Towards a grounded deep learning paradigm for music modeling

Jordi Pons

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

Motivation

Scope: deep learning for music signals

- **Influence from the computer vision** field:
  (i) assume similar architectures: CNNs.
  → efficient filters configuration?
  (ii) assume similar hierarchy of concepts
  → *frequency*: note, chord; note, motive.
  → *time*: onset, rhythm; onset, tempo.
  (iii) assume *seeing* spectrograms
  → phase is not considered.
  → goal is machine listening!
- **Better performance?** By ***fully exploiting the capacity*** of deep learning for music.
  (i) *Waveforms*: end-to-end learning.
  (ii) *Spectrograms*
- **Better understanding?** By introducing some ***intuition*** during the design process.

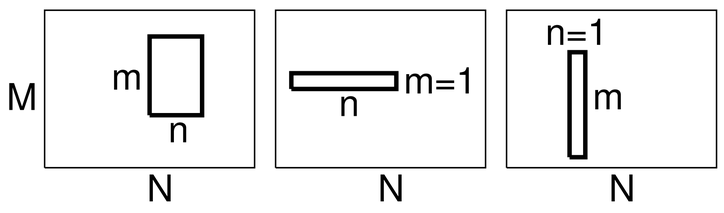
Filter shapes

Scope: music spectrograms and CNNs

- **Musically motivated filter shapes** [1]
  (i) **Squared/rectangular filters** (m-by-n)
  - kick, notes: $m \ll M$ and $n \ll N$.
  - snare, cymbals: $m = M$ and $n \ll N$.
  - music motives: $m < M$ and $n < N$.

  (also chords, harmonic/rhythmic patterns)
  → *CQT: filters are pitch and time invariant.*
  (ii) **Temporal filters** (1-by-n)
  - onsets, patterns. *...very efficient!*

  (iii) **Frequency filters** (m-by-1)
  - timbre, chords. *... NMF?*



Discussion in agreement with other studies [3]: auralisation of 3x3 filters, genre classification.

- ***Layer 1***: onsets.
- ***Layer 2***: onsets, bass, harmonics, melody.
- ***Layer 3***: onsets, melody, kick, percussion.
- ***Layer 4***: harmonic structures, notes, vertical lines, long horizontal lines.
- ***Layer 5***: textures, harmo-rhythmic patterns.

- **Impact of using small squared filters?**
  - Small rectangular filters can **limit** the representational power of **the first layer**.
  - **Non-musical hierarchy** of concepts.
    → maybe a shapes combination?

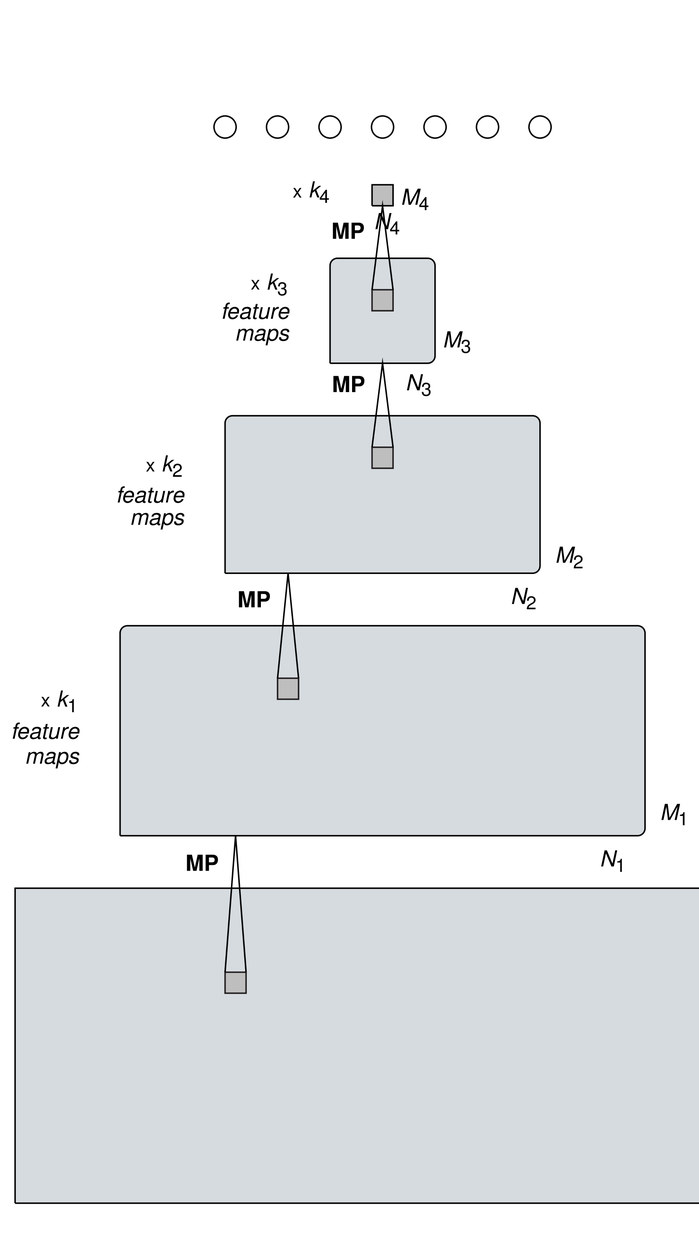
References

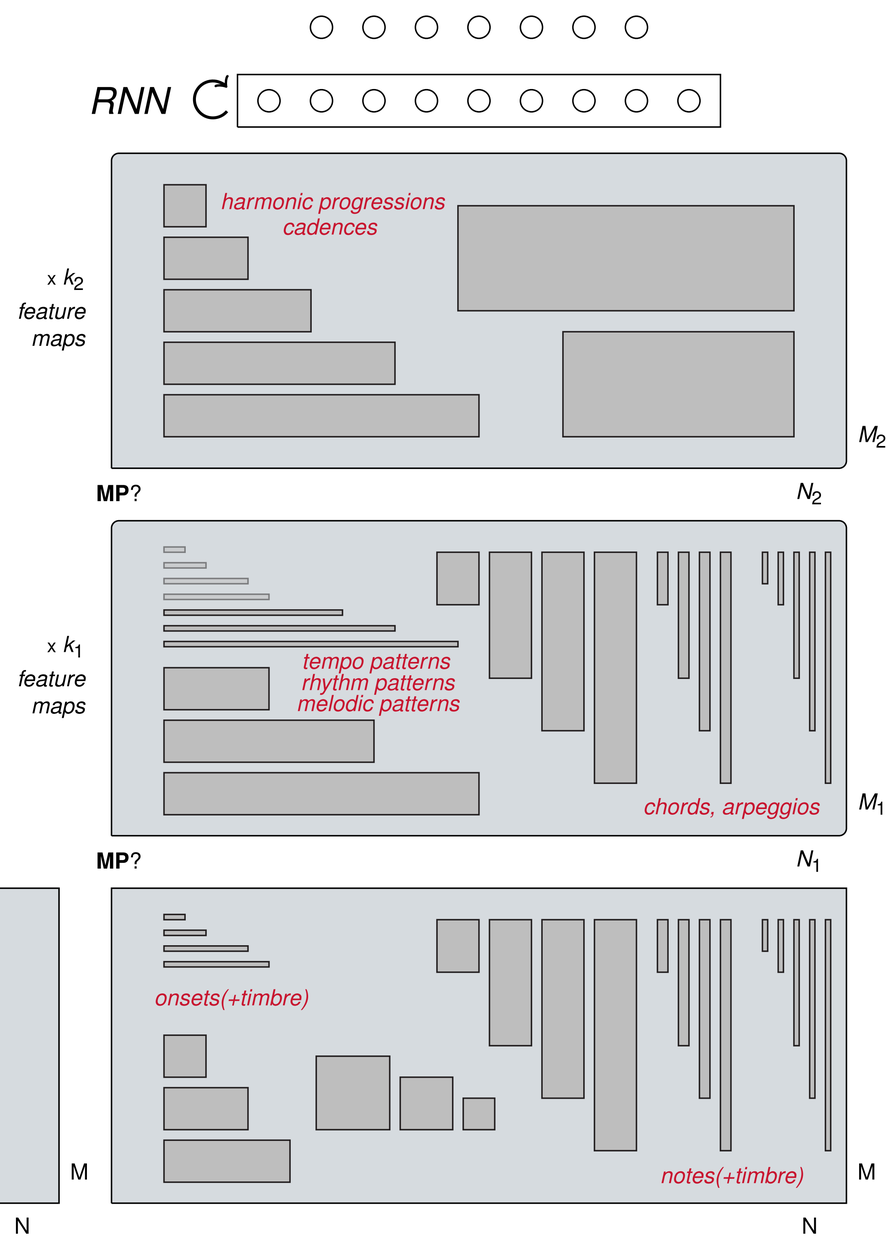
[1] Pons *et al. Experimenting with musically motivated CNNs*. CBMI, 2016.

[2] Pons *et al. Designing efficient architectures for modeling temporal features with CNNs*. ICASSP, 2017.

[3] Choi *et al. Auralisation of deep CNNs: listening to learned features*. ISMIR, 2015.

Architectures comparison





Architectures

Scope: deep learning architectures for modeling music spectrograms

- **Efficiently exploiting the representational power of the first layer** [2]

  Observations:
  (i) more efficient and interpretable.
  (ii) Hebbian principle.

  Proposal:
  → Many different shapes in the first layer.
  → Efficient modeling of different contexts.
  → Interesting also for 1D data: waveforms.

- **Considering music hierarchy**
  - ***Layer 1***: onsets(+timbre), notes(+timbre).
  - ***Layer 2***: rhythm/tempo patterns, chords, arpeggios, melodic patterns.
  - ***Layer 3***: cadences, harmonic progress.
  - ***Layer 4***: structure.
  - ***Layer 5***: *classifier*.

- **Modeling music structure with RNNs**
  - CNNs can model short time-scale.
  - RNNs can model short/long time-scale.
    → input the whole song!
  - CNN + RNN - *Layers 1, 2, 3 + Layer 4*

- **Implications?**
  - Potentially **more expressive** models.
  - **Intuitive**, more understandable.

Waveforms

Scope: end-to-end learning from raw data

- **Advantages of using waveforms?**
  - Directly considering the music signal.
    → allows minimizing the assumptions.
  - Recent studies *show* that it is feasible:
    → *Wavenet*, Van den Oord *et al.*
    → *Soundnet*, Torralba *et al.*

- **Disadvantages of using waveforms?**
  - Data demanding.
  - Computationally demanding.

- **Considered applications:**
  - Classification.
  - Speech denoising.
  - Source separation.
    → currently: phase information from the original mixture to estimate a source.

Filters example