Research in NLP has yielded increasingly capable systems: able to perform various tasks across different domains and modalities, sometimes even without having explicitly trained on them beforehand. One of the key factors in these achievements has been transfer learning. The emergence of transferable skills in the model during training 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 NLP systems whose capabilities and limits we understand, I've had the opportunity to work on projects in **transfer learning** and **model analysis** at the Allen Institute for AI (AI2) and Stanford University. During my PhD, I aspire to broaden my vision and continue working towards this goal.

Transfer Learning: Transfer learning has enabled great leaps in progress towards systems that model language: through the transfer of knowledge gained at pre-training time to downstream tasks, as well as through transferring knowledge between various tasks when training on them jointly. 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 every combination of concept (defined here as a noun, verb or adjective, e.g. "cheetah", "running") and task (e.g. QA, captioning). 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 AI2, advised by Aniruddha Kembhavi, I dove into the vision-language field to study these hypotheses. In collaboration with Derek Hoiem at UIUC, my team and I 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* (e.g. text, bounding box), rather than per task, thus unifying the output format of tasks that share a modality, sharing more parameters between such tasks than previous multi-task models, and making the model architecture task-agnostic. I 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" questions about images after joint training with a detection task, and showing good zero-shot performance to new 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 vision-language research community and is currently **under submission at CVPR 2022** (Gupta et al., 2021).

Inspired by our model's ability to transfer concepts across tasks, I led a project that enabled our model to perform these tasks on over 125 times as many concepts as before, at a very low cost. Noting that image search engine results tend to return clean, classification-style centered images for even uncommon concepts (e.g. "hyacinth") while remaining inexpensive, I carefully curated a large list of queries and obtained 1M image-query pairs from search engine data (for only $150!) and converted them into a QA-style dataset. When added to the original training data, the web 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). We also scaled up the model capabilities and evaluation. This project was exciting in several ways: the low cost of web data shows significant potential to scale up even further; and the new evaluation mechanism allows more work to evaluate on "tail" concepts, which will result in systems more useful in the real world. 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, other recent research works towards this goal by (1) using new learning algorithms; and (2) blurring the lines between tasks through the unification of output modalities or the use of textual descriptions of tasks – the latter of which we do in our work as well. All of this work seems to be making steady progress 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: e.g., rather than adding 1M web data examples to training, could we learn to select a small subset of the data that would be most useful for a given input, and either finetune on it or use it for in-context learning (as in Min et al., 2021)? This would result in a more lightweight way to expand model vocabulary to tail concepts.

Model Analysis: The ability of our models to transfer knowledge can be seen as properties that emerge during pre-training or training that are helpful in downstream tasks or in joint training with other tasks. This makes model analysis crucial, to understand model behavior and have a better sense of how to capitalize on model properties. And as models scale up in 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 when the test distribution differs from the training distribution (as often happens in real life), since models tend to fail more often and become poorly calibrated. The traditional 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 (OOD) questions, while maintaining high accuracy. I trained a calibrator to predict when the QA model would err on a given input, establishing that exposure to OOD data (even if from a different distribution than the test data) improved its performance significantly. 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 (Kamath, 2020), advised by Percy Liang and Chris Manning. Follow-up work in this line has improved model calibration out-of-distribution, while other research that analyzes model failure cases does so using contrastive examples.

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, and one I wish to explore further for NLP and vision-language.

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 courses inspired me to reach out to and begin working with Percy, eventually extending my MS by a year solely to continue my research. In my time since then at AI2, I have continued to learn about NLP and ML, as well as how to select and tackle research problems. My numerous experiences as a teaching assistant – getting students excited about course material, and encouraging their interest in research by mentoring course projects – have also 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 Stanford: At Stanford, I'm enthusiastic to resume working with Percy Liang, whose group has recently shown interesting findings about selective classification, and a lightweight way to maximize transfer from pre-training to various downstream tasks (Jones et al., 2021; Li and Liang, 2021). I'm also keen on working with Chris Manning and Noah Goodman, whose groups have done exciting work in how NLP systems model language and conceive concepts. Per my strong track record of success in academics, teaching, and leading research projects, as well as my past experience being part of the Stanford NLP group and the knowledge I've gained since then, I am confident that I would be a good fit, and that returning to Stanford for a PhD would allow me to gain additional exposure to grow as a researcher and as a mentor.

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

Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. 2021. Selective classification can magnify disparities across groups. In *ICLR*.

Amita Kamath. 2020. Selective prediction under domain shift for question answering. *MS Thesis*.

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

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

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*.

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