Designing efficient architectures for modeling temporal features with CNNs

https://github.com/jordipons/ICASSP2017

Jordi Pons and Xavier Serra

Music Technology Group, Universitat Pompeu Fabra, Barcelona
jordi.pons@upf.edu - www.jordipons.me - @jordiponsdotme

Introductory discussion

Use case: music classification

- **Approach**: single-layer CNNs.
- **Input**: log-mel spectrograms.
- **Dataset**: Ballroom. Interesting because:
  - Small dataset size.
  - Tempo and rhythm features.

Basic observations

- Filters dimensions **interpretable**.
- **Unreasonable assumption**:
  - Large data for training large models.
- Influence from the **computer vision** field:
  - *seeing* spectrograms.
  - 3x3 filters – ***limit*** the representational power of the first layer!
  - *Seeing* a semantically irrelevant context.
  - Need to combine. Hebbian principle:

Proposal: *design strategy*

Motivation

- Domain knowledge available.
- Intuition → interpretability.
- Semantically relevant contexts are expressed in different scales.
- Minimize CNN number of parameters:
  - Efficiency.
  - Over-fitting.

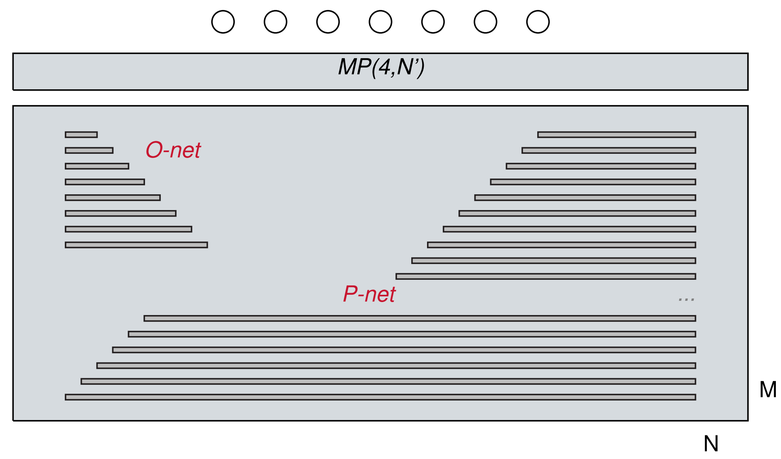
Design strategy

Promotes an ***efficient use*** of the representational capacity of the first layer by using ***different*** musically motivated filter shapes that model several contexts.

Underlying hypothesis

CNNs can benefit from a design oriented towards learning musical features rather than *seeing* spectrograms.

Architecture: *1-by-n filters*



- **O**(nsets)**-net** (5x)
  → where $n \in [6, 11, 16, 21, 26, 31, 36, 41]$.
- **P**(atterns)**-net** (1x)
  → where $46 \leq n \leq 216$ ($n \in 46 + 5 \cdot f$).
- **Max-pool layer** interpretation:
  → sub-bands analysis.

STFT-window: 2048 (50% hop) at 44.1 kHz:

$$n_O \equiv \Delta Fr|_{bpm=60} = \frac{44100 \times 60_{(sec)}}{60_{(bpm)} \times 2048 \times 0.5} = 43$$

$$n_P \equiv 1 + 5 \cdot \Delta Fr|_{bpm=60} = 216$$

Tempo $\in [60, 224]$ M=40 N=250

Results

Model:	hop	# params	accuracy	Model:	hop	# params	accuracy
O-net	250/80	4,188	76.66/85.24 %	*4x O-net + 4x P-net*	250/80	46,408	88.82/91.55 %
P-net	250/80	7,428	83.95/89.26 %	*8x O-net + 8x P-net*	250/80	92,808	88.68/**92.27** %
2x O-net	250/80	8,368	81.53/86.54 %	*Time-freq* [2]	80	196,816	**87.68** %
O-net + P-net	250/80	11,608	87.25/89.68 %	*Time* [2]	80	7,336	81.79 %
2x P-net	250/80	14,848	85.67/89.11 %	*Black-box* [2]	80	3,275,312	**87.25** %
2x O-net + 2x P-net	250/80	23,208	87.25/91.27 %	*Marchand et al.* [1]	-	-	**96** %

Fair comparison: → When setting *hop* = *N* = 250, no input spectrograms overlap is used as in [2].
→ When setting *hop* = 80 as many training examples as in [2], although overlapping data is used.

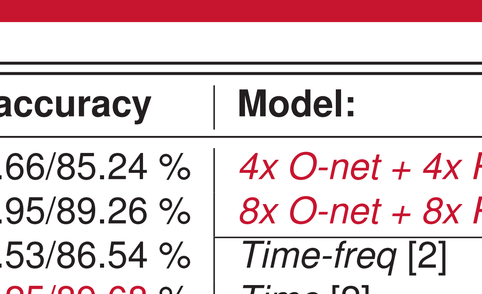
Conclusions

- ***O-net + P-net***: **best** (presented) models.
  → having the most diverse combination of filter shapes.
  → **validates design strategy!**
  → **efficient adaptation of CNNs for music!**
- ⇓ data – ⇑ useful the design strategy.
- Adding different filter shapes:
  → increasing the representational capacity of the first layer at a very **low cost**.
  → cheaper than *2x* CNN's capacity.
- **P-net** most succesful ones!
  → as was **intuitively** designed.
  → step towards **interpretability of CNNs!**

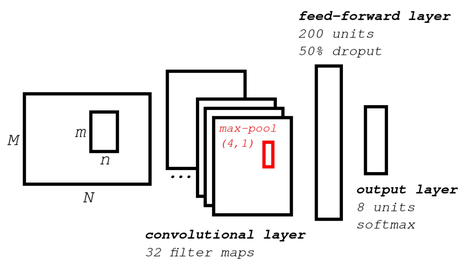
Experimental constraints
→ Only 1-by-n filters are used
→ Only a single layer is used.
→ Limited amount of data.

Baselines

- *Time-frequency* and *Time*:

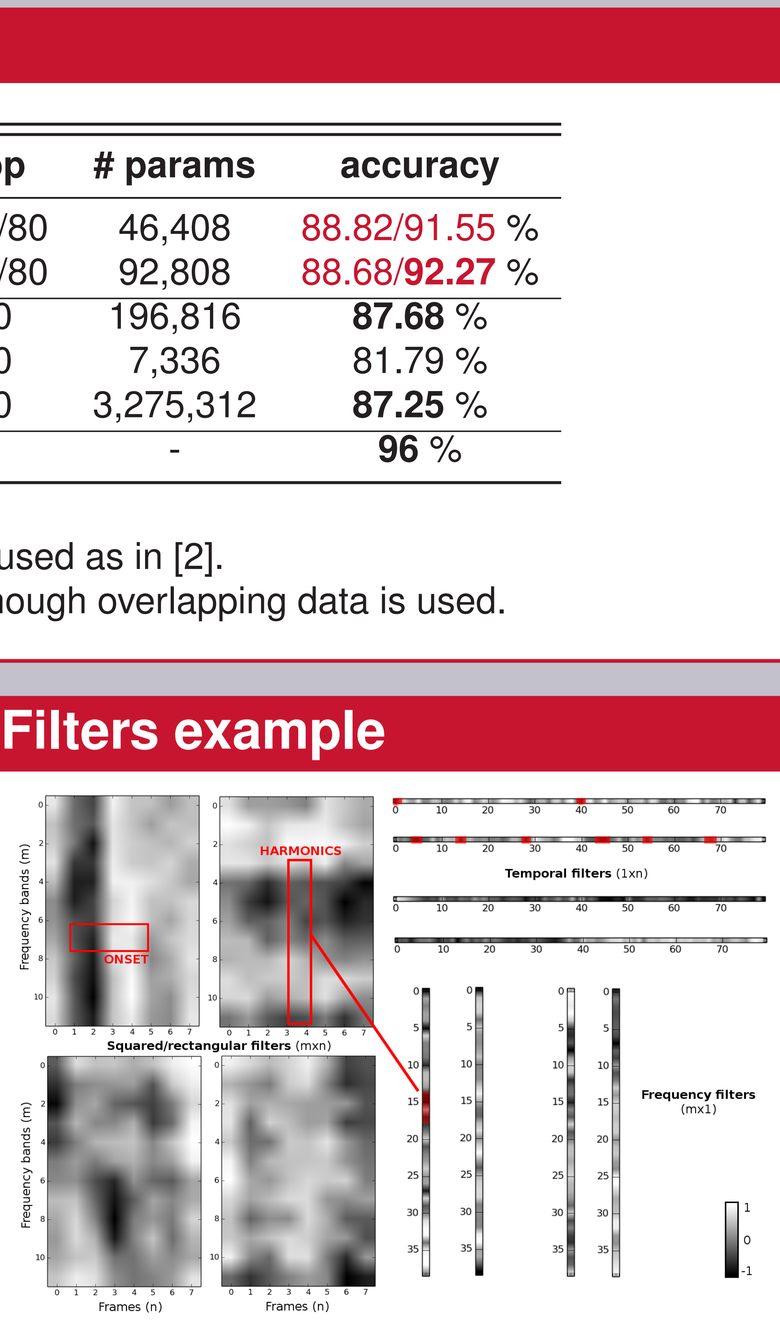


- *Black-box*: 12×8 filters and MP(4,1).



- Marchand *et al.*: based on a scale and shift invariant time/frequency representation that uses auditory statistics, not deep learning.

Filters example



References

[1] Pons *et al.* ***Experimenting with musically motivated CNNs***. CBMI, 2016.

[2] Marchand *et al.* ***Scale and shift invariant time/frequency representation using auditory statistics: application to rhythm description***. MLSP, 2016.