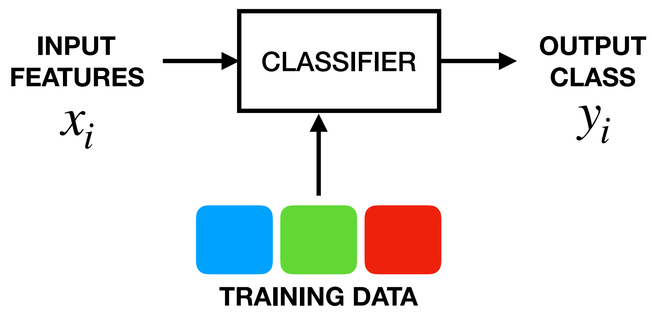
# Deep Embeddings for Rare Audio Event Detection with Imbalanced Data

IIT Kanpur

*Vipul Arora, Ming Sun and Chao Wang*

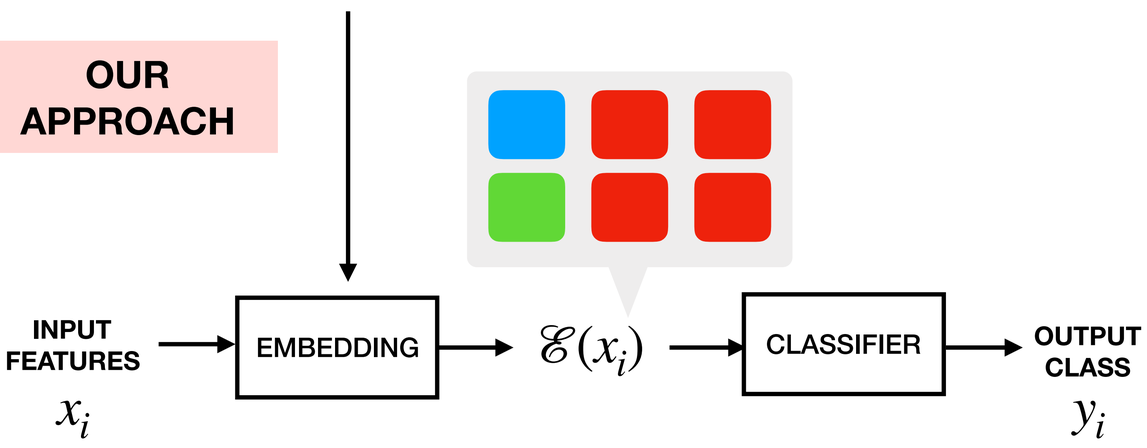
amazon alexa

## SUPERVISED LEARNING OF CLASSIFIERS




## WHAT IF THE DATA IS IMBALANCED OVER DIFFERENT CLASSES?

## OUR APPROACH

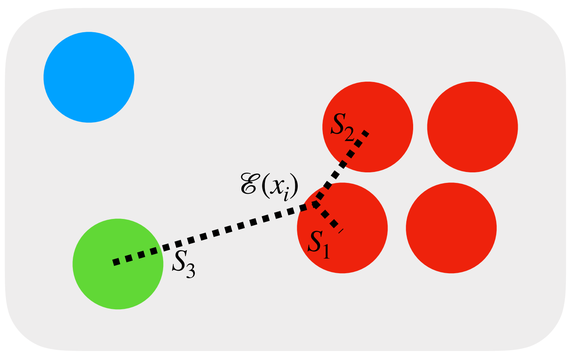


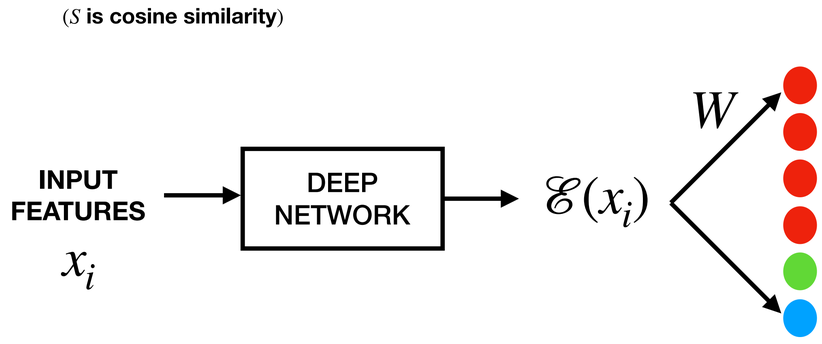

## EMBEDDING LEARNING

LEARN THE EMBEDDING SUCH THAT

$$S_1 > S_2 > S_3$$

($S$ is cosine similarity)




**If $W^\intercal$ and $\mathscr{E}(x_i)$ are L2 normalized, then $h_i = S$**

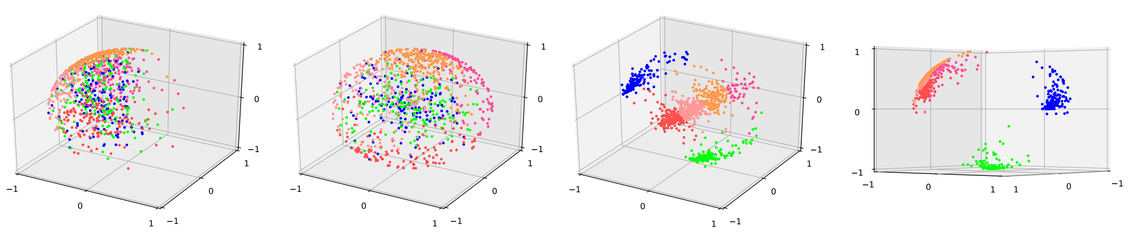
$$v_i[k_1] = \frac{e^{\alpha S_1}}{e^{\alpha S_1} + e^{\alpha S_2 - \beta'} + e^{\alpha S_3 - \beta''}}$$

($k_1$ is the desired cluster for sample $i$)

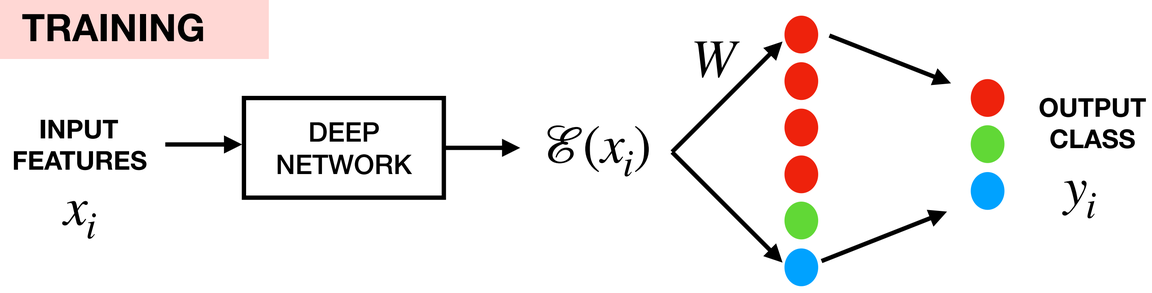
## The loss function is

(derived from weighted categorical cross entropy loss)

$$\mathscr{L} = -\sum_i \log v_i[k_1]$$

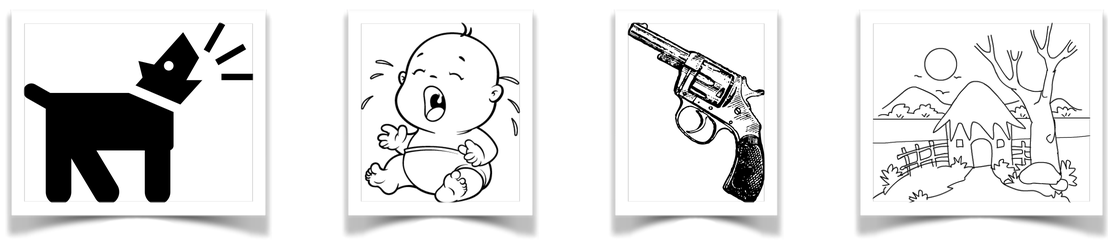


## CLASSIFIER TRAINING




Train last layer, and fine tune end-to-end

## AUDIO EVENT DETECTION



| Dogbark | Babycry | Gunshot | Background |
|---|---|---|---|
| 13.5k | 2.3k | 4.1k | 36.0k |

**Dataset:**

- 10s audio samples from AudioSet
- Weakly labeled with 1 event or background
- Training (70%), validation (20%), testing (20%)

**Input Features from audio:**

- Frame length of 25ms and hop size of 10ms
- 64 dimensional log mel filter bank energies
- Mean and variance normalization

**Embedding Networks:**

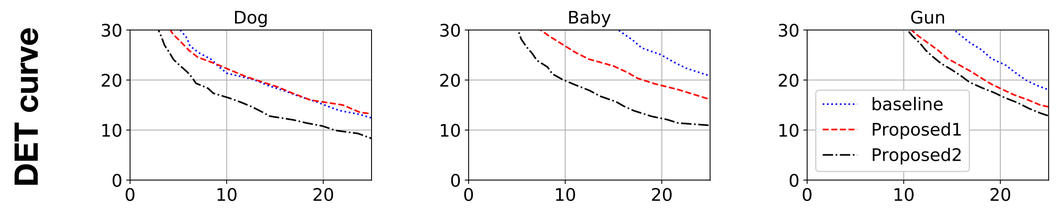
1. **LSTM model**
   - Single LSTM layer with 128 nodes
2. **CNN model**
   - First layer: 32 7x7 conv filters, ReLU activation
   - Batch normalization
   - 5x4 max pooling, 30% dropout
   - Second layer: 64 7x7 conv filters, ReLU activation
   - Batch normalization
   - 100x4 max pooling, 30% dropout

Trained with Adam, batch size 64 with 8 parallel GPUs

**Baseline:**

Class-weighted loss function, same network architecture

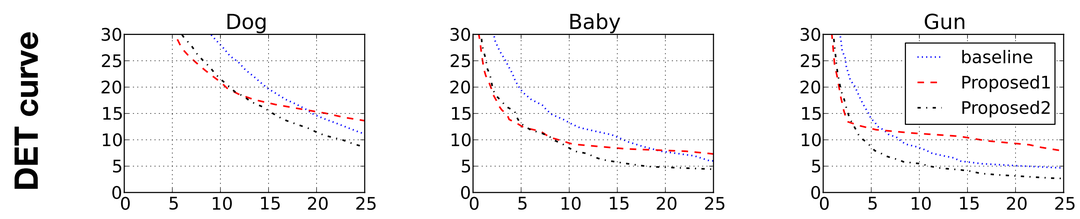
## LSTM MODEL WITH 6:1:2:16 DATA RATIO




| EER | Baseline | Proposed1 | Proposed2 |
|---|---|---|---|
| Dog | 21.5 | 19.1 | **18.3** |
| Baby | 22.4 | 19.3 | **15.5** |
| Gun | 17.0 | 17.1 | **13.6** |
| Overall | **20.3** | **18.5** | **15.8** |

(Proposed2 does final end-to-end tuning of Embedding+Classifier, Proposed1 does not)

## CNN MODEL WITH 2:2:1:26 DATA RATIO




| EER | Baseline | Proposed1 | Proposed2 |
|---|---|---|---|
| Dog | 17.3 | 16.5 | **15.2** |
| Baby | 12.0 | 9.6 | **9.2** |
| Gun | 9.0 | 11.1 | **6.8** |
| Overall | **12.8** | **12.4** | **10.4** |