Discursive Datasets: Situated Social Knowledge for Multimodal Modeling

The most revolutionary breakthroughs in science come not from gradually working out the answers to questions, but from figuring out new types of questions to ask, argues Thomas Kuhn in *The Structure of Scientific Revolution*.[1] Kuhn's insight is deeply felt in machine learning, a field known for its recent "revolutionary" progress in many perceptual and cognitive tasks. Fei-Fei Li et al.'s 2009 ImageNet database[2] was revolutionary because it moved from asking "How can we identity and *build* the fundamental structures of vision into algorithms?" to "How can algorithms *learn* the fundamental structures of vision from data?" Recent work in multimodal modeling[3] is also on the cusp of such a "revolution" because it moves from asking "How can we get humans to *give us* the data we want (to train models)?" to "How can we harness the vast amount of socially produced data *already available* (to train models)?" Here, data is "socially produced" if it is produced by agents (humans) in response to or in context of other agents' information, such as in comment threads or hashtags. Indeed, rapid recent advances in image understanding and generation owe much of their success to large-scale datasets[4] constructed by extracting pairs of images and associated text scraped from all across the web in social contexts like social media and forums.[5]

The attention of researchers in multimodal modeling has been directed largely towards *harnessing* these giant sources of data. But I think there are compelling

[1] Thomas S. Kuhn, *The Structure of Scientific Revolutions* (Chicago: University of Chicago Press, 1962), 264.

[2] Jia Deng et al., "ImageNet: A Large-Scale Hierarchical Image Database," 2009 conference on Computer Vision and Pattern Recognition.

[3] *Multimodal modeling*: The modeling of multiple modalities simultaneously, such as images and text. Examples include image captioning and text-to-image generation (e.g. DALL-E, Stable Diffusion, GPT-4V).

[4] See: Karan Desai et al., "RedCaps: Web-Curated Image-Text Data Created by the People, for the People," eprint arXiv:2111.11431 (2021); Christoph Schuhmann et al., "LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models," eprint arXiv:2210.08402 (2022); Wanrong Zhu et al., "Multimodal C4: An Open, Billion-scale Corpus of Images Interleaved with Text," eprint arXiv:2304.06939 (2023).

[5] See: Robin Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models," eprint arXiv:2112.10752 (2022); Alec Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," eprint arXiv:2103.00020 (2021).

concerns directing us towards the second half of the question: understanding the properties of *socially produced* data (which we are quick to harness for models).

To articulate this concern, I enlist the philosopher Donna Haraway, who argues in her seminal “Situated Knowledges” that many of our ways of understanding the world express “*unlocatable*, and so *irresponsible*, knowledge claims.”[6] Haraway provides an instructive example: when we gaze across those dazzling photos of colorful planets and stars leaping across the cosmos, we may conclude that we are but a small and ultimately insignificant part of its grand beauty. Certainly, a beautiful sentiment – but one which is *unlocatable*, because it neglects the *human* choice of frequencies to study, the procedure for colorizing waves which have no visible color, etc. And therefore, it is *irresponsible*, because one may claim to have seen the universe in its “objective” beauty only because they forgot the conditions for their sight. My concern is that those large-scale image-text datasets are like giant collections of planet photos: each image-text pair extracted from a rich context of social discourse but ultimately cut off from it, just like those pictures of planets where the telescopes and rendering that produced them are hidden. When models are trained on these isolated pairs, they certainly do learn nuanced concepts, but in a messy and unclear way — because they don’t have access to the social context from which those pairs are produced (see Fig. 1). Hence, these models exhibit “irresponsible” behavior, which actualize as technical problems for computer vision researchers, such as “hallucination” or “misalignment” with human intentions and behaviors[7] — just as our claim to see the universe “objectively” was a “hallucination”, an irresponsible claim.

[6] Donna Haraway, "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective," Feminist Studies 14, no. 3 (Autumn 1988): 575-599, https://doi.org/10.2307/3178066, 583.

[7] For a few instructive examples, see: Sandro Pezzelle, "Dealing with Semantic Underspecification in Multimodal NLP," eprint arXiv:2306.05240 (2023); Yifan Li et al., "Evaluating Object Hallucination in Large Vision-Language Models," eprint arXiv:2305.10355 (2023); Zhiqing Sun et al., "Aligning Large Multimodal Models with Factually Augmented RLHF," eprint arXiv:2309.14525 (2023).

In my view, these concerns motivate more attention towards the second half of that revolutionary question — “How can we harness the vast amount of *socially produced data* already available?” The project I am proposing in this direction has three components. Firstly, we need a way of representing the social context which allows us to *locate* knowledge in the conditions it was produced from.[8] Secondly, we need to create a dataset implementing such a representation — to my knowledge, the first of its kind. Thirdly, we need to design, train, and evaluate models which can learn from this sort of representation. The contribution to the field of multimodal modeling is both *practical*, in providing a novel representation, dataset, and model; and *theoretical*, in pointing deep learning towards greater engagement with concerns from social philosophy.

This is, bluntly, a huge project. But I have many reasons to believe I can do it.

Firstly, my mentors — professors Amy X. Zhang and Ranjay Krishna — have deeply relevant expertise. Amy’s work on social interaction in digital spaces will guide the design of social context representations,[9] and Ranjay is intimately familiar with bringing theoretical representations to fruition — one of his important contributions to computer vision being the operationalization of scene graphs to represent structured logical entities in scenes.[10] I have worked wonderfully with both for over two years now.

Secondly, I have a history of successful independently-led research projects,[11] all of which have investigated similar themes on the relations between models and human cognitive, perceptual, and social structures. I intend to lead this project similarly:

[8] This representation would be a specialized graphs with distinct edge types and specialized traversals.

[9] See: Amy X. Zhang et al., "Characterizing Online Discussion Using Coarse Discourse Sequences," Proceedings of the Conference on Web and Social Media (2017), Amy X. Zhang et al., "Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization," in Proceedings of the 2017 ACM Conference on CSCW, ACM, 2017, pp. 2082–2096.

[10] See: Ranjay Krishna et al., "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," eprint arXiv:1602.07332 (2016); and earlier works.

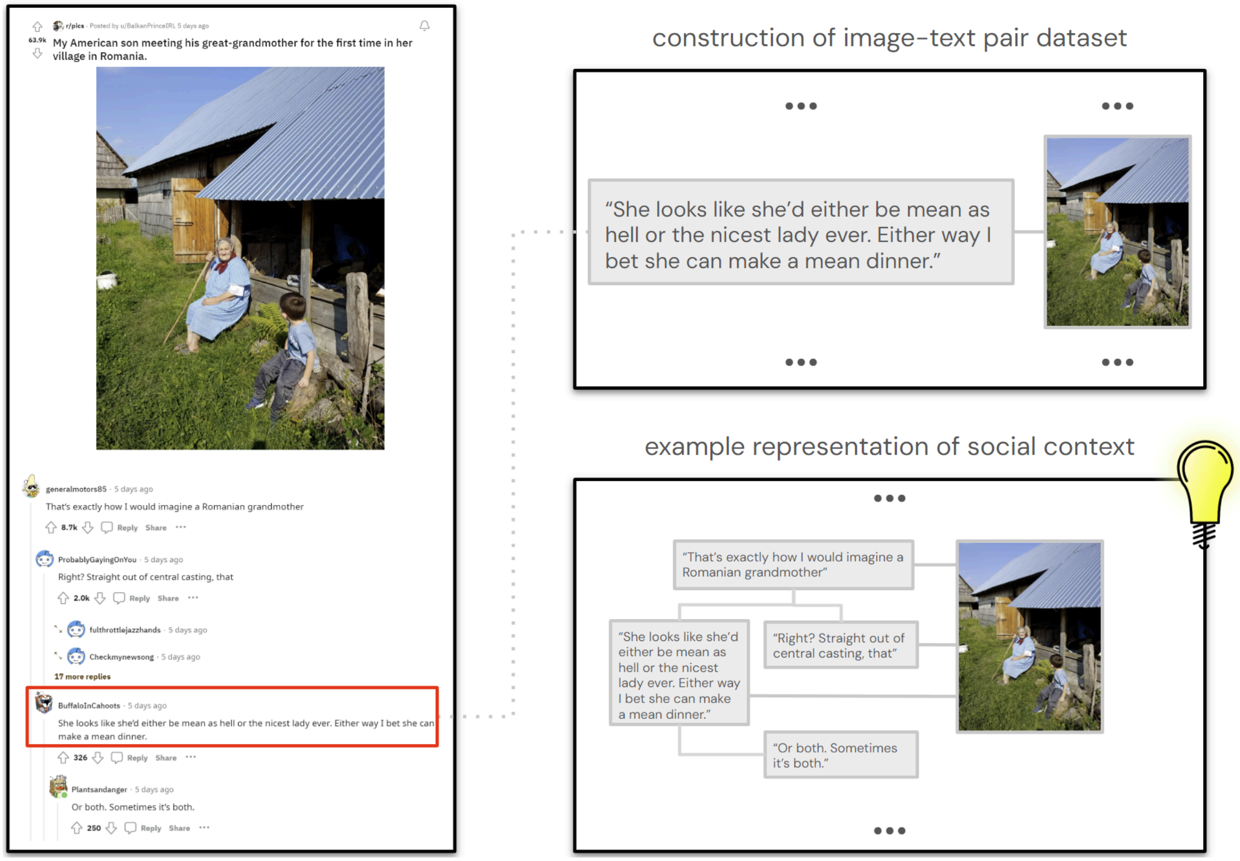
[11] For examples, see: Andre Ye, Quan Ze Chen, and Amy Zhang, "Confidence Contours: Uncertainty-Aware Annotation for Medical Semantic Segmentation," eprint arXiv:2308.07528 (2023) (best paper honorable mention at HCOMP 2023); Andre Ye, Sebastin Santy, Jena D. Hwang, Amy X. Zhang, and Ranjay Krishna, "Cultural and Linguistic Diversity Improves Visual Representations," eprint arXiv:2310.14356 (2023).

explicating the research direction, running experiments, re-explicating and re-running, writing the paper — soliciting help as appropriate. Ranjay and Amy have introduced me to a wonderful community of grad students and faculty in the computer science department who provide me with technical and theoretical resources as I need them.

Thirdly, this project is of great personal interest to me. As a computer science and philosophy double major, I see many rich connections between the two disciplines which I believe are not being sufficiently investigated in either area. This project was motivated by many, many hours of reading philosophy in the classroom and discussing ideas with philosophy faculty. I am excited by the chance to contribute in a small but concrete way towards a bridge between the computer science and philosophy. After graduation, I intend to pursue a doctoral degree with a research program directed broadly at formulating and addressing philosophical concerns in machine learning. I've already done some theoretical work in this space,[12] but want to pursue a larger-scale and more concrete project. My hope is that this project will serve as a "stepping-stone" environment to feel out what this kind of underexplored research would look like.

The Mary Gates scholarship is instrumental towards actualizing this environment. I currently work two teaching assistant positions to cover living costs. While I enjoy TAing, it is very time consuming and I would like to dedicate more time towards this research project. I've come to see that great exploratory research takes an extreme willingness to pursue pathways of inquiry which may not bear fruit. The Mary Gates scholarship will give me the room to do just this kind of genuine exploration.

[12] See Mark Pock*, Andre Ye*, and Jared Moore, "LLMs Grasp Morality in Concept," eprint arXiv:2311.02294 (2023). *equal contribution.



r/pics · Posted by u/BalkanPrinceIRL 5 days ago

63.9k

My American son meeting his great-grandmother for the first time in her village in Romania.

generalmotors85 · 5 days ago

That's exactly how I would imagine a Romanian grandmother

8.7k Reply Share ···

ProbablyGayingOnYou · 5 days ago

Right? Straight out of central casting, that

2.0k Reply Share ···

fulthrottlejazzhands · 5 days ago

Checkmynewsong · 5 days ago

17 more replies

BuffaloInCahoots · 5 days ago

She looks like she'd either be mean as hell or the nicest lady ever. Either way I bet she can make a mean dinner.

326 Reply Share ···

Plantsandanger · 5 days ago

Or both. Sometimes it's both.

250 Reply Share ···

"She looks like she'd either be mean as hell or the nicest lady ever. Either way I bet she can make a mean dinner."

"That's exactly how I would imagine a Romanian grandmother"

"She looks like she'd either be mean as hell or the nicest lady ever. Either way I bet she can make a mean dinner."

"Right? Straight out of central casting, that"

"Or both. Sometimes it's both."

Figure 1. Depiction of how an image-pair dataset sample (right, top) and representation which captures social context (right, bottom) might be constructed from a Reddit post on r/pics (left). Observe that all information in the example representation is situated in relation to other information produced in relation to it in.