CONTENT-BASED REPRESENTATIONS OF AUDIO USING SIAMESE NEURAL NETWORKS

Pranay Manocha*, Rohan Badlani, Anurag Kumar, Ankit Shah, Benjamin Elizalde, Bhiksha Raj

Motivation

- Need effective ways to browse by content through audio databases of growing sizes
- Relate two audio based on their semantic content

Problem

-To create a Query by example retrieval system

- So how do we encode semantic content of audio?

Our Approach

Train a Siamese Neural Network which encode the audio-class specific information in a vector representation.

Challenges

-Searching audio events is hard because the video metadata focus mainly on images.

-The metadata does not guarantee the presence of an audio event.

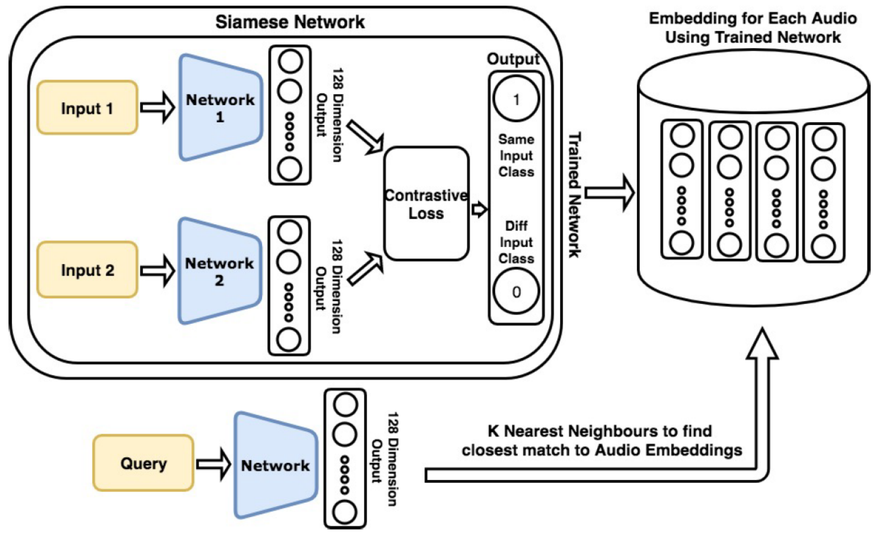
-The audio of user-generated videos , even within the same class, is unstructured and highly variable

Fingerprinting and Similarity Matching

-Fingerprinting is useful in finding the exact match i.e. for finding multiple videos of the same event.

-Our aim is similarity matching -> to retrieve all semantically similar files together

Approach and Model



Loss Function

$$L(W,Y,X_1,X_2) = \frac{(Y)\,(D_W{}^2)}{2} + \frac{(1-Y)max\{0,(m-D_W)\}^2}{2}$$

Custom Dataset Information

-We consider list of sound events from 3 databases- ESC-50, US8K and TUT 2016. Overall, we consider 76 sound classes

-We work on audio recordings from YouTube. For each of 76 classes, obtain 100 recordings from YouTube.

Similarity Measures

- Cosine Similarity
- Euclidean Distance

Training

- Balanced **(N_B)**
- Unbalanced Train Sets **(N_U)**

Metrics

- Mean Average Precision across all queries (**MAP**)
- Precision at 1 (**MP_1**)
- Precision of Top K Retrieval (**MP_K**)

Results

We obtain the best results with unbalanced training and Euclidean distance as the distance measure.

	Euclidian Distance		Cosine Similarity	
Measures	N_B	N_U	N_B	N_U
MAP	*0.0241*	*0.0342*	*0.0186*	*0.0133*
MP_1	*0.314*	*0.436*	*0.132*	*0.333*
MP_{25}	*0.099*	*0.177*	*0.105*	*0.133*

Conclusions

- Euclidian Distance performs better than Cosine Distance
- Took K=25 as we found the highest value of MP_K at K=25.
- N_U performs better than N_B as it is able to learn the discriminative features better.
- MP_1 values are fairly high, meaning that the first positive class hit is obtained at Rank 3 (for 0.436).

Audio Class	MP_{25}
Wind Blowing	0.784
Sheep	0.753
Pig	0.724
Water Drop	0.711
Clock Tick	0.708
Brushing Teeth	0.699

- Multiplying MP_K by K would give us the average number of correct class matches in top K retrievals. This value is nearly 20 for class 'wind blowing', meaning that 20 out of the 25 retrieved samples were of the correct class.

*Contact Author- pranaymnch@gmail.com