BEAT TRANSFORMER: Demixed Beat and Downbeat Tracking with Dilated Self-Attention

Jingwei Zhao Gus Xia Ye Wang

NUS National University of Singapore
Sound & Music Computing Lab
NYU Shanghai
MBZUAI
Music X Lab

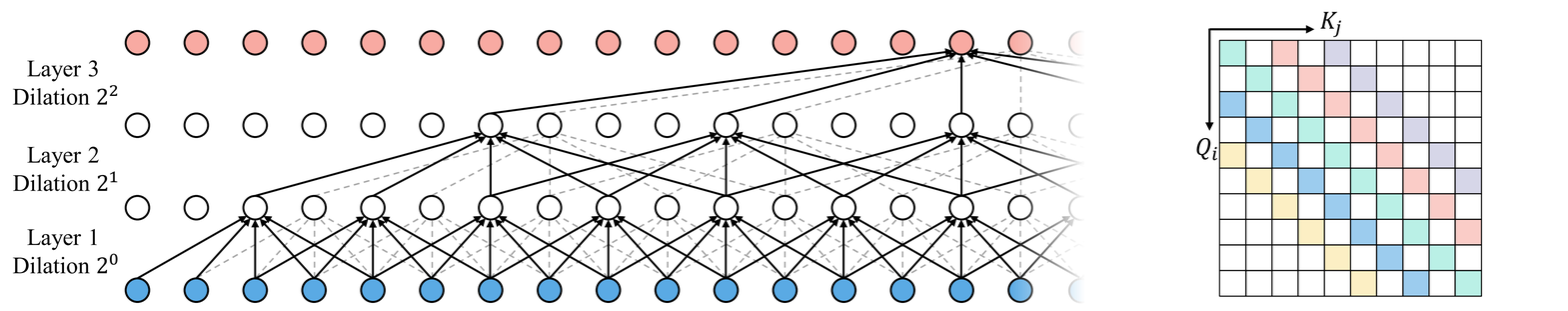
Key Idea

- When humans perceive music beats, instead of relying solely on the spectral energy, we can "understand" the music in a deeper musical context.
- To reflect our musical understanding, we resort to *demixed audio input* and excavate metrical cues with both temporal attention and instrumental attention.
- We formalize *dilated self-attention* mechanism for temporal attention, which demonstrates superior sequential modelling power with linear complexity.

Resource

Paper Colab Notebook GitHub

Dilated Self-Attention and Demixed Beat Tracking



(a) Layer-wise view of dilated self-attention (Adapted from TCN structures [9, 27]) (b) Attention matrix view at layer 2

Figure 1: Illustration of dilated self-attention (with a non-causal short window of size 5) over a three-layer Transformer. Part (a) shows the hierarchical connectivity across layers, which shares the same pattern as TCN in [9]. Part (b) shows the attention matrix at layer 2, with colours indicating relative position. The white colour indicates unattainable positions.

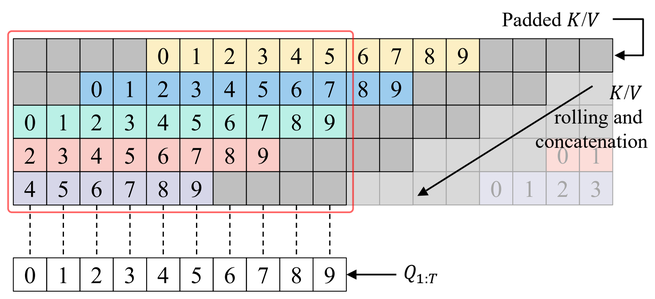


Figure 2: Illustration of efficient DSA implementation, with dilation rate $r = 2$ and window size $l_{\text{win}} = 5$.

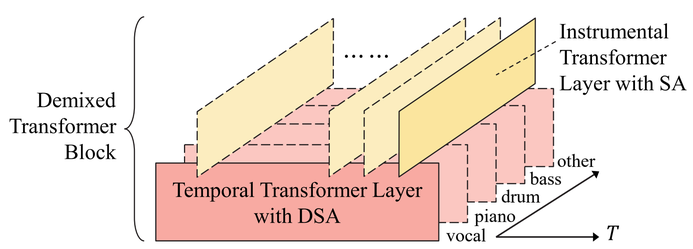
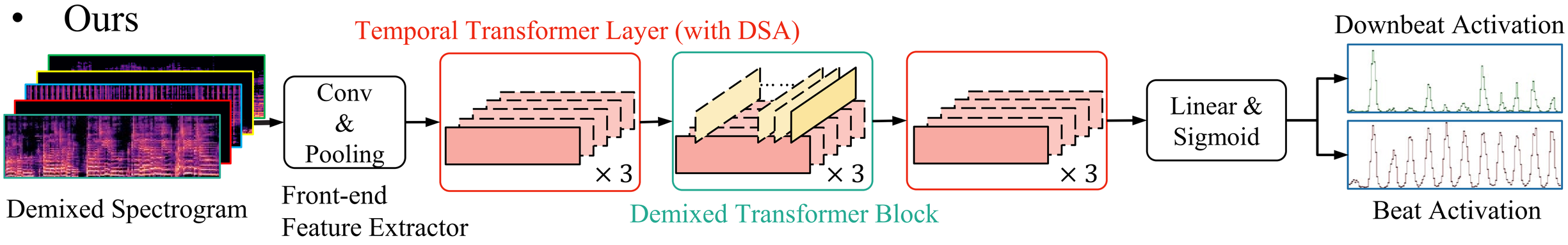


Figure 3: Demixed Transformer block. Two Transformer layers are stacked "orthogonally", each handling time-wise dilated self-attention and instrument-wise self-attention.

Complete Architecture



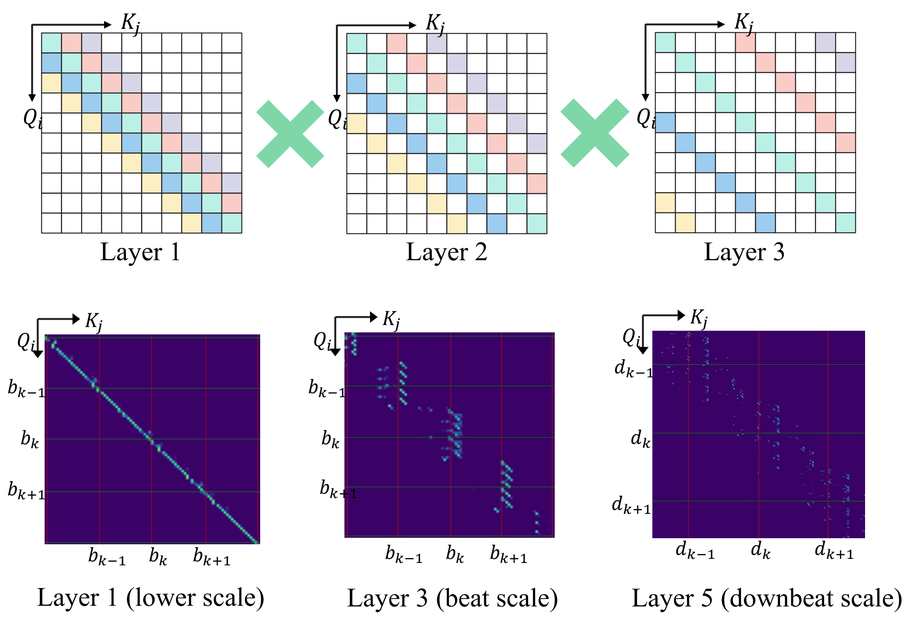
Models in Comparison

- ***Ours w/o Demix***: Use Temporal Transformer Layers only, while input is a single spectrogram of audio mixture.
- ***TCN+Demix***: Each Temporal Transformer Layer is replaced by a TCN layer with same dilation rate & feature size.
- ***Ours+Aug.***: During training, we apply partial-demix augmentation to further improve our model's performance.
- ***Bock et. al.*** [1]: SOTA beat tracking architecture based on TCN .
- ***Hung et. al.*** [2]: SOTA model based on SpecTNT, a variant of Transformer which models time-frequency relation.

Model Performance

Dataset	Model	Beat Accuracy			Downbeat Accuracy		
		F-Measure	CMLt	AMLt	F-Measure	CMLt	AMLt
Ballroom	TCN+Demix	0.960	0.942	0.961	0.927	0.926	0.957
	Ours w/o Demix	0.968	0.946	0.965	0.930	0.925	0.963
	Ours	0.967	0.949	**0.967**	0.928	0.931	0.958
	Ours+Aug.	**0.968**	**0.954**	0.966	**0.941**	**0.944**	**0.969**
	Bock *et al.*	0.962	0.947	0.961	0.916	0.913	0.960
	Hung *et al.*	0.962	0.939	0.967	0.937	0.927	0.968
GTZAN	TCN+Demix	0.873	0.780	0.907	0.700	0.646	0.842
	Ours w/o Demix	0.876	0.787	0.914	0.686	0.633	0.834
	Ours	0.881	0.797	0.921	0.703	0.653	0.845
	Ours+Aug.	0.885	0.800	0.922	0.714	0.665	0.844
	Bock *et al.*	0.885	**0.813**	**0.931**	0.672	0.640	0.832
	Hung *et al.*	**0.887**	0.812	0.920	**0.756**	**0.715**	**0.881**

- Interpretation with Markov chain rules
  - Each attention matrix represents a one-step transition on a finite-state Markov chain
  - Product of attention matrices through layers represents a multi-step transition
- Attention Visualization
  - *Cumulative product* of attention matrices
  - Our model gathers information from both local and global scales with an organized hierarchy, which resembles the way we human detect music beats and downbeats.



References

[1] S. Böck and M. E. P. Davies, "Deconstruct, analyse, reconstruct: How to improve tempo, beat, and downbeat estimation," in *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR*, pp. 574–582, 2020.

[2] Y. Hung, J. Wang, X. Song, W. T. Lu, and M. Won, "Modeling beats and downbeats with a time-frequency transformer," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pp. 401–405, IEEE, 2022.