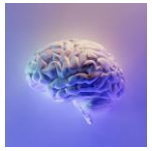
Emergent Language: Independent AI Development of a Language-Like Syntax



Alec Bunn
Paul G. Allen School of Computer Science & Engineering
Interactive Intelligence

Amelia Johnson
Paul G. Allen School of Computer Science & Engineering
Interactive Intelligence

Andre Ye
Paul G. Allen School of Computer Science & Engineering
Interactive Intelligence

Eric Xia
University of Washington Department of Mathematics
Interactive Intelligence

Yegor Kuznetsov
Paul G. Allen School of Computer Science & Engineering
Interactive Intelligence

Context and Motivation

Natural language models have been rapidly getting more powerful... or so it seems. Modern NLP models, while capable of demonstrating incredibly sophisticated behavior, suffer to grasp the semantics of the text they represent. **To master language, one must not only manipulate references (syntactics) convincingly but understand the meaning behind references (semantics)**. However, modern language models have no good way to access the semantics, and therefore end up learning brittle syntax webs.

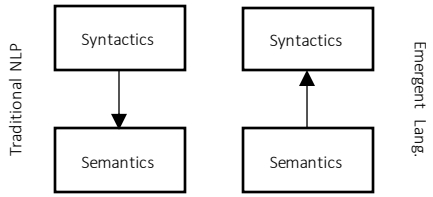
Syntactics

Semantics

It has been well-documented that large language models, like the well-known GPT-3 model demonstrate inability to reason through physical, experiential, and societal semantics needed to sensibly navigate language.

> ***Prompt***: You poured yourself a glass of cranberry juice, but then you absentmindedly poured about a teaspoon of grape juice into it. It looks okay. You try sniffing it, but you have a bad cold, so you can't smell anything. You are very thirsty. So you drink it.
>
> ***GPT-3 Continuation***: You are now dead.

To develop a more robust understanding of language, we highlight the relevance of the subfield of *emergent language*: the development of syntactics (language) *from* meaning/semantics (bottom-up approach), in contrast to traditional NLP (attempting to understand semantics through a web of references) (top-down approach).

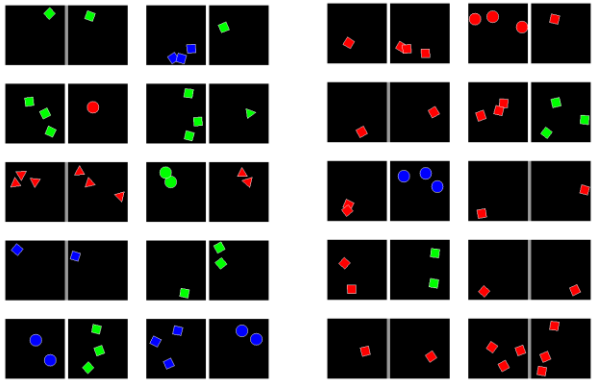


Such a syntactics emerges from optimizing under conditions of meaning; we do not define what symbols refer to and how they are used. Rather, we set conditions for the structure of language and allow the system to optimize for the syntax within the constraints. We propose three of such, which we build into our benchmark task and model design.

- *Discrete*. The language must be comprised of symbol units.
- *Sequential*. The language must be read and generated in sequence.
- *Variable-length*. Sequences can differ in length to reflect differences in meaning/content.

Benchmark Task

We need to define a problem that provides meaning upon which syntax can emerge. After iterating over several tasks, we converged upon the *geometric scene similarity* task. In the geometric scene similarity task, the model is presented with two images of geometric scenes. Each scene can between 1 and 3 objects (inclusive); all objects share the same color (either red, green, or blue) and shape (square, triangle, circle). Two scenes are considered the same if they feature the same number, color, and shape of objects. However, the objects may be in various locations or states of rotation and overlap. The primitive model objective, therefore, is to evaluate the similarity between two scenes by abstracting each image to these three essential characteristic dimensions.



Left: in-distribution samples. Right: out-of-distribution samples. A gray bar joining a scene pair indicates a positive label (i.e. same); its absence indicates a negative label (i.e. different).

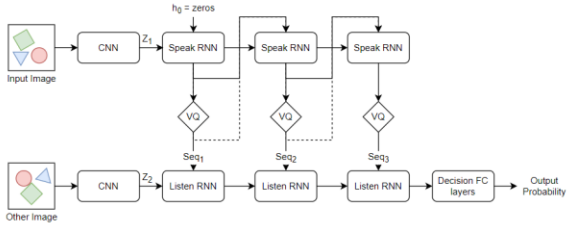
The model must encode each scene into a 'sentence', or a sequence of discrete tokens. The hope is that the model can develop an encoding like 'three green triangles' – demonstrating the emergence of adjectives and nouns – completely organically without being explicitly told of the presence of these three characteristics. The generated language is then compared to the other image to determine how representative or descriptive the sentence is of the image's content.
(See the "Model Design" section for more details on the model.)

Importantly, we define an *in-distribution* and *out-of-distribution* dataset. That is, we exclude a particular color-shape pair characterizing each object during training (in-distribution data) but evaluate the model on its ability to understand that excluded pair (out-of-distribution). If the model has properly acquired language, it will be able to separate and transfer adjective references to successfully solve the problem – to imagine the existence of objects that haven't been explicitly seen yet. Restated, we are optimizing for model *abstraction* via *linguistic reference transfer*.

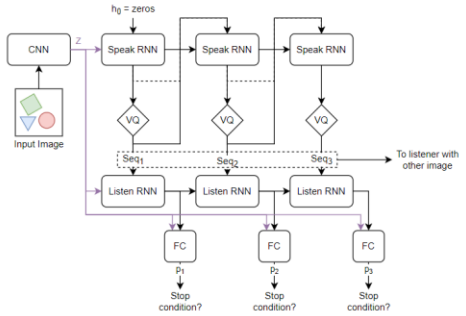
Model Design

To match the simplicity of the geometric scene similarity task, our model design is lightweight. It is comprised of the following core components:

- The speaker and listener map their respective images to a sparse embedding with a shallow Convolutional Neural Network (CNN).
- The speaker's image embedding is passed through a recurrent layer to generate sequence vectors.
- Sequence vectors are converted into a 'sentence' with the Vector-Quantization (VQ) layer, which snaps each vector to the nearest of a fixed-size set of learned embeddings. These represent the words in the 'language'. After quantization, the speaker has generated a discrete sentence-like encoding of the viewed image.
- The listener processes the sequence using its own recurrent facilities and decides as to if it received the same scene or not.

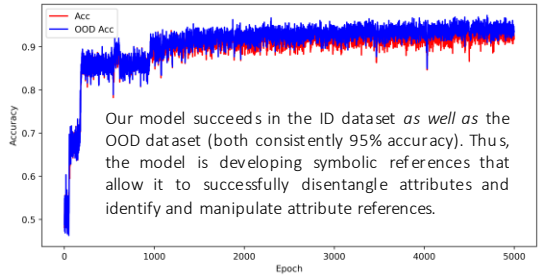


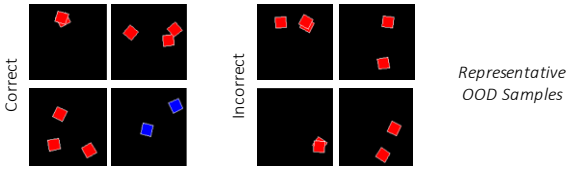
The true architecture was complicated with a variable-length mechanism, in which the speaker 'listens' to its own output at each step. We can utilize the final output prediction as a measure of how complete the sentence is. If the speaker can label its own image correctly with high confidence, then the sentence describes the image. This is done for each generated token. Once this self-understanding reaches a set threshold, generation is complete, and the sequence is sent to the listener.



Additional Mechanisms: Gumbel Softmax Sampling (in lieu of Vector Quantization), Sparse Visual Unit, Stop Token Estimation, Expanded Language Sizes, Recurrent Dropout. *Ask us about these!*

Results



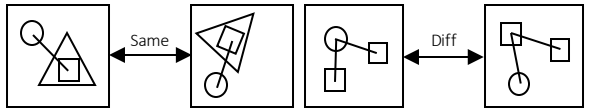


Representative OOD Samples

Next Steps & Future Work

Our current work has been preliminary. Our primary interest moving forward is in advancing the complexity of the emergent language by increasing the complexity and generality of the benchmark task.

One such task is to move from a geometric scene similarity task, which varies only along three axes, to a more complex *relational* scene similarity task. Objects can either be enclosed within other shapes (hierarchy) or connected to other shapes via a line or arrow (linkage). A model that successfully models this task must develop representations of verb-like relational tokens, which are transferable – like adjectives – across individual objects (nouns) and groups of objects (abstract nouns).



Broadly, our group is moving towards researching emergent language in reinforcement-learning contexts. Reinforcement learning is a more natural environment of semantics/meaning to develop language upon than standard supervised learning, since agents engage in adaptive 'decisions' and 'experience'. We are developing two such environments: a collaborative swarm search task in which agents must communicate to efficiently locate landmarks in a two-dimensional world and a predatory-prey system in which agents must communicate to survive predators and optimally maintain health.