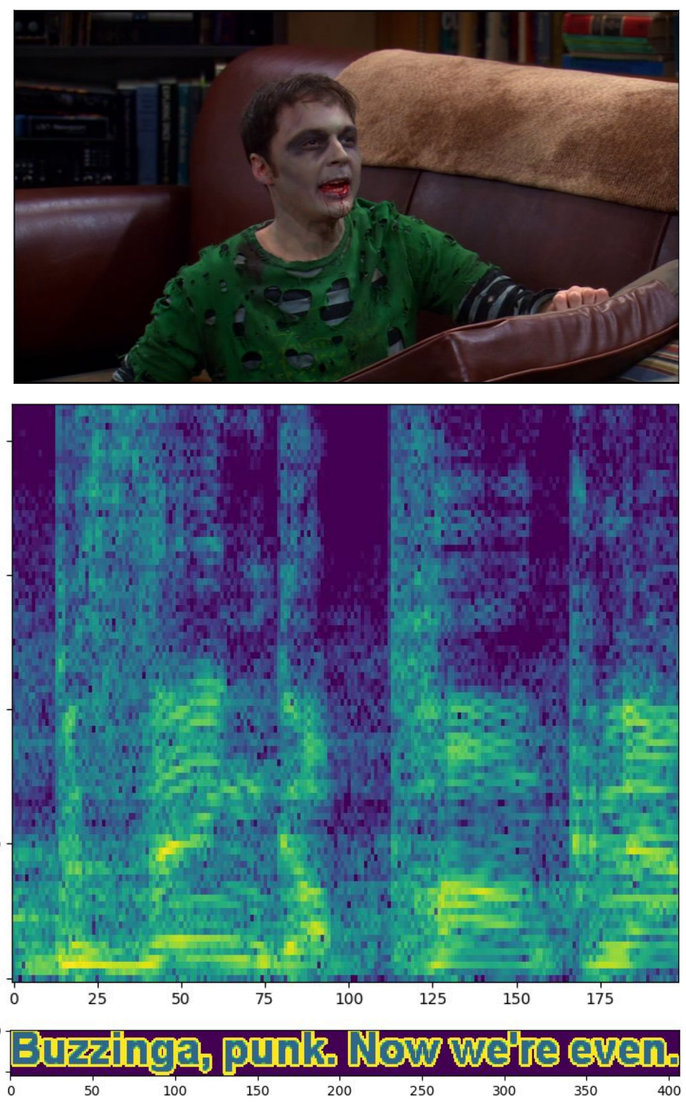
Laughter Anticipation in TV shows using Deep Neural Networks

Joseph Young
Supervised by Dima Damen and Evangelos Kazakos
University of Bristol, Department of Computer Science

Introduction

It is common for TV shows to produce many episodes, resulting in hours of video content. Often this content is annotated with subtitles, and also includes multi-channel audio which allows for the extraction of features such as audience responses, like laughing. Given the large amount of labelled data, this suggests applications within deep learning. This project will produce and investigate the performance of neural networks which predict whether or not a segment of video will be followed by the audience laughing.

The dataset used for training and testing will be 10 seasons of The Big Bang Theory, amounting to over 75 hours of video content. The project will investigate approaches to network design and training from the field of action recognition, as we believe the methods used within action recognition are applicable to the challenges that are prominent in the proposed problem. While the direct application of the solution would be to automatically insert laughing into content, it may also be able to score content given how often the program believes the audience would laugh. There is also the possibility of transfer learning, reusing learnt features in other areas, such as enabling AI assistants to respond to humour.



Modalities

The project will investigate the effectiveness of network architectures on multiple modalities both separately and combined with fusion.

The image stream consists of RGB frames at 720x576 resolution, 25 times a second. These images will both be used on their own as individual frames representing a short segments, and to construct optical flow as this may be useful to the network.

The audio stream consists of 5.1 channel audio, with the actors voices in the main center channel, with laughing in both the center and side channels. However, the side channels are almost entirely used for the audiences laugh and so can be used to extract when the audience is laughing. We will investigate representing the audio in raw form, spectral form, and possibly also in wavelet transform.

Subtitles in are in image form, requiring optical character recognition word embedding to be suitable for use. However, we may also investigate directly using the subtitle images in convolution networks to compare performance.

Inspiration from Action Recognition

The field of action recognition in videos has a large amount of work both complete and on-going, so it is useful to liken the problem at hand to the field as often as possible.

One might view the problem as detecting the action of actors being humorous or comedic within a video segment, and in this view the link is obvious.

Some of the work we can make use of from action recognition is the on data-fusion for the multiple modalities of video, and temporal segment networks which do not depend on the length of a video by sparsely sampling the video in a uniform manner.