ML in 2024: Ocean Deep

July 5, 2024

The following discussion is deeply inspired by the talk from Qi Lu and the book "A Thousand Brains".

I. Foreword

The AI hypes emerge and fade, but they never stop. In the past ten years, the advancements in convolutional neural network and reinforcement learning have empowered the pursuits of autonomous driving and embodied robot, but we can hardly see such applications, still. In retrospect, these techniques are more than marginal improvement but less than paradigm shift.

Given the prompt above, will the fate of the recent hype in large language models be different?

II. The triplets of intelligence

In general, an intelligent system consists of three modules: perception, comprehension with planning, and execution. One such example can be found in autonomous driving, where perception, planning, and control are the three core parts of the self-driving system. Mathematically, the relationship of these modules can be simplified as:

$$\hat{y} = f(\hat{x})$$

where $\hat{x}, f(\cdot), \hat{y}$ corresponds to perception, comprehension with planning and execution, respectively. The notation of $\hat{\text{hat}}$ indicates the module requires some pre- and post-processing on x and y.

According to the talk from Qi, the previous tech hype was mainly fueled by breakthroughs in the perception module, where models can see the world as we see it. The latest rise in large language models, however, have not only demonstrated the capabilities of comprehension with planning, but also enlightened further exploration on the execution module. In other words, the most recent hype gives rise to the end-to-end deployment of intelligent systems like never before.

III. The missing masterpiece

Although GPT and its variants have achieved remarkable performance on certain tasks, their generalizability is far from perfect. Whereas marginal improvements have been attained by imitating human behaviors, momentous leap might require the imitation of human brain. In the book of "A thousand brains", the author explains how brain works and also points out that many of the mechanisms can shed light on the development of artificial intelligence. Interestingly, I found many of these ideas map to the triplets of intelligence. Please allow me to introduce a minimal background before drawing the connections.

A thousand brains

our brain is composed of neocortex and the older parts (allocortex). While the older parts controls primal activities, the neocortex is responsible for the sophisticated functionalities. Our (physical) actions are directly controlled by the older parts, and indirectly by the neocortex as the latter governs the former. In short, the neocortex is pertinent to perception and comprehension, and the allocortex corresponds to execution.

In finer grain, the neocortex is composed of identical unit (cortical column) that follows the same mechanism. In other words, the neocortex process vision, touch, language and logic indifferently. As a result, these heterogeneous signals are abstracted and stored in the same format to build up our world models.

The world models are complementary models of the objects cognized through our interaction with the world. Although located in different cortical columns, these models are all updated by the same intrinsic principle in neocortex: the prediction mechanism. Unconsciously, our brain is constantly making predictions on the imminent sensations. Essentially, we might not realize these activities unless the prediction contradicts with the output of the current world models.

Perception $\leftrightarrow$ unified tokenizer $\leftrightarrow \hat{x}$

To recap, our brain consumes inputs indifferently through the cortical column, regardless of the heterogeneous format. The recent advance in ML, however, does not follow this principle. For example, in multi-modal learning images and texts are firstly tokenized by different vision and language encoders. Subsequently, these encoded representations are aligned by pairwise similarities before they could be consumed by the model. The procedure is identical for other modalities but the complexity compounds.

A unified tokenizer that directly projects varied signals into a unified representation space could be a better design. The raw signals are nothing but a

combination of zeros and ones. In other words, the view, sense, and textual description of the same object are orthogonal in format but collateral in essence. The alignment would be much easier or even not required if more effort is put on the tokenization.

Comprehension and planning $\leftrightarrow$ predictive world models $\leftrightarrow f(\cdot)$

Our brain updates and corrects itself by constantly predicting the futures. Its underlying units are world models capturing the complementary aspects of object and concept. Majority voting is utilized when making the decisions.

Does the mechanisms above remind you of self-supervised learning and mixture of experts? Model training requires feedback from third party, and our brain actively obtains such assessment through predictions. This procedure is identical to the current practice of autoregressive models where the model parameters are updated through predicting the next token. However, self-correction is still missing as the model becomes stagnant after deployment. We should never stop learning.

Diversified models with unified representation is another critical implication. Such framework not only guarantees better redundancy from a control perspective, but also allocate bias-free adaptivity from a learning perspective. The recent wide adoption of mixture of experts, unintentionally, has followed this path. However, the scaling of the number of experts and the design of voting requires further priority.

Execution $\leftrightarrow$ hierarchical control $\leftrightarrow \hat{y}$

The duality of neocortex and allocortex unveils a hierarchical architecture of executing actions from thoughts. Explicitly, execution might require two coupled and synchronized modules, where the high-level one rolls out plan and the low-level one takes primal action. The two modules can either be trained jointly in an end-to-end fashion, or independently given shared signals. The recent advance of FSD V12, however, evinces the plausibility of the former over the latter.

This design recipe can be directly leveraged by the development of embodied agents. The multi-modal (language) models have unprecedentedly enabled the agent to comprehend and plan, and now could be the time to double down to make the agent move. Nevertheless, the duality gap between the two parties needs to be filled by experts with interdisciplinary background.

IV. Forward

It is a new world. It is a new start.