Audio Based Multimedia Event Classification with Convolutional Neural Networks and Transfer Learning

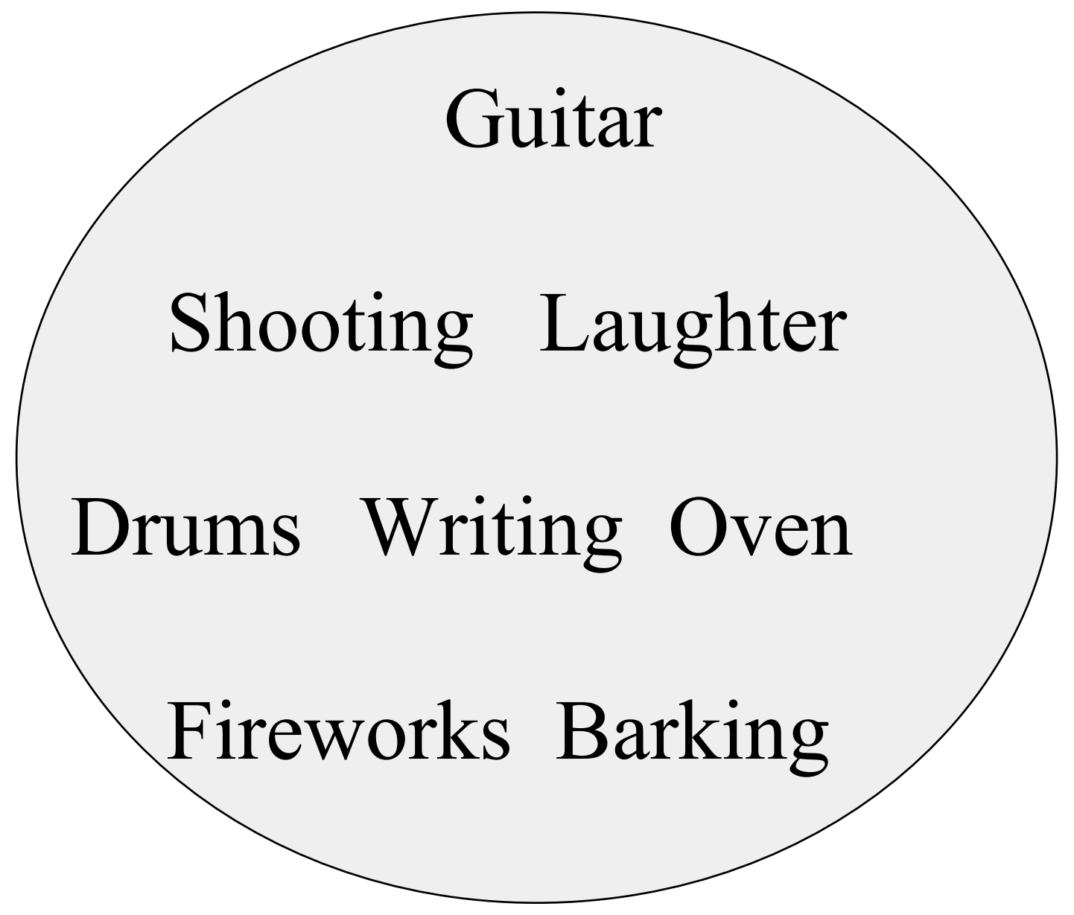
Soham Kelkar, Nishant Gurunath, Kevin Chon
Carnegie Mellon University

Problem

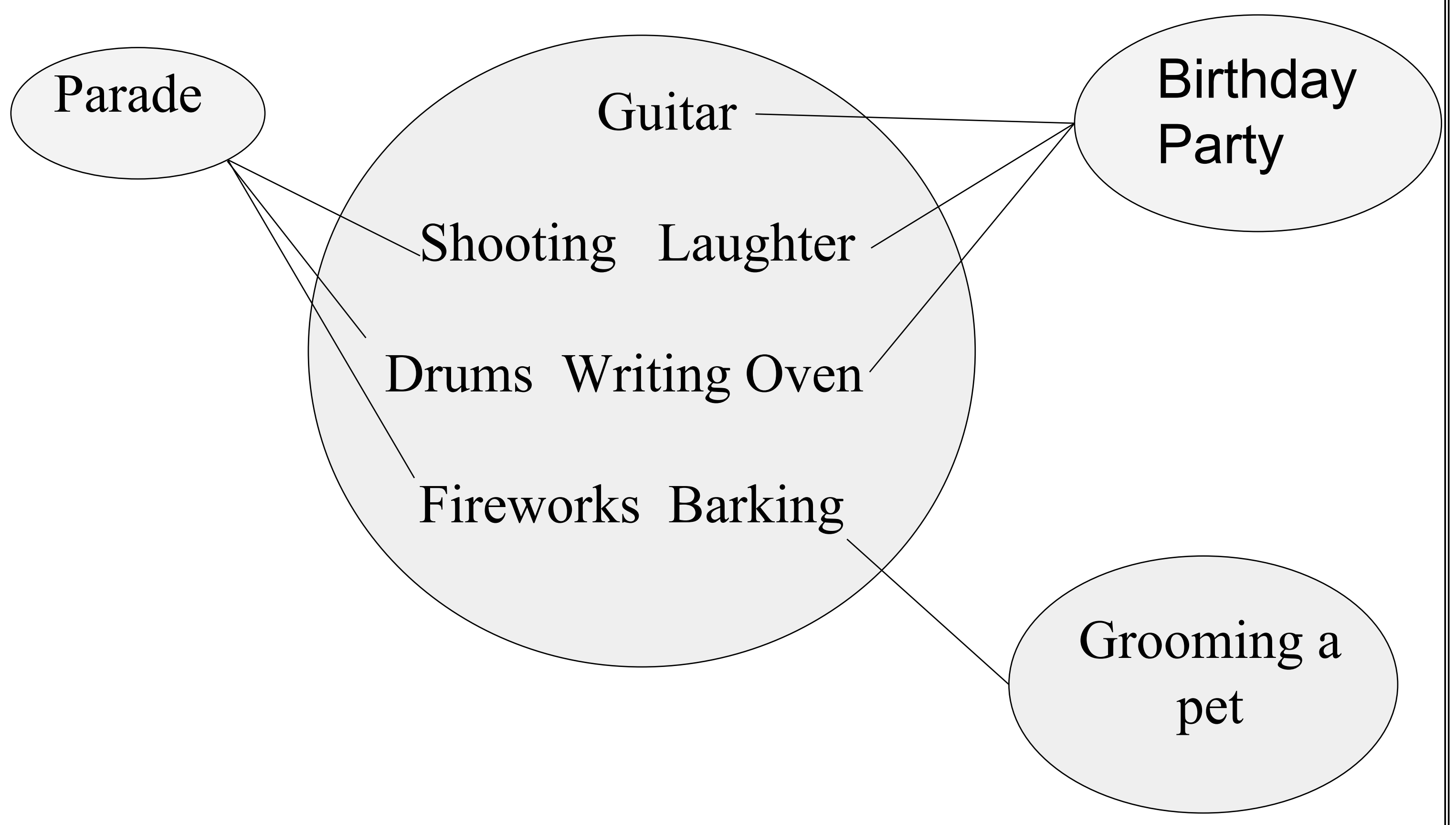
- Classification of YouTube Videos into events only based on its audio content.
- Due to real-world noisy conditions, a lot of unwanted features are learned and classification becomes random.

Solution

- Training the model to learn basic audio features from videos on a different dataset.



- After this we then train our model to classify the audio events based on these the basic features which are inherited by respective events.



Previous Work

- Dense Networks
- Convolutional Architectures

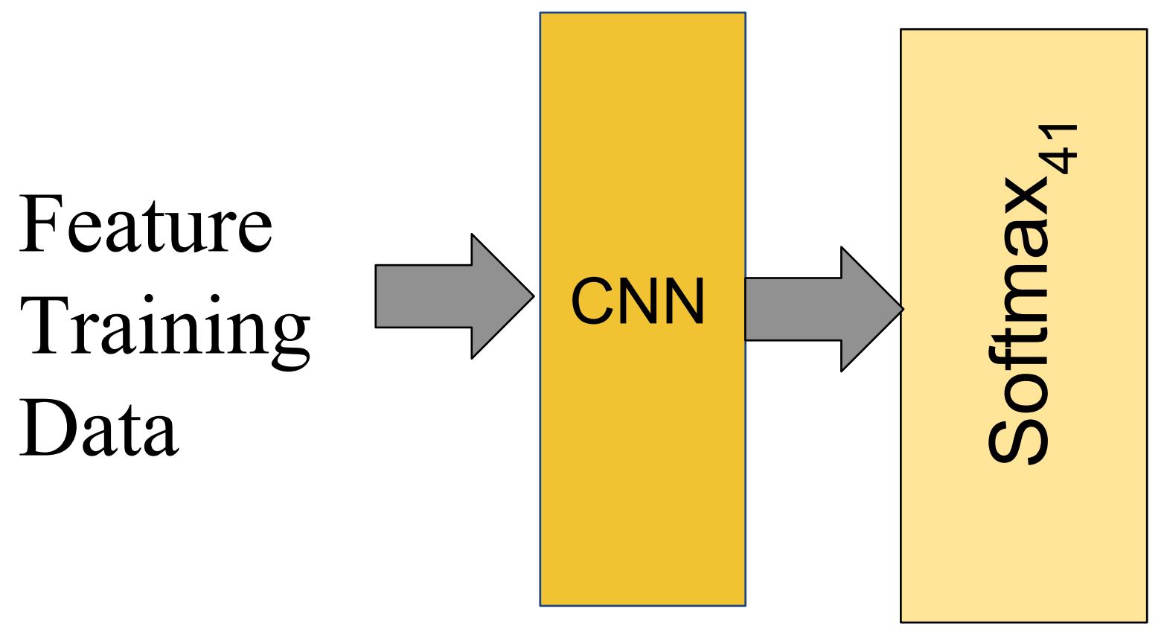
Number to beat: Accuracy = 37.4 %

Methods

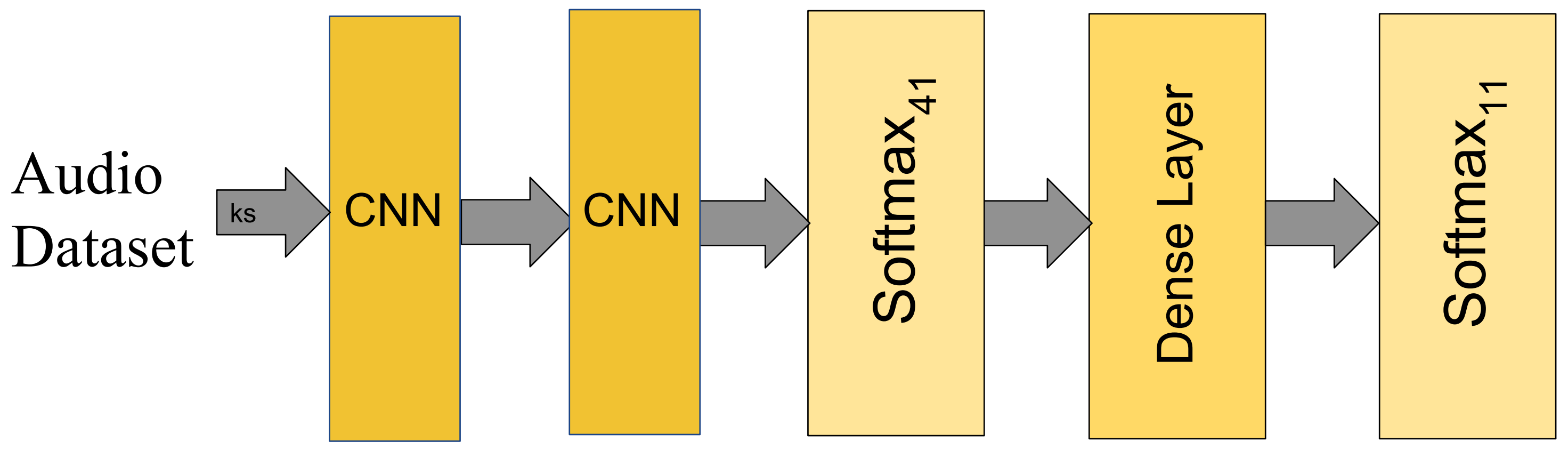
1. ResNet with Boosting

Layer	Type	Channels
1	Conv2D (3x3)	32
2	ResNet Block (3x3)	32
3	Conv2D (3x3)	64
4	ResNet Block (3x3)	64
5	Conv2D (3x3)	128
6	ResNet Block (3x3)	128
7	Avg Pool	-

2. Transfer Learning



Training the CNN with Basic Features



Training on the Event Dataset

Further Work

- Try other datasets so that the model learns more features relevant to the events.
- Try Recurrent Architectures which have proven to be more effective with speech recognition problems

Datasets

1. Feature Training: Freesound General Purpose Audioset

Training Samples	Test Samples	Events
5609	40879	11

2. Multimedia Training Data: YLI-MED

Training Samples	Number of basic features
9740	41

Results

Event	Accuracy	mAP
Birthday Party	76.06	0.8
Flash Mob	15.29	0.3
Getting a Vehicle Unstuck	12.31	0.27
Parade	38.81	0.39
Person attempting a board trick	11.61	0.2
Person grooming an animal	14.29	0.21
Person feeding an animal	21.19	0.24
Person landing a fish	20.63	0.26
Wedding Ceremony	32.46	0.33
Woodworking Project	16.13	0.32

Current Test Accuracy - 31.6%

References

•Khalid Ashraf, Benjamin Elizalde, Forrest Iandola1, Matthew Moskewicz1, Julia Bernd2, Gerald Friedland2, Kurt Keutzer1. *Audio-Based Multimedia Event Detection with DNNs and Sparse Sampling*

•Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, Kevin Wilson, Google, Inc., New York, NY, and Mountain View, CA, USA. *CNN Architectures for Large-Scale Audio Classification*

•Rohan Badlani, Ankit Shah, Benjamin Elizalde, Anurag Kumar, Bhiksha Raj. *FRAMEWORK FOR EVALUATION OF SOUND EVENT DETECTION IN WEB VIDEOS.*