

Group 15: What's that sound?

Machine Learning for Urban Sound Classification

Yuhan Zhang, Jingyang Li, Hao Zhu, Yinuo Wang
 {yuz053, jil050, h8zhu, yiw010}@ucsd.edu

Predicting

The classification of environmental sound has lots of applications in large scale and content-based multimedia indexing and retrieval. Nowadays, because of multimedia sensor networks and large quantities of online multimedia contents, people pay more attention to sound classification in urban environments.

In this project, by using urban sound datasets, we will train different machine learning models, such as neural network and boosting, to classify common sounds in urban environments. Also, we will compare the performance of these models.

Data

8732 labeled sound slices ($\leq 4s$) of urban sounds from 10 classes: [1]

air_conditioner	car_horn	children_playing	dog_bark	drilling
engine_idling	gun_shot	street_music	jackhammer	siren

Table1. Labels of urban sound data

File name format:

[fsID]-[classID]-[occurrenceID]-[sliceID].wav

Available at:

<https://urbansounddataset.weebly.com/urbansound8k.html>

Features

There are 4 extracted features:

Mel-Frequency cepstrum: The mel-frequency cepstrum (MFC) is a nonlinear mel-scaled representation of the linear cosine transformation of short-term power spectrum of a sound. [4] **MFCC:** Mel-frequency cepstral coefficients are coefficients of MFC, which are widely used features in sound recognition. [4] **Chroma:** Chroma-based features refers to the "color" of a pitch, which is an important tool for analyzing music. [2][5] **Contrast:** It represented the relative spectral distribution instead of average spectral envelope. [6]

Models

Neural Network

Neural Network model uses four features: Mel-frequency cepstrum, MFCC, Chroma, and Contrast. Then, a two-hidden-layer Neural Network with 4096 cells in each layer is used to make predictions.

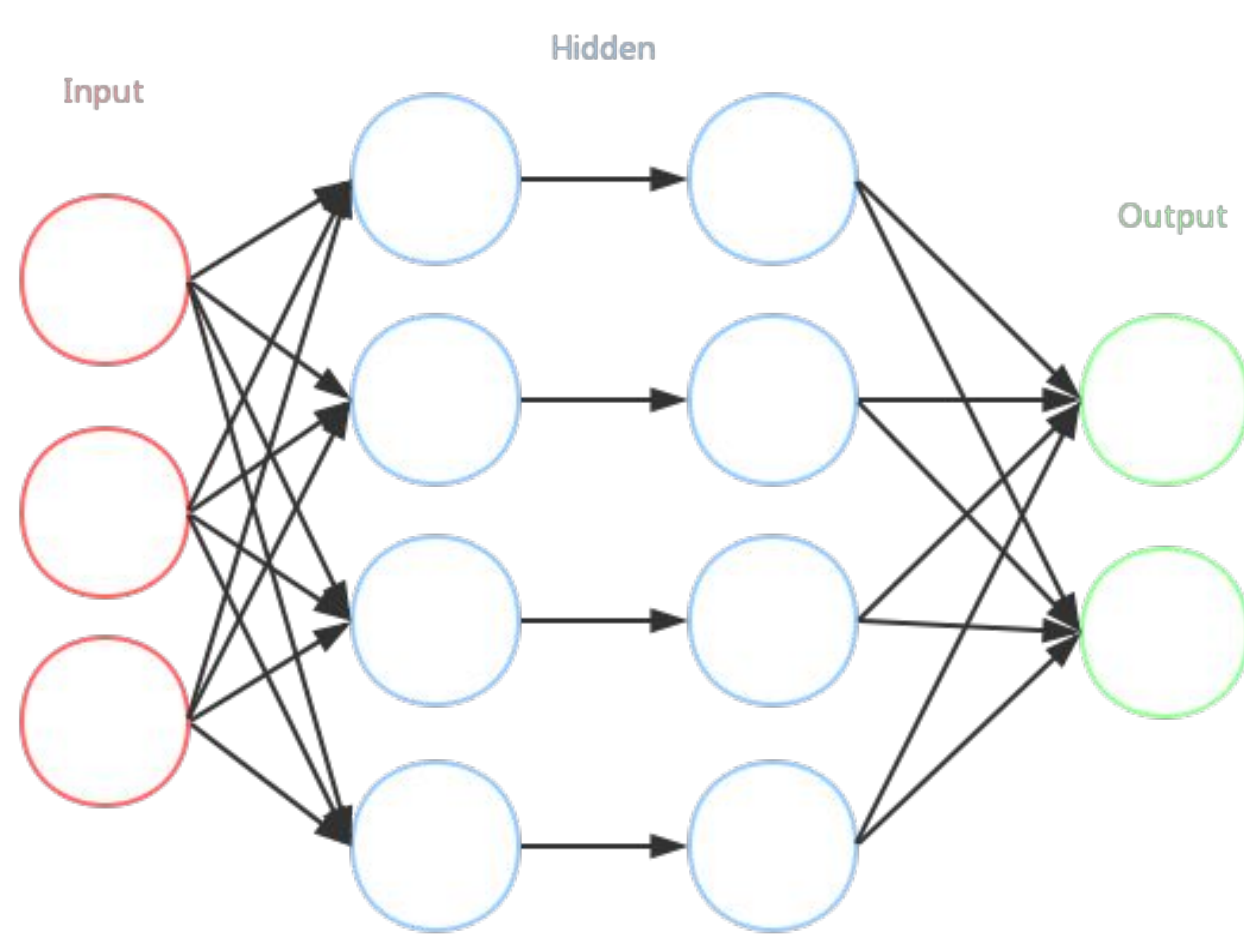


Fig 1. Sample of a Two-hidden-layer Neural Network [7]

Convolutional Neural Network

Extract mel spectrum feature from audio files and then display the power of the mel spectrum. Saving all these mel spectrum features as pictures. Then, using a 34 layer ResNet to train the model based on those pictures and classify audios based on those pictures.

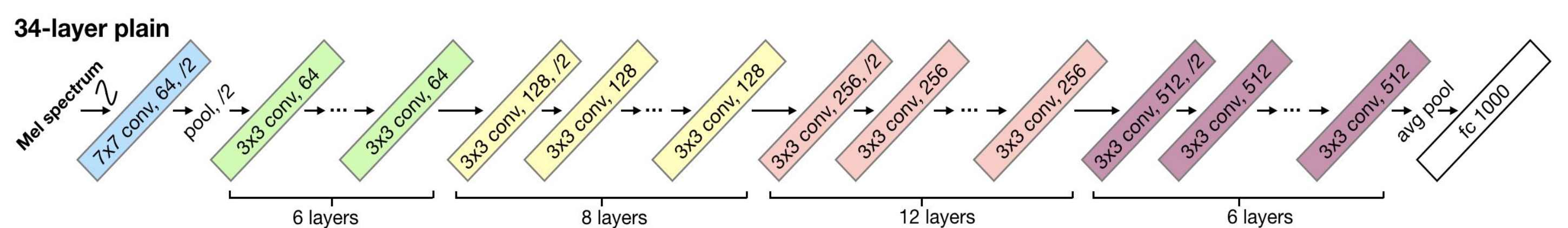


Fig 2. Sample of a 34-layer of Convolutional Neural Network [8]

Recurrent Neural Network

Extract MFCC features with overlapped windows from audio data and encodes labels of corresponding features by one-hot. Use long short term memory(LSTM) in RNN to train the model with features and labels.

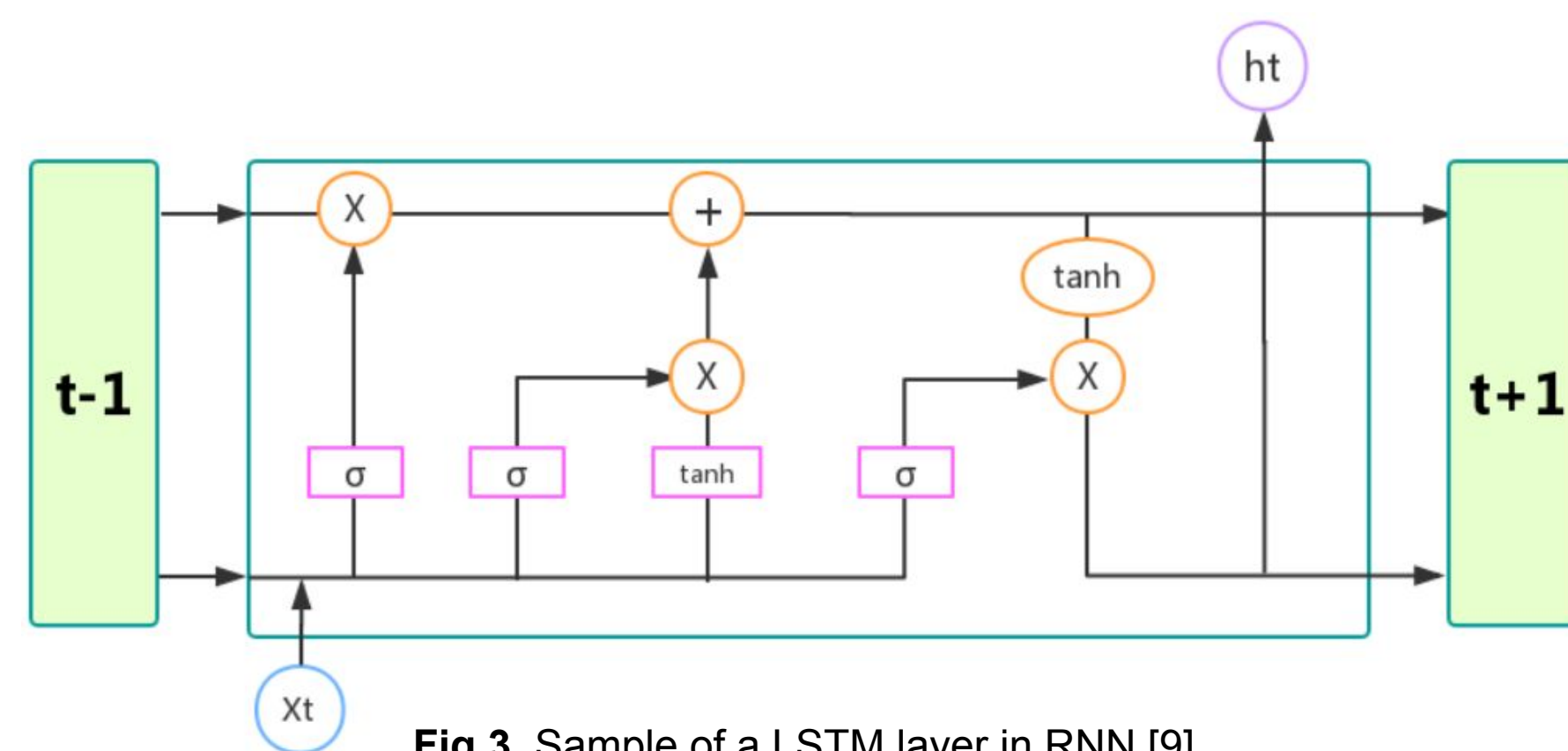


Fig 3. Sample of a LSTM layer in RNN [9]

Results

Neural Network model uses four features: Mel-frequency cepstrum, MFCC, Chroma, and Contrast. Then, a two-hidden-layer Neural Network with 4096 cells in each layer is used to make predictions.

	Training set	Testing set	Training accuracy	Testing accuracy
NN	80%	20%	99.9%	89.8%
CNN	80%	20%	96.3%	87.8%
RNN	80%	20%	91.6%	82.5%

Table2. Results of NN, CNN and RNN

Future

Ensemble all models currently. We will try to ensemble different models to improve overall performance.
 Make a modern webpage to demonstrate our result for better user interaction.

Discussion

Meaning of urban sound classification

With the help of sound classification, while giving recommendations or providing services, applications in mobile devices can not only base on location information but also based on surrounding sound information.

Also, it can improve the life quality of the disabled who suffers from hearing problems. If a device can provide a description of the surrounding environment based on the sound classification to the disabled, it will definitely help them avoid some dangers and make their life more convenient.

Meaning of the results

From the above result table, we can tell that our results are quite great. It's because we did many analysis and experiments for extracting proper features and finding suitable methods. For each feature, we try to understand their physical meanings and choose a suitable one for different using purposes. For each neural network, we explored it thoroughly and grid-searched for best hyper-parameters.

Long-term audio recognition

For audio longer than 4 seconds, we use sliding windows to divide them into 4-second pieces and recognizing those pieces to classify the entire audio. Also, we transfer to predicting results into .srt files, which means we can recognize live sound in videos and display them as captions.

References

[1] Kaggle.com. (n.d.). *Urban Sound Classification*. [online] Available at: <https://www.kaggle.com/pavansanagapati/urban-sound-classification> [Accessed 28 Apr. 2019].
 [2] M Kattel, Araju Nepal, Ayush Shah, and Dev Shrestha. Chroma feature extraction. 01 2019.
 [3] Juncheng Li, Wei Dai, Florian Metz, Shuhui Qu, and Samarjit Das. A comparison of deeplearning methods for environmental sound. CoRR, abs/1703.06902, 2017.
 [4] En.wikipedia.org. (2019). *Mel-frequency cepstrum*. [online] Available at: https://en.wikipedia.org/wiki/Mel-frequency_cepstrum [Accessed 1 Jun. 2019].
 [5] En.wikipedia.org. (2019). *Chroma feature*. [online] Available at: https://en.wikipedia.org/wiki/Chroma_feature [Accessed 1 Jun. 2019].

[6] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. Music type classification by spectral contrast feature, Aug 2002.
 [7] En.wikipedia.org. (2019). *Artificial neural network*. [online] Available at: https://en.wikipedia.org/wiki/Artificial_neural_network [Accessed 1 Jun. 2019].
 [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Deep residual learning for image recognition*. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.
 [9] Colah.github.io. (2019). *Understanding LSTM Networks -- colah's blog*. [online] Available at: <http://colah.github.io/posts/2015-08-Understanding-LSTMs/> [Accessed 3 Jun. 2019].