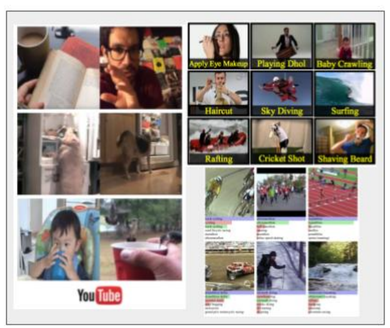
CS229: Machine Learning for Human Activity Recognition from Video

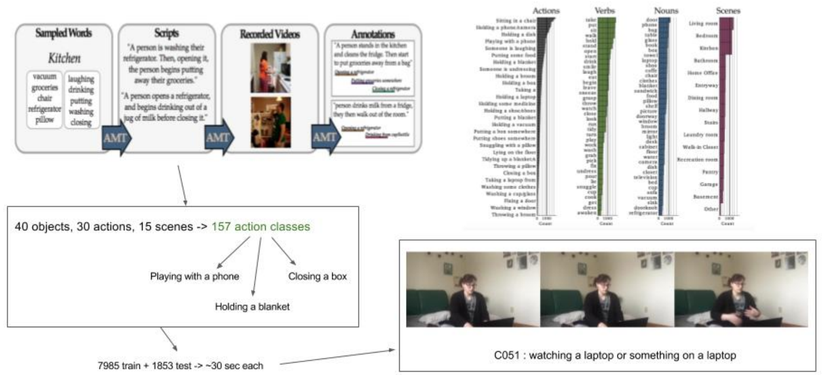
Shikhar Shrestha (shikhars@stanford.edu)

Motivation - Real-world Activity Recognition

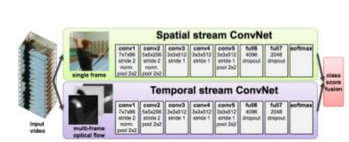


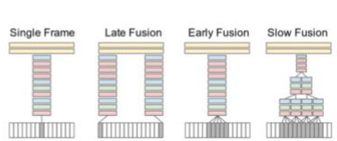
Explore deep representation to recognize activities in real-world video data.

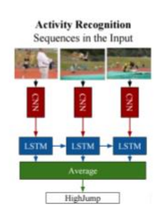
Dataset - Charades from Allen Institute



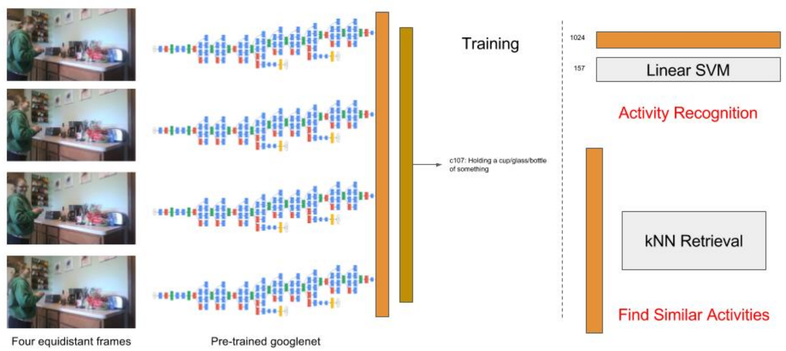
Video Activity Recognition - Available Methods





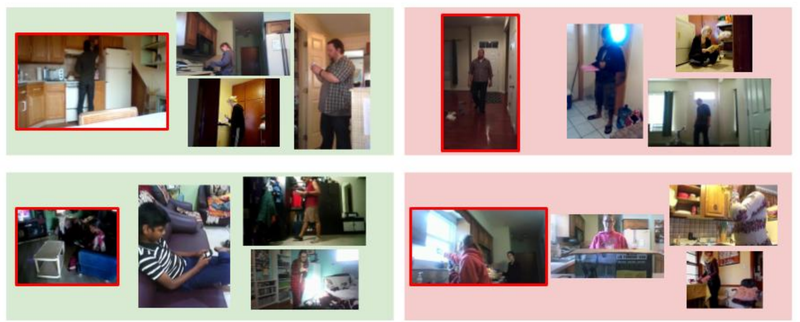


Chosen Method - Multi-Stream Late Fusion

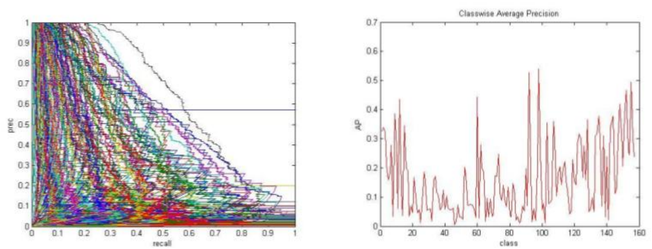


Results - Qualitative [Recognition]

Results - Qualitative [Retrieval]



Results - Quantitative



Method	C3D	Two-Stream	IDT	THIS
mAP (%)	10.9	11.9/14.3	17.2	15.6

Conclusions

Key Insights :

- **~static CNN methods do well for discriminative object/activity interactions**
- **Handling video datasets is very hard and time-consuming**
- **HW bottleneck become very significant (slow inference, out-of-memory)**

Future Directions :

- **Try out the Structure-RNN method on the dataset -- could be interesting!**
- **Look into handling class imbalance**

References

Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." *Advances in Neural Information Processing Systems*. 2014.

Karpathy, Andrej, et al. "Large-scale video classification with convolutional neural networks." *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2014.

Tran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015.

Sigurdsson, Gunnar A., et al. "Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding." *arXiv preprint arXiv:1604.01753*(2016).

Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhutdinov. "Unsupervised learning of video representations using lstms." *CoRR, abs/1502.04681* 2 (2015).