RANJAY KRISHNA RESEARCH STATEMENT

As modern machines struggle to fully conceptualize the visual world, my research looks toward human learning for potential solutions. Inspired by Computer Vision's original basis in human perception, my research methodology identifies learning mechanics from cognitive and social sciences to bootstrap machine learning. From this process I developed two lines of work: (1) representations and models inspired by Cognitive Science and (2) optimizations and training paradigms spurred by Human-Computer Interaction.

Drawing on Cognitive Science, specifically from human visual processing and visual memory, I designed representations and models that have led to improvements on a variety of core Computer Vision tasks. Specifically, I introduced scene graphs, a dense, detailed, computational representation of visual information, via the Visual Genome dataset [30, 20, 37, 5, 3, 25, 17, 32]. Since our introduction of scene graphs, the Computer Vision community has developed hundreds of scene graph models and utilized scene graphs to achieve state-of-the-art results across multiple core tasks, including object localization, image captioning, generation, question answering, 3D understanding, and spatio-temporal action recognition. Visual Genome is also now the de-facto dataset for pre-training object detectors for any downstream task. The three scene graph workshops I organized at international conferences have each attracted between 400-600 attendees [26, 17, 22].

Drawing on Human-Computer Interaction, together with Gricean Maxims from Linguistics and Social Development Theory from Social Psychology, I have created training paradigms that go beyond traditional Computer Vision tasks and enabled learning from human interactions. Concretely, I created a real-world deployment of an agent, which learned new visual information over a course of 8 months by interacting with over $230K$ people on social media by asking questions [21, 35, 29]. This work demonstrates the possibility of agents that can uncover social norms to learn how to interact, and improve their capabilities over time by engaging people in existing social environments. Building towards this deployment, I have also introduced tasks and models that can express visual knowledge using language and, more importantly, ask questions to learn new visual information [28, 24, 23]. I have organized 3 competitions at international conferences where teams from over 30 institutions and 8 countries have doubled the state-of-the-art in these tasks in just 3 years [7, 8].

PAST RESEARCH

I will summarize these two broad areas of my past research, highlighting key ideas and contributions.

Teaching machines to see with scene graphs. When I started my Ph.D., Computer Vision models were primarily focused on object classification and detection, leaving all other visual information unattended. For example, these methods could not differentiate between images of a person *riding*, *pushing*, or even *falling off* of the bicycle. To realize vision applications which people can use, it is vital that our computational visual representations reflect human cognition. Work in Cognitive Science and human visual memory [2, 39] have identified that when people process visual stimuli, they identify objects and simultaneously encode the *relationships* between them. In an effort to represent visual scenes similar to human cognition, we developed a new representation — scene graph [30, 20], which encodes objects in the scene as nodes and relationships as edges. To promote research on scene graphs, we collected the Visual Genome repository [30], a large-scale dataset containing scene graphs with millions of objects, attributes, and relationships. To collect Visual Genome, we developed new data collection interfaces inspired by Psychophysics to speed up crowdsourcing data by an order of magnitude [27].

With Visual Genome, we introduced the task of scene graph generation, which expects an image as input and detects objects in the image with relationships connecting them [32]. Along with the task, we also developed a new model since existing predictions models tended to overfit to this difficult structured prediction task. Inspired by Cognitive Science findings that show humans processing relationships in parallel to identifying objects [2], we designed a model architecture with two parallel branches: one for object detection and another for relationships classification [32]. This separation between between objects and relationships affords the prediction of novel scene graphs unseen during training. For instance, our model could identify "people - sitting on - fire hydrants", a composition it had never seen before by individually recognizing the objects, "person", "fire hydrant", and the relationship, "sitting on".

Our subsequent work [25, 5] on scene graph generation improves upon this original model by adding a graph convolution layer, which iteratively fine-tunes each prediction by considering all the other objects and relationships in the image. This model was inspired by work in Neuroscience which postulates that visual processing is an iterative process [9]. Over the last 4 years, the computer vision community has published hundreds of scene graph models using a variety of recent deep learning advances. Yet, one design decision has persisted: the parallel object and relationship branches introduced by our original model.

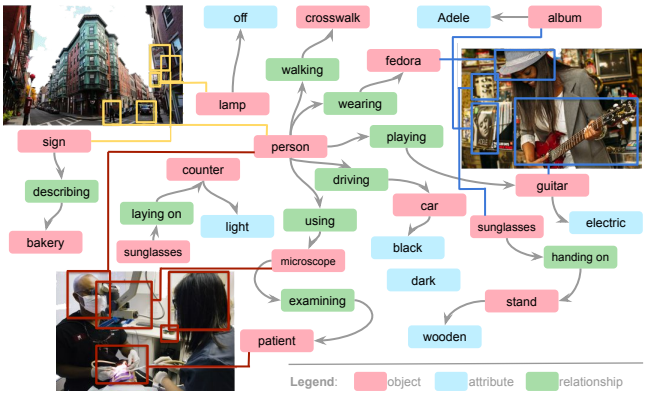

Figure 1: Scene graphs in Visual Genome.

Scene graphs provide a comprehensive intermediate visual representation that other downstream tasks can use to reason over visual data. Using scene graphs as this intermediate representation, we improved numerous core vision tasks, including image retrieval [20, 37], object localization [25], action recognition [17] and learning with few training examples [3, 5]. Likewise, the community has utilized scene graphs to create new diagnostic datasets [18, 16, 41], supervised learning techniques [15, 33, 19, 14], interactive learning methods [11], and evaluation protocols [1].

Teaching machines to interact with people using language. Social development theory has demonstrated that human development is a socially mediated process in which children acquire new beliefs and cultural norms through inquisitive dialogues with more knowledgeable members of society. Yet, the ability to learn from other people remains an open challenge limiting artificial intelligence agents. The space of possible interactions is combinatorially vast and feedback from people is often implicit, making learning from interaction intractable outside of simple domains or simulation.

We introduced a framework that learns from interactions with people by simultaneously learning how to interact with people [29]. To navigate the combinatorial space of interactions, we drew an insight from numerous social science fields: while the action space of all possible behaviors is vast, and intractable to explore within a reasonable number of interactions, most human-human interactions lie on a low-dimensional manifold [36, 6]. We developed an algorithm that discretizes the space of all possible interactions into a lower dimensional manifold, and uses it as a surrogate tractable action space for reinforcement learning. We enable learning from human interactions by

designing two rewards: one to uncover interactions that elicit people's interests and another to generate interactions that result in new visual information.

With this framework, we deployed an agent that learns to recognize new visual concepts by interacting with people using natural language questions. Through an 8-month deployment on a photo-sharing social network, our system interacted with $230K$ people, learned through implicit feedback to ask questions that increased response rates, and improved recognition of new visual concepts. Unlike prior work, we did not train participants to provide explicit rewards. Instead, our system explored the space of possible interactions to infer the social norms already established within the social network [34]. It learned questions that can be easily interpreted and answered, avoided ambiguous utterances, and demonstrated social proof by mentioning recognizable concepts. We further improved the system's ability by teaching it a number of request strategies [35] (e.g. using compliments or justifications) identified by Social Psychology [4]. This work expands possibilities for socially-capable AI to continuously improve and adapt in whatever domain it is launched in.

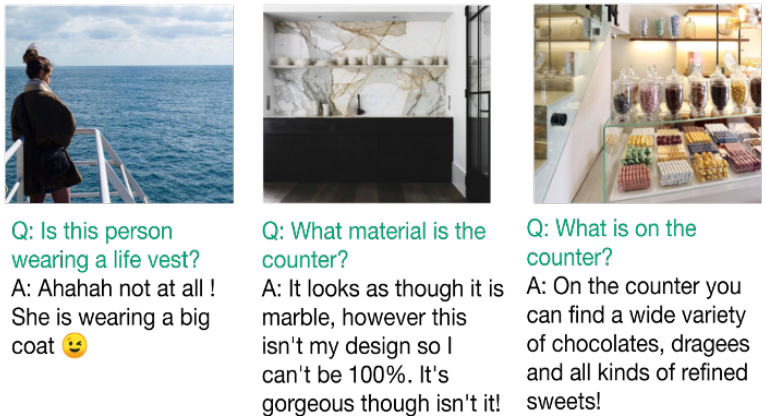

Figure 2: Example interactions with real people on social media.

Building up to this deployment, my team and I have been developing vision models that can communicate using language [23, 28, 24]. In two such projects, use the maxim of quantity from Gricean Maxims, to generate dense captioning models. In one, we introduced a hierarchical recurrent network to generate paragraph descriptions of images [23]. In another, we introduced a hierarchical temporal detection model to detect and caption events in videos [28]. Next, using the maxim of relation and quality, we developed an information maximizing question generation model that can generate different questions depending on whether it wants to learn about objects ("What is the person throwing?"), attributes ("What color is the frisbee?"), or relationships ("Did the person on the left throw or catch the frisbee?") [24]. Since the introduction of these tasks, the community has more than doubled the state-of-the-art on paragraph generation and dense video captioning [10, 40].

FUTURE RESEARCH AGENDA

I envision a future where artificial intelligence agents are akin to adaptable co-workers taught "on-the-job" to assist anyone in any task. These machines are capable of learning from people, and empower anybody, including the vast majority without computing experience, to tailor AI applications in domains like knowledge acquisition, healthcare services, education, sustainability, automation, and countless others. To realize this goal, I will continue to draw ideas from human learning to pursue the following concrete future directions.

Compositional spatio-temporal representations. As video understanding applications have matured, we need representations that characterize visual events in the compositional manner that people understand them. Recently, inspired from Event Segmentation Theory [31], we introduced a representation that decomposes actions into spatio-temporal scene graphs — representing actions

has changes in relationships between objects [17]. We utilized this new representation to train a model that improves action recognition while also being more interpretable [17]. My current work is utilizing this representation to develop a new benchmark to test whether models exhibit compositional spatio-temporal generalization [12].

Improving collaboration by inferring human intentions. While people are uniquely adept at inferring others' intentions, delegating tasks, and coordinating their behavior on the fly, collaboration remains an open challenge for artificial intelligence. I have recently embarked upon two projects related to improving human-AI collaboration. In one, we are drawing on the Theory of Mind to develop a self-supervised multi-agent exploration approach that leads to emergent intention inference. In other words, machines learn to infer human intentions by practicing in simulation with a copy of itself. A new ongoing project is exploring how different AI explainability methods affect our mental models and in turn, impact human-AI collaboration.

A science for human-AI interaction. Human-Computer Interaction has formalized the process of observing how machines (mis-)behave when deployed in real-world situations and proposing design decisions to improve them. Yet, incorporating those decisions as inductive biases or data generation strategies or evaluation protocols is largely an ad hoc process within the AI community. I intend to standardize and validate processes to directly convert human observations and interactions into machine learning changes. I have taken some steps towards this direction [13, 27, 38, 42]. In one such project, we discovered that generative models were being evaluated using heuristics because of the difficulty in attaining consistent human judgements. In response, we developed a human evaluation protocol for generative adversarial models that is consistent and grounded in Psychophysics [42].

Designing pro-social interactions. Many non-technical factors affect how people perceive and interact with machines. I intend on exploring various avenues of impression formation to study how factors that induce pro-social and avoid anti-social human-AI interactions. My initial explorations have drawn on the Stereotype Content Model from Psychology to suggest how machines should be introduced, resulting in a theory that provides a new lens on why consumers have abandoned or refused to interact with many commercial AI agents [21].

In sum, my research draws on frameworks originating in cognitive and social sciences to improve machine learning. It aims to realize a world in which computing systems learn from interactions with people. By propelling the evolution of real-world systems, my research is taking steps to empower people to take charge of their machine's learning.

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. "Spice: Semantic propositional image caption evaluation". In: *European Conference on Computer Vision*. Springer. 2016, pp. 382–398.

[2] Irving Biederman, Robert J Mezzanotte, and Jan C Rabinowitz. "Scene perception: Detecting and judging objects undergoing relational violations". In: *Cognitive psychology* 14.2 (1982), pp. 143–177.

[3] Vincent Chen, Paroma Varma, **Ranjay Krishna**, Michael Bernstein, Christopher Re, and Li Fei-Fei. "Scene Graph Prediction with Limited Labels". In: *International Conference on Computer Vision*. 2019.

[4] Robert B Cialdini and Robert B Cialdini. *Influence: The psychology of persuasion*. Collins New York, 2007.

[5] Apoorva Dornadula, Austin Narcomey, **Ranjay Krishna**, Michael Bernstein, and Li Fei-Fei. "Visual Relationships as Functions: Enabling Few-Shot Scene Graph Prediction". In: *ArXiv*. 2019.

[6] Ernst Fehr and Urs Fischbacher. "Social norms and human cooperation". In: *Trends in cognitive sciences* 8.4 (2004), pp. 185–190.

[7] Bernard Ghanem, Juan Carlos Niebles, Cees Snoek, Fabian Caba Heilbron, Humam Alwassel, Victor Escorcia, Ranjay Khrisna, Shyamal Buch, and Cuong Duc Dao. "The activitynet large-scale activity recognition challenge 2018 summary". In: *arXiv preprint arXiv:1808.03766* (2018).

[8] Bernard Ghanem, Juan Carlos Niebles, Cees Snoek, Fabian Caba Heilbron, Humam Alwassel, **Ranjay Krishna**, Victor Escorcia, Kenji Hata, and Shyamal Buch. "Activitynet challenge 2017 summary". In: *arXiv preprint arXiv:1710.08011* (2017).

[9] Charles D Gilbert and Wu Li. "Top-down influences on visual processing". In: *Nature Reviews Neuroscience* 14.5 (2013), pp. 350–363.

[10] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. "COOT: Cooperative Hierarchical Transformer for Video-Text Representation Learning". In: *Advances in Neural Information Processing Systems* 33 (2020).

[11] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. "Iqa: Visual question answering in interactive environments". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4089–4098.

[12] Madeleine Grunde-McLaughlin, **Ranjay Krishna**, and Maneesh Agrawala. *AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning*. 2021.

[13] Kenji Hata, **Ranjay Krishna**, Li Fei-Fei, and Michael Bernstein. "A Glimpse Far into the Future: Understanding Long-term Crowd Worker Quality". In: *CSCW: Computer-Supported Cooperative Work and Social Computing*. 2017.

[14] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. "Relation networks for object detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3588–3597.

[15] Drew A Hudson and Christopher D Manning. "Compositional Attention Networks for Machine Reasoning". In: *International Conference on Learning Representations*. 2018.

[16] Drew A Hudson and Christopher D Manning. "Gqa: A new dataset for real-world visual reasoning and compositional question answering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 6700–6709.

[17] Jingwei Ji, **Ranjay Krishna**, Ehsan Adeli, Juan Carlos Niebles, Olga Russakovsky, and Li Fei-Fei. *Compositionality in Computer Vision*. `http://cicv.stanford.edu`. 2020.

[18] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 2901–2910.

[19] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. "Inferring and executing programs for visual reasoning". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2989–2998.

[20] Justin Johnson, **Ranjay Krishna**, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. "Image retrieval using scene graphs". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3668–3678.

[21] Pranav Khadpe, **Ranjay Krishna**, Li Fei-Fei, Jeff Hancock, and Michael Bernstein. "Low Expectations Lead to Better Experiences: The Effect of Conceptual Metaphors on Human-AI Collaboration". In: *ACM Conference on Computer-Supported Cooperative Work and Social Computing*. 2020.

[22] Kazuki Kozuka, **Ranjay Krishna**, Jingwei Ji, Alec Hodgkinson, Olga Russakovsky, Juan Carlos Niebles, and Li Fei-Fei. *International Challenge on Compositional and Multimodal Perception*. `https://camp-workshop.stanford.edu/`. 2020.

[23] Jonathan Krause, Justin Johnson, **Ranjay Krishna**, and Li Fei-Fei. "A Hierarchical Approach for Generating Descriptive Image Paragraphs". In: *Computer Vision and Patterm Recognition (CVPR)*. 2017.

[24] **Ranjay Krishna**, Michael Bernstein, and Li Fei-Fei. "Information Maximizing Visual Question Generation". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

[25] **Ranjay Krishna**, Ines Chami, Michael Bernstein, and Li Fei-Fei. "Referring Relationships". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[26] **Ranjay Krishna**, Jia Deng, Michael Bernstein, and Li Fei-Fei. *Scene Graph Representation and Learning*. `http://sgrl.stanford.edu`. 2019.

[27] **Ranjay Krishna**, Kenji Hata, Stephanie Chen, Joshua Kravitz, David A Shamma, Li Fei-Fei, and Michael S Bernstein. "Embracing error to enable rapid crowdsourcing". In: *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM. 2016, pp. 3167–3179.

[28] **Ranjay Krishna**, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. "Dense-Captioning Events in Videos". In: *International Conference on Computer Vision (ICCV)*. 2017.

[29] **Ranjay Krishna**, Donsuk Lee, Li Fei-Fei, and Michael Bernstein. "Socially Situated Artificial Intelligence: Learning to Interact and Interacting to Learn". In: *in submission* (2021).

[30] **Ranjay Krishna**, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations". In: *International Journal of Computer Vision* 123.1 (2017), pp. 32–73.

[31] Christopher A Kurby and Jeffrey M Zacks. "Segmentation in the perception and memory of events". In: *Trends in cognitive sciences* 12.2 (2008), pp. 72–79.

[32] Cewu Lu, **Ranjay Krishna**, Michael Bernstein, and Li Fei-Fei. "Visual Relationship Detection with Language Priors". In: *European Conference on Computer Vision*. 2016.

[33] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. "The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision". In: *arXiv preprint arXiv:1904.12584* (2019).

[34] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. "What do people ask their social networks, and why? A survey study of status message Q&A behavior". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2010, pp. 1739–1748.

[35] Junwon Park, **Ranjay Krishna**, Pranav Khadpe, Fei-Fei Li, and Michael Bernstein. "AI-based Request Augmentation to Increase Crowdsourcing Participation". In: *AAAI Conference on Human Computation and Crowdsourcing*. 2019.

[36] Roger C Schank and Robert P Abelson. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press, 2013.

[37] Sebastian Schuster, **Ranjay Krishna**, Angel Chang, Li Fei-Fei, and Christopher D Manning. "Generating semantically precise scene graphs from textual descriptions for improved image retrieval". In: *Proceedings of the fourth workshop on vision and language*. 2015, pp. 70–80.

[38] Rajan Vaish, Snehalkumar Neil S Gaikwad, Geza Kovacs, Andreas Veit, **Ranjay Krishna**, Imanol Arrieta Ibarra, Camelia Simoiu, Michael Wilber, Serge Belongie, Sharad Goel, et al. "Crowd research: Open and scalable university laboratories". In: *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. ACM. 2017, pp. 829–843.

[39] Jeremy M Wolfe. "Visual memory: What do you know about what you saw?" In: *Current biology* 8.9 (1998), R303–R304.

[40] Xu Yang, Chongyang Gao, Hanwang Zhang, and Jianfei Cai. "Hierarchical Scene Graph Encoder-Decoder for Image Paragraph Captioning". In: *Proceedings of the 28th ACM International Conference on Multimedia*. 2020, pp. 4181–4189.

[41] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. "Clevrer: Collision events for video representation and reasoning". In: *arXiv preprint arXiv:1910.01442* (2019).

[42] Sharon Zhou, Mitchell Gordon, **Ranjay Krishna**, Austin Narcomey, Durim Morina, and Michael S Bernstein. "Hype: Human eye perceptual evaluation of generative models". In: *Thirty-third Conference on Neural Information Processing Systems* (2019).