# MT-GCN For Multi-Label Audio Tagging With Noisy Labels

*Shrivaslava, Harsh, et al.*

** ICASSP 2020*

음악오디오 연구실

이재준

*International Conference on Acoustics, Speech and Signal Processing

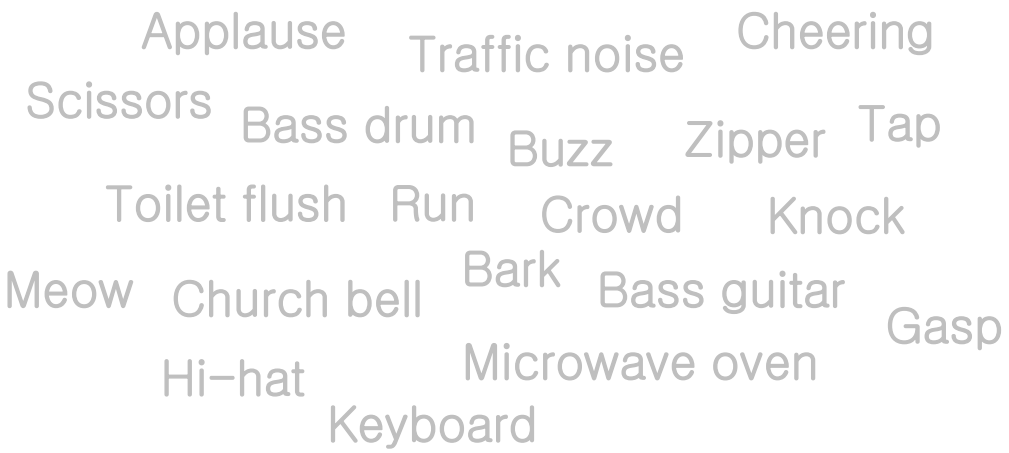
# Contents

# 1. Preliminaries & Introduction

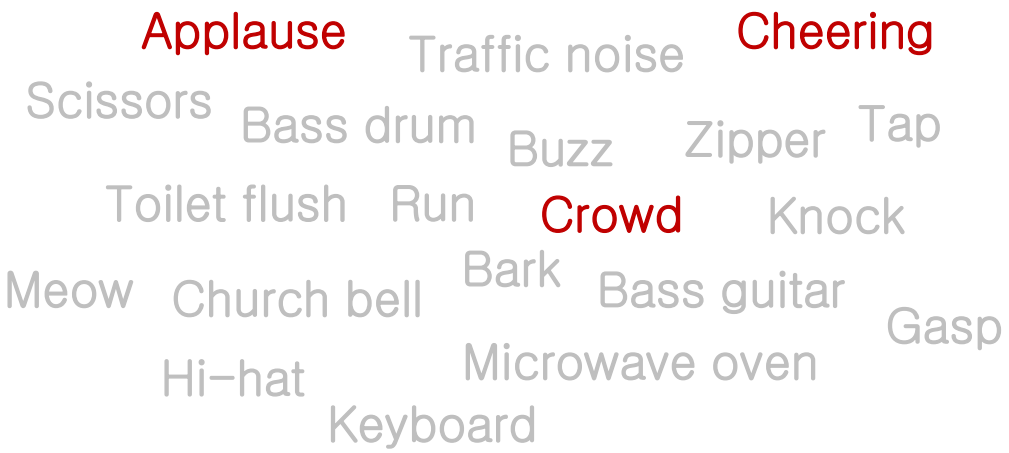
# MT-GCN For Multi-Label Audio Tagging With Noisy Labels

# MT-GCN For Multi-Label Audio Tagging With Noisy Labels

# MT-GCN For Multi-Label Audio Tagging With Noisy Labels



# MT-GCN For Multi-Label Audio Tagging With Noisy Labels

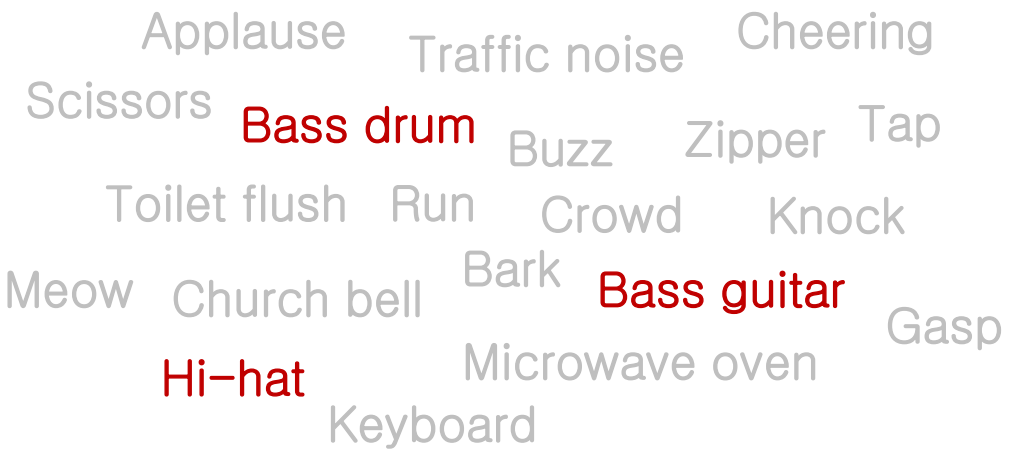


# MT-GCN For Multi-Label Audio Tagging With Noisy Labels

Applause
Traffic noise
Cheering
Scissors
Bass drum
Buzz
Zipper
Tap
Toilet flush
Run
Crowd
Knock
Bark
Meow
Church bell
Bass guitar
Gasp
Hi-hat
Microwave oven
Keyboard



# MT-GCN For Multi-Label Audio Tagging With Noisy Labels



# MT-GCN For Multi-Label Audio Tagging With Noisy Labels

## Motivation

1. Release of large-scale datasets
2. Inevitable label noise
3. Using minimal supervision of labels
   (**DCASE 2019* challenge task2)

# MT-GCN For Multi-Label Audio Tagging With Noisy Labels

## Motivation

1. Release of large-scale datasets
2. Inevitable label noise
3. Using minimal supervision of labels
   (**DCASE 2019* challenge task2)

# MT-GCN For Multi-Label Audio Tagging With Noisy Labels

## Motivation

1. Release of large-scale datasets
2. Inevitable label noise
3. Using minimal supervision of labels
   (**DCASE 2019* challenge task2)

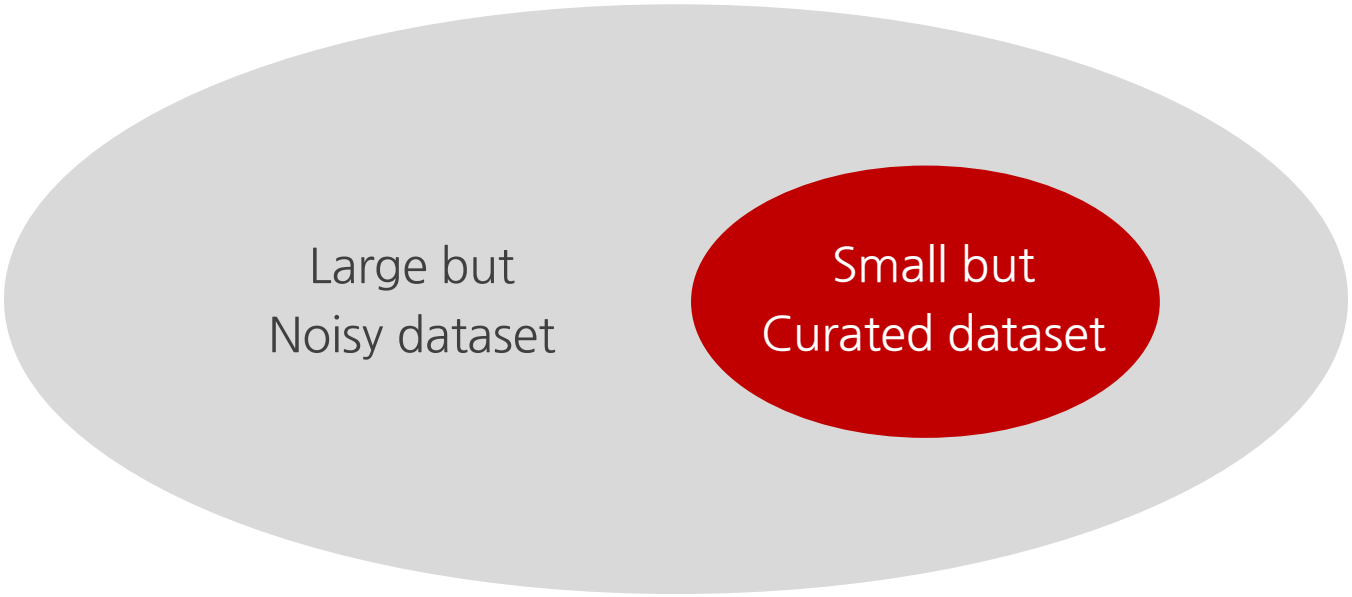
# MT-GCN For Multi-Label Audio Tagging With Noisy Labels

## Motivation

1. Release of large-scale datasets
2. Inevitable label noise
3. Using minimal supervision of labels
   (**DCASE 2019* challenge task 2)

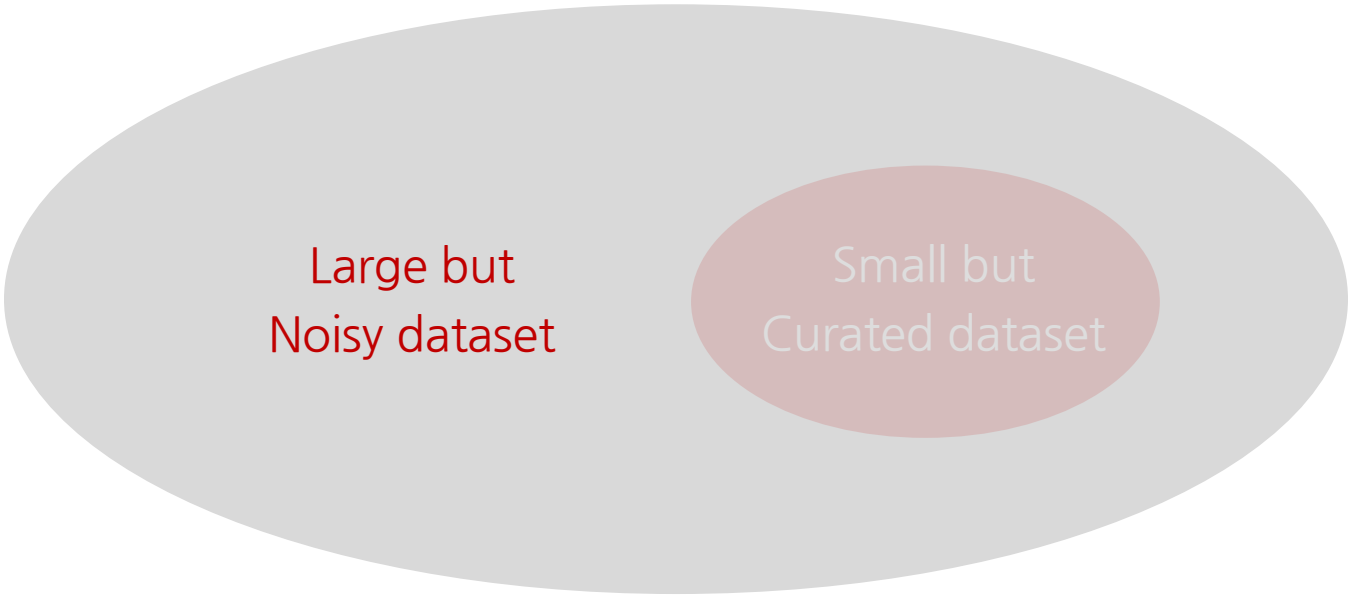
# MT-GCN For Multi-Label Audio Tagging With Noisy Labels

## Motivation



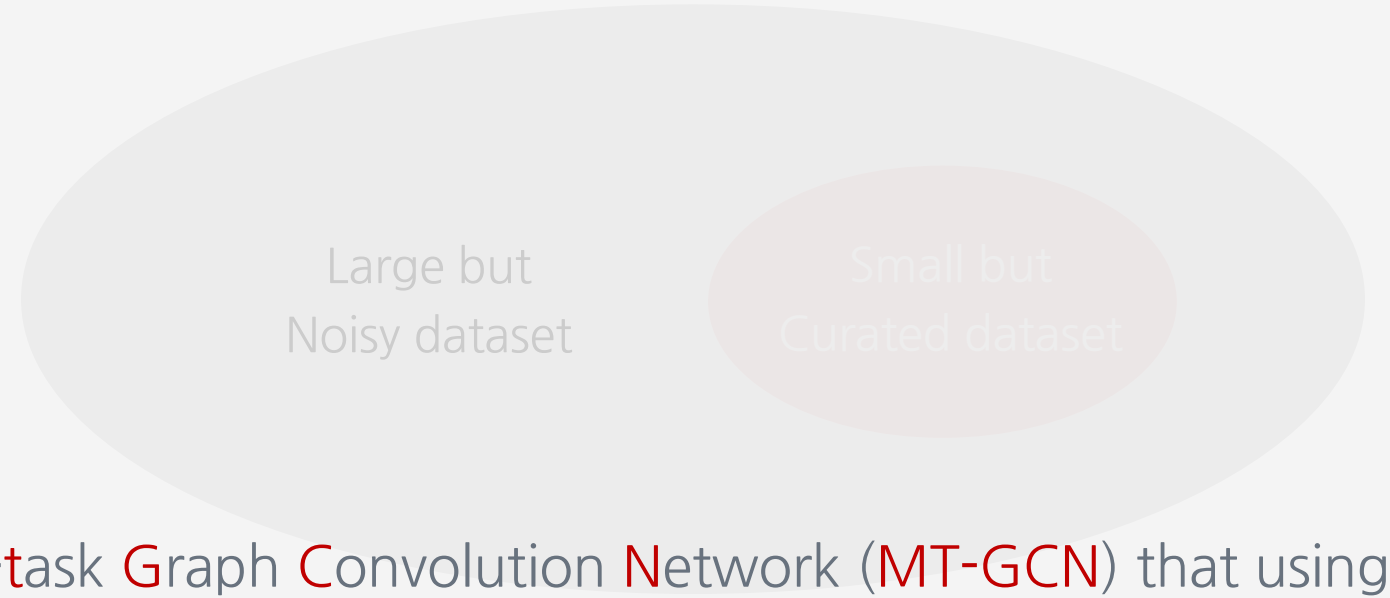
# MT-GCN For Multi-Label Audio Tagging With Noisy Labels

## Motivation



# MT-GCN For Multi-Label Audio Tagging With Noisy Labels

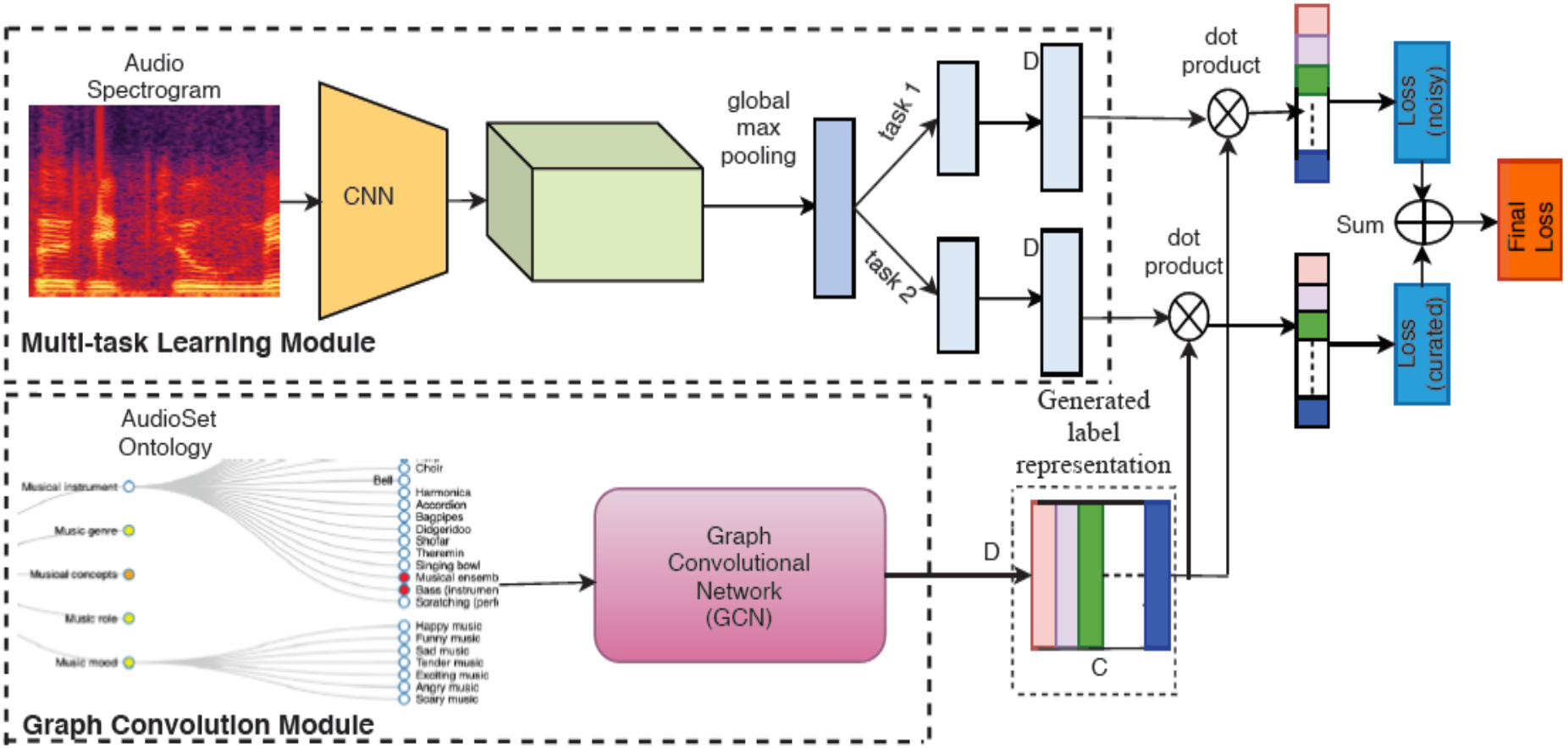
Motivation



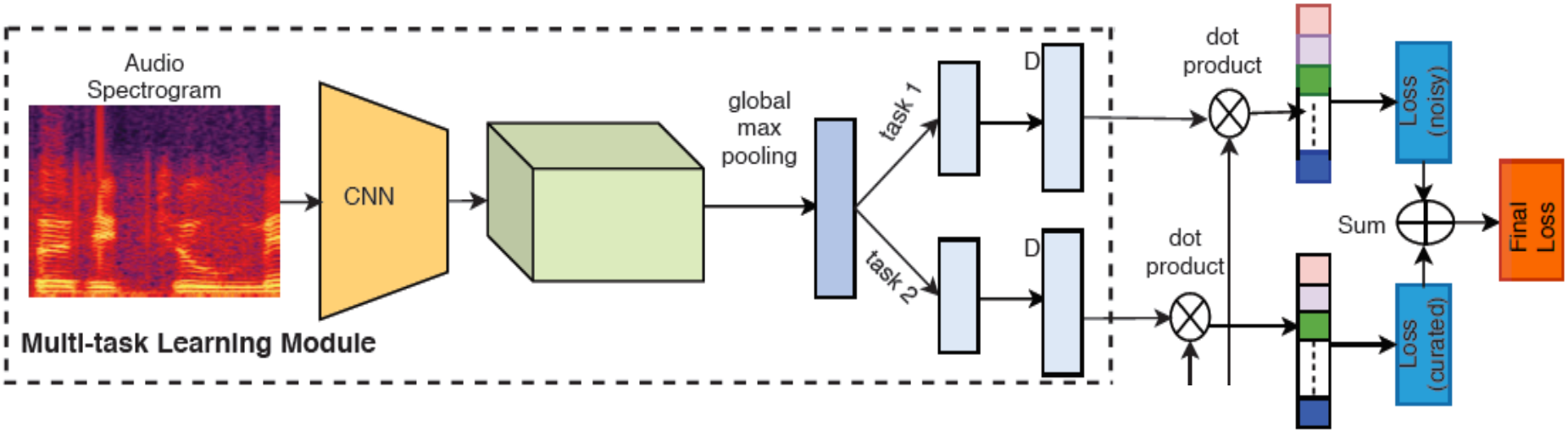

Multi-task Graph Convolution Network (MT-GCN) that using Ontology-based domain knowledge as a regularization in a multi-task learning setup to deal with label noise

# 2. Methodology

# Block diagram of MT-GCN



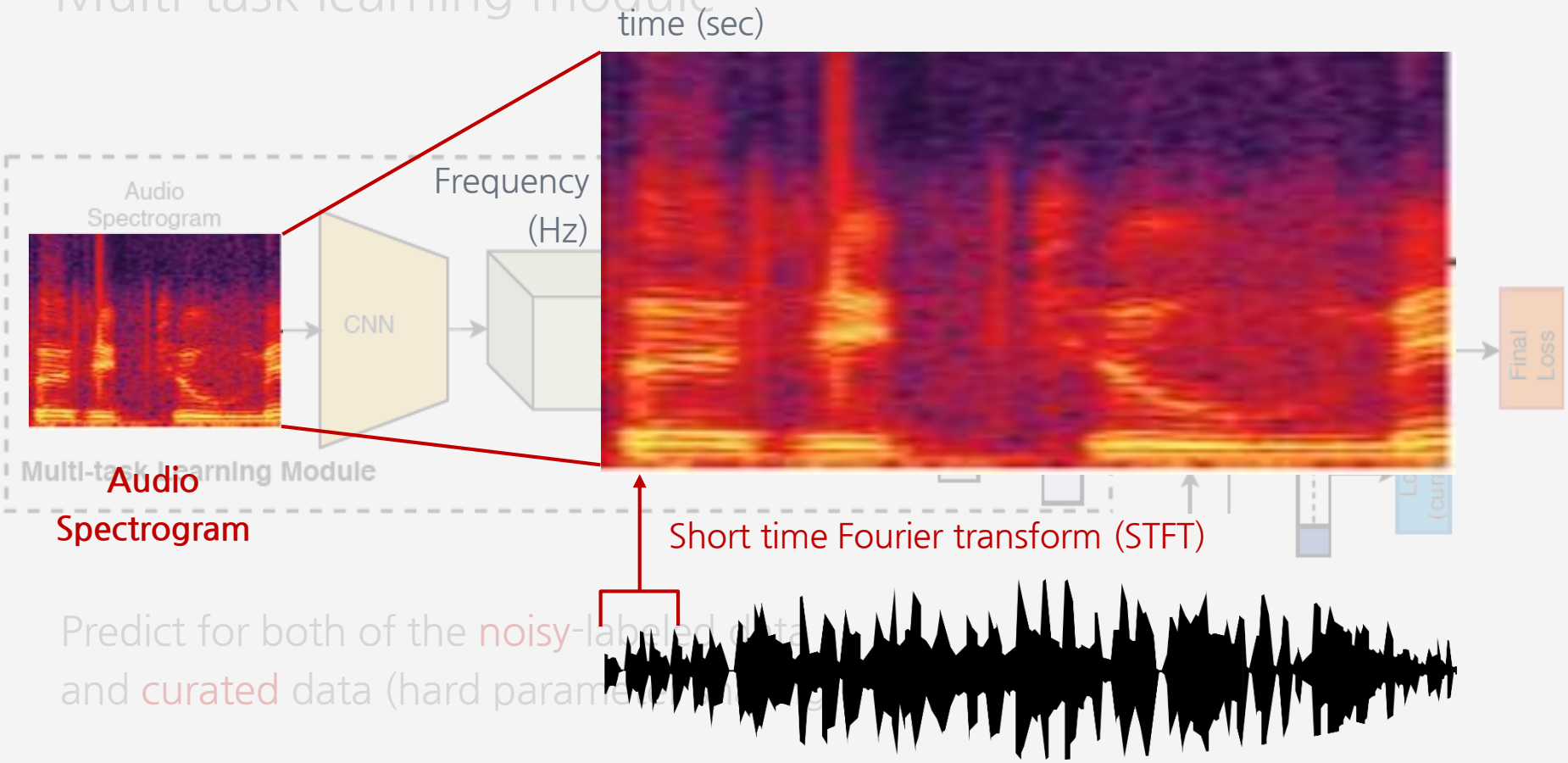
# Multi-task learning module




Predict for both of the noisy-labeled data
and curated data (hard parameter sharing)

Shared network architecture : ResNet-101
Network input : audio spectrogram

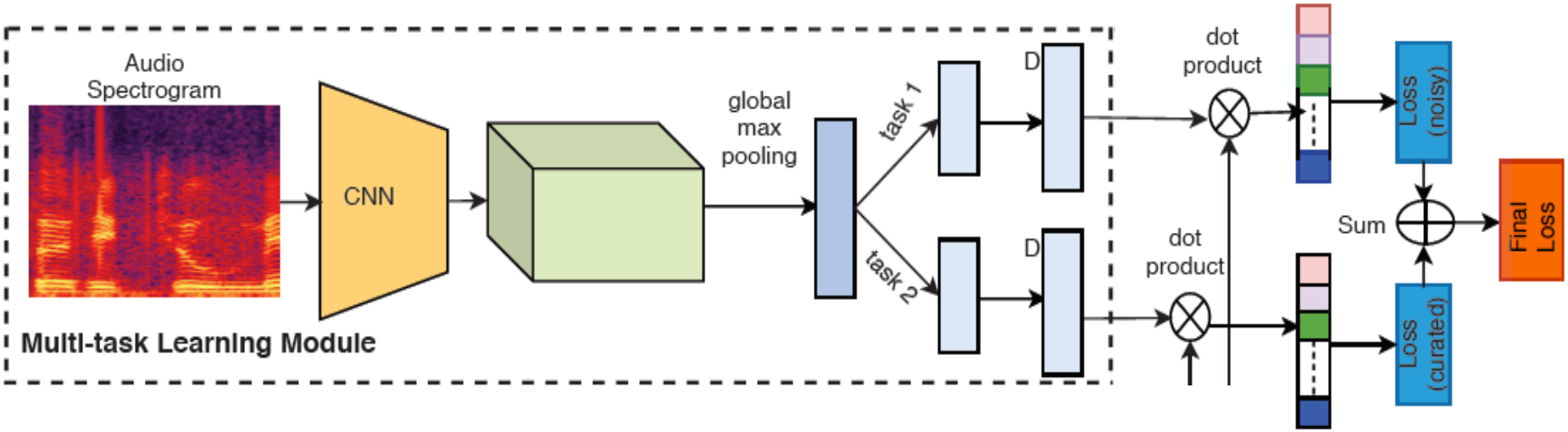
## Multi-task learning module



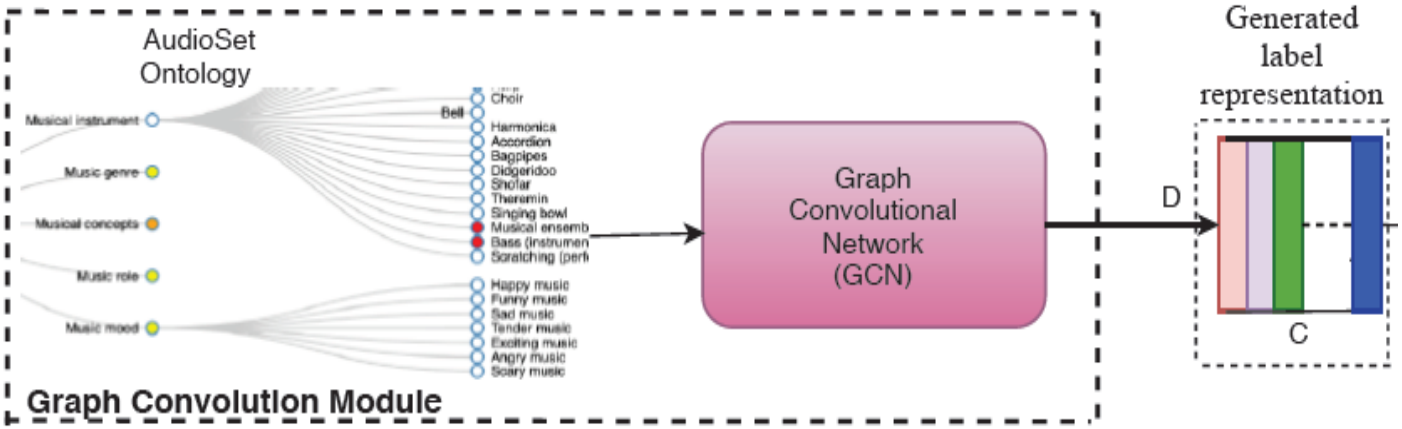

Predict for both of the noisy-labeled data and curated data (hard parameter sharing)

Shared network architecture : ResNet-101
Network input : audio spectrogram

# Multi-task learning module



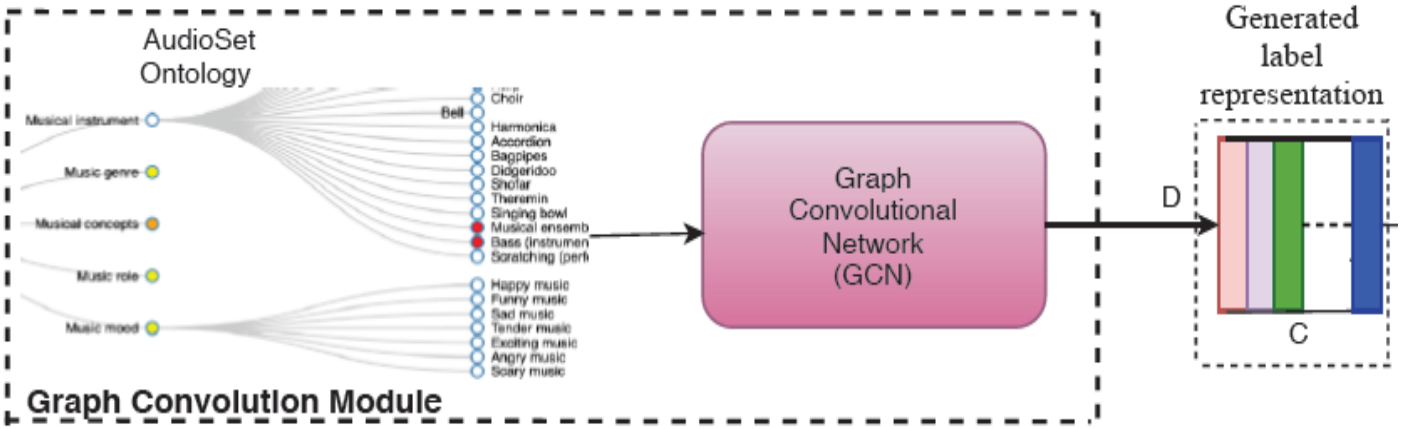
# Graph Convolution module



# Graph Convolution module - Ontology Based Affinity Matrix

Using Google *AudioSet

Very large dataset in audio domain
632 audio event classes
2.1 million of annotated video

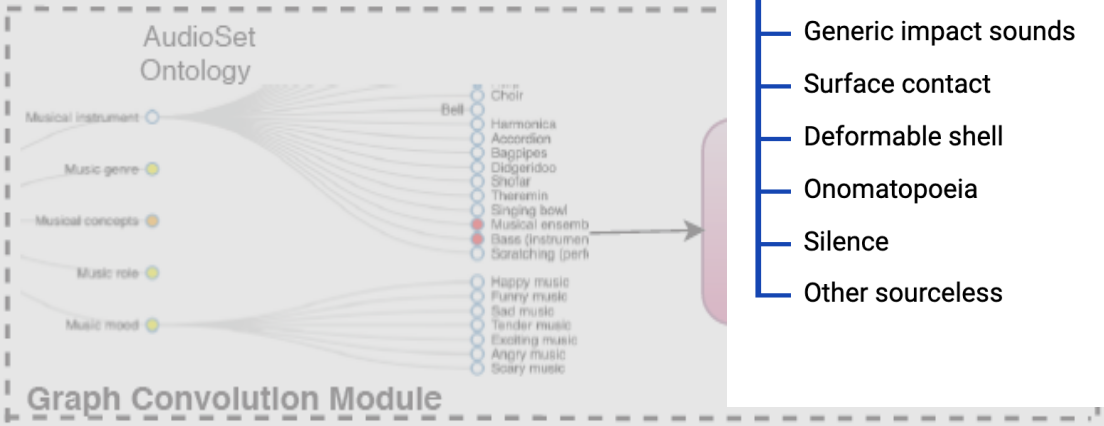



* Gemmeke, Jort F., et al. "Audio set: An ontology and human-labeled dataset for audio events." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

# Graph Convolution mo

Using Google *AudioSet

Very large dataset in audio domain
632 audio event classes
2.1 million of annotated video




**Human sounds**
- Human voice
- Whistling
- Respiratory sounds
- Human locomotion
- Digestive
- Hands
- Heart sounds, heartbeat
- Otoacoustic emission
- Human group actions

**Source-ambiguous sounds**
- Generic impact sounds
- Surface contact
- Deformable shell
- Onomatopoeia
- Silence
- Other sourceless

**Animal**
- Domestic animals, pets
- Livestock, farm animals, working animals
- Wild animals

**Sounds of things**
- Vehicle
- Engine
- Domestic sounds, home sounds
- Bell
- Alarm
- Mechanisms
- Tools
- Explosion
- Wood
- Glass
- Liquid
- Miscellaneous sources
- Specific impact sounds

**Music**
- Musical instrument
- Music genre
- Musical concepts
- Music role
- Music mood

**Natural sounds**
- Wind
- Thunderstorm
- Water
- Fire

**Channel, environment and background**
- Acoustic environment
- Noise
- Sound reproduction

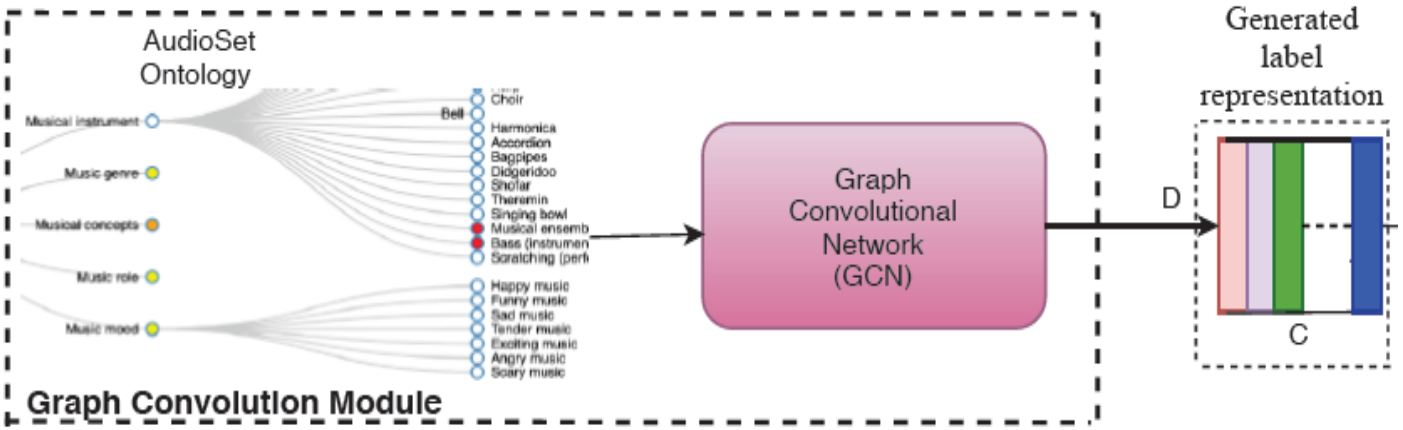
* Gemmeke, Jort F., et al. "Audio set: An ontology and human-labeled dataset for audio events." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

# Graph Convolution module - Ontology Based Affinity Matrix

Using Google *AudioSet

Very large dataset in audio domain
632 audio event classes
2.1 million of annotated video

Benchmark dataset in this paper (FSDKaggle2019 - 80 classes)
followed subset of label ontology of AudioSet.

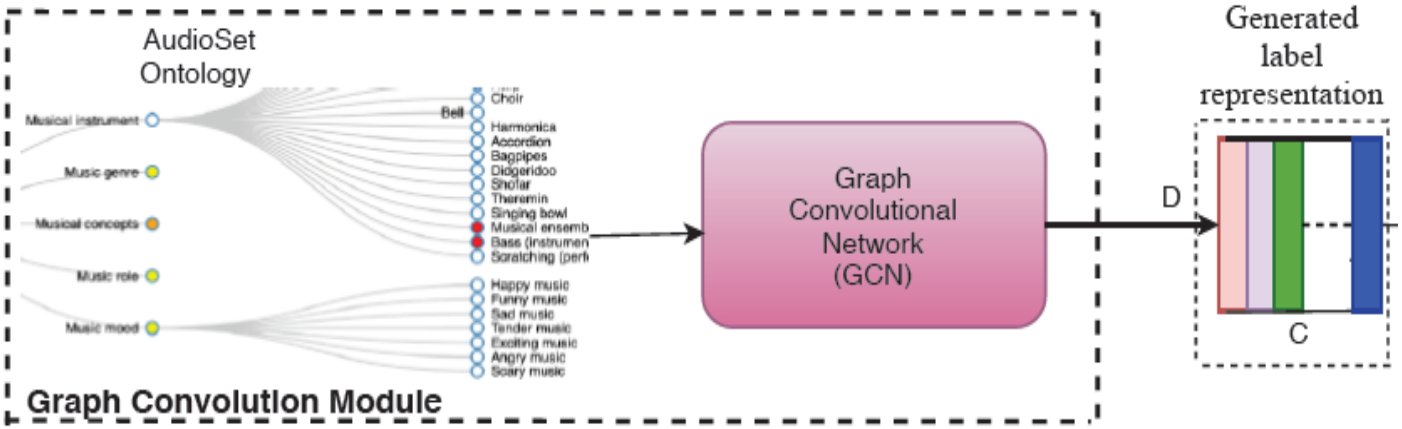



* Gemmeke, Jort F., et al. "Audio set: An ontology and human-labeled dataset for audio events." 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.

# Graph Convolution module - Ontology Based Affinity Matrix

- Method 1

- Method 2

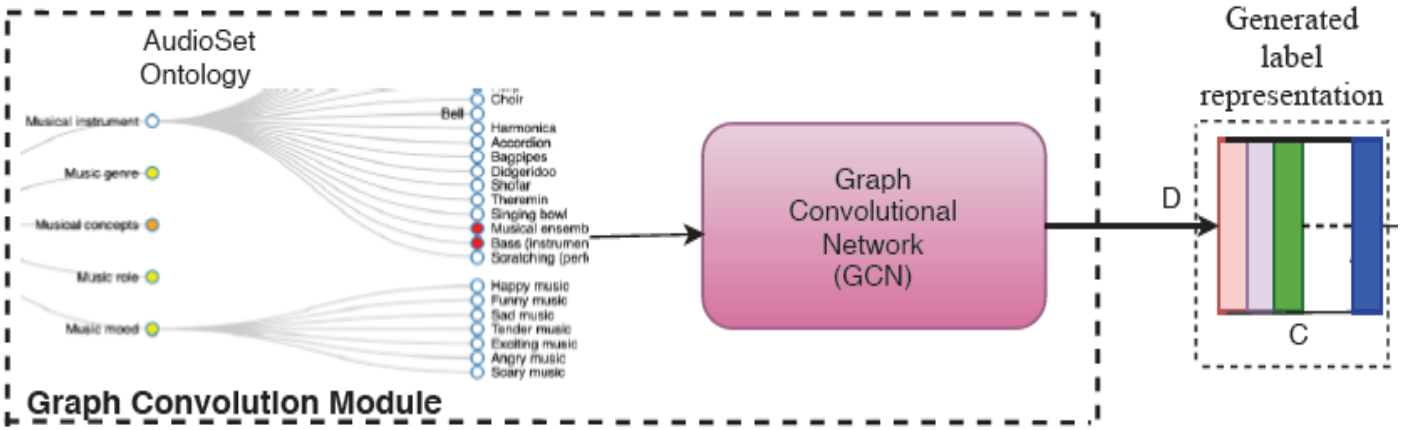


# Graph Convolution module - Ontology Based Affinity Matrix

- Method 1

  Edge is 1 if two labels in benchmark dataset have same parents in ontology

- Method 2



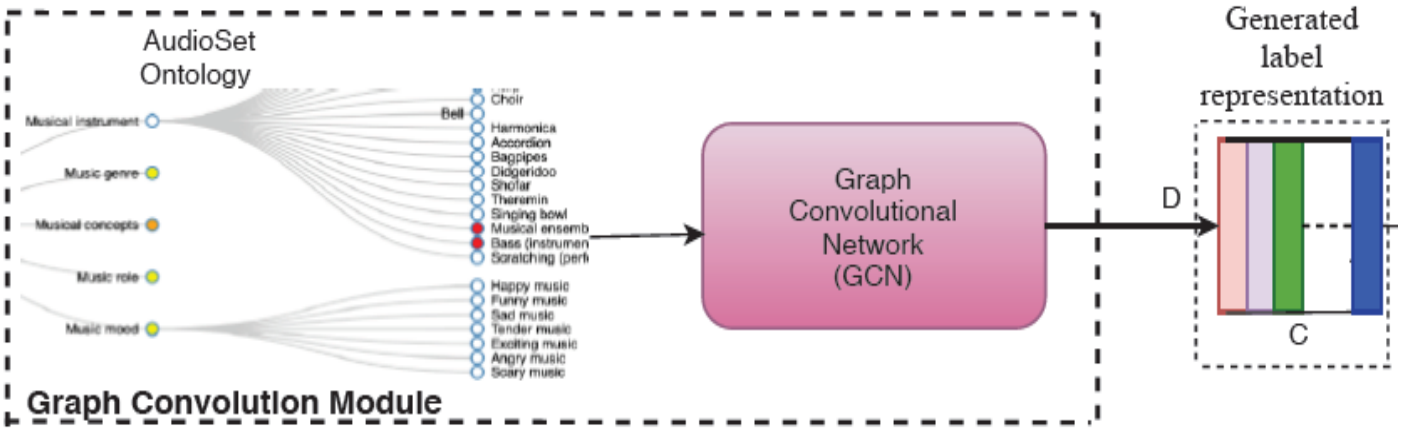
# Graph Convolution module - Ontology Based Affinity Matrix

- Method 1

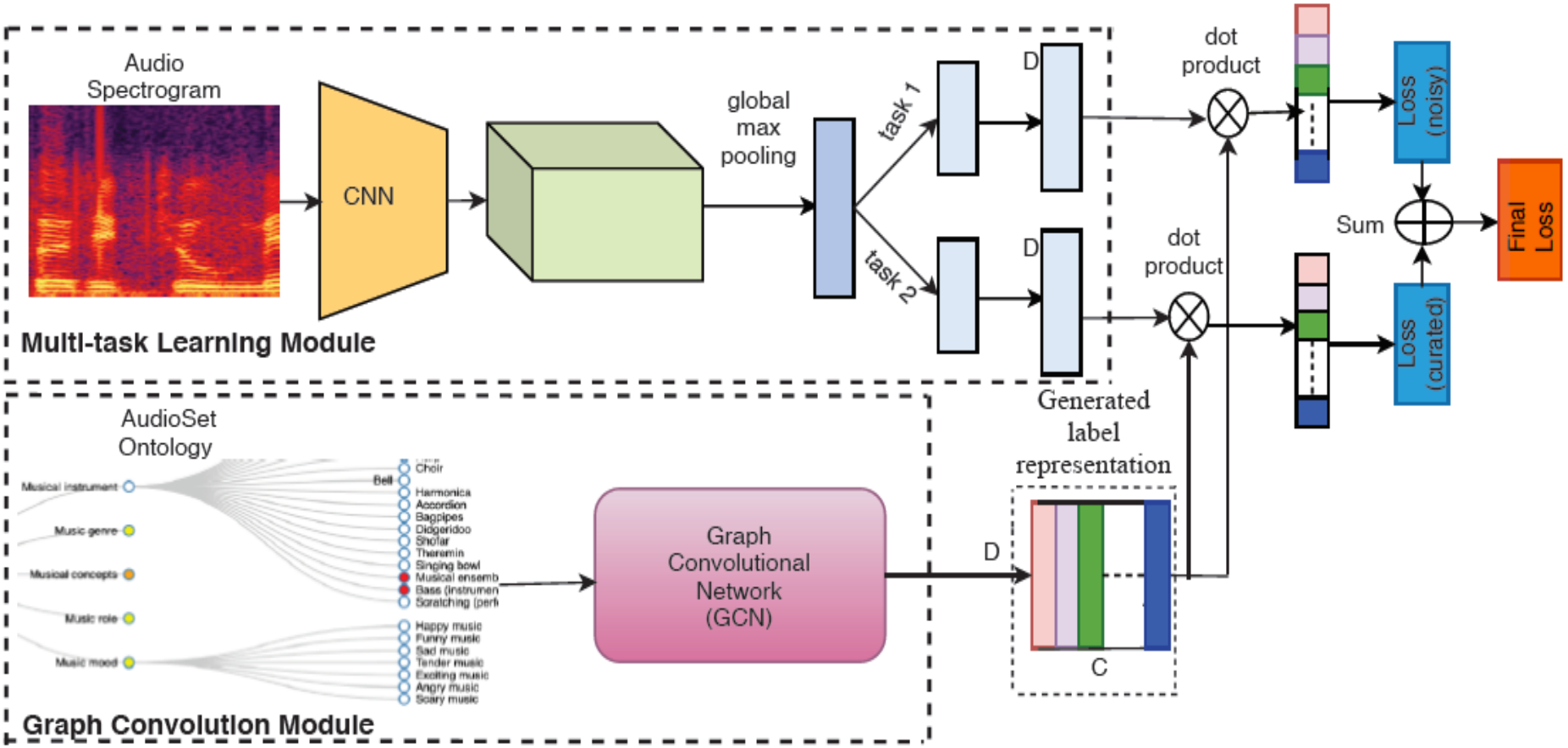
  Edge is 1 if two labels in benchmark dataset have same parents in ontology

- Method 2

  Train GCN using all (~632) classes in Ontology, then slice out only 80



# Block diagram of MT-GCN



# 3. Experiments and Results

# Experimental Settings

1) **MT-GCN_1** : co-occurrence based method of **Chen, Zhao-Min, et al.*
=> Use only curated dataset

2) **MT-GCN_2** : co-occurrence based method of **Chen, Zhao-Min, et al.*
=> Use curated and noisy dataset

3) **MT-GCN_3** : Ontology-based method one (Using Google AudioSet)
=> Train GCN using only 80 label classes in benchmark dataset

4) **MT-GCN_4** : Ontology-based method two (Using Google AudioSet)
=> Train GCN using all label classes in AudioSet and slices out 80 classes

* *Chen, Zhao-Min, et al.* "Multi-label image recognition with graph convolutional networks", CVPR 2019 (홍채은님 발표)

# Results

| Methods | Overall Lwlrap |
|---|---|
| MTN | 0.6794 |
| MT-GCN_1 | 0.6941 |
| MT-GCN_2 | 0.7178 |
| MT-GCN_3 | 0.7244 |
| MT-GCN_4 | **0.7405** |

**Lwlrap** : label-weighted label-ranking average precision (0 bad ~ 1 good)

- MTN (baseline) ~~ Others
- MT-GCN_1 ~~ MT-GCN_2
- MT-GCN_1, MT-GCN_2 ~~ MT-GCN_3, MT-GCN_4
- MT-GCN_3 ~~ MT-GCN_4

## Results

| Methods | Overall Lwlrap |
|---|---|
| MTN | 0.6794 |
| MT-GCN_1 | 0.6941 |
| MT-GCN_2 | 0.7178 |
| MT-GCN_3 | 0.7244 |
| MT-GCN_4 | **0.7405** |

**Lwlrap** : label-weighted label-ranking average precision (0 bad ~ 1 good)

- MTN (baseline) ~~ Others
- MT-GCN_1 ~~ MT-GCN_2
- MT-GCN_1, MT-GCN_2 ~~ MT-GCN_3, MT-GCN_4
- MT-GCN_3 ~~ MT-GCN_4

# Results

| Methods | Overall Lwlrap |
|---|---|
| MTN | 0.6794 |
| MT-GCN_1 | 0.6941 |
| MT-GCN_2 | 0.7178 |
| MT-GCN_3 | 0.7244 |
| MT-GCN_4 | **0.7405** |

Co-occurrence based
MT-GCN_1 (only curated ~~ 5,000 data)
MT-GCN_2 (curated + noisy ~~ 25,000 data)

- MTN (baseline) ~~ Others
- MT-GCN_1 ~~ MT-GCN_2
- MT-GCN_1, MT-GCN_2 ~~ MT-GCN_3, MT-GCN_4
- MT-GCN_3 ~~ MT-GCN_4

# Results

| Methods | Overall Lwlrap |
|---|---|
| MTN | 0.6794 |
| MT-GCN_1 | 0.6941 |
| MT-GCN_2 | 0.7178 |
| MT-GCN_3 | 0.7244 |
| MT-GCN_4 | **0.7405** |

Co-occurrence based
~~
Proposed Ontology based

- MTN (baseline) ~~ Others
- MT-GCN_1 ~~ MT-GCN_2
- MT-GCN_1, MT-GCN_2 ~~ MT-GCN_3, MT-GCN_4
- MT-GCN_3 ~~ MT-GCN_4

# Results

| Methods | Overall Lwlrap |
|---|---|
| MTN | 0.6794 |
| MT-GCN_1 | 0.6941 |
| MT-GCN_2 | 0.7178 |
| MT-GCN_3 | 0.7244 |
| MT-GCN_4 | **0.7405** |

**Proposed Ontology based**
MT-GCN_3 (train using 80 classes of benchmark dataset)
MT-GCN_4 (train all (~632) classes of AudioSet dataset)

- MTN (baseline) ~~ Others
- MT-GCN_1 ~~ MT-GCN_2
- MT-GCN_1, MT-GCN_2 ~~ MT-GCN_3, MT-GCN_4
- MT-GCN_3 ~~ MT-GCN_4

# 4. Reproducing Experiments

# Limitation of reproducing

Unfortunately there is no reproducible code for this paper (MT-GCN), and further, there is no detail experimental settings such as 'batch size', 'input audio length', 'model selection method', or even any training hyper parameters.

Therefore, all the experimental settings are chosen naively, but partially (especially for baseline model) followed the technical report [1] (*Akiyama et al*) which ranked 1st in DCASE2019 challenge task2. The title of the DCASE2019 task2 is '*Audio tagging with noisy labels and minimal supervision*', which used the same dataset with this paper. Note that DCASE model utilized various machine learning techniques such as ensembles, but did not used any graph based techniques.

DCASE : Detection and Classification of Acoustic Scenes and Events

[1] Akiyama, Osamu, and Junya Sato. *Multitask learning and semisupervised learning with noisy data for audio tagging*. DCASE2019 Challenge, Tech. Rep, 2019.

# References for Reproducing

## For preprocessing and baseline model (=DCASE model) - partially followed

Akiyama, Osamu, and Junya Sato. Multitask learning and semisupervised learning with noisy data for audio tagging. DCASE2019 Challenge, Tech. Rep, 2019.
https://github.com/OsciiArt/Freesound-Audio-Tagging-2019

## For extracting embedding of graph data (GCN) - partially followed

Chen, Zhao-Min, et al. "Multi-label image recognition with graph convolutional networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
https://github.com/Megvii-Nanjing/ML-GCN

## Google AudioSet Ontology data

https://github.com/audioset/ontology

# Experimental Settings - Preprocessing (Following DCASE model [1])

## Waveform

sampling rate 44.1 kHz (original data), input audio length 4 sec (slicing part is randomly selected from various length of audio).

## Log mel spectrogram

128 mel frequency channel, 347 STFT hop size, 128 Hz time resolution, converting power to dB, normalized by the mean and standard deviation of each data.

## Simple augmentation techniques

MixUp (*H. Zhang et al*), SpecAugment (*D. S. Park et al*), gain augmentation.

[1] Akiyama, Osamu, and Junya Sato. *Multitask learning and semisupervised learning with noisy data for audio tagging*. DCASE2019 Challenge, Tech. Rep, 2019.

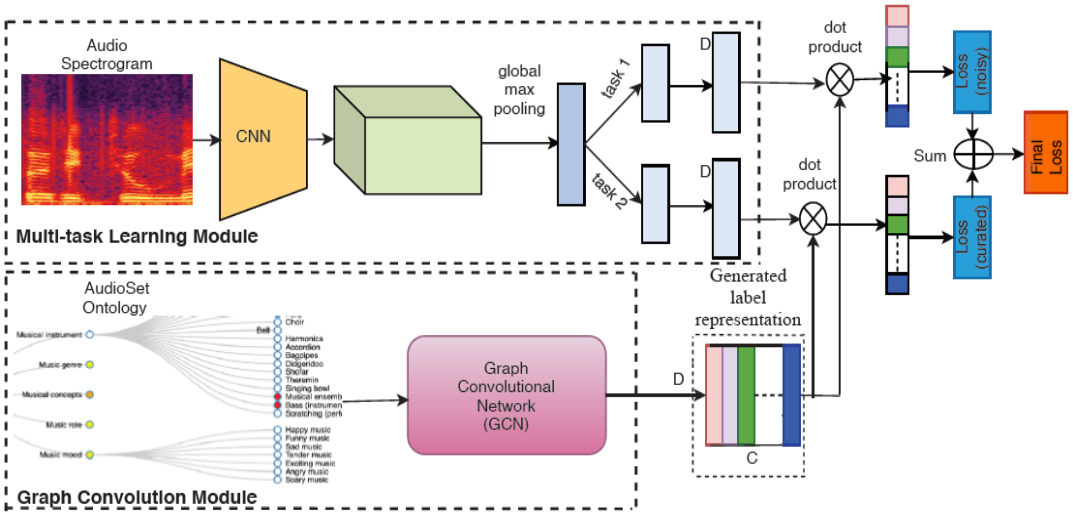
# Experimental Settings - Models

## Baseline model

As mentioned in MT-GCN paper, ResNet101 architecture is used. But detail settings such as dropout or feature dimension are partially similar to DCASE model.

## MT-GCN model

For extracting embedding of graph data using Graph Convolutional Network (GCN), parameters partially followed *Chen, Zhao-Min, et al*[1].




[1] Chen, Zhao-Min, et al. "Multi-label image recognition with graph convolutional networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

# Experimental Settings - training

Loss : Binary Cross Entropy with Logit loss
Batch Size : 16
Maximum epoch : 512 (early stopped if no performance enhance in 100 epochs)
Evaluation : 5-fold cross validation
Model selection : Choose model that has the best validation performance.
Optimizer : Adam
Learning Scheduler : Cosine annealing learning rate schedule
Evaluation metric (performance) : Lwlrap [1]
(label-weighted label-ranking average precision)
Training time : about 4 days for 1 model training (with GTX 1080 ti)

[1] http://dcase.community/challenge2019/task-audio-tagging

# Results

## Original (MT-GCN paper)

| Model | Lwlrap |
|---|---|
| MTN (base) | 0.6794 |
| MT-GCN1 | 0.6941 |
| MT-GCN2 | 0.7178 |
| MT-GCN3 | 0.7244 |
| MT-GCN4 | 0.7405 |

## Reproduced

| Model | Lwlrap |
|---|---|
| MTN (base) | 0.5697 |
| MT-GCN1 | 0.5670 |
| MT-GCN2 | 0.5610 |
| MT-GCN3 | 0.5474 |
| MT-GCN4 | - |

- Reproduced performances are much less than paper performances
- Also according to the reproduced performance, there is no performance gain through proposed MT-GCN methods

# Review for low performance (of reproduced)

- Various hyper-parameter settings are not released in paper. (input settings, train hyper parameter, details of model architecture.. etc)
- Many related work uses batch size 64 (including DCASE model), but we used batch size 16 due to memory issue.
- Affinity matrix settings are not clear (For co-occurence based affinity matrix, we followed method similar to *Chen, Zhao-Min, et al*, but for ontology based affinity matrix which is core proposed technique of the paper, we followed the light instructions in paper. But details like post-processing or normalizing strategies are not released.

- May have code mistake..

# Github link for reproducing code

https://github.com/jaejunL/GCN_final

(Q&A : jjlee0721@snu.ac.kr)

## Contribution

1. Present a novel approach for building general audio tagging systems with noisy labels by utilizing the ontology-based domain knowledge.

2. Propose two effective methods to build the label correlation graph based on the ontology.

## Personal question

1. Is there really a regularization effect

2. Should be multi-tasking learning

Thank you