Video Classification: WLASL Dataset

The main thing that separates videos from images is that videos have a temporal structure in addition to the spatial structure found in images. Video is just a collection of images operating in a specific temporal resolution i.e. frames per second.

WLASL Dataset

Classes: 2,000 common different words in ASL.
Dataset
Total number of videos for training: 18216
Total number of videos for testing: 2879

PREPROCESSING :

1. Convert to mp4, extract Youtube frames and create video instances,

Algorithm Used:

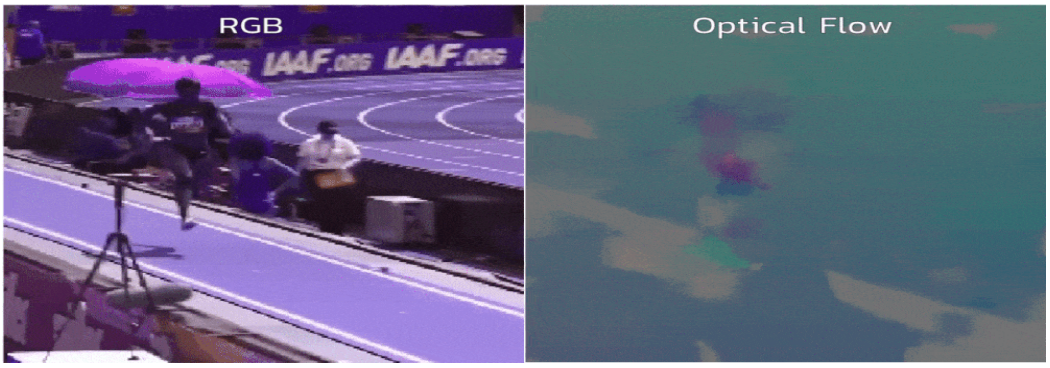
I3D is a 3D video classification network: Uses 3D Convolution to learn spatiotemporal information directly from videos.

- It expands the filters and pooling layers to 3D.
- Transfer Learning: The weights and biases that a model uses to detect features in one domain will often work well for detecting features in a different domain, if the two are similar. This is called **transfer learning**.
    - ❖ Weights and biases were taken from Inception v3 model(pre-trained from ImageNet Dataset). When inflating the 2D model into 3D, they simply took the weights and biases of each 2D layer and “stacked them up” to form a 3rd dimension
    - ❖ The final I3D architecture was trained on the Kinetics dataset, a massive compilation of YouTube URLs for over 400 human actions and over 400 video samples per action.

Model Used:

Mode used: RGB Frames [Other option: optical flow] Optical
Model Set Up:
Original model: InceptionI3d (400, in_channels=3) : 400 classes and 3 input channels // In the original kinetic dataset
https://github.com/deepmind/kinetics-i3d

Tuned Model : The class number varies in our WLASL datset, only the last classification layer is modified in accordance with the class number.So, the last layer is replaced from 400 to 2000 neurons.



Imp. Files:
Configuring Parameters: ‘`configfiles/asl2000.ini`’
Pre-trained weights :
`archived/asl2000/FINAL_nslt_2000_iters=5104_top1=32.48_top5=57.31_top10=66.31.pt`

LINK to COLAB:
https://colab.research.google.com/drive/1eGksOOM6U01e-Sryyl8wqImHPU5pgKj2?usp=sharing

Ref:
https://towardsdatascience.com/exercise-classification-with-machine-learning-part-ii-d60d1928f31d

Word-level Deep Sign Language Recognition from Video

The original I3D network is trained on ImageNet and fine-tuned on Kinetics-400. In order to model the temporal and spatial information of the sign language, such as focusing on the hand shapes and orientations as well as arm movements, we need to fine-tune the pre-trained I3D. In this way, the fine-tuned I3D can better capture the spatio-temporal information of signs. Since the class number varies in our WLASL subsets, only the last classification layer is modified in accordance with the class number.