Academic Statement of Purpose

Amita Kamath
https://amitakamath.github.io

Research using machine learning techniques for NLP has yielded increasingly capable systems: able to perform various tasks across different domains and modalities, sometimes even without having trained on them beforehand. One of the key factors in these achievements has been transfer learning: from pretraining to downstream tasks, or across multiple tasks at training time. The emergence of properties in the model during training that results in this transfer is interesting and occasionally unexpected. This makes model analysis critical to better understand what properties our models have (or don't) and how they emerge.

As steps towards the longer-term goal of creating general systems that we understand, I've had the opportunity to work on projects in **transfer learning** and **model analysis** at the Allen Institute for AI and Stanford University. I aspire to broaden my vision and continue working towards general, interpretable NLP systems during my PhD.

Transfer Learning: Transfer learning has enabled great leaps in progress towards systems that understand language: through the transfer of knowledge gained at pre-training time to downstream tasks, as well as through transferring knowledge between various tasks at training time. Models that are able to leverage annotations on one task to do well on other tasks could overcome the inherent intractability of obtaining annotations for the cartesian product of every concept and every task. These models would also be able to perform well on downstream tasks with only a small number of (or even no) labeled examples.

As a Predoctoral Young Investigator at the Allen Institute for AI, I dove into the multimodal NLP field to study these hypotheses. We built a model that could transfer concepts across tasks, as well as a corresponding evaluation mechanism. To facilitate this transfer, we designed an architecture that had one output head per *modality*, rather than per task, so tasks that share a modality have even more parameters in common than previous multitask models. This also made our model able to support any task within the supported modalities without any architectural changes.

I aided in the design of our model and evaluation mechanisms, trained the visual backbone of the model, and conducted an in-depth analysis of the model performance on our new evaluation benchmark, showing interesting evidence of transfer of skills between tasks (e.g. becoming better at "where" VQA questions after training with the localization task, or showing good zero-shot performance on tasks related to our training tasks) as well as transfer of concepts from one task to others (e.g. answering questions about a cat, after only having seen classification data for cats and VQA data about non-cats). This work was received positively by the technical community and is currently **under submission at CVPR 2022** (Gupta et al., 2021).

Inspired by the ability we saw of our model to transfer concepts between tasks, I co-led a project to push this idea much further. Noting that image search engine results tend to return clean, classification-style centered images for even tail concepts (e.g. "hyacinth") while remaining very inexpensive, I carefully curated a large list of queries and obtained 1M image-query pairs from search engine data (for only $150, which shows significant potential to scale up!) and converted them into a QA-style dataset. The data introduced over 10,000 new concepts to our previously proposed model, which was thereafter able to leverage this knowledge in other tasks (e.g. now captioning an image of a hyacinth). I also ran zero-shot experiments on action- and attribute- focused datasets to evaluate our design decisions of adding verbs and adjectives to the web data. My co-authors scaled up the model capabilities and the evaluation mechanism, to facilitate transfer as well as the evaluation of this transfer. This work is currently **under submission at CVPR 2022** (Kamath et al., 2021).

Rather than changing model architectures to facilitate transfer, as our work does above (as do Cho et al., 2021 and Jaegle et al., 2021), other recent research in this space works towards this goal by using new learning algorithms (Min et al., 2021), or by blurring the lines between tasks through the unification of output modalities (Cho et al., 2021) or through the use of textual descriptions of tasks (Mishra et al., 2021) – the latter of which we do in our work as well. All of this work seems to be making small steps towards a longer-term goal: to have models that do not need explicit supervision for each task/domain/concept, but can leverage their existing knowledge to scale up and do well in a wide variety of scenarios. I would like to work in this broad area: how far can we push these different approaches? How can we think bigger?

Model Analysis: The ability of our models to transfer knowledge in various ways can be seen as properties that emerge during pretraining or training that are helpful down the line in downstream tasks or in joint training with other tasks. Many of these emergent properties are useful – some are surprising. This makes model analysis critical, to understand model behavior and have a better sense of how to capitalize on model properties. Of course, as models scale up in terms of parameters, training data, and tasks, it becomes correspondingly more challenging to interpret them, and understand what they do know, as well as what they don't.

Predicting failure cases through model analysis is of even more importance for the realistic phenomenon of distribution shift, under which models tend to fail more often, and become poorly calibrated. The ML problem of selective classification (deciding when to abstain from prediction) had not yet been tackled for complex NLP models. As an MS student in the Stanford NLP group with Percy Liang, I developed a method to perform selective classification for QA under domain shift – answering as many questions as possible from a mixture of in-distribution and out-of-distribution questions, while maintaining high accuracy. I established that exposure of the method to out-of-distribution data helps significantly, even when from a different distribution than the target. I presented this work at **ACL 2020** (Kamath et al., 2020), and was awarded a Distinction in Research from Stanford CS for my masters thesis based on this project. Follow-up work in this line has improved model calibration out-of-distribution (Jiang et al., 2021; Ye and Durrett, 2021), while other research that analyzes model failure cases does so through studying or generating counterfactuals (Jacovi et al., 2021; Gardner et al., 2020; Ross et al., 2021).

Beyond predicting model failures, model analysis is essential to learn more about what models *do* know; particularly because training paradigms have changed so much over the past few years. What do our new training techniques teach models implicitly? This has become a growing field of study in NLP (Tenney et al., 2019; Hewitt and Manning, 2019), vision (Zamir et al., 2018) and embodied AI (Weihs et al., 2019) – and one I wish to explore further for NLP and multimodal NLP. How can we best evaluate existing representations; and how can we use that knowledge to, in turn, inform development of model architectures and training techniques to arrive at better representations – effectively closing the loop?

Career Aspirations: My time at Stanford served as a turning point in my research career – both in terms of my area of interest (my undergraduate research revolved primarily around systems security) as well as my goals. My interactions with faculty, PhD students and post-docs inspired me to pursue research (an aspiration that further grew during my time at AI2), and my numerous experiences as a teaching assistant – hosting discussion sections, organizing office hours, and mentoring student research projects – kindled in me the passion to teach and advise. Working towards a PhD would give me the opportunity to explore these passions further, as well as allow me to gain a deeper understanding of the field and eventually become faculty, or find a suitable combination of industry research and mentorship programs.

Fit at Cornell: At Cornell, I'm excited to work with Professors Yoav Artzi and Sasha Rush, who have done very interesting research involving transfer learning and model analysis, in the NLP and multimodal NLP spaces (Zhang et al., 2021; Kojima et al., 2020; Sanh et al., 2021). I am confident that Cornell would be a great place for me to continue to grow as a researcher during my PhD.

References

Jaemin Cho, Jie Lei, Haochen Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *ICML*.

Matt Gardner, Yoav Artzi, Jonathan Berant, Ben Bogin, Sihao Chen, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Eric Wallace, Ally Quan Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *FINDINGS*.

Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. 2021. Towards general purpose vision systems. *ArXiv*, abs/2104.00743.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *NAACL*.

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive explanations for model interpretability. In *EMNLP*.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Andrew Brock, Evan Shelhamer, Olivier J. H'enaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. 2021. Perceiver io: A general architecture for structured inputs & outputs. *ArXiv*, abs/2107.14795.

Zhengbao Jiang, J. Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Amita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. 2021. Webly supervised concept expansion for general purpose vision models. *preprint*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *ACL*.

Noriyuki Kojima, Hadar Averbuch-Elor, Alexander M. Rush, and Yoav Artzi. 2020. What is learned in visually grounded neural syntax acquisition. In *ACL*.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. Metaicl: Learning to learn in context. *ArXiv*, abs/2110.15943.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions.

Alexis Ross, Ana Marasović, and Matthew E. Peters. 2021. Explaining nlp models via minimal contrastive editing (mice). In *FINDINGS*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang A. Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M SAIFUL BARI, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Stella Rose Biderman, Leo Gao, T. G. Owe Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization. *ArXiv*, abs/2110.08207.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *ArXiv*, abs/1905.05950.

Luca Weihs, Aniruddha Kembhavi, Kiana Ehsani, Sarah Pratt, Winson Han, Alvaro Herrasti, Eric Kolve, Dustin Schwenk, Roozbeh Mottaghi, and Ali Farhadi. 2019. Learning generalizable visual representations via interactive gameplay. *ArXiv*, abs/1912.08195.

Xi Ye and Greg Durrett. 2021. Can explanations be useful for calibrating black box models? *ArXiv*, abs/2110.07586.

Amir Roshan Zamir, Alexander Sax, Bokui (William) Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling task transfer learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3712–3722.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. 2021. Revisiting few-sample bert fine-tuning. *ArXiv*, abs/2006.05987.