PH.D Proposal for Human-like Compositional Understanding of World

Le Zhang

1. Introduction

I am Le Zhang, with an expertise in visual intelligence and vision-language learning. Under the mentorship of Prof. Aishwarya Agrawal at UdeM and Mila, I've delved deep into the realm of imparting machines with a compositional understanding akin to human cognition. My contributions in both natural language processing [2, 5] and vision-language learning [1,3] stand testament to my qualifications and motivation to confront the multifaceted challenges of a PhD journey.

2. Background

The recent ascendancy of Large Language Models in encapsulating compositionality in language is commendable. Yet, the visual domain still poses intricate challenges. Notably, many state-of-the-art models are hampered by their myopic focus on individual objects, rather than holistically deciphering a scene's intricate composition. My past endeavors include innovative approaches to ameliorate these shortcomings. For instance, my research has propelled advancements like the conceptualization of novel loss functions and meticulous model fine-tuning on select datasets for nuanced comprehension [4]. Additionally, I've probed into the pivotal role of prompts in vision-language models [1].

3. Research Proposal

For my doctoral research, I envision a trajectory committed to refining machines' human-like compositional cognition and bolstering their ability to autonomously gather world knowledge. At its core, this trajectory seeks to learn intricate object representations and unravel the web of their interrelations within multifaceted real-world scenes, all without a crutch of specific annotations. Delving deeper into this vision, I propose exploration across several pivotal dimensions:

- **Unsupervised Segmentation:** The foundation lies in sculpting an unsupervised segmentation methodology that gleans both pixel-level and object-level insights purely from pixel data.
- **Pretraining with Real-world Videos:** Leveraging real-world video datasets will be paramount, capturing the intricate dance of object interactions and the stories they narrate. With my past pretraining experience [4], I'm confident to fulfill the goal.
- **Grounding in Language:** A harmonious blend of visual wisdom with textual sagacity is quintessential. This phase seeks to seamlessly align these modalities, forging a richer tapestry of understanding.

This proposed research blueprint stands on the vanguard of innovation. While the domain lacks endeavors striving to amalgamate the facets I envisage, I am driven by an unyielding spirit of innovation and challenge. My substantial groundwork in compositional understanding combined with an undying passion augments my confidence in sculpting a paradigm shift.

Regularly, I find my muse in human cognitive processes, drawing parallels and insights for machine learning. This human-centric philosophy, juxtaposed with my rich research tapestry and unwavering dedication, uniquely qualifies me to chart novel territories and indelibly mark the annals of this exhilarating 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