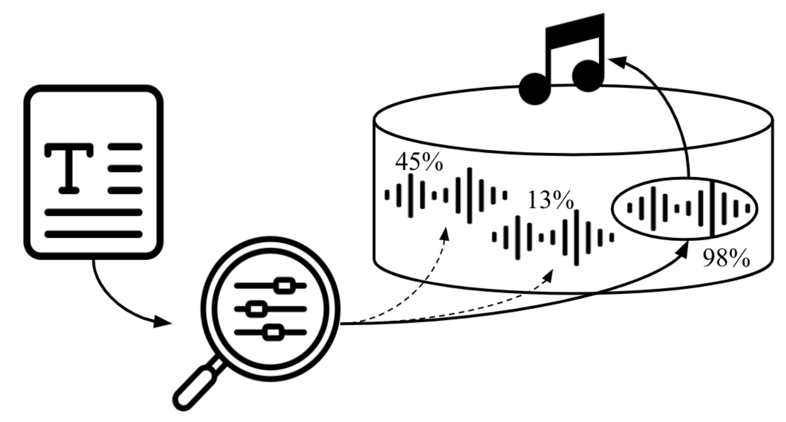
CONTRASTIVE AUDIO-LANGUAGE LEARNING FOR MUSIC

Ilaria Manco[1, 2], Emmanouil Benetos[1], Elio Quinton[2], George Fazekas[1]

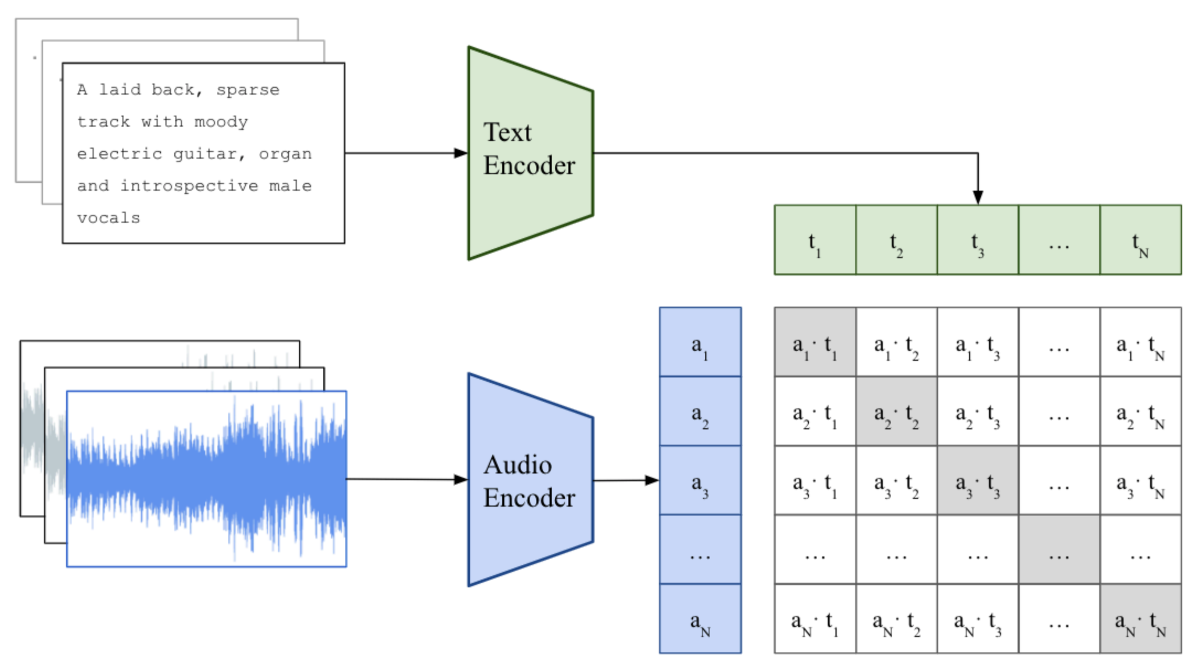
[1] Queen Mary University of London, [2] Universal Music Group

1 Bridging audio and language in the music domain via cross-modal learning

- Natural language queries offer a **convenient** and **human-friendly** way of searching for music, but they are not commonly supported by MIR systems



- With **MusCALL**, we propose to learn **audio-text correspondences** via **contrastive learning**, and successfully apply this to **cross-modal retrieval** for music, mapping natural language to audio and vice versa
- This ability to align audio and text can be **transferred** to music classification tasks such as genre classification and tagging in a **zero-shot** setting
- We adopt a **dual-encoder architecture** to process modalities independently and ensure scalability



2 Extending multimodal contrastive learning to music and language

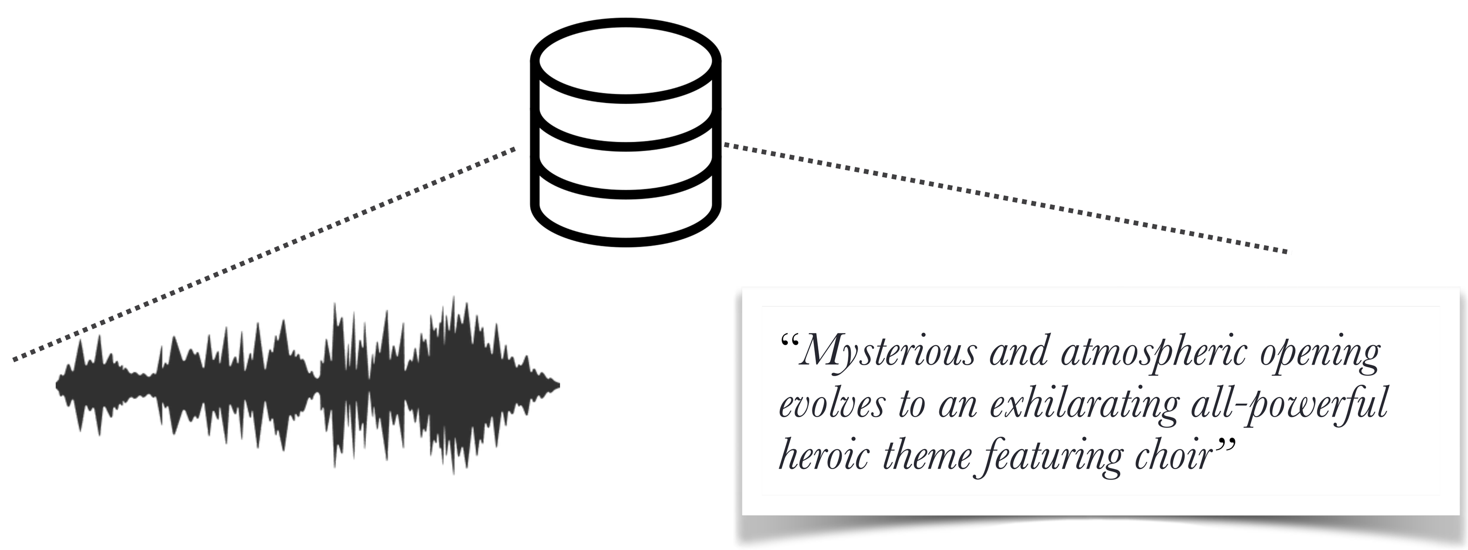
- MusCALL is trained via **multimodal contrastive learning**, where each component of the loss (audio-to-text and text-to-audio) is of the form

$$\mathcal{L}_{a\to t} = -\frac{1}{N}\sum_{i}^{N} \log \frac{\exp\left(z_{a,i}\cdot z^{+}_{t,i}/\tau\right)}{\sum_{z\in\{z^{+}_{t,i}, z^{-}_{t,i}\}} \exp\left(z_{a,i}\cdot z/\tau\right)}$$

(N: batch size; $z^{+}_{t,i}$ & $z^{-}_{t,i}$: embeddings of positive and negative text samples for item a_i; τ: temperature parameter)

- To mitigate some of the limitations in the data, we explore two variants:
  (1) Content-aware **loss weighting**
  (2) Combining a **self-supervised learning objective** with the multimodal contrastive loss

- MusCALL is trained on a dataset of ~250k **(audio, caption) pairs**
  - Audio: 20 seconds
  - Text: ~ 1 sentence describing the track



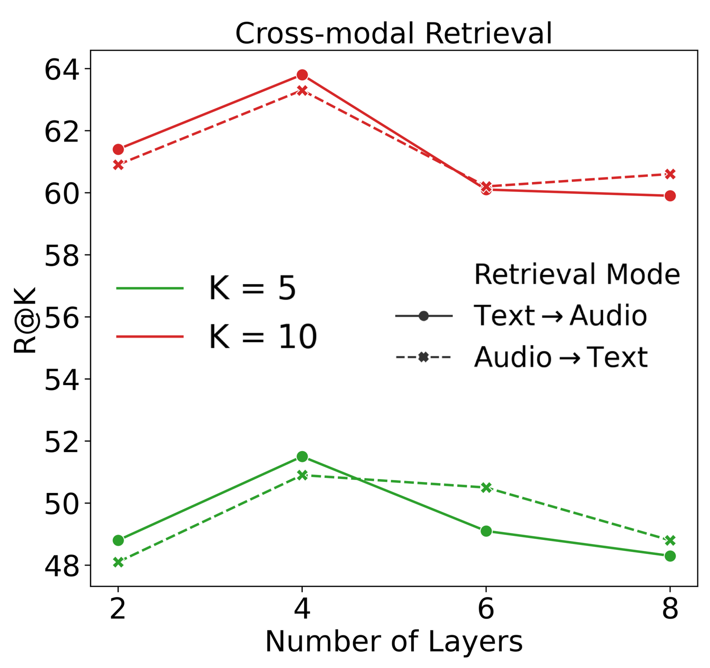
- Architecture: ResNet50 as the audio backbone and a lightweight (4-layer) Transformer as the text encoder

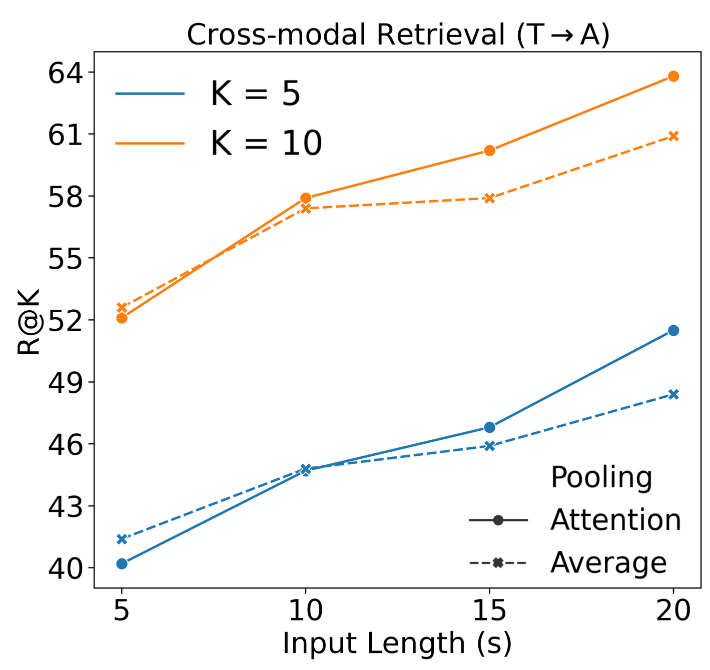
3 Experiments & Results

Text-to-Audio & Audio-to-Text Retrieval

Method	Text → Audio					Audio → Text				
	R@1	R@5	R@10	mAP10	MedR ↓	R@1	R@5	R@10	mAP10	MedR ↓
DCASE [56]	2.3	10.4	17.4	5.5	50	1.1	5.6	10.1	3.0	84
DCASE + CL	3.9	12.4	18.1	6.8	81.5	2.0	8.6	16.4	4.5	64
MusCALL (ours)	**25.9**	**51.9**	**63.3**	**36.0**	**5**	**25.8**	**53.0**	**63.0**	**35.9**	**5**

- MusCALL outperforms the DCASE baseline by a considerable margin even when using the same contrastive loss (DCASE + CL)



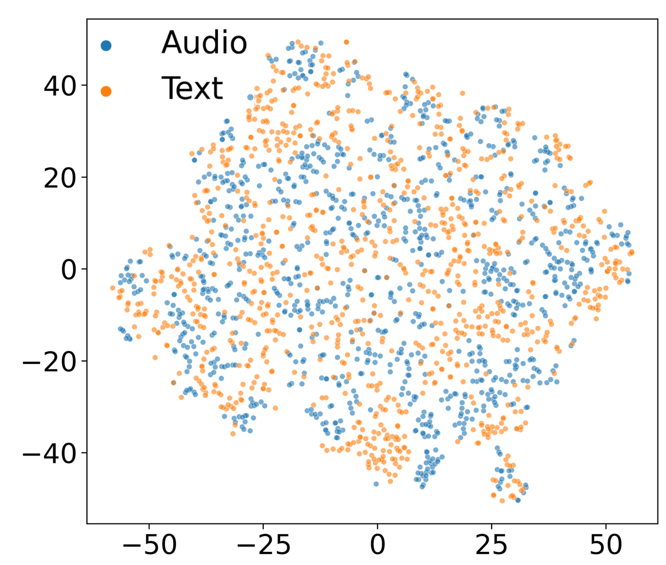


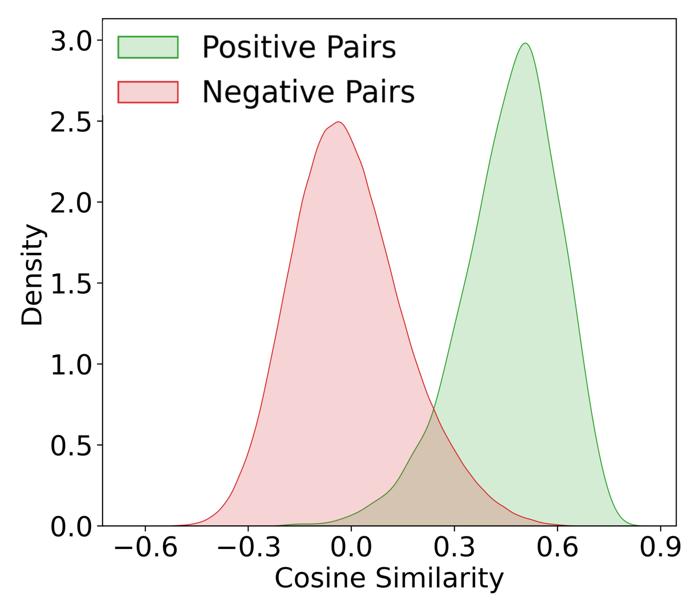
- Analysing the effect of the main design choices in MusCALL, we find that the model capacity needs to be adequately chosen: adding layers to the text encoder does not help past a certain point; providing the model with longer audio sequences is beneficial, particularly when this is done alongside using attention pooling

Zero-shot Transfer

Method	Prompt	Genre	Tagging	
		Acc.	ROC	PR
MusCALL$_{BASE}$	✗	55.5	**78.0**	28.3
MusCALL$_{BASE}$	✓	52.0	72.0	21.0
MusCALL$_{SSL}$	✗	58.2	**77.4**	**29.3**
MusCALL$_{SSL}$	✓	**62.0**	73.4	23.2

Qualitative Results





Query Text	Text of the Top-1 Audio
An atmospheric and introspective orchestral track featuring strings, piano, and synth.	*An inspirational and moody orchestral track featuring strings and choir.*
Deep chilled out space jazz with crisp beats and lush electronics.	*Jaunty swing featuring trumpet.*
Up tempo, pumping dance pop with female vocals.	*Quirky, fun, positive disco party music.*

This work was supported by UK Research and Innovation [grant number EP/S022694/1] and Universal Music Group

ISMIR 2022 BENGALURU
UK Research and Innovation

Contact: i.manco@qmul.ac.uk
arXiv: 2208.12208
Code: github.com/ilaria-manco/muscall