Neural Network Linguistics

Marco Baroni

2019/2020 CCiL Master Annual Lecture
The FAIR Paris EViL team

Emmanuel Dupoux

Diane Bouchacourt

Rahma Chaabouni

Eugene Kharitonov
Abstract: We argue that rich data gathered in experimental primatology in the last 40 years can benefit from analytical methods used in contemporary linguistics. Focusing on the syntactic and especially semantic side, we suggest that these
Abstract: We argue that rich data gathered in experimental primatology in the last 40 years can benefit from analytical methods used in contemporary linguistics. Focusing on the syntactic and especially semantic side, we suggest that these...
Neural network linguistics

• Why
• What
• How
Deep neural networks

Deep neural networks

![Deep neural network diagram](image)

- **Input layer**
- **Hidden layer 1**
- **Hidden layer 2**
- **Output layer**

**yorkshire terrier**
The deep learning revolution

https://cs.stanford.edu/people/karpathy/cnnembed/
The deep learning revolution

I would like to request a stapler for my office.

Voldria demanar una grapadora per a la meva oficina.
KING LEAR:
O, if you were a feeble sight, the courtesy of your law,
Your sight and several breath, will wear the gods
With his heads, and my hands are wonder'd at the deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.
The empiricist stance

Nativism vs. *Tabula Rasa* [edit]

Connectionist models are often viewed as inherently accepting the theory that our minds are *tabula rasa*, blank slates. This view makes sense due to the fact that initial weights are generally randomized. If they were not, it would both take a long time to program, and then also not show the emergent phenomena. Some practitioners of connectionism feel that connectionism can solve the tabula rasa learning issue. My view is that eventually, we will see some compromise between the position of restricted domain knowledge as an expert system and that of connectionism modeling to remove the brittleness of those systems. Nor-

Walter Schneider's presidential address at the Society for Computers in Psychology, 1986

KING LEAR:

O, if you were a feeble sight, the courtesy of your law,
Your sight and several breath, will wear the gods
With his heads, and my hands are wonder'd at the deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.
Tabula rasa?

https://www.mihaileric.com/posts/transformers-attention-in-disguise/
Tabula rasa?

THE TRANSFORMER

Layer 1
- Add + Norm
- Multi-head Attention
- Input Embedding
- El, gato, está, feliz

Layer 6
- Same Operations
- ...
Tabula rasa?
Models of the mind/brain?
Models of the mind/brain?

NETWORKS AND THEORIES: The Place of Connectionism in Cognitive Science
by Michael McCloskey

A ROLE FOR CONNECTIONIST MODELING

Am I suggesting, then, that connectionist modeling has no role to play in cognitive science? Definitely not. In my view connectionist models hold substantial promise as tools for developing cognitive theories, if viewed from an appropriate perspective. Specifically, it may prove fruitful to think of connectionist models as akin to animal models of human functions or disorders (e.g., an animal model of working memory, or an animal model of attention deficit disorder).
A broader comparative perspective on language and communication

A richer language than any other species this side of humans

if you were a feeble sight, the courtesy of your law,
Your sight and several breath, will wear the gods
With his heads, and my hands are wonder'd at the deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.

Neural network linguistics

• Why
• What
• How
Are we really studying communication?

KING LEAR:
O, if you were a feeble sight, the courtesy of your law,
Your sight and several breath, will wear the gods
With his heads, and my hands are wonder'd at the deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.
Are we really studying communication?

O, if you were a feeble sight, the courtesy of your law,
Your sight and several breath, will wear the gods
With his heads, and my hands are wonder'd at the deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.

Communication games

http://www.publicdomainfiles.com/

Lazaridou et al. ICLR 2017,
Havrylov and Titov NIPS 2017,
Kottur et al. EMNLP 2017,
Evtimova et al. ICLR 2018,
...
Neural network linguistics

• Why
• What
• How
Our setup

- Two networks must jointly solve a task
- *Sender* network sees some input (e.g., a target image) and sends a *message* to *Receiver* network
- Receiver sees some input, including Sender message, and performs an action (e.g., point to target image) to complete the task
- The message is a single *discrete* symbol or a sequence of *discrete* symbols from a fixed alphabet ("discrete bottleneck")
- Networks rewarded for task success only, no supervision on the messages generated by Sender
Learning words for natural categories

Lazaridou et al. ICLR 2017
Learning words for natural categories

dog
Learning words for natural categories

At training time...

dog

Bouchacourt and Baroni EMNLP 2018
Learning words for natural categories???

At test time!

dog

Bouchacourt and Baroni EMNLP 2018
What would you talk about if all you had to do in life was to discriminate picture pairs?

"dog" (= "category 157 of 500")

9 bits

"larger average intensity in 16 pixels at image center"

1 bit

dsaves time and effort!
Efficient encoding of input information

• The cost function of human language?
  \[ \text{Cost} = -\alpha \ \text{Accuracy} + \beta \ \text{Complexity} \]

  evidence from naming in domains such as color and kinship (Zavlasky/Regier/Kemp/Tishby), might even account for Zipf’s law about word frequency distributions (Ferrer-i-Cancho)

• Are the agents in emergent language simulations optimizing for a similar trade-off (without need for explicit complexity minimization)?

• Kharitonov et al. 2019: Entropy minimization in emergent languages
The number game
Complexity (=entropy) minimization in the number game

Agents consistently develop the simplest (lowest-entropy) code necessary to solve the task.
Robustness to massive noise in training with the discrete channel bottleneck

Agents communicating through discrete channel are minimally affected by noise.

Comparable *continuous* agent interface does not show entropy minimization effect, and it overfits training data noise.
Zipf's law of abbreviation (ZLA)

In human languages, word frequency and length are inversely correlated

<table>
<thead>
<tr>
<th>Top 10 Most Frequent Words</th>
<th>10 Random Rare(st) Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>anadromous</td>
</tr>
<tr>
<td>of</td>
<td>barmaster</td>
</tr>
<tr>
<td>and</td>
<td>cruddy</td>
</tr>
<tr>
<td>to</td>
<td>gemstone</td>
</tr>
<tr>
<td>in</td>
<td>gonzo</td>
</tr>
<tr>
<td>I</td>
<td>idolization</td>
</tr>
<tr>
<td>that</td>
<td>pigling</td>
</tr>
<tr>
<td>was</td>
<td>sanguinity</td>
</tr>
<tr>
<td>his</td>
<td>unpredictability</td>
</tr>
<tr>
<td>he</td>
<td>walkman</td>
</tr>
</tbody>
</table>

Source: Gutenberg books frequency list, from Wiktionary
ZLA and efficiency

Word frequencies tend to be distributed according to a power-law frequency-rank relation

Assigning shorter forms to more frequent words is thus an efficient encoding strategy

I have been occupied with this story, during many working hours of two years. I must have been very ill employed, if I could not leave its merits and demerits as a whole, to express themselves on its being read as a whole. But, as it is not unreasonable to suppose that I may have held its threads with a more continuous attention than anyone else can have given them during its desultory publication, it is not unreasonable to ask that the weaving may be looked at in its completed state, and with the pattern finished.

Charles Dickens, *Little Dorrit*
ZLA and efficient encoding

important caveat:
"monkey typing" also obeys ZLA!
The Power-law Referent Game

Chaabouni et al. Anti-efficient encoding in emergent communication
to appear at NeurIPS 2019

inputs ("referents") sampled from 1k numbers (encoded as 1-hot vectors),
with sampling probability obeying power-law distribution
Most frequent "words":
English vs emergent language

<table>
<thead>
<tr>
<th>English</th>
<th>Emergent language</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>llmuhhmummmmmmmzuuuyyyzzvqzplan</td>
</tr>
<tr>
<td>of</td>
<td>naauuhhhhhhhhhhpucczzpbqoaqpln</td>
</tr>
<tr>
<td>and</td>
<td>shmhhhhhxxxxxxzummyyyttlctlgl</td>
</tr>
<tr>
<td>to</td>
<td>jahhhhhhuxxxxxzzzzzzzzzzzzzceyoawb</td>
</tr>
<tr>
<td>in</td>
<td>I</td>
</tr>
<tr>
<td>I</td>
<td>that</td>
</tr>
<tr>
<td>that</td>
<td>was</td>
</tr>
<tr>
<td>was</td>
<td>his</td>
</tr>
<tr>
<td>his</td>
<td>he</td>
</tr>
</tbody>
</table>
Anti-efficient message encoding in emergent language

more frequent messages significantly longer, unique case compared to human languages and animal communication systems (Ferrer-i-Cancho et al., Cognitive Science 2013)
The origin of anti-efficient encoding

Before training, Sender obeys ZLA... but it has no "articulatory" pressure to stick to it
The origin of anti-efficient encoding

The "anti-efficient" messages are already easier to discriminate for the Receiver before any training, so the latter is likely the culprit!

![Diagram showing L2 distance vs alphabet size (log scale)]

avg distances in untrained Receiver hidden space
Explicitly penalizing longer messages
Conclusion

• Modern neural networks are not *tabulae rasa* and probably they are not realistic models of the human brain

• Still, as cognitive scientists, we can extract precious insights from their linguistic behaviour *in a genuine communicative setting* by taking a comparative linguistics perspective

• Some initial findings:
  • Complexity minimization pressure in meaning encoding akin to the one found in human language
    • Leading to independent beneficial noise robustness effect
  • Distribution of "word" lengths in emergent language is *anti*-efficient, in a way that is profoundly different from human/animal communication systems
    • Showing that presence of ZLA in latter is not trivial!
Play with us!

https://github.com/facebookresearch/EGG

**EGG 🐔: Emergence of lanGuage in Games**

**Introduction**

EGG is a toolkit that allows researchers to quickly implement multi-agent games with discrete channel communication. In such games, the agents are trained to communicate with each other and jointly solve a task. Often, the way they communicate is not explicitly determined, allowing agents to come up with their own 'language' in order to solve the task. Such setup opens a plethora of possibilities in studying emergent language and the impact of the nature of task being solved, the agents' models, etc. This subject is a vibrant area of research often considered as a prerequisite for general AI. The purpose of EGG is to offer researchers an easy and fast entry point into this research area.
thank you

https://github.com/facebookresearch/EGG