Statement of Purpose

Bingbin Liu
Applying to PhD in Computer Science

I am intrigued by the problem of video understanding. Humans understand and interact with the world through continuous imagery unrolling in time, and I hence believe videos should be where vision algorithms extract information as well. As a master's student on the video team and healthcare team in Stanford Vision and Learning Lab, I got the opportunities to work on technical projects that were turned into publications at top vision conferences, as well as studies on ICU patient mobility that led to papers at MLHC'18, ML4H'18 and a submission to Nature Digital Medicine. These experiences on both tackling technical challenges and addressing real-world applications highlighted the importance of **model interpretability** and **data efficiency**. Another challenge is the lack of ability for generalization and continuous learning, which I think could be largely alleviated with a better **memory mechanism**. My goal is hence to build video understanding models that are reliable, efficient and versatile enough for real-world applications by addressing these three points.

Interpretability is of great importance especially in applications for practical use such as healthcare, since it provides the transparency that gives confidence to the reliability of a model. For example when understanding complex composite activities such as "A person walks to a bed and sits on its edge," it is intuitive for a human to decompose it into two parts of "walking" and "sitting". In a realistic setting of natural language video retrieval, common RNN-based language models generate sentence-level embeddings, which are black boxes that are hard to interpret and debug. In my ECCV 2018 project, I proposed a compositional approach that used the structure in natural language queries to match with the temporal structures of complex activities hierarchically. These tree-structured *temporal modular networks* proved to be more effective than previous holistic approaches. Moreover, we were able to better understand how information got propagated and aggregated in the model by visualizing the attention vectors in network modules. I later did another study on fine-grained action understanding, as an attempt to better understand each unit in these structured activities. With the intuition that manipulating verbs and objects were usually interdependent, I chose to use object-level reasoning to help with action understanding and designed a *verb-object graph* that grounded its decisions in both visual appearances and prior knowledge extracted from the dataset. The model showed promising results and was submitted to CVPR 2019. Moreover, the connection weights between verbs and objects could be visualized, which not only increased interpretability but also helped identify challenging scenarios for future improvements.

Both projects also take data efficiency as an important motive. For instance, the *temporal modular network* can learn more efficiently with hierarchical information aggregation and reused network modules, while RNN-based models are usually not as efficient since they collapsed most of the language structures. For the *verb-object graph*, reasoning over individual objects allows the model to better leverage prior knowledge in the dataset and leads to enhanced few-shot performance. Another example is a video prediction project I published at NeurIPS 2018, where we proposed to decompose a video into components with individual objects and to disentangle time-invariant appearance from time-varying motion. The proposed *Decompositional*

Disentangled Predictive Auto-Encoder hence only needs to predict low-dimensional pose vectors, greatly reducing the amount of data required for training. All these projects exposed data efficiency as a challenging problem especially for video tasks. This not only makes computation resources a barrier to research, but also means that the models may readily overfit to artifacts that hinder generalizability. This is even more problematic for real-world applications where the problem at hand can evolve over time. On the healthcare team where I build algorithms to automatically monitor mobility activities of ICU patients, lack of data poses a constant bottleneck to performance, such as when migrating models to data collected at a different venue or incorporating new clinically interesting activities. During these projects, I have also looked into other directions such as few-shot and continual learning as well as unsupervised representation learning, which I hope to dive deeper into in my PhD study.

Finally, the memory mechanism, associated strongly with the ability to abstract, relate and reason, is another topic that I am keen on studying, since I think it might be the key to various challenges such as long-horizon reasoning, learning from limited data and generalization. To my knowledge, video understanding research mainly focuses on one of two aspects: convolutional architectures for extracting better features capturing spatiotemporal information, and structures operating on extracted features to address specific tasks, with the second direction relatively less explored. I hope to find a versatile and elegant mechanism that is widely applicable to various tasks to replace some current highly specific solutions that are somewhat heavily engineered. One particular venue I would like to draw inspirations from is neuroscience. At Stanford, I have taken a course on theoretical neuroscience where my team looked into reinforcement learning for dopamine reward prediction error, and certainly look forward to learning more in the future.

I am truly enthralled by and passionate about video understanding, as I have been ever since my first project on video object recognition in the senior year of my undergraduate study. The past year working at the Stanford Vision and Learning Lab has strengthened my technical ability and presented me with more appealing challenges for future research. I therefore hope to pursue my PhD study at Stanford to continue working on model interpretability and data efficiency with **Prof. Fei-Fei Li, Dr. Juan Carlos Niebles, Dr. Chelsea Finn** or **Dr. Serena Yeung**, and also look forward to opportunities collaborating with the **NeuroAILab** for inspirations from neuroscience research. Teaching is another thing that I am passionate about. I have gained experience as a teaching assistant for *CS231N - CNN for Visual Recognition* and *CS337 - AI-Assisted Care*, as well as a mentor for Girls teach Girls to Code and AI4ALL. In the future, I hope to stay in the academia, where I can continue working on problems I am interested in, learn from fresh perspectives working with people with different talents, and get the opportunity to teach and inspire.