Automatic Tagging Using Deep Convolutional Neural Networks

Keunwoo.Choi
@qmul.ac.uk

Centre for Digital Music, Queen Mary University of London, UK

@keunwoochoi

11 Aug 2016, ISMIR 2016, NY

Problem definition

What is auto-tagging?

Tags

Descriptive keywords that people (just) put on music

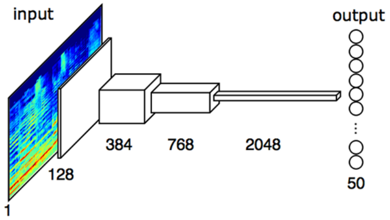
- Multi-label nature
  - E.g. {*rock, guitar, drive, 90's*}
- Music tags include Genres (rock, pop, alternative, indie), Instruments (vocalists, guitar, violin), Emotions (mellow, chill), Activities (party, drive), Eras (00's, 90's, 80's).
- Collaboratively created (Last.fm) → noisy and ill-defined (of course)
  - false negative
  - synonyms (vocal/vocals/vocalist/vocalists/voice/voices. guitar/guitars)
  - popularity bias
  - typo (harpsicord)
  - irrelevant tags (abcd, ilikeit, fav)

Problem definition

What is auto-tagging?

- Multi-label classification
- Criteria: AUC-ROC (Area Under an ROC Curve)
  - $0.5 <= \text{AUC-ROC} <= 1.0$
  - Robust to unbalanced datasets
  - Higher if lower false positive rate
  - Higher if higher true positive rate

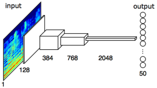
The proposed architectures



- $1 \times 96 \times 1366$ melgram $\rightarrow$ conv's/pooling's $\rightarrow$ $2048 \times 1 \times 1$
- All ReLU
- All 3x3 convolutions
- 2048 feature maps at the end
- 3,4,5,6,7 layers

Assumptions

Why (I think) would it work?



conv-MP-conv-MP-conv-MP..

- $N \times M$ Convolution: There are some useful patterns in input and feature maps that are local, location-invariant, and equal or smaller than $N \times M$.
- $L \times K$ Max-Pooling: We are generous up to $L \times K$ so we allow variances within this range.

Which means,

We see *big picture*, some macroscopic patterns

...assuming/hoping that they are related to *tag*

Experiments and discussions

	MTT	MSD
# tracks	25k	214K (out of total 1M)
# songs	5-6k	214K (out of total 1M)
Length	29.1s	30-60s
Benchmarks	10+	0
Labels	Tags, genres	Tags, genres, EchoNest features, bag-of-word lyrics,...

Experiments and discussions

For	Dataset	Specificaions
Input representation	MTT	STFT/MFCC/Melgram
# Layers	MTT	3/4/5/6/7
Benchmark	MTT	FCN-4 vs 5 previous methods
# Layers[1]	MSD	3/4/5
# Layers[2]	MSD	3/4/5, Narrower structure

[1] Different from the paper

[2] Not in the paper

- Same depth (l=4), melgram>MFCC>STFT
  - melgram: 96 mel-frequency bins
  - STFT: 128 frequency bins
  - MFCC: 90 (30 MFCC, 30 MFCCd, 30 MFCCdd)

Methods	AUC
FCN-4, mel-spectrogram	**.894**
FCN-4, STFT	.846
FCN-4, MFCC	.862

- Still, ConvNet may outperform frequency aggregation than mel-frequency (if there's more data). But not yet.
- ConvNet outperformed MFCC

Experiments and discussions

MagnaTagATune - Number of layers

Methods	AUC
FCN-3, mel-spectrogram	.852
FCN-4, mel-spectrogram	**.894**
FCN-5, mel-spectrogram	.890
FCN-4, STFT	.846
FCN-4, MFCC	.862

- FCN-4>FCN-3: Depth worked!
- FCN-4>FCN-5 by .004
  - Deeper model might make it equal after ages of training
  - Deeper models requires more data
  - Deeper models take more time *(deep residual network[4])*
  - *4 layers are enough vs. matter of size(data)*?

Experiments and discussions

MagnaTagATune

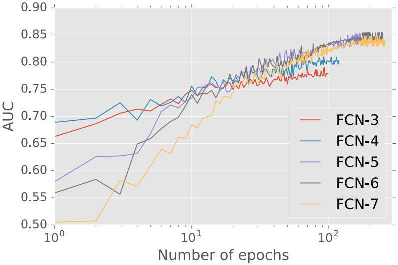
Methods	AUC
The proposed system, FCN-4	.894
2015, Bag of features and RBM [5]	.888
2014, 1-D convolutions[2]	.882
2014, Transferred learning [6]	.88
2012, Multi-scale approach [1]	.898
2011, Pooling MFCC [3]	.861

- All deep and NN approaches are around .88-.89
- Are we touching the glass ceiling?
  - Perhaps due to the noise of MTT, but tricky to prove it
  - 26K tracks are not enough for millions of parameters

Experiments and discussions

Million Song Dataset - on the paper

Methods	AUC
FCN-3, mel-spectrogram	.786
FCN-4, —	.808
FCN-5, —	.848
FCN-6, —	**.851**
FCN-7, —	.845



WARNING!

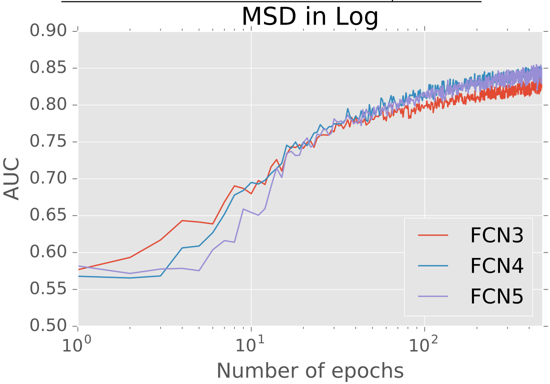
- The MSD results are not reproduced.
    - I suspect a incorrect learning rate controlling
        - and this is why we shouldn't rush before deadline..
- Ran the experiments again
    - without weird learning rate controlling,
    - and more epochs (240→480)

Experiments and discussions

Million Song Dataset - re-run

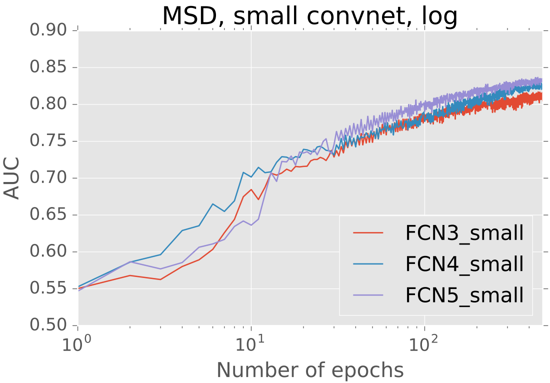
Keunwoo.Choi @qmul.ac.uk

Methods	AUC
FCN-3, mel-spectrogram	.839
FCN-4, —	.852
FCN-5, —	.855



Smaller (narrower) convnet

No. of feature maps: [128@1 – 2048@5] $\rightarrow$ [32@1 – 256@5], i.e.*narrower* network, because there's no difference between FCN-4 and FCN-5.



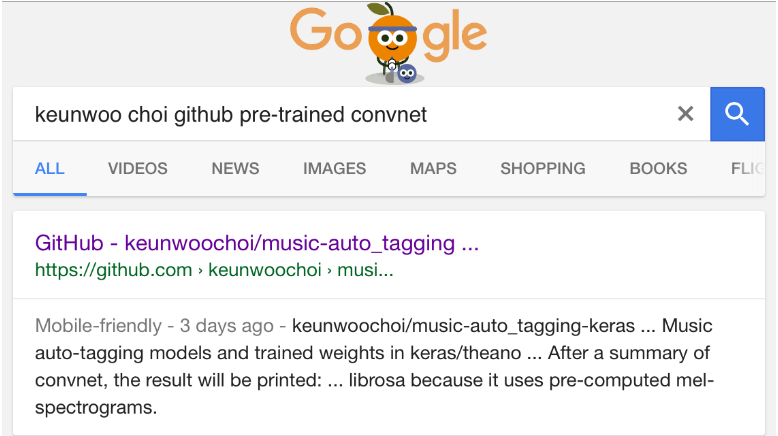
Conclusions

- Assumptions - about macroscopic view seems fine
- In general, the behaviour agrees with computer vision community, which are..
  - the deeper, the better (or equal)
  - the wider, the better (or equal), but not as much as depth
- Melgram+feature learning > MFCC
- Melgram > STFT
  - At some point, we will argue STFT + learning > melgram
- MTT is too small, even MSD might be small
- Future work: More investigation, variable input length, better dataset, re-thinking the problem...

Thank you for listening and...

You can *plug-and-predict*

The pre-trained weights and model is open!



https://github.com/keunwoochoi/music-auto_tagging-keras

References I

- Dieleman, S., Schrauwen, B.: Multiscale approaches to music audio feature learning. In: ISMIR. pp. 3–8 (2013)
- Dieleman, S., Schrauwen, B.: End-to-end learning for music audio. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. pp. 6964–6968. IEEE (2014)
- Hamel, P., Lemieux, S., Bengio, Y., Eck, D.: Temporal pooling and multiscale learning for automatic annotation and ranking of music audio. In: ISMIR. pp. 729–734 (2011)
- He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)

References II

Nam, J., Herrera, J., Lee, K.: A deep bag-of-features model for music auto-tagging. arXiv preprint arXiv:1508.04999 (2015)

Van Den Oord, A., Dieleman, S., Schrauwen, B.: Transfer learning by supervised pre-training for audio-based music classification. In: Conference of the International Society for Music Information Retrieval (ISMIR 2014) (2014)