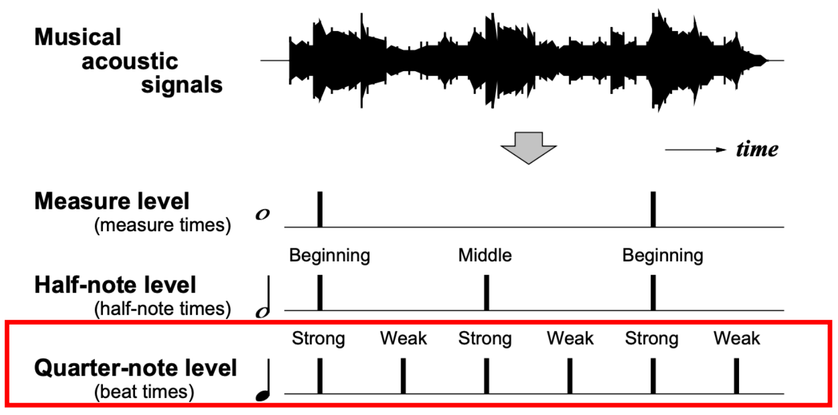
Transformer-Based Beat Tracking with Low-Resolution Encoder and High-Resolution Decoder

Tian Cheng and Masataka Goto
National Institute of Advanced Industrial Science and Technology (AIST), Japan

1. Introduction

We address the beat tracking task which is to predict beat times corresponding to the input audio.



[Goto&Muraoka1997]

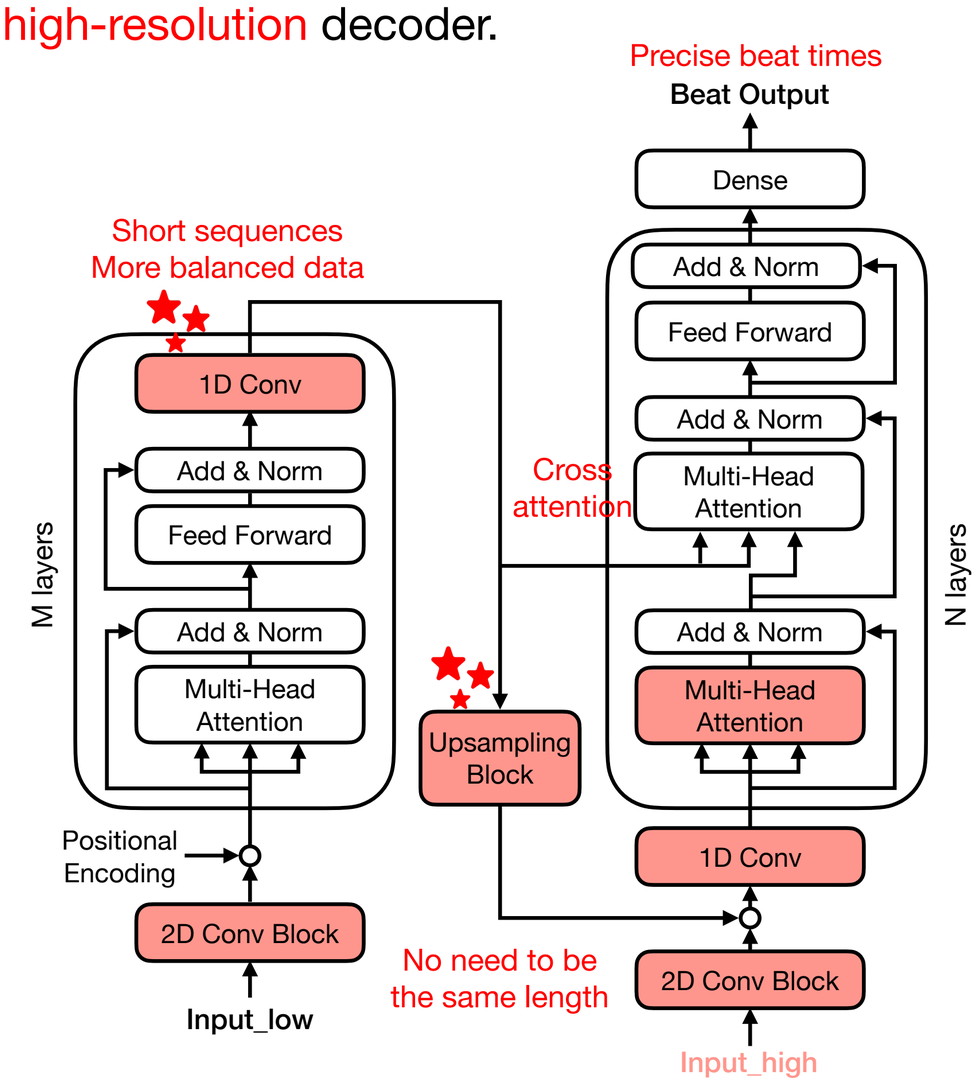
Motivation

- To produce good results, the model needs to consider both local timing and global consistency.
- This brings a contradiction on choosing the temporal resolution.

Low resolution	short sequences, more balanced data, no precise beat times
High resolution	long sequences, imbalanced data, precise beat times

2. Proposed model

A novel beat tracking model based on the Transformer with low-resolution encoder and high-resolution decoder.



Main modifications

1D Conv & Upsampling Block

In comparison to previous models,

our model uses both the encoder and decoder.

- Multi-scale features
- A more reasonable resolution for sequence modelling

3. Experiment

Data augmentation

- Tempo-wise
- Triple data using HPSS

Training

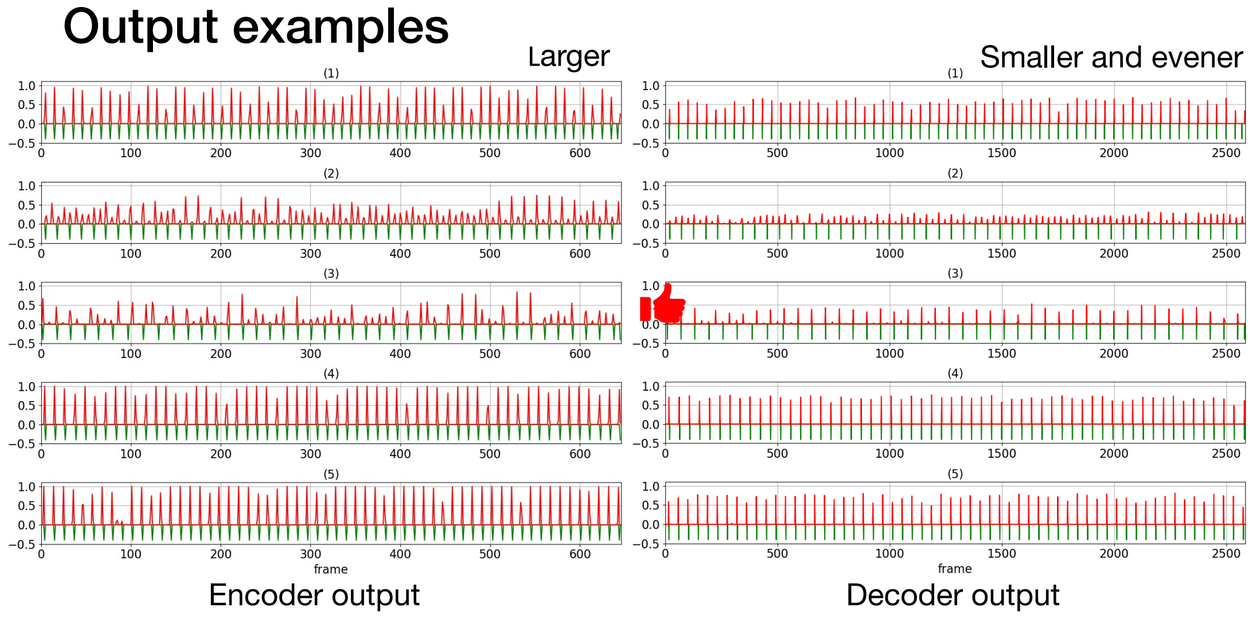
- Train the model with the pre-trained encoder

Results

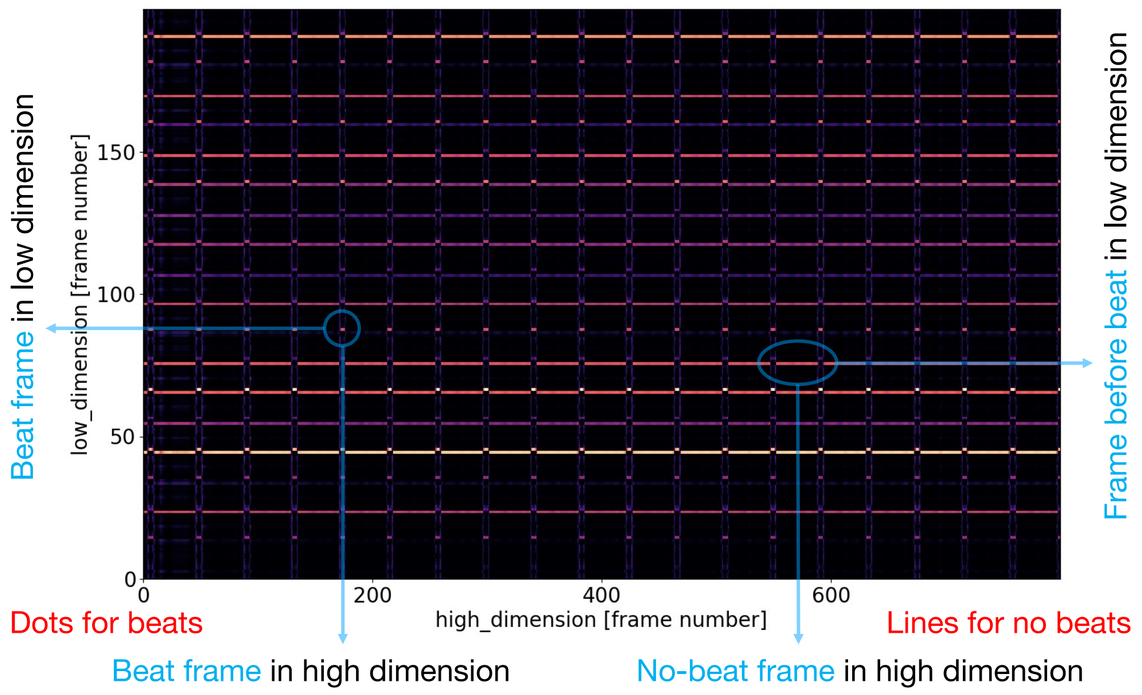
- Impressive results on the encoder by thresholding only.
- The decoder improves results of the Ballroom and GTZAN datasets.
- Post-processing (DBN) improves results especially for continuity-based ones.

Method	F-measure	CMLt	AMLt
Dataset: Ballroom			
Encoder (Th)	90.7	80.1	85.7
Encoder	93	87.4	96.1
Decoder (Proposed)	95	91.1	96.4
Beat trans [9]	96.8	95.4	96.6
TF trans [8]	96.2	93.9	96.7
TCN [7]	96.2	94.7	96.1
Dataset: Hainsworth			
Encoder (Th)	84.4	66.7	81.8
Encoder	88.2	81	93.4
Decoder (Proposed)	87	76.2	93.6
Beat trans [9]	90.2	84.2	91.8
TF trans [8]	87.7	86.2	91.5
TCN [7]	90.4	85.1	93.7
Dataset: SMC			
Encoder (Th)	53.9	32.9	45.6
Encoder	55	45.8	64.1
Decoder (Proposed)	55.4	45.1	65.6
Beat trans [9]	59.6	45.6	63.5
TF trans [8]	60.5	51.4	66.3
TCN [7]	55.2	46.5	64.3
Dataset: GTZAN			
Encoder (Th)	87.1	72.8	85.5
Encoder	87.8	78.5	93.7
Decoder (Proposed)	88.4	80.8	94
Beat trans [9]	88.5	80	92.2
TF trans [8]	88.7	81.2	92
TCN [7]	88.5	81.3	93.1

Output examples



Cross Attention Visualisation



4. Conclusions

- We present a novel Transformer-based model for beat tracking with the encoder and decoder of different resolutions.
- It provides a new framework for handling multi-scale features and learns features jointly by the cross attention in the decoder.
- It enables us to sample the features with more reasonable time resolutions, which helps to model the sequences more efficiently.
- We believe that the above advantages are beyond beat tracking and can be useful for other tasks too.