Mark Yatskar
Research Statement

My research is motivated by a desire to create intelligent machines that we can work collaboratively with in a diverse set of environments. My work primarily looks at the interplay between natural language and visual perception for the purpose of building such a machine.

I study how language can be used to structure visual understanding in AI systems.

I am designing new machine learning approaches that enable tight coupling between how people express themselves in language and how machine behavior is specified. Coupling language to behavior can create explicit representations that are auditable and debuggable, allowing for more reliable models that are more ethical to deploy. For example, my lab, in collaboration with University of Pennsylvania radiologists, recently introduced Knowledge-enhanced Bottlenecks (KnoBo) [1], a class of inherently interpretable models that can reason with clinically relevant factors found in PubMed (Figure 1). KnoBo integrates strong priors from verified research articles into prediction on medical images by leveraging enhanced controlability. It can accurately avoid age, sex , race and other hospital specific confounds commonly found in medical datasets that are not relevant to target predictions. The approaches my lab is developing can form the basis of safer model behavior in a broader range of environments.

A severe limitation of existing AI systems is that they are largely opaque, making it difficult to understand their limitations and how to improve them. My research philosophy is that scientific insight and new methodology should go hand-in-hand. My work focuses on both establishing limitations of current systems and exploring new designs. Such an approach allows for the creation of long-lived insight while simultaneously contributing new methodology. In many cases, this approach has resulted in the creation of large scale datasets defining areas where models currently fail that can serve as benchmarks for the future [2; 3; 4; 5].

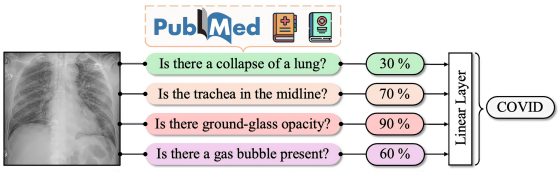

Figure 1: An excerpt of a concept bottleneck that KnoBo [1] derives from PubMed, expressed as a set of questions about a scan. Structuring medical image understanding around explicit knowledge from PubMed builds inherent robustness. KnoBo is significantly more resilient to confounding than fine-tuned vision transformers, improving on average 32% across two medical imaging modalities. KnoBo enables novel transfer scenarios, such as developing a model in one hospital and deploying it in another.

With equal emphasis on model development and analysis, my group has discovered new ways of predicting catastrophic failures in models. Models often adopt shortcuts (rules of thumb), that result in biased or incorrect predictions. In encountering such limitations, my research has sought to deepen our understanding, forming connections to psychological research. A fundamental challenge in this area is establishing relationships between how models are trained and their resulting brittleness. A new finding from my lab, in collaboration with a University of Pennsylvania psychologist, is that if data annotators are more likely to use cognitive heuristics then models trained on their data are more brittle [6]. Such heuristic use is often not detectable on individual data samples but instead via psychological testing of data creators. People who are more likely to engage in rule of thumb thinking transfer such behavior to models via data in aggregate.

My lab's research is highly interdisciplinary, spanning the areas of natural language processing, computer vision, and machine learning while making connections to other fields like medicine and psychology. Concretely, my lab has explored the following themes:

- Scaffolding visual intelligence with natural language [7; 8; 2; 9; 10; 11; 12; 13; 14; 15; 1]
- Avoiding and understanding bias to build resilience to unintended data correlations. [16; 1; 17; 18; 6]

- Evaluating limitations of existing systems. [1; 3; 4; 19; 20; 5]

The following sections provide a high level summary of my recent advances in these three areas and then highlight some future work where we can make further progress.

1 Scaffolding Visual Intelligence with Natural Language

Natural language provides a rich signal for what people can perceive. It also works as a cultural store of information: important aspects of the world are named, described, and repeated. My research has explored how to build systems that rely on language as an explicit representation to accomplish a diverse set of tasks, including object classification [12], event recognition [8; 2], visual question answering [13], motion editing [15] and medical image understanding [1]. I have hypothesized that explicit structure within systems will allow for more systematic evaluation, a higher degree of control, and greater interpretability that can be used for verification. A central challenge with explicit representations is showing that they can reach similar levels of accuracy as those implicitly constructed in deep neural networks. Realizing the practical benefits of explicit systems is hard because scenarios where they might be preferred are challenging for all models (i.e. high stakes environments). In recent years, my group has addressed both of these challenges.

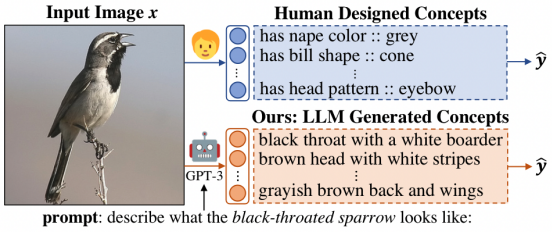

Figure 2: An excerpt of LaBo generated concepts, where each statement defines a factor for classifying an object. LaBo alleviates the need for human-designed concepts by prompting large language models such as GPT-3 to generate rich spaces of possible bottleneck models. Optimizing over this space allows for the joint selection of both high accuracy and highly interpretable explicit models. In a broad set of experiments, LaBo outperforms models that do not have explicit factorizations in low data regimes by over 10%, while performing comparably when dataset are large.

At CVPR 2023, my group proposed Language Model Guided Concept Bottlenecks (LaBo) and showed we could construct explicit language based models that perform similarly to implicit ones in a broad set of image classification tasks [12]. Our proposal builds on Concept Bottleneck Models (CBM), a class of inherently interpretable models that factor model decisions into humanreadable concepts [21]. Like many explicit models, CBMs require manual specification and often under-perform their counterparts. We address these shortcomings and are first to show how to construct high-performance CBMs without manual specification of similar accuracy to black box models. As seen in Figure 2, LaBo leverages a language model, GPT-3 [22], to define a large space of possible bottlenecks. Given a problem domain, LaBo uses GPT-3 to produce factual sentences about categories to form candidate concepts. LaBo efficiently searches possible bottlenecks through a novel submodular utility that promotes the selection of discriminative and diverse information. Ultimately, GPT-3's sentential concepts can be aligned to images using CLIP [23], to form a bottleneck layer. LaBo is a highly effective prior for concepts important to visual recognition, outperforming models without explicit language layers in low data regimes, and performing comparably otherwise. Beyond being highly performant, LaBo's design is simple. We were able to show its efficacy on over 10 problem settings, and it has been adopted broadly as a baseline for other explicit approaches.

LaBo forms the conceptual foundation of our recent preprint, Knowledge-enhanced Bottlenecks (KnoBo), that shows how to leverage explicit language based representations to achieve increased robustness to domain shift in medical images analysis [1]. While deep networks have achieved broad success in analyzing natural images, when applied to medical scans, they often fail in unexcepted situations. We investigated this challenge and focused on model sensitivity to domain shifts, such as data sampled from different hospitals or data confounded by demographic variables such as sex, race, etc, in the context of chest X-ray

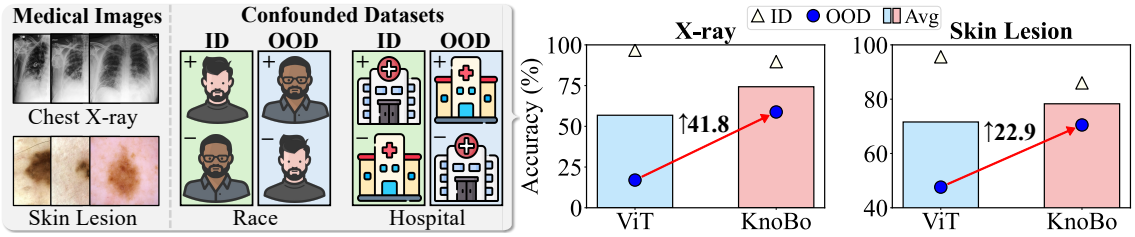

Figure 3: In-domain (ID), out-of-domain (OOD), and average of ID and OOD (Avg) performance on confounded medical image datasets. Our interpretable Knowledge-enhanced Bottlenecks (KnoBo) are more robust to domain shifts (e.g., race, hospital, etc) than fine-tuned vision transformers (ViT) [24].

and skin lesion images. A key finding we show empirically is that existing visual backbones lack an appropriate prior from the architecture for reliable generalization in these settings. Our main insight is that explicit interpretable models can play this role. Taking inspiration from medical training, we propose giving deep networks a prior grounded in explicit medical knowledge communicated in natural language. To this end, we introduced KnoBo, a class of concept bottleneck models that incorporates knowledge priors that constrain it to reason with clinically relevant factors found in medical textbooks or PubMed. KnoBo uses retrieval-augmented language models to design an appropriate concept space paired with an automatic training procedure for recognizing the concept. We evaluated different resources of knowledge and recognition architectures on a broad range of domain shifts across 20 datasets. As seen in Figure 3, KnoBo outperforms fine-tuned models on confounded datasets and substantially reduces the impact of confounding.

Future Work Labo and KnoBo show that explicit and interpretable language based representations can be accurate on a broad range of tasks. They are also sufficiently controllable with high quality priors to work well in low-data or confounded settings. In KnoBo, we showed that scientific literature in PubMed, combined with our explicit models, can be used to construct high quality predictors. **Scientific articles is an under-explored resource of real world knowledge that can be used to form priors for a broad range of problems.** Such articles are inherently multimodal and can be used to as resources for targeted search for specific problems or as general training corpora for constructing foundation models. Knowledge derived from articles needs to be integrated into end-models and our language based predictors are a natural route. There is on-going work in my lab for evaluating the efficacy of existing language model tools for retrieving and summarizing scientific literature and efforts to pretrain foundation models on drug studies for predicting novel compounds with properties specifiable with prompts.

2 Avoiding and Understanding Biases

Existing machine learning approaches require large quantities of data that make it difficult to understand what exactly is being learned. This lack of transparency can result in unpredictable behavior where models fail to generalize in unnatural ways when conditions deviate slightly from training. My research has sought to both lend scientific insight to these phenomenon and develop methods that make models more robust. I have developed methods that leverage machine learned models of undesirable behavior to avoid such failures [17; 18]. These have been applied to avoid generalization errors arising in adversarial splits in vision and language systems. I have also developed methods for reducing a model's incorrect dependence on people's gender in images [16]. KnoBo, described above, is another example, where my group developed an interpretable mechanism for incorporating knowledge priors to achieve robustness to domain shift.

While it is understood that brittleness to data shifts arise from problems in training, identifying specific samples of data that are the source of the problem is challenging. My group has recently proposed to

instead study this issue from the human perspective and ask the question: if we cannot identify samples that are the problem, can we identify factors contributing to problematic annotators? In collaboration with psychologists at University of Pennsylvania, we have answer affirmatively [6]. Cognitive psychologists have documented that humans use cognitive heuristics, or mental shortcuts, to make quick decisions while expending less effort. While performing annotation work on crowdsourcing platforms, we hypothesized that such heuristic use among annotators cascades on to data quality and model robustness. We studied cognitive heuristic use in the context of annotating multiple-choice reading comprehension datasets. We were able to tangibly measure multiple low-effort annotation strategies that were indicative of usage of various cognitive heuristics. We find evidence that annotators might be using multiple such heuristics, based on correlations with a battery of psychological tests. Importantly, heuristic use among annotators determines data quality along several dimensions: (1) known biased models more easily solve examples authored by annotators that rate highly on heuristic use, (2) models trained on annotators scoring highly on heuristic use don't generalize as well, and (3) heuristic-using annotators tend to create qualitatively less challenging examples. Our findings suggest that tracking heuristic usage among annotators can potentially help with collecting challenging datasets and predicting model biases.

Future Work Our work on cognitive heuristic use is the first to establish a connection between psychological properties of people and model behavior after training. While models rely heavily on pretraining, instruction tuning and learning from preference data are increasingly important. My group has several ongoing efforts studying implicit biases in learning from preference data. Prompted queries are often underspecified, and people use their cognitive biases to to fill in details when making judgements about model outputs. Work has already shown that preference data has biases for aspects like length [25] and we hypothesize that many other biases exist. For example, models appear overconfident and vague, and we suspect that this combination makes the cognitive load of falsifying their output too high for annotation scenarios. Our goals are to evaluate if such affects exist and propose annotation frameworks for soliciting judgements on long textual output that avoids these pitfalls.

3 Evaluating Model Limitations and Benchmarking

As models achieve more complex capabilities they have become increasingly opaque. My work has sought to shed light on their limits, focusing on proposing datasets that could be used for future evaluation. In the past, I have established benchmarks for limitations in grounded domains, like event reasoning [9] and activity recognition [2], and communicative settings like question answering [26; 27] and coreference resolution [3]. With the rapid rise of large language models (LLMs) we have been left with many blind spots in our understanding of their behavior in the scenarios they being called upon to handle. LLMs are now being experimented with broadly across many disciplines and areas of society. For example, doctors are using LLMs for differential diagnosis or fast information recall [28], scientists broadly as discovery aides [29], lawyers are using them to prepare court documents [1], and business people to craft emails. People appear willing to adopt these systems despite warnings from researchers about risks. My group has hypothesized that this is in part due to a disconnect between applications researchers study and those that people are considering. Existing evaluations lack **ecological validity** – they do not reflect real world situations and therefore provide little practical guidance for society at large. Recent efforts in my group have sought to address this failing.

As language models are adopted by a more sophisticated and diverse set of users, the importance of guaranteeing that they provide factually correct information supported by verifiable sources is critical across fields of study. This is especially the case for high-stakes fields, such as medicine and law, where the risk of

[1] https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html

propagating false information is high and can lead to undesirable societal consequences. In recent work from my lab, ExpertQA [4], we conduct human evaluation of responses from a few representative systems along various axes of attribution and factuality, by bringing domain experts in the loop. Specifically, we collect expert-curated questions from 484 participants across 32 fields of study, and then ask the same experts to evaluate generated responses to their own questions. In addition, we ask experts to improve upon responses from language models. The output of our analysis a high-quality long-form QA dataset with 2177 questions spanning 32 fields, along with verified answers and attributions for claims in the answers. A central output of this effort has been establishing that across the fields we study, model output is largely already useful, but is rated as depending on unreliable sources by experts. Our effort is reusable: ExpertQA forms a benchmark for evaluating model attribution quality in ecologically valid settings constructed by experts.

Future Work Broadly, my group has been focusing on several new efforts to establish benchmarks closer to realistic use cases for models today. Continuing the theme of ecological validity, in collaboration with Google, we recently proposed Dolomites [5]. Like ExpertQA, Dolomites uses field specialists to define tasks for language models and asks them to evaluate responses. Dolomites expands ExpertQA beyond information seeking behavior to methodical tasks: writing tasks requiring synthesis and judgement that professionals do regularly. For example, a medical expert could use a model to propose physical therapy plans based on reports of symptoms. Our evaluation again shows current utility, but showing that models lack specificity required to accomplish such tasks. Explicitly asking experts to annotate data is challenging and expensive. My group has ongoing projects trying to find mechanisms for constructing ecologically valid datasets using found data. For example, we are constructing question answering datasets from scientific literature surveys. Such a resource can be used to evaluate and build information aides for scientists.

References

[1] Y. Yang, M. Gandhi, Y. Wang, Y. Wu, M. S. Yao, C. Callison-Burch, J. C. Gee, and M. Yatskar, "A textbook remedy for domain shifts: Knowledge priors for medical image analysis," arXiv preprint arXiv:2405.14839, 2024.

[2] A. Sadhu, T. Gupta, M. Yatskar, R. Nevatia, and A. Kembhavi, "Visual semantic role labeling for video understanding," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5589–5600, 2021.

[3] Y. Yuan, C. Malaviya, and M. Yatskar, "Ambicoref: Evaluating human and model sensitivity to ambiguous coreference," arXiv preprint arXiv:2302.00762, 2023.

[4] C. Malaviya, S. Lee, S. Chen, E. Sieber, M. Yatskar, and D. Roth, "Expertqa: Expert-curated questions and attributed answers," arXiv preprint arXiv:2309.07852, 2023.

[5] C. Malaviya, P. Agrawal, K. Ganchev, P. Srinivasan, F. Huot, J. Berant, M. Yatskar, D. Das, M. Lapata, and C. Alberti, "Dolomites: Domain-specific long-form methodical tasks," arXiv preprint arXiv:2405.05938, 2024.

[6] C. Malaviya, S. Bhatia, and M. Yatskar, "Cascading biases: Investigating the effect of heuristic annotation strategies on data and models," arXiv preprint arXiv:2210.13439, 2022.

[7] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," arXiv preprint arXiv:1908.03557, 2019.

[8] S. Pratt, M. Yatskar, L. Weihs, A. Farhadi, and A. Kembhavi, "Grounded situation recognition," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, pp. 314–332, Springer International Publishing, 2020.

[9] Y. Yang, A. Panagopoulou, Q. Lyu, L. Zhang, M. Yatskar, and C. Callison-Burch, "Visual goal-step inference using wikihow," arXiv preprint arXiv:2104.05845, 2021.

[10] Y. Yang, J. Kim, A. Panagopoulou, M. Yatskar, and C. Callison-Burch, "Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval," arXiv preprint arXiv:2111.09276, 2021.

[11] Y. Yang, A. Panagopoulou, M. Apidianaki, M. Yatskar, and C. Callison-Burch, "Visualizing the obvious: A concreteness-based ensemble model for noun property prediction," arXiv preprint arXiv:2210.12905, 2022.

[12] Y. Yang, A. Panagopoulou, S. Zhou, D. Jin, C. Callison-Burch, and M. Yatskar, "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19187–19197, 2023.

[13] X. Fu, B. Zhou, S. Chen, M. Yatskar, and D. Roth, "Interpretable by design visual question answering," arXiv preprint arXiv:2305.14882, 2023.

[14] C. Clark, J. Salvador, D. Schwenk, D. Bonafilia, M. Yatskar, E. Kolve, A. Herrasti, J. Choi, S. Mehta, S. Skjonsberg, et al., "Iconary: a pictionary-based game for testing multimodal communication with drawings and text," arXiv preprint arXiv:2112.00800, 2021.

[15] Y. Huang, W. Wan, Y. Yang, C. Callison-Burch, M. Yatskar, and L. Liu, "Como: Controllable motion generation through language guided pose code editing," arXiv preprint arXiv:2403.13900, 2024.

[16] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez, "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019.

[17] C. Clark, M. Yatskar, and L. Zettlemoyer, "Don't take the easy way out: Ensemble based methods for avoiding known dataset biases," arXiv preprint arXiv:1909.03683, 2019.

[18] C. Clark, M. Yatskar, and L. Zettlemoyer, "Learning to model and ignore dataset bias with mixed capacity ensembles," arXiv preprint arXiv:2011.03856, 2020.

[19] J. Magnus Ludan, Q. Lyu, Y. Yang, L. Dugan, M. Yatskar, and C. Callison-Burch, "Interpretable-by-design text classification with iteratively generated concept bottleneck," arXiv e-prints, pp. arXiv–2310, 2023.

[20] C. Malaviya, S. Lee, D. Roth, and M. Yatskar, "Pachinko: Patching interpretable qa models through natural language feedback," arXiv preprint arXiv:2311.09558, 2023.

[21] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models," in International Conference on Machine Learning, pp. 5338–5348, PMLR, 2020.

[22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., "Language models are unsupervised multitask learners," OpenAI blog, vol. 1, no. 8, p. 9, 2019.

[23] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," arXiv preprint arXiv:2110.04544, 2021.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[25] P. Singhal, T. Goyal, J. Xu, and G. Durrett, "A long way to go: Investigating length correlations in rlhf," ArXiv, vol. abs/2310.03716, 2023.

[26] E. Choi, H. He, M. Iyyer, **M. Yatskar**, W. tau Yih, Y. Choi, P. Liang, and L. Zettlemoyer, "Quac: Question answering in context," in Proceedings of Empirical Methods in Natural Language Processing (EMNLP), 2018.

[27] M. Yatskar, "A qualitative comparison of coqa, squad 2.0 and quac," arXiv preprint arXiv:1809.10735, 2018.

[28] P. Lee, S. Bubeck, and J. Petro, "Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine," New England Journal of Medicine, vol. 388, no. 13, pp. 1233–1239, 2023.

[29] B. Owens, "How nature readers are using chatgpt," Nature, 2023.