Research Proposal: concrete challenges on the way of Artificial General Intelligence

Zied Ben Houidi

Background

Neuroscience and Computational neuroscience have made considerable progress in understanding how single neurons and populations of neurons encode information in the human brain. On the other hand, deep learning, although it lacks biological plausibility, has achieved spectacular progress on hard tasks like computer vision and speech recognition. This progress was boosted by a jump in the availability of labeled data but also an increase of affordable computational power (GPUs). Although not biologically plausible, the intermediate simple representations that deep neural networks learn as well as experiments like the deepdream computer vision program [4] [1], show that there is a functional similarity between deep artificial neural networks and biological neural networks.

This research project will sit on the boundaries between these two disciplines to bring some computational neuroscience models, intuitions and problems into deep learning frameworks and eventually feed back computational neuroscience with insights and experiences learned on the way.

The project will address none, one, some, or all of the challenges below, as the understanding of the challenges might change with time and experience.

1 Extraction of solid invariant representations from unstructured input

The human vision system seeks invariance out of extremely varying input sensory data. The first challenge is how to build solid and invariant intermediate representations of the structure behind input sensory data (vision with motion (videos) ideally coupled with speech data, as a start). In current deep Convolutional Neural Networks (CNNs), such representations today are forced or (mis?)guided by the input training data and the specific recognition task at hand. Indeed, fitting the neural networks weights is done today by the backward propagation of errors. Hence, the learned weights end up in forming representations that are optimal with respect to

[1] which generated psychedelic images similar to human hallucinations provoked by certain drugs

the underlying task at hand, instead of seeking the solidity and invariance of the intermediate representations.

A lot is known today about the neural coding or intermediate representations that are made in the primate (and human) visual cortex (see e.g. [3] for a comprehensive overview). Such visual cortex representations could serve as a guideline/baseline for finding the right biologically plausible intermediate representations for an artificial neural network. Biological plausibility is not a goal for its own sake, but we believe that it is important to have machines equipped with similar intermediate representations as humans to be able to communicate effectively with them.

Finally, this challenge can build on recent advances that use more biologically plausible spike timing dependent plasticity (STDP) in deep convolutional networks to "unsupervisedly" extract intermediate representations and successfully use them for classification tasks [2].

Hypothesis/New ideas

My hypothesis/new idea in this space is that besides Hebbian learning (the "fire together wire together" rule), I suspect that there is a novelty-based learning to reinforce certain representations. Any novel representation (e.g. a first time a neuron fires) that causes a surprise (information is surprise) attracts more attention (a neuromodulating[2] effect) that reinforces the connections that fired/activated for the first time. One angle of attack here is to explore the use of such surprise/novelty-seeking reinforcement mechanism to achieve the highest levels of invariance and hence solid representations.

2 Augmenting AI machines with the ability to describe, explain and reconstruct complex objects in terms of intermediate simpler representations

A human brain can detect a face, from the retina on, until the detection, in a single feed-forward pass. It can also however picture a face by the "simple" pronunciation of its name. The second challenge is hence to exploit the formed intermediate representations to augment deep neural networks with the ability to describe/explain a classified/recognized item in terms of the elementary representations that led to its activation. This challenge is related to a variant of the (somewhat mysterious) neural binding problem [1]. A neural network is a large distributed system where each single neuron activates in response to a single simple feature. In the human brain, different areas in the human brain process different aspects like motion color and shape. The binding problem is the question of how a single apparently unified object can be traced back to these elementary features.

[2] As opposed to synaptic transmission, it is a process where a neuron affects remotely a population of neurons instead of a single one.

Hypothesis/New idea

The same neurons that activate when the object is classified/recognized can be used by means of a retrograde/backward signaling, from upstream layers to downstream layers, to reconstruct the object (Upstream layers are the layers that are farther from the input).

3 Augmenting AI machines with a declarative memory

Current deep learning networks lack a declarative explicit memory where objects are retrieved explicitly. This third challenge, also related to the second, consists in augmenting artificial neural networks with a long term memory store where a complex priorly observed stimuli (e.g. image of a cat) is reconstructed from the unique intermediate representations that led to its classification/recognition.

3.1 Hypothesis/New idea

Some neurons of a specialized network act as memory pointers. The activation of such pointers coupled with a retrograde signaling allows to recollect memories. My intuition is that such a model seems inline with recognition memory that is based on two distinct sub-processes, familiarity and recollection [5]. Familiarity is knowing something without necessarily being able to get/reconstruct its details. Recollection is reconstructing the details.

4 Augment AI with offline conceptual simulation or reasoning

The fourth challenge is to build on 2 and 3 to augment such deep neural networks with *understanding* and *imagination*. We define the former as the ability to reinvoke/reconstruct mental visual images from natural language symbolic stimuli (see challenge 2 and 3). The latter is defined as the ability to mix acquired intermediate representations to form new never-seen-in-the-real-world/input data stimuli. We believe that solving this ultimate challenge can offer answers to the unsolved variable binding variation of the neural binding problem [1] and open the first concrete door to Artificial General Intelligence: an intelligence that conceptually simulates the world offline, recognizes and understands the result of the conceptual simulation (thanks to challenge 2)

Hypothesis/New idea

My hypothesis/theory here is that variable binding will come out as an emergent property of well-trained networks, similarly to how well-trained word embeddings (like word2vec) result in nice semantic properties of the resulting vector space.

5 The fifth challenge

At least of fifth challenge is missing and I need more time thinking and reading to be able to describe it. It roughly concerns what I think a neuro modulating effect that triggers and guides the conceptual simulation. This aspect is related to planning and goal oriented behavior.

References

[1] Feldman, J. The neural binding problem (s). *Cognitive neurodynamics 7*, 1 (2013), 1–11.

[2] Kheradpisheh, S. R., Ganjtabesh, M., Thorpe, S. J., and Masquelier, T. Stdp-based spiking deep convolutional neural networks for object recognition. *Neural Networks* (2017).

[3] Kruger, N., Janssen, P., Kalkan, S., Lappe, M., Leonardis, A., Piater, J., Rodriguez-Sanchez, A. J., and Wiskott, L. Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *IEEE transactions on pattern analysis and machine intelligence 35*, 8 (2013), 1847–1871.

[4] Mordvintsev, A., Olah, C., and Tyka, M. Deepdream-a code example for visualizing neural networks. *Google Res 2* (2015).

[5] Wixted, J. T., and Squire, L. R. The role of the human hippocampus in familiarity-based and recollection-based recognition memory. *Behavioural brain research 215*, 2 (2010), 197–208.