Understanding Multimodal Understanding

Xuejun Zhang

The fundamental question of whether artificial intelligence can achieve true understanding has become increasingly central as language models demonstrate ever more impressive capabilities. While large language models can generate coherent text and solve complex problems, the challenge lies in how we evaluate and verify their true understanding. Much like how a person can follow a recipe without understanding cooking, or plug numbers into formulas without grasping mathematical concepts, models can exhibit superficial competence through sophisticated pattern matching without developing deeper comprehension. This reflects what calls "cookbook understanding" - the ability to follow procedures and produce correct outputs without grasping the underlying principles or being able to creatively apply knowledge in novel contexts.

While assertions (like those in unit tests) can enable semantic emulation for languages with strong transparency, the task becomes provably impossible for languages where expressions can take different values in different contexts - a property essential to most practical programming and natural languages [8]. These limitations point to a critical gap: achieving true understanding may require additional forms of grounding beyond basic symbols of the languages. Multimodal learning offers a promising direction to bridge this gap, as it provides opportunities for grounding language understanding in visual and other perceptions, similar to how humans develop understanding through multiple sensory inputs and experiences.

To address these challenges in artificial understanding, I approach this problem from three angles: (1) developing systematic evaluation methodologies to expose limitations in multimodal understanding, (2) employing post-training alignment toward different forms of understanding, and (3) understanding how models understand through a mechanistic approach.

Evaluation for Multimodal Understanding

My journey into understanding multimodal systems began during my first research project in the SLED lab at my senior year. Observing how current evaluation methods often oversimplified model capabilities by focusing on surface-level metrics, I was driven to develop more specific evaluation approaches to expose specific failures in model understanding - revealing where and how models fall short of true undestanding. This led to my Recognition-based Object Probing Evaluation (ROPE) framework, which was published at **NeurIPS 2024 as a co-first author** contribution [1].

The project emerged from a critical observation: while existing benchmarks [4, 6] showed promising results for simple tasks like single-object recognition, they failed to capture deeper limitations in multimodal understanding. Rather than developing another comprehensive benchmark, I chose to focus on **revealing specific model understanding limitations**. This approach was inspired by how scientific understanding often advances through careful examination of edge cases and limitations rather than broad performance metrics.

Working with MSCOCO-Panoptic and ADE20K datasets under both seen and unseen cases, I developed novel data processing pipelines that specifically targeted multi-object scenarios. The challenge was not just technical implementation, but conceptual: how could we meaningfully evaluate "understanding when the concept itself is so broad and ill-defined? This led me to design both student-forcing and teacher-forcing evaluation regimes that probed models' reliance on textual patterns versus visual understanding, testing this across state-of-the-art models like LLaVA, GLAMM, and Groundhog.

Our findings were revealing and somewhat surprising: models that appeared highly capable in standard benchmarks showed fundamental limitations in their multimodal understanding. Through careful analysis of logit patterns and model behavior, I discovered that visual information contributed less than 20% to model predictions - a stark indication that these systems were relying more on statistical patterns than true visual understanding. Particularly in multi-object scenarios, models exhibited strong biases toward textual patterns rather than building robust cross-modal connections.

These insights pointed to a deeper issue: the limitations we uncovered weren't simply matters of scale or data, but a fundamental rethinking of how models develop and maintain connections between language and visual understanding. This realization has shaped my subsequent research direction, pushing me to explore not just how to evaluate models, but how to fundamentally understand how models understand.

Alignment Towards Pluralistic Understanding

Model understanding does not have a single fixed standard. Model understanding needs to be pluralistic, capable of representing diverse perspectives and capabilities just as human understanding manifests differently across various contexts and tasks [9]. My empirical investigations revealed an intriguing challenge: models often face a trade-off when fine-tuned for specific behaviors- when optimized for instruction following or 3H(helpful, honest, harmless) preference, etc., they tend to lose their visual grounding capabilities. This observation led me to question the conventional sequential fine-tuning approach and explore a question: how can we align models toward specific desired forms of understanding while preserving their capabilities?

This investigation led to the development of our Aligning Vision-Language Models for Multi-Objective Coordination with Adaptive Optimization(AVOCADO) framework, which reframes alignment as a multi-objective optimization

problem. Rather than using traditional sequential fine-tuning that often leads to capability conflicts, we are developing techniques combining active learning with gradient-based optimization methods like PCGrad. The framework enables more precise control over model behavior, allowing us to guide models toward specific desired characteristics while maintaining their performance across multiple objectives - essentially achieving a balance between different aspects of model understanding rather than sacrificing one for another.

Understanding Multimodal Model Understanding

My earlier evaluation work revealed a critical question: beyond identifying what models fail to understand, could we uncover how they construct understanding in the first place? Inspired by recent advances in neural network "reverse engineering" [5], I embarked on a deeper investigation into the nature of model understanding itself. This wasn't just about improving performance metrics - it was about fundamentally understanding how artificial systems process and integrate multimodal information.

This investigation led me to develop an opensource integrated analysis framework that bridges multiple interpretability approaches. I use hooks to implement toolkit synthesizes techniques including analysis of anchor tokens [3], examination of information flow patterns [10], and causal tracing through targeted neuron ablation [7].

My mechanistic analysis uncovered a pattern: image regions triggering high backpropagation saliency often correspond to background areas rather than salient objects, while attention maps show unexpected peaks in less informative regions. This phenomenon, which aligns with recent findings in vision transformers [2], suggests that models repurpose background features for internal computations rather than building understanding around salient objects.

These insights highlight a challenge in multimodal understanding: models need to learn to attach more importance to the right tokens. While models can achieve high performance by exploiting statistical patterns in background features, true understanding requires the ability to identify and focus on task-relevant objects and relationships.

Through this research, I've come to view interpretability as a fundamental approach to understanding how we understand models themselves. Just as biologists study cellular mechanisms to understand life processes, examining the internal mechanisms of neural networks reveals how these systems construct their own unique ways of processing multimodal information - ways that might differ from human cognition and that we can never fully grasp by only studying external behavior. This deeper investigation into model understanding, combining rigorous evaluation methods with mechanistic interpretability, shows us that understanding a model's behavior requires more than observing its outputs; we must examine how it constructs its own representations and processing patterns, bridging the gap between what models do and how they actually do it.

Future Plan

My research experiences have deepened my interest in natural language processing and multimodal learning. Building upon this foundation, I am eager to pursue a PhD to explore new challenges in artificial intelligence and contribute to our understanding of how these systems process and represent information.

References

[1] Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David F. Fouhey, and Joyce Chai. Multi-object hallucination in vision-language models. *ArXiv*, abs/2407.06192, 2024.

[2] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024.

[3] Qidong Li, Qingbo Li, Chenxiang Chen, Qin Jin, and Yandong Feng. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2312.14534*, 2023.

[4] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023.

[5] Jack Lindsey, Adly Templeton, Jonathan Marcus, Thomas Conerly, Joshua Batson, and Christopher Olah. Sparse crosscoders for cross-layer features and model diffing. *Anthropic Technical Report*, 2024.

[6] Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. Negative object presence evaluation (nope) to measure object hallucination in vision-language models, 2024.

[7] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *arXiv preprint arXiv:2202.05262*, 2023.

[8] William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A Smith. Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? *arXiv preprint arXiv:2104.10809*, 2021.

[9] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. A roadmap to pluralistic alignment, 2024.

[10] Zihao Wei, Yibing Liu, Yixuan Fan, Jiaqi Liu, Yunze Wang, Xiang Li, Guanzhen Wu, Tingting Xu, Yiqing Wang, and Haizhou Lin. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160*, 2023.