Emergent Graphical Conventions in a Visual Communication Game

Shuwen Qiu*[1], Sirui Xie*[1], Lifeng Fan[2], Tao Gao[1], Jungseock Joo[1], Song-Chun Zhu[1,2,3], Yixin Zhu[3]

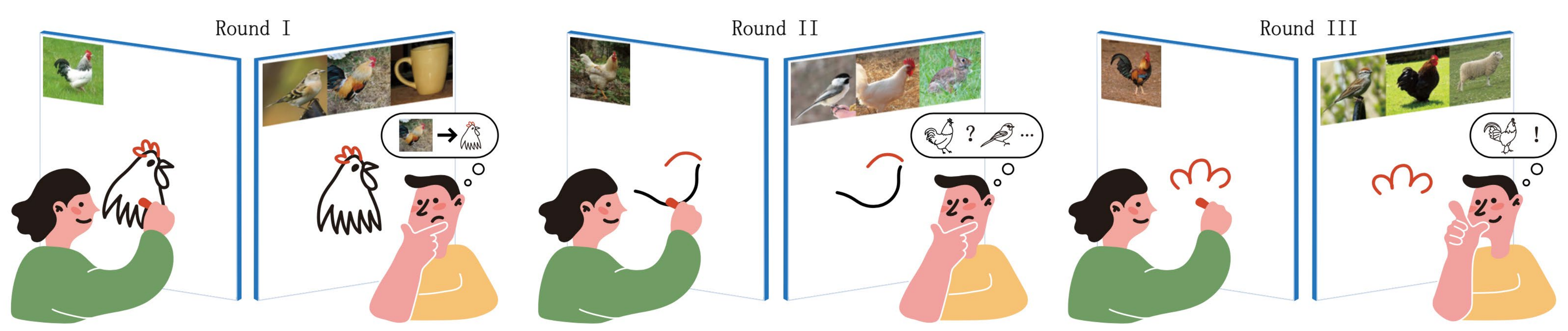
[1]University of California, Los Angeles, [2]Beijing Institute for General Artificial Intelligence, [3]Institute for Artificial Intelligence, Peking University, *Equal contribution

UCLA

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

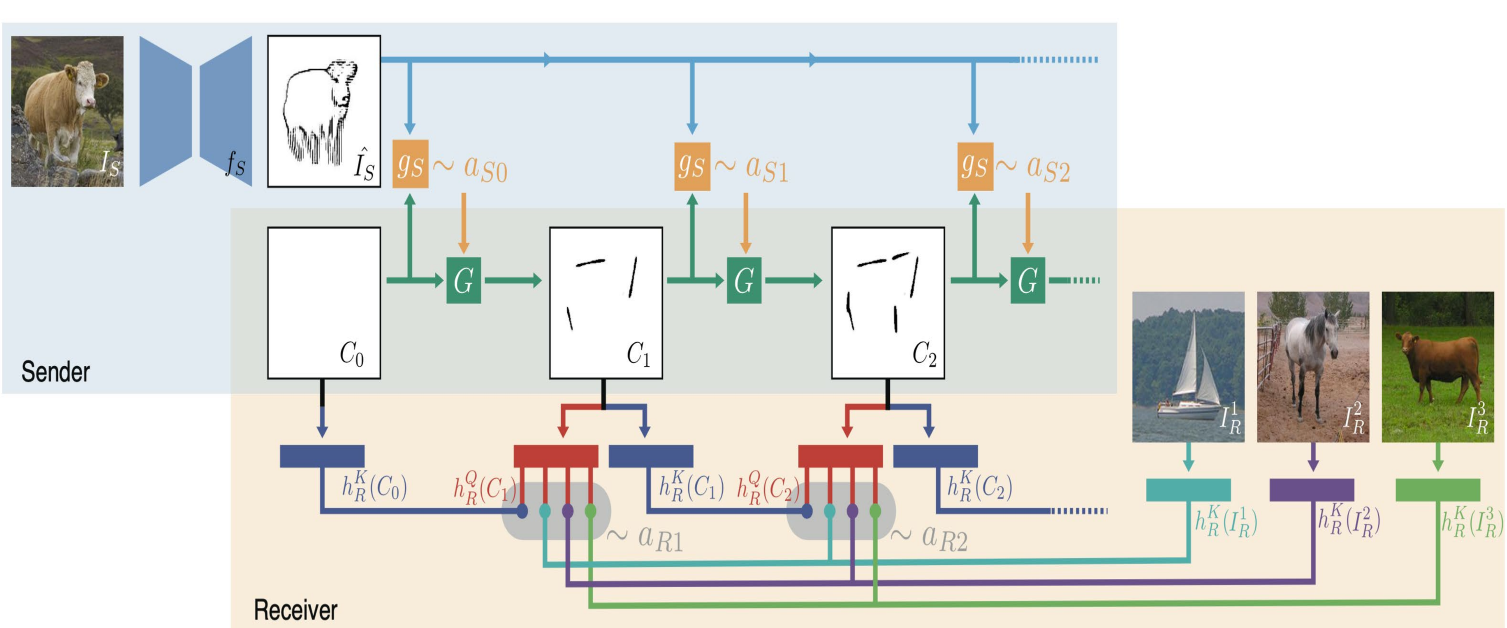
Humans first depicted the natural scene using drawings. In this way, the visual concept was grounded into iconic signs. After iterated use in communication, these signs gradually become abstract. The iconicity drops while the symbolicity rises. We aim to model this evolution process via two neural agents playing a visual communication game; the sender communicates with the receiver by sketching on a canvas.



The visual communication game

Communication process

In our visual communication game, a sender S and a receiver R only share the observation of the canvas C. The sender converts the target image I_S to a pixel-level sketch $\hat{I}_S$. At each step, the sender first draws five strokes a_S through the renderer G, which updates the canvas to C_{t+1}. Next, the receiver uses the updated canvas C_{t+1} to query from the context images $\{I_R^1, \dots, I_R^M\}$ and the last canvas C_t, deciding the action a_R at this step. The game continues if the receiver chooses to wait. A game round terminates when the receiver chooses one image as the target ($t = T_{choice}$). The agents will receive a shared temporally decayed reward or penalty $\gamma^t / -\gamma^t$, depending on if the receiver makes the right choice.



Agents are trained jointly to maximize the objective

$$\pi_S^*,\ \pi_R^* = \underset{\pi_S, \pi_R}{\operatorname{argmax}}\, \mathbb{E}_{\tau \sim (\pi_S, \pi_R)} \left[\sum_t \gamma^t r_t\right]$$

Value functions for an optimization surrogate

$$V(X_t) = \mathbb{E}_{\pi_S(a_{St}|I_S, C_{t-1}), \pi_R(a_{Rt}|\hat{X}_t)}[(r_t + \gamma \delta(a_{Rt}) V_\lambda(X_{t+1})]$$

An eligibility trace estimation: mixing Monte Carlo estimate at different roll-out lengths

$$V_\lambda(X_t) = \begin{cases} (1-\lambda) \sum_{n=1}^{H-1} \lambda^{n-1} V_N^n(X_t) + \lambda^{H-1} V_N^H(X_t) & \text{if } t \le T_{choice} \\ v_\phi(X_t) & \text{otherwise} \end{cases}$$

$$V_N^k(X_t) = \mathbb{E}_{(\pi_S, \pi_R)}\left[\sum_{n=t}^{h-1} \gamma^{n-t} r_n + \gamma^{h-t} \delta(a_{Rh}) v_\phi(X_h)\right] \quad h = \min(t+k, T_{choice})$$

Last step value estimate is trained by regressing the value returns

$$\phi^* = \underset{\phi}{\operatorname{argmax}}\, \mathbb{E}_{\pi_S, \pi_R}\left[\sum_t \left\| v_\phi(X_t) - V_\lambda(X_t) \right\|^2\right]$$

Experiments

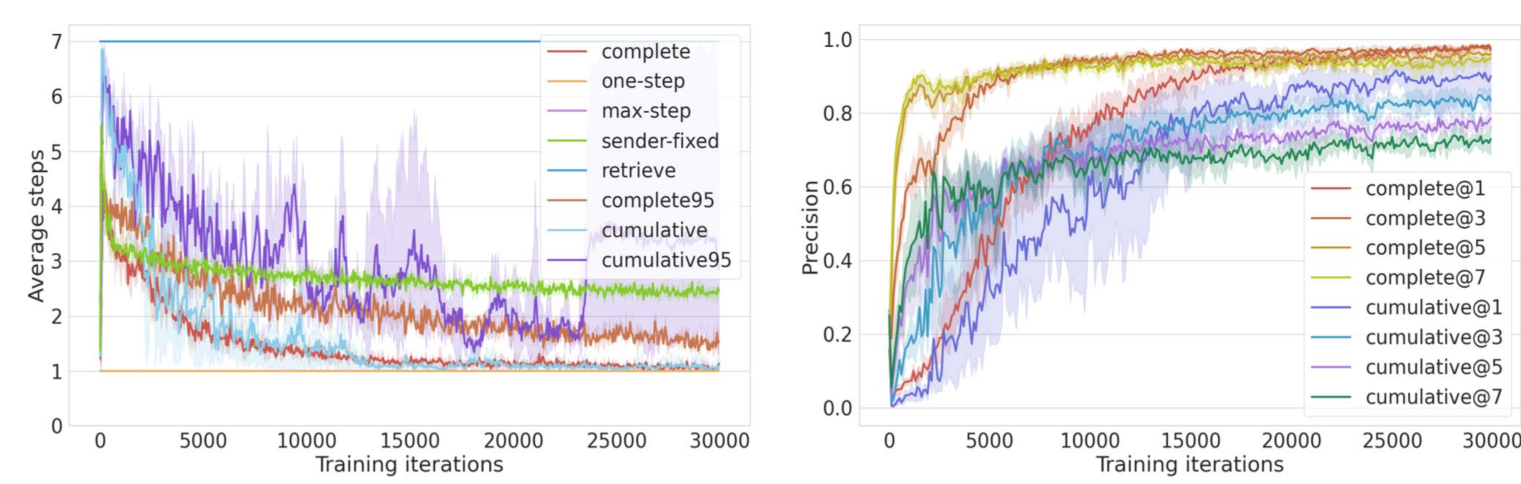
Settings

We consider three factors as crucial environmental drivers. To isolate each factor, we have one experimental setting and four control settings.

Game Settings				
early decision	update sender	max/one step	description	setting names
yes	yes	max	our experimental setting	complete
no	yes	max	control setting for early decision	max-step
yes	no	max	control setting for evolving sender	sender-fixed
yes	yes	one	control setting for sequential game	one-step
no	no	max	baseline for all settings above	retrieve

Communication efficacy and sketch abstraction

- All pairs except one-step can communicate successfully.
- Sketches are simplified along training.
- Agents trained in our framework can actively pursue the efficiency bound of accuracy and complexity compared with the REINFORCE baseline.



Iconicity: generalizing to unseen image

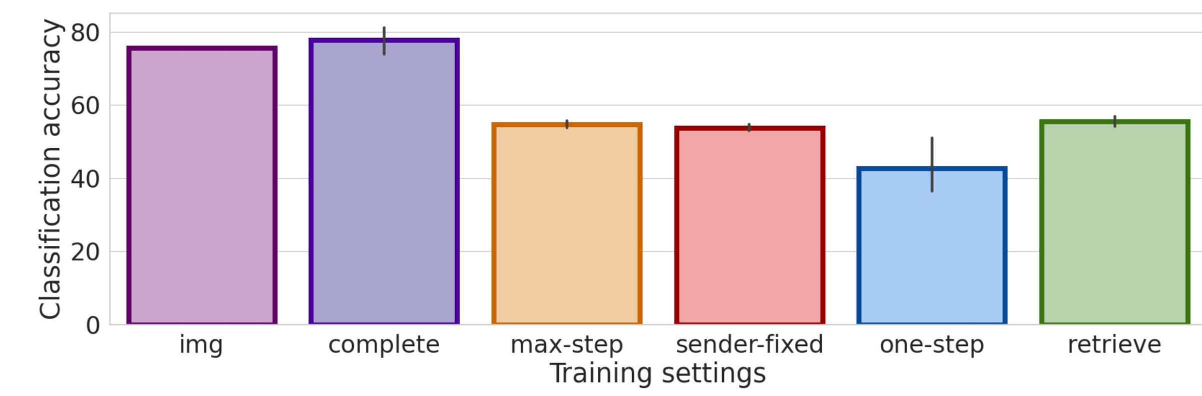
- Definition: drawings being proximal to the corresponding images on the high-level embedding space.
- Measure: agents' generalization ability on unseen images.
- Results: agents in the complete and sender-fixed setting can return to iconic communication when facing concepts not covered by established conventions (Table 1).

Communication Accuracy (%) ± SD (avg. step)			
setting names	seen	unseen instance	unseen class
complete	98.07 ± 0.01(1.03)	70.37 ± 0.04(2.36)	39.40 ± 0.05(3.76)
max-step	86.27 ± 0.03(7.00)	67.93 ± 0.02(7.00)	38.40 ± 0.04(7.00)
sender-fixed	99.60 ± 0.01(2.41)	71.80 ± 0.02(3.83)	45.40 ± 0.02(4.75)
one-step	22.87 ± 0.23(1.00)	14.07 ± 0.15(1.00)	9.60 ± 0.09(1.00)
retrieve	99.47 ± 0.01(7.00)	76.80 ± 0.02(7.00)	48.00 ± 0.02(7.00)

Table 1

Symbolicity: separating evolved sketches

- Definition: drawings being consistently separable in the high-level visual embedding.
- Measure: accuracy of finetuning a VGG16 to classify the final sketches into their corresponding categories.
- Results: agents in the complete setting can consistently highlight some features across all training instances in each category (bar plot).



setting	correlation
complete	0.43 ± 0.02
max-step	0.41 ± 0.13
sender-fixed	0.36 ± 0.06
one-step	0.31 ± 0.08
retrieve	0.55 ± 0.00

Table 2

Semanticity: correlating category embedding

- Definition: topography of the high-level embedding space of the drawings being strongly correlated to that of images.
- Measure: the correlation between distances of all vector pairs extracted using word2vec and all pairs in visual space extracted by the trained VGG16.
- Results: semanticity can be better retained in the complete setting .
- Visualization: in the complete setting, different concepts have a clearer boundary and the semantically similar concepts lie close to each other.

Visualizing evolution process

The sketches become abstract through iterations. The first several strokes in the same category consistently highlight the salient parts of the concept.