Semantics and Spatiality of Emergent Communication

Rotem Ben Zion, Boaz Carmeli, Orr Paradise, Yonatan Belinkov,
NeurIPS 2024

What is emergent communication (EC)

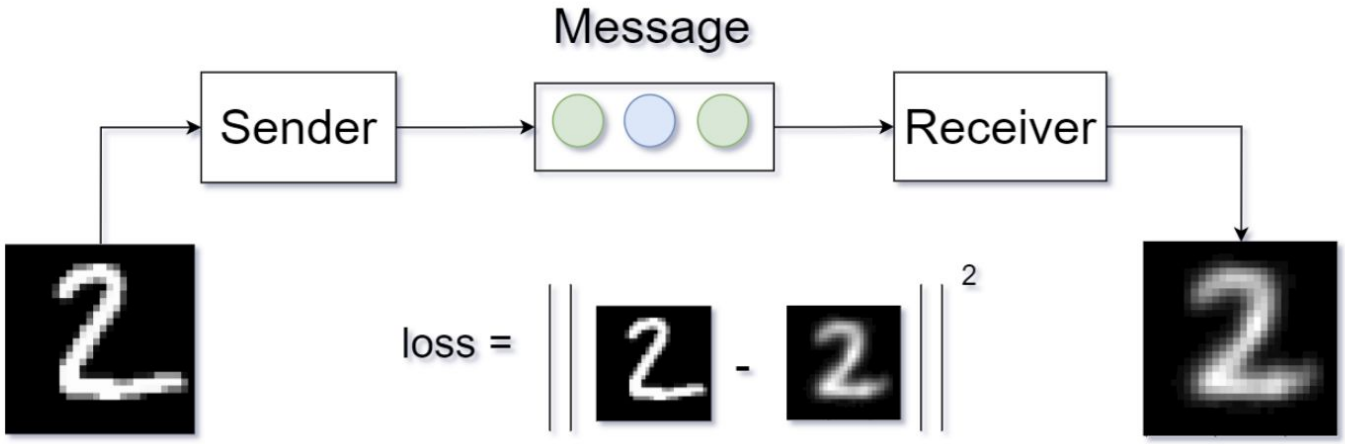
Emergent communication (**EC**) is an interdisciplinary field that

- raises questions like “what is communication and how does it work?”
  - “what is a message? what properties does it have?”
  - “how communication protocols emerge? can different objectives lead to different protocols with different properties? what objectives lead to the natural communication protocol?”
- and tries to design experiments (setups) that could answer these questions
  - communication is required to pass **information** between **agents** who need it for some **reason**
  - EC explores who are the agents and what reasons they have to communicate, and what information they want to pass given the reason

What emergent communication (EC) is dealing with

This is an example of a EC setup — an autoencoder:

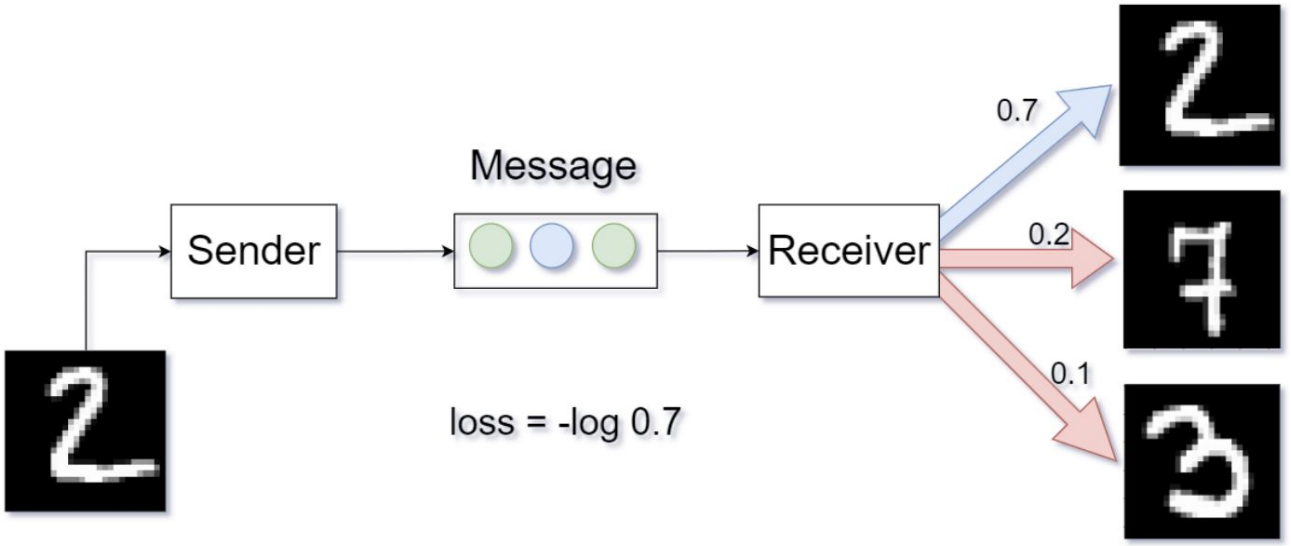
- Encoder (sender) sees a picture, computes its latent representation and sends it to a decoder
- Decoder (receiver) tries to reconstruct the original image.



What emergent communication (EC) is dealing with

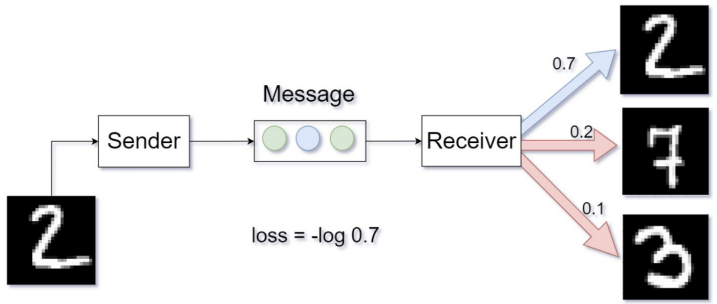
Another example — digit classifier:

- Encoder (sender) sees a picture, extracts its latent representation and sends it to a classifier
- Classifier (receiver) tries to map the incoming message to a class



What this paper presents

- The authors try to define and test the *semantic consistency* of the messages in the emerging protocols
  - In a setup like below input images can be mapped to the same message →
  - Inputs that are close in the message space should be semantically similar
  - At least that's true for the natural communication
- On a higher level, they try to interpret the properties of natural languages into formal constraints, and then analyze whether the EC objectives create protocols that follow them.



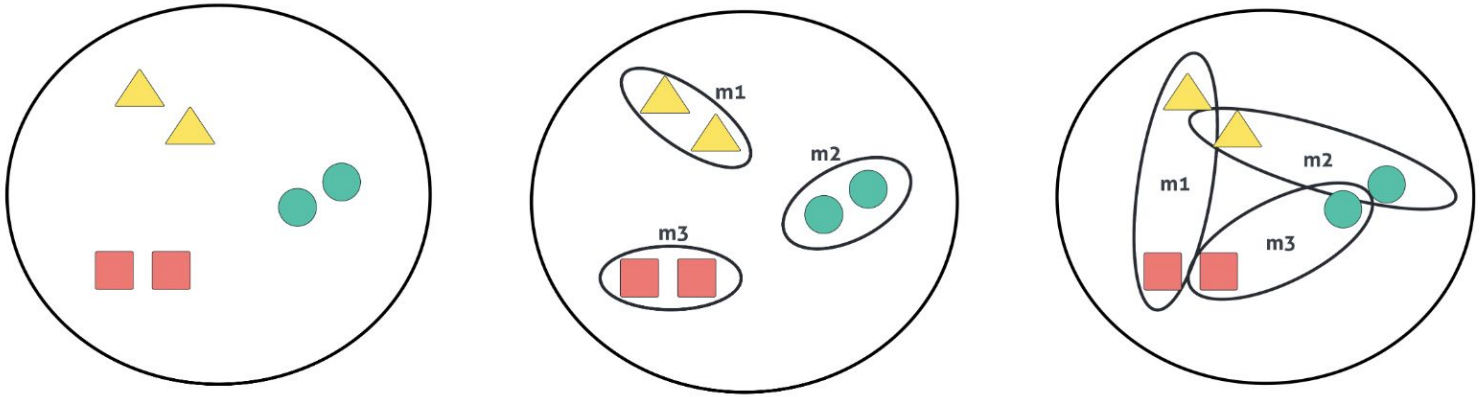
The idea of semantic consistency

Definition 3. A communication protocol S_θ is *semantically consistent* if

$$\mathop{\mathbb{E}}_{m \sim S_\theta(X)} \left[\mathrm{Var}\left[X \mid S_\theta(X) = m \right] \right] < \mathrm{Var}\left[X \right].$$

S_θ is the sender agent who makes a message *m* given input *X*.

In other words, the variance of inputs mapped to the same message should be less than the overall variance.



(a) Input space.

(b) Semantically consistent mapping.

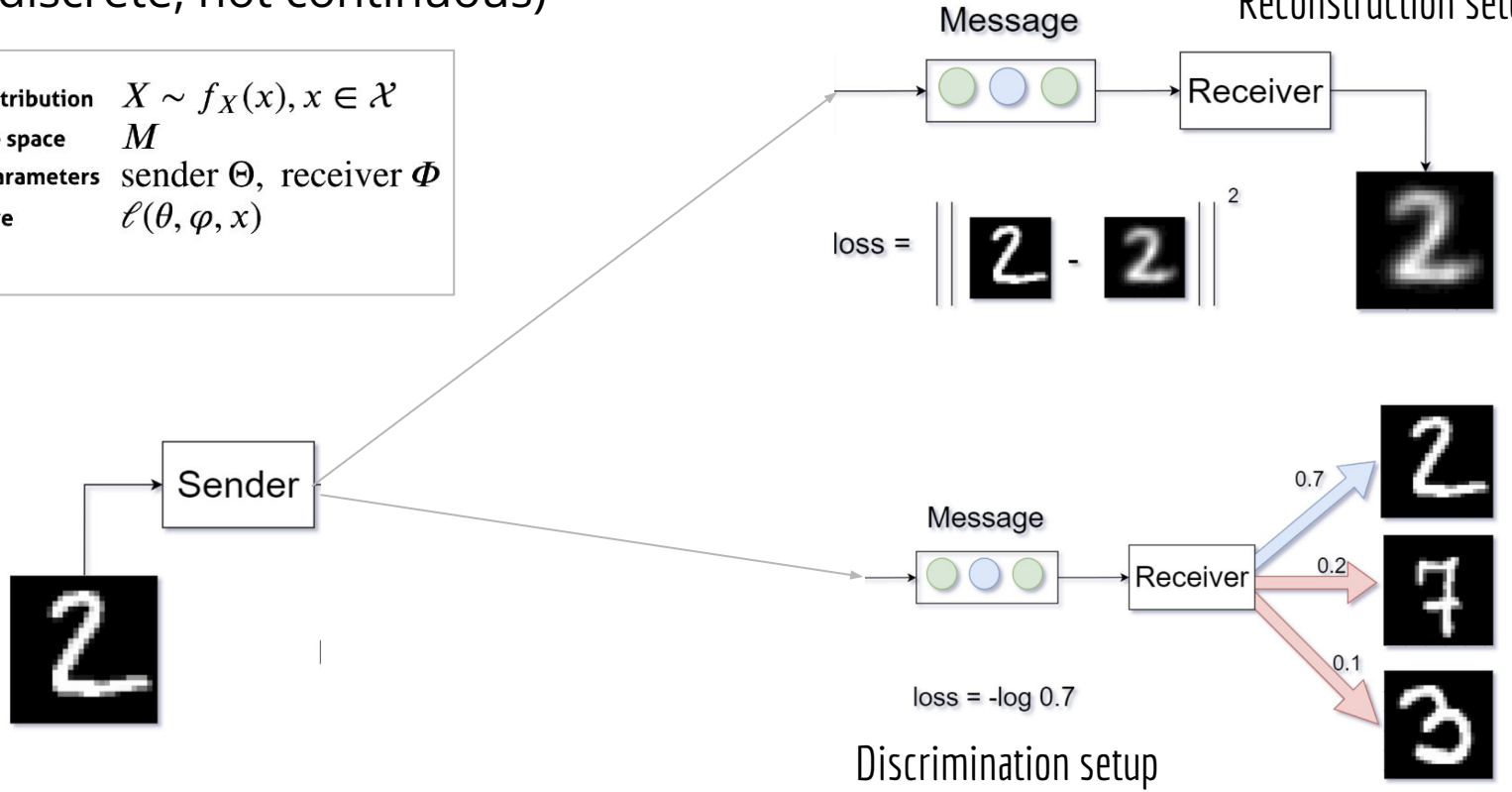
(c) Semantically inconsistent mapping.

Setups for testing semantic consistency

- They considered a simple encoder-decoder architecture (note: messages are discrete, not continuous)

EC setup

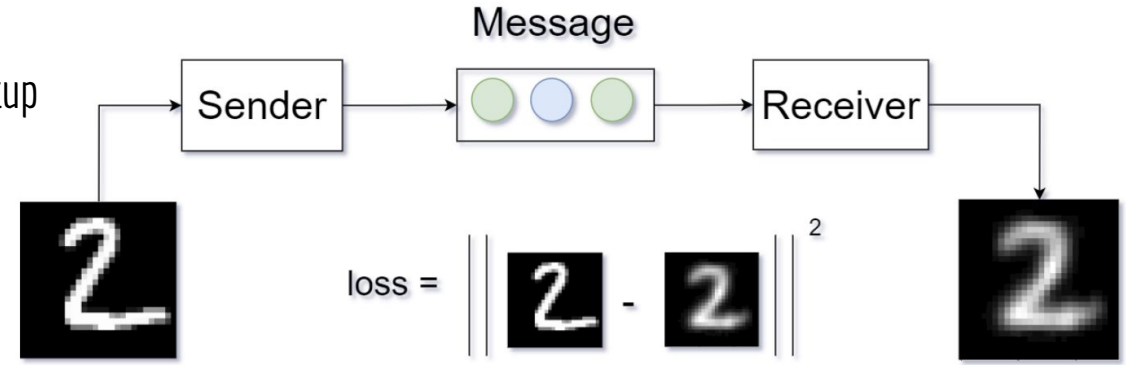
Input distribution	$X \sim f_X(x), x \in \mathcal{X}$
Message space	M
Agent parameters	sender Θ, receiver Φ
Objective	$\ell(\theta, \varphi, x)$



Reconstruction setup

Reconstruction. The receiver maps a message $S_\theta(x)$ to a prediction in the input space. The loss is the Euclidean distance between that prediction and the target x.

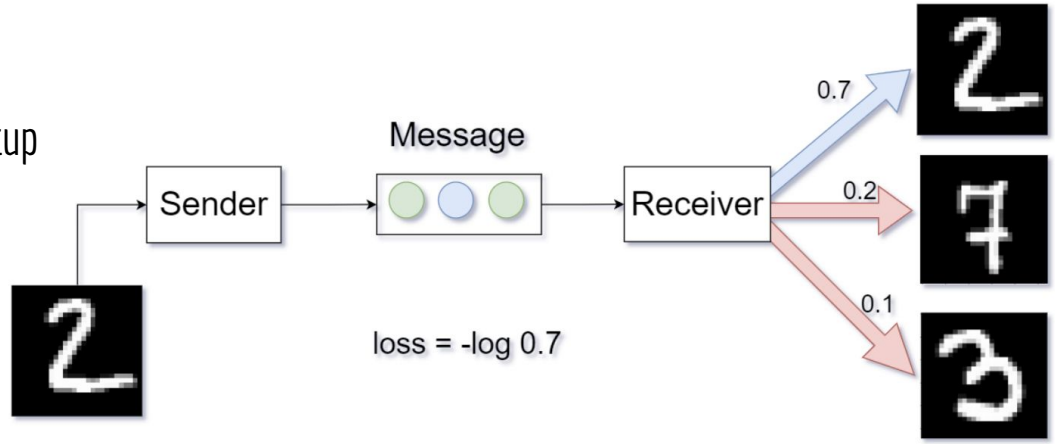
Reconstruction setup



Discrimination setup

Discrimination. In addition to a message $S_\theta(x)$, the receiver sees a set of candidates $\{x_1, \ldots, x_d\}$, which contains the target at a random position t, i.e. $x_t = x$, and the rest are $d - 1$ independently sampled distractors. The receiver outputs a probability distribution over the candidates. The loss is the negative log-likelihood of the correct position t according to this distribution, averaged over the target position and distractors.

Discrimination setup



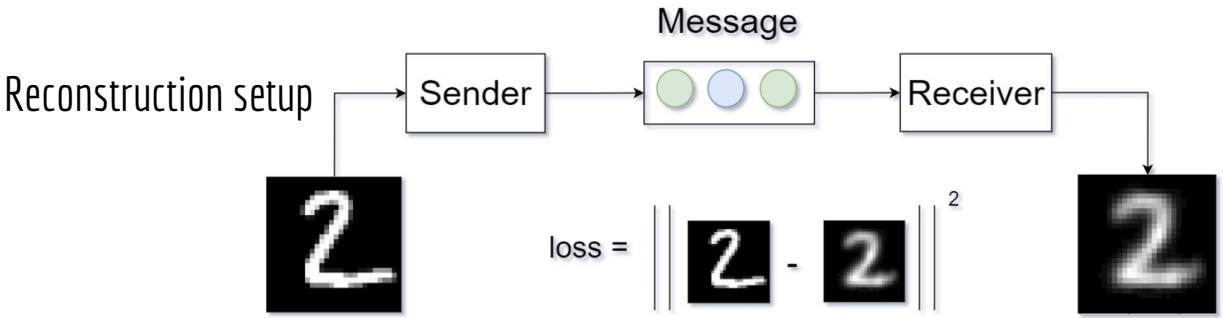
Reconstruction setup promotes semantic consistency

They prove that when a model trained to do reconstruction, the sender is trained to minimize the variance of inputs mapped to the same message

→ Such setup promotes semantically consistent communication

Lemma 5.1. [proof in page 16] *Let* $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ *be a reconstruction game. Assuming* Φ *is unrestricted, a sender* S_θ *is optimal if and only if it minimizes the following objective:*

$$\sum_{m \in M} P(S_\theta(X) = m) \cdot Var\left[X \mid S_\theta(X) = m\right] \tag{2}$$



cation protocol S_θ is *semantically consistent* if

$$\mathop{\mathbb{E}}_{m \sim S_\theta(X)} \left[\mathrm{Var}\left[X \mid S_\theta(X) = m\right]\right] < \mathrm{Var}\left[X\right].$$

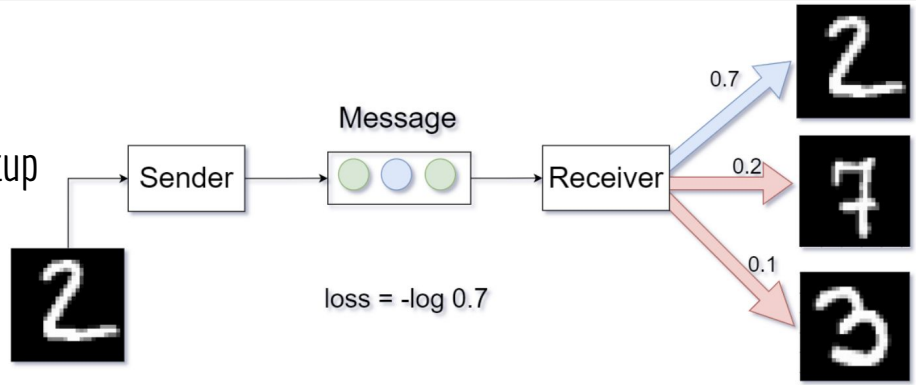
Discrimination setup DOES NOT promote semantic consistency

Semantic consistency is not a part of the discrimination objective. It seems the objective is to utilize as much of the message space as possible.

Lemma 5.3. [proof in page 17] *Let* $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ *be a* d*-candidates discrimination game. Assuming* Φ *is unrestricted, a sender* S_θ *is optimal if and only if it minimizes the following objective:*

$$\sum_{m \in M} P(S_\theta(X) = m) \cdot \mathbb{E} \log \Big(1 + \text{Binomial}\big(d-1, P(S_\theta(X) = m)\big)\Big) \tag{3}$$

And when $d = 2$ *(a single-distractor game), this simplifies into:* $\sum_{m \in M} P(S_\theta(X) = m)^2$.



cation protocol S_θ is *semantically consistent* if

$$\mathop{\mathbb{E}}_{m \sim S_\theta(X)} [\text{Var}[X \mid S_\theta(X) = m]] < \text{Var}[X].$$

Discrimination setups can be more diverse (there are no pictures past this slide)

The minimization objective of discriminative games can change depending on the setup.

Global discrimination In Appendix A.1, we analyze a version of the discrimination game where the receiver outputs a distribution over the entire data rather than a small set of candidates. We find that optimal communication protocols in this setting maximize mutual information between the inputs and messages, as shown in Rita et al. [42].

Supervised discrimination Appendix A.2 explores a variant of the discrimination game that incorporates labels by selecting distractors with labels different from the target. We find that optimal communication strategies in this setting are both diverse (the sender is encouraged to output messages uniformly) and pure (labels distribute with low entropy given a message).

Classification discrimination In Appendix A.3, we consider a format of the discrimination game where the target is excluded from the candidate set, requiring the receiver to identify a candidate that matches the target's label. We find that optimal solutions in this setup maximize mutual information between messages and labels.

Semantic consistency can be turned into a stronger metric

Semantic consistency, as defined in this paper, is a fairly loose property.

E.g., 2 groups of similar inputs can be mapped to very different messages, yet the protocol will be semantically consistent.

Definition 3. A communication protocol S_θ is *semantically consistent* if

$$\mathop{\mathbb{E}}_{m \sim S_\theta(X)} \left[\mathrm{Var}\left[X \mid S_\theta(X) = m \right] \right] < \mathrm{Var}\left[X \right].$$

So they introduce a concept of spatial meaningfulness.

Spatial meaningfulness

For a protocol to be spatially meaningful it must map similar-but-somewhat-different inputs into a close neighborhood in the message space.

Definition 4. For $\varepsilon_0 \geq \varepsilon_M$, a communication protocol S_θ is ε_0-*spatially meaningful* if $\forall 0 < \varepsilon \leq \varepsilon_0$

$$\mathop{\mathbb{E}}_{x_1,x_2\sim X}\left[\|x_1 - x_2\|^2 \mid \|S_\theta(x_1) - S_\theta(x_2)\| \leq \varepsilon\right] < \mathop{\mathbb{E}}_{x_1,x_2\sim X}\left[\|x_1 - x_2\|^2\right]$$

A communication protocol S_θ is *spatially meaningful* if this definition holds for $\varepsilon_0 = \varepsilon_M$.

Semantic consistency only requires similar inputs to be mapped to the same message.

Definition 3. A communication protocol S_θ is *semantically consistent* if

$$\mathop{\mathbb{E}}_{m\sim S_\theta(X)}\left[\mathrm{Var}\left[X \mid S_\theta(X) = m\right]\right] < \mathrm{Var}\left[X\right].$$

Spatial meaningfulness and different setups

The authors prove that the conclusions they got for semantic consistency are the same for spatial meaningfulness:

Theorem 6.1. [proof in page 21] *Let $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$ be a reconstruction game, let $\varepsilon_0 \geq \varepsilon_M$ and let $\varphi \in \Phi$ such that R_φ is (X, M, ε_0)-simple and non-degenerate. Every sender that is synchronized with R_φ is ε_0-spatially meaningful.*

Theorem 6.2. [proof in page 22] *There exists a discrimination game $(\mathcal{X}, f_X, M, \Theta, \Phi, \ell)$, $\varepsilon_0 \geq \varepsilon_M$ and a receiver $\varphi \in \Phi$ which is (X, M, ε_0)-simple and non-degenerate, where a synchronized sender matching R_φ is not ε-spatially meaningful for any ε.*

This is all great, but is it proved empirically?

They trained a few encoder-decoder models under different setups.

Unique messages — the number of unique messages made by the sender agent (the messages are discrete, so they can be counted)

- There is a discrepancy here — theory predicted that the discrimination setup should utilize more of the message space, yet it falls behind the reconstruction results.

Table 2: Empirical results on MNIST, averaged over three randomly initialized training runs.

				Msg Var ↓	
EC setup	Uniqe Msgs	Disc. accuracy ↑	TopSim ↑	Trained	Rand
Reconstruction	2523.66 ±30.0	88.00 ±0.6	0.365 ±0.042	371.63 ±2.1	1029.57 ±13.9
Discrimination	402.00 ±70.4	78.76 ±10.4	0.360 ±0.036	1226.14 ±42.4	1784.59 ±24.3
Supervised disc.	287.33 ±28.5	87.10 ±5.4	0.269 ±0.044	1381.72 ±41.6	1821.53 ±12.2

Discrimination accuracy — basically an accuracy. In the reconstruction setup the model decision was derived based on the distance between the reconstructed image and the candidates.

- On shapes the discrimination setup works better. On digits it's the same.

Table 1: Empirical results on Shapes, averaged over five randomly initialized training runs.

EC setup	Uniqe Msgs	Disc. accuracy ↑	TopSim ↑	Msg Var ↓	
				Trained	Rand
Reconstruction	306.60 ±28.52	31.64 ±2.51	0.34 ±0.02	1334.38 ±78.05	2554.77 ±108.19
Discrimination	251.60 ±29.53	61.96 ±4.78	0.09 ±0.01	2280.24 ±157.38	2793.65 ±115.45

Table 2: Empirical results on MNIST, averaged over three randomly initialized training runs.

EC setup	Uniqe Msgs	Disc. accuracy ↑	TopSim ↑	Msg Var ↓	
				Trained	Rand
Reconstruction	2523.66 ±30.0	88.00 ±0.6	0.365 ±0.042	371.63 ±2.1	1029.57 ±13.9
Discrimination	402.00 ±70.4	78.76 ±10.4	0.360 ±0.036	1226.14 ±42.4	1784.59 ±24.3
Supervised disc.	287.33 ±28.5	87.10 ±5.4	0.269 ±0.044	1381.72 ±41.6	1821.53 ±12.2

Topological similarity — evaluates the correlation between distances in the input space and the corresponding distances in the message space.

- Kind of like semantic consistency, but it takes into account all pairs; semantic consistency only cares about pairs from the same messages

Table 1: Empirical results on Shapes, averaged over five randomly initialized training runs.

EC setup	Uniqe Msgs	Disc. accuracy ↑	TopSim ↑	Msg Var ↓ Trained	Msg Var ↓ Rand
Reconstruction	306.60 ±28.52	31.64 ±2.51	0.34 ±0.02	1334.38 ±78.05	2554.77 ±108.19
Discrimination	251.60 ±29.53	61.96 ±4.78	0.09 ±0.01	2280.24 ±157.38	2793.65 ±115.45

Table 2: Empirical results on MNIST, averaged over three randomly initialized training runs.

EC setup	Uniqe Msgs	Disc. accuracy ↑	TopSim ↑	Msg Var ↓ Trained	Msg Var ↓ Rand
Reconstruction	2523.66 ±30.0	88.00 ±0.6	0.365 ±0.042	371.63 ±2.1	1029.57 ±13.9
Discrimination	402.00 ±70.4	78.76 ±10.4	0.360 ±0.036	1226.14 ±42.4	1784.59 ±24.3
Supervised disc.	287.33 ±28.5	87.10 ±5.4	0.269 ±0.044	1381.72 ±41.6	1821.53 ±12.2

Message variance — a way to measure the semantic consistency.

- This is the equation to compute it
- This is a relative metric, so it needs to be compared against something to make sense.

$$\frac{1}{2N}\sum_{m\in M}\frac{1}{|[m]|}\sum_{x_1,x_2\in[m]}\|x_1-x_2\|^2$$

- The authors compared a trained sender against a sender that was assigning messages to inputs randomly

EC setup	Msg Var ↓ Trained	Rand
Reconstruction	1334.38 ±78.05	2554.77 ±108.19
Discrimination	2280.24 ±157.38	2793.65 ±115.45

Shapes

EC setup	Msg Var ↓ Trained	Rand
Reconstruction	371.63 ±2.1	1029.57 ±13.9
Discrimination	1226.14 ±42.4	1784.59 ±24.3
Supervised disc.	1381.72 ±41.6	1821.53 ±12.2

MNIST

That's all!

Conclusions, limitations, topics for discussion:

- Different training objectives induce different communication between ML model layers
- Euclidean distance is used as a measure of semantic similarity of messages, but it is debatable
  - last time I presented a representational alignment paper, it had some metrics for measuring the alignment of messages/whole models
- Their analysis assumes that the messages are discrete
  - So it's not immediately clear if all this is true for the floating-point vectors
- I wonder how similar the discrimination and classification setups are.
  - I think classification is a subset of discrimination. What do you think?
  - They considered a classification discrimination setup, where the target and 9 distractors were sampled according to their label. It is close to ML classification, but it's not quite it.