Sanjay Subramanian: Statement of Purpose

I was drawn to the field of natural language processing (NLP) in 2017 by the significant shortcomings of state-of-the-art models in terms of accuracy on seemingly simple tasks like coreference resolution (Subramanian and Roth, 2019; Agarwal et al., 2019). In 2020, the gap between human and model performance has nearly disappeared for many datasets, so the shortcomings of models are not as visible as they were three years ago. However, a newcomer to the field of NLP will still observe major gaps between the language understanding capabilities of humans and machines.

First, most NLP research focuses on text alone, whereas language in the real world is grounded in sensory experiences. Authors of children's books and science papers alike use language and images jointly to communicate. While visual grounding has been explored, my recent work suggests that current vision-language systems exhibit only shallow understanding (Subramanian et al., 2019; Gardner et al., 2020). I am interested in building systems that possess a deeper understanding of language and vision.

Second, today's high-performing deep learning systems generally require a large amount of labeled training data. In contrast, humans learn tasks from much less data, which is important for application domains (e.g. medicine) and structured prediction tasks where data collection is challenging. I am broadly interested in developing methods for learning from limited labeled data and am particularly interested in using interaction between the system and the annotator to enable systems to generalize from less data.

These two topics are closely related to my long-term goal of developing AI systems capable of assisting humans across a wide range of domains. In addition to these two specific topics, to which I have devoted a great deal of time and effort, I am broadly interested in research that makes progress toward this goal or advances our understanding of why methods fall short of this goal.

Deepening Visually Grounded Language Understanding

Recent work shows that self-supervised pre-training with paired images and text yields impressive accuracy on visual question answering and related tasks (Tan et al., 2019). My work has explored whether these models are doing compositional reasoning (rather than relying on shortcuts) and how to visually ground each step of reasoning in these models, which is important for interpretability.

Upon seeing the impressive performance of these pre-trained multimodal models, I was interested in the question of whether they are doing compositional reasoning. In particular, I studied whether relations between objects were necessary for the questions that the model answers correctly. In one of the associated experiments, I dropped prepositions and verbs (which encode many such object relations) from the input sentences and found that the model's performance stayed nearly the same (Subramanian et al., 2019). Therefore, current pre-training techniques and data are effective in grounding nouns and adjectives but are shallow in that they seem to rely on priors based on object co-occurrences to determine relations.

To address the problem of grounding each step of reasoning in the image, I co-led a project in which we combined the representations from these pre-trained models with neural module networks (NMN)s (Andreas et al., 2016), which provide a list of steps taken to solve each problem as well as the output of each step. In prior work, NMNs were successful in tasks with synthetic language, but we found that on a task with real images and natural compositional language, module outputs were not necessarily *faithful* to their intended function. For instance, the output of "find[dog]" did not necessarily assign high probability to boxes in the image containing dogs and low probability to other boxes. I developed three techniques of improving the faithfulness of module outputs to their intended function -- e.g., using a different architecture for the count module and decontextualizing word representations (Subramanian et al., 2020a). Through both qualitative and systematic quantitative analyses, we found that these methods improve faithfulness and that faithfulness remains difficult for relations and rare objects and attributes.

Motivated by my past work, I aim to improve visual grounding for object relations and rare objects and attributes (few-shot visual grounding). This work will involve creating not only novel algorithms, but also new datasets -- e.g., a dataset in which relations are not implied by object co-occurrences. Finally, I am interested in how images can augment machine understanding of documents such as news or science articles. In this vein, I recently led a project in which we introduced a dataset linking figures and text in medical papers, and I formulated the task of subfigure-subcaption alignment, which is a step toward reasoning over images and text in science papers (Subramanian et al., 2020b).

Interactive learning for more efficient and robust generalization

Deep learning systems require lots of supervision to train and lack robustness to small changes in inputs. I am interested in leveraging interaction -- two-way feedback between the system and the annotator during training -- to address both of these issues. This is inspired in part by the importance of interaction in human language learning (Kuhl, 2004). The most elementary form of interaction is active learning (Cohn et al., 1996), where models can repeatedly select unlabeled examples to label. In one of my first research projects, I found that even this restricted form of interaction is powerful. Specifically, I designed correlation clustering algorithms that can query annotators for whether any two nodes in a graph should be in the same cluster. I proved theorems showing that using a bounded number of these same-cluster queries, the algorithms obtain a significantly better clustering, and I empirically validated these claims in experiments as well (Saha and Subramanian, 2019).

Active learning is also applicable to NLP, and recent work shows that active learning with pre-trained language models (LMs) can improve data efficiency for some NLP tasks (Yuan et al., 2020). However, active learning is a restricted form of interaction, limited by the available unlabeled data and label space.

I am excited about using recent advances in language understanding and generation, such as pre-trained LMs, to facilitate richer interaction than what classical active learning offers. As a first step, I will apply a pre-trained LM to generate prompts for the annotator to request input examples that are maximally informative to the model. I am also interested in enabling annotators to give (and models to solicit) input beyond labels. For example, annotators can give instructions that the model must follow or explanations for specific examples.

This interactive learning paradigm is applicable not only to NLP tasks, but also to other areas of AI, such as vision and robotics. I am interested in using interactive learning to improve data efficiency in these other areas as well as in NLP problems or domains where annotation is expensive.

My second goal for interactive learning is to make NLP models more robust. My motivation stems in part from work in which my collaborators and I showed on many NLP datasets (my contribution was specifically on a visual reasoning dataset) that small changes to test set examples lead to large accuracy drops for state-of-the-art models (Gardner et al., 2020). Using the generated prompts proposed above, the model can request during training minimally contrastive examples, potentially mitigating this issue.

[Concluding and School-specific paragraph. I listed four professors at UC Berkeley who I was interested in working with and described how my interests aligned with theirs. Other research-related reasons you want to go to this university can also be listed here. If you have interacted with these professors or their groups before, you can include that here too.]

References (* denotes equal contribution, ** denotes alphabetical ordering)

Oshin Agarwal*, Sanjay Subramanian*, Ani Nenkova, Dan Roth. “Evaluation of Named Entity Coreference.” _CRAC Workshop at NAACL 2019._

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, Dan Klein. “Neural module networks.” _CVPR 2017_.

David A. Cohn, Zoubin Ghahramani, Michael I. Jordan. "Active learning with statistical models." _JAIR (1996)_.

Matt Gardner, et al. “Evaluating Models' Local Decision Boundaries via Contrast Sets.” _EMNLP-Findings 2020_.

Barna Saha and Sanjay Subramanian**. “Correlation Clustering with Same-Cluster Queries bounded by Optimal Cost.”_ESA 2019_.

Sanjay Subramanian*, Ben Bogin*, Nitish Gupta*, Tomer Wolfson, Sameer Singh, Jonathan Berant, Matt Gardner. “Obtaining Faithful Interpretations from Compositional Neural Networks.” _ACL 2020a_.

Sanjay Subramanian and Dan Roth.“Improving Generalization in Coreference Resolution via Adversarial Training.” _*SEM 2019_.

Sanjay Subramanian,et al.“MedICaT: A Dataset of Medical Images, Captions, and Textual References.” _EMNLP-Findings 2020b_

P. K. Kuhl. (2004). “Early language acquisition: cracking the speech code.” _Nature reviews neuroscience_, 5(11), 831-843.

Sanjay Subramanian, Sameer Singh, Matt Gardner. “Analyzing Compositionality of Visual Question Answering.” _ViGIL Workshop at NeurIPS 2019 (spotlight)_.

Hao Tan and Mohit Bansal. “LXMERT: Learning Cross-Modality Encoder Representations from Transformers.” _EMNLP 2019._

M. Yuan, H. Lin, J. Boyd-Graber. “Cold-start Active Learning through Self-supervised Language Modeling.” _EMNLP 2020_.