Research Statement: Tejas Gokhale

tejasgokhale.com

My mission is to research and develop robust and reliable AI systems by leveraging the complex interactions between images and natural language. My research towards this goal lies at the wonderful intersection of machine learning, computer vision, and natural language processing. My research has two central goals: (a) to design algorithms powered by semantic data engineering techniques to improve robustness, interpretability, calibration, and reliability, and (b) to develop benchmarks, evaluation protocols, and to discover and mitigate emerging failure modes of semantic vision models. This mission is directly aligned with the clarion call for safe and robust AI systems – from DARPA[1], White House OSTP's "AI Bill of Rights"[2], ACL[3], and AAAI[4].

Investigate Failure Modes

- *New types of distribution shift*
- *Test-time perturbations and attacks*
- *Grounded Reasoning Abilities*

Leverage Complex Relations between Vision and Language

- *Utilize semantic and linguistic variations of image descriptions for robust optimization*
- *Build benchmarks using semantic knowledge*

Discover Transformations for Boosting Robustness

- *Use information from models to discover optimal data augmentation functions.*
- *Data engineering cannot be static –develop model-in-the-loop augmentations*

In the past decade, we have witnessed a paradigm shift in computer vision – the connection between language and vision *("semantic vision")* is now an integral part of AI, with deep impact on NLP, vision, robotics, graphics, and direct industrial implications for software, arts, media, and journalism. Recent studies in ML have revealed alarming failures due to distribution shift and dataset biases and threats of adversarial attacks; fairness concerns and undesirable societal implications have also emerged [1]. As semantic vision models become widely adopted, new types of challenges and failure modes will emerge (as I have shown through my research) – I aim to be at the forefront of enhancing the reliability of these high-impact systems.

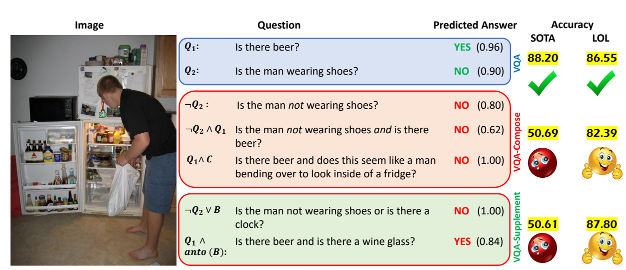
A series of my work has addressed robustness and generalization of image classification **[AAAI'21, ACL'22, WACV'23]**, visual question answering **[ECCV'20, EMNLP'20, ACL'21, ICCV'21, EMNLP'22]**, vision-language inference **[ACL'22]**, video captioning **[EMNLP'20]** natural language understanding **[NAACL'21, ACL'22, AAAI'22]**. These publications have led to successful grant proposals that I helped in writing; for eg. a funded NSF Robust Intelligence grant[5]) which builds on my work on robustness in vision and language. I am part of several collaborative projects with ASU, Lawrence Livermore National Laboratory, Microsoft Research, Carnegie Mellon, and Adobe Research.

Selected work is described below. Section (A) describes work in multimodal (V+L) understanding, while Section (B) addresses image classification. The underlying common theme is the combination of active design and discovery of data transformations and adversarial training algorithms for improving robustness.

(A) Robust Multimodal (Vision+Language) Perception

▶ Robustness to Logical Transformations in Visual Question Answering [ECCV 2020] [2]

Multi-modal tasks involving both vision and language (V&L) inputs, such as visual question answering (VQA), open up intriguing domain discrepancies that can affect model performance of test time. For the VQA task, models are trained to predict the answers to questions about images. My first paper, VQA-LOL [2], discovered that existing VQA models fail when logical transformations such as negation, conjunction, and disjunction are introduced in the questions. I built on this surprising finding to develop a data augmentation tool that produces logical combinations of multiple questions in the source dataset. I also designed a logic-inspired training objective that is based on Frechet inequalities to guide the predicted probabilities of answers to questions with logical connectives. VQA-LOL was quickly appreciated by the V&L

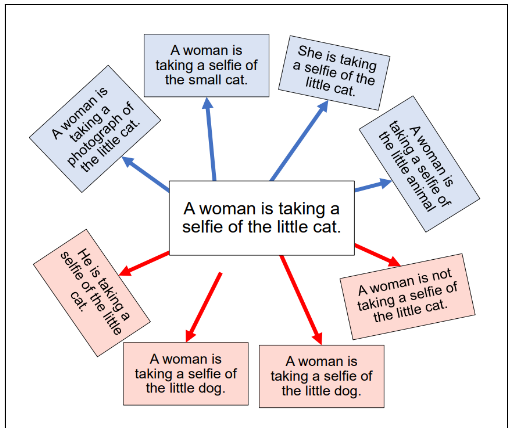


VQA-LOL revealed that SOTA VQA models answer questions from the VQA dataset (Q_1, Q_2) correctly, but fail to answer logical compositions including negation, conjunction, disjunction.

community and was adopted as part of a compendium of datasets for testing VQA robustness [3], and led to a series of papers [4, 5, 6] that adopted linguistic and semantic transformations. With collaborators, I have also built weakly-supervised VQA models that learn with limited or synthetic data [7, 8], and video QA benchmarks for reasoning about implicit physical properties of objects [9].

▶ Semantically Distributed Robust Optimization Improves Vision–Language Inference [ACL 2022] [5]

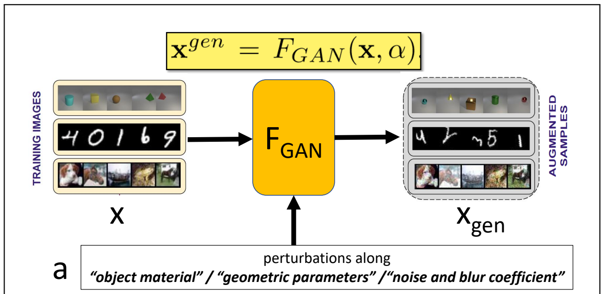
I identified that knowledge of linguistic transformations can inform the design of algorithms for improving performance on V+L tasks. A carefully designed set of probing experiments led me to develop the "SISP transformation" suite – a controlled method to semantically manipulate text to generate augmented data that is semantics-inverting (SI) or semantics-preserving (SP). I showed that these SISP transformations can be leveraged to train robust models by developing a new algorithm called *Semantically Distributed Robust Optimization (SDRO)*. The combination of SISP (data engineering) and SDRO (robust optimization) led to improvements on image-based reasoning, video-based reasoning, and visual question answering, along several dimensions of robustness – in-domain and out-of-domain accuracy, adversarial robustness, and calibration. Our method also improved performance on my previous VQA-LOL benchmark [2].



Training with SISP transformations improves accuracy, robustness & calibration of V+L models.

(B) Robust Image Classification and Domain Generalization

▶ Robustness under Attribute Shift [AAAI 2021] [10]



AGAT, an adversarial training algorithm, supports a wide-range of attribute specifications and improves robustness.

Previous work on robust image classification focuses on pixel-level adversarial attacks. However, in real world scenarios, test examples can vary along known attributes such as size, shape, colors, and geometric transformation. Unfortunately, these cannot be covered by methods that utilize norm-bounded and additive pixel-level perturbations. We consider a setting where information about the target domain is available only in terms of a set of attributes that are known to differ at test time – there is no access to a target validation set, or knowledge about the magnitudes and combinations of attributes at test time. As such, standard data augmentation and pixel-level adversarial training is ineffective. I developed a new form of adversarial training: *Attribute-Guided Adversarial Training (AGAT)* that parameterizes the input space in terms of attributes, and adversarially perturbs image attributes to maximize exposure of the classifier to previously unobserved variations. AGAT supports a wide-range of attribute specifications, which we demonstrate with large gains in three different use-cases: (1) object-level attribute-shift (2) geometric transformations (3) common natural corruptions.

▶ Improving Diversity with Adversarially Learned Transformations [WACV 2023] [11] *Single source domain generalization (SSDG)* is a challenging setting, where the model has access only to a single training domain (eg. real photos), and is expected to generalize to multiple testing domains with domain shift (eg. sketches and cartoons). Unlike the setting for AGAT, there is no access to attributes or external knowledge about the nature or magnitude of domain shift. Success of SSDG depends on maximizing diversity of training data; this naturally implies that data augmentation is one of the main sources of diversity! But what augmentation method should we choose? We found that pre-specified augmentations [12, 13] cannot model large domain shift in SSDG effectively. This led to our novel framework that discovers adversarially learned transformations (ALT), by perturbing the parameter space of an "adversity" network to model plausible yet hard image transformations. ALT offers a synergy between diversity and adversity, exposing the model to increasingly unique, challenging, and semantically diverse examples – ideally suited for SSDG. ALT's ability of improving the training diversity resulted in performance gains over all existing techniques, including standard data augmentation and pixelwise adversarial training, on multiple domain generalization benchmarks.

▶ Future Research Agenda

Over the last decade, the AI community has barely scratched the surface when it comes to the leveraging the complex interactions between the visual world and the meaning humans assign to it via language. V+L research is like an iceberg – we have only scratched the surface; there are even more facets to robust V+L submerged under the water, waiting to be discovered. My future research agenda will run two parallel programs: (1) developing systems that combine explicit knowledge and data-driven learning to perform complex reasoning and interact with humans, and (2) through these interactions improving the reliability, transparency, and robustness of machine learning models. Target funding sources will include: NSF's Robust Intelligence (RI) and Human-Centered Computing (HCC) and SaTC, relevant DARPA programs such as ECOLE, IARPA's BETTER, HIATUS, TrojAI, and relevnt future programs announced by DoE, NIH, ONR, etc.

Towards Reliable Visual Reasoning. The link between vision and language is much more complex than simple image–text similarity. Language is ideally suited for developing reasoning capabilities beyond the visible – physical and spatial reasoning, embodied reasoning, and commonsense reasoning (see my collaborations **[EMNLP'20]** [14] for commonsense video captioning, and **[CVPRW'21]** [15] for reasoning about object co-occurence). My core research agenda is to develop robust methods for visual reasoning. I have taken concrete steps in this direction:

- Spatial reasoning is a fundamental aspect of computer vision. With collaborators from Microsoft Research, I have developed an evaluation framework called *"VISOR"* for quantifying the fidelity of text-to-image synthesis models in generating spatial relationships between objects. VISOR reveals the surprising finding that although recent SOTA models like DALLE exhibit high photorealism, they are ineffective in composing images with two or more distinct objects. We curated a dataset to enable future research in this direction.
- I am investigating how V+L models like CLIP [16] can be used for reasoning about everyday actions and the linked commonsense aspects (eg. people often kick footballs, but rarely kick brick walls). A parallel effort will be dedicated to revealing spurious biases between text, actions, and objects by studying the topology of multimodal models via counterfactually augmented images, such as when objects are replaced, removed, or moved within the image.

Human-Computer Interaction and Augmented Reality. In the last five years, the nature of work in V+L has evolved from research prototypes to bringing about a paradigm shift in AI. I am convinced that the use of language has immense potential in changing the way we interact with AI, and it is the way forward for democratizing and simplifying access to graphics and robotics. This will engender exciting new technologies, but they will be accompanied by risks and threats due to proliferating deep-fakes [17]. I plan to expand my work into the domains of Human-Computer Interaction and Augmented/Virtual Reality to develop techniques that make the fullest use of language, while mitigating the security threats and failure modes.

Connections between Adversarial and Distributional Robustness. In the long term, I am interested in understanding theoretical connections between adversarial and distributional robustness, especially when multiple modalities such as images, videos, text, and audio are involved. Distribution shift and "OODnesss" can often manifest in different ways in images vs in language. ML theory is often limited to single modalities, but the effect of complex interactions between different modalities remains unexplored. I recently pursued an empirical investigation **[ACL'2022]** [18] which found that data filtering methods with good intentions of removing spurious correlations in training data, can end up hurting adversarial robustness. This finding has been recently corroborated by other researchers [19]. I plan to pursue this direction and expect theory and empirical evidence to lead to actionable design considerations for building robust ML models.

References

[1] Lu Cheng, Kush R Varshney, and Huan Liu. Socially responsible ai algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research*, 71:1137–1181, 2021.

[2] **Tejas Gokhale**, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Vqa-lol: Visual question answering under the lens of logic. In *European conference on computer vision*. Springer, 2020.

[3] Linjie Li, Zhe Gan, and Jingjing Liu. A closer look at the robustness of vision-and-language pre-trained models. *arXiv preprint arXiv:2012.08673*, 2020.

[4] **Tejas Gokhale**, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 878–892, Online, 2020.

[5] **Tejas Gokhale**, Abhishek Chaudhary, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Semantically distributed robust optimization for vision-and-language inference. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1493–1513, 2022.

[6] Neeraj Varshney, Pratyay Banerjee, **Tejas Gokhale**, and Chitta Baral. Unsupervised natural language inference using phl triplet generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2003–2016, 2022.

[7] Pratyay Banerjee, **Tejas Gokhale**, Yezhou Yang, and Chitta Baral. Weaqa: Weak supervision via captions for visual question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3420–3435, 2021.

[8] Pratyay Banerjee, **Tejas Gokhale**, Yezhou Yang, and Chitta Baral. Weakly supervised relative spatial reasoning for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1908–1918, October 2021.

[9] Maitreya Patel, **Tejas Gokhale**, Chitta Baral, and Yezhou Yang. Counterfactual reasoning about implicit physical properties via video question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

[10] **Tejas Gokhale**, Rushil Anirudh, Bhavya Kailkhura, Jayaraman J Thiagarajan, Chitta Baral, and Yezhou Yang. Attribute-guided adversarial training for robustness to natural perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7574–7582, 2021.

[11] **Tejas Gokhale**, Rushil Anirudh, Jayaraman J Thiagarajan, Bhavya Kailkhura, Chitta Baral, and Yezhou Yang. Improving diversity with adversarially learned transformations for domain generalization. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

[12] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[13] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. In *International Conference on Learning Representations*, 2020.

[14] Zhiyuan Fang, **Tejas Gokhale**, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Video2Commonsense: Generating commonsense descriptions to enrich video captioning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 840–860, Online, 2020. Association for Computational Linguistics.

[15] Kuldeep Kulkarni, **Tejas Gokhale**, Rajhans Singh, Pavan Turaga, and Aswin Sankaranarayanan. Halluci-net: Scene completion by exploiting object co-occurrence relationships. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021. https://arxiv.org/abs/2004.08614.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[17] Eric Horvitz. On the horizon: Interactive and compositional deepfakes. *arXiv preprint arXiv:2209.01714*, 2022.

[18] **Tejas Gokhale**, Swaroop Mishra, Man Luo, Bhavdeep Sachdeva, and Chitta Baral. Generalized but not robust? comparing the effects of data modification methods on out-of-domain generalization and adversarial robustness. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2705–2718, 2022.

[19] Mazda Moayeri, Kiarash Banihashem, and Soheil Feizi. Explicit tradeoffs between adversarial and natural distributional robustness. *NeurIPS 2022*, 2022.

---

[1] https://www.darpa.mil/work-with-us/ai-next-campaign

[2] https://www.whitehouse.gov/ostp/ai-bill-of-rights/

[3] https://2023.aclweb.org/calls/main_conference/#theme-track-reality-check

[4] https://aaai.org/Conferences/AAAI-23/safeandrobustai/

[5] https://nsf.gov/awardsearch/showAward?AWD_ID=2132724