Thesis proposal review

Sunit Bhattacharya: Multimodal Machines from a Perspective of Humans

The thesis proposal of Sunit Bhattachary is an interdisciplinary work from the area of cognitive science which combines ideas and methods of deep learning and neuroscience. It aims to study deep learning models for Natural Language Processing (NLP) tasks by exploiting cognitive data collected from brains of humans solving the same tasks.

Although the artificial neural networks were originally inspired by biological neural systems and some of the existing architectures do exhibit certain "deeper" similarities with human neural systems (e.g. in Computer Vision (CV) tasks), the design of the state-of-the art models is still motivated by engineering needs only and brain-inspired attempts are very rare. Existing literature comparing biological and artificial systems is also limited and datasets to allow such research are very scarce. Therefore, the proposed research direction is up-to-date but also very challenging.

Content

The proposal is written in English, spanning 11 pages of the main text plus a rich bibliography. The work is structured into 6 sections. The introduction in Section 1 puts the proposed work into a large context. The author first reviews some of the recent achievement of deep learning methods (e.g. in NLP and CV), some of their main drawbacks (e.g. lack of generalization), and discuss their biological plausibility. Then, the author presents the concept of multimodality, how it differs in different areas and which of the existing definitions is applied in his work. Finally, the author presents three research questions he would like to study: 1) how to compare human and machine performance in multimodal tasks; 2) how to use deep learning systems to predict human brain activity when handling the same multimodal tasks; 3) whether biological plausibility can help in designing systems that exhibit human learning abilities. Section 2 provides an overview of related work from several areas: human vs. machine learning, human vs. machine representations, and multimodal learning.

Sections 3, 4, and 5 present the main content of the proposal. In Section 3, the author presents three main research tasks he would like to tackle: 1) comparison of machine representations; 2) assessing the capability of neural models to predict cognitive data; 3) comparions of human vs. machine performance on multimodal tasks. The tasks are planned to be studied in two use-cases (tasks): language modeling and machine translation, both in the traditional text-based and also multimodal setting.

In Section 4, the author presents results of his work so far. The first achievement is the dataset collected during an experiment when human participants were asked to read and translate sentences from English to Czech and optionally revise the translation after considering a visual signal (image). The audio, EEG and eye-tracking data were recorded and included in the dataset. Section 4.1 presents some interesting findings from this experiment (e.g. ambiguous sentences take longer to translate). A paper describing the dataset has been published on Arxive.org. Section 4.2 is devoted to participation of the author's team in a shared task on predicting eye-tracking data by pre-trained language models. Other work of the author is briefly presented in Section 4.4.

The future plans are described in Section 5 and the text is concluded in Section 6.

Evaluation

The text is readable, well structured, with some infrequent grammatical errors and typos. Some technical errors occur in the text too (e.g. in Equation 1). The introduction and motivation is very interesting and puts the research plans and goals into a motivating context. The review of related work and overview of the most relevant concepts and methods is rich and complete (the bibliography includes about 150 referred papers!). A special subsection is devoted to the problem of catastrophic forgetting, however, it is not very clear how this is relevant to the proposed research plan. The author probably hypothesizes that biological plausible methods can help to reduce this problem (catastrophic forgetting), but this is not clearly reflected in the methodology and future plans.

Similar inconsistencies are noticeable also later in the text. For instance, it is not clear how "comparison of machine representations" and analysis of "how linguistic features are encoded in such representations" (the first task in the list in Section 3) would contribute to answer (any of) the research questions formulated in Section 1. Similarly, the second task is to "assess the capability of neural models to predict human multimodal behaviour", but it is not clear how this would help to "identify if biological plausibility translates to better performance for machines".

The methodology described in Section 3 is unfortunately limited to the third task only (comparison of human vs. machine performance). The first two tasks are not addressed and is not clear how the author wants to tackle them.

Section 4 presents very interesting work and I would like to see it published at a good conference. I understand that the thesis proposal cannot provide all the details of the experiments but i would like to know how the data instances (sentences and pictures) were constructed, what kind of ambiguity is present in the ambiguous examples and whether it can be resolved from the picture. Also, it is not clear, how the 4 stages were distinguished. For instance, it is not clear how the duration of ”see“ was measured.

Throughout the text, the author deals with some important concepts that are not well explained. For instance, ”biological plausibility“. The author frequently distinguishes between models that are biologically plausible and those that are not. It is not clear how this quality is assessed. Another important concept is multimodal modeling (the **C** in Equation 4 and 5), however this is is not discussed in the text at all. It is a very large research area and plenty of options exist, especially for image and video modalities (that are highly relevant to this work).

Conclusion

Despite the above mentioned issues, the work plan described in Section 5 seems reasonable. The author has already got some achievements that he can build on and direct his future steps accordingly.

I consider the presented research topic and the research plan sufficient for the author's dissertation.

doc. RNDr. Pavel Pecina, Ph.D. In Prague, September 8, 2022.