National College of Ireland

Categorization of Audio/Video Content using Spectrogram-based CNN

MSc Research Project
Data Analytics

Nishant Rajput
Student ID: x17170508

School of Computing
National College of Ireland

Supervisor: Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Nishant Rajput
Student ID:	x17170508
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Vladimir Milosavljevic
Submission Due Date:	12/08/2019
Project Title:	Categorization of Audio/Video Content using Spectrogram-based CNN
Word Count:	6511
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	10th August 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Categorization of Audio/Video Content using Spectrogram-based CNN

Nishant Rajput
x17170508

Abstract

Understanding the content in the videos or audios has been an important task since a long time now. By developing the understanding of content the world could actually become a safer place for example if the hatred videos could be blocked before it spreads or the adults content can be prevented to be shared with the children. Imagine if the systems were smart enough to stop the hatred spread during Christchurch shootings, then maybe christchurch massacre wouldn't have happened as the main motive behind the shootings was to spread the hatred. This research steps towards the identification of content in the audio and videos using the embedded audio present in the audio. In this research a state-of-the-art audio CNN is developed which could even run on the basic machine and do not necessarily require high end devices to run on the cost of loosing the accuracy by 3-4%. This model can run up to the accuracy of 94.88% and this model is able to classify the content even in the presence of the background noise though the performance get hampered a bit by the induction of noise but it is still feasible enough to run and obtain output from the model.

Keywords– CNN, Audio Classification, Video Classification, VGG

1 Introduction

The new era of multimedia has given rise to an enormous number of videos getting captured and uploaded on different websites such as YouTube, Facebook, etc. Given the exponential growth of videos files, the traditional text-based search method of retrieving the videos needs to be replaced by content analysis. Thus, understanding the content of the videos for multimedia indexing and retrieval is perhaps a key factor in the analysis of the video to absorb insights.

Most of the video classifications are based on picture frames which fail if the video quality is poor or below optimal standard for Image classification model to work. While visual content in the video contains key elements for event detection many of the researchers are encouraging attention towards more concrete audio effects with high-level semantics. Ever since the evolution of digital era and easy availability of web-scale data exchanges, there have been various kinds of development by researchers in these fields to design a sophisticated algorithm to identify, index, retrieve and organize the video files by discrete features that are embedded in the video.

With the advancement in the field of Deep learning comes the feasibility for researchers across the globe to learn from the videos. The top software giants of the world also

consider automatic retrieval and management of the multimedia files as a major research area. Human like ability to identify and relate sounds from audio is a nascent problem in machine learning audio event detection. Unlike image classification, audio classification faces another problem as the audios in the video last for a very short span compared to the objects in the image classification which occupy dominant part of the Image. The key factor in maximizing the advancements of deep learning models ability to self-learn depends on the labeled dataset.

Google recently launched labeled Youtube 8m challenge dataset to expedite the research by using the multimedia files from worlds most popular video sharing site on the web, YouTube. Also, Audioset and UrbanSound dataset contains labeled audio files which can be used to train and test the model. The key idea of the research is to identify sound for each frame, forcing the network to pay attention to acoustic details in the video clip and categorize the contents by tagging.

Research Question:

How the Accuracy in understanding the content of video be enhanced using deep Convolution neural network model trained on the spectrogram of the Audio ?

The Model developed here generates the spectrogram from the audio and further analysis of the image is done in VGG fashion. VGG is a model developed by Google to understand the image classification. A sample spectogram is pasted in the Figure 1 below.

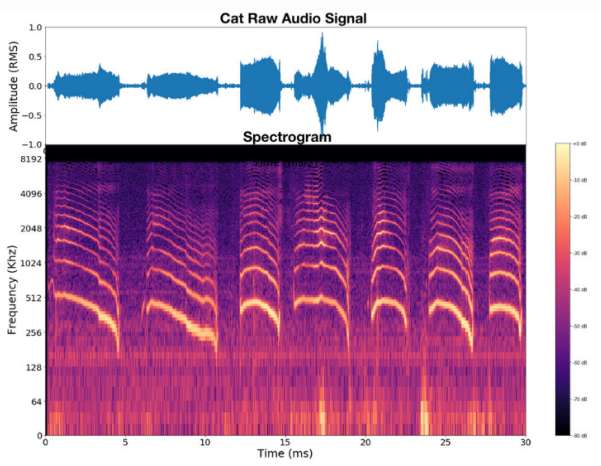

Figure 1: A sample Spectogram

With the development of web and multimedia technology, the Internet has become ubiquitous with videos. The objective of the research is to use Audio from the video file to intensify the understanding of the content in the video, facilitating content identification and classification. Another significant importance of this research is developing a better solution for the classification of videos without using a lot of computation power. The research ideals in computing the classification of video with low CPU usage and minimal loss of accuracy. The research methodologies used are quantitative and research type is secondary. It is inductive research as it involves building a classification model with low computational power on the dataset, the classification model then can be applied to categorize and classify any available accessible video or audio.

Furthermore, This model follows an approach which involves reading the data, generating the spectrogram and analyzing the spectrogram to train the model and later making a prediction on the trained model. The Workflow is similar to the described in the Figure 2

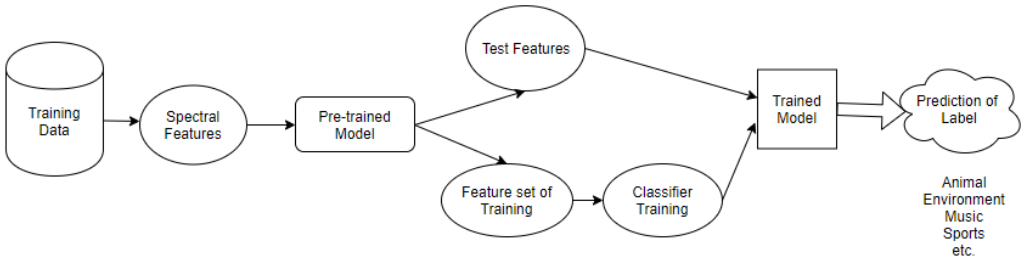

Figure 2: WorkFlow for Audio Classification

In comparison to the image data domain, very little work is observed in using high caliber deep learning models for video classification. A state-of-the-art approach has been applied in this paper in which the video classification is done using the audio content embedded in the video and keeping the accuracy in mind the main motive behind this research is to reduce the computational power and make the model accessible for the cheaper machines as well. The developed model has a tendency in which the model can be trained easily 40 EPOCH and get to the accuracy of almost 93% however the highest accuracy is achieved by training the model up to 150 EPOCH which is approximately 96.8%.

2 Related Work

Classification of video and audios content had been a difficult task for a lot of researchers and many researchers have performed the searches and provide solutions towards the classification/ categorization or identification of the available content. The video can mainly consist of the three different types of content, it generally consists of the audio, a series of picture frames and the captioned text and hence the researchers have segregated the content by analyzing any of the three information. Majority of the research is conducted on the picture present in the content. Using the available picture in the video at 1 picture per frame the researchers have trained the models which could identify the content with the over the time span of the video. The major identification is done in such a way that researchers are able to identify what is the content of the data provided in the audio or video. In the analysis done by the (Brezeale and Cook; 2008) a clear and critical analysis is conducted on the segregation of the content. The related work for the classification of content is divided majorly in three ways which are discussed further in this section.

2.1 Identification of the content using the information present in the picture

A research performed by (Yue-Hei Ng et al.; 2015) helps in understanding how can we use and train the model by stabilizing the video content on 1 picture per scale and training the model with labeled data and further with multiple epochs to increase the accuracy.

The researcher has suggested multiple ways to identify the content of the video, the better model could be build using the Convoluted Neural Network and comparatively easy and handier model could be built using the LSTM. The author has used a process called as independently feature pooling networks in which the author allows the model to get trained using the picture frame data available in the content of the video. After this, Yue-Hei Ng et al. (2015) suggests that the output of CNN can be merged with the network of LSTM and the engine could further keep on adapting and increasing the accuracy by learning from the output of both the models. Technically the researchers have majorly used the AlexNet and GoogLeNet which are a type of CNN.

2.2 Identification of the content using the Multimodal- picture, and text

Another novel approach which is adopted by researchers explains how they managed to train the model using the videos which are 2 minutes in length and they broke down the videos to 1 picture per second that is 120 pictures were used to train the model. Moreover they took the LSTM output and Feature pooling output which is merged further to increase the accuracy and precision of the model and they were able to increase the accuracy from 60.8% to 73% which is a good increase in the accuracy with comparison to the previous models. Though the accuracy of this model is high but on the contrary it requires a lot of hardware power, High CPU usage, High graphical usage and hence it is an expensive model to train. Apart from being expensive if the model has a little bit of blurry images the model starts to fail and wouldn't be able to perform the task it is being assigned.

The research done by (Lin and Hauptmann; 2002) tries to classifies the videos which consist of news. Here the researchers used a multi-modality modal which combines the image processing with the text processing. The output was combined using the SVM and apart from this the researchers used the probabilistic approach to combine the output from both the models. It takes a lot of efforts to build a classifier with really high accuracy so the researchers opted for a different approach and they used the combination of different models. This is one of the best ways to increase the accuracy by combining the output of different models and merging them using probability or other approaches to significantly increase the accuracy. It is definitely a better approach because to increase the accuracy for a single classifier than after a certain point it becomes constant and even if the accuracy is increased it increases with a very low rate.

Output statistics from this research depicts that classifier built on the image derived from the video is more accurate when compared to the text analyzing classifier but when both the classifier has merged the output of both the classifier is even better. The research shows clearly that by using the multi modalities model we can easily reduce the noise and get better results even in the noisy environment. This model is comparably good but the only problem with this model is that it is restricted to News which is a very small horizon of the practical world. This approach needs to broaden and application of this approach is a must in a different real-life scenario as well.

2.3 Analyzing the Dynamics of Video using picture

(Roach et al.; 2001) In this particular research the technique of video dynamics is used, this technique involves the analyzing of dynamic of video and authors have used it to

classify the videos in three major genres which are Sports, Cartoons and News. This technique involves the comparison and analysis of the images in the foreground and the background of the video over a span of time and here the comparison on done over the 30 seconds of the video which helps in understanding the video and classifying it.

The authors have used a model which is known as the Gaussian Mixture Model and in this model, the comparison of the image is done at the pixel level and a proper flow of image and pixels is drawn over the mathematical axes 3D plane. Further, the results showed that the percentage of error was decreased to merely 17% for the background image dynamics and the error percentage for the foreground was merely 8%. However, when researchers combined or merged the output of both the motions the resultant error percentage was degraded to merely 6% and which is indeed a very accurate classifier. Further, this model could become better by adding audio or textual reference and moreover this classifier just consist of the three classes and in the real world scenario we have way more scenarios than these three scenarios and so more data can be fed to this model to increase the number of the classes.

A lot of various other methods have been used by the researchers, in this paragraph, some of the techniques will be discussed. A state-of-the-art approach is used by (Dimitrova and Agnihotri; 2000) the patterns have been used and further analysis of text trajectories and faces is done to identify the video's class. the HMM i.e. Hidden Markov Model is used to the classification. In this article, the researchers have used a high end developed algorithm which is designed by (Agnihotri and Dimitrova; 1999) and this algorithm majorly functions on the textual context of the data. The data that has been acquired from the video's face and text is the measure of development that has occurred, size and to what extent it was available and by examining every aspect a lot of researchers have attempted to put the recordings in 4 classes which they named as sitcoms, Soaps, Commercial, and news. Also, to identify the Face the researchers have used the algorithm designed by (Wei and Sethi; 1999) and to identify the text the algorithm designed by (Agnihotri and Dimitrova; 1999) is used. As discussed in the last paragraph this study also takes into account the change in the contrast of the pixel of the picture and this is how the calculation of image is done. Textbox, detection of Edges, filtration, etc. in this study is done using the work of (Agnihotri and Dimitrova; 1999) and as mentioned above the face detection is used with the help of (Wei and Sethi; 1999) and this algorithm has the tendency of comparing the pixel and hence it can differentiate in the face color and background or greenfield and object etc due to change of value in contrast of every pixel and it can then understand what the object is by drawing the boundaries. Further, with the help of the labeled dataset, it can learn the object's size and shape, the basic foundation model of face detection lies on this research and with this algorithm they were able to get an accuracy of 85%. However, the advancement in the CPU power and complex algorithm this study has been assumed to be an obsolete study and a lot of better and powerful models are available in today's world.

2.4 Analyzing the Dynamics of Video using Text and Picture

The author (Gibert et al.; 2003) has extended the research done by (Dimitrova and Agnihotri; 2000) and tried to create a model with the HMM and the author have used this approach to classify the different types of sports which are Soccer, Football, Ice Hockey and Basketball. The approach behind this research is to identify the color of in the background that is the color of the field and the movement of the camera and the iden-

tification of text has also been taken into consideration so that they would comprehend the difference between a sport and non-sport video. The authors have combined the features extracted from the Motion and color of the video and they have further used the HMMs to combine the output of Motion Features and color features for the analysis. In the implementation, the authors implemented this model and processed a video of 220 minutes which comprised of all the different four sports. Individually accuracy for motion and color is pretty low which is 53% and 77% but when on the other hand the accuracy of both the output is added we drastically high accuracy of 93% . Though it has good results but this is a very restricted research as it concerned with only four sports but the approach is amazingly good.

A state-of-the-art approach is used in (Zha et al.; 2015) where the author has assessed and worked on the Image trained CNN model which has been already trained. So, to perform this task the author used a CNN model which is called ImageNet but the inadequacy of this architecture is that it does not take motion into the consideration. Hence another approach was developed by (Lin et al.; 2009) where they captured this motion information using the Optical Flow Descriptors. This inadequacy is also challenged by distinguishing the Optical Stream-Based Descriptors (e.g. (Lin et al.; 2009)), descriptors from Spatio-temporal interest points (e.g., (Laptev; 2005), (Dalal and Triggs; 2005), (Willems et al.; 2008), (Laptev et al.; 2008), (Wang et al.; 2009)) or estimated movement directions (e.g., (Wang et al.; 2011), (Jiang et al.; 2012), (Jain et al.; 2013), (Wang et al.; 2013), (Wang and Schmid; 2013)).

Convoluted Neural Networks (CNN) over the last decade has developed to be the best approach to understand machine learning and it is shinning on top of all the other algorithms. All the methodologies which are based on the CNN are providing really high accuracy and efficiency especially since the publication of the ImageNet there has been a significant increase in the accuracy of the Image researches. For instance, the model which is developed in research by (Krizhevsky et al.; 2012) is transformed to a different dataset which is called as PASCAL VOC dataset. The top algorithm has been developed by Google which is known as the GoogleNet and VGG architecture. These model are the most accurate models present in the world as of this moment and they are developed by the team of Google. Now researchers used these CNNs and using them they exploit the variables which can be changed to get the better accuracy or performance and moreover some researchers even feed the output of these CNNs to a classifier like SVM. After evaluating the results it has been observed by the researchers that the CNN based model is the highest performing models and they can easily beat any ordinary Spatio-temporal and motion models. But in spite of being such a great model, these models also fail when the quality i.e. the pixels of the pictures is not good or the image is blurry and hence there's a need to analyze the content using other methodologies as well or multi-modalities model. The researchers (Karpathy et al.; 2014) has published a report which is quite same as the approaches that we are discussing here but they took this research to a whole new different level. Before this research people used to classify the data into 3-4 classes only but these researchers have performed this task on the data of 1million videos which consists of 487 classes.

2.5 Classifications using the Audio content in the Video

As early as this research (Lu et al.; 2001) the stepping stone were kept for this type of Audio classification. In this research, the author explains and demonstrates how the

information can be extracted from the audio of the video. Generally, the building blocks for researches like this functions on the base that whether the audio present in the video is filled with speech or no-speech elements. This high level of differentiation is done using the famous KNN (K Nearest Neighbour) and LSP VQ (Linear spectral Pair Vector Quantization) and after differentiating the primary sounds the further step is to understand the sound whether it background noise or music or any kind of environmental sound. Moreover, another task is to understand whether there is a silence or not and to understand this part the researchers have used the zero-crossing rate and short-term energy and analyzed it if this is less than the threshold or not and if this is less than the threshold it is considered as silence. Now, the remaining part is the non-silence and to understand this last important part the researchers have used spectrum J and X (SF) and Band periodicity (BP), noise frame ratio (NFR) to differentiate between whether it is music playing or environment sound and this is how in this research the authors were able to differentiate between the noise, environmental sound, music sounds, and silence.

With this approach, the accuracy percentage was very high and it reached up to 98.03% for segregating the music and speech and also the accuracy for all different classes which are environment sounds, silence, music, the speech was scored up to 96.51%. This research certainly has an amazing accuracy but it lacks the variety and hence more classes are needed to be added to run it in the real world scenario.

Another study on the natural sound was conducted by (Aytar et al.; 2016), this research is more sophisticated than the previous researches as they have taken 2 million unlabeled videos into consideration for the analysis of wild. They developed a new network of sound which they called SoundNet and this network has gained quite a lot of attention in the research world. By the research (Lee et al.; 2009) a CNN model was generated in which the characteristic is obtained from the audio and convolutional-restricted-Boltzmann-machines (CRBMs) is used for convolutional-deep-belief-networks (CDBNs) which is an advanced way for RBM. Further, PCA is employed to cope up with the higher dimensionality of the model and spectrogram. Also, various audio classification on music genre is conducted and build on top of this model, and it has been observed that after employing more classifications method the accuracy certainly boosts up and (Li and Guo; 2000) have used SVM to perform the same classification. In research done by (Takahashi et al.; 2017) the authors have performed and created a new network of analysis which they named as AENet and this is used extensively for the analysis of the audio features. More research on the same dataset is done by (Wu et al.; 2015), (Girgensohn and Foote; 1999), and particularly on youtube 8m video these researches have been done, (Abu-El-Haija et al.; 2016),(Na et al.; 2017)(Li et al.; 2017),(Wang et al.; 2017)

2.6 Summary of Related Work

After reading through and analyzing the literature mentioned above thoroughly it can be seen that there has been a great effort put in to the understanding of video and audio content and the researchers have majorly distributed the entire understanding the content on the basis of three pillars which they call as modalities and the modalities are Video, Audio, and text where video is nothing but the analysis of a series of pictures at different level. Researchers in (Dimitrova and Agnihotri; 2000) and (Lin and Hauptmann; 2002) research have demonstrated a technique long back where they have compared the contrast of the picture at the pixel level and tried to draw the image at the pixel level to

understand the objects in the video. This research has been one of the finest technique to identify the objects in the video. Some researchers like (Lin and Hauptmann; 2002) presented the classification using a classifier built using SVM while others have used a different approach and built the classifier using the HMM. However, using any classifier or technique which is based on the image has its flaws like when it'll be dark the pixels won't be able to detect and if the image is blurry the pixels will fail again to identify the content, moreover, when high-efficiency devices like high performing GPU and powerful CPU are needed for analyzing these type of data. Also, text depending model identify the text by using the picture in the video and which is again will become problematic in case of bad pixels or blurred image. Another research on classification is done on the basis of the audio i.e. the embedded audio the AENet is introduced by (Li and Guo; 2000) and the SoundNet is introduced by (Aytar et al.; 2016) and these both are effective model of classification but they are not good when compared with the latest innovation that has been done in the latest years where GoogleNet, Alexnet, VGG have been introduced and a research by (Hershey et al.; 2016) has been done which involves the latest developments.

Also, another aspect or angle that has been considered by the authors is regarding the dynamics of the video which involves the foreground and background motion analysis. These researches were discussed and conducted by (Wu et al.; 2015) and (Roach et al.; 2001). After analysing all the methods available it can be observed that whenever there's been an involvement of more than one classifier the accuracy has always been improved and the best way to design a classifier is to increase the accuracy not just by perfecting the same classifier rather including another classifier and increasing the efficiency with the help of both the classifier. (Lin and Hauptmann; 2002) and (Gibert et al.; 2003) would be the complex models where the accuracy is improved by designing some state-of-the-art complex solutions for the classification.

3 Research Methodology & Design Specification

For the development of any project, the way it is being developed is one of the most crucial parts, for the development of this project the famous CRISP-DM approach has been used. The CRISP-DM approach is one of the widely known approaches to execute the Data mining projects and this approach provides a structured way to perform things and it provides the practicality, feasibility, and flexibility. This approach is consist of majorly 6 steps which involves understanding the business, understanding of Data, further preparing the Data, and then modeling Evaluation and Deployment.

3.1 Dataset Preparation - Preprocessing & Availability

The datasets that have been used for this research is freely and openly available. Since the big conglomerates have been trying to build better models, the company google released multiple datasets for the training of models.

Youtube-8-m dataset: This dataset is probably the largest dataset in the world for the classification purpose. This dataset consist of around 8 million videos and this dataset is partitioned into the different parts which are called as shards. Like in the related work we have seen people have worked on the hardly 3-4 classes but this dataset consists of the 3862 classes which are a huge amount of data. Moreover, every video is approximately 2 minutes long and hence the total size of data is approximately 1.52 TB. The dataset consists of a CSV file as well which will have the video name and the and their class or

label. This file can be loaded into memory to assign the label to the videos and training the model. Similarly, it can be used to test the model with the information present in the CSV file. The Dataset can be found at the link mentioned below.[1]

index	mid	display_name
0	/m/09x0r	Speech
1	/m/05zppz	Male speech, man speaking
2	/m/02zsn	Female speech, woman speaking
3	/m/0ytgt	Child speech, kid speaking
4	/m/01h8n0	Conversation
5	/m/02qldy	Narration, monologue
6	/m/0261r1	Babbling
7	/m/0brhx	Speech synthesizer
8	/m/07p6fty	Shout
9	/m/07q4ntr	Bellow
10	/m/07rwj3x	Whoop
11	/m/07sr1lc	Yell

Figure 3: Sample data - Class

The sample of Data with Class name is demonstrated in figure Figure 3 and with the class code is demonstrated in figure Figure 4

# Segments csv created Sun Mar 5 10:54:25 2017					
# num_ytids=20371	num_segs=20371	num_unique_labels=527	num_positive_labels=51804		
# YTID	start_seconds	end_seconds	positive_labels		
--4gqARaEJE	0	10	"/m/068hy	/m/07q6cd_	/m/0bt9lr
#NAME?	20	30	"/m/03l9g"		
#NAME?	0	10	"/m/01b_21"		
#NAME?	0	9	"/m/04rlf	/m/09x0r	/t/dd00004
-0BIyqJj9ZU	30	40	"/m/07rgt08	/m/07sq110	/t/dd00001"
-0CamVQdP_Y	0	6	"/m/04rlf	/m/07pbtc8	/m/09x0r"
-0GjB-vB1q4	30	40	"/m/0140xf	/m/02cjck	/m/04rlf"
-0RWZT-miFs	420	430	"/m/03v3yw	/m/0k4j"	
-0YUDn-1yII	30	40	"/m/02cjck	/m/04rlf"	
-0jeONf82dE	21	31	"/m/03k3r	/m/04rlf	/m/07q5rw0
-0nqfRcnAYE	370	380	"/m/04brg2"		
-0p7hKXZ1ww	30	40	"/g/122z_qxw	/m/09x0r"	
-0vPFx-wRRI	30	40	"/m/025_jnm	/m/04rlf"	
-0xzrMun0Rs	30	40	"/m/01g90h	/m/04rlf"	
-0yRKSOzyTI	30	40	"/m/07pzfmf	/m/09x0r"	
-116CjQ3MAg	160	170	"/m/015p6	/m/0chx_	/t/dd00128"
-1EXhfqLLwQ	150	160	"/m/03dnzn	/m/068hy	/m/07p7b8y
-1Hub6Ps_cc	10	20	"/m/0130jx	/m/02jz0l	/m/09x0r
-1II0Di9Hkc	30	40	"/m/04rlf	/m/06bz3"	

Figure 4: Sample data - class & code

Audioset- This is another dataset which consists of the Audio files, This audioset is developed by Google too and they have used almost 2 million videos to construct this dataset and this dataset consist of 527 classes. The length of the sound in this audio is around 10 seconds which makes it more feasible to understand the model and train the model. This dataset is also consist of the audio files and CSV file, CSV file again contains the information about the file name and its genre or class. This dataset can be downloaded from the link here [2]

Urban sound Dataset: This dataset is most suitable dataset for the research purpose as this dataset is not that huge like other youtube or audioset dataset as the other dataset requires huge servers with powerful CPU and GPUs to train the model, even then it'll take days to train the model. However, in this Urban sound dataset, the data is consist of 8732 files and we can easily divide the data for test and train purpose. These 8732 files are divided into the 10 classes which makes it preferably easy to work on for the research purpose as this dataset can be handled on a local machine like laptop or desktop. Link to the dataset is here [3]

[1]Dataset 1: `https://research.google.com/youtube8m/index.html`
[2]Dataset 2: `https://research.google.com/audioset/dataset/index.html`
[3]Dataset 3: `https://urbansounddataset.weebly.com/urbansound.html`

Pre-processing: The pre-processing of data is done as there are some problems in the CSV file. So the CSV file is read and the special character and blank columns and columns with unexpected data is removed since there is enough data to train so during pre-processing all the data which is corrupted or missing is removed from the file and the correct data is loaded and similarly the data files i.e. the audio files were also ignored since the class value is NULL there is no point in reading the data for that particular file.

3.2 Modelling Approach - Transfer Learning

The most efficient way to develop models is to use the already built high performing model and tweak some changes or enhance these models by combining the output of two different models, this approach would be the best way. Transfer learning approach suggests from the name that it is the transfer of the learning and hence in this approach an already trained architecture which is popularly known as VGG. This model is designed to train the machine learning algorithm and it is made on the concept of Deep Machine Learning and this model has a lot of variables which are configurable and can be used to train the new model. The best thing about this approach is that with this approach the already trained model is trained again and this would reduce the consumption of computational power significantly. The architecture of VGG looks like in the image shown in Figure

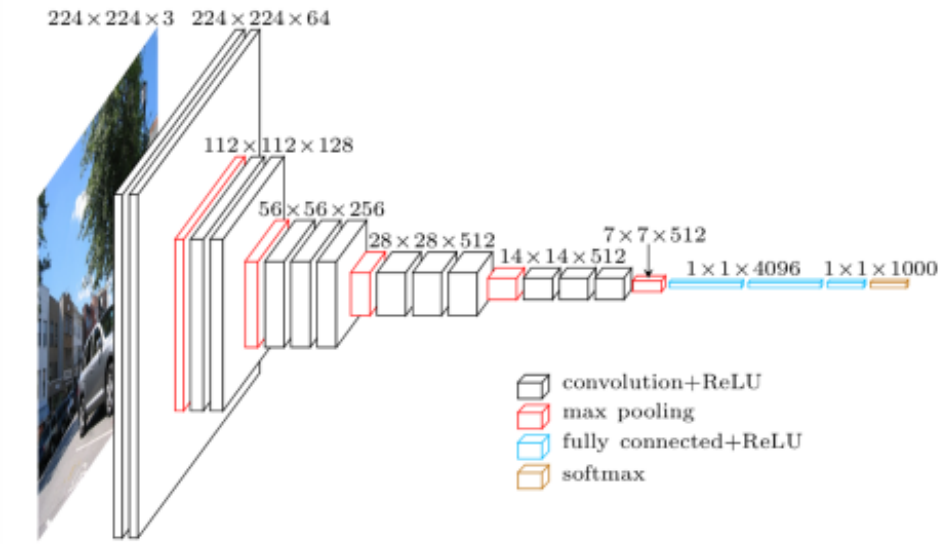

Figure 5: A Visualization of VGG Architecture (Source)

3.3 Data Modeling - Working of Model

The critical part and detail about the working of the model are discussed in this section. In this research, the primary focus is on audio since the audio contains information which can be easily understood and comprehend and also it doesn't get affected a lot by the introduction of the noise. To achieve this, the noise has been introduced to the same audio and then the output was tested and it was seen that the model was still able to perfectly classify the sound to the Dog's sound. This procedure is consist of a model which is entirely based on the architecture of the Convolutions Neural Network

and it works solely on the spectrograms generated from the audio data. Further, this spectrogram is passed through highly sophisticated layers of CNN and this pre-trained model would be able to understand the deep features to assign the related class for that audio spectrogram. The spectrogram is generated using the audio data of urbansound and every single wave is generated using the 960 milliseconds of sound and further, the dimension feature that is generated for this generated spectrogram image is 64 and hence the entire spectrogram image would be around 96*64. Later, this processed image is fed to the already trained model of VGG and the higher dimensionality can be obtained up to 128.

The mathematical expressions can be explained using the research published by (Aytar et al.; 2016) and (Xu and Li; 2003).

let

$$x_i \in R^D$$

be the audio wave and,

$$y_i \in R^{3\times T\times W\times H}$$

be the audio/video and the duration of the video is from 1 to N and T, shows the total samples and H represents height and W shows width.

During the initial part of model i.e training the topmost part of the model optimizing the coefficient is a mandatory task which can be written as

$$min_\theta \sum_{k=1}^{K}\sum_{i=1}^{N} D_{KL}(gk(y_i)\| \ f_k(x_i;\theta)$$

where

$$D_{KL}(P\|Q)=\sum_j P_j \log(Pj/Qj)$$

is the K-L divergence. (Aytar et al.; 2016)

3.3.1 Reducing the Dimensionality - PCA

The higher dimensions in the images makes it difficult to run the model as it requires the higher configuration devices and moreover the higher configuration devices will have to utilize a lot of power to work through so much of data and hence it is mandatory to reduce the dimensionality and PCA is accepted as a great technique worldwide to reduce the dimensionality. The google itself has launched the starter code for the calculation of the matrix in late 2018 and the mathematical equation behind that code is explained by (Xu and Li; 2003) and demonstrated below.

Letś say the maximum value of vectors is N then,

$$x_i, i = 1,2, \ N$$

the value of Covariance matrix can be written as:

$$i = (1/N)\sum_{i=1}^{N} x_i and C = (1/N)\sum_{i=1}^{N}(x_i - i)(x_i - i)^T$$

where C can represent the covariance matrix and i represents the mean vector.

PC stands for the principal components and the first K relatively significant eigenvectors for C V, i=1,2.K, further the eigenmatrix is constructed with d*k dimensions, and that is demonstrated further as K dimensional vector with

$$y = U^T(x - i)$$

where K demonstrates smaller than N and D, K$\ll NandK \ll D$.

This is how PCA helps in reducing the dimensionality and making it feasible to run on the machine with comparatively low configuration and helps in performing faster training of models.

3.4 Gantt Chart

The Figure 6 shows that how the project development timelines were for the Implementation and execution of this Research Project.

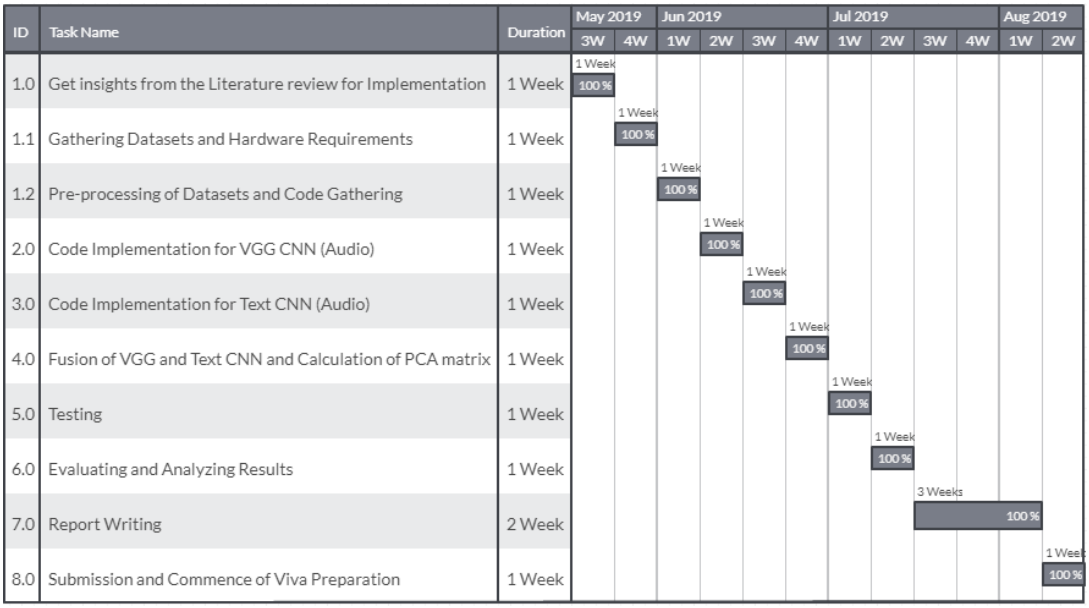

Figure 6: Gantt Chart

4 Implementation

The implementation plays a crucial role in the execution of any project. The implementation of this project is done using python. Python is preferred over other languages because python have an advantage over the other languages in terms of the availability of the library and moreover the library like librosa have functions like create_spectrogram which can directly create the spectrogram from the audio files. Also, the VGG architecture model is available in the python library. Cleaning and loading the data to python and further the running and testing of the model is also easy in python as a lot of tutorials are available for python.

4.1 Technical Details

Python will be used as a language and Spyder is used as the IDE on top of the python and spyder is selected because it is convenient and help in providing the proper indentation for the code.

Dataset, VGG model starter code is provided by Google and it is available on Github to start the building of the model.gc,

The libraries which were used in this project were memory_usage, os, pandas, glob, numpy, keras, librosa, pylab, matplotlib, gc and PIL from the Image library. They need python 3.7 to run all the libraries.

Keras based on the TensorFlow and it can be used to process the audio signals. Librosa was used to capture and display the generated spectrogram from the audio files.

Matplotlib is used for the generation of spectrogram using the short-time Fourier transform.

Numpy is used for certain mathematical calculations and determine the accuracy and mathematical calculation required for the model.

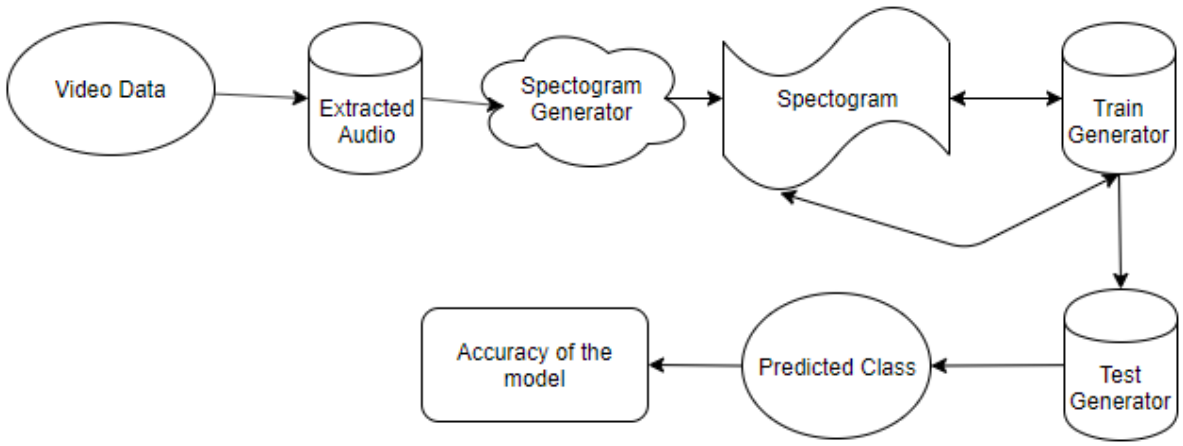

Figure 7: A Visualization of Implementation

The model is developed using the Google VGG with a similar approach and the model is run again and again to improve the accuracy. For the first epoch, the accuracy of the model was around 30% and further, the model is re-run, again and again, to improve the accuracy until the accuracy percentage stopped increasing and it became constant. During the execution of this project, the exact value of epochs when the model accuracy became constant was around 150 which is a really high number. Hence to ease the project the urbansound dataset is used as the Youtube8m and Audioset dataset was very huge and model ran for over 2 days without giving significant results. Further, keeping this aspect in mind and to perform the analysis the data is switched to just urbansound dataset. approximately 8000 labeled files were present for analysis and these were divided into training and testing dataset. Around 5500 files were used with all the 10 mixed classes and rest 2600 files were used for the testing purpose. The model took around 15 hours for training and after 150 epochs the model achieved an approximate of 94.6% of accuracy. The flow of the implementation is depicted in figure Figure 9

4.2 Inducing Noise in the Data

Further, the noise is induced in the audio data manually. The barking of a dog is taken into consideration and the noise is induced the same file and changes in results were noticed. Even after adding the noise the model was still able to successfully identify the class of the Audio. Below spectrogram of audio where the noise is induced and when noise is not induced.

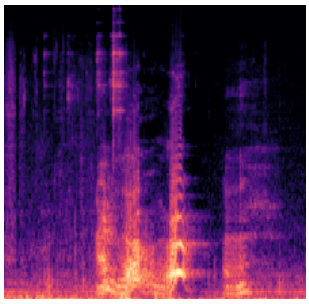

Figure 8: Dog Barking with No Noise

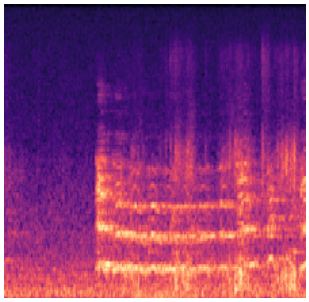

Figure 9: Dog Barking with Added Noise

Hence this model can be applied in real noise as it is able to counter the noise generated in the model and it was still able to understand the class in the audio.

5 Evaluation

The evaluation of a model is a mandatory part as it tells us about the performances and the drawbacks of the models. The performance of this model is measured on majorly two aspects one is that how the model is reacting to the number of runs and how many runs will be needed to make the model stable with the accuracy. The second aspect is how the time and processing power took to train the model increases or decreases with the induction of noise in the Training and Testing datasets.

5.1 Comparison of Loss, Accuracy, and Epochs

To evaluate this aspect the loss, acc, val_loss, val_acc is measured along with the number of the epoch. As per the keras, this is the only way to evaluate the model and check

the accuracy of the model along with train the model until it becomes a fit model and prevents the model to become over-fit.

Epochs	loss	acc	val_loss	val_acc
1	0.7546	0.7509	0.6563	0.7866
2	0.6459	0.7877	0.5994	0.8051
3	0.5764	0.8055	0.5665	0.8175
4	0.4906	0.8316	0.7519	0.7881
5	0.4582	0.8466	0.4758	0.8439
6	0.4097	0.8625	0.7281	0.7707
7	0.3830	0.8717	0.4593	0.8612
8	0.3440	0.8817	0.4878	0.8537
9	0.3067	0.8990	0.5290	0.8462
10	0.2743	0.9081	0.3832	0.8786
25	0.1307	0.9624	0.3237	0.9253
35	0.1101	0.9660	0.3133	0.9299
45	0.1035	0.9742	0.3511	0.9321
55	0.0959	0.9729	0.3402	0.9261
65	0.0929	0.9722	0.3363	0.9389
150	0.0734	0.9974	0.3065	0.9487
151	0.0711	0.9975	0.3011	0.9479
152	0.0686	0.9982	0.2988	0.9488

As per the keras theory, we should keep training the model until val_acc is, even if acc is still increasing as if the model is being run again and again the model will most likely to become an overfit model. The models give a pretty good accuracy around 50 epochs but to make it extremely good the model can be run again until 150 epochs but after 150 epochs the model starts to become overfit as acc is still increasing, however, val_acc has stopped increasing.

5.2 Analysis of Impact of Noise Induction in Audio

In real life scenario, we never receive a file which does not contain any noise or distortion and hence it is imperative that the model should be able to recognize the noises or distortions and it should be able to remove it or recognize it. The model should be able to bifurcate between the information stored in the file with respect to the noise that has been mixed.

Also, another important aspect is that the model should be able to perform well with the noise and should be able to predict the class. Moreover, along with recognizing the class the model should be able to perform swiftly, it should never be the case that model will fail or will start taking a lot of time when compared to isolated classes audio. Hence it is mandatory to run and analyze the performance of the model when induced noise is brought into the data.

Figure Figure 10 shows how the processing time gets impacted with the induction of noise and this is measured against a number of epochs to generate a graphical representation.

On the other hand, Figure Figure 11 shows that how the processing power gets impacted with the induction of noise and this is also measured against a number of epochs to generate a graphical representation.

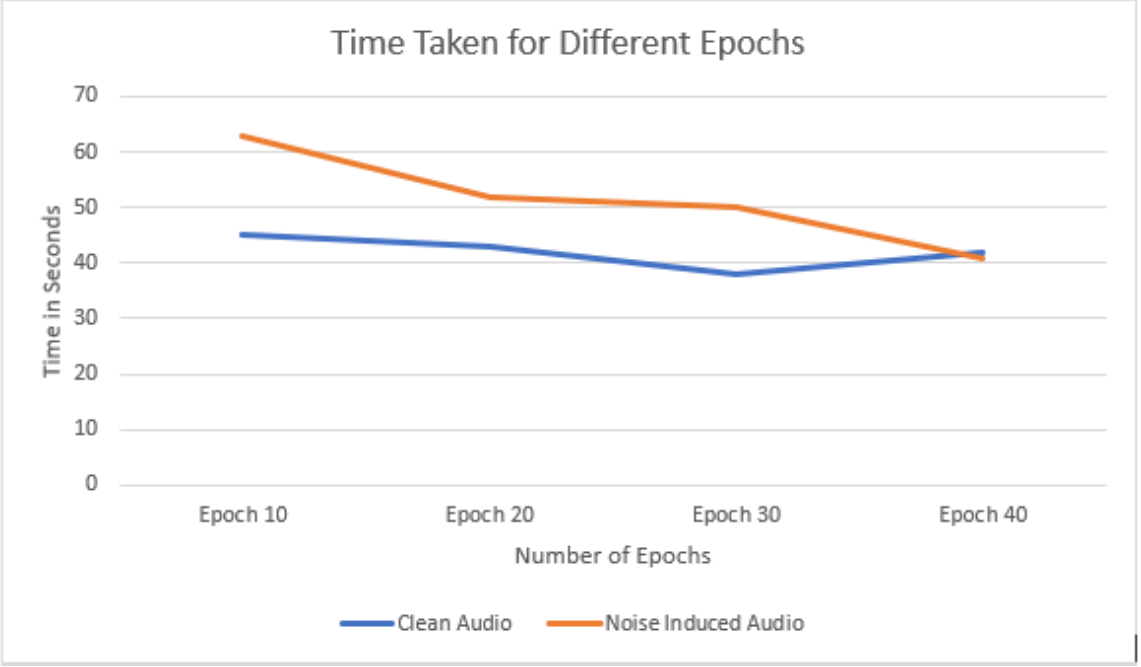

Figure 10: Impact on Processing Time after Inducing the Noise

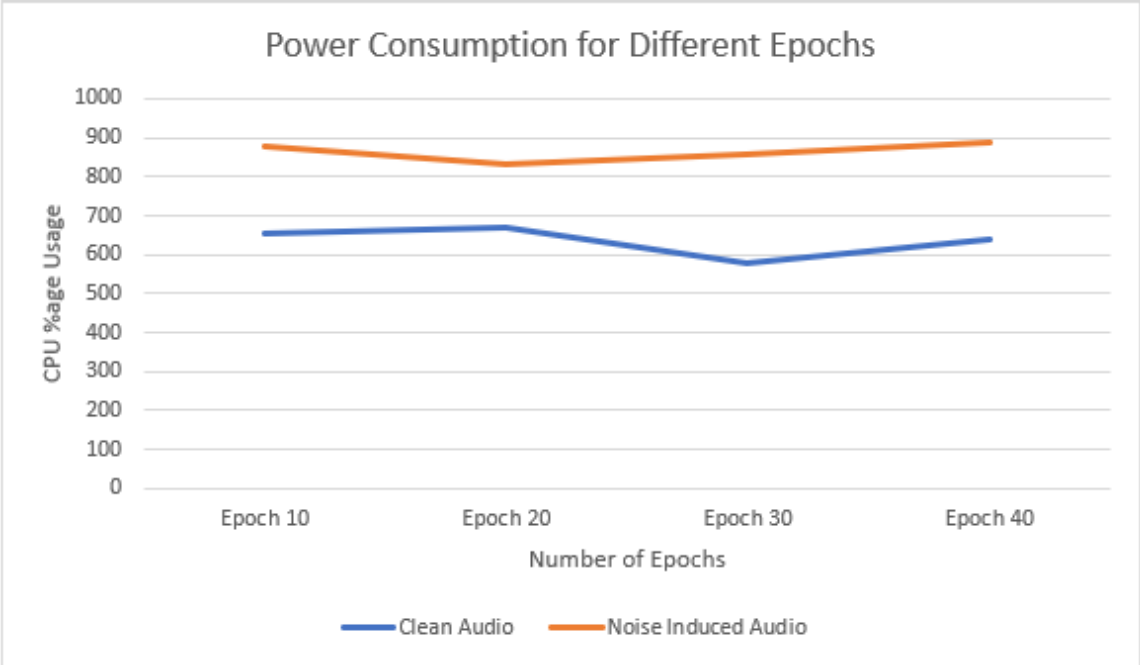

Figure 11: Impact on Processing Power after Inducing the Noise

We can see from the results that the model has ana accuracy of 94.88% and this model also doesn't get highly impacted with the induction of the background noise. Though the performance does get hampered with the induction of noise it does not affect majorly and hence this change in processing time and power is acceptable.

6 Conclusion and Future Work

This research tries to build a model with good accuracy and a state-of-the-art solution to prevent the model from becoming overfit or underfit. Also, during this research, it has been kept in mind to reduce the computational power and along with that, there shouldn't be a steep decline in the accuracy of the model. The accuracy of this model is 94.88% which is comparatively low with respect to some models like (Hershey et al.; 2016). However, this model requires comparatively very less computation power when compared to other models as a majority of the model works on the picture pixel analysis and this model works on the audio present the videos and hence require comparatively less computational power.

Apart from training testing the model with the valid data the noise has also been introduced to the model to measure the extent of change or degradation in the performance of the model. After introducing the noise the model is compared with the previous results in terms of the computational power consumption and time taken to train in one epoch that is the performance of the model for every epoch is compared.

The results clearly depicts that the model is trained to a great percentage at around 50 epoch, so if someone does not want to waste a lot of computational resources and can work with a 2-3% more error can run the model after 50epoch, however if someone requires a model which should perform better with a low error rate that person could run the model for 150 epoch to gain better accuracy but this will impact the computational power. Also, after 150 epochs the model starts to become over-fit and hence it should be never trained after the 150 epochs. Also, the graphical results show that the model works fine even after induction of the background noise however the performance is decreased and the time taken for processing is increased which also does not have a major impact on the time and performance. Hence this model can work in the real-life scenario for the prediction of the genre or class of the Videos or Audios.

In terms of Future development, better models can be developed by introducing more modalities and combining the output of the modalities to get better results. Also, more data and classes can be fed to the model to further expand the horizon of the classes. Apart from expanding the classes, if the percentage bifurcation is introduce to this model that would be a major improvement like let us assume in an audio there is the voice of dog and cat at the same time so if the model would be able to bifurcate and give a percentage like the audio contains 60% dog sound and 40% cat sound that would improve the efficiency and clarity in the model drastically. So by introducing more modalities and getting percentages for different voices the output of the model can be improved and a better model can be built.

References

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B. and Vijayanarasimhan, S. (2016). Youtube-8m: A large-scale video classification benchmark,

CoRR **abs/1609.08675**.
URL: *http://arxiv.org/abs/1609.08675*

Agnihotri, L. and Dimitrova, N. (1999). Text detection for video analysis, *Proceedings IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL'99)*, pp. 109–113.

Aytar, Y., Vondrick, C. and Torralba, A. (2016). Soundnet: Learning sound representations from unlabeled video, *CoRR* **abs/1610.09001**.
URL: *http://arxiv.org/abs/1610.09001*

Brezeale, D. and Cook, D. J. (2008). Automatic video classification: A survey of the literature, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **38**(3): 416–430.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection, *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1, pp. 886–893 vol. 1.

Dimitrova, N. and Agnihotri, L. (2000). Video classification based on hmm using text and faces, *European Signal Processing Conference* **2015**.

Gibert, X., Li, H. and Doermann, D. (2003). Sports video classification using hmms, pp. II– 345.

Girgensohn, A. and Foote, J. (1999). Video classification using transform coefficients, *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, Vol. 6, pp. 3045–3048 vol.6.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J. and Wilson, K. W. (2016). CNN architectures for large-scale audio classification, *CoRR* **abs/1609.09430**.
URL: *http://arxiv.org/abs/1609.09430*

Jain, M., Jgou, H. and Bouthemy, P. (2013). Better exploiting motion for better action recognition, *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2555–2562.

Jiang, Y.-G., Dai, Q., Xue, X., Liu, W. and Ngo, C.-W. (2012). Trajectory-based modeling of human actions with motion reference points, pp. 425–438.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks, *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1725–1732.

Krizhevsky, A., Sutskever, I. and E. Hinton, G. (2012). Imagenet classification with deep convolutional neural networks, *Neural Information Processing Systems* **25**.

Laptev, I. (2005). On space-time interest points, *Int. J. Comput. Vision* **64**(2-3): 107–123.
URL: *http://dx.doi.org/10.1007/s11263-005-1838-7*

Laptev, I., Marszalek, M., Schmid, C. and Rozenfeld, B. (2008). Learning realistic human actions from movies, *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8.

Lee, H., Pham, P., Largman, Y. and Ng, A. Y. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks, *in* Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams and A. Culotta (eds), *Advances in Neural Information Processing Systems 22*, Curran Associates, Inc., pp. 1096–1104.
URL: *http://papers.nips.cc/paper/3674-unsupervised-feature-learning-for-audio-classification-using-convolutional-deep-belief-networks.pdf*

Li, F., Gan, C., Liu, X., Bian, Y., Long, X., Li, Y., Li, Z., Zhou, J. and Wen, S. (2017). Temporal modeling approaches for large-scale youtube-8m video understanding, *CoRR* **abs/1707.04555**.
URL: *http://arxiv.org/abs/1707.04555*

Li, S. and Guo, G. (2000). Content-based audio classification and retrieval using svm learning.

Lin, W.-h. and Hauptmann, A. (2002). News video classification using svm-based multimodal classifiers and combination strategies.

Lin, Z., Jiang, Z. and Davis, L. (2009). Recognizing actions by shape-motion prototype trees, pp. 444 – 451.

Lu, L., Jiang, H. and Zhang, H. (2001). A robust audio classification and segmentation method, p. 203.

Na, S., Yu, Y., Lee, S., Kim, J. and Kim, G. (2017). Encoding video and label priors for multi-label video classification on youtube-8m dataset, *CoRR* **abs/1706.07960**.
URL: *http://arxiv.org/abs/1706.07960*

Roach, M. J., Mason, J. D. and Pawlewski, M. (2001). Video genre classification using dynamics, *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, Vol. 3, pp. 1557–1560 vol.3.

Takahashi, N., Gygli, M. and Gool, L. V. (2017). Aenet: Learning deep audio features for video analysis, *CoRR* **abs/1701.00599**.
URL: *http://arxiv.org/abs/1701.00599*

Wang, H., Klser, A., Schmid, C. and Liu, C. (2011). Action recognition by dense trajectories, *CVPR 2011*, pp. 3169–3176.

Wang, H., Klser, A., Schmid, C. and Liu, C.-L. (2013). Dense trajectories and motion boundary descriptors for action recognition, *International Journal of Computer Vision* **103**.

Wang, H. and Schmid, C. (2013). Action recognition with improved trajectories, *2013 IEEE International Conference on Computer Vision*, pp. 3551–3558.

Wang, H., Ullah, M., Klser, A., Laptev, I. and Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition.

Wang, H., Zhang, T. and Wu, J. (2017). The monkeytyping solution to the youtube-8m video understanding challenge, *CoRR* **abs/1706.05150**.
URL: *http://arxiv.org/abs/1706.05150*

Wei, G. and Sethi, I. K. (1999). Face detection for image annotation, *Pattern Recognition Letters* **20**(11-13): 1313–1321.

Willems, G., Tuytelaars, T. and Van Gool, L. (2008). An efficient dense and scale-invariant spatio-temporal interest point detector, pp. 650–663.

Wu, Z., Wang, X., Jiang, Y., Ye, H. and Xue, X. (2015). Modeling spatial-temporal clues in a hybrid deep learning framework for video classification, *CoRR* **abs/1504.01561**.
URL: *http://arxiv.org/abs/1504.01561*

Xu, L.-q. and Li, Y. (2003). Video classification using spatial-temporal features and pca, pp. III– 485.

Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R. and Toderici, G. (2015). Beyond short snippets: Deep networks for video classification, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zha, S., Luisier, F., Andrews, W., Srivastava, N. and Salakhutdinov, R. (2015). Exploiting image-trained CNN architectures for unconstrained video classification, *CoRR* **abs/1503.04144**.
URL: *http://arxiv.org/abs/1503.04144*