PH.D Proposal

Le Zhang

1. Introduction

I am Le Zhang, currently pursuing my Master's degree in Computer Science at UdeM and Mila under the guidance of Prof. Aishwarya Agrawal. My primary research focus lies at the intersection of visual intelligence and vision-language learning. I am particularly intrigued by the challenge of empowering machines with a human-like compositional understanding of the world.

In my past research endeavors, I have delved into several works related to compositional understanding, both in the realms of natural language processing and vision-language learning. For instance, I initially applied compositional data augmentation to language, resulting in two papers published in NAACL@2022. The focus was on text understanding tasks [5] and semantic parsing tasks [2].

Given the emergence of Large Language Models (LLMs), I am convinced that the issue of compositionality in language has been addressed. However, the visual domain continues to present significant challenges. To illustrate, the current *state-of-the-art* models grapple with identifying the number of objects, understanding their spatial relationships, and discerning object attributes. I have striven to augment vision-language models with these compositional understanding abilities [3] and have submitted the resultant work to NeurIPS@2023.

Beyond these areas, my research interests span language models and their applications. For example, I have contributed to a project on pretraining language models for protein sequence generation [4] and investigated the impact of prompts in vision-language models (VLMs) on downstream visual question answering tasks [1]. Both of these papers have been submitted to NeurIPS@2023.

In an effort to better align LLMs and VLMs with real-world applications, I have explored a work that utilizes LLM to integrate information and perform multi-modal open-domain question answering. This research has been submitted to EMNLP@2023 for review.

2. Background

The ultimate goad of my research is to pioneer a visio-linguistic system enriched with worldly knowledge, proficient in perceiving, comprehending, and interacting with humans. One of the main hurdles we face is enabling the system to grasp the compositional understanding of intricate, real-world scenes while simultaneously automating the collection of world knowledge.

From a language processing perspective, current LLMs demonstrate emergent ability through next-token-prediction. However, reaching our ambitious goal requires overcoming challenges associated with vision processing. Contemporary vision models often struggle to merge pre-training tasks with downstream tasks, lacking a standardized formulation akin to next-token-prediction in language. Therefore, there is a critical need for a universal vision model, capable of performing tasks across varying levels of granularity and trained through unsupervised learning methods. Currently, models are overly specialized, requiring fine-tuning to accomplish particular tasks.

Another issue stems from the inherent inability of existing vision models to glean compositional features from complex scenes. In lieu of this, they are prone to learning a single, entangled representation, aiming to condense all information into a singular representation without explicit modeling of the image's compositional information hierarchically. This shortcoming originates from the limitations imposed by datasets like ImageNet, which primarily center on single objects. When applied to complex, naturally occurring images, the classic training methodology collapses. This, in turn, impacts the compositional understanding of vision models adversely, since these models do not require the composition information within their pre-training objectives. Therefore, it's crucial to design and implement solutions that can address these challenges, improving the overall effectiveness of our visio-linguistic system.

3. Research Proposal

The primary aim of my doctoral research is to construct a novel vision foundation model capable of learning object-specific representations and their interrelations within complex scenes, in an unsupervised manner. Ideally, this model should be capable of performing all vision tasks without the need for specific annotations, while autonomously learning world knowledge.

Achieving this ambitious goal necessitates several steps. First, we plan to develop an unsupervised segmentation method, capable of learning pixel-level and object-level representations purely from pixel data. Following this, we aim to pretrain the model on real-world video datasets. The rationale being that world knowledge can be gleaned from observing object interactions within these videos, such as the *apple falling from the tree* or a *bird flying to the sky*.

This approach will result in a model capable of learning hierarchical representations and performing tasks at all granularity levels. As the model is geared to learn object-centric representations hierarchically, it necessitates compositional understanding during pretraining, thereby enhancing its capability in downstream tasks.

The subsequent step involves grounding this model with language, thereby aligning visual knowledge with textual understanding. This interweaving of visual and textual information would significantly bolster the model's effectiveness, bridging the gap between vision and language in a manner that is conducive to comprehensive understanding.

To date, there are no existing works that endeavor to unify all the components I have proposed above, rendering this novel idea both ambitious and challenging. Nevertheless, I am confident in my capacity to undertake this task, bolstered by my extensive research experience in compositionality and my genuine motivation for this line of work.

I find it incredibly enriching to reflect on the human cognitive processes and draw inspiration from the field of cognitive science. This human-centric approach is instrumental in guiding my research and innovation in machine learning. With the skills I have honed through my prior academic pursuits, along with the unwavering dedication I bring to my work, I am ready to take on the challenges of doctoral research. I believe that my unique perspective, coupled with my drive for discovery and understanding, will significantly contribute to our shared goal of advancing the frontiers of knowledge in this exciting field.

4. Introduction

I am Le Zhang, specializing in the area of visual intelligence and vision-language learning. My current research under Prof. Aishwarya Agrawal at UdeM and Mila is focused on equipping machines with a human-like compositional understanding of the world. Given my established track record, having worked on compositional understanding in both natural language processing and vision-language learning, I am prepared and motivated to tackle the challenges of my PhD journey.

5. Background

In recent years, Large Language Models (LLMs) have shown advancements in the domain of compositionality in language. The challenges in the visual domain, however, remain more profound. Current state-of-the-art models often fall short in capturing compositional features from scenes. Their training often revolves around individual objects rather than understanding the scene as a whole. My prior research has sought to bridge these gaps, which culminated in multiple papers, including contributions to the pretraining of language models for protein sequence generation and examining the role of prompts in vision-language models.

6. Research Proposal

My PhD research is committed to the construction of compositional and fine-grained understanding methods. The methods aim at learning not just object-specific representations, but also their relations within scenes, without relying on specific annotations. To be more specific, several following aspects could be explored under this direction:

- **Unsupervised Segmentation** I plan to initiate with an unsupervised segmentation technique that derives pixel-level and object-level representations solely from pixel data.

- **Pretraining with Real-world Videos** Integrating the visual knowledge with textual understanding is the subsequent challenge, intending to align both modalities for an enriched understanding.

- **Grounding in Language** Integrating the visual knowledge with textual understanding is the subsequent challenge, intending to align both modalities for an enriched understanding.

This proposed methodology is pioneering. No existing works have aimed to unify the components I envisage, making my research path both innovative and demanding. Nevertheless, with my background in compositionality and genuine enthusiasm for this research, I'm confident of the positive impact of my endeavor.

I often find myself inspired by the cognitive processes in humans and believe that this human-centric perspective, combined with my research experience and dedication, positions me uniquely to contribute meaningfully to the field.

References

[1] Rabiul Awal, Le Zhang, and Aishwarya Agrawal. Investigating prompting techniques for zero- and few-shot visual question answering, 2023. 1

[2] Jingfeng Yang, Le Zhang, and Diyi Yang. Subs: Subtree substitution for compositional semantic parsing, 2022. 1

[3] Le Zhang, Rabiul Awal, and Aishwarya Agrawal. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic fine-grained understanding, 2023. 1

[4] Le Zhang, Jiayang Chen, Tao Shen, Yu Li, and Siqi Sun. Enhancing the protein tertiary structure prediction by multiple sequence alignment generation, 2023. 1

[5] Le Zhang, Zichao Yang, and Diyi Yang. TreeMix: Compositional constituency-based data augmentation for natural language understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5243–5258, Seattle, United States, July 2022. Association for Computational Linguistics. 1