Moments in Time Summary

Elliot Holtham[1], Moumita Roy Tora[1], Keegan Lensink[1], David Begert[1], Lili Meng[1], Megan Holtham[1], Eldad Haber[1], Lior Horesh[2], Raya Horesh[2]

[1] – Xtract AI

[2] – IBM Research

Before tackling the 339 class challenge, the full dataset was split into 20 classes to create a smaller test problem upon which different solution methods could be examined more quickly. For the 20 class subset, a variety of actions were chosen that would require different data streams for effective classification (for example barking & clapping for audio), bowling and rafting for RGB images, and ascending for motion. After looking through several of the training videos, it was apparent that the content and action of the video could abruptly change throughout several of the videos. Each 3 second video was split into 3 x 1s segments with hope that at least one of the second segments would capture the main action of the video.

For the mp4 videos which contained audio, .wav files were extracted from the video. Initially the .wav files were converted into spectrograms which were then trained using ResNet 101 network. Because of lack of time on the final 339 class problem, a pre-trained VGG on Audioset was used for the audio files. The features from each second for each video was extracted and then the three consecutive feature vectors were passed into an LSTM for the classification and the creation of the first data stream. The features for the videos with no audio files were zero padded such that the dimensions matched the other streams.

Static images were extracted from the videos at 5 fps using ffmpeg. The images were used to fine-tune a pre-trained ResNet 101 model from ImageNet. The trained ResNet 101 model was used in two ways. Firstly, the features from each frame were extracted and one random frame feature from each second used to train a LSTM to create a separate stream. Secondly, as in the MIT/IBM paper, the logits from 6 equidistant frames were averaged to produce a separate data stream.

Motion from the videos was extracted in three ways. Firstly, a pre-trained (ImageNet then Kinetics) I3D model was fine-tuned on the Moments data at 15 fps. Unfortunately our team was running out of time so didn't manage to fully fine-tune this model to the level that was certainly possible. Secondly, temporal slices from the videos were extracted from the 3D volume and used to fine-tine a pre-trained ResNet 101 model. For the 20 class subproblem, we had originally worked with our Leap-Frog network architectures (https://arxiv.org/abs/1705.03341) and had gotten better results than the ResNet 101 network, but didn't have time to train from scratch on the full 339 class problem. Thirdly, a pre-trained TRN model was also used that was provided by the "MIT-IBM Watson AI Lab and IBM Research" group. This provided model was run on the validation and test datasets.

The audio, temporal and spatial stream features were combined together with an LSTM before all of the logits from each stream was ensembled together in a weighted average of the logits

where the weights were based on the accuracy of that data stream on the validation dataset. All of the computations were run internally on desktop computers running GTX 1080Ti video cards and 500 GB M2 SSDs. The above workflow gave 34.00% top 1 and 61.75% top 5 on the validation set.