Studying the syntax and semantics of emergent languages

Timothée Bernard (LLF, Université Paris Cité)

2023/11/09, Research seminar in language technology, University of Helsinki

Prologue

Personal anecdote

- Agents in a 2D world, behaviour encoded by an neural network (the "DNA"), variation during reproduction $\rightarrow$ optimisation via natural selection/evolution

Personal anecdote

- Agents in a 2D world, behaviour encoded by an neural network (the “DNA”), variation during reproduction $\rightarrow$ optimisation via natural selection/evolution
- Can language emerge?

Personal anecdote

- Agents in a 2D world, behaviour encoded by an neural network (the “DNA”), variation during reproduction $\rightarrow$ optimisation via natural selection/evolution
- Can language emerge?
- First experiment around language (a signalling game).

Personal anecdote

- Agents in a 2D world, behaviour encoded by an neural network (the “DNA”), variation during reproduction $\rightarrow$ optimisation via natural selection/evolution
- Can language emerge?
- First experiment around language (a signalling game).
- But no interesting language, because of trivial winning strategies.

Personal anecdote

- Agents in a 2D world, behaviour encoded by an neural network (the “DNA”), variation during reproduction $\rightarrow$ optimisation via natural selection/evolution
- Can language emerge?
- First experiment around language (a signalling game).
- But no interesting language, because of trivial winning strategies.
- ACL in Firenze, Timothee Mickus informs me that there is a field around such questions; we decide to collaborate.

Language emergence

Game theoretical experiments about the evolution of language

- Goal: understanding of how agents can develop a language.
- Into consideration: collaboration, competition, noise, cost, benefit, evolving environment, evolving population of agents, etc.
- Language games: experimental setups designed to test hypotheses about language emergence with human or artificial agents (Kirby 2002; Kirby, Cornish, and Smith 2008).

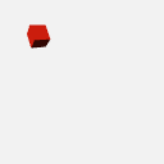
Signalling games are cooperative language games

- Signalling game (Lewis 1969):
  - two agents: a sender and a receiver,
  - a mapping from world state to correct action,
  - at each round:
    1. a world state is selected, only the sender knows which,
    2. the sender produces a signal, sent to the receiver,
    3. the receiver selects an action,
    4. both are informed of whether it is the correct action $\rightarrow$ common goal
- Neural implementations are possible (e.g. Lazaridou, Peysakhovich, and Baroni 2017).

Making signalling game work

Structured images with clear non-trivial semantics

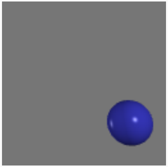
- Artificial dataset of images (Bernard and Mickus 2023):
  - object on a grey background (with varying shade),
  - variation: shape (*cube* or *sphere*), size (*large* or *small*), colour (*red* or *blue*), and vertical (*top* or *bottom*) and horizontal position (*left* or *right*).
  - $\rightarrow$ 32 categories (background is irrelevant)
- Examples:



$\in C_1$

$\in C_2$

$\in C_3$



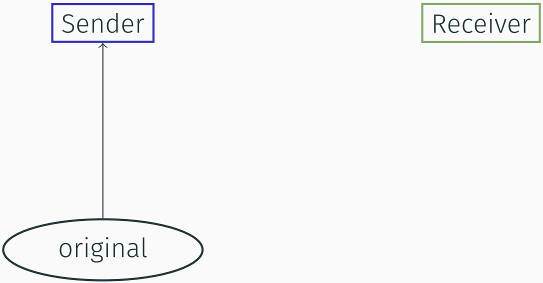
$\in C_3$

A round of the game

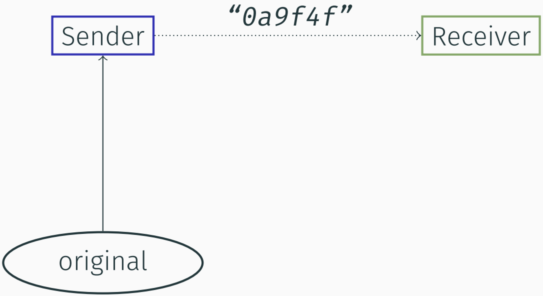
Sender

Receiver

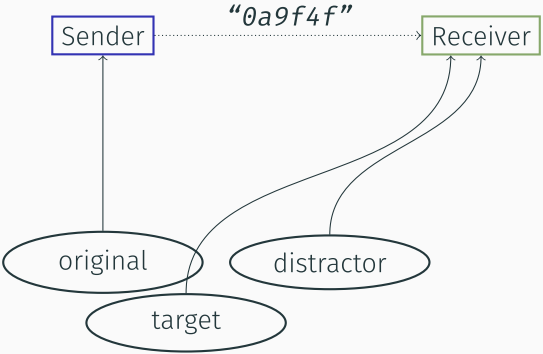
A round of the game



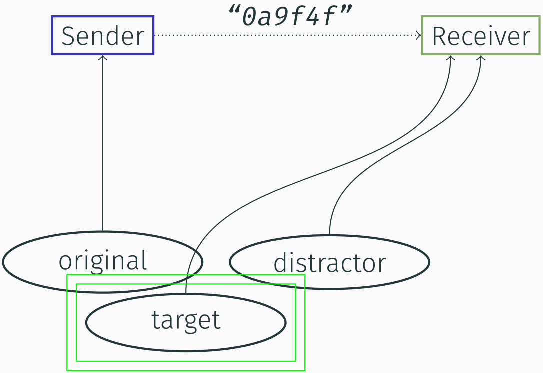
A round of the game



A round of the game



A round of the game



Architecture

- Sender:
  - CNN enc.: img $\mapsto$ vec
  - LSTM dec.: vec $\mapsto$ msg
- $|\text{msg}| \leq 10$, $|\text{alphabet}| = 16$
- Receiver:
  - CNN enc.: img $\mapsto$ vec
  - LSTM enc.: msg $\mapsto$ vec
  - dot product: (img vec, msg vec) $\mapsto$ compatibility score
  - softmax: compatibility scores $\mapsto$ probability distribution

Training

- Sender:

Training

- Sender:
  - ~~standard supervised training~~

Training

- Sender:
    - ~~standard supervised training~~
    - REINFORCE (Williams 1992):
        - one action per symbol generated, $(a_t)_{1 \leq t \leq |\mathrm{msg}|}$
        - same reward r for all actions: $r = 1$ if success, $r = -1$ otherwise.
        - loss:

$$\mathcal{L} = -r \sum_{1 \leq t \leq |\mathrm{msg}|} \log p(a_t)$$

Training

- Sender:
  - ~~standard supervised training~~
  - REINFORCE (Williams 1992):
    - one action per symbol generated, $(a_t)_{1 \leq t \leq |\text{msg}|}$
    - same reward r for all actions: $r = 1$ if success, $r = -1$ otherwise.
    - loss:

$$\mathcal{L} = -r \sum_{1 \leq t \leq |\text{msg}|} \log p(a_t)$$

- Receiver:

Training

- Sender:
  - ~~standard supervised training~~
  - REINFORCE (Williams 1992):
    - one action per symbol generated, $(a_t)_{1 \leq t \leq |\mathrm{msg}|}$
    - same reward r for all actions: $r = 1$ if success, $r = -1$ otherwise.
    - loss:

$$\mathcal{L} = -r \sum_{1 \leq t \leq |\mathrm{msg}|} \log p(a_t)$$

- Receiver:
  - (standard supervised training)
  - REINFORCE; one action (pointing), same r.

A bit of work is required for reliable training

- category(original) $=$ category(target) $\neq$ category(distractor)

A bit of work is required for reliable training

- category(original) $=$ category(target) $\neq$ category(distractor)
- small grid-search for the learning rate

A bit of work is required for reliable training

- category(original) $=$ category(target) $\neq$ category(distractor)
- small grid-search for the learning rate
- 10 runs $\times$ (100 000 batches $\times$ 128 instances) $\rightarrow$ 3 runs fail

A bit of work is required for reliable training

- category(original) $=$ category(target) $\neq$ category(distractor)
- small grid-search for the learning rate
- 10 runs $\times$ (100 000 batches $\times$ 128 instances) $\rightarrow$ 3 runs fail
- Super effective: + *baseline term* in the loss,

$$\mathcal{L} = -(r - b) \sum_{1 \leq t \leq |\mathrm{msg}|} \log p(a_t)$$

$b =$ average of r over the last 1000 batches

A bit of work is required for reliable training

- category(original) $=$ category(target) $\neq$ category(distractor)
- small grid-search for the learning rate
- 10 runs $\times$ (100 000 batches $\times$ 128 instances) $\rightarrow$ 3 runs fail
- Super effective: + *baseline term* in the loss,

$$\mathcal{L} = -(r - b) \sum_{1 \leq t \leq |\mathrm{msg}|} \log p(a_t)$$

  $b =$ average of r over the last 1000 batches
- 10 runs $\rightarrow$ 0 run fails

Main evaluation metric

- Categorical communication efficiency:

$$\text{c.c.e.} = \mathop{\mathbb{E}}_{\substack{\text{categories } c \neq c' \\ I_o, I_t \in c,\, I_d \in c'}} [p(I_t \mid I_t, I_d, msg_{I_o})]$$

Main evaluation metric

- Categorical communication efficiency:

$$\mathrm{c.c.e.} = \mathop{\mathbb{E}}_{\substack{\text{categories } c \neq c' \\ I_o, I_t \in c,\, I_d \in c'}} [p(I_t \mid I_t, I_d, msg_{I_o})]$$

- Dataset:
  - for each category: training and evaluation images
  - partition: *base* (train.+eval.) and *generalisation* (eval. only) categories

Main evaluation metric

- Categorical communication efficiency:

$$\text{c.c.e.} = \mathop{\mathbb{E}}_{\substack{\text{categories } c \neq c' \\ I_o, I_t \in c,\, I_d \in c'}} [p(I_t \mid I_t, I_d, msg_{I_o})]$$

- Dataset:
  - for each category: training and evaluation images
  - partition: *base* (train.+eval.) and *generalisation* (eval. only) categories
- Twist: 2 base cat. differ by at least two features (same for gen.).
  $\rightarrow$ one feature can be ignored

Agents learn (very) well, depending on many aspects of the game

- With baseline term:
    - c.c.e.: 0.963;
    - base c.c.e.: 0.982; gen. c.c.e.: 0.980; mixed c.c.e.: 0.950.

  (max. c.c.e., median over all runs)

Agents learn (very) well, depending on many aspects of the game

- With baseline term:
  - c.c.e.: 0.963;
  - base c.c.e.: 0.982; gen. c.c.e.: 0.980; mixed c.c.e.: 0.950.

  (max. c.c.e., median over all runs)
- Better perf. when training with *hard distractors*:
  - c.c.e.: 0.981;
  - base c.c.e.: 0.999; gen. c.c.e.: 0.997; mixed c.c.e.: 0.967.
- With pretraining, regularisation, etc.; higher perf. is possible (Bernard and Mickus 2023).

Agents are lazy when they can

- mixed c.c.e.: 0.967 $\rightarrow$ compatible with one feature (e.g. size) being systematically ignored

Agents are lazy when they can

- mixed c.c.e.: 0.967 $\rightarrow$ compatible with one feature (e.g. size) being systematically ignored
- Is it the case?

Agents are lazy when they can

- mixed c.c.e.: 0.967 $\rightarrow$ compatible with one feature (e.g. size) being systematically ignored
- Is it the case? Not exactly, but almost.
- Agents tend to focus significantly less on shape (`cube`|`sphere`).

Agents are lazy when they can

- mixed c.c.e.: 0.967 $\rightarrow$ compatible with one feature (e.g. size) being systematically ignored
- Is it the case? Not exactly, but almost.
- Agents tend to focus significantly less on shape (*`cube`*|*`sphere`*).
- How do we know?

Grammar in emergent languages

We'd like to understand how compositionality emerges

- Long-term goal: human language-like features in emergent languages.

We'd like to understand how compositionality emerges

- Long-term goal: human language-like features in emergent languages.
- Compositional language (Carnap 1947; Montague 1974):
    - syntax,
    - semantics,
    - principle of compositionality: the meaning of a compound structure is a function only of the meaning of its (direct) components and of the syntactic rule that binds them.

Compositionality is what makes a language productive

- Some consequences:
    - replacing a component with a paraphrase (irrespectively of their structure) has no impact on the meaning of the whole,
    - semantics = one semantic combination rule per syntactic rule + lexical semantics,
    - once one knows the meaning of a lexical item, they can use it in any structure/context.

Compositionality is what makes a language productive

- Some consequences:
  - replacing a component with a paraphrase (irrespectively of their structure) has no impact on the meaning of the whole,
  - semantics = one semantic combination rule per syntactic rule + lexical semantics,
  - once one knows the meaning of a lexical item, they can use it in any structure/context.
- $\Rightarrow$ *productivity* of natural language, which “can (in Humboldt’s words) ‘make infinite use of finite means’” (Chomsky 1965).

Compositionality is what makes a language productive

- Some consequences:
  - replacing a component with a paraphrase (irrespectively of their structure) has no impact on the meaning of the whole,
  - semantics = one semantic combination rule per syntactic rule + lexical semantics,
  - once one knows the meaning of a lexical item, they can use it in any structure/context.
- $\Rightarrow$ *productivity* of natural language, which "can (in Humboldt's words) 'make infinite use of finite means'" (Chomsky 1965).
- Interpreted formal languages are usually compositional (e.g. simply-typed λ-calculus, first-order logic, positional numeral systems, the language of arithmetic expressions).

Compositionality conflicts with contextuality and holism

- Stronger or less depending on the kind of composition rules and semantic entries that one is ready to accept.
- Can be seen as a methodological principle. E.g.,
  - to draw the line between semantics and pragmatics;
  - to define multiword expressions (*to kick the bucket*, *ivory tower*).
- Not compositional: any algorithm of the form

$$\text{msg} \mapsto \begin{cases} \text{case string}_1 \Rightarrow \text{meaning}_1 \\ \text{case string}_2 \Rightarrow \text{meaning}_2 \\ \dots \end{cases}$$

- Our goal:
    1. observe a compositional language;

- Our goal:
    1. observe a compositional language;
    2. know when we do.

Grammar in emergent languages

Toward measuring compositionality

There is no easy-to-use formula to quantify compositionality

- *Meaning-form correlation* (or “topographic similarity”; Brighton and Kirby 2006) is interesting but not perfect (Mickus, Bernard, and Paperno 2020).

There is no easy-to-use formula to quantify compositionality

- *Meaning-form correlation* (or "topographic similarity"; Brighton and Kirby 2006) is interesting but not perfect (Mickus, Bernard, and Paperno 2020).
- Bernard and Mickus (2023):
    - c.c.e. and variants,

There is no easy-to-use formula to quantify compositionality

- *Meaning-form correlation* (or "topographic similarity"; Brighton and Kirby 2006) is interesting but not perfect (Mickus, Bernard, and Paperno 2020).
- Bernard and Mickus (2023):
    - c.c.e. and variants,
    - *abstractness*,
    - *scrambling resistance*,
    - *semantic probes*.

Abstractness: images-specific information

$$\text{abs.} = 2 \underset{\substack{\text{category } c \\ I_o, I_t \in c}}{\mathbb{E}} [p(I_t \mid I_o, I_t, msg_{I_o})]$$

- Quantifies sensitivity to intra-category differences.

Abstractness: images-specific information

$$\text{abs.} = 2 \mathop{\mathbb{E}}_{\substack{\text{category } c \\ I_o, I_t \in c}} [p(I_t \mid I_o, I_t, msg_{I_o})]$$

- Quantifies sensitivity to intra-category differences.
- Natural language?
  - Hard to say (categories for sentences?).
  - But think about how the same caption may suit two different pictures.

Scrambling resistance: bag-of-words semantics

$$\text{s.r.} = \mathop{\mathbb{E}}_{\substack{\text{categories } c \neq c' \\ I_o, I_t \in c,\ I_d \in c' \\ \text{permutation } \sigma}} \left[\frac{p(I_t \mid I_t, I_d, \sigma(msg_{I_o}))}{p(I_t \mid I_t, I_d, msg_{I_o})} \right]$$

- Quantifies sensitivity to symbol order.

Scrambling resistance: bag-of-words semantics

$$\text{s.r.} = \mathop{\mathbb{E}}_{\substack{\text{categories } c \neq c' \\ I_o, I_t \in c,\, I_d \in c' \\ \text{permutation } \sigma}} \left[\frac{p(I_t \mid I_t, I_d, \sigma(msg_{I_o}))}{p(I_t \mid I_t, I_d, msg_{I_o})} \right]$$

- Quantifies sensitivity to symbol order.
- Natural language?
    - Below 1 in general (*Achilles beat the turtle* vs *the turtle beat Achilles*, *fake Malaysian ivory* vs *Malaysian fake ivory*).
    - But high in our case (*cube on blue the left corner big a image top of*).

Semantic probes: high-level concept

- Messages are converted into bag-of-symbols vectors ($\in \mathbb{N}^{16}$).
- For each of the five features, we train a decision tree to predict the corresponding value.

These metrics tell us a coherent story

- Baseline term, hard distractors, no pretraining (median values):

abs.	s.r.	semantic probes				
		shape	size	colour	h. pos.	v. pos.
0.997	0.903	0.531	0.992	0.999	0.999	0.999

- Shape:
    - either encoded in an exotic way
    - or ignored. $\rightarrow$ coherent with the c.c.e. values

These metrics tell us a coherent story

- Baseline term, hard distractors, no pretraining (median values):

abs.	s.r.	semantic probes				
		shape	size	colour	h. pos.	v. pos.
0.997	0.903	0.531	0.992	0.999	0.999	0.999

- Shape:
    - either encoded in an exotic way
    - or ignored. $\rightarrow$ coherent with the c.c.e. values
- With (auto-encoder) pretraining of the vision CNN: up to 0.651 for shape (and higher c.c.e.).

Grammar in emergent languages

Toward producing compositionality

Concepts emerge because they are needed

- Complex languages are made necessary by complex environments.
  $\rightarrow$ need for structured images as stimuli

Concepts emerge because they are needed

- Complex languages are made necessary by complex environments.
  $\rightarrow$ need for structured images as stimuli
- Maybe not enough...
- Hypothesis: Without pressure towards high-level semantics, agents in a signalling game communicate about low-level, unstructured features of their stimuli.
  $\rightarrow$ in our case, e.g. background colour

Concepts emerge because they are needed

- Complex languages are made necessary by complex environments.
  $\rightarrow$ need for structured images as stimuli
- Maybe not enough...
- Hypothesis: Without pressure towards high-level semantics, agents in a signalling game communicate about low-level, unstructured features of their stimuli.
  $\rightarrow$ in our case, e.g. background colour
- Something of the sort has been observed by Bouchacourt and Baroni (2018).

Overview of the experiments

- Three setups around this hypothesis:

Overview of the experiments

- Three setups around this hypothesis:
    1. topline models: category(original) $=$ category(target) (as before)

Overview of the experiments

- Three setups around this hypothesis:
    1. topline models: category(original) $=$ category(target) (as before)
    2. baseline models: original $=$ target
       $\rightarrow$ we don't expect high-level semantics to emerge

Overview of the experiments

- Three setups around this hypothesis:
    1. topline models: category(original) $=$ category(target) (as before)
    2. baseline models: original $=$ target
       $\rightarrow$ we don't expect high-level semantics to emerge
    3. adversarial models: original $=$ target, + a third agent is introduced in order to foster the emergence of high-level semantics

Overview of the experiments

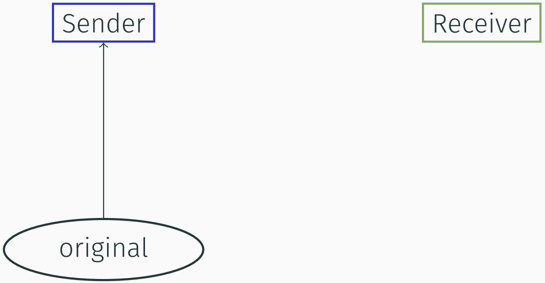
- Three setups around this hypothesis:
  1. topline models: category(original) $=$ category(target) (as before)
  2. baseline models: original $=$ target
     $\rightarrow$ we don't expect high-level semantics to emerge
  3. adversarial models: original $=$ target, + a third agent is introduced in order to foster the emergence of high-level semantics
- (Categories are partitioned differently so as to ensure that some pairs of base categories differ by only one feature.)

Baseline models

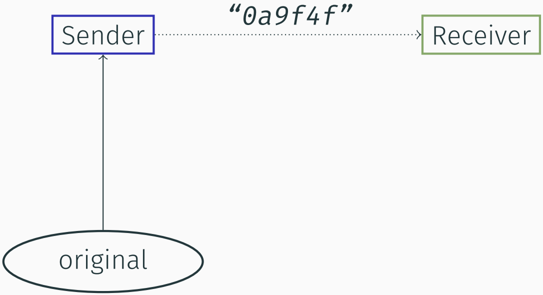
Sender

Receiver

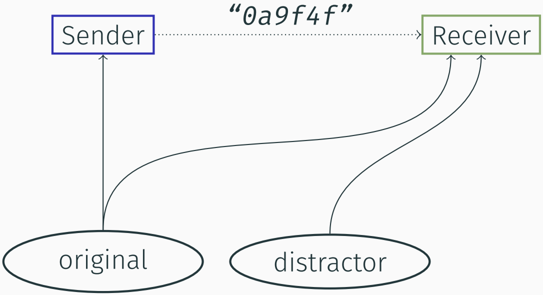
Baseline models



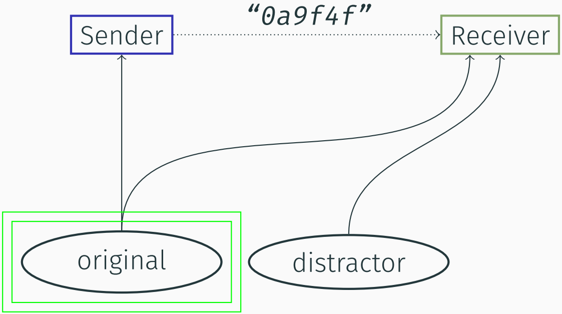
Baseline models



Baseline models



Baseline models



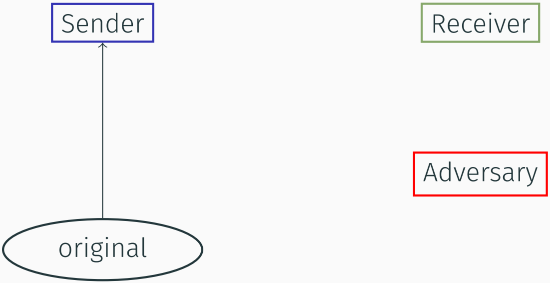
Adversary models

Sender

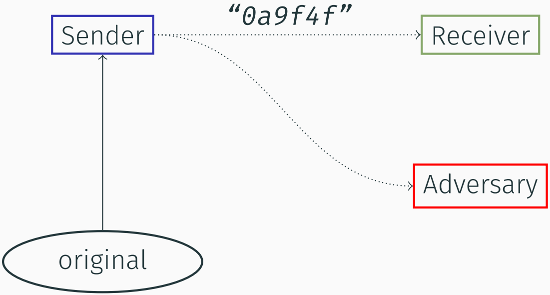
Receiver

Adversary

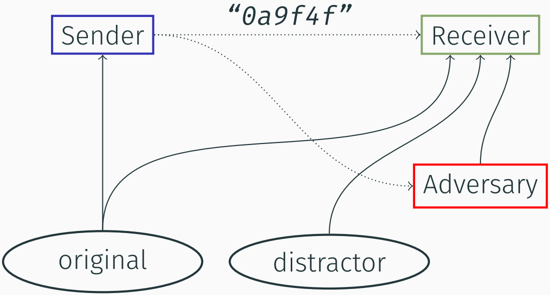
Adversary models



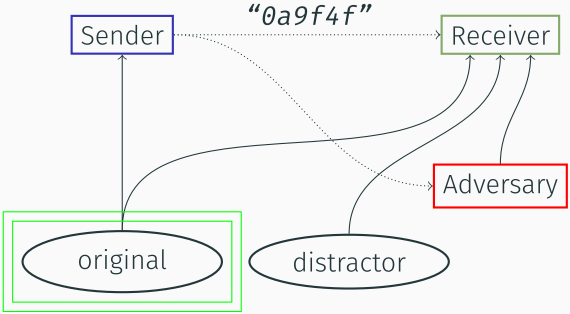
Adversary models



Adversary models



Adversary models



Architecture of the adversary

- Adversary:
  - LSTM enc.: msg $\mapsto$ vec
  - CNN dec.: vec $\mapsto$ img

Training with the adversary

- Sender: REINFORCE, $r = 1$ if the receiver retrieves the original against the distractor.
- Receiver: standard supervised learning, original against distractor and adversary.
- Adversary: adversarial training (Goodfellow et al. 2014), uses the receiver's loss to maximise the probability of the adversary image against the original.

Training with the adversary

- Sender: REINFORCE, $r = 1$ if the receiver retrieves the original against the distractor.
- Receiver: standard supervised learning, original against distractor and adversary.
- Adversary: adversarial training (Goodfellow et al. 2014), uses the receiver's loss to maximise the probability of the adversary image against the original.

Intuition

sender communicates a low-level feature $\rightarrow$ adversary easily learns to replicate it $\rightarrow$ receiver tries to rely on other features $\rightarrow$ sender tries to communicate about other features

What the metrics tell us about the messages' content

- Low abstractness $\times$ low c.c.e.: only image-level information that does not generalise to other images of the same category.

What the metrics tell us about the messages' content

- Low abstractness $\times$ low c.c.e.: only image-level information that does not generalise to other images of the same category.
- Low abstractness $\times$ high c.c.e.: at least image-specific information (might be enough to achieve high c.c.e.).

What the metrics tell us about the messages' content

- Low abstractness $\times$ low c.c.e.: only image-level information that does not generalise to other images of the same category.
- Low abstractness $\times$ high c.c.e.: at least image-specific information (might be enough to achieve high c.c.e.).
- High abstractness $\times$ low c.c.e.: no image-specific nor category-level information. $\rightarrow$ failed run

What the metrics tell us about the messages' content

- Low abstractness $\times$ low c.c.e.: only image-level information that does not generalise to other images of the same category.
- Low abstractness $\times$ high c.c.e.: at least image-specific information (might be enough to achieve high c.c.e.).
- High abstractness $\times$ low c.c.e.: no image-specific nor category-level information. $\rightarrow$ failed run
- High abstractness $\times$ high c.c.e.: no image-specific information but then category-level information.

What the metrics tell us about the messages' content

- Low abstractness $\times$ low c.c.e.: only image-level information that does not generalise to other images of the same category.
- Low abstractness $\times$ high c.c.e.: at least image-specific information (might be enough to achieve high c.c.e.).
- High abstractness $\times$ low c.c.e.: no image-specific nor category-level information. $\rightarrow$ failed run
- High abstractness $\times$ high c.c.e.: no image-specific information but then category-level information.

- + semantic probes to complete the picture (if high s.r.)

Results

- 40 runs of each models; trained on 1 000 000 batches.
- Metrics obtained at max c.c.e., median over the 40 runs.
- REINFORCE with baseline term; auto-encoder pretraining of the vision CNNs.

Model	c.c.e.	abs.	s.r.	semantic probes				
				shape	size	colour	h. pos.	v. pos.
Topline	0.986	0.992	0.822	0.642	0.996	0.998	0.999	0.999
Baseline	0.992	0.853	0.949	0.818	0.993	0.993	0.999	0.999
Adversarial	0.991	0.876	0.937	0.806	0.995	0.992	0.999	0.999

Results

- 40 runs of each models; trained on 1 000 000 batches.
- Metrics obtained at max c.c.e., median over the 40 runs.
- REINFORCE with baseline term; auto-encoder pretraining of the vision CNNs.

Model	c.c.e.	abs.	s.r.	semantic probes				
				shape	size	colour	h. pos.	v. pos.
Topline	0.986	0.992	0.822	0.642	0.996	0.998	0.999	0.999
Baseline	0.992	0.853	0.949	0.818	0.993	0.993	0.999	0.999
Adversarial	0.991	0.876	0.937	0.806	0.995	0.992	0.999	0.999

- Topline models:
  - High abs. $\times$ high c.c.e.: category-level information only.
  - (still low shape? low s.r.?)

Baseline models work better that expected

Model	c.c.e.	abs.	s.r.	semantic probes				
				shape	size	colour	h. pos.	v. pos.
Topline	0.986	0.992	0.822	0.642	0.996	0.998	0.999	0.999
Baseline	0.992	0.853	0.949	0.818	0.993	0.993	0.999	0.999
Adversarial	0.991	0.876	0.937	0.806	0.995	0.992	0.999	0.999

- Baseline models:
  - Abs., c.c.e. and probe accuracy all higher than expected: category-level information and not too much image-specific information. $\rightarrow$ against our hypothesis
  - Caused by the pretraining? (we can show that pretraining helps) but is not necessary

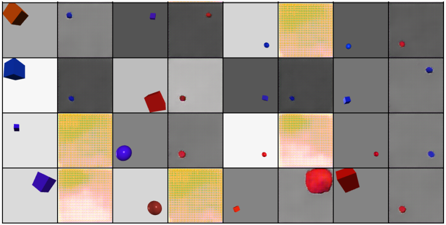
The adversary does not seem to make a big difference

Model	c.c.e.	abs.	s.r.	semantic probes				
				shape	size	colour	h. pos.	v. pos.
Topline	0.986	0.992	0.822	0.642	0.996	0.998	0.999	0.999
Baseline	0.992	0.853	0.949	0.818	0.993	0.993	0.999	0.999
Adversarial	0.991	0.876	0.937	0.806	0.995	0.992	0.999	0.999

- Adversarial models:
  - similar to baseline models
  - not working? or baseline models are already too good?

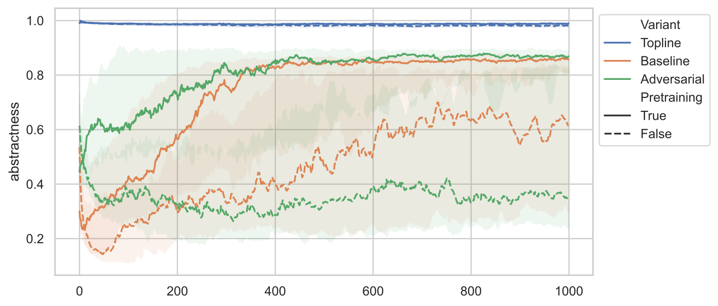
Adversary images are revealing

- Each image in an even column is an adversary image corresponding to the original image immediately on its left:



- The adversary replicates the background colour. → Communicating this information is a strategy developed by sender-receiver systems.

The adversary impacts the dynamics of learning



- Dynamics (solid green and orange lines): the adversary boosts abstractness early in the game.

Conclusion

We’ve been surprised to see category-level information emerge

- Evidence of baseline models developing high-level semantic concepts, even though this is not required.
- More training $\Rightarrow$ less sensitivity to intra-category differences.
- The agents learn the concept of background colour $\rightarrow$ easy strategy; so why?
- Maybe category(original) $\neq$ category(distractor) is enough for the agents to induce the categories.

Future work

- More effective training of the adversary $\Rightarrow$ stronger impact on the emergent language?
- Reconstruction of the emergent grammar (grammatical inference, machine translation, etc.).
- Emergence of numerical systems?
- Emergence of pragmatics?
- Use of more structured input (multiple objects, subsequent frames, natural images, etc.).

References

Carnap, Rudolf (1947). ***Meaning and Necessity: A Study in Semantics and Modal Logic.*** University of Chicago Press. ISBN: 978-0-226-09347-5. URL: *http://www.press.uchicago.edu/ucp/books/book/chicago/M/bo3638630.html*.

Chomsky, Noam (1965). ***Aspects of the Theory of Syntax.*** Massachusetts Institute of Technology. Research Laboratory of Electronics. Special technical report 11. The MIT Press. ISBN: 978-0-262-52740-8. URL: *https://www.jstor.org/stable/j.ctt17kk81z*.

- Lewis, David K. (1969). ***Convention: a philosophical study.*** eng. Cambridge, MA, USA: Harvard University Press. ISBN: 978-0-631-23257-5 978-0-631-23256-8.
- Montague, Richard (1974). ***Formal Philosophy. Selected Papers of Richard Montague.*** Ed. by Richmond H. Thomason. New Haven: Yale University Press.
- Williams, Ronald J. (May 1992). **“Simple statistical gradient-following algorithms for connectionist reinforcement learning”.** en. In: *Machine Learning* 8.3-4, pp. 229–256. ISSN: 0885-6125, 1573-0565. DOI: *10.1007/BF00992696*. URL: *https://link.springer.com/article/10.1007/BF00992696*.

- Kirby, Simon (Apr. 2002). **"Natural Language From Artificial Life".** In: *Artificial Life* 8.2, pp. 185–215. ISSN: 1064-5462. DOI: *10.1162/106454602320184248*. URL: *https://doi.org/10.1162/106454602320184248* (visited on 07/31/2019).
- Brighton, Henry and Simon Kirby (Mar. 2006). **"Understanding Linguistic Evolution by Visualizing the Emergence of Topographic Mappings".** In: *Artificial Life* 12.2, pp. 229–242. ISSN: 1064-5462. DOI: *10.1162/106454606776073323*. URL: *https://doi.org/10.1162/106454606776073323*.

- Kirby, Simon, Hannah Cornish, and Kenny Smith (Aug. 2008). "Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language". en. In: *Proceedings of the National Academy of Sciences* 105.31. 00803, pp. 10681–10686. ISSN: 0027-8424, 1091-6490. DOI: *10.1073/pnas.0707835105*. URL: *https://www.pnas.org/content/105/31/10681*.
- Goodfellow, Ian et al. (2014). "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., pp. 2672–2680. URL: *http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf*.

Lazaridou, Angeliki, Alexander Peysakhovich, and Marco Baroni (2017). **"Multi-Agent Cooperation and the Emergence of (Natural) Language".** In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. URL: *https://openreview.net/forum?id=Hk8N3Sclg*.

Bouchacourt, Diane and Marco Baroni (Oct. 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*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, pp. 981–985. DOI: *10.18653/v1/D18-1119*. URL: *https://aclanthology.org/D18-1119* (visited on 11/08/2023).

Mickus, Timothee, Timothée Bernard, and Denis Paperno (Dec. 2020). **"What Meaning-Form Correlation Has to Compose With: A Study of MFC on Artificial and Natural Language".** In: *Proceedings of the 28th International Conference on Computational Linguistics.* Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 3737–3749. URL: *https://www.aclweb.org/anthology/2020.coling-main.333*.

Bernard, Timothée and Timothee Mickus (July 2023). **"So many design choices: Improving and interpreting neural agent communication in signaling games".** In: *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, pp. 8399–8413. URL: *https://aclanthology.org/2023.findings-acl.531*.