Harnessing the Power of Distributions: Probabilistic Representation Learning on Hypersphere for Multimodal Music Information Retrieval

Takayuki Nakatsuka, Masahiro Hamasaki, Masataka Goto
takayuki.nakatsuka@aist.go.jp
National Institute of Advanced Industrial Science and Technology (AIST)

Key Contributions

A) We leverage a **von Mises-Fisher (vMF) distribution** for multimodal probabilistic representation learning.

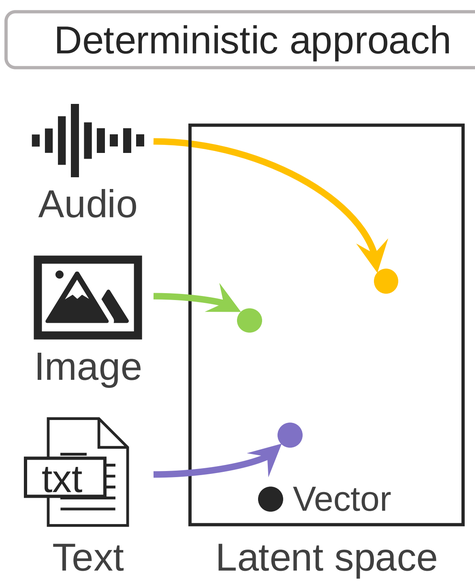
B) We design a **probabilistic contrastive loss function** and a loss function based on the **Optimal Transport (OT) distance** to facilitate multimodal probabilistic representation learning.

C) We confirm the effectiveness of integrating the probabilistic contrastive loss function with the OT-based loss function through **quantitative evaluations**.

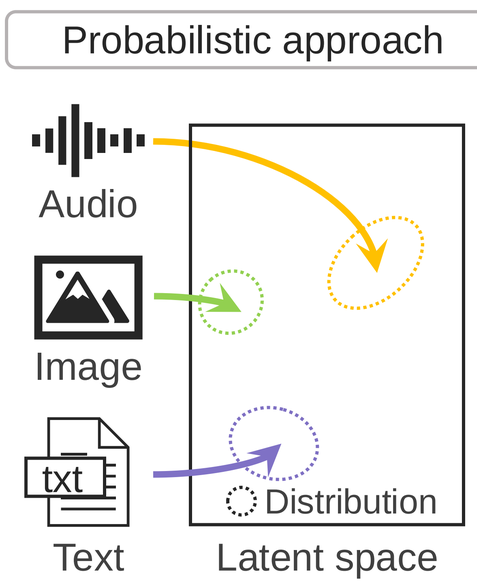
A) Harnessing the Power of Distributions

- Multimodal representation learning of music content has been an important topic of research, given its wide applications to Music Information Retrieval (MIR) tasks.
- To achieve such learning, we propose **multimodal probabilistic representation learning**, in which each content item is represented as a probability distribution in a latent space, for multimodal MIR.

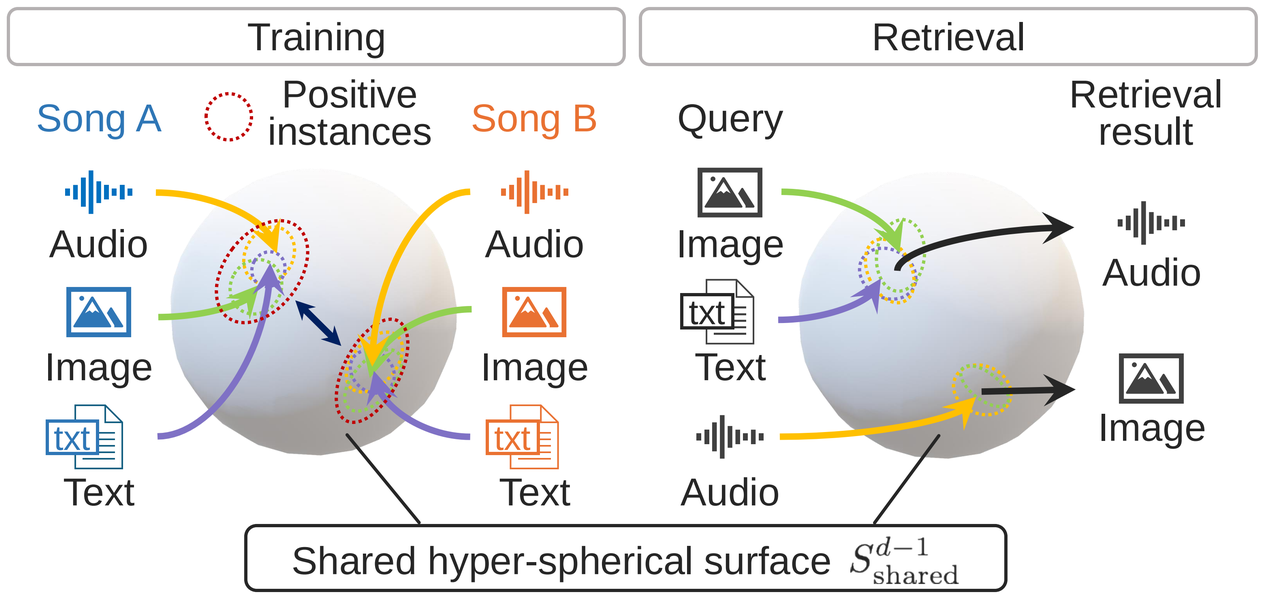
Deterministic approach



Probabilistic approach



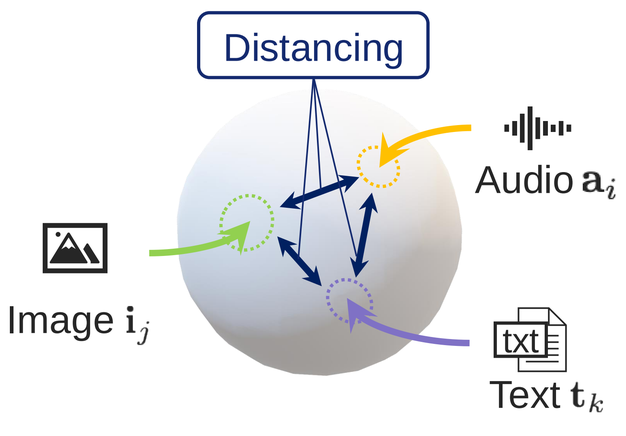
- We leverage the **vMF distribution**, which is a probability distribution defined on the hypersphere S^{d-1} in $\mathbb{R}^d$.
  - ✓ Methods using the vMF distribution outperform those using the Gaussian distribution (Li et al., 2021).
- We design encoders that map each content item to a hypersphere as the vMF distribution.
  - ◆ [Training] Encoders are trained so that the vMF distributions of the positive instances are close to each other on S^{d-1}_{shared}, while those of irrelevant instances are far apart.
  - ◆ [Retrieval] Given a single-modal query or a multimodal query such as a query that combines an image and text, our method can retrieve content items that match the query by calculating the distance between their distributions.



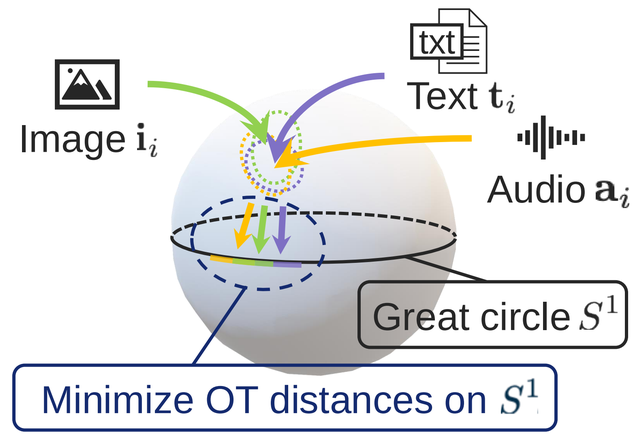
B) Multimodal Probabilistic Representation Learning

- We propose two novel loss functions, one based on **probabilistic contrastive learning** (Kirchhof et al., 2023) and the other on **Spherical Sliced-Wasserstein (SSW) p-distance** (Bonet et al., 2023), to be used together for multimodal MIR on a hypersphere.

Probabilistic contrastive learning



Optimal transport (SSW)

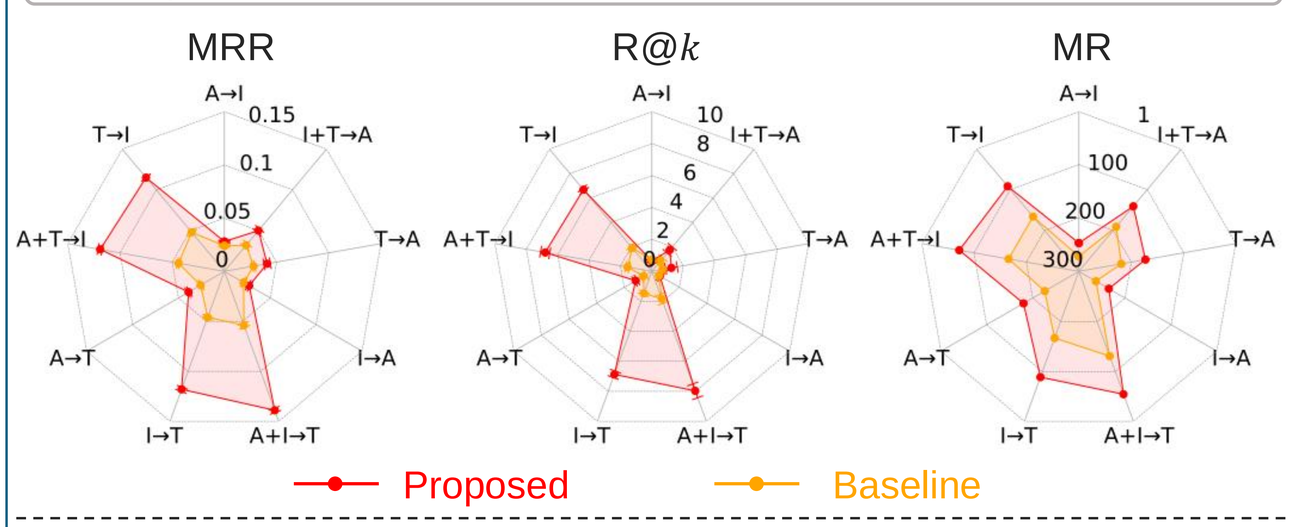


- Contrastive learning is an effective tool to jointly map each content item of multiple modalities to a shared latent space.
  - ✓ Methods using the angular distance between distributions has been shown to be more effective than those using the Euclidean distance (Scott et al., 2021).
- OT offers a robust and effective tool for calculating distances between probability distributions.
  - ✓ It allows the encoders to bring the probability distributions of a positive pair closer together, thus ensuring a more accurate representation learning.

C) Quantitative Evaluations

- We used three standard evaluation metrics for retrieval tasks: the mean reciprocal rank (MRR), the recall@k (R@k), and the median rank (MR).
- The results show that our method (Proposed) outperforms a baseline method (Baseline; solely used the loss function based on probabilistic contrastive learning) on MIR tasks.

Comparison on YT8M-MusicVideo test dataset for Multimodal MIR



Comparison on AS5M test dataset for Multimodal MIR