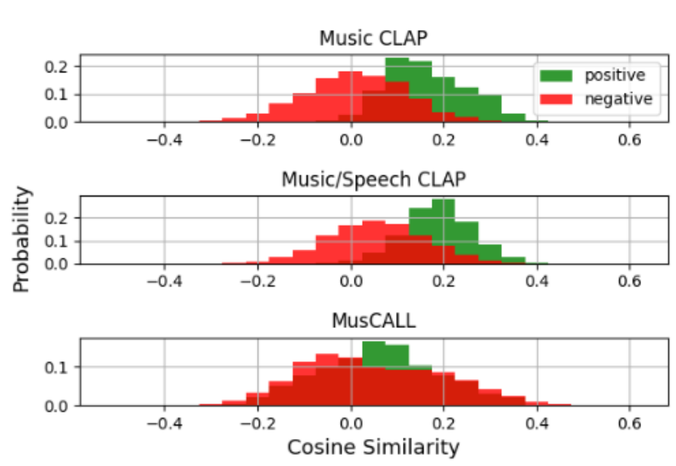
AIM Summer School 2024

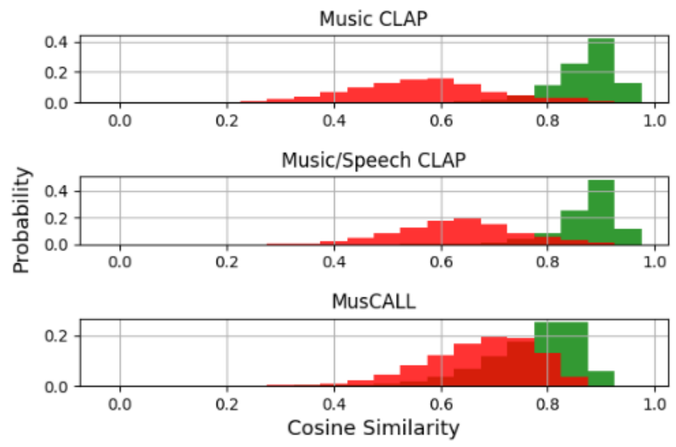
Multimodal User Adaptation

Yannis Vasilakis

2022



(a) Histogram of cosine similarity between TinySOL data and MusCALL prompts in joint audio-text space.



(b) Histogram of cosine similarity between TinySOL audio data and the mean of intra-class embeddings in joint audio-text space.

Finding

Alignment problems: 1) Use of captions as is without any augmentation leads to overfitting 2) BERTs inability to capture music related semantics compared to a taxonomy of concepts We proposed to apply stochastic augmentation (masking, paraphrasing, adding phrases) and musical fine-tuning The Audio encoder provides with informative representations.

Question

How can we effectively utilise external information for tag dependency for better alignment? How can we enforce balance between the modalities? Is intra-modal queries symmetric? How can we adapt the systems to a users concept definition with minimal annotations?

Supervisor(s): Johan Pauwels, Rachel Bittner