Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Andrew McIsaac
Název práce	Temporal Reasoning in Vision and Language Models
Rok odevzdání	2024
Studijní program	Computer Science
Studijní obor	Language Technologies and Computational Linguistics
Autor posudku	Mateusz Krubiński
Role	Oponent
Pracoviště	Ústav formální a aplikované lingvistiky

Review text

In his Master thesis, Andrew McIsaac aims to provide an answer to an important question regarding the temporal reasoning abilities of Vision and Langauge Models (VLMs). He realizes it through the lens of (retrieval-based) Video Question Answering, probing how the model behaves under the *temporal* perturbations (e.g., by substituting "before" with "after" in text or shuffling video frames) of the input. There are two main contributions of this work. The first one is the creation of a new video-based dataset with annotated actions (descriptions and timestamps) based on the Charades (Sigurdsson et al., 2016) dataset that includes newly generated hard negatives. The second one is the evaluation framework built upon the calculus of temporal intervals (Allen, 1983), extending the "before/after" reasoning to other temporal relations, such as "overlaps" or "during".

The thesis is structured into six Chapters. Chapter 1 introduces (some of) the relevant Vision and Language Modeling background and establishes the definition of temporal reasoning. Chapter 2 presents how previous works applied contrastive training to VLMs and highlights a number of methods developed to improve robustness. Chapter 3 describes (too) briefly two datasets: STAR (Wu et al., 2021) and NExT-QA (Xiao et al., 2021) and two recent VLMs: Merlot Reserve (Zellers et al., 2022) and VideoCLIP (Xu et al., 2021), that are explored in the experimental part. In Chapter 4, we look at the zero-shot probing results. Chapter 5 presents the preparation of the novel dataset and describes the post-pretraining procedure. Finally, Chapter 6 is dedicated to comparing the off-the-shelf model and the fine-tuned one, including both qualitative and quantitative measures. The relevant code base is provided as an electronic attachment. Unfortunately, due to the issues discussed below, I was not able to verify whether the novel dataset (Chapter 5) – or at least the code required to re-create it – is attached.

Overall, I think that the research question approached by the student is fundamental and

well-motivated. In recent years, thanks to more efficient usage of unlabeled data and various improvements to the training paradigm, novel VLMs are leading popular benchmarks and show great empirical results on down-stream tasks. It was shown, however, that they tend to make shortcuts, leading to a line of works trying to break the models, and a separate one trying to improve their robustness. It should be highlighted that this thesis builds upon a very recent TEST OF TIME work by Bagad et al. (CVPR 2023), extending it in a non-trivial way – authors of that work consider only the "before/after" temporal relations. From that perspective, I consider the dataset/model choices and the applied techniques (probing and post-pretraining) as very appropriate. On the other hand, the thesis is, in my opinion, imbalanced and, thus, sometimes difficult to follow. I think that the experimental setup is described too briefly (the thesis has less than 45 pages of content), and key technical aspects are not discussed. Some of the hyperparameter choices (e.g., a learning rate of 5e-6) differ from the commonly used ones and deserve some explanation.

Despite those fallbacks, I fully recommend the thesis be defended.

Major comments:

1. **Chapter 1** – The term "Vision and Language Model" (VLM), understood in a broader sense, applies to (neural) machine learning models capable of consuming visual (vision, usually image or video) and textual (language) information. MERLOT RESERVE and VIDEOCLIP, probed in this work, fall into this category. However, they were trained with the *Masked*, not *Casual* Langauge Modeling objective – they do not possess the generative capabilities. I am highlighting this as I believe that some of the introductory sections (e.g., Section 1.1 or Section 1.3) dive too deep into the partially relevant context (e.g., generative, textual LLMs) while missing important background on Question Answering – for example, the difference between *generative*- and *retrieval*-based (video) question answering is not highlighted enough.

2. In my opinion, Section 5.1.1, which describes a core contribution of this work – a temporal-aware dataset used for fine-tuning (post-pretraining) – would vastly benefit from a few additional explanatory paragraphs. Without Figure 5.1, it is difficult to grasp the idea of empty annotations, which, to my understanding, differs from the original implementation in MERLOT RESERVE. The extremely brief Section 6.4, which introduces an alternative approach to segmentation, would probably benefit even more from additional text, as it is missing an example of an annotated instance.

3. While I appreciate the code base provided, the fact that the Author extends the original repositories (provided as a whole) with his own scripts makes it difficult to differentiate between the public code base and personal contribution.

Minor comments:

1. **Section 3** – When introducing the models and the datasets used for evaluation (STAR and NExT-QA), not enough information is provided regarding the technical aspect of videos. Without the knowledge regarding the distribution of lengths, frame sampling techniques, or the frame rate (fps) used, it is difficult to contextualize information such as *"(...) for Merlot Reserve this involves shuffling the order of the 8 frames presented to the model "*. For example, 8 frames from a video lasting 60 seconds at 60fps carry different information than 8 frames from a 5-second video sampled at 24fps.
2. **Table 5.2** – The notation used in the "Annotation" column is not explained. Please compare with Table 5.1, and how $m(\cdot)$, $s(\cdot)$, $e(\cdot)$, and $t(\cdot)$ are introduced to the reader.
3. **Table 6.3** – I believe the correct form to be "Merlot Reserve results on NExT-QA" instead of *"NExT-QA Results on Merlot Reserve"*. Also, two rows are both named *Shuffled Frames*. Based on the text, one can guess that the upper one comes from the off-the-shelf Merlot Reserve and the lower one from the fine-tuned version, but it is not clear.

Questions for the defense:

1. What was the motivation for probing only the retrieval-based VideoQA models, as opposed to testing also the generative (decoder-based) ones?
2. When perturbing the videos by frame shuffling (Section 4.4 and Section 6.5), are the reported numbers the outcome of a single run or an aggregated result based on several experiments? If the latter, how does the variance compare to the difference (e.g., the accuracy of 43.31 vs 43.60 for frame-shuffled vs ordered VideoCLIP on sequence questions from STAR) between ordered and shuffled frames?
3. In Section 6.1, the Author reports the performance of the Merlot Reserve model on the NExT-QA dataset – why are the results of VideoCLIP not included, similarly to how it was done in Chapter 4?

Práci doporučuji **k obhajobě.**
Práci nenavrhuji **na zvláštní ocenění.**

Praha, 31. 01. 2024

Podpis: