Deep Learning for Musical Instrument Recognition

Mingqing Yun, Jing Bi
Department of Electrical and Computer Engineering, University of Rochester, NY, 14627
{myun5,jbi5}@ur.rochester.edu

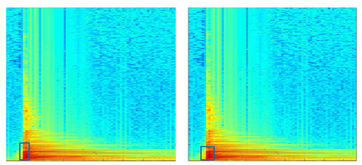
Abstract

The focus of this paper is to compare a convolutional neural network (CNN) and a recurrent neural network (RNN) in the particular task of instrument classification with log magnitude spectrogram. We first choose to use a simple but efficient CNN architecture– LeNet to verify the validity of using CNN for instrument classification. We propose a design strategy meant to capture the relevant time-frequency contexts for learning timbre, which permits using domain knowledge for designing architectures. In addition, another goal of this paper is to use one of RNN structure called Long-Short Term Memory to realize instrument recognition. After comparing different network structure, we can make a conclusion that the LENET learns faster and more accurate when doing instrument classification.

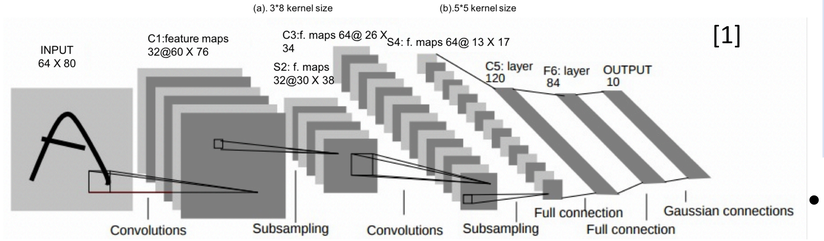
Proposed Method

- Convolutional Neural Network

$$x_l^j = \mathrm{f}\left(\sum_{i \in M_j} x_i^{l-1} * k_{ij}^l + b_j^l\right)$$

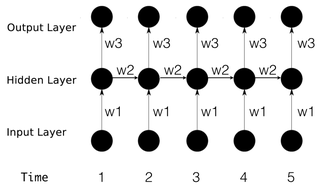


(a). 3*8 kernel size (b).5*5 kernel size

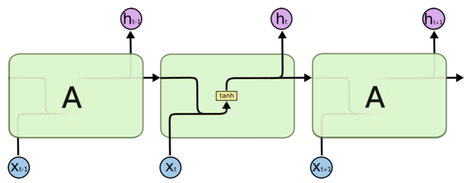


[1]

- Recurrent neural network

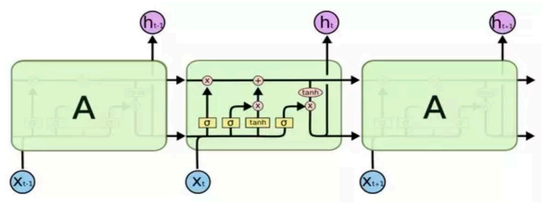


[2]



Vanishing Gradient Problem

- Long Short-Term Memory Recurrent Neural Networks



[3]

Experiment

Step 1: Collect dataset

- 14 instruments out of 25 with 200 training and 120 test audio .
- Each audio is trimmed into 1 second.
- The start point of the clip is choose from the maximum of the derivative of the signal power.

Step 2: Preprocess

- Compute short-time Fourier transform (STFT) of the recordings with 1024 fft length and 50% overlap.
- Filterbank with 64 and 128 bands spanning 0 to 22050Hz, which is the Nyquist rate. *
- Finally we computing dB relative to peak power and nominalized the data

Step 3:Training the network:

- Convolutional neural network(CNN)
- Recurrent neural network (LSTM)

Step 4: compare the result

CNN: 5*5 filter vs. 3*8 filter*

LSTM:

1. 1 layer with 64 units vs. 128 *
2. 1 layer vs. 2 layer*
3. 2 layer, the second layer 128 units vs. 64 units*

CNN vs. LSTM*

Lenet with 3*8 filter vs. 2 layer LSTM system with 128*128 units in each layer.

Result

Fig 1: filterbank with 64 and 128 bands

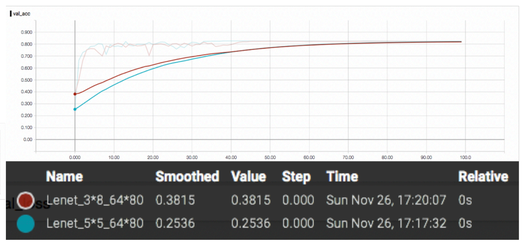


Fig 2:Lenet with 5*5 and 3*8 filter

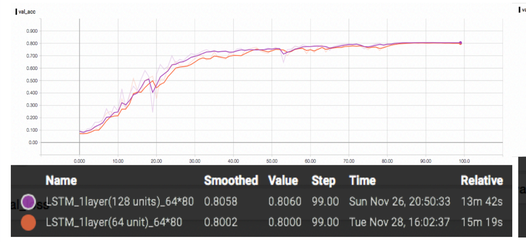


Fig 3: 1 layer LSTM with 64 and 128 units

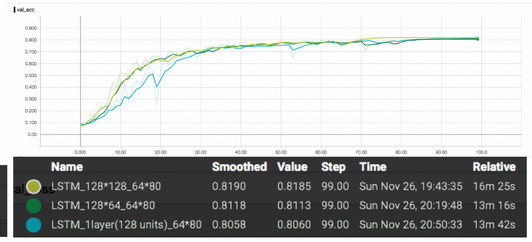


Fig 4:LSTM with 1 layer and 2 layers

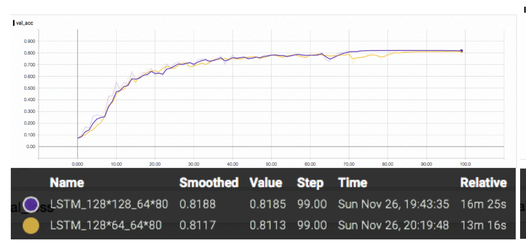


Fig 5:LSTM with 128 units and 64 units in the second layer

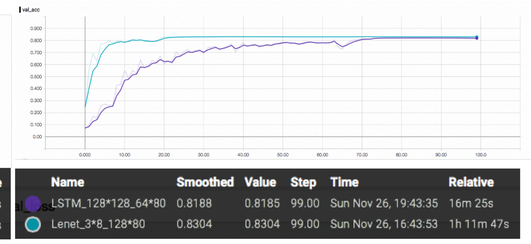


Fig 6:Lenet compare with LSTM

Conclusion

1. In a LENET system, the 3*8 kernel size can get a better solution.
2. In a LSTM system, not only the number of units, but also the number of layers can affect the system. A two layer LSTM system with 128 units in each layer gets the best solution.
3. The Lenet is more suitable with larger input size(128*80).But the LSTM get better result with 64*80 input size
4. The learning rate from Lenet are faster than from LSTM system. And after 100 epochs, the result from Lenet is a little bit higher than from LSTM system.
5. In this case, we can make a conclusion that Lenet system is more effective when classifying music instrument.

Reference

[1]Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11,
[2]D. Britz, "Recurrent Neural Networks Tutorial, Part 1 – Introduction to RNNs," *WildML*, 17-Sep-2015.
[3]Eugenio Culurciello, *7.1 Recurrent Neural Networks RNN*. .