The Youtube-8M dataset is a large-scale labeled video dataset that is compromised of millions YouTube video ID’s and labels from over 4700 classes of labels. The data is precomputed so as to reduce the size from Terabytes of data, to merely Gigabytes. This dataset is especially exciting to be working with because of how recently it has been released to the public. Google and YouTube first announced the dataset in September of 2016 and has even released new feature extraction code in November of this year, that we were able to work with.

The dataset comes in two major flavors, video level and frame level features. Despite the fact that over seven million videos have been used to build this set, the fully compressed dataset is still smaller than two terabytes.

The labels, that are coupled with the video ID’s, represent different categories of video. For example a video of Lebron James dunking in the middle of a basketball game would receive the labels, sports, game, and possibly celebrity. The average number of labels per video is about three and a half.

The dataset also contained audio features that we did not manipulate in any way but we hypothesized about how coupling those features along with frame and/or video level features could lead to interesting results.