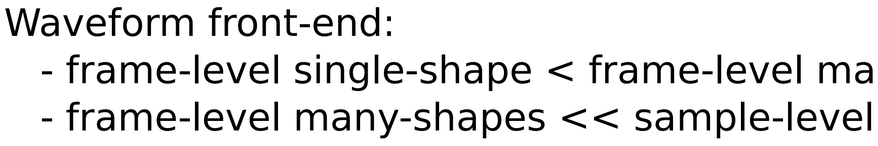
Music Technology Group
EXCELENCIA MARÍA DE MAEZTU
pandora®

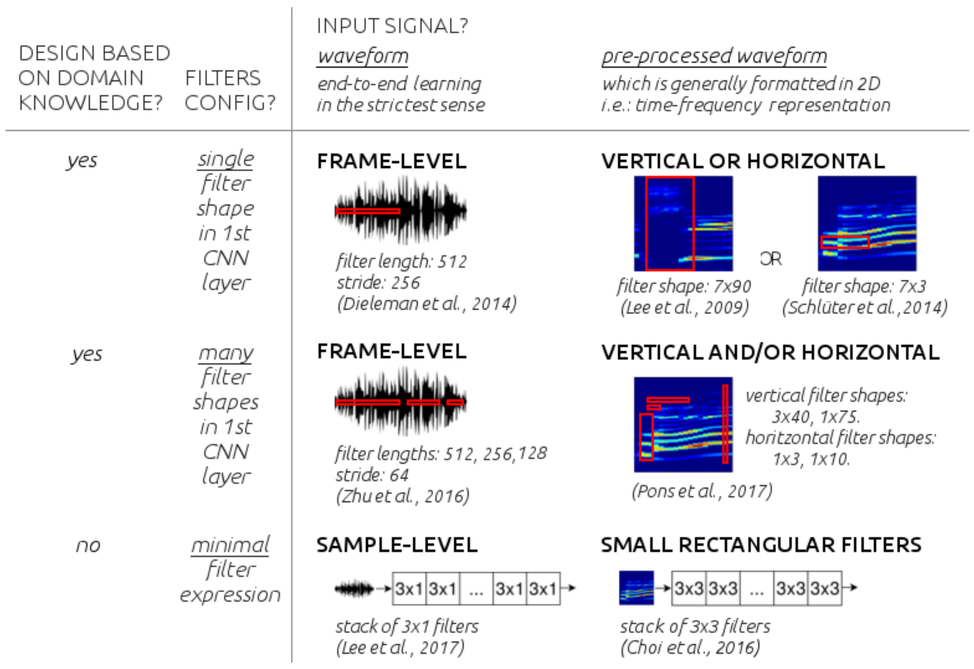
**Jordi Pons, Oriol Nieto, Matt Prockup,
Erik Schmidt, Andreas Ehmann and Xavier Serra**
contact: jordipons.me - @jordiponsdotme

End-to-end learning for music audio tagging at scale

Which deep learning architecture shall we use for audio tagging?



Many CNN frontends!



Waveform front-end:
- frame-level single-shape < frame-level many-shapes
- frame-level many-shapes << sample-level

Specrogram front-end:
- domain knowledge intuitions are a valid guide for designing your front-end

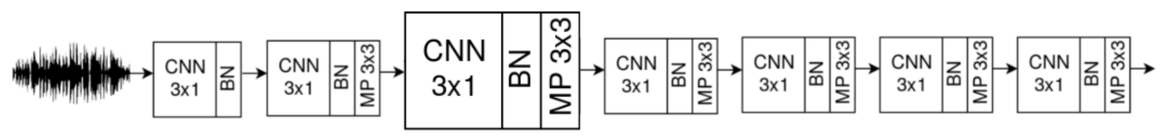
How much data is available?

A human-annotated corpus of **1.5M songs**

Can waveform front-ends achieve better performance than spectrogram front-ends?

Waveform front-end

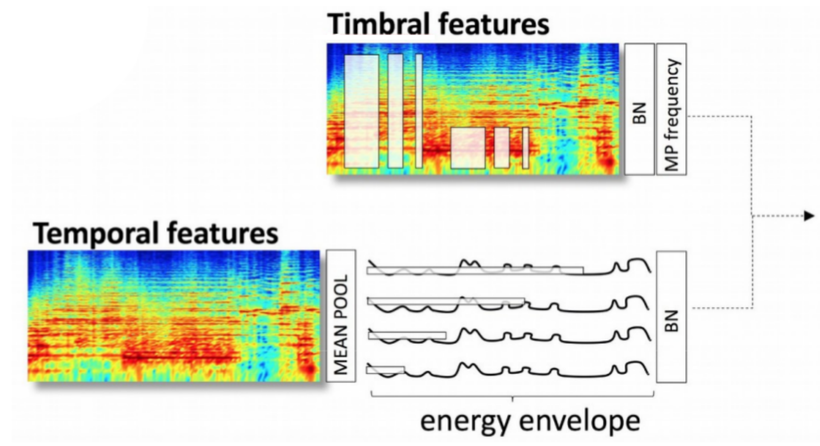
(Lee et al, 2017) an assumption-free model



#filters in every layer: 64, 64, 64, 128, 128, 128, 256.
waveform down-sampled to 16kHz.

Spectrogram front-end

(Pons et al, 2017) heavily based on domain knowledge



Many possible backends!

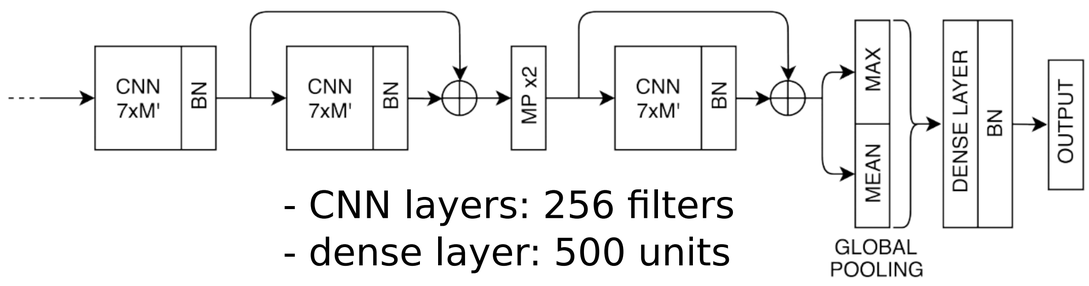
Variable-length input back-end:
- max and average pooling
- attention models
- RNNs

Fixed-length input back-end:
- fully convolutional models
- DNNs

Note: music is of variable length!

But: most models assume a fixed-length input!

(Dieleman et al, 2014) Shared back-end



- CNN layers: 256 filters
- dense layer: 500 units

Quantitative results

Models	*#training examples*	*ROC AUC*	*PR AUC*	$\sqrt{MSE}$	Δ*ROC AUC*	Δ*PR AUC*	$\Delta\sqrt{MSE}$	training time
GBT+features	1.2M	91.61%	54.27%	0.1569	-	-	-	-
Waveform	1M	91.54%	57.86%	0.1501	0.6%	1.49%	0.0021	<2 weeks
Spectrogram	1M	**92.14%**	**59.35%**	**0.1480**				
Waveform	500k	91.23%	56.15%	0.1537	0.54%	1.75%	0.0044	≈ 1 week
Spectrogram	500k	91.76%	57.90%	0.1493				
Waveform	100k	89.16%	49.25%	0.1591	0.97%	2.83%	0.0049	few days
Spectrogram	100k	90.13%	52.08%	0.1542				

Audio segments of 15 sec.
Song-level predictions:
- averaging windowed predicions

Annotations, 2 distributions:
- bi-modal, classification tags
  ROC-AUC and PR-AUC
- uniform, regression tags
  Error

Qualitative results

Bias towards predicting popular tags
"lead vocals", "English" or "male vocals"

Predicting each tag independently vs. predicting all tags together
"East Coast", "West Coast" / "baroque period", "classic period"

Reproduce this experiment online:
jordipons.me/apps/music-audio-tagging-at-scale-demo

Conclusions

Better performance than GBT+features baseline

spectrogram front-ends > waveform front-end
..but the gap has been reduced!
with more training data and Lee et al. front-end

Models' implementation in tensorflow:
github.com/jordipons/music-audio-tagging-at-scale-models