Detection of Cut-Points for Automatic Music Rearrangement

Daniel Stoller[1], Vincent Akkermans[2] and Simon Dixon[1]

[1]Queen Mary University of London [2]MXX London

Summary

- Task: Given a music piece, rearrange it according to some user constraints
- Most successful music rearrangement methods so far cut between sections as smoothly as possible
- "Jumps" are noticed when musical expectations are violated at cut points

⇒ Rate cut candidates according to musical features, but these are numerous and hard to describe

- We propose a data-driven approach at finding cut points by using cut annotations
- Model learns automatically to attend to rhythm and instrument activity

Motivation

User wants to change a music piece's

- Duration
- Musical structure
- Instrument presence (remove vocals)

Previous approaches

Mainly **cut-based** approaches [3, 2]:

- Find time points t_1, t_2 so that skipping from t_1 to t_2 is least noticeable, ensuring that
- the resulting rearrangement fulfils the given user demands

Main problem for cut-based approach: Selection of cut points. Melodic expectations of the listener have to be met regarding

- Melody
- Rhythm
- Instrument activity
- etc.

Many handcrafted features were used to try and capture some of these aspects, but they are numerous and hard to define

Acknowledgements

This work was partially funded by EPSRC grant EP/L01632X/1.

Key concept

Apply deep learning on a dataset with cut annotations to automatically **learn** what makes a good cut, instead of making hypotheses using handcrafted features

Dataset

- 300 Western Pop songs
- Musical structure annotated
- Note onsets marked as entry or exit cuts

After randomly sampling negative examples, we obtain 38796 music snippets for training.

Feature sets

- Handcrafted features (baseline):
  - MFCCs (12-dim.)
  - Chroma features (12-dim.)
  - Tempogram (12-dim.)
- Gammatone (GT) spectrogram (75 filters)
  - Absolute time/beat-aligned
- Constant-Q (CQT) (12 bins/oct., 8 oct.)

Classification models

Train two models to classify for the central frame of a music snippet if it is

- Exit or no exit?
- Entry or no entry?

After bad results with fully connected networks, we used three architectures:

1. 1D CNNs
2. 2D CNNs
3. U-Net adaptation [1]

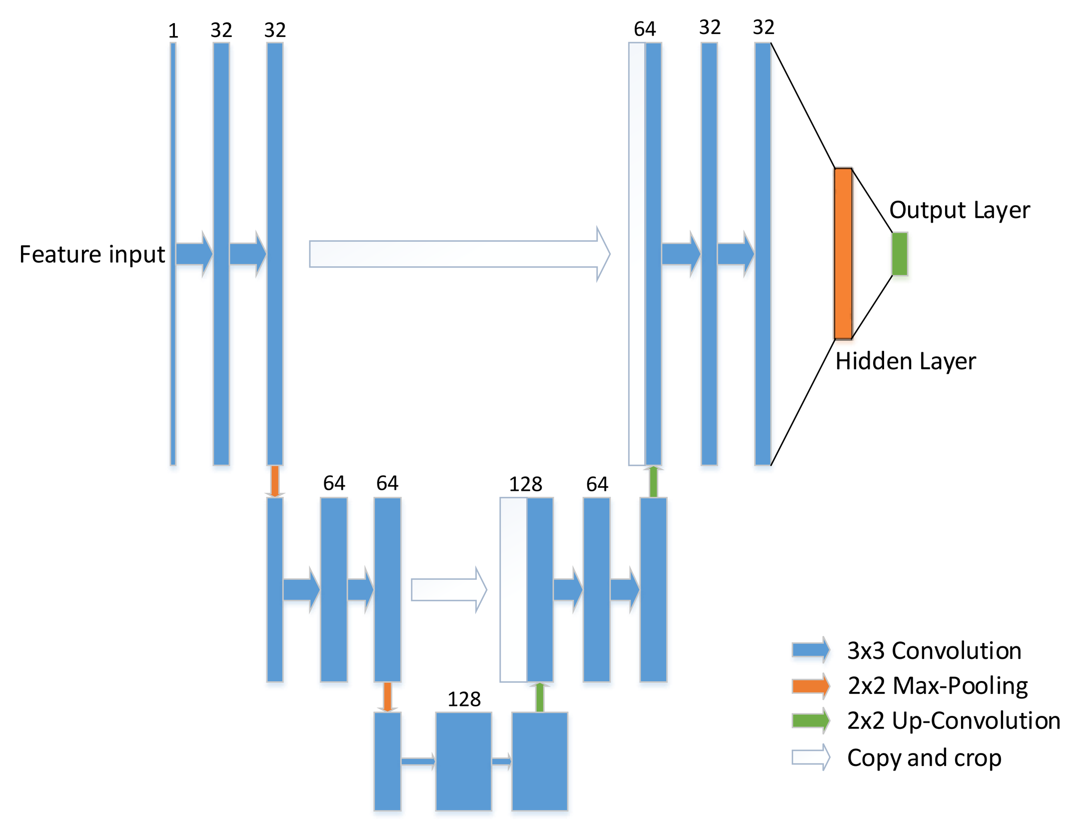


Figure 1: The adapted U-Net architecture

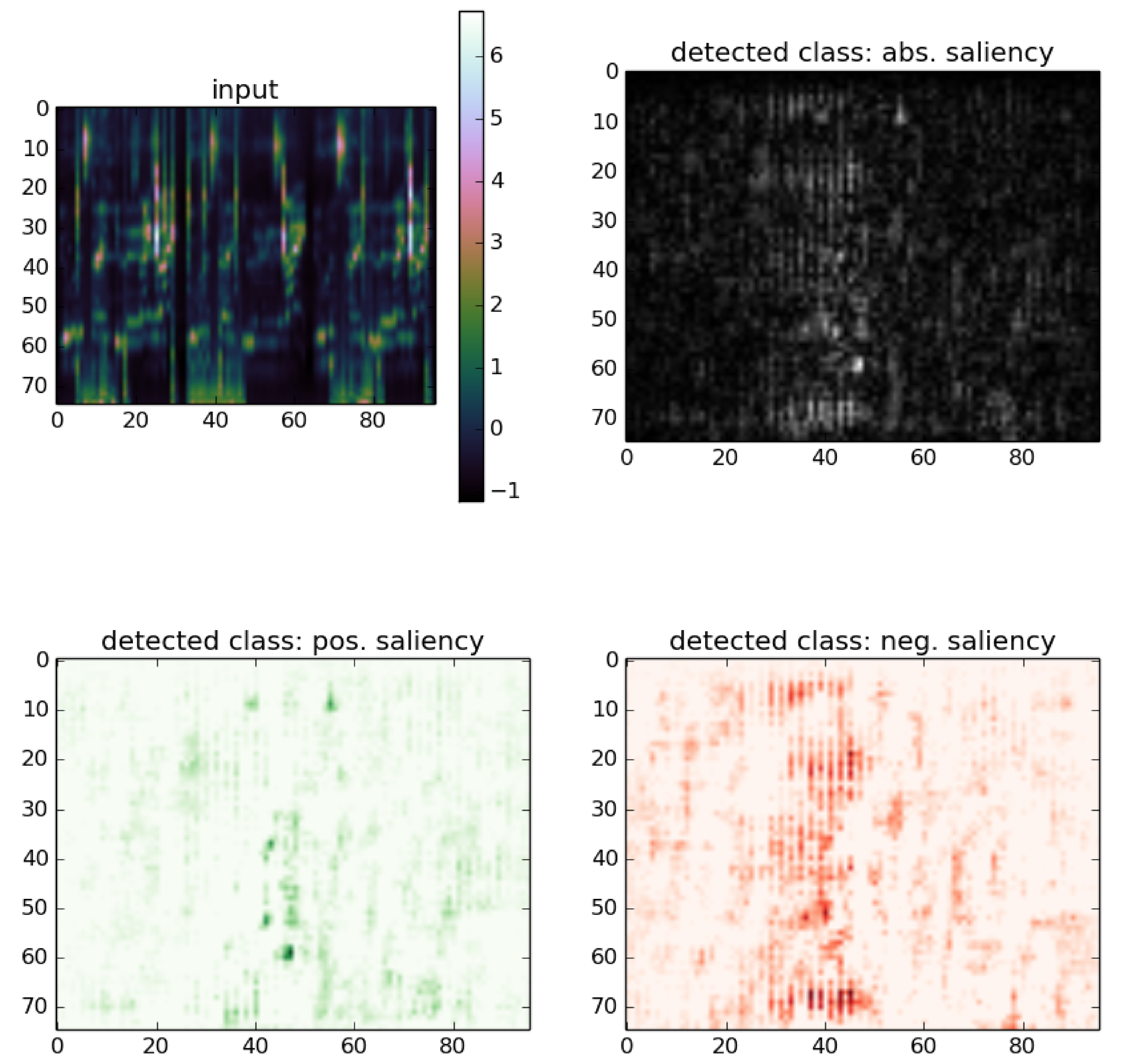
Results

Analysis with balanced classification rate

- 2D-CNN > 1D-CNN (0.656 > 0.610)
- U-Net > 1D-CNN, but < 2D-CNN (0.637)
- 1D-CNN: GT much better than CQT (0.643 > 0.600), 2D-CNN: GT still better, but only slightly (0.658 > 0.653)
- Beat-aligned GT slightly better than absolute-time GT (0.678 > 0.670)
- 4-10 sec. long inputs work best

What did the model learn?

Saliency map shown for true neg. entry:



- Vocals present in pos. saliency map
- Neg. saliency shows more sound before the cut would lead to predicting entry

Error analysis

- Ask annotator about confidence of label for 65 randomly chosen false positives
- 33.8% accepted as true positive predictions

⇒ Dataset bias due to only few suitable onsets being labelled, limits model performance

References

[1] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[2] D. Stoller, I. Vatolkin, and H. MÃijller. Intuitive and efficient computer-aided music rearrangement with optimised processing of audio transitions. *Journal of New Music Research*, 0(0):1–22, 2018.

[3] S. Wenger and M. Magnor. A genetic algorithm for audio retargeting. In *ACM Multimedia (ACMMM)*, pages 705–708, 2012.