Symbol Grounding in the Age of LLMs

Benjamin Gaskin

1. What is the symbol grounding problem?
2. How can we move from symbols to the world?
3. How can we move from the world to symbols?
4. What does this mean for us?

1. What is the symbol grounding problem?

What is the symbol grounding problem?

- **Searle's Chinese room argument—**
  - Individual in a room is provided with a **string of Chinese characters**
  - They have a **ruleset** which indicates the **proper manipulation** of these
  - They use this to **transform one string of characters and return another**
  - This is **received by those outside as a perfectly meaningful sentence**
- Where do large language models come into this?
  - The structure of the network can be seen as a **connectionist 'ruleset'**
  - This provides for the **transformation of one string into another**
  - Searle's argument indicates that this is **possible without any grounding**

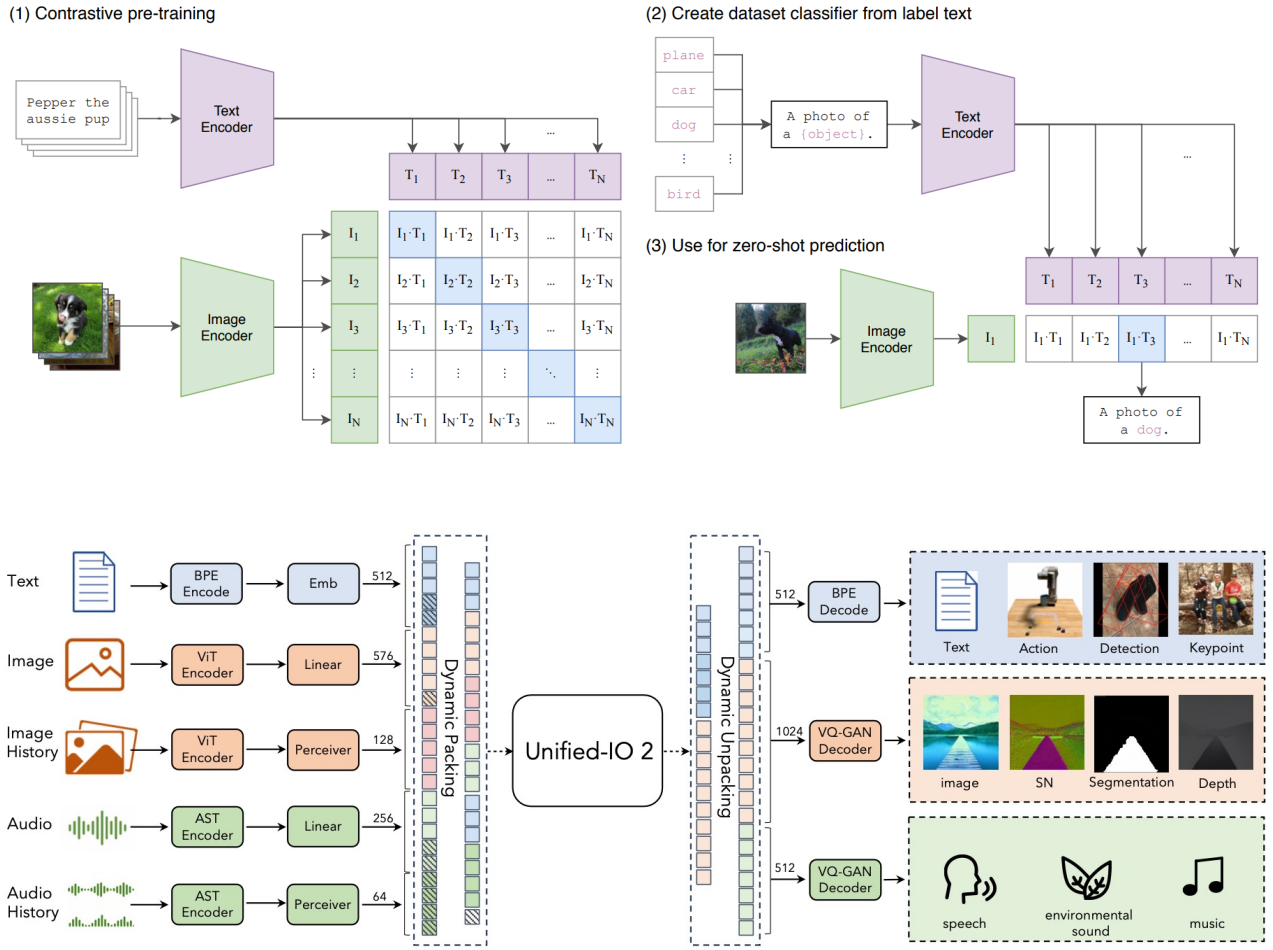
Extending the Chinese room

- Suppose that we allowed a **greater variety of inputs** to the room
  - The inputs are meaningless to the individual, so their nature is arbitrary
  - We might provide **tokens** (numbers corresponding to a discrete dictionary)
  - We might provide **n-dimensional embeddings** (continuous meaning space)
    - These vectors might equally refer to **text, images, sound, or some combination thereof**
- Similarly, suppose that we allowed the individual **further outputs**
  - Some might correspond to **language** (e.g., strings of Chinese characters)
  - Others might correspond to **more complex actions** (e.g., moving an object)
- Taken together, this would provide **a set of methods for grounding**
  - But it is nevertheless **parasitic upon the meanings given by its creators**

2. How do we move from symbols to the world?

Multi-modality

- STT, etc.
  - Translated into **text**
- CLIP, Flamingo
  - Translated into **embeddings**
- Unified-IO 2, GPT-4o
  - **End-to-end multi-modality**

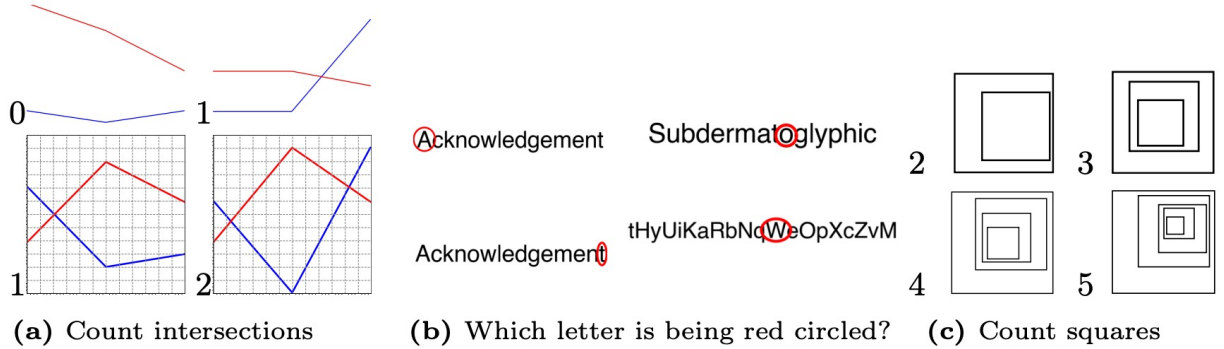


Robotics

- **RT-2** (Google DeepMind)
  - Brohan et al., 2023
  - These word by outputting action tokens, akin to any other output
- **Figure 01** (Figure and OpenAI)

Limitations in multi-modality and robotics

- Vision language models are **blind** (Rahmanzadehgervi et al., 2024)
  - Struggle with **low-level vision tasks** (e.g., whether close lines intersect)
  - While GPT-4o does well on VLM benchmarks, performs as poorly here
    - **End-to-end multi-modality is not the answer**, still structured by language



(a) Count intersections (b) Which letter is being red circled? (c) Count squares

- Action models are **limited to motions observed in the training data**
  - Generalise in applying these to **unseen objects, environments, and backgrounds**

3. How do we move from the world to symbols?

Language learning in children and machines

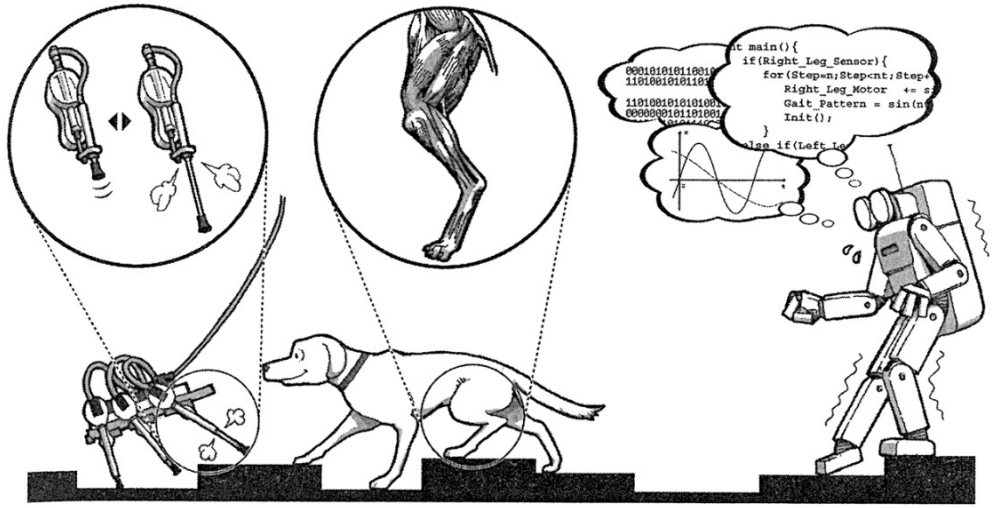
- LLMs are trained on **text data**, primarily taken from the internet
  - Decomposed into ‘tokens,’ which can be **words, sub-words, or even phrases**
- The exact quantity is unknown, but estimates are LLMs likely require **three orders of magnitude more than even a highly literate human**
  - Frank (2023)
- How do LLMs learn?
  - Simply put, they model the **probability distribution of sequences of tokens**
    - They are trained by having them predict the next token, given a prior sequence
    - When they make a mistake, the network elements responsible for the error are tuned
- What do LLMs learn?
  - They learn **to predict the next token**, by repeating this they infer larger ‘units’

Language and the structure of experience

- **External** symbols, for instance: **a tree**
  - Trees are an **entity** that is **encountered perceptually** in the world
  - The **structure of this experience constrains the use of the word** ‘tree’
  - **Metaphor extends this into abstraction**: family trees, tree search, etc.
- **Internal** symbols, for instance: **to kick**
  - For the **child**—
    - The **action as embodied** comes before the verb
    - This structure **constrains the use of the word** ‘kick’
  - For a **language model**—
    - The word is converted to a **token**, then an **n-dimensional vector**
    - Its meaning (i.e., usage) is **derived probabilistically from text data**

Morphological computation

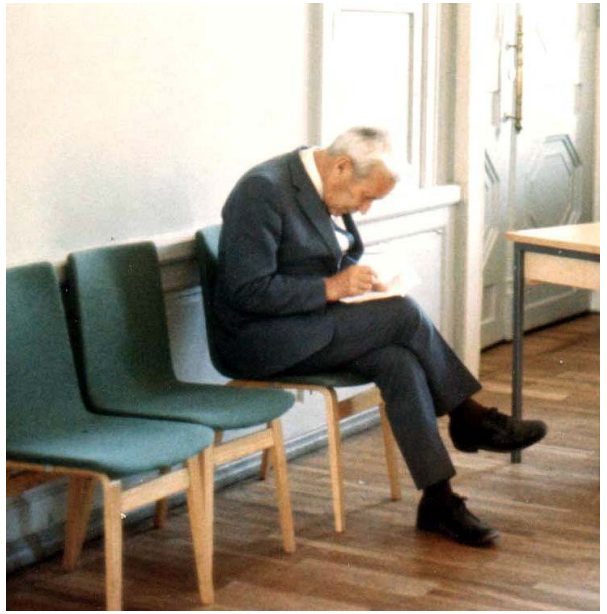
- "By '**morphological computation**' we mean that **certain processes are performed by the body that otherwise would have to be performed by the brain**."
  - Pfeiffer & Bongard, 2007
- Müller & Hoffman (2017) define these cases as "**morphology facilitating control**"
  - Similarly, that the structure of perceptual experience acts as "morphology facilitating control"



Kolmogorov complexity

- What is the **shortest possible program** that would **reproduce a given mathematical object?**
  - This is its Kolmogorov complexity

- 14159265358979323846264338327950288419 71



Conditional Kolmogorov complexity

- "The conditional Kolmogorov complexity … measures the amount of constructive information h′ contains about h—**how much information h′ contains for the purpose of constructing h**."
  - Mahmud & Ray, 2007
- We can thus formalise our argument as follows:

$$K(L|C) < K(L)$$

  - K = the number of tokens required for linguistic aptitude
  - L = a given level of linguistic aptitude, roughly that of an adult
  - C = integrative access to sensory information

Multi-modal language learning in LLMs

- **Wang et al., 2024**
  - "The visual representation produced by the vision encoder is used to **initialize the hidden state** of the uni-directional LSTM. …the captioning network shares the same LSTM architecture [as the text-only LSTM] for language processing and is trained to optimize the same objective, next token prediction."
  - "The improvements for most syntactic categories are statistically significant, but in particular, **nouns and verbs benefit the most** from additional visual information."
- **Zhuang et al., 2024**
  - "… **when only a small amount of data is available**, **Visual + Word models are more efficient than Language-Only models** in learning to relate words and predict semantic features."

Implications

- Multi-modality may well improve the efficiency of training in LLMs
  - **Reasonable evidence** for cross-modal grounding of language
  - Currently limited to visual input, **cross-modal rather than truly multi-modal**
  - Children also have access to **aural, proprioceptive, etc.**
    - **No reason in principle to doubt this extension**, perhaps requiring architectural advances
- More broadly, however—**what problem do we want to solve?**
  - We can imagine an **experimental philosophy** which supports theoretical work
    - "**Truth is verified only by creation or invention**," per Vico
  - Whether we want intelligent behaviour or something more human-like

4. What does this mean for us?

Language and the world

- Humans start out by moving **from the world to language**
  - This provides us with a **minimal ontology**
  - Of course, **language then alters our relation to the world**
  - **Humans also make use of statistical methods**
    - As where we infer the meaning of a word based on its context
- LLMs, however, build **from language to the world**
  - Some of their limitations may relate to this lack of grounding
    - The **inefficiency** of language learning
    - **Surprising failure modes in vision and action**
- Finally, **what about consciousness and creativity?**

In the beginning of heaven and earth there were no symbols. Symbols came out of the womb of matter.

—Lao Tzu