

Kwisthout, J., Vogt, P., Haselager, W. P., & Dijkstra, T. (2008). Joint attention and language evolution. *Connection Science*, *20*, 155 - 171.

Lazaridou, A., & Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. *ArXiv*, *abs/2006.02419*.

Lazaridou, A., Hermann, K. M., Tuyls, K., & Clark, S. (2018). Emergence of linguistic communication from referential games with symbolic and pixel input. In *International conference on learning representations.*

Lazaridou, A., Peysakhovich, A., & Baroni, M. (2017). Multi-agent cooperation and the emergence of (natural) language. In *International conference on learning representations.*

Lewis, D. K. (1969). *Convention: A philosophical study*. Cambridge, MA, USA: Wiley-Blackwell.

Li, F., & Bowling, M. (2019). Ease-of-teaching and language structure from emergent communication. In *Advances in neural information processing systems* (Vol. 32).

Lin, J. (1991). Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, *37*, 145-151.

Liu, F., Ren, X., Wu, X., Ge, S., Fan, W., Zou, Y., & Sun, X. (2020). Prophet attention: Predicting attention with future attention. In *Advances in neural information processing systems.*

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

Lowe, R., Foerster, J. N., Boureau, Y.-L., Pineau, J., & Dauphin, Y. (2019). On the pitfalls of measuring emergent communication. In *Aamas.*

Luong, T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 conference on empirical methods in natural language processing.*

Mordatch, I., & Abbeel, P. (2018). Emergence of grounded compositional language in multi-agent populations. In *The thirty-second aaai conference on artificial intelligence.*

Ren, Y., Guo, S., Labeau, M., Cohen, S. B., & Kirby, S. (2020). Compositional languages emerge in a neural iterated learning model. In *International conference on learning representations.*

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition*, *7*, 17 - 42.

Rita, M., Strub, F., Grill, J., Pietquin, O., & Dupoux, E. (2022). On the role of population heterogeneity in emergent communication. In *International conference on learning representations.*

Rita, M., Tallec, C., Michel, P., Grill, J.-B., Pietquin, O., Dupoux, E., & Strub, F. (2022). Emergent communication: Generalization and overfitting in lewis games. In *Advances in neural information processing systems.*

Rodríguez Luna, D., Ponti, E. M., Hupkes, D., & Bruni, E. (2020). Internal and external pressures on language emergence: least effort, object constancy and frequency. In *Findings of the association for computational linguistics: EMNLP 2020.*

Ryo, U., Taiga, I., Koki, W., & Yusuke, M. (2022). Categorial grammar induction as a compositionality measure for emergent languages in signaling games. In *Proceedings of emergent communication workshop at iclr 2022.*

Słowik, A., Gupta, A. K., Hamilton, W. L., Jamnik, M., Holden, S., & Pal, C. J. (2020). Structural inductive biases in emergent communication. In *Proceedings of the 43rd annual meeting of the cognitive science society.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (Vol. 30).

Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP).*

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, *8*(3–4), 229–256.

Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *ArXiv*, *abs/1708.07747*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., ... Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the international conference on machine learning* (Vol. 37).

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. In *IEEE conference on computer vision and pattern recognition (CVPR).*

Łukasz Kuciński, Korbak, T., Kołodziej, P., & Miłoś, P. (2021). Catalytic role of noise and necessity of inductive biases in the emergence of compositional communication. In *Advances in neural information processing systems* (Vol. 34).