Cross Modal Audio Search and Retrieval with Joint Embeddings based on Text and Audio

Benjamin Elizalde, Shuayb Zarar, Bhiksha Raj

Microsoft Research, Carnegie Mellon University

ICASSP 2019

Introduction

- existing audio search engines : matching text-text, or audio-audio
- 팝콘이 터지는 소리 & 불꽃놀이 소리 -> 음향적으론 비슷하지만 어휘적으론 아님
- 바이올린 연주 & 바이올린 파괴 -> 어휘적 의미론 비슷하지만 음향은 아님
- no tag, 혹은 noisy한 audio는?

Proposed Solution - Cross Modal Search

- shared latent space에 text와 audio를 임베딩

- Shared space는 audio와 text 간의 semantic similarity 반영

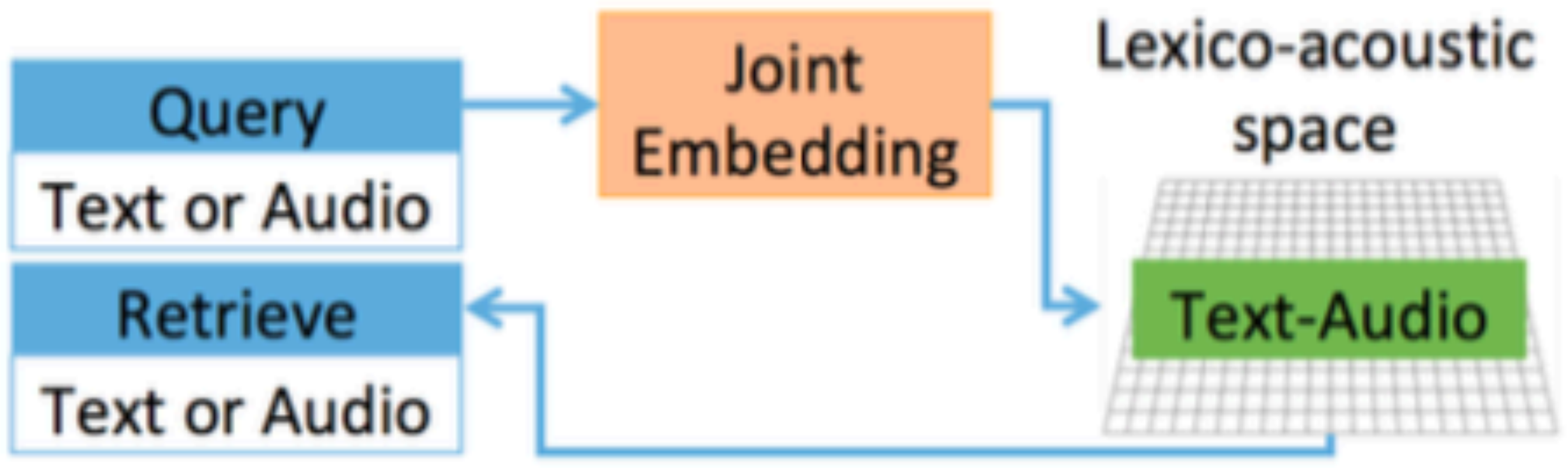
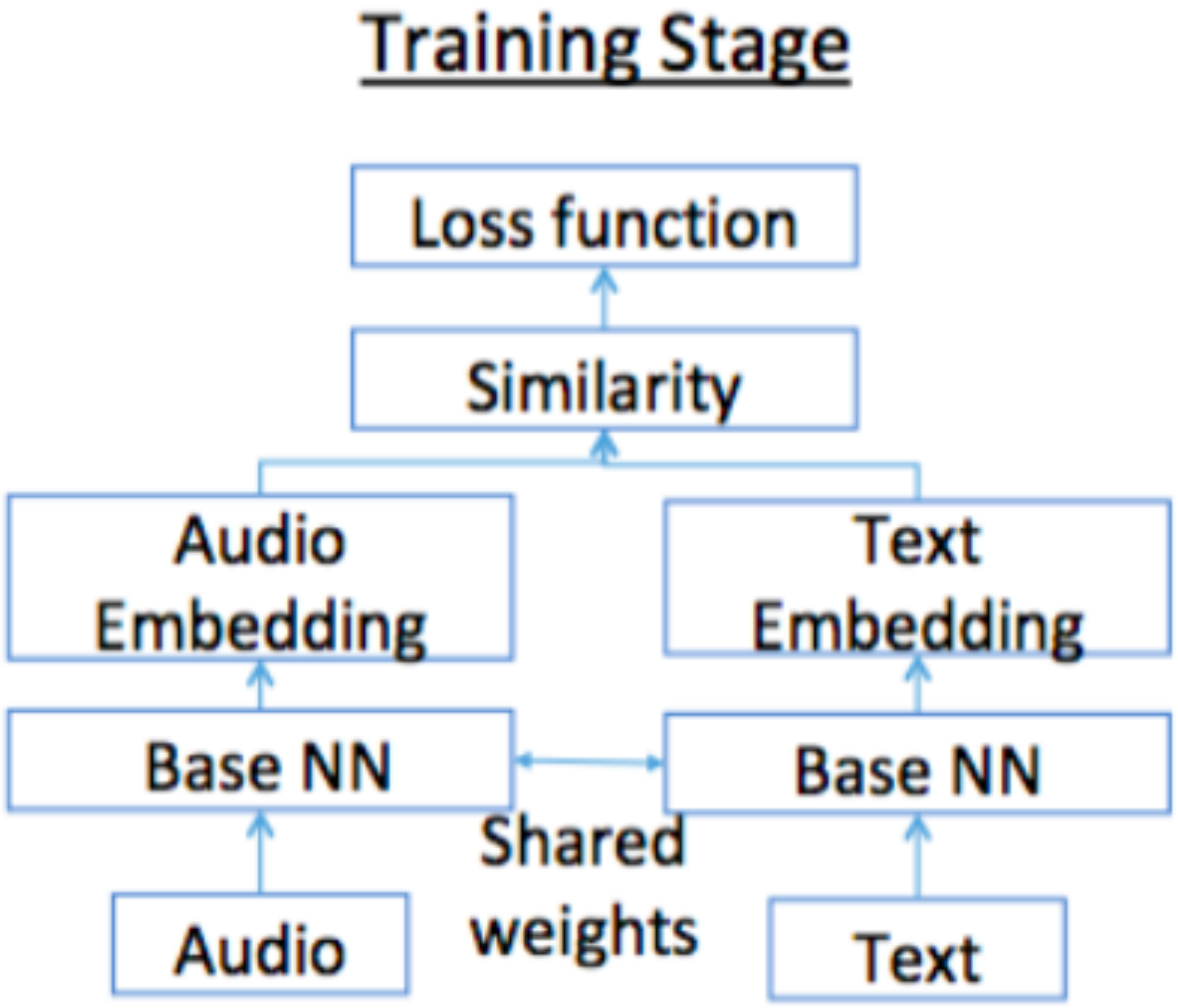


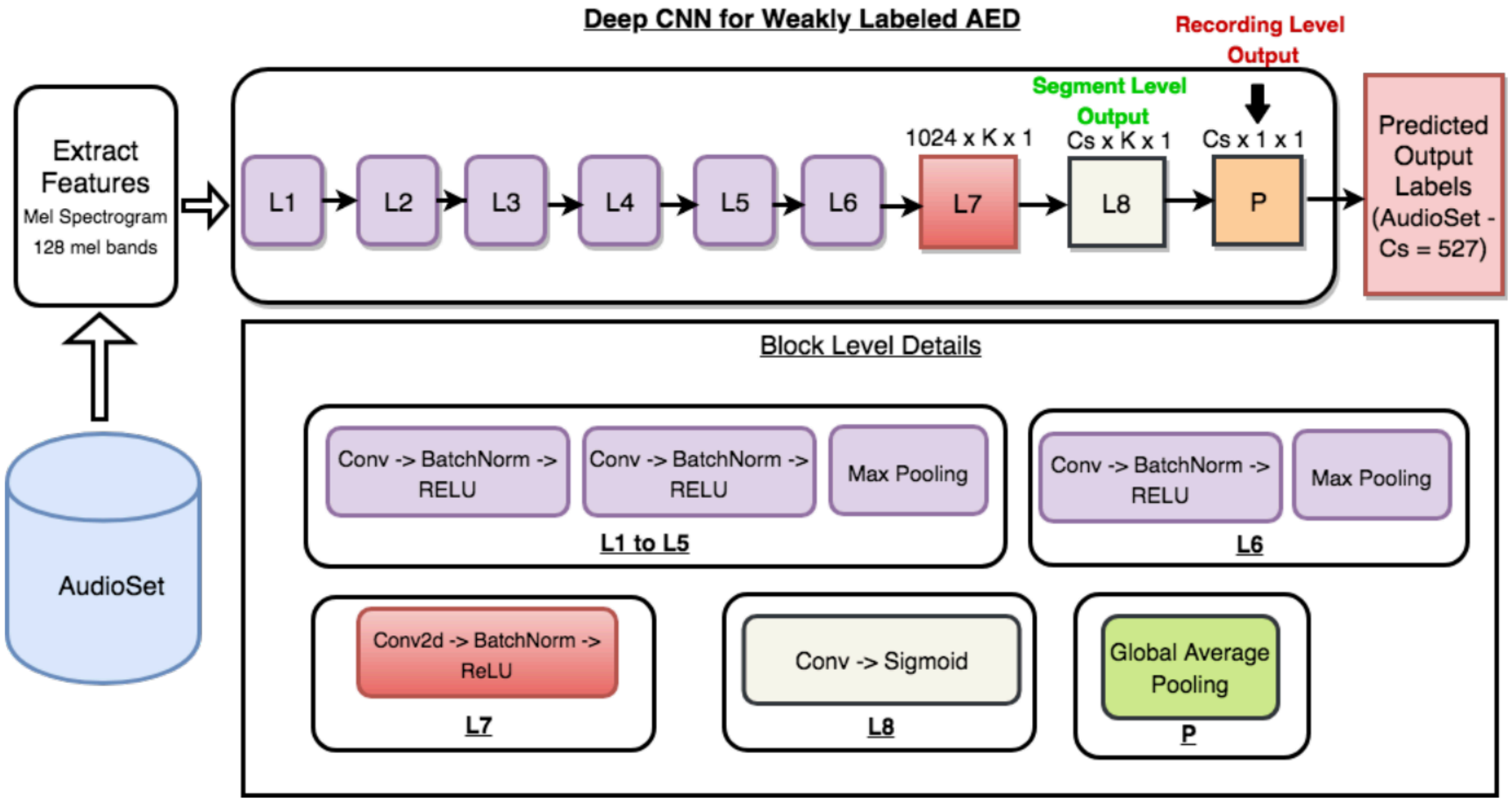
Fig. 1. Proposed framework enables cross-modal search and direct comparison of audio and text modalities. Shared latent space fuses lexical semantics with acoustic similarity.

Proposed Architecture

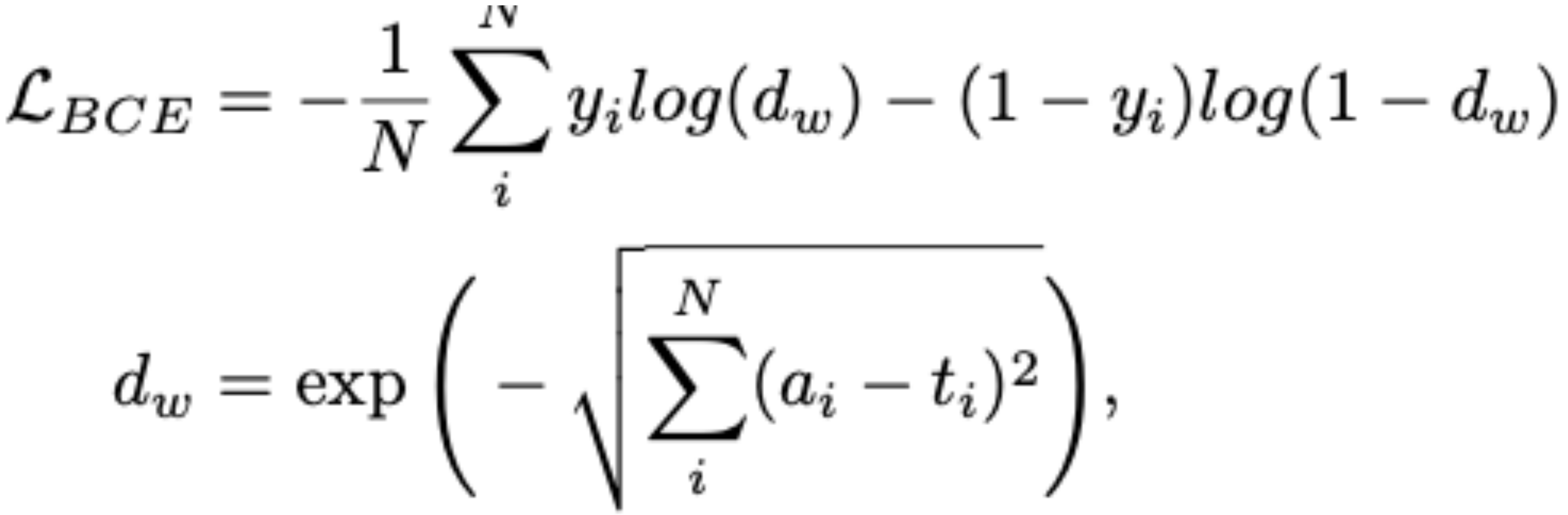


- Siamese network 사용
- Text features : GloVe (별도 학습)
- Audio features : MFCC
  WAL-Net (별도 학습)
- 데이터 당 300차원의 features 하나씩
- Base NN : 4 fc layers
  300(input) - 1024 - 512 -512 - **1024**

Wal-Net


Deep CNN for Weakly Labeled AED
Extract Features
Mel Spectrogram
128 mel bands
AudioSet
L1
L2
L3
L4
L5
L6
L7
L8
P
1024 x K x 1
Segment Level Output
Cs x K x 1
Recording Level Output
Cs x 1 x 1
Predicted Output Labels (AudioSet - Cs = 527)
Block Level Details
Conv -> BatchNorm -> RELU
Conv -> BatchNorm -> RELU
Max Pooling
L1 to L5
Conv -> BatchNorm -> RELU
Max Pooling
L6
Conv2d -> BatchNorm -> ReLU
L7
Conv -> Sigmoid
L8
Global Average Pooling
P

Similarity & Loss



$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i}^{N} y_i log(d_w) - (1 - y_i) log(1 - d_w)$$

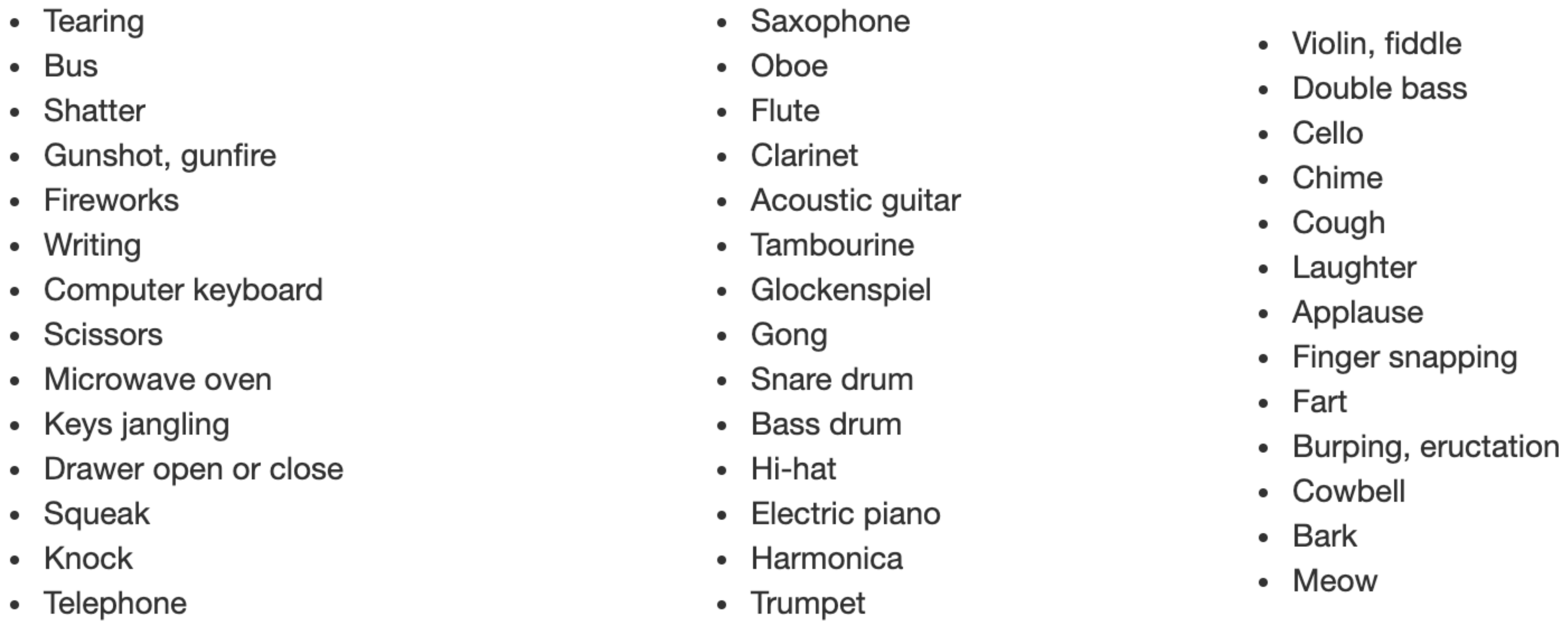
$$d_w = \exp\left(-\sqrt{\sum_{i}^{N}(a_i - t_i)^2}\right),$$

- 임베딩 후 결과가 sparse하고, 0에 가까운 값을 가짐
  -> negative pair에 대해서도 loss가 작음

- y_i : audio-text의 positive pair 여부

- d_w : audio embedding-text embedding의 negative Euclidean distance

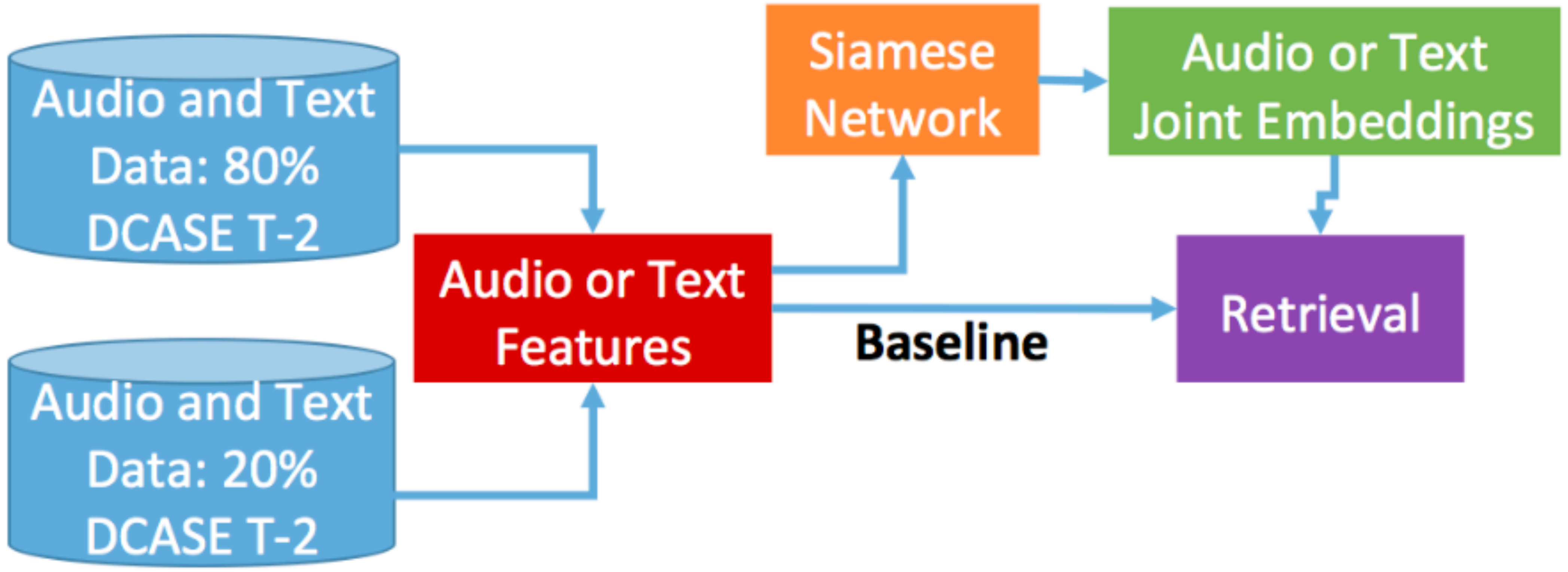
Dataset



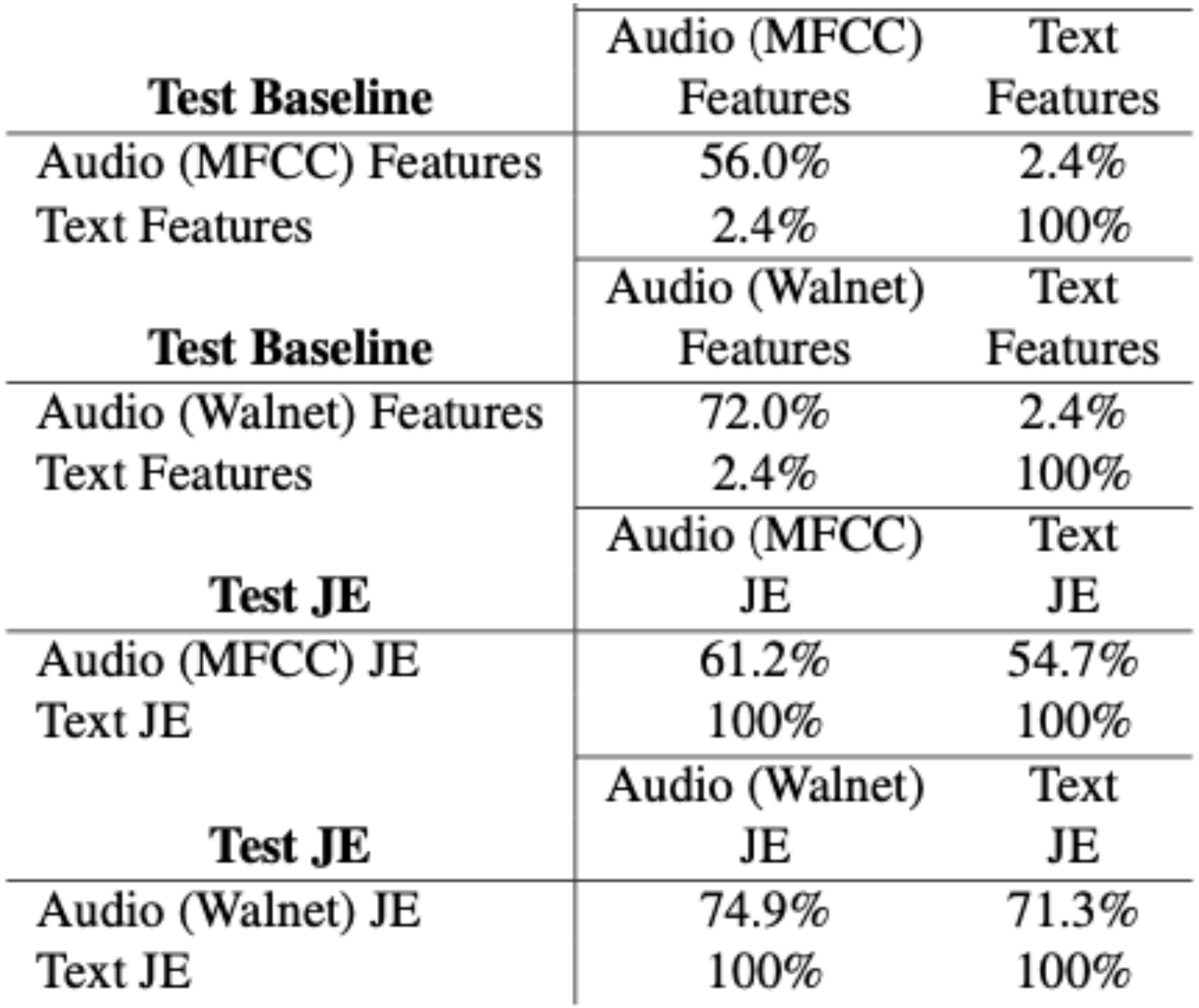
- Tearing
- Bus
- Shatter
- Gunshot, gunfire
- Fireworks
- Writing
- Computer keyboard
- Scissors
- Microwave oven
- Keys jangling
- Drawer open or close
- Squeak
- Knock
- Telephone
- Saxophone
- Oboe
- Flute
- Clarinet
- Acoustic guitar
- Tambourine
- Glockenspiel
- Gong
- Snare drum
- Bass drum
- Hi-hat
- Electric piano
- Harmonica
- Trumpet
- Violin, fiddle
- Double bass
- Cello
- Chime
- Cough
- Laughter
- Applause
- Finger snapping
- Fart
- Burping, eructation
- Cowbell
- Bark
- Meow

- task-2 of the 2018 DCASE challenge (train 9.5k / test 1.6k)
- 41 classes (unequally distributed, 94-300)
- length : 0.3 ~ 30s

Experiments


Audio and Text Data: 80% DCASE T-2
Audio and Text Data: 20% DCASE T-2
Audio or Text Features
Siamese Network
Audio or Text Joint Embeddings
Retrieval
Baseline

classifier : K-nn (k=25)
metrics : mAP



Test Baseline	Audio (MFCC) Features	Text Features
Audio (MFCC) Features	56.0%	2.4%
Text Features	2.4%	100%
Test Baseline	Audio (Walnet) Features	Text Features
Audio (Walnet) Features	72.0%	2.4%
Text Features	2.4%	100%
Test JE	Audio (MFCC) JE	Text JE
Audio (MFCC) JE	61.2%	54.7%
Text JE	100%	100%
Test JE	Audio (Walnet) JE	Text JE
Audio (Walnet) JE	74.9%	71.3%
Text JE	100%	100%

Experiments

sample results

- **query : <gunshot>**
  glove : gunshot, tearing, applause, cough
  proposed : gunshot, fireworks, microwave oven, knock

- **query : <meow>**
  glove : meow, fart, cough
  proposed : meow, bark, trumpet

- acoustic 정보가 반영되었음을 확인

Experiments

Out of Vocabulary

- **query : <house>**
  GloVe: drawer, telephone, writing, gunshot, double bass
  proposed : meow, cough, finger snapping, laughter, computer keyboard,

- **query : thunderstorm (sound file)**
  WAL-net : fire-works, applause, tearing, fart
  proposed : fire-works, cough, drawer open or close, gunshot

- **query : orchestra (sound file)**
  WAL-net : applause, cello, acoustic guitar, filte, fireworks, violin, clarinet
  proposed : violin, trumpet, saxophone, flute, double bass, clarinet, cello

.

Conclusions

- audio와 text 쿼리 모두 같은 모델을 사용하여 검색하는 corss-modal 모델을 제안
- shared latent space으로 매핑된 벡터는 audio와 text사이에 semantic similarity를 보존

Reference

- Wal-Net : https://arxiv.org/pdf/1804.09288.pdf
- DCASE2018 : http://dcase.community/challenge2018/task-general-purpose-audio-tagging
- Paper : https://www.microsoft.com/en-us/research/uploads/prod/2019/04/MartinezZararRaj_ICASSP_2019.pdf