Master Thesis Review

Faculty of Mathematics and Physics, Charles University

Thesis author	Andrew McIsaac
Thesis title	Temporal Reasoning in Vision and Language Models
Year of submission	2024
Study programme	Computer Science
Study branch	Computer Science – Language Technologies and Computational Linguistics

Review author doc. RNDr. Ondřej Bojar, Ph.D. **Role** supervisor
Department Institute of Formal and Applied Linguistics

Review:

The master thesis submitted by Andres McIsaac was done jointly at the University of Trento and Charles University. The thesis focuses on a very timely topic, namely the reasoning ability in neural vision and language models (VLM). With the rise of multimodal models and with the high expectations and promising but also sometimes controversial results in various types of commonsense reasoning in these models, carefully evaluating a well-defined question on reasoning is very useful.

Andrew used data perturbation techniques starting with the STAR text annotations of Charades video dataset (indoors, single activity) and the NExT-QA questions for the YFCC-100M video dataset (outdoors, greater variety of scenes, more types of temporal relations). The results confirm that two VLM models, Merlot Reserve and VideoCLIP, are insensitive to time-related perturbations. In particular, (1) the models' ability to predict the verb in an action description is not decreased when the action description is made inadequate by swaps of temporal expressions before/after, and (2) shuffling the frames based used by the models to make their prediction does not cause any big harm the prediction.

Aiming to improve the temporal reasoning ability, Andrew then fine-tuned the Merlot Reserve VLM based on the Charades dataset. The additional data Andrew constructs provide the model with contrastive, correct and wrong (temporal relation damaged), examples. The model thus sees some training material where the words expressing the temporal relation explicitly play an important role in the task. The results are promising but far from fully satisfactory, demonstrating an improvement in some of the tests but inconclusive with respect to the general ability of temporal reasoning.

The thesis text is structured into six chapters plus introduction and conclusion: Chapter 1 focuses on the theoretical background (language modeling, image recognition, then visual and video language modelling, concluding with the topic of the thesis, temporal reasoning). Chapter 2 reviews work related to Andrew's goals: techniques for evaluation and improvement of reasoning in VLMs. Data are described in Chapter 3. Chapter 4 summarizes the perturbation experiments. Chapter 5 and Chapter 6 cover the attempts to improve temporal reasoning, describing the method and results, respectively.

The thesis is somewhat shorter than the average for excellent master theses (with the Conclusion

on page 48 and the total of 67 printed pages). This does indicate that the set of experiments would be insufficient. Quite on the contrary, I like the range of quantitative as well as qualitative probes Andrew did. The succinctness is harmful primarily in the discussion of the results where a more verbose style would make the lessons learn much more accessible and easier to verify.

Overall the text quality is excellent. The presented figures come, with three exceptions, from previous work and all correctly cited.

I have two additional questions:

- In Section 1.5.2, you cite Moens and Steedman (1988) who argue for having a good account of event mutual dependecies rather than just their sequential ordering for 'when' questions. I would assume that a considerable part of temporal reasoning performance would be coming from the fact that textual data used in the training of VLMs follow this and represent primarily event pairs which do have causal or other dependencies between them, not just sequential ones. To what extent do your test sets respect this? Could the presence or absence of not-just-sequential dependency between events be distributed differently in your positive and negative examples, skewing the results?
- What is the utility of having the sound spectrograms as input to the models? I know you do not have an ablation study of this, but what would you guess based on your knowledge of the data?

As clearly documented by the submitted thesis, Andrew McIsaac can conduct research on his own, design and carry out complex of experiments, obtain comparable scores from them and interpret and present the results concisely. A more verbose discussion would have made the thesis much easier to follow but already the submitted version is undoubtedly sufficient to meet the requirements for master theses at Charles University. I thus recommend the thesis to be accepted.

I recommend the thesis to be accepted.

I do not propose the thesis for special recognition.

In Prague, February 5, 2024

Signature: