Team Member(s): Brandon Peck (bjp9pq)

Project: Human Action Recognition training with Synthetic Data

Background: Based on your feedback that project was just benchmarking, I would like to consider modifying it to be more interesting. My project has been to try different Human Action Recognition (HAR) models on different datasets (like HMDB51 and UCF101) and match the accuracies of the models I implement with those I found in the corresponding papers.

Over winter break, a friend and I implemented a Reinforcement Learning simulation in which a synthetic person attempts to learn different movements. By modifying the reward function we were able to get it to learn how to sit up, roll around, and stand. In an effort to make my project more interesting I would like to create a repository of three HAR models that are trained with synthetic data generated from the simulations and tested with corresponding real world data. I'll gather the synthetic data from the simulations we made and try different techniques to train the model to not overfit to synthetic data. The test data will be comparatively small and either from webcam footage I take of myself doing the actions or from videos I scrape from online (likely the latter). My goal is to design three HAR models which classify with real world data, given that they have only trained with synthetic data.

Note: this might seem like a lot of work for one student but I've finished two model implementations and started collecting data.

Deliverable:

	HMDB51 - Train/Val Set Real Data - Test Set	Synthetic - Train/Val Set Real Data - Test Set	
2D CNN + LSTM	Validation/Test Accuracy		
3D CNN			
Two Streams			

Datasets:

Train Set - Synthetic Dataset gathered from Reinforcement Learning simulation. We used Unity3D's MLAgents (which used OpenAI Gym) to train a human, SMPL model to: stand, sit up, and lying down. These will be the three actions to classify over and each class will have 5,000 unique videos. Each video action takes place over 30 frames.

Test Set - Real Dataset gathered from webcam of a single room or scraped from youtube and human action recognition datasets. The action classes will also be stand, sit up, and lying down. The size of this dataset will be smaller than the Train Set since scraping data will be more difficult. I will attempt to gather 100 unique videos per classification. Each video action takes place over 30 frames.

Style Transfer Dataset - LSUN Bedroom dataset https://www.yf.io/p/lsun

Data Preparation Technique:

In an attempt to make the Synthetic Data more realistic I will try Neural Style Transfer between the LSUN Bedroom dataset and the synthetic person dataset and possibly between the real person dataset and LSUN Bedroom dataset. I'll attempt to find and equilibrium across the real and synthetic dataset texturing. I am not sure how effective this will be but I will try it.

Goals:

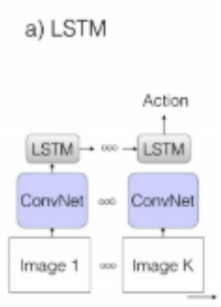
1. Explore HAR literature
2. Gather synthetic person dataset
3. Implement repository of 3+ HAR models
4. Train each model with synthetic data and test on real data
5. Report train, validation, and test accuracy table and confusion matrix for each model

Models to implement:

Specifications I found different approaches to try from this blog post:
http://blog.qure.ai/notes/deep- learning-for-videos-action-recognition-review

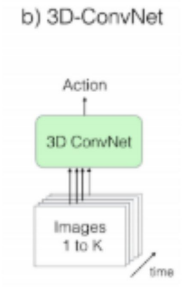
CNN + LSTM

- Run each video frame through a pretrained CNN (vgg, alexnet, resnet, etc.) then use the features generated for the LSTM.



3D-ConvNet

- Look for “spacio-temporal” features in video frames by running a 3D convolution over the video.
- Try different 3D convolution architectures that I find in papers like DenseNet3D and Inception3D found here: https://github.com/MohsenFayyaz89/T3D/tree/master/models
- Build one architecture of my own (applying something like SNIPER would probably be useful)



Two Streams

- There are a bunch of different Two-Stream architectures to try but the goal is use one stream of optical flow to capture temporal features and one stream of convnets to capture special features. These features collected can be fused together at various layers
- Try different architectures
  - Two Stream Fusion
    - This is a model that uses 3D convolutions over optical flow frames generated from a video to find temporal relationships in the video. Then a 2D convolution is applied over a single frame to look for specific features in the image. These two temporal and special features are then fused. https://arxiv.org/abs/1604.06573
  - I3D This model uses two 3D streams instead of one and fuses at the end https://arxiv.org/abs/1705.07750