Invariances and Data Augmentation for Supervised Music Transcription

Published in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

7107018026

統研一 劉俊廷

Contents

Methods

本文考慮的模型可以分成三大類:

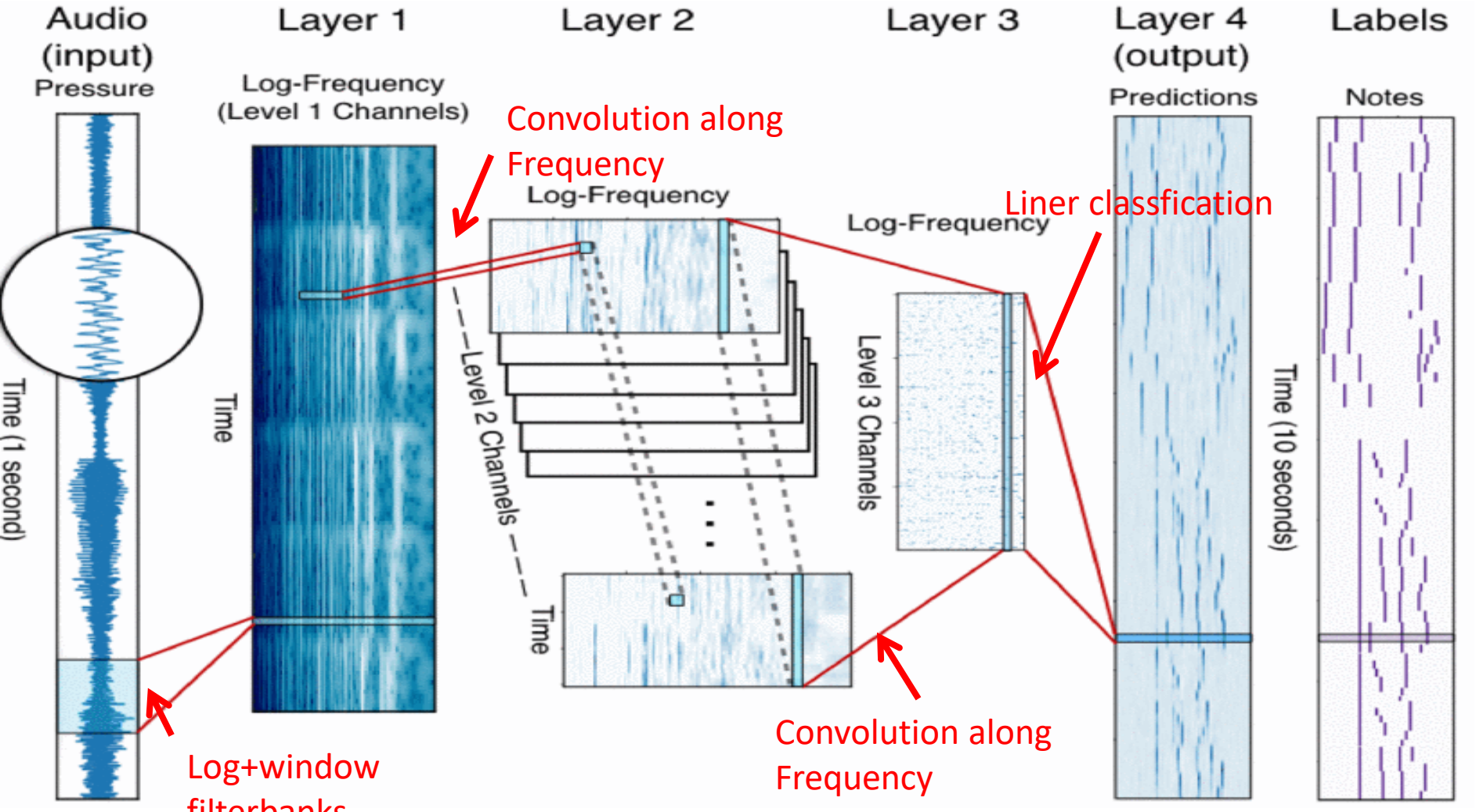
Filter bank:

概念:藉由filter去篩選特徵值，以減少網路參數

1. **Short-time Fourier transform**
2. **Log-spaced filterbank**
3. **Windowed filterbank(cosine window:1-cos(t))**

$$filter_k = \left(w_{k,\sin}^T x_t\right)^2 + \left(w_{k,cos}^T x_t\right)^2$$

Methods


Audio (input)
Pressure
Time (1 second)
Layer 1
Log-Frequency (Level 1 Channels)
Time
Convolution along Frequency
Layer 2
Log-Frequency
Level 2 Channels
Time
Layer 3
Log-Frequency
Level 3 Channels
Liner classfication
Layer 4 (output)
Predictions
Time (10 seconds)
Labels
Notes
Log+window filterbanks
Convolution along Frequency

Training

- dataset:

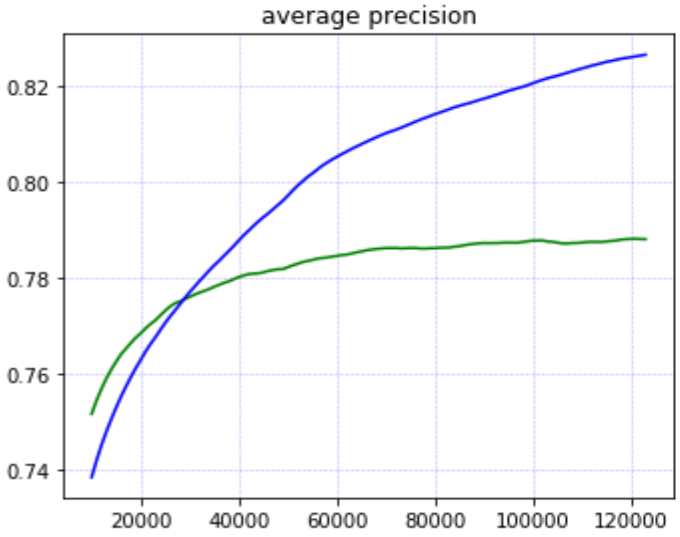
  MusicNet(331首古典音樂)
- 學習模式:

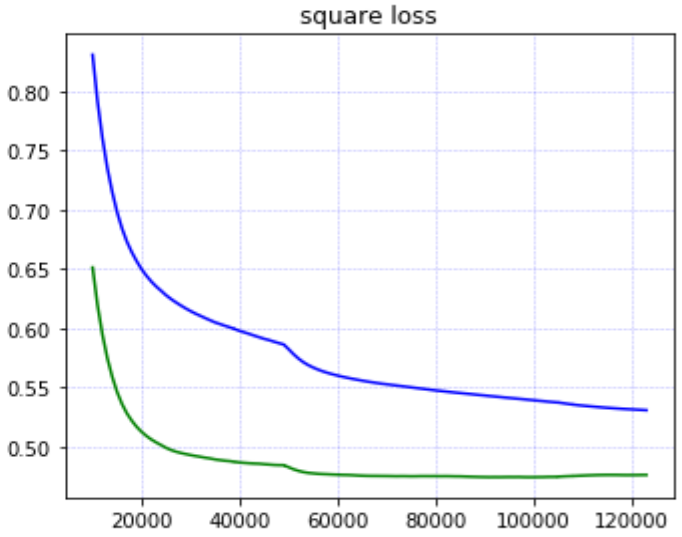
  momentum(ρ=.95)

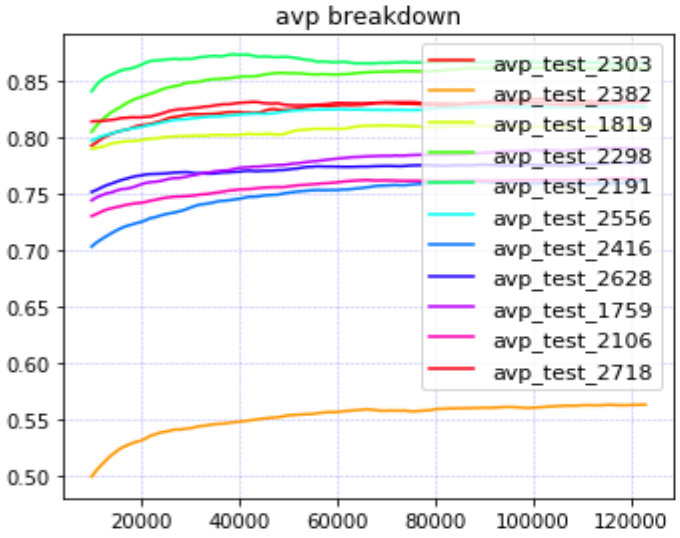
Results

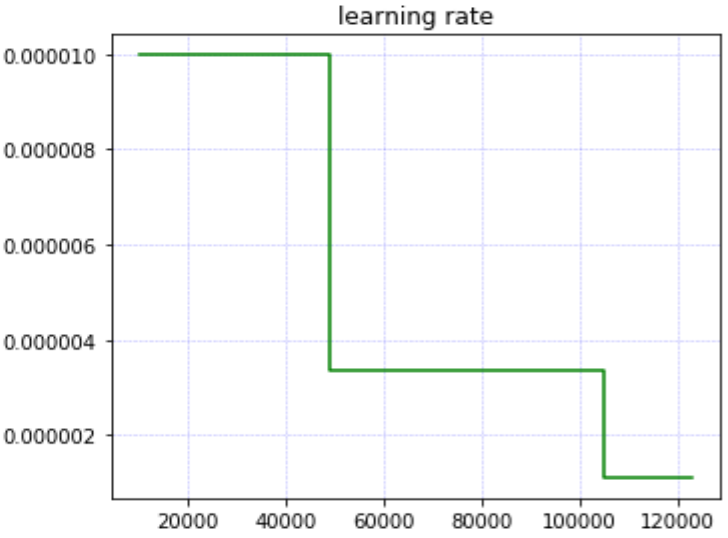
Model	Avg. Prec.	Acc.	Err.
filterbanks			
STFT (no compress)	40.4	15.9	.860
STFT	60.4	36.2	.681
log frequencies	62.7	39.8	.646
cosine windows	66.1	38.7	.637
log + windows	66.7	38.9	.633
three layer network	73.8	51.4	.541
end-to-end			
learned filterbank [7]	67.8	48.9	.634
three layer network	70.8	48.8	.558
deep complex [12]	72.9	-	-
channel convolution	73.3	50.4	.531
translation-invariant			
baseline	76.5	53.2	.496
pitch-shift	77.1	54.5	.482
wide layer 3	**77.3**	**55.3**	**.474**
commercial software			
Melodyne [22]	58.8	41.0	.760

Log+cos windows









遇到困難

1.資料太大，電腦跑不動

----------嘗試分出一部分小檔案來訓練

2.對程式的架構不太熟練

Thanks for listening