Audio-Video Correspondence Study with Deep Learning Networks

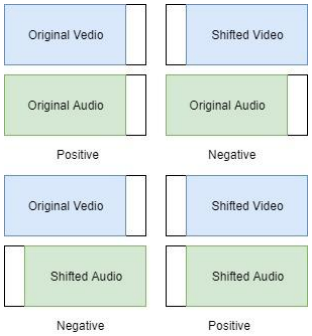
Han Chen (chczy@stanford.edu), Pei-Chen Wu (pcwu1023@stanford.edu)

Introduction

Human's vision and hearing system are closely intertwined through a process known as *multisensory integration*[1]. In this project, we are trying to simulate this process with a deep learning network. The objective is to solve Audio-Video Correspondence problem-determine whether a given audio-video pair is aligned.

Data

AudioSet from Google Research
1 clip -> **full-combination** data:



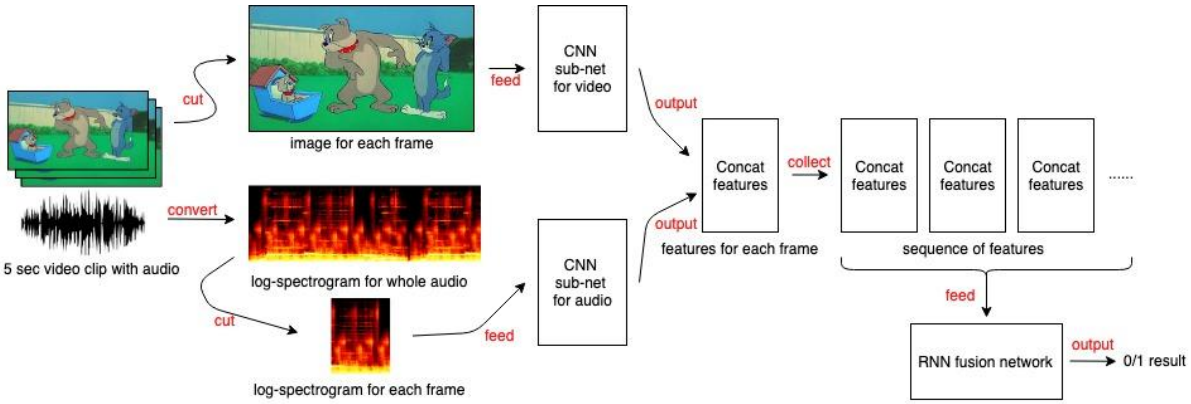
Features

Each frame image has features like color, shape, posture. Videos would have features include movement, transform, color shifting.

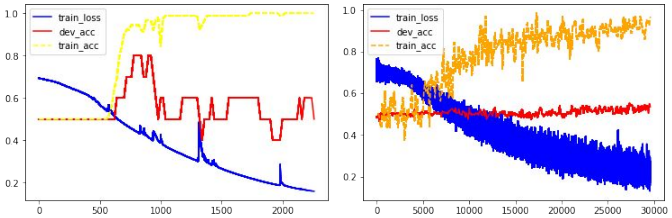
Audio signal has features amplitude. Audio wave has features such as waves, impulse, enhance, fade-out.

The combination of audio and video features are the target to learnt by this model.

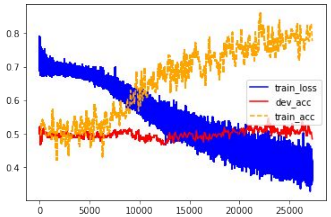
Model Architecture



Results



CNN-NN-ManyToOne **CNN-CNN-ManyToOne**



CNN-CNN-ManyToMany

Most models we trained can finally fit the training set. But the variance between training set and dev set are all huge.

Discussion

- Extend final prediction into multi-dimension, calculate MSE instead of sigmoid entropy loss
- Limited by computation power, can't train on huge data
- Model is not deep enough
- Noise in training data
- Data distribution issue, can limit the training set into specific categories
- Different wave of audio processing
- Transfer learning for video or audio sub-network

Reference

Owens, A., Efros, A. A.: *Audio-Visual Scene Analysis with Self-Supervised Multisensory Features*. arXiv preprint arXiv:1804.03641v2 (2018)
Arandjelovic, R., Zisserman, A.: *Objects that sound*. arXiv preprint arXiv:1712.06651 (2017)
Arandjelovic, R., Zisserman, A.: *Look, Listen and Learn*. arXiv preprint arXiv:1705.08168v2 (2017)