Research Statement

Videos are a huge amount of the visual data in our world today. 70% of internet traffic was videos in 2016, and this is projected to grow to over 80% by 2020[1]. 400 hours of video are uploaded to YouTube every minute[2]. While platforms such as YouTube and television are pervasive sources of video, recent years has also seen the rise of continuous, first-person video recording such as the GoPro and other wearable cameras. Furthermore, ambient video such as surveillance recording comprises a large part of video data, and will only continue to increase with the growth in smart buildings and smart spaces.

My research focuses on using artificial intelligence and machine learning approaches to understand what is happening in videos, so that we can index, catalog, and make use of these large amounts of video in downstream applications. In particular, while the computer vision field has made large progress in recent years on image classification, detection, and richer tasks such as captioning, the state-of-the-art in video understanding has lagged behind. A large factor in this is challenges of scale in video understanding, across multiple axes: labeling training data, modeling the temporal dimension, and inference. In my research, I have sought to develop algorithms to address these challenges of scale in video understanding, and I will discuss some of these works below.

I am also passionate about the use of artificial intelligence to improve healthcare delivery, and have in parallel worked on a number of projects to apply video understanding algorithms to enable AI-assisted smart hospitals. I will also discuss this further below.

Towards Scaling Video Understanding

Machine learning approaches to video understanding present challenges of scale across multiple axes including labeling training data, modeling the temporal dimension, and inference. In order to tackle the challenge of efficient video inference, I worked on an approach for action detection in videos that uses reinforcement learning to learn policies for selectively observing frames in videos to detect actions, instead of densely observing all frames, and therefore gains in computational efficiency [5]. In particular, our intuition is that the process of detecting actions is naturally one of observation and refinement: observing moments in video, and refining hypotheses about when an action is occurring. Based on this insight, we formulate our model as a recurrent neural network-based agent that interacts with a video over time. The agent observes video frames and decides both where to look next and when to emit a prediction. Since backpropagation is not adequate in this non-differentiable setting, we use the REINFORCE algorithm to learn the agent's decision policy. Our model achieves state-of-the-art results on the THUMOS'14 and ActivityNet datasets while observing only a fraction (2% or less) of the video frames.

In another recent work, I tackle the challenge of modeling the temporal dimension of videos for dense recognition of human activity in videos, in particular labeling every video frame with possibly multiple co-occurring actions [4]. To study this problem we introduced a new dataset, MultiTHUMOS, consisting of dense labels over unconstrained internet videos. Modeling multiple, dense labels benefits from temporal relations within and across classes. We therefore defined define a novel variant of long short-term memory (LSTM) deep networks for modeling these temporal relations via multiple input and output connections. We show that this model improves action labeling accuracy and further enables deeper understanding tasks ranging from structured retrieval to action prediction.

Finally, video recognition presents the challenge of obtaining sufficient labeled training data, since videos are much more expensive and labor-intensive to annotate than images. Manually labeling training videos is feasible for some action classes but doesn't scale to the full long-tailed distribution of actions. A promising way to address this is to leverage noisy data from web

[1] Cisco Visual Networking Index: Forecast and Methodology, 2016–2021, 2017.

[2] http://tubularinsights.com/hours-minute-uploaded-youtube/

queries to learn new actions. Towards this direction, I introduced a reinforcement learning-based formulation for selecting the right examples for training a classifier from noisy web search results [3]. Our method uses Q-learning to learn a data labeling policy on a small labeled training dataset, and then uses this to automatically label noisy web data for new visual concepts. Experiments on the challenging Sports-1M action recognition benchmark as well as on additional fine-grained and newly emerging action classes demonstrate that our method is able to learn good labeling policies for noisy data and use this to learn accurate visual concept classifiers.

AI-Assisted Healthcare Delivery

Video understanding algorithms have potential for many useful applications in smart spaces. I am particularly excited about their potential for AI-assisted hospitals, where they can continuously sense what is happening in hospitals and be used to improve patient care. In my PhD, I have led collaborations with a number of partner hospitals towards this objective [2, 1]. We have equipped hospital units with ceiling-mounted depth sensors that record privacy-safe depth video (distance from sensor measurements). Video understanding algorithms can then be used to detect events of interest such as hand hygiene compliance for infection control, or clinical care activities such as line insertions and turning ICU patients in bed. Detection of such activities can provide important benefits such as automated clinical care documentation (relieving nursing documentation burden in critical areas such as the ICU), ensuring that appropriate care guidelines are being followed, and providing data that can be correlated with outcomes.

Future Work

While we have been able to show promising results in using video understanding algorithms for applications such as AI-assisted healthcare delivery, these have still required the effort of significant manual data labeling to obtain training data. However there is a rich variety of events and actions which would be valuable to detect. I am therefore very interested in continuing research efforts in the direction of training video recognition models with few and weak labels, and using this to quickly scale the capacity of AI-assisted healthcare delivery. I would like to create an environment where physicians can specify a new event of interest, and with only a few rough examples be able to obtain a recognition model for that event that can then be used to study and improve patient care.

References

[1] Albert Haque, Michelle Guo, Alexandre Alahi, Serena Yeung, Zelun Luo, Alisha Rege, Jeffrey Jopling, Lance Downing, William Beninati, Amit Singh, et al. Towards vision-based smart hospitals: A system for tracking and monitoring hand hygiene compliance. *Machine Learning for Healthcare (MLHC)*, 2017.

[2] Serena Yeung, Alexandre Alahi, Albert Haque, Boya Peng, Zelun Luo, Amit Singh, Terry Platchek, Arnold Milstein, and Li Fei-Fei. Vision-based hand hygiene monitoring in hospitals. In *American Medical Informatics Association (AMIA)*, 2016.

[3] Serena Yeung, Vignesh Ramanathan, Olga Russakovsky, Liyue Shen, Greg Mori, and Li Fei-Fei. Learning to learn from noisy web videos. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision (IJCV)*, 2017.

[5] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.