EMERGENT COMMUNICATION WITH ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

To develop computational agents that can better communicate with others 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 a thing or scene as a composite of concepts and those concepts are further mapped onto words. We implement this intuition as 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 helps us understand 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.

1 INTRODUCTION

We endow computational agents with an ability to focus their attention on a particular concept in the environment and communicate the concept using emergent language. Emergent language refers to the communication protocol based on discrete symbols developed by computational agents to achieve a specific task goal (Nowak & Krakauer, 1999; Lazaridou et al., 2017). One important goal in the study of emergent language is clarifying conditions that lead to more compositional emergent languages. Seeking compositionality provides insights into the origin of the compositional natures of human language and helps develop efficient communication protocols in multi-agent systems that are interpretable and easily controllable by humans.

Much recent work studies emergent language with generic deep agents (Lazaridou & Baroni, 2020) with minimum assumptions on the inductive bias of the model architecture. For example, a typical speaker agent encodes information into a single fixed-length vector to initialize the hidden state of a RNN decoder and generate symbols (Lazaridou et al., 2017; Mordatch & Abbeel, 2018; Ren et al., 2020). Only a few studies have explored some architectural variations (Słowik et al., 2020) to improve the compositionality of emergent language, and there remains much to be discussed about the effects of the inductive bias provided by different architectures.

We posit that towards more human-like emergent language we need to explore other modeling choices that reflect the human cognitive process. In this study, we focus on the attention mechanism. Attention is one of the most successful neural network architectures (Bahdanau et al., 2015; Xu et al., 2015; Vaswani et al., 2017) that have an analogy in psychology (Lindsay, 2020). The conceptual core of attention is an adaptive control of limited resources, and we hypothesize that this creates pressure for learning more compositional emergent languages. Compositionality entails a whole consisting of subparts. Attention allows the agents to dynamically highlight different subparts of an object when producing/understanding each symbol, which potentially results in clear associations between the object attributes and symbols.

Another reason to explore the attention mechanism is its interpretability. Emergent language is optimized for task success 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 et al., 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.,

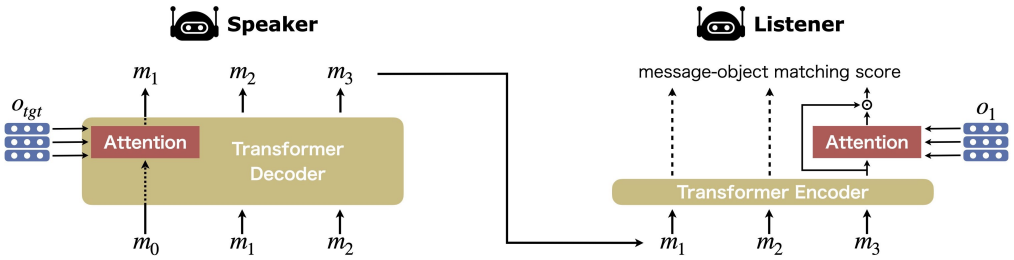


Figure 1: Illustration of the attention agents in the referential game.

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 referential game (Lewis, 1969; Lazaridou et al., 2017), which involves two agents: *Speaker* and *Listener*. The goal of the game is to convey the type of object that Speaker sees to Listener. We conduct extensive experiments with two types of environments: one-hot vector games where the referents are given in the form of binary vectors and more realistic settings where the referents are pixel images. We compare our attention agents against non-attention counterparts with various metrics of compositionality and show that adding attention mechanisms with disentangled inputs to either/both Speaker or/and Listener helps develop a more compositional language. We also analyze the attention weights and explore how they can help us understand the learned language. After we confirm that the agents learn to selectively attend to the inputs, we visualize the learned symbol-concept mapping and demonstrate how the emergent language can deviate from human language and, with experiments with pixel images, how the emergent language can be affected by the visual similarity of referents. We also investigate the potential of the alignment between Speaker's and Listener's attention weights as a proxy for the establishment of common understanding and show that the alignment correlates with task success.

2 Experimental Framework

2.1 Referential Game

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

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

2.2 Agent Architectures

Our goal in this paper is to test the effect of the attention mechanism on emergent language. The attention mechanism in its general form takes a query vector Q and key vectors K to compute attention weights over the keys. Then the attention weights are used to produce a weighted sum vector of value vectors V. A key feature of attention is that it allows the agents to selectively attend to a part of disentangled vector inputs. Our intuition is that modeling direct associations between the symbol representations as Q and disentangled input representations as K and V will bias agents toward packing each symbol with the information of a meaningful subpart of the inputs rather than with opaque and non-compositional information. To test these hypotheses, we design non-attention and attention agents for both Speaker and Listener (Figure 1).

2.2.1 OBJECT ENCODER

The objects are presented to the agents in the form of a set of real-valued vectors, e.g., attribute-wise one-hot vectors or patch-wise pretrained CNN feature vectors.

All the 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). The baseline non-attention agents average the transformed vectors into a single vector for the subsequent computations, whereas the attention agents leave the transformed vectors intact and attend to the set of vectors.

2.2.2 SPEAKER AGENTS

Speaker has an auto-regressive message decoder that takes the transformed vectors of the target object o_{tgt} as input and generates a multi-symbol message m. There is a number of possible choices for the decoder architecture especially with attention (Bahdanau et al., 2015; Luong et al., 2015; Vaswani et al., 2017). In this paper, we adopt the Transformer decoder (Vaswani et al., 2017) because Transformer naturally incorporates the attention mechanism in its basic architecture and is also a current de-facto standard and well-studied in a range of fields (Devlin et al., 2019; Schneider et al., 2019; Dosovitskiy et al., 2021). Throughout this paper, the number of the Transformer source-target attention heads is set to one for the sake of interpretability and ease of analysis.

Non-attention (NoAT) Speaker is a baseline agent that encodes an object into a single vector and uses the vector as a source vector for the decoder. The source-target attention in the Transformer decoder always attends to that single vector and does not change where to focus during message generation.

Attention (AT) Speaker, in contrast, encodes an object into a set of vectors and the source-target attention dynamically changes its focus at each time step.

Our agent design is motivated by a fair comparison of non-attention and attention agents. They have the same number of parameters and only differ in whether they can dynamically change their focus when generating each symbol.

2.2.3 LISTNER 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 and choosing the object with the maximum score.

Non-attention (NoAT) Listener encodes the objects and message into a single vector respectively and computes the matching score by taking the dot product of the object and message vectors. The message vector is produced by feeding the message symbols to a Transformer encoder (Vaswani et al., 2017) and then averaging the outputs.

Attention (AT) Listener encodes the objects and message into a set of vectors respectively and the matching score is computed by averaging the symbol-object matching scores. For each symbol, the model produces an attention vector with the symbol vector as a query and the object vectors as keys and values. The symbol-object matching scores are then calculated as the dot product of the symbol vector and the attention vector.

2.3 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 for the task success and 0 otherwise.

As an auxiliary loss function, we employ an entropy regularization loss $L_H(\theta_S) = -\sum_t H(\pi_{\theta_S}(\cdot|t))$, where H is the entropy of a probability distribution, to encourage exploration.

We also add a KL loss $L_{\text{KL}}(\theta_S) = \sum_t D_{\text{KL}}(\pi_{\theta_S}(\cdot|t)\|\pi_{\bar{\theta}_S}(\cdot|t))$, 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 $\mathcal{L}(\theta_L) = -\log p_{\theta_L}(o_{tgt}|C)$. The details of hyperparameter settings and training procedures can be found in Appendix A.3.

2.4 Evaluation Metrics

We quantitatively evaluate emergent languages from two perspectives: (1) how well the language can be used to solve the task and (2) how well the language exhibits compositionality.

Generalization accuracy (Acc) 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 et al., 2008), is one of the most commonly used metrics to assess the compositionality of emergent language (Brighton & Kirby, 2006; Lazaridou et al., 2018; Ren et al., 2020; Chaabouni et al., 2020). Intuitively, TopSim checks if similar objects have similar messages assigned. To compute TopSim, we enumerate all the object-message pairs $(o^1, m^1), ..., (o^{|\mathcal{O}|}, m^{|\mathcal{O}|})$ with a trained Speaker and define a distance function for objects $d_{\mathcal{O}}(o^i, o^j)$ 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 $d_{\mathcal{O}}(o^i, o^j)$, 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.

Positioanl disentanglement (PosDis) measures if each position in the message is associated with a particular attribute of the object (Chaabouni et al., 2020). For example, when an English speaker describes objects with color and shape attributes, they would say "red circle" and "blue triangle" with consistent word order, not "red circle" and "triangle blue". PosDis assigns high scores for such order-dependent languages.

Suppose each object o has n attributes $\{a(1, o), ..., a(n, o)\}$. Given the object-message pairs, we compute the mutual information between the i-th attributes of the objects $a(i; o)$ and symbols at the j-th position $s(j; m)$ in the messages: $\mathcal{I}(a(i; o), s(j; m))$. PosDis of the position j is computed by finding the attributes with the highest mutual information $\mathcal{I}^j_{(n)} = \max_i \mathcal{I}(a(i; o), s(j; m))$ and the second highest $\mathcal{I}^j_{(n-1)}$, taking the difference between them, and then normalizing the difference by the entropy of j-th position: $\frac{\mathcal{I}^j_{(n)} - \mathcal{I}^j_{(n-1)}}{H(s(j;m))}$. The final PosDis score is the average of the position-wise values over the positions with the values with zero entropy ignored.

Bag-of-symbols disentanglement (BosDis) captures position-agnostic correspondence between message symbols and object attributes (Chaabouni et al., 2020). For each message symbol s_k in the vocabulary, we count its occurrence in each message m: $c(s_k; m)$.

BosDis of the symbol s_k is computed by finding the attributes with the highest mutual information $\mathcal{I}^k_{(n)} = \max_i \mathcal{I}(a(i; o), c(s_k; m))$ and second highest $\mathcal{I}^k_{(n-1)}$, taking the difference between their mutual information, and then normalizing the difference by the entropy of the count: $\frac{\mathcal{I}^k_{(n)} - \mathcal{I}^k_{(n-1)}}{H(c(s_k;m))}$. The final BosDis score is the average of the symbol-wise values over vocabulary with symbols with the values with zero entropy ignored.

3 EXPERIMENTAL SETUP

3.1 ONE-HOT GAME

Our first experimental setting is the referential game with objects consisting of one-hot vectors (Lazaridou et al., 2017; Kottur et al., 2017; Chaabouni et al., 2020; Guo et al., 2022). The objects are represented as a combination of two attributes (e.g., shape and color), which is a popular setting in the literature (Li & Bowling, 2019; Ren et al., 2020). Each attribute has 8 possible values, which results in 64 distinct objects. To evaluate generalization to unseen objects, we randomly split the 64 objects into training and evaluation sets at a ratio of 48/16. The number of candidates is set to 16, which is the maximum number of objects available in the evaluation set.

Input Representation. The objects are disentangled in terms of attributes and represented as a set of one-hot vectors where each dimension indicates a different value of a different attribute (i.e., two 16-dim one-hot vectors).

Agent Configurations. The vocabulary size of the agents is set to 16 and the message length of Speaker is 2. In this setting, a perfectly compositional language that scores the maximum score for all the compositionality metrics would assign different symbols to each attribute value, and symbols at each position consistently refer to a single attribute.

3.2 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 are able to develop a 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 to test if the agents develop attention patterns for each symbol that focus on a single item in a given image. The items in each image are taken from the Fashion-MNIST dataset[1] (Xiao et al., 2017). The images for the game are created by placing two item images, each of which is rescaled to the size of 48×48, on a 224×224 black canvas (Figure 3). The places of the items are randomly sampled so that the items never overlap.

As we would like to let the agents capture concepts that are understandable by humans, we design the game so that the agents have to communicate the classes of the items in each image rather than distinguish among specific instances of images to avoid degenerated solutions (Lazaridou et al., 2018; Bouchacourt & Baroni, 2018). The objects in the game are distinguished solely by the classes of items e.g., two images with different pairs of a T-shirt and sneaker in different positions are considered to be the same object. The Fashion-MNIST dataset has 10 item classes and thus the game has ${}_{10}\mathrm{C}_2$ = 45 different types of objects. Also, the Speaker and Listener are presented with different images within the same type as the target to facilitate learning a robust communication protocol (Rodríguez Luna et al., 2020). The 45 types are randomly divided into training and evaluation sets at a ratio of 30/15, and the number of candidates is set to 15.

Input representation. Each image is encoded into $7 \times 7 \times 768$-dim feature vectors with a pretrained ConvNet[2] (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.

More details of the experimental setup are found in Appendix A.

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

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

	One-hot				Fashion-MNIST		
	Acc↑	TopSim↑	PosDis↑	BosDis↑	Acc↑	TopSim↑	BosDis↑
NoAT-NoAT	89.4±7.4	55.1±6.4	12.9±9.3	49.3±8.4	45.6±1.8	29.6±1.7	20.0±3.9
AT-NoAT	**95.9±5.3***	68.8±7.7*	39.3±19.5*	54.7±6.8*	52.1±6.8*	**51.8±6.1***	29.5±5.3*
NoAT-AT	93.6±4.4*	58.2±4.8	21.4±12.0*	**67.5±11.5***	44.3±4.7	39.9±6.0*	24.3±6.8
AT-AT	**95.9±4.2***	**73.5±12.0***	**53.5±25.2***	62.6±12.2*	**54.0±4.0***	51.1±3.6*	**31.9±8.6***

Table 1: The results of the referential game. The leftmost column indicates the architectures of Speaker-Listener. The scores show the average and standard deviations of 20 different runs for the One-hot game and 10 runs for the Fashion-MNIST game. The highest average scores in each column are boldfaced. $*$ indicates scores with statistically significantly improvement ($p < 0.05$ in Welch's t-test) over NoAT-NoAT.

4 Results

4.1 Attention agents find more compositional solutions

We evaluate non-attention agents and attention agents where either/both Speaker and Listener have dynamic attention. The generalization accuracy and compositionality metrics in the one-hot and Fashion-MNIST games are presented in Table 1. The PosDis metric cannot be computed for the Fashion-MNIST game as the game defines only a single attribute for the objects.

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 (Acc) and compositionality metrics (TopSim, PosDis, and BosDis) with different degrees of significance. This observation provides evidence in support of the hypothesis that attention creates pressure for learning more compositional emergent languages.

When looking at Acc and TopSim, AT-NoAT performs better than NoAT-AT, suggesting the effect of the attention mechanisms implemented in this study is stronger at the Speaker side. When looking at more fine-grained metrics in the one-hot game, we see that AT-NoAT provides a significantly better PosDis score than that of NoAT-AT (39.3±19.5 vs. 21.4±12.0, $p < 0.05$ in Welch's t-test) but not for BosDis (54.7±6.8 vs. 67.5±11.5, $p \geq 0.05$), suggesting that the Speaker attention helps developing a language with consistent word order (e.g., the first symbols usually refer to the shape attribute) more than the Listener attention.

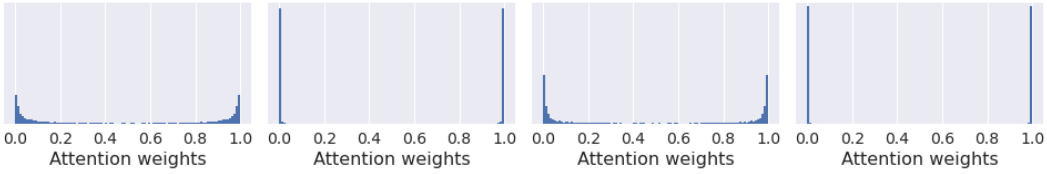
AT-AT provides a performance mostly comparable to AT-NoAT as a statistically significant difference between them is only observed in BosDis (62.6±12.2 vs. 54.7±6.8, $p < 0.05$) in the one-hot game. Although having attention at both Speaker and Listener does not provide significant improvements in this setting, we will show in §4.3 that the double-attention agents enable us to inspect the alignment of Speaker's and Listener's attention weights, which is an informative indicator for communication success.

4.2 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.

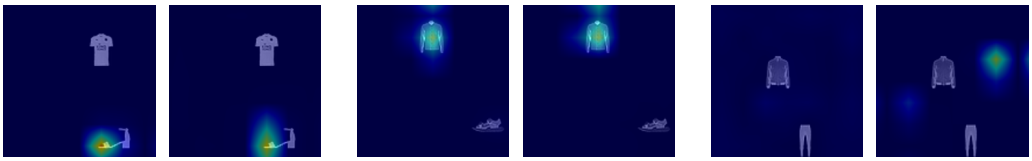
Figure 2 shows the distributions of the attention weights assigned to each input attribute vector in the one-hot game. We observe that most attention weights are around either 0 or 1, indicating that the attention agents generally learn to focus attention on a single value, which may have contributed to more compositional languages. We also inspect the attention weights in the Fashion-MNIST game and confirm that attention agents generally learn to focus on a single object when generating a symbol as in Figure 3(a)-(d), although there are some failure cases as shown in Figure 3(e)-(f).

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 between each symbol in the vocabulary and each concept from a pair of AT-AT agents in Figure 4 to inspect the mapping



(a) AT-NoAT: S (b) NoAT-AT: L (c) AT-AT: S (d) AT-AT: L

Figure 2: The distribution of attention weights from attention agents in the one-hot game. The attention weights are obtained by feeding all the possible objects to the agents from 20 different runs for each setting.



(a) Speaker (b) Listener (c) Speaker (d) Listener (e) Speaker (f) Listener

Figure 3: Images with the attention maps produced by a AT-AT agent pair.

patterns developed by the agents. A symbol is considered to be associated with a concept when the attribute value has the largest attention weight in the one-hot game or the center of gravity of the attention weights is within the bounding box of the item in the Fashion-MNIST game. Overall, we observe that the mappings learned by Speaker and Listener are mostly aligned except for the symbol `c` in the one-hot game and the symbol `i` in the Fashion-MNIST game. We identify three types of symbol-to-concept mapping patterns.

Monosemy. A single symbol always refers to a single concept, e.g., `a`, `b`, and `d` in the one-hot game; `a`, `e`, and `g` in the Fashion-MNIST game. This is a desired mapping pattern that allows an unambiguous interpretation of symbols. However, we also observe "synonyms" where a single concept is referred to by multiple symbols, e.g., `1-3` by `d` and `e` in the one-hot game.

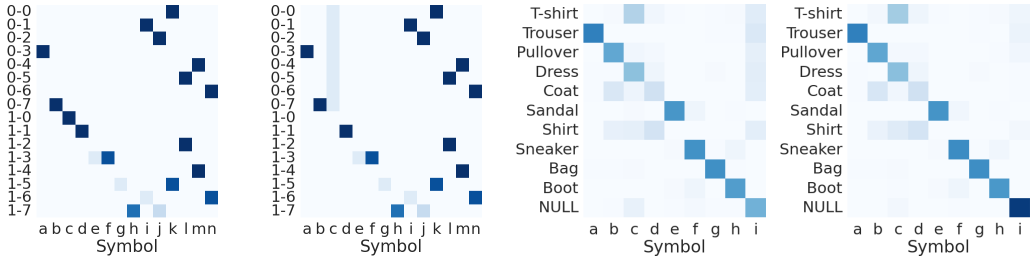
Polysemy. A single symbol refers to multiple concepts, e.g., `j-n` in the one-hot game; `b`, `c`, and `d` in the Fashion-MNIST game. These symbols are somewhat ambiguous, but it is likely that the agents develop a way to resolve the ambiguity, especially in the one-hot game. In the one-hot game, agents achieve more than 90% average accuracy scores with unseen objects and relatively high PosDis scores (Table 1). Moreover, these polysemous symbols always refer to two values of the different attributes. Given these observations, the agents are likely to have learned a consistent word order in terms of the object attributes and disambiguated the polysemous symbols by their position.

The polysemous symbols in the Fashion-MNIST game 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 is in contrast to the one-hot game, where the object attribute representations are initialized randomly, which probably led to an arbitrary assignment of polysemy. The polysemous patterns in the Fashion-MNIST game demonstrate 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 in the Fashion-MNIST game, e.g., the symbol `i`. These gibberish symbols could have conveyed something informative but uninterpretable to humans, but we confirmed that the communication success rate with the gibberish symbols is much lower (19.4%) than the overall success rate (45.3%), indicating that the gibberish symbols are the results of optimization failure.

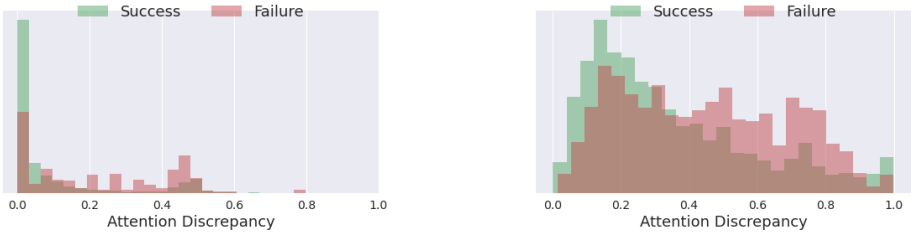
4.3 The coordination of Speaker and Listener attentions predicts the task performance

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



(a) One-hot: Speaker (b) One-hot: Listener (c) Fashion-MNIST: Speaker (d) Fahion-MNIST: Listener

Figure 4: The mappings between symbols and concepts derived from the attention weights of a AT-AT agent pair in the one-hot and FashionMNIST game. The darker a cell, the more frequently the symbol and concept are associated.



(a) Onehot (b) Fashion-MNIST

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

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.

We define a metric called *attention discrepancy*, which measures the difference between the attention weights of Speaker and Listener given the same inputs. The metric is calculated by computing the Jensen–Shannon Divergence (Lin, 1991) between the Speaker's and Listener's attention maps.

Figure 5 shows the frequency of attention discrepancy scores within successful and failed communications.[3] We can see that successful communications can be characterized by lower attention discrepancy compared to failed communications, indicating that the alignment of the attention weights does correlate with communication success. This suggests that the attention agents largely develop intuitive communication protocols that require a shared understanding of symbols.

5 Related Work

5.1 Emergent Language

One of the desired properties of emergent language is *compositionality* because more compositional languages are supposed to be intuitive, interpretable, generalizable, and easier to learn (Li & Bowling, 2019; Ren et al., 2020). Compositional languages in this context usually have symbols each of which corresponds to one of the primitive concepts in the environment and combine them in a straightforward manner to form a message. Existing studies have investigated various environmental pressures to improve the compositionality of emergent language: learning across generations (Li & Bowling, 2019; Ren et al., 2020), learning with a population (Rita et al., 2022), applying noise to communication channel (Kuciński et al., 2021).

[3]We observe small peaks around 0.5 in the one-hot game for both successful and failed communications, and at a closer look, this corresponds to cases where the Speaker and Listener attend to a different attribute at one of the two symbols in the message.

These recent studies have been mainly conducted with simple RNN-based Speaker and Listener agents, where the inputs (e.g., attribute values or raw images) are encoded into a single input vector for subsequence computations. Other agent architectures have been employed for other specific motivations. Chaabouni et al. (2019) and Ryo et al. (2022) used the LSTM sequence-to-sequence (with attention) architecture to study how agents transduce a sequential input generated from a certain grammar to their own emergent language. Gupta et al. (2021) developed a patch-based Speaker architecture that produces messages by focusing important image patches.

One of our goals in this study is to investigate how the inductive biases of the Speaker and Listener architectures affect the compositionality of the emergent language. In a similar spirit, Słowik et al. (2020) compared Speaker agents which process inputs in the form of a graph, sequence, and bag-of-words and show that the graph architecture exhibits more compositional emergent languages. Our study offers additional empirical evidence in this direction with the attention mechanism.

5.2 Attention Mechanism

The attention mechanism has been shown very effective in supervised learning (Xu et al., 2015; Vaswani et al., 2017) and has become an indispensable modeling piece in modern neural networks. Conceptually, attention models pairwise associations between a query vector and key-value vectors. We hypothesize that this can provide an inductive bias for the agents in the referential game to learn a one-to-one mapping between symbols and concepts.

Some studies have shown that attention can help the model reach compositional solutions. Hupkes et al. (2019) and Baan et al. (2019) investigated a technique called *attentive guidance*, which leads the model to converge to compositional solutions by adding supervision for the attention weights. Korrel et al. (2019) also showed that a model with attention converges to more compositional solutions compared to its non-attention counterpart in a lookup table task. Our study demonstrates that the agents with the attention mechanism can find more compositional languages in the referential game without any additional supervision.

Another potential benefit of attention is interpretability. Attention has been shown to provide plausible alignment patterns between inputs and outputs, for example, 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). On the other hand, the alignment patterns may need careful interpretation. The patterns are not always consistent with human intuitions (Alkhouli et al., 2018; Liu et al., 2020) and there is an ongoing debate about to what extent the attention weights can be used as explanations for model prediction (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Bibal et al., 2022). In this paper, our agents show relatively straightforward attention patterns that allow unambiguous interpretations but the interpretability of attention weights with more complex models and environments needs further investigation in the future.

6 Discussion and Conclusion

When we face a complex object, we are likely to dynamically change our focus on its subpart to describe the whole (Rensink, 2000). Hearing a word helps us quickly recognize its referent (Boutonnet & Lupyan, 2015) and affects where to focus in an image (Estes et al., 2008). Motivated by these psychological observations, we implemented agents with the dynamic interaction between symbols and input representations in the form of the attention mechanism. We showed that the attention agents develop more compositional languages than the non-attention agents. This implies that the human ability to focus might have contributed to developing compositional language.

Overall, our results suggest that future work should explore more architectural variations that originate from human cognitive processing to gain insights into the effect of the cognitive property on shaping a language. Given the observation that the alignment of Speaker's and Listener's attention is somewhat indicative of successful communication in §4.3, for example, future work can explore how incorporating joint attention (Kwisthout et al., 2008) into training helps to learn a language.

REPRODUCIBILITY STATEMENT

The experiments described in this paper are ensured to be reproducible in the following ways:

- We describe the task setup and models in §2 together with more specific descriptions of each task in §3. The additional details and hyperparameters to reproduce the results are given in appendix A.
- The artifacts used in the experiments (the Fashion-MNIST dataset and pretrained ConvNet encoder in §3.2) are all publicly available.
- The source code will be released upon acceptance.

REFERENCES

Tamer Alkhouli, Gabriel Bretschner, and Hermann Ney. On the alignment problem in multi-head attention-based neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, 2018.

Joris Baan, Jana Leible, Mitja Nikolaus, David Rau, Dennis Ulmer, Tim Baumgärtner, Dieuwke Hupkes, and Elia Bruni. On the realization of compositionality in neural networks. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2019.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 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)*, 2022.

Diane Bouchacourt and Marco Baroni. 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*, 2018.

Bastien Boutonnet and Gary Lupyan. Words jump-start vision: a label advantage in object recognition. *Journal of vision*, 15 12:11, 2015.

Henry Brighton and Simon Kirby. Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial Life*, 12:229–242, 2006.

Rahma Chaabouni, Eugene Kharitonov, Alessandro Lazaric, Emmanuel Dupoux, and Marco Baroni. Word-order biases in deep-agent emergent communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. Compositionality and generalization in emergent languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Rahma Chaabouni, Florian Strub, Florent Altché, Eugene Tarassov, Corentin Tallec, Elnaz Davoodi, Kory Wallace Mathewson, Olivier Tieleman, Angeliki Lazaridou, and Bilal Piot. Emergent communication at scale. In *International Conference on Learning Representations*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Zachary Estes, Michelle Verges, and Lawrence W. Barsalou. Head up, foot down. *Psychological Science*, 19:93 – 97, 2008.

Simon Garrod and Martin J. Pickering. Why is conversation so easy? *Trends in Cognitive Sciences*, 8(1):8 – 11, 2004. ISSN 1364-6613. doi: DOI:10.1016/j.tics.2003.10.016.

Shangmin Guo, Yi Ren, Kory Wallace Mathewson, Simon Kirby, Stefano V. Albrecht, and Kenny Smith. Expressivity of emergent languages is a trade-off between contextual complexity and unpredictability. In *International Conference on Learning Representations*, 2022.

Kamal Gupta, Gowthami Somepalli, Anubhav Gupta, Vinoj Jayasundara, Matthias Zwicker, and Abhinav Shrivastava. Patchgame: Learning to signal mid-level patches in referential games. In *NeurIPS*, 2021.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *ArXiv*, abs/1606.08415, 2016.

Dieuwke Hupkes, Anand Singh, Kris Korrel, Germáan Kruszewski, and Elia Bruni. Learning compositionally through attentive guidance. In *20th International Conference on Computational Linguistics and Intelligent Text Processing*, 2019.

Sarthak Jain and Byron C. Wallace. 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)*, 2019.

Kris Korrel, Dieuwke Hupkes, Verna Dankers, and Elia Bruni. Transcoding compositionally: Using attention to find more generalizable solutions. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2019.

Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. Natural language does not emerge 'naturally' in multi-agent dialog. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2, 2008.

Ł ukasz Kuciński, Tomasz Korbak, PawełKoł odziej, and Piotr Mił oś. Catalytic role of noise and necessity of inductive biases in the emergence of compositional communication. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Johan Kwisthout, Paul Vogt, W.F.G Pim Haselager, and Ton Dijkstra. Joint attention and language evolution. *Connection Science*, 20:155 – 171, 2008.

Angeliki Lazaridou and Marco Baroni. Emergent multi-agent communication in the deep learning era. *ArXiv*, abs/2006.02419, 2020.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations*, 2017.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. Emergence of linguistic communication from referential games with symbolic and pixel input. In *International Conference on Learning Representations*, 2018.

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

Fushan Li and Michael Bowling. Ease-of-teaching and language structure from emergent communication. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37:145–151, 1991.

Grace W. Lindsay. Attention in psychology, neuroscience, and machine learning. *Frontiers in Computational Neuroscience*, 14, 2020.

Fenglin Liu, Xuancheng Ren, Xian Wu, Shen Ge, Wei Fan, Yuexian Zou, and Xu Sun. Prophet attention: Predicting attention with future attention. In *NeurIPS*, 2020.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Ryan Lowe, Jakob N. Foerster, Y-Lan Boureau, Joelle Pineau, and Yann Dauphin. On the pitfalls of measuring emergent communication. In *AAMAS*, 2019.

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. In *The Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Martin A. Nowak and D C Krakauer. The evolution of language. *Proceedings of the National Academy of Sciences of the United States of America*, 96 14:8028–33, 1999.

Yi Ren, Shangmin Guo, Matthieu Labeau, Shay B. Cohen, and Simon Kirby. Compositional languages emerge in a neural iterated learning model. In *International Conference on Learning Representations*, 2020.

Ronald A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 7:17 – 42, 2000.

Mathieu Rita, Florian Strub, Jean-Bastien Grill, Olivier Pietquin, and Emmanuel Dupoux. On the role of population heterogeneity in emergent communication. In *International Conference on Learning Representations*, 2022.

Diana Rodríguez Luna, Edoardo Maria Ponti, Dieuwke Hupkes, and Elia Bruni. Internal and external pressures on language emergence: least effort, object constancy and frequency. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

Ueda Ryo, Ishii Taiga, Washio Koki, and Miyao Yusuke. Categorial grammar induction as a compositionality measure for emergent languages in signaling games. In *Proceedings of Emergent Communication Workshop at ICLR 2022*, 2022.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pretraining for speech recognition. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*, 2019.

Agnieszka Słowik, Abhinav Kumar Gupta, William L. Hamilton, Mateja Jamnik, S. Holden, and Christopher Joseph Pal. Structural inductive biases in emergent communication. In *Proceedings of The 43rd Annual Meeting of the Cognitive Science Society*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Sarah Wiegreffe and Yuval Pinter. 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)*, 2019.

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

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *ArXiv*, abs/1708.07747, 2017.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*, volume 37, 2015.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

APPENDIX

A ADDITIONAL DETAILS ON EXPERIMENTAL SETUP

A.1 MINI-BATCH CREATION

Each mini-batch in our referential game is specified via two parameters: the batch size b and candidate size $|C|$. We first sample b objects from the set of the objects available in the game O. Each object in a mini-batch serves as the target object in a round of the game, and we sample $|C| - 1$ distractors for each target object. Importantly, we sample the objects so that every object in a round of the game has a different object type.

A.2 DATA PROCESSING IN THE FASHION-MNIST GAME

The Fashion-MNIST game (§3.2) is conducted on the level of object type rather than object instance, i.e., the agents have to communicate the classes of the items are on an image rather than those specific items and positions. Instead of generating the image and input features on the fly during training, we precompute the image features of 3000 instances for each object type and sample inputs from the pre-computed features during training.

A.3 HYPERPARAMETERS

The training setups and hyperparameters are shown in Table 2. The hyperparameter tuning is conducted for the weight of entropy loss α, and the scores in the paper are computed from the hyperparameter with the best generalization accuracy score for each agent.

	One-hot	Fashion-MNIST
Speaker Transformer decoder size	64	256
Listener Transformer encoder size	64	256
Training Batch size	48	480
Training steps	15K	50K
Evaluation Rounds	1500	15000
Entropy loss weight α	[0.1, 0.01, 0.001]	[0.1, 0.01, 0.001]
KL loss weight β	0.1	0.1
Learning rate	1e-4	1e-4

Table 2: Hyper-parameters of the experiments.