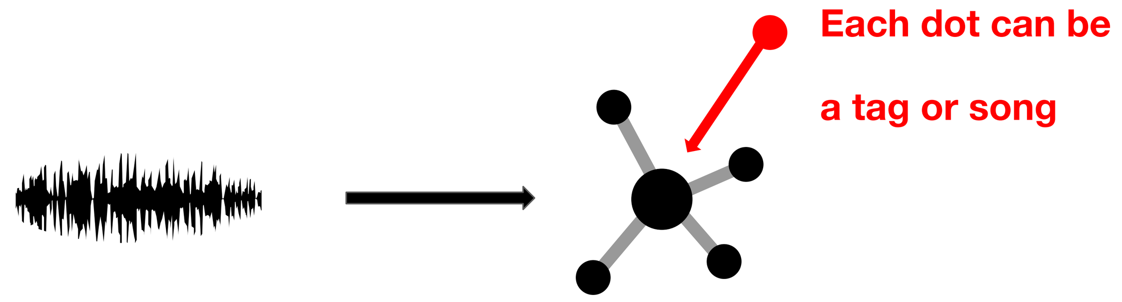
Metric Learning VS Classification for Disentangled Music Representation Learning

Jongpil Lee[1], Nicholas J. Bryan[2], Justin Salamon[2], Zeyu Jin[2], Juhan Nam[1]

[1]KAIST, Daejeon, South Korea
[2]Adobe Research, San Francisco, United States

https://jongpillee.github.io/metric-vs-classification/

Representation Learning

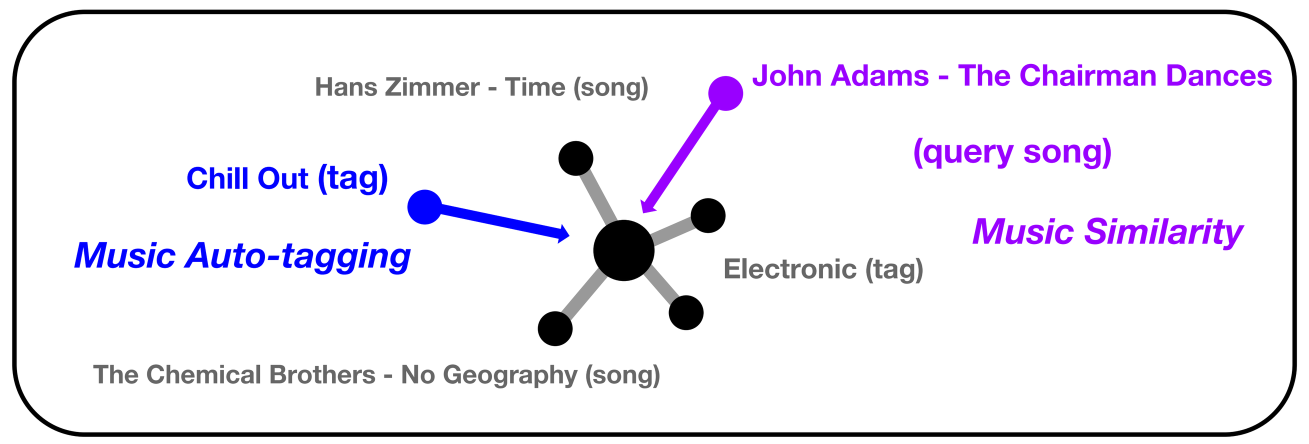


- **Deep representation learning** offers a powerful paradigm for mapping input data onto **an organized embedding space** and is useful for many music information retrieval tasks.

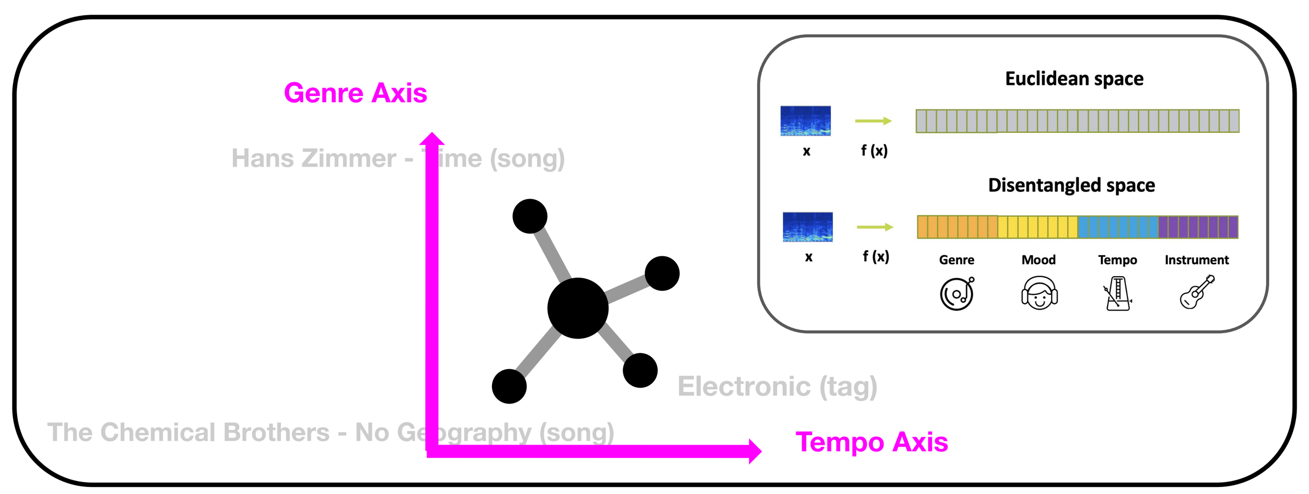
Content-based Music Retrieval

Tag-based Retrieval	Classification
Example-based Retrieval	Metric Learning
Multidimensional Retrieval	Disentanglement

Tag-based Retrieval and Example-based Retrieval



Multidimensional Retrieval



Summary

- Two central methods for representation learning include **deep metric learning** and **classification**.
- The emerging concept of **disentangled representations** is also of great interest, where multiple semantic concepts (e.g., genre, mood, instrumentation) are learned jointly but remain **separable in the learned representation space**.
- In this paper, we present **a unified representation learning framework** that can perform **example-based retrieval**, **tag-based retrieval**, and **multidimensional retrieval** in a holistic manner.
- (1) we first outline past work on **the relationship between metric learning and classification**.
- (2) then, we **extend this relationship to multi-label data** by exploring three different learning approaches and **their disentangled versions**.
- (3) Finally, we **evaluate all models on four tasks** (training time, similarity retrieval, auto-tagging, and triplet prediction).
- As a result, we find that **classification-based models** are generally advantageous for **training time, similarity retrieval, and auto-tagging**, while **deep metric learning** exhibits better performance for **triplet-prediction**.
- At last, we show that our proposed approach yields **state-of-the-art results** for **music auto-tagging** and **similarity-based retrieval**.

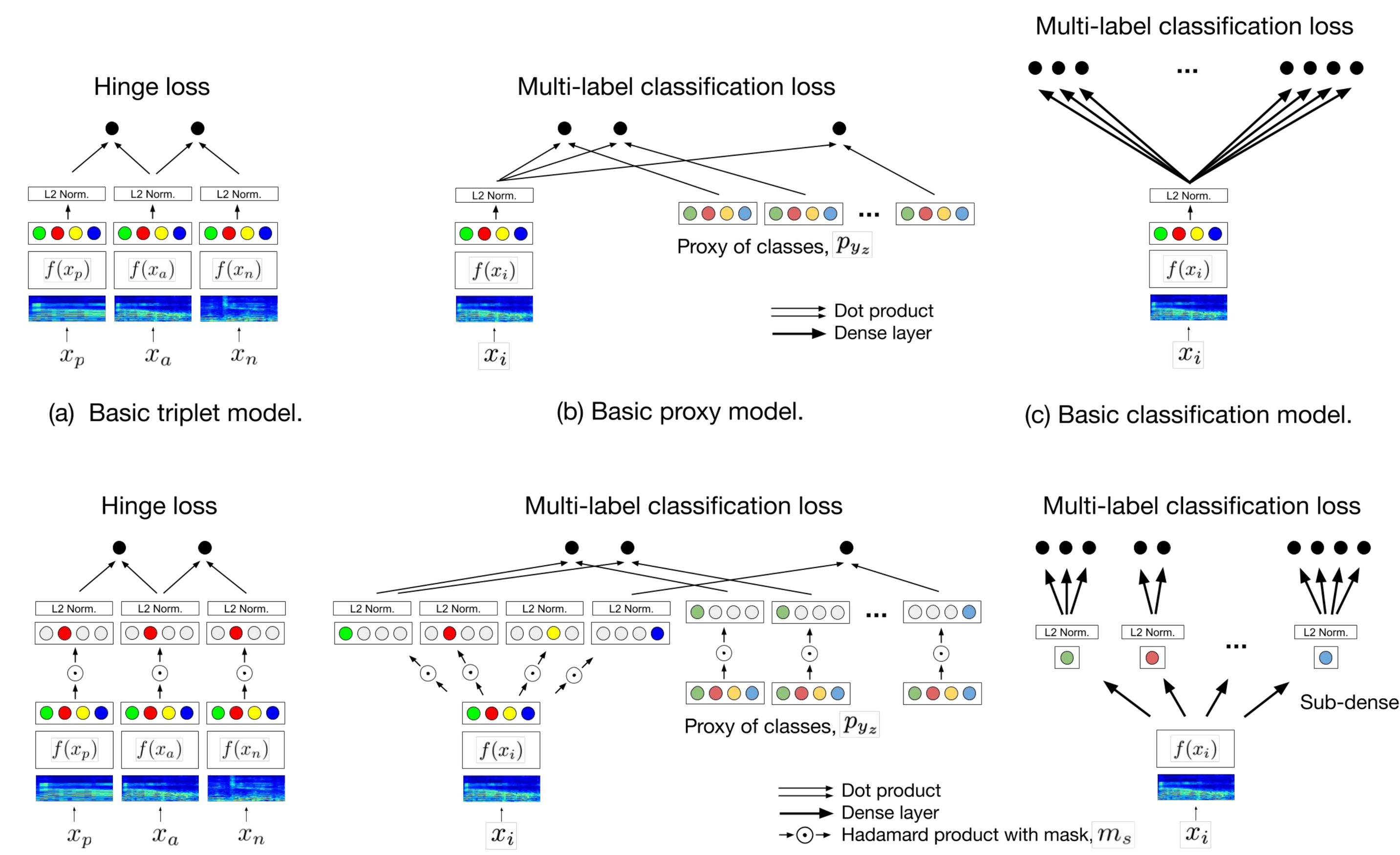
Experiments

- **Million Song Dataset (MSD) with Last.FM tag annotations**
- **50 tags (28 genres, 12 moods, 5 instruments, 5 eras)**
- **201680, 11774, 28435 tracks for train, validation, and test sets**
- **3 second excerpts based deep inception backbone model**

Models

	Basic	Disentangled Version
Metric Learning	✓	✓
Proxy-based Metric Learning	✓	✓
Classification	✓	✓

- We connect the relationship between **classification** and **metric learning** using **proxy-based metric learning**. Then, we develop their **disentangled version of the models** to perform **all the three retrieval cases** and then we compare the models.

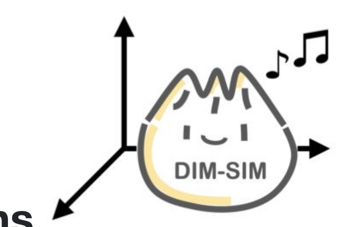


(a) Basic triplet model. (b) Basic proxy model. (c) Basic classification model.
(d) Disentangled triplet model. (e) Disentangled proxy model. (f) Disentangled classification model.

Dim-Sim Dataset

https://jongpillee.github.io/multi-dim-music-sim/
https://zenodo.org/record/3889149#.X3gtaJMzbyW

- **A user-annotated music similarity triplet ratings**
- **linked to the Million Song Dataset (MSD)**
- **4,000 3-second triplets, 39,440 human annotations**



Evaluations

Tasks	Evaluation Metrics	
Tag-based Retrieval	Tagging Performance	AUC
Example-based Retrieval	Similarity Performance	R@K
Multidimensional Retrieval	Triplet Prediction	Acc.
Training Efficiency	Training Time	Ratio

Results

Results for training time, similarity-based retrieval, and auto-tagging

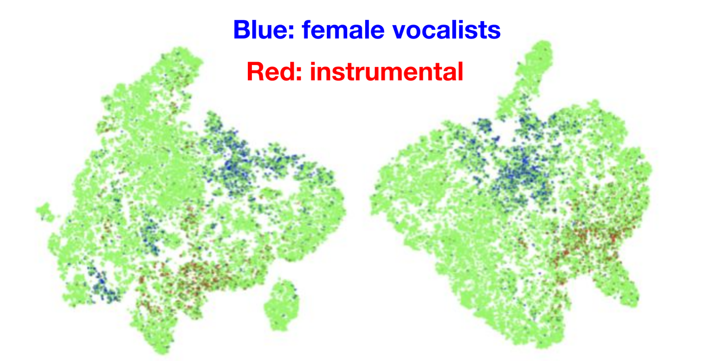
Models	Normalization	Disentanglement	Training time ratio	Similarity-based retrieval R@1	R@2	R@4	R@8	Auto-tagging AUC
Triplet	✓	✗	1.87	31.8	45.2	59.9	73.0	0.815
Triplet	✓	✓	2.37	36.5	50.5	64.1	76.0	0.825
Triplet + track reg.	✓	✓	3.05	33.9	47.5	61.9	74.3	0.813
Proxy	✓	✗	1.11	**45.0**	**58.5**	**71.0**	**80.9**	**0.890**
Proxy	✓	✓	1.29	44.7	58.2	70.7	80.6	**0.890**
Classification	✗	✗	**1.00**	6.1	11.5	21.1	35.9	0.887
Classification	✓	✗	**1.00**	43.8	57.8	70.3	80.3	0.887
Classification	✓	✓	1.27	44.7	58.4	70.7	**80.9**	**0.890**

Results for triplet prediction

Embedding space	Models	Normalization	Disentanglement	Genre	Mood	Instruments	Era	Overall
Complete space	Triplet	✓	✗	0.771	0.725	0.653	0.701	0.712
	Triplet	✓	✓	0.762	0.744	0.696	0.733	0.733
	Triplet + track reg.	✓	✓	0.757	0.733	0.673	0.715	0.720
	Proxy	✓	✗	0.774	0.742	0.645	0.693	0.714
	Proxy	✓	✓	0.762	0.742	0.660	0.716	0.720
	Classification	✗	✗	0.783	0.745	0.659	0.723	0.728
	Classification	✓	✗	0.776	0.747	0.647	0.704	0.719
	Classification	✓	✓	0.758	0.742	0.659	0.715	0.719
Sub-space	Triplet	✓	✓	**0.790**	**0.785**	**0.798**	**0.797**	**0.792**
	Triplet track reg.	✓	✓	0.775	0.748	0.743	0.742	0.752
	Proxy	✓	✓	0.777	0.740	0.734	0.700	0.738
	Classification	✓	✓	0.775	0.739	0.732	0.701	0.737

Visualization of Disentangled Space

- The highlighted samples are **relatively scattered** when considering **all dimensions**, but **well clustered** when considering only **the instrument sub-space.**



(a) All-dimensions (b) Instruments-dimensions