# Contextual flexibility in visual communication

Judith E. Fan[a], Robert X.D. Hawkins[a], Mike Wu[b], and Noah D. Goodman[a,b]

[a]Department of Psychology, Stanford University
[b]Department of Computer Science, Stanford University

November 26, 2017

Visual modes of communication are ubiquitous in modern life — from maps to data plots to political cartoons — and form the foundation for the cultural transmission of knowledge and higher-level reasoning. Here we investigate drawing, the most basic form of visual communication. Communicative uses of drawing pose a core challenge for theories of vision and communication alike: they require a detailed understanding of how sensory information is encoded and how social context guides what information is relevant to communicate. Here we meet this challenge by providing a unified task paradigm and computational framework for investigating contextual flexibility in real-time visual communication tasks.

We developed a drawing-based reference game involving two players: a *sketcher* who has the goal of producing drawings so that the *viewer* picks out a target object from a set of distractor objects. Participants (N=192) were paired in an online environment and interacted in real time. On each trial, both participants were shown an array of the same four objects, which appeared in different positions for each participant. One object was highlighted on the sketcher's display to designate it as the target. Stimuli consisted of 3D renderings of 32 objects belonging to 4 categories (i.e., birds, chairs, cars, dogs), containing 8 objects each. For each pair, objects were grouped into eight quartets: Four contained objects from the same category (*close*); the other four contained objects from different categories (*far*). Each quartet was presented four times, such that each object in the quartet served as the target exactly once. We found that people exploited information in common ground with their partner to efficiently communicate about the target: on far trials, participants achieved 99.7% accuracy while applying fewer strokes, using less ink, and spending less time ($p$s $< 0.001$) on their drawings than in the close condition, where accuracy was still high (87.7%).

We hypothesized that humans succeed at this task by recruiting three core competencies: (1) **visual abstraction**, the capacity to perceive the correspondence between an object and a drawing of it; (2) **informativity**, the drive to produce drawings that distinguish between the target and distractors; and (3) **cost sensitivity**, a preference to avoid using more ink than necessary. We instantiated these three competencies in a model architecture that combines a convnet visual encoder with a Bayesian model of social reasoning during communication (Goodman & Frank, 2016).

The visual encoder maps images (3x224x224) to a fixed-length feature vector (1000-dimensional). To capture humans' ability to perceive resemblance across image domains, we adapt higher-layer features from the pretrained VGG-19 convnet (Simonyan & Zisserman, 2014) to learn a multimodal feature embedding that captures image-level correspondences between drawings and photorealistic renderings of objects. This embedding minimizes the distance between matched renderings and drawings while preserving higher-order category structure, so we use distances in the embedding as a proxy for the perceptual similarity between images. The social reasoning module accepts a 4-tuple containing the distances between the human sketch and each of the four objects in the array, and outputs a probability distribution over drawings, reflecting the joint contribution of informativity and economy. We found that the model combining all three competencies outperformed lesioned variants of the model.

Together, these findings provide a unified model of how perception and social reasoning are integrated to support contextual flexibility in human visual communication. In the long run, investigating the computational basis of visual communicative flexibility may shed light on the sources of variation in pictorial style and the emergence of graphical conventions.

Word Count: 574 of 300 [way over]

Methodology/Approach: Behavior/Psychophysics

Primary Topic Descriptor: Perception and action: other

Secondary Topic Descriptor: XX

Funding sources: XXX

Presentation Preference: XXX

Suggested Reviewer 1: XXX

Suggested Reviewer 2: XXX

Suggested Reviewer 3: XXX