SAMPLE-LEVEL CNN ARCHITECTURES FOR MUSIC AUTO-TAGGING USING RAW WAVEFORMS

Taejun Kim[1], Jongpil Lee[2], Juhan Nam[2]

[1]School of Electrical and Computer Engineering, University of Seoul, Republic of Korea. ktj7147@uos.ac.kr

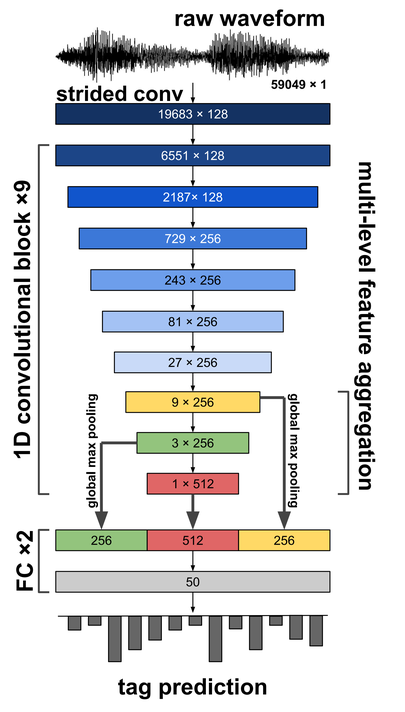
[2]Graduate School of Culture Technology, KAIST, Republic of Korea. {richter, juhannam}@kaist.ac.kr

1 Summary

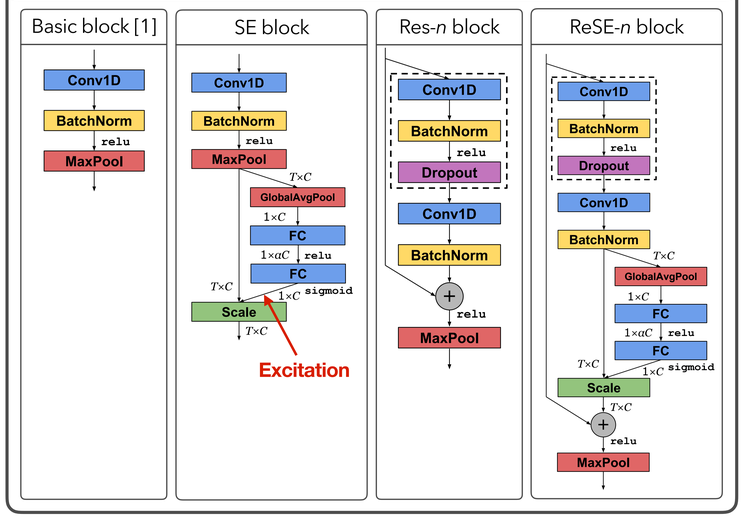
- We explore **end-to-end Convolutional Neural Network (CNN) architectures** for music auto-tagging tasks.
- We adopt architectures from state-of-the-art image classification networks (ResNets & SENets).
- Our models achieve **state-of-the-art results on MagnaTagATune** dataset and **comparable results on Million Song Dataset**.
- We analyze and visualize that the **SE blocks tend to normalize loudness of audios.**

2 Architectures

Base Architecture



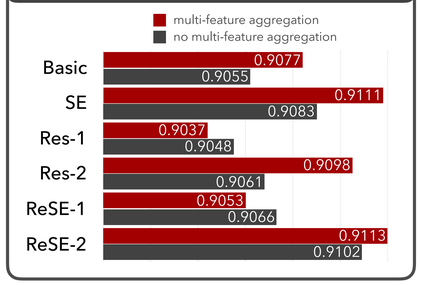
1D Convolutional Blocks



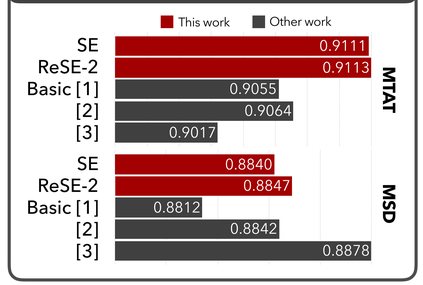
- Basic block is from our previous work [1], Res-n block from ResNets, SE block from Squeeze-and-Excitation Nets (SENets), and ReSE-n block is a combination of them.
- ***n*** denotes the number of convolution layers. If the ***n*** is one, then layers in the dotted lines are removed.
- A **training time** of the ***ReSE-2*** block is **1.7x longer** than the ***SE*** block.
- The **SE block** aggregates a global temporal information, and **recalibrates feature maps** from the basic block.

3 Experiments

AUCs of the Blocks



AUCs of the State-of-the-arts

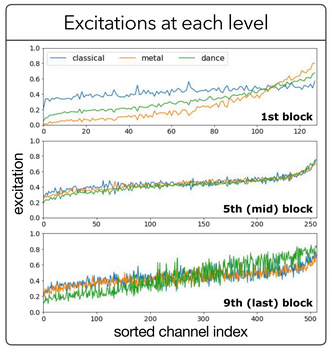


- **MagnaTagATune (MTAT)**
  - 21,105 songs
  - 170h long (29.1s for each song)
  - Split: 15,244 / 1529 / 4332
  - Top 50 most frequent tags
- **Million Song Dataset (MSD)**
  - 241,889 songs
  - 1955h long (29.1s for each song)
  - Split: 201,680 / 11,774 / 28,435
  - 50 tags (Last.FM tag annotations)

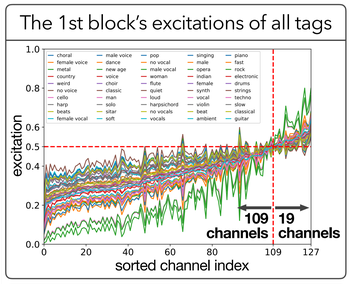
4 Analysis of Excitations from the SE Blocks

Excitations of Tags

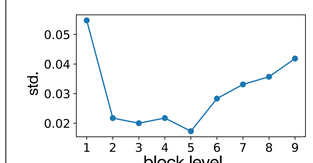
Excitations at each level



The 1st block's excitations of all tags

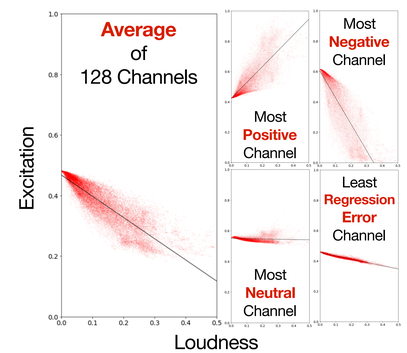


Standard deviations across tags



- **The 1st SE block's excitations vary according to tags.**
- The mid blocks tend to process audios generally whereas the later blocks discriminatively.

The 1st Block's Excitations & Loudnesses



- **The 1st block's excitations tend to normalize the loudness of audios.**
- Black lines are linear regression lines.
- **109** channels have **positive** regression lines whereas **19** have **negative** ones
- Each dot represents a 2.68s long segment from MTAT test set.

https://github.com/tae-jun/resemul

References

- [1] Jongpil Lee, Jiyoung Park, Keunhyoung Luke Kim, and Juhan Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," in Sound and Music Computing Conference (SMC), 2017.
- [2] Jongpil Lee and Juhan Nam, "Multi-level and multi-scale feature aggregation using sample-level deep convolutional neural networks for music classification," Machine Learning for Music Discovery Workshop, International Conference on Machine Learning (ICML), 2017.
- [3] Jongpil Lee and Juhan Nam, "Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging," IEEE Signal Processing Letters, vol. 24, no. 8, pp. 1208–1212, 2017.