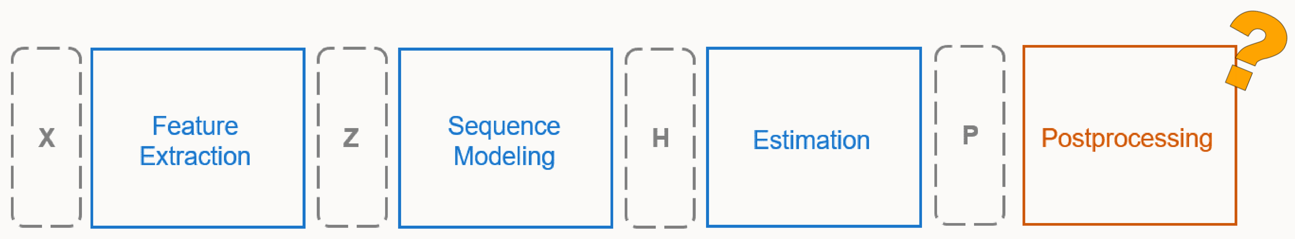
Toward Postprocessing-free Neural Networks for Joint Beat and Downbeat Estimation

Tsung-Ping Chen and Li Su

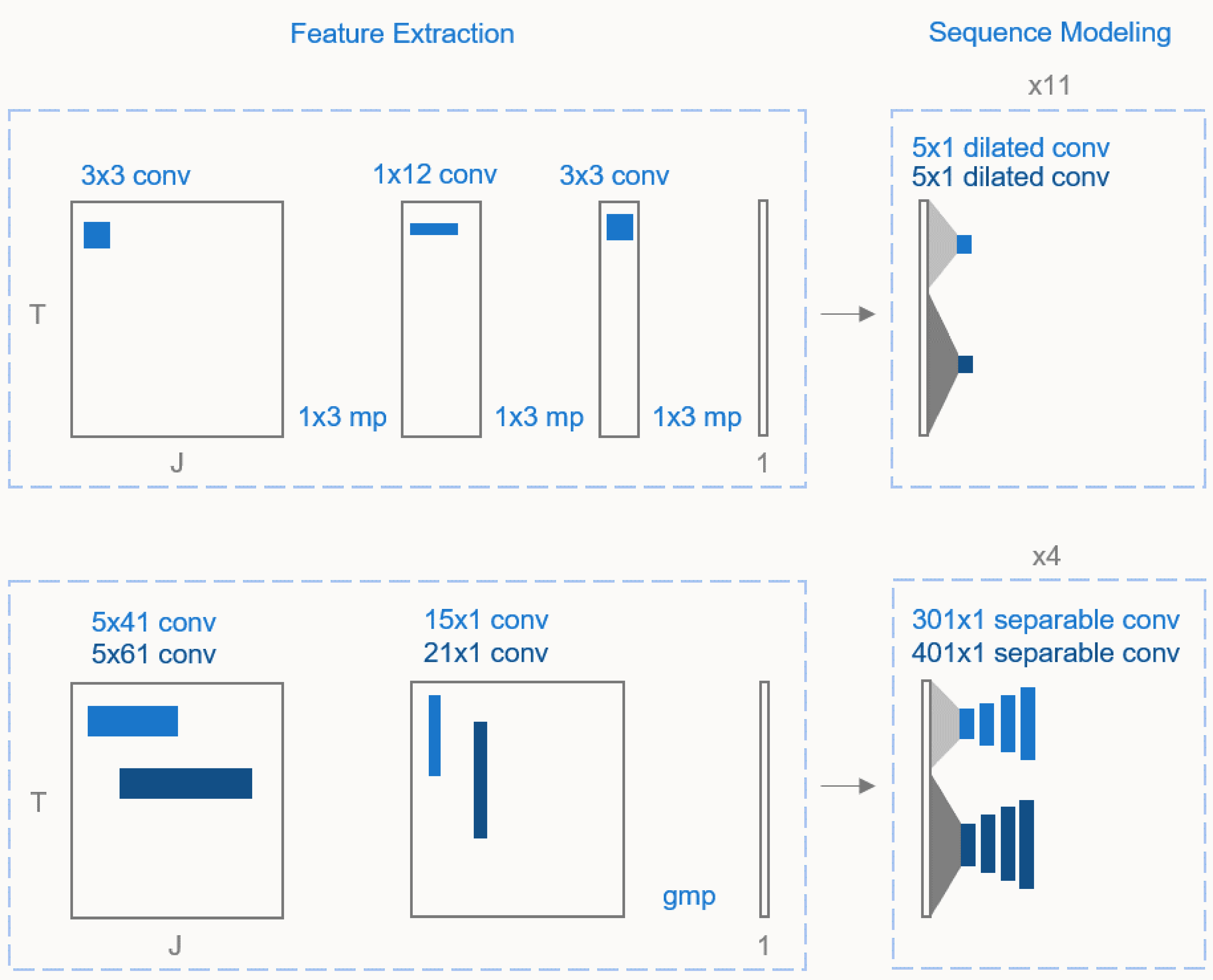
Institute of Information Science, Academia Sinica, Taiwan

Introduction

A dedicatedly-designed network based on hidden Markov models or dynamic Bayesian networks is often combined with deep neural networks in various sequence modeling systems for postprocessing. Ideally, deep neural networks might be able to forgo such a post-processing stage. In this work, we attempt to tackle the joint beat and downbeat estimation task without incorporating a postprocessing network. By inspecting a state-of-the-art approach, we propose several reformulations regarding the network architecture and the loss function. We evaluate our model on various music data and show that the proposed methods are capable of improving the baseline approach without the aid of a post processing approach.

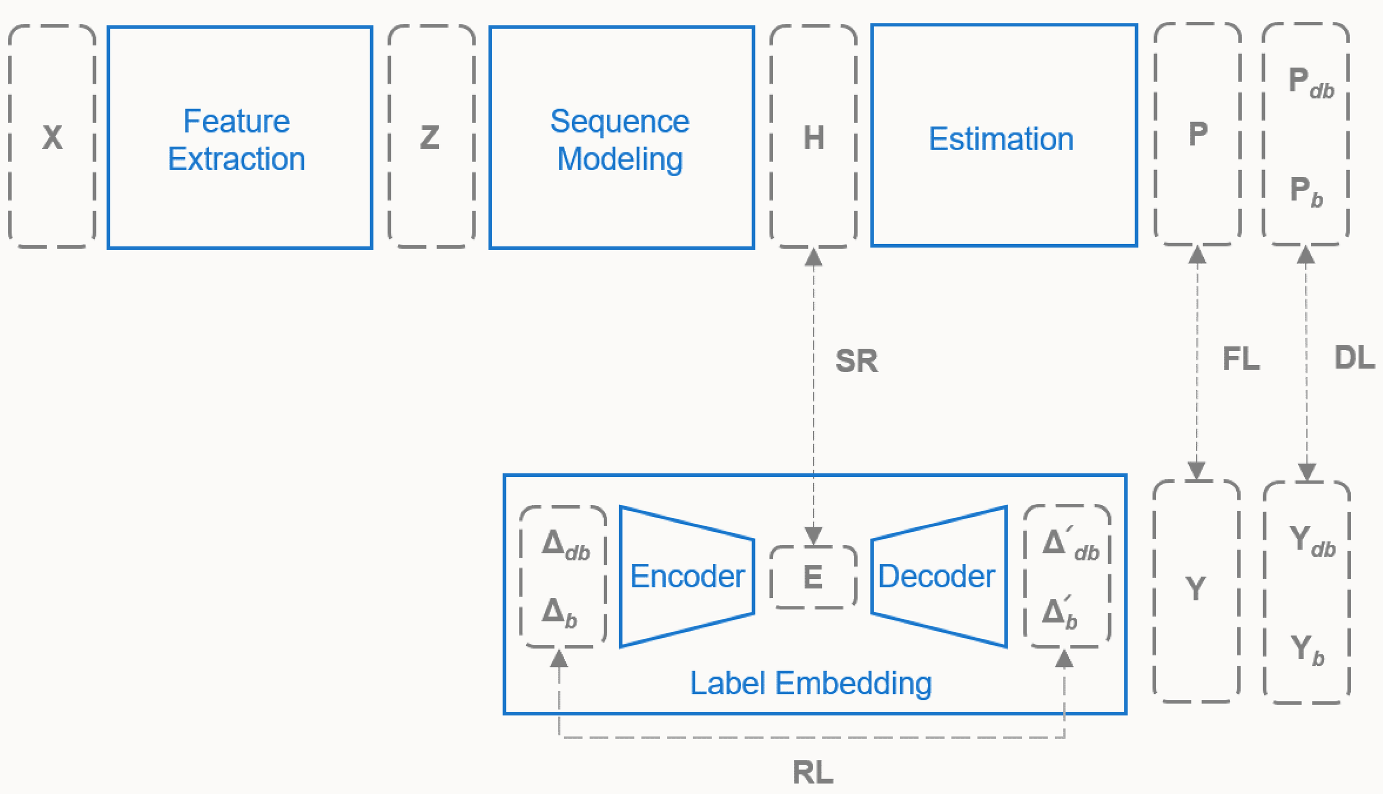


Reformulation of Architecture



	Feature Extraction	Sequence Modeling
Baseline	Small kernel size	Small Kernel size
	Intermediate local max pooling (mp)	Sparse sampling via dilated convolution
Proposed	Larger kernel size	Larger Kernel size
	Endmost global max pooling (gmp)	Dense sampling via separable convolution
Motivation	The spectro-temporal patterns might not be well-captured with convolutions of small kernel size followed by immediate pooling.	Convolutions with small kernels and high dilation rates relate sparsely distributed elements which might lack of correlations.

Reformulation of Loss Function



	Estimation Loss	Structural Loss
Baseline	Cross Entropy	
Proposed	Focal loss (FL) $\mathrm{FL} = -\frac{1}{N}\sum_{i=1}^{N} m_i \times y_i \times \log(p_i)$	Structural regularization (SR)
	Dice loss (DL) $\mathrm{DL} = 1 - \frac{2\sum_{i=1}^{N} p_i \times y_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} y_i^2}$	Reconstruction loss (RL)
Motivation	FL accentuates individual hard samples whereas DL underlines collective similarity of the minority samples.	We leverage a label embedding approach for learning the periodic structure of beat/downbeat sequences.

Evaluation

Task	Model	Ballroom	Hainsworth	GTZAN	ASAP (Audio)	ASAP (Midi)
Beat	Baseline	96.60	84.18	85.12	**71.43**	72.45
	Proposed	**96.81**	**86.28**	**88.50**	71.40	**75.70**
Downbeat	Baseline	92.06	66.18	61.96	49.46	57.34
	Proposed	**94.21**	**69.11**	**67.56**	**63.92**	**67.43**

Evaluation results in terms of F1 score (%)

Ablation	Ballroom (Beat)	Ballroom (Downbeat)	Hainsworth (Beat)	Hainsworth (Downbeat)
Feature extraction	96.08 (-0.73)	**92.16 (-2.05)**	85.49 (-0.79)	**62.27 (-6.84)**
Sequence labeling	96.98 (+0.17)	94.16 (-0.05)	85.96 (-0.32)	69.73 (+0.62)
Estimation loss	95.27 (-1.54)	**91.99 (-2.22)**	86.52 (+0.24)	**66.79 (-2.32)**
Structural loss	96.54 (-0.27)	93.61 (-0.60)	86.83 (+0.55)	69.58 (+0.47)

Ablation performance and relative improvement against the proposed model (in parentheses)

Conclusion

- The experiment results show that we can further the performance of a deep neural network by reconsidering the architecture and the loss function.
- The ablation study indicates that the reformulations of the feature extraction and the Loss function have notable positive effect especially on estimating downbeats.
- The inclusion of the structural loss might have negative effect due to the regularization (SR) on the output of the sequence modeling module (H).
- While involving a post-processing network often leads to an improvement over the preceding deep learning models, it hinders the formulation of end-to-end training and indicates a necessity to reconsider the employed neural networks.