Downbeat Tracking with Tempo Invariant Convolutional Neural Networks

Bruno Di Giorgi bdigiorgi@apple.com, Matthias Mauch mmauch@apple.com, Mark Levy mark_levy@apple.com

Abstract

The human ability to track musical downbeats is robust to changes in tempo, and it extends to tempi never previously encountered.

We propose a deterministic time-warping operation that enables this skill in a Convolutional Neural Network (CNN) by allowing it to learn rhythmic patterns independently of tempo.

Conventional deep learning approaches learn rhythmic patterns at the tempi present in the training dataset.

The patterns learned in our model are tempo-invariant, leading to better tempo generalisation and more efficient usage of the network capacity.

Contributions

The aim of this work is to factorise the space of rhythmic music signals into tempo and tempo-invariant representations in a way that is amenable for machine learning.

Technical contributions:

- The introduction of a scale-invariant convolutional layer that learns temporal patterns irrespective of their scale
- The application of the scale-invariant convolutional layer to CNN-based downbeat tracking to explicitly learn tempo-invariant rhythmic patterns

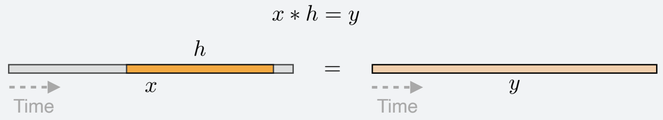
Advantages:

- Better generalisation
- Lower capacity requirements
- No data augmentation needed

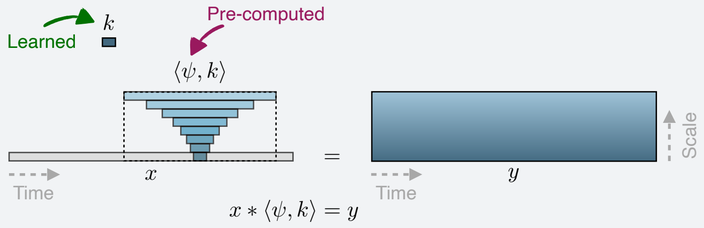
Method

Scale-invariant convolution

Standard convolution

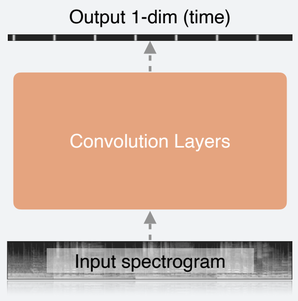


Scale-invariant convolution

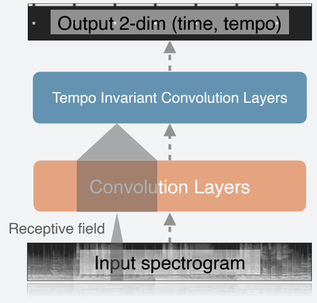


Network

Standard network



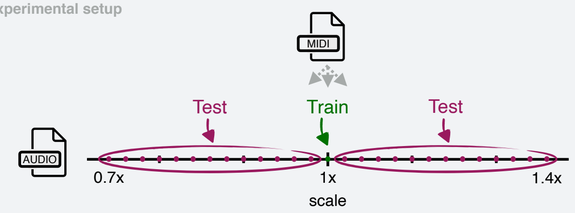
Tempo-invariant network



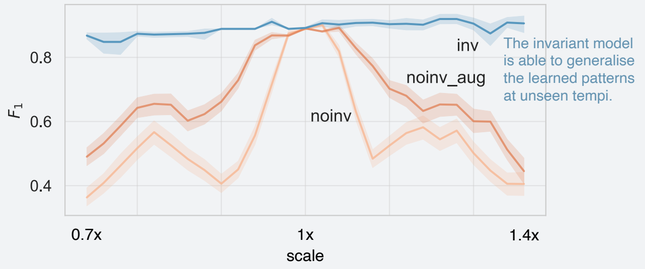
Results

Validation on synthesised drum patterns

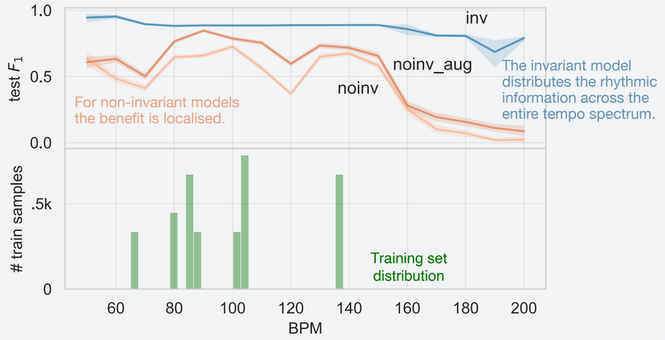
Experimental setup



Results in terms of scale



Results in terms of tempo



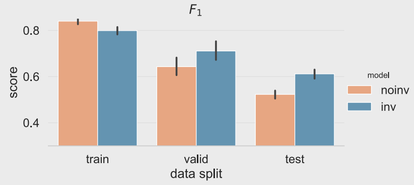
Validation on music datasets

Datasets:

- Train: Internal, RWC
- Test: Ballroom, Gtzan

Advantages with respect to the non invariant architecture:

- Better generalisation over unseen data
- Efficient usage of the network capacity, half as many parameters required



Conclusions

The proposed architecture:

- Generalises to unseen tempi by design
- Achieves higher accuracy compared to a standard CNN architecture
- Requires no data-augmentation, while implicitly providing its advantages

The learned rhythmic patterns are tempo-invariant:

- Efficient usage of the network capacity
- Potential application for rhythm classification.