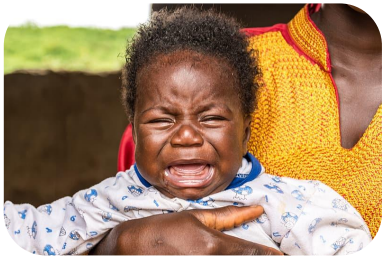
Audio Event Recognition

Alessandro Goller, Andrea Favia, Irene Buffa

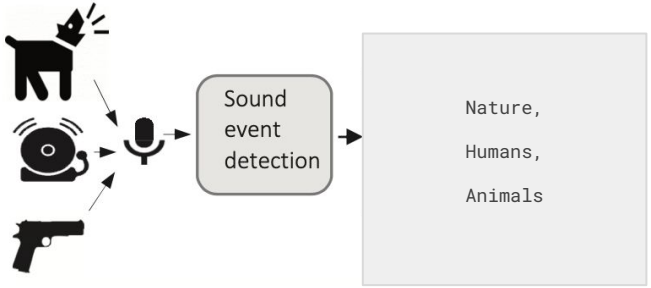
Why Audio Events?



- It is possible to distinguish different sources within an audio clip:
  - Human speech
  - Natural phenomena
  - Animal noise

- Audio Event Detection (AED) is becoming popular because it revolutionizes:
  - Surveillance
  - Security (e.g. gunshot)
  - Monitoring (e.g. baby crying)
  - Speech recognition

OUR SOLUTION



- Google AudioSet as training dataset
- Extraction of audio features
- Training of supervised model with constrained resources in terms of memory.
  - SVM
  - RNN
  - TRANSFORMERS
- Classify sound in an audio clip originated by:
  - Human speech
  - Animals
  - Natural phenomena

Google Audio Set

- A large-scale dataset of manually annotated audio events
- Collection of **+2M** human-labeled **10-second sound clips** drawn from YouTube videos
- There are about **5.8K hours** of audio and **527 classes** of annotated sound

AudioSet

Our training data

- We aim to identify **three** major groups by using the ontology offered:
    - **Humans**: speech, shout, laughter
    - **Animals**: domestic animals, livestock, wild animals
    - **Natural phenomena**: wind , thunderstorm, water, fire
- **1200** audio files per class: 60-20-20 train-val-test
- Data preparation:
    1. Download youtube video
    2. Extract 10-seconds audio clips
    3. Extract features from each clip and save on Mega
    4. Remove video & audio files

	url	start_time	end_time	class_label
5569	eLmWZL2_r80	30.0	40.0	Humans
11420	lK2fHm0V04o	410.0	420.0	Natural
9246	XsZ9ndrYPxI	0.0	10.0	Humans
6712	ehmd1gJM890	10.0	20.0	Humans
6348	ZUgJXeToltg	200.0	210.0	Humans
10565	1zoYJVhzDGc	30.0	40.0	Natural
336	R3C1n611idY	90.0	100.0	Animal
1278	QkpNIF8xzEE	370.0	380.0	Animal
10623	01pUDNKXK9c	220.0	230.0	Natural
6706	lcwMqeA7fT0	30.0	40.0	Humans

Audio Features

- Set of features extracted from the input audio signal, where each feature represents a vector element in the feature space

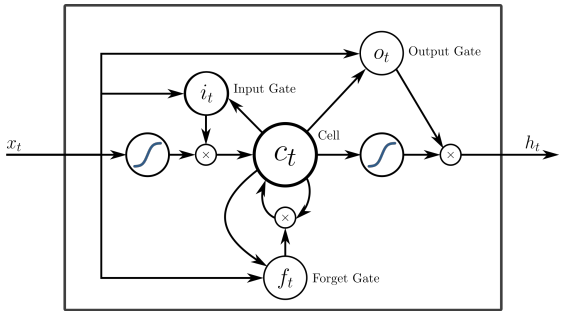
- Audio features' characteristics are:
  - easy adaptable
  - robust against noise
  - easy to implement

- **Log Mel scaled energies**: presentation of the short-term power spectrum of a sound
- **Mel-frequency cepstral coefficients** (MFCCs): cepstral representation of audio signals

Models: Recurrent Neural Networks (**RNNs**)

- Family of Neural Networks for processing sequential or temporal data.

- Advanced architectures for AED:
  - Bidirectional Long-Short Term Memory (**BLSTM**)
  - Bidirectional Gated Recurrent Unit **(BGRU)**

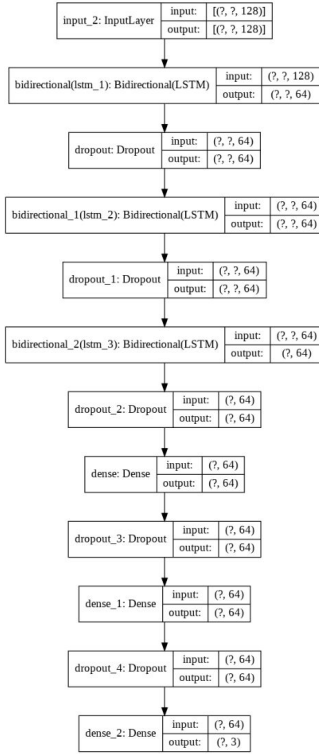


BLSTM - BGRU: Configuration

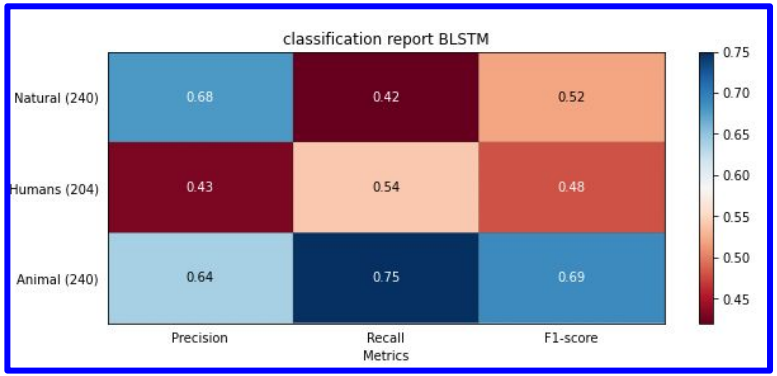
- **Encoder**
  - Three-layers Bidirectional LSTM / GRU
  - 32 units per layer (**activation** = Tanh)
- **Decoder**
  - 2-layers MLP
  - 64 units per layer (**activation** = ReLU)

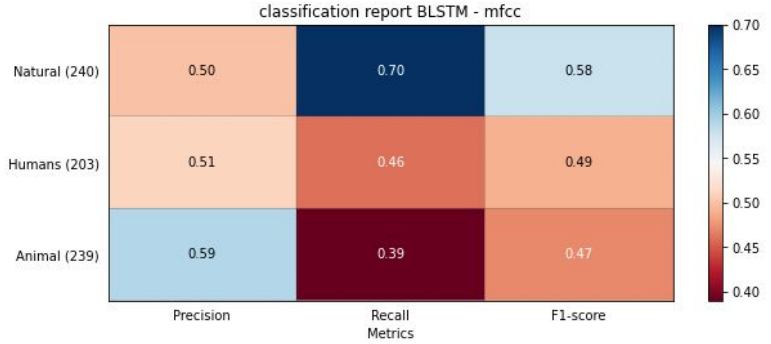
- **Regularization:**
  - Dropout

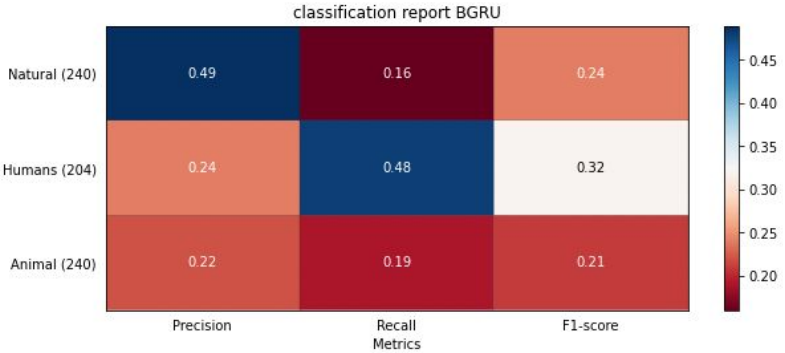
- **Loss:** categorical cross entropy
- **Optimizer**: Adam with Nesterov momentum
  (epochs = 200, learning rate = 0.01, batch_size = 64)

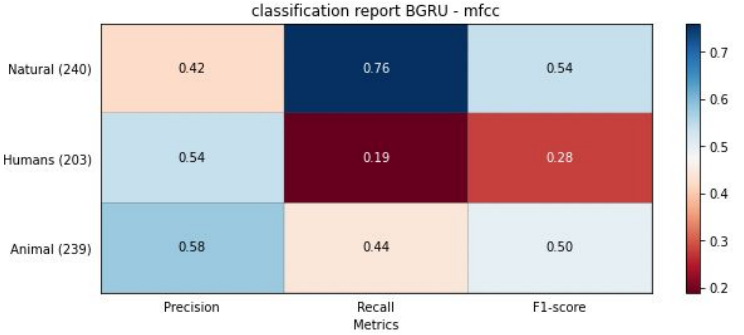


BLSTM - BGRU: Results

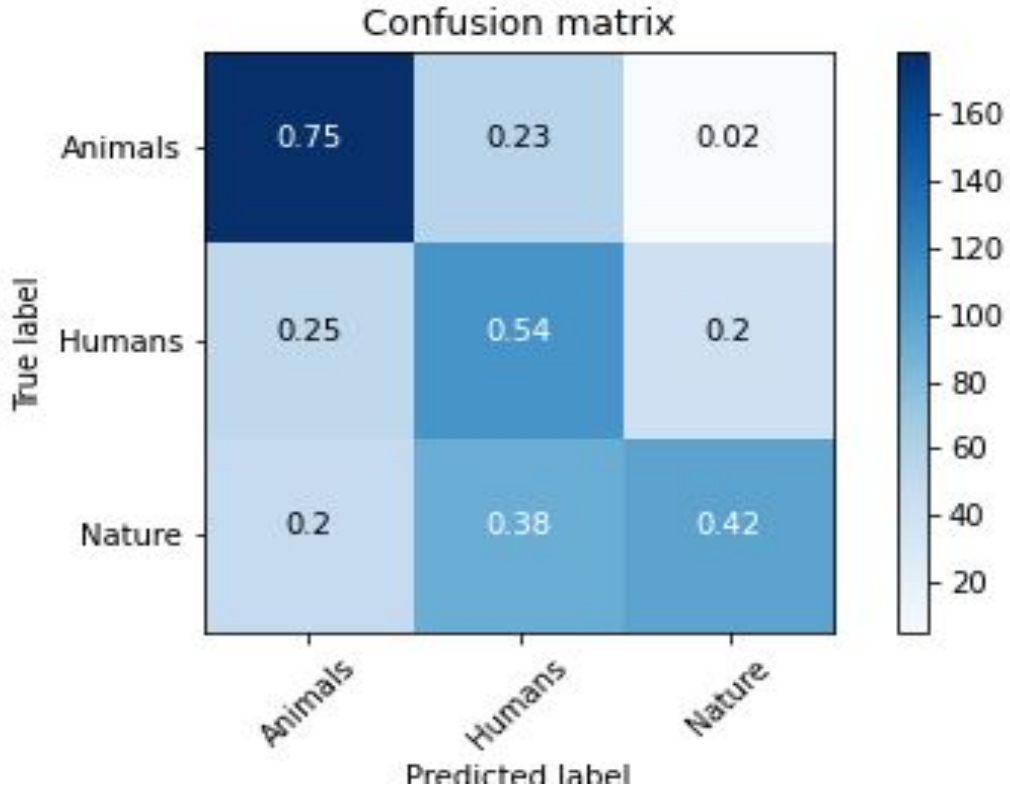








BLSTM with Mel-Spec energies



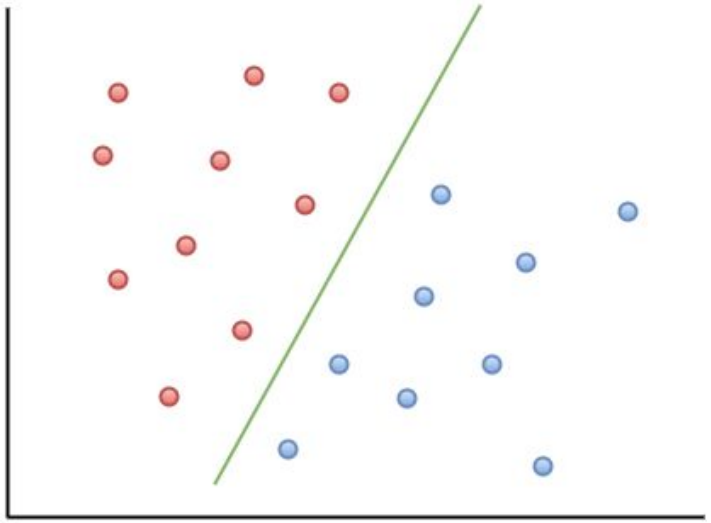
- F-score = **0.57**

Models: Support Vector Machines **(SVM)**

- Separate classes through hyperplane
- Maximizes the margin distance

SVM

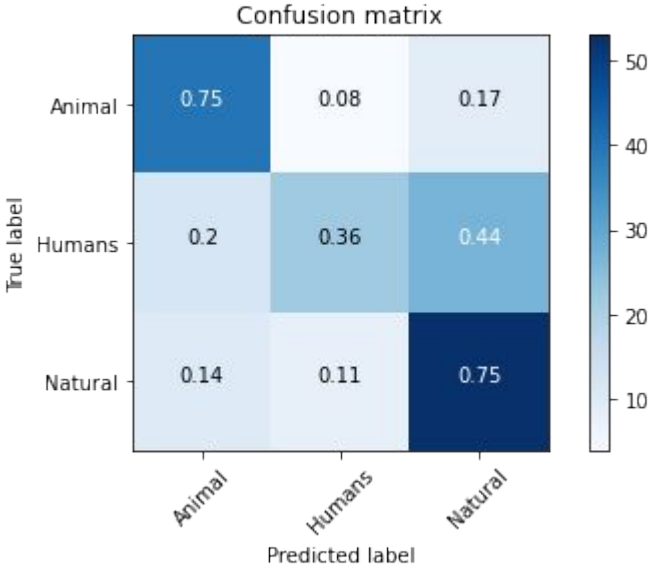
- Radial Basis Function (RBF) kernel
- One-VS-One strategy



SVM: Configuration & Results

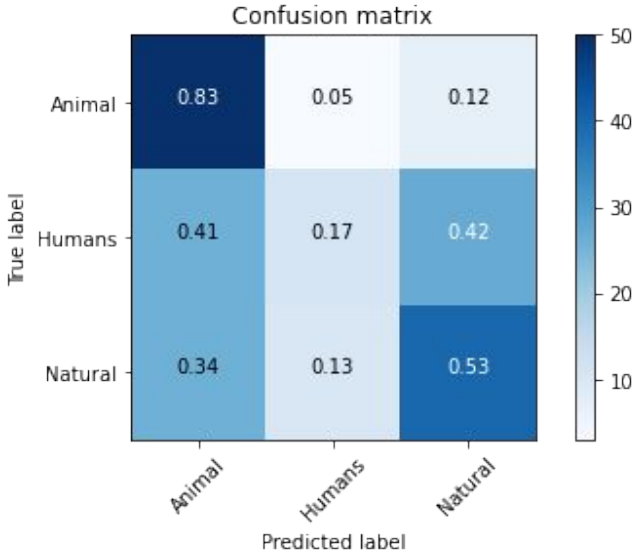
MFCC

- Accuracy=**0.62**
- F-score=**0.61**
- ROC AUC = 0.786



MEL spectrogram

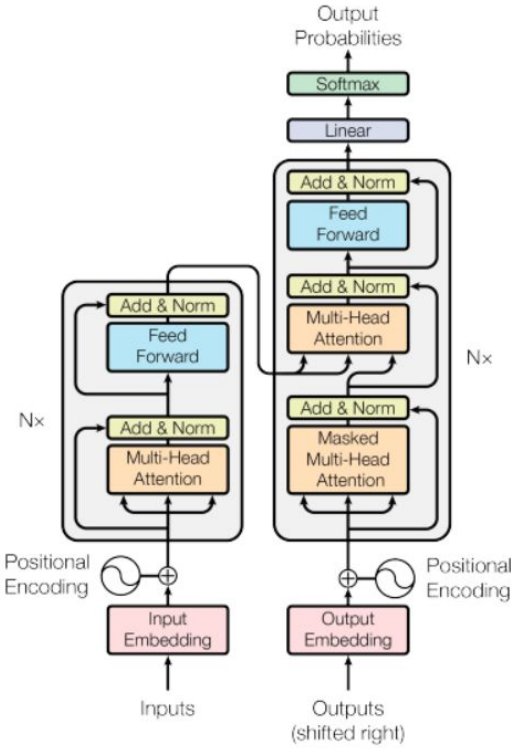
- Accuracy=0.51
- F-score=0.47
- ROC AUC = 0.729



Models: Transformers

Only Encoder

- Input (Padding Mask)
  - Projection
- Positional Encoder
- Encoding Layer
  - Multi Head Attention (4 Heads)
    - Padding mask
    - Scaled dot product attention
  - 1 Dense Layer (activation = ReLU)
  - 2 Dense Layer
  - Dropout (rate = 0.1)
  - Layer Normalization



Transformers: Configuration & Results

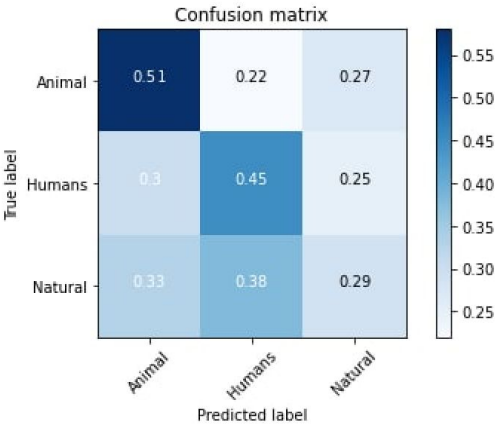
Epoch → 10

Projection → Linear

Number Layers → 2

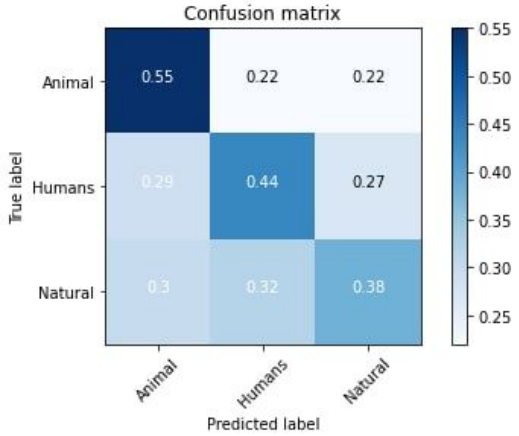
MFCC

- Accuracy = 0.41
- F-score = 0.41



MEL Spectrogram

- Accuracy = 0.46
- F-score = 0.45



Discussion and Conclusion

- Deep Neural Networks require huge amount of data
- Data augmentation may help but we also have small RAM memory
- Training requires performant GPUs
- human sounds is the most unclassified class
- RNNs capture time-based information but we probably need CNN to extract frequency-based features as well (CNN-RNN)
- Simple models like SVM reach good accuracy with constrained resources

Question?