Towards Robust Visual Understanding: from Recognition to Reasoning

Tejas Gokhale

University of Maryland, Baltimore County
gokhale@umbc.edu

The connection between vision and language (V+L) is now an integral part of AI, with deep impact in not only in vision, but in adjacent fields such as robotics and human-computer interaction. I refer to this paradigm as *semantic vision*, where *meaning*, and by proxy, natural language, serves as a critical source of knowledge for modern computer vision algorithms that seek to understand the visual world. While previous decades in vision were dominated by *"the three R's"* (reconstruction, recognition, and reorganization) (Malik et al. 2016), the success of semantic vision has created a fourth *"R"* of computer vision: *reasoning*. The ability to reason requires not only prediction, but a combination of active perception, language grounding, world knowledge, and speculation beyond the observable.

Models that learn from data are widely and rapidly being deployed today for real-world use, but they suffer from unforeseen failures due to distribution shift, adversarial attacks, noise and corruption, and data scarcity. But many failures also occur because many modern AI tasks require reasoning beyond pattern matching – and such reasoning abilities are difficult to formulate as data-based input-output function fitting. The reliability problem has become increasingly important under the new paradigm of semantic "multimodal" learning. My research provides avenues to develop robust and reliable computer vision systems, particularly by leveraging the interactions between vision and language.

In the AAAI New Faculty highlights talk, I will cover three thematic areas of my research, described below. Readers are encouraged to refer to my doctoral dissertation Gokhale (2023) and publications (www.tejasgokhale.com).

(1) Knowledge-Guided Adversarial Discovery of Transformations for Robust Image Classifiers. Our efforts towards improving the robustness of image classifiers have focused on leveraging domain knowledge in order to discover diverse data augmentations that can expose the classifier to a larger distribution during training. This theme has led to improved robustness in several settings such as attribute-level shift, style-shift, and domain-shift including applicability to domain generalization for satellite images and medical images. I will discuss important dimensions of machine learning reliability and our recent finding of curious trade-offs between them when using data modification.

(2) Open-Domain Reliability for Visual Reasoning. Multi-modal tasks involving both vision and language (V&L) inputs combined with the quirks, complexity, and ambiguities of human language, open up intriguing domain discrepancies that can affect model performance at test time. I will demonstrate our findings of V&L models when dealing with logical compositions, semantic and syntactic perturbations of questions and sentences. This will be followed by a description of techniques to mitigate these failures through knowledge-guided regularization and data engineering, resulting in improvements along several dimensions of robustness. in image-based reasoning, video-based reasoning, and visual question answering tasks.

(3) Evaluation of Generative Vision Models: Challenges and Opportunities. Text-to-image (T2I) models have seen exceptional advances in less than a decade. But when it comes to evaluation of generated images, metrics of photorealism dominate the discourse. T2I models have much more to offer than photorealism; to do justice to this potential, we need extensive evaluation tools and benchmarks. My lab is actively pursuing the development of evaluation frameworks to quantify the reliability and abilities of T2I – this includes benchmarking spatial reasoning abilities via a challenging dataset ($\mathrm{SR}_{\mathrm{2D}}$) and automated evaluation metrics (VISOR) and evaluation of the ability to learn, reproduce, and compose visual concepts, offered by ConceptBed.

Future Directions: My lab is exploring applications of vision-language techniques in mission-critical domains where large-scale image data is unavailable, expensive, or unlabeled, but expert knowledge about images is available in language form. Integrating this natural language knowledge will help guide important decisions while also improving their reliability. Rigorous, exhaustive, holistic, and trustworthy evaluation and benchmarking is possibly the biggest challenge in multimodal systems today, and I intend to be at the forefront of this community effort.

References

Gokhale, T. 2023. *Towards Reliable Semantic Vision*. Ph.D. thesis, Arizona State University.

Malik, J.; et al. 2016. The three R's of computer vision. *Pattern Recognition Letters*, 72: 4–14.