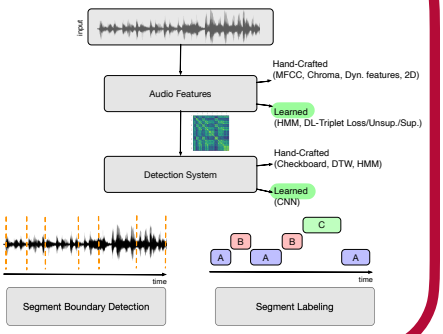
Self-Similarity-Based and Novelty-Based Loss for Music Structure Analysis

Geoffroy Peeters LTCI, Télécom-Paris, IP-Paris
https://github.com/geoffroypeeters/ssmnet_ISMIR2023

Introduction

Music structure analysis ?
- identify musical segments that compose a music track (also known as Segment Boundary estimation)
- - label them based on their similarity (also known as segment labelling).

Over the years and the accessibility of annotated datasets, systems have switched from
- hand-crafted audio features used as input to hand-crafted detection systems
- hand crafted audio features used as input to deep learning detection
- deep learned features used as input to hand-crafted detection systems



Proposal

We propose a system which relies on deep learned audio features used as input to deep learning detection system

We propose a straightforward training paradigms based on:
- a direct comparison of estimated and ground-truth SSM
- a direct comparison of estimated and ground-truth novelty-curve

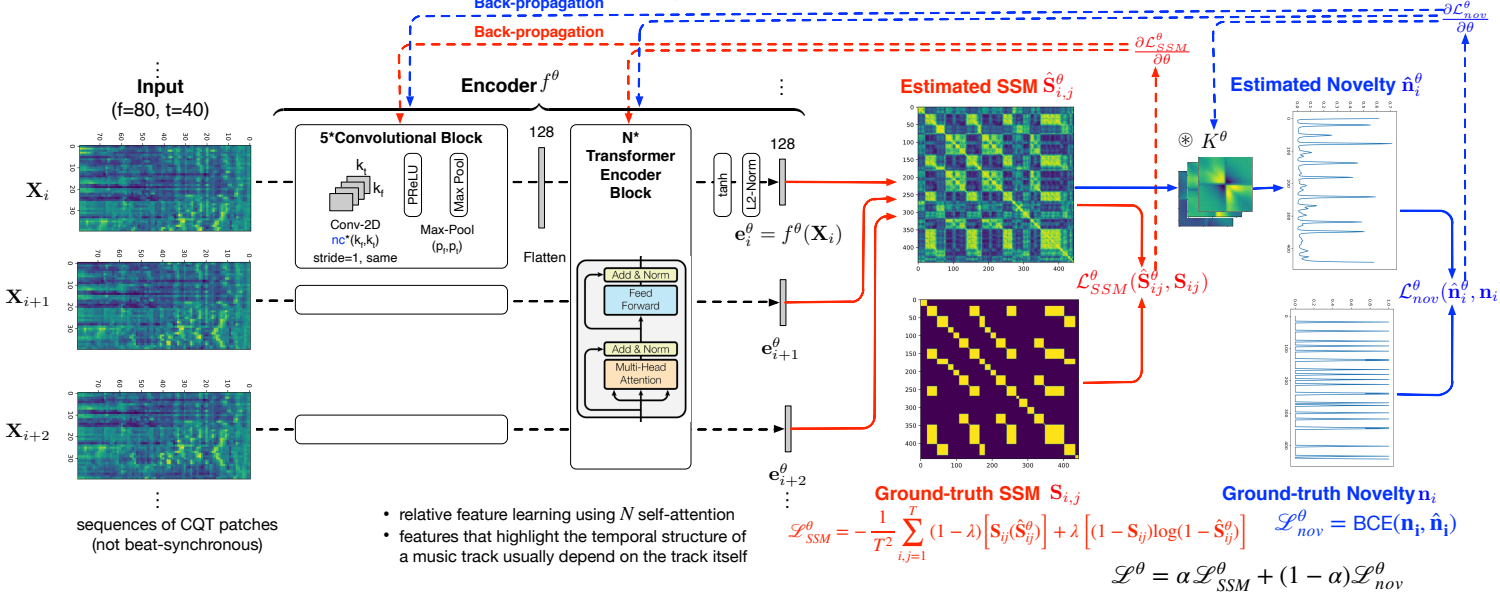
- We train a deep encoder f^θ and a set of convolution kernels K^θ such that
  - the Self-Similarity-Matrix (SSM) $\hat{\mathbf{S}}^\theta_{ij}$ resulting from the learned features $\mathbf{e}^\theta_i = f^\theta(\mathbf{X}_i)$ approximates a ground-truth SSM $\mathbf{S}_{ij}$
  - the Novelty-curve $\hat{\mathbf{n}}_i$ resulting from convolving $\hat{\mathbf{S}}^\theta_{ij}$ with learned kernels K^θ approximates a ground-truth Novelty $\mathbf{n}_i$

$$\frac{\partial \mathcal{L}^\theta_{SSM}}{\partial \theta} = \sum_{i,j=1}^{T} \frac{\partial \mathcal{L}^\theta_{SSM}}{\partial \hat{\mathbf{S}}^\theta_{ij}} \left(\frac{\partial \hat{\mathbf{S}}^\theta_{ij}}{\partial \mathbf{e}^\theta_i} \frac{\partial \mathbf{e}^\theta_i}{\partial \theta} + \frac{\partial \hat{\mathbf{S}}^\theta_{ij}}{\partial \mathbf{e}^\theta_j} \frac{\partial \mathbf{e}^\theta_j}{\partial \theta} \right)$$

Back-propagation

$$\hat{\mathbf{S}}^\theta_{ij} = 1 - \frac{1}{4}\|\mathbf{e}^\theta_i - \mathbf{e}^\theta_j\|_2^2 \quad \in [0,1]$$

K^θ
- 3 kernels of (41,41) corresponds to 20s + (1x1) convolution
- Initialised using Foote chck kernels / using randn



- relative feature learning using N self-attention
- features that highlight the temporal structure of a music track usually depend on the track itself

$$\mathcal{L}^\theta_{SSM} = -\frac{1}{T^2}\sum_{i,j=1}^{T}(1-\lambda)\left[\mathbf{S}_{ij}(\hat{\mathbf{S}}^\theta_{ij})\right] + \lambda\left[(1-\mathbf{S}_{ij})\log(1-\hat{\mathbf{S}}^\theta_{ij})\right] \qquad \mathcal{L}^\theta_{nov} = \mathrm{BCE}(\mathbf{n}_i, \hat{\mathbf{n}}_i)$$

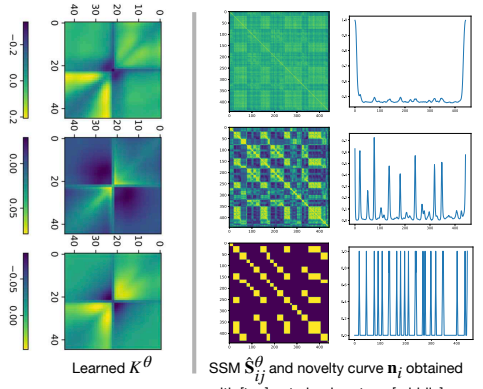
$$\mathcal{L}^\theta = \alpha\mathcal{L}^\theta_{SSM} + (1-\alpha)\mathcal{L}^\theta_{nov}$$

Evaluation

Previous results
Our results
Ablation study

	RWC-Pop-AIST HR.5F	RWC-Pop-AIST HR3F	SA-Pop HR.5F	SA-Pop HR3F	SA-IA HR.5F	SA-IA HR3F	SA-Two HR.5F	SA-Two HR3F	Annotation
Grill [23,42] GS1	.506	**.715**	-	-	-	-	.541	.623	Up./An-*
McCallum [25] Unsynch.	-	-	-	-	-	.497	-	-	
Beat-synch.	-	-	-	-	-	**.535**	-	-	
Salamon [30] DEF$_{\beta 5.05/*_\mu H_{\sigma}H}$	-	-	-	-	-	-	.337	.563	Up./An-*
Wang [27] scluster/D/eu/mul	.438	.653	.447	.623	-	-	.356	.553	Up./An-*
Buisson [29] HE$_0$/HE$_1$	-	.681	-	-	-	-	-	**.597 / .595**	Up./An-1/2
			-	-	-	-	-	**.611 / .600**	Low./An-1/2
Ours (best conf.)	.399	**.713**	.298 / .295	**.631/ .624**	.250 / .261	**.520 / .511**	.231 / .237	.521 / .530	Up./An-1/2
			.296 / .318	.570 / .610	.302 / .336	.547 / .612	.287 / .287	.589 / .589	Low./An-1/2
Ablation study N									
N=3/α=0.5/K:train-Init:chck		**.713**		.532		.472		.448	Up./An-1
N=2/α=0.5/K:train-Init:chck		.701		.535		.474		.449	Up./An-1
N=1/α=0.5/K:train-Init:chck		.677		**.631**		**.520**		**.521**	Up./An1
N=0/α=0.5/K:train-Init:chck		.696		.535		.459		.443	Up./An-1
Ablation study α									
N=3/α=1/K:train-Init:chck		.154		.121		.102		.111	Up./An-1
N=3/α=0/K:train-Init:chck		.007		.120		.026		.095	Up./An-1
Ablation study K^θ									
N=3/α=0.5/K:train-Init:randn		**.713**		.543		.470		.457	Up./An-1
N=1/α=0.5/K:train-Init:randn		.709		.547		.470		.457	Up./An-1
N=3/α=0.5/K:fix-Init:chck		.330		.250		.199		.196	Up./An-1

Illustration



SSM $\hat{\mathbf{S}}^\theta_{ij}$ and novelty curve $\mathbf{n}_i$ obtained with [top] untrained system, [middle] trained system and [bottom] ground-truth

[23] T. Grill and J. Schlüter, "Music Boundary Detection Using Neural Networks on Combined Features and Two-Level Annotations," in Proc. of ISMIR, Malaga, Spain, 2015.
[25] M. C. McCallum, "Unsupervised Learning of Deep Features for Music Segmentation," in Proc. of IEEE ICASSP, Brighton, UK, May 2019.
[42] T. Grill and J. Schlüter, "Structural segmentation with convolutional neural networks MIREX submission," 2015.
[30] J. Salamon, O. Nieto, and N. J. Bryan, "Deep embed- dings and section fusion improve music segmentation," in Proc. of ISMIR, Online, November, 8–12 2021.
[27] J.-C.Wang, J.B.L. Smith, W.-T. Lu, and X. Song, "Supervised metric learning for music structure features," in Proc. of ISMIR, Online, 2021.
[29] M. Buisson, B. McFee, S. Essid, and H.-C. Crayencour, "Learning multi-level representations for hierarchical music structure analysis," in Proc. of ISMIR, 2022.