All our knowledge begins with the senses, proceeds then to the understanding, and ends with reason. There is nothing higher than reason.

– Immanuel Kant

Assimilating perceptual knowledge into understanding, and applying it for reasoning remains an enduring enigma of the human mind, from the era of enlightenment to the current age of artificial intelligence (AI). Intriguingly, despite using internet-scale data, large-scale human feedback, and hundreds of billions of model parameters – reasoning remains an elusive frontier for AI. My goal is to craft the foundational gears, and levers of artificial intelligence that can reason about the physical world as well as the realm of ideas in a seamless manner.

Concretely, I will study the 3 Rs of anthropic multimodal artificial intelligence that is designed to share our cognitive underpinnings, emulate humanlike reasoning, and be behaviorally trustworthy:

- Representation: How to represent multimodal input signals to derive latent "understanding"?
- Reasoning: What computational mechanisms underlie abstract machine reasoning?
- Robustness: How can we ensure that AI systems are broadly deployable?

"Understanding" from Representation. Abstraction enables humans to distill the seemingly infinite complexity of the world into meaningful, transferable chunks. Consider an image of the Brooklyn bridge taken during the day. Depending upon the context, the underlying premise might be the generic concept "day" (*e.g.,* if we want to decide whether to use sunglasses) or the specific concept "photo of the brooklyn bridge" (*e.g.,* if we want to decide which train to take next). Representing input instances at the right level of abstraction can help transfer solutions from one instance to the other, and help tame the exploding complexity of the world by sharing statistical strength.

In MERU [1], I learn vision-language (VL) embeddings that capture long-range visual-semantic dependencies. The idea is to structure the embeddings of dual encoder VL representations like CLIP [2] using hyperbolic geometry, which is particularly well suited to model hierarchies due to an exponential growth of space that mirrors the branching of trees. Intriguingly, with this inductive bias one can learn visually-grounded hierarchies simply by observing correlations of visual patterns, and paired sentences – without access to annotations of hierarchy (which is both intractable, and expensive to collect). At inference, MERU enables us to "read-off" long-range semantic dependencies (*e.g.,* "day") better than CLIP, and is more performant – especially for small embedding dimensions – suggesting it more organically captures the world's implicit visual-semantic structure.

My work has kindled interest in the vision, and language community on the topic of hierarchies, with follow up works proposing benchmarks [3], and scaling MERU-like models up to billions of parameters [4]. MERU won a best paper award at the ICLR workshops in 2023, was highlighted as a featured publication by Meta AI at ICML 2023[1], and has received 44 citations over the past year.

Taking a step back, can machines can intuit an abstraction of the intricate interactions between people, and objects in visual scenes – essentially representing a "typical" point in the hierarchy? In [5], I demonstrated that humans often converge on intuitive, typical sentences when describing images, and developed a novel metric, CIDEr for evaluating typicality of novel sentences. First introduced at CVPR 2014, CIDEr has significantly shaped the field of image captioning by providing inexpensive, and reliable feedback on modeling iterations (and has since been cited over 5,200 times).

[1] https://ai.meta.com/events/icml-2023/

Contextual, and Coherent Reasoning. Consider a second order reasoning problem, where one has to not only describe a target image with a sentence, but ensure that it uniquely refers to it, in context of a closely related distractor image. What inferential mechanisms can help achieve this task? In [6], I developed a novel decoding algorithm for conditional language models (LMs), called emitter-suppressor (ES) beam search, which implicitly reasons about how a listener might identify the target image, based on the generated sentence while decoding it token by token – without access to any context-aware ground truth sentences at training time. This work was published as a spotlight at CVPR, 2018 (top 8% of submissions), and has been cited 168 times till date.

In follow up work Prob-NMN [7], I proposed a probabilistic scaffolding for neural-symbolic models on reasoning-oriented visual question answering [8] tasks, and showed how to probe the coherence of their reasoning (beyond simply predicting the right answer). Interestingly, inferential coherence [9] remains a significant challenge for reasoning in state-of-the-art large language models (LLMs) [10]. Prob-NMN, presented as a long oral at ICML 2021 (top 4.2%), has garnered 104 citations to date.

Mechanistic, and Certifiable Robustness. AI systems can often be fragile, and fail spectacularly when deployed in the wild, leading to unforeseen consequences, and a loss of public trust[2]. Thus, it is of crucial importance to build a mechanistic understanding of generalization in AI.

In this vein, I address feature learning for supervised learning [11], by theoretically characterizing the optimal features for a chosen model class (*e.g.,* two layer MLP) that will enforce generalization (*i.e.,* low gap between train, and test accuracy); and provide practical algorithms to learn features that generalize even in adversarial settings. This fundamental work was published at NeurIPS, 2020 as a Spotlight (top 4% of all submissions), and has garnered 41 citations to date.

For end user trust, is crucial to certify when AI models are expected to work. In [12] (Gen-OOD), I measured how well a model trained on a domain (sunny weather) will work on a novel domain (snow). I contributed a benchmark to evaluate progress on this problem, comprising generalization statistics of 12,000 models, an evaluation of standard theory [13], and a number of baseline measures. First published at NeurIPS 2021, Gen-OOD has been cited 42 times. In [14], I applied these ideas to robust visual question answering (VQA) [15] with an option to abstain, enabling safety critical deployment (*e.g.,* to users who are visually impaired). This work was published at CVPR 2023.

Future Work

Grounding. State-of-the-art VLMs often fail on tasks that require grounding [16, 17], exhibiting a textual bias – where they primarily rely on the textual signal for giving answers, even in a multimodal context. To address this, we must understand the extent to which visual, and textual token patterns are synergestic (as opposed to antagonistic) in VLMs. Moreover, on a broader level we need a first principles approach to address textual bias in VL models [18–23]. To this end, I will extend my previous foundational work on optimal supervised learning [11] to the multimodal setting. I will also investigate connections between the unimodal bias and the simplicity bias [24].

Compositionality. It is prudent to explore how the advances in computer vision, and machine learning over the last two decades can help improve reasoning in large scale models [25]. A particularly promising direction that I will explore is integrating more object centric [26] primitives into visual tokenization schemes (building on my previous work [27] on object-centric features). Further,

[2] https://hbr.org/2022/03/why-ai-failed-to-live-up-to-its-potential-during-the-pandemic

I will also study inference in agentic VLMs (*e.g.,* chain-of-thought [28]) through carefully designed prompting schemes, and instruction tuning to integrate computer vision artifacts (*e.g.,* objects, poses) into the prompts. More broadly, I will disentangle whether it is our current representational or inferential choices that bottleneck compositional reasoning, by crafting representations especiallity tuned for reasoning tasks through meta learning [29] under constraints on inference.

Abstraction. Representing a perceptual input at the right level of abstraction ("dog" *vs* "dalmatian") has potential to massively improve the reasoning abilities of Multimodal AI [17]. As a first step, I will explore the use of hyperbolic embeddings like MERU for abstract reasoning tasks such as Raven's matrices [30], and the more complex CURI [12] task that I proposed at ICML 2021. Next, I will analyze how to integrate context cheaply into VLM tokenization without incurring an expensive forward pass through language conditioning of the entire visual backbone. In this direction, I am curious about advancing in hierarchical tokenization schemes for vision, and combining them with inference techniques such as self-notes [31], and scratchpads [32]. Overall, integrating abstraction into multimodal AI remains an engaging long term research direction in machine reasoning.

Robustness. I will study robustness of grounded reasoning systems with respect to input distribution shifts (building on my initial work in robust VQA [14]), and performance on sets of related reasoning problems, by evaluating for inferential coherence [9]. An exciting direction is to represent the same reasoning problem across multiple input modalities (extending my initial efforts in [33]). This will help ensure that grounding, and robust reasoning progress hand in hand.

Foundations of Reasoning. While deep networks today minimize the average empirical risk (ERM) [34] (that scales well to large datasets), we are often interested in reasoning to help us with truly novel, unforseen situations (that are in the tail). I will study this fundamental tension, by analyzing the interplay between learning curves of models (including phenomena like grokking [35,36]), model classes (*e.g.,* neural symbolic models [37]), and emergent properties (*e.g.,* inferential coherence) to understand the models (and hyperparameters) can that can unlock reasoning by simply minimizing the empirical risk. Simultaneously, I will pursue novel paradigms in ML for reasoning (*e.g.,* using meta learning [29], sample reweighting [38], learning with side information [39]) to address alternatives to ERMs, that can deal more robustly with problems of reasoning in the tail.

Datasets. I will use the data as the clue to understand deep learning, by learning synthetic datasets with desired properties. Firstly, I will create high-quality multimodal datasets that enable rapid learning, by extending my recent work in scalable dataset distillation [40] to multimodal settings. Also, I will address learnability, and establish architecture design on a firmer footing by understanding which functions one is now able to learn, by say adding a head to a transformer, enabling more systematic modeling choices for practitioners. Finally, I will work backwards from learning curves, and craft synthetic reasoning datasets (by leveraging GenAI models) that demonstrate certain patterns in learning curves (*e.g.,* grokking) to gain insights into the underlying phenomena.

Conclusion

Crafting the building blocks of anthropic multimodal artificial intelligence – that closes the loop between perception and reasoning, while being trustworthy, and well understood mechanistically – has potential to drive far-reaching developments in a number of application areas such as personal assistants, medicine, education, scientific discovery, and eventually, benefit society as a whole.

References

[1] K. Desai, M. Nickel, T. Rajpurohit, J. Johnson, and R. Vedantam, "Hyperbolic image-text representations," *arXiv [cs.CV]*, Apr. 2023. 1

[2] A. Radford, I. Sutskever, J. W. Kim, G. Krueger, and S. Agarwal, "CLIP: Connecting text and images." https://openai.com/blog/clip/, Jan. 2021. Accessed: 2021-12-9. 1

[3] M. Alper and H. Averbuch-Elor, "Emergent visual-semantic hierarchies in image-text representations," *arXiv [cs.CV]*, July 2024. 1

[4] P. Mandica, L. Franco, K. Kallidromitis, S. Petryk, and F. Galasso, "Hyperbolic learning with multimodal large language models," *arXiv [cs.LG]*, Aug. 2024. 1

[5] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2014. 1

[6] R. Vedantam, S. Bengio, K. Murphy, D. Parikh, and G. Chechik, "Context-aware captions from context-agnostic supervision," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1070–1079, 2017. 2

[7] R. Vedantam, K. Desai, S. Lee, M. Rohrbach, D. Batra, and D. Parikh, "Probabilistic neural-symbolic models for interpretable visual question answering," *arXiv [cs.LG]*, Feb. 2019. 2

[8] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[9] M. Nye, M. H. Tessler, J. B. Tenenbaum, and B. M. Lake, "Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning," *arXiv [cs.AI]*, July 2021. 2, 3

[10] S. Mishra, A. Mitra, N. Varshney, B. Sachdeva, P. Clark, C. Baral, and A. Kalyan, "NumGLUE: A suite of fundamental yet challenging mathematical reasoning tasks," *arXiv [cs.CL]*, Apr. 2022. 2

[11] Y. Dubois, D. Kiela, D. J. Schwab, and R. Vedantam, "Learning optimal representations with the decodable information bottleneck," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, (Red Hook, NY, USA), Curran Associates Inc., 2020. 2

[12] R. Vedantam, A. Szlam, M. Nickel, A. Morcos, and B. Lake, "CURI: A benchmark for productive concept learning under uncertainty," in *International Conference on Machine Learning*, 2021. 2, 3

[13] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, pp. 151–175, May 2010. 2

[14] C. Dancette, S. Whitehead, R. Maheshwary, R. Vedantam, S. Scherer, X. Chen, M. Cord, and M. Rohrbach, "Improving selective visual question answering by learning from your peers," *arXiv [cs.CV]*, June 2023. 2, 3

[15] S. Whitehead, S. Petryk, V. Shakib, J. Gonzalez, T. Darrell, A. Rohrbach, and M. Rohrbach, "Reliable visual question answering: Abstain rather than answer incorrectly," *arXiv [cs.CV]*, Apr. 2022. 2

[16] F. B. Baldassini, M. Shukor, M. Cord, L. Soulier, and B. Piwowarski, "What makes multimodal in-context learning work?," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1539–1550, IEEE, June 2024. 2

[17] Y. Zhang, H. Bai, R. Zhang, J. Gu, S. Zhai, J. Susskind, and N. Jaitly, "How far are we from

intelligent visual deductive reasoning?," *arXiv [cs.AI]*, Mar. 2024. 2, 3

[18] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," *arXiv [cs.CV]*, Dec. 2016. 2

[19] R. Cadene, C. Dancette, H. Ben-younes, M. Cord, and D. Parikh, "RUBi: Reducing unimodal biases in visual question answering," *arXiv [cs.CV]*, June 2019. 2

[20] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal networks hard?," *arXiv [cs.CV]*, May 2019. 2

[21] I. Gat, I. Schwartz, and A. Schwing, "Perceptual score: What data modalities does your model perceive?," *arXiv [cs.LG]*, Oct. 2021. 2

[22] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," *arXiv [cs.CV]*, Mar. 2022. 2

[23] Y. Zhang, P. E. Latham, and A. Saxe, "Understanding unimodal bias in multimodal deep linear networks," *arXiv [cs.LG]*, Dec. 2023. 2

[24] H. Shah, K. Tamuly, A. Raghunathan, P. Jain, and P. Netrapalli, "The pitfalls of simplicity bias in neural networks," *arXiv [cs.LG]*, June 2020. 2

[25] Z. Ma, J. Hong, M. O. Gul, M. Gandhi, I. Gao, and R. Krishna, "CREPE: Can vision-language foundation models reason compositionally?," *arXiv [cs.CL]*, Dec. 2022. 2

[26] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, "Object-centric learning with slot attention," *arXiv [cs.LG]*, June 2020. 2

[27] S. Xie, A. S. Morcos, S. Zhu, and R. Vedantam, "COAT: measuring object compositionality in emergent representations," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 24388–24413, PMLR, 2022. 2

[28] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," *arXiv [cs.CL]*, Jan. 2022. 3

[29] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning (ICML)*, 2017. 3

[30] D. G. T. Barrett, F. Hill, A. Santoro, A. S. Morcos, and T. Lillicrap, "Measuring abstract reasoning in neural networks," in *International Conference on Machine Learning (ICML)*, 2018. 3

[31] J. Lanchantin, S. Toshniwal, J. Weston, A. Szlam, and S. Sukhbaatar, "Learning to reason and memorize with self-notes," *arXiv [cs.LG]*, May 2023. 3

[32] M. Nye, A. J. Andreassen, G. Gur-Ari, H. Michalewski, J. Austin, D. Bieber, D. Dohan, A. Lewkowycz, M. Bosma, D. Luan, C. Sutton, and A. Odena, "Show your work: Scratchpads for intermediate computation with language models," *arXiv [cs.LG]*, Nov. 2021. 3

[33] S. R. Vedantam, D. Lopez-Paz, and D. J. Schwab, "An empirical investigation of domain generalization with empirical risk minimizers," in *Advances in Neural Information Processing Systems*, May 2021. 3

[34] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 10, pp. 988–999, Sept. 1999. 3

[35] Z. Liu, O. Kitouni, N. Nolte, E. J. Michaud, M. Tegmark, and M. Williams, "Towards under-

standing grokking: An effective theory of representation learning," *arXiv [cs.LG]*, May 2022. 3

[36] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra, "Grokking: Generalization beyond overfitting on small algorithmic datasets," *arXiv [cs.LG]*, Jan. 2022. 3

[37] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," *arXiv [cs.CV]*, Nov. 2015. 3

[38] H. Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *J. Stat. Plan. Inference*, vol. 90, pp. 227–244, Oct. 2000. 3

[39] R. Jonschkowski, S. HÃ¶fer, and O. Brock, "Patterns for learning with side information," *arXiv [cs.LG]*, Nov. 2015. 3

[40] Y. Feng, S. R. Vedantam, and J. Kempe, "Embarrassingly simple dataset distillation," *Int Conf Learn Represent*, Oct. 2024. 3