Less is more: Faster and better music version identification with embedding distillation

Furkan Yesiler, Joan Serrà, Emilia Gómez

furkan.yesiler@upf.edu

upf. Music Technology Group · DOLBY · European Commission JOINT RESEARCH CENTRE · ISMIR MTL2020

Motivation

Goal:
Scale up to industrial-size databases

Problem:
Models w/ larger embeddings perform better
More storage space!!
Longer retrieval times!!!

Our research question:
Pre-trained models w/ large embeddings → Models w/ smaller embeddings + same accuracy?

Da-TACOS training set

Pre-extracted cremaPCP features and metadata for +97k songs!

Training partition (+83k songs) and validation partition (14k songs)

Annotations from secondhandsongs.com

CC BY-NC-SA 4.0

Re-MOVE on GitHub

https://github.com/furkanyesiler/re-move

Pre-trained models

Instructions to download Da-TACOS training set

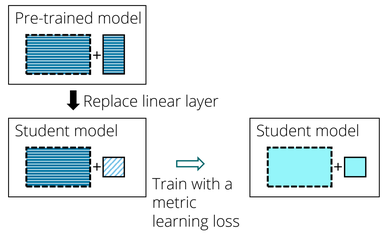
Supplementary materials

MIPFrontiers · This work is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068, and by the Horizon 2020 project 770376-2. · TROMPA

Embedding Distillation

Reducing the embedding size of pre-trained models while maintaining the accuracy

Latent space reconfiguration



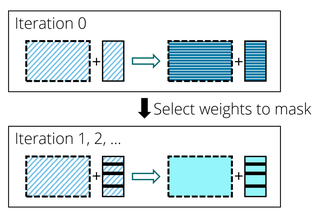
"Reconfigure" a learned distance metric with a new loss

Represent the semantic relations in a more compact space

Re-MOVE (Reduced-MOVE)
NormalizedSoftmax

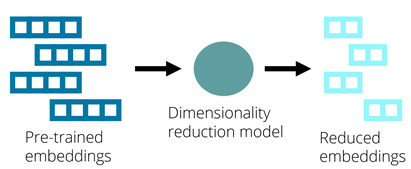
	Transfer learning	Latent space reconfiguration
Underlying idea	Using learned priors for better optimization	Using learned priors for better optimization
Tasks	Different	Same
Input distributions for tasks	Different	Same
Main purpose	To improve generalization on a new setting	To facilitate learning a compact latent space

Neural network pruning



1. Store initial weights
2. Train for N epochs
3. Select rows to mask
4. Restore initial weights
5. Train for N epochs
6. Repeat 3-5

Unsupervised dimensionality reduction

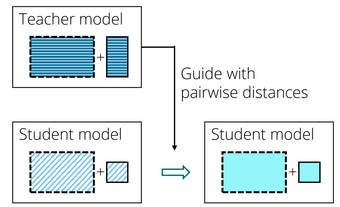


Principal Component Analysis

Independent Component Analysis

Gaussian Random Projections

Knowledge distillation

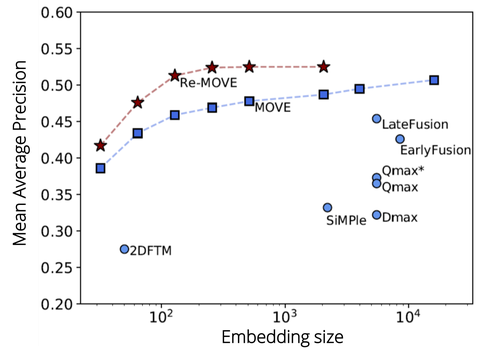


Distance-based KD
Match distances from teacher and student

Cluster quality-based KD
Match Davies-Bouldin Index from teacher and student

Results vs State of the Art

Evaluation on Da-TACOS



Method	Embedding size 128	256	512	2048
Baselines (no reduction, training from scratch)				
Triplet	0.459	0.469	0.478	0.487
ProxyNCA	0.168	0.185	0.212	0.206
NormalizedSoftmax	0.445	0.470	0.475	0.422
Group	0.265	0.271	0.269	0.271
Unsupervised				
PCA	0.494	**0.507**	**0.507**	**0.507**
ICA	0.456	0.425	n/a	n/a
GRP	0.429	0.465	0.485	**0.502**
Knowledge distillation				
Distance matching + Triplet	0.492	**0.499**	**0.503**	**0.500**
Cluster matching + Triplet	0.424	0.471	0.465	0.455
Latent space reconfiguration				
Triplet	0.485	0.491	0.494	**0.506**
ProxyNCA	0.424	0.465	0.485	**0.502**
NormalizedSoftmax	**0.513**	**0.524**	**0.525**	**0.525**
Group	0.465	0.483	**0.495**	**0.511**