

# I CAN LISTEN BUT CANNOT READ: AN EVALUATION OF TWO-TOWER MULTIMODAL SYSTEMS FOR INSTRUMENT RECOGNITION

Yannis Vasilakis

Queen Mary University of London

i.vasilakis@qmul.ac.uk

Rachel Bittner

Spotify

rachelbittner@spotify.com

Johan Pauwels

Queen Mary University of London

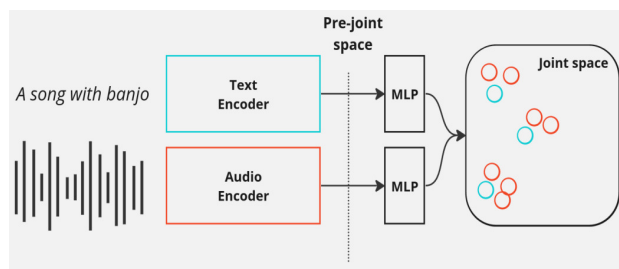
j.pauwels@qmul.ac.uk

## ABSTRACT

Music two-tower multimodal systems integrate audio and text modalities into a joint audio-text space, enabling direct comparison between songs and their corresponding labels. These systems enable new approaches for classification and retrieval, leveraging both modalities. Despite the promising results they have shown for zero-shot classification and retrieval tasks, closer inspection of the embeddings is needed. This paper evaluates the inherent zero-shot properties of joint audio-text spaces for the case-study of instrument recognition. We present an evaluation and analysis of two-tower systems for zero-shot instrument recognition and a detailed analysis of the properties of the pre-joint and joint embedding spaces. Our findings suggest that audio encoders alone demonstrate good quality, while challenges remain within the text encoder or joint space projection. Specifically, two-tower systems exhibit sensitivity towards specific words, favoring generic prompts over musically informed ones. Despite the large size of textual encoders, they do not yet leverage additional textual context or infer instruments accurately from their descriptions. Lastly, a novel approach for quantifying the semantic meaningfulness of the textual space leveraging an instrument ontology is proposed. This method reveals deficiencies in the systems' understanding of instruments and provides evidence of the need for fine-tuning text encoders on musical data.

## 1. INTRODUCTION

Multiclass classification has been a heavily researched topic in Music Information Retrieval (MIR) with many concrete applications such as genre, instrument and emotion recognition [1–4]. Despite the success of Deep (DL) systems for such tasks, recurring deficiencies persist among these problems. These are: (1) the limited availability of large-scale annotated datasets curated by experts, (2) the restricted capability of these systems to infer only a set of predefined classes.



**Figure 1:** Figure of a pipeline for two-tower multimodal systems. A separate model for each modality is used and their individual representations are projected to a joint audio-text space through a Multi-Layer Perceptron (MLP). This enables direct comparison between audio and textual data. We refer to embeddings obtained before joint-space projection as pre-joint space embeddings.

Music is an ever-evolving art form and as a result, there is an inherent need to make these systems adaptable to new terms/classes [5–7] and infer task-agnostic representations that can be useful for a plethora of downstream tasks with representation learning [8–10]. Zero-shot learning (ZSL) is focused on estimating a classifier capable of inferring unseen, new classes without annotated examples [11–14]. ZSL is often achieved in either of two ways: (1) by decomposing each class into attributes and inferring unseen classes through their related attributes (e.g. genres decomposed into presence or absence of instruments [12]) or (2) using word embeddings from Language Models (LM) [12–14]. The success of contextualized Large Language Models (LLM) has driven the research community predominantly toward the second solution, as it doesn't require experts to define attributes and the mapping between classes and attributes [15–17].

As a result, ZSL for audio classification is primarily focused on connecting the audio and semantic representation spaces. This interconnection can happen in 2 main ways: (1) mapping the audio representations to text space [12, 13], or (2) mapping both of the spaces to a new, joint audio-text space [14, 18–20]. The systems of the second category are named two-tower multimodal systems, where pre-trained audio models and LMs are used as the audio and text encoders respectively. Representations obtained from each modality are then mapped to a joint audio-text space and systems are jointly optimized such



that the audio and text representation are close in the joint space (e.g. the phrase “A rock song track” is similar to the recording of a “rock” song). We will call such representations as embeddings from here on.

This work aims to better understand the properties of existing two-tower systems. We use instrument classification as a case-study to provide insights into the presence (or absence) of semantic properties in the audio, text or joint spaces in addition to reporting classification metrics. Concretely, we consider 3 systems: MusCALL [18], a CLAP [20,21] model trained on speech and music datasets, and a CLAP model trained on music data [21]. We evaluate the performance of these systems on instrument classification using the TinySOL dataset [22]. We would like to highlight that multimodal DL models typically excel at simple tasks and datasets like this.

Furthermore, a novel approach for quantifying the semantic meaningfulness of textual encoders for instrument recognition is proposed.

For reproducibility, our experiments are performed on open-source datasets and the code of our experiments is made publicly available<sup>1</sup>, such that they can be reproduced.

## 2. RELATED WORK

### 2.1 Zero-shot transfer

ZSL focuses on estimating classifiers for novel, unseen classes without annotated examples. Two-tower systems are not primarily optimized for ZSL but due to the pre-trained textual encoder, novel words or phrases can be interpreted during inference. This property is known as zero-shot transfer (ZST) [23].

Side-information can be used in multiple ways that fall into two categories: (1) decomposing classes into shared attributes and (2) using LMs to represent this information as a text embedding. Despite its success, the first solution requires experts to effectively estimate the relevance of attributes and several classes and is a costly activity. As a result and due to the remarkable results obtained through contextualized LLMs, research has focused on the second option.

Generally, the methodology can be broken down into 3 components: (1) an audio encoder, (2) a textual encoder and (3) a projection to a common space. General purpose audio DL models that have been used as the audio encoder include VGGish [24], PANN [8], HTS-AT [9] and Audio Spectrogram Transformers [25]. For the textual encoder, distributional LMs like GloVe [26], Word2Vec [27] and contextualized LMs like BERT [28] have been thoroughly tested. As each modality produces heterogeneous representations, different methods of establishing comparability have been tested. This is predominantly achieved through projecting audio to text/semantic space [11–13] or a novel, joint audio-text space [18–20].

<sup>1</sup> [https://github.com/YannisBilly/i\\_can\\_listen\\_but\\_cannot\\_read](https://github.com/YannisBilly/i_can_listen_but_cannot_read)

### 2.2 Two-tower multimodal systems

Multimodal systems aim to represent data with additional knowledge from multiple modalities. Examples are audio combined with images [29], text [12, 30] or a combination thereof.

Two-tower multimodal systems focus on combining the textual and audio modalities by projecting them in a joint audio-text space. In that space, words that are relevant to a specific song will produce embeddings that will be close in terms of some similarity metric. An illustration of a two-tower system is presented in Figure 1. Information flowing through the audio encoder or textual encoder is referred to as the audio and textual branch respectively. We are also interested in the embeddings obtained through the encoders before projecting them into the joint audio-text space. We will call these the pre-joint spaces from now on.

The text used during training is usually a description of a song and will be referred to as a caption. The text used during inference will be referred to as a prompt.

MusCALL [18] combined a ResNET-50 for audio [31] with a Transformer for text encoding [32], optimized jointly over InfoNCE contrastive loss [33]. Additionally, a weighing mechanism based on caption-caption similarity was incorporated between negative audio-caption pairs. This is based on the premise that similar captions will be given to similar audio. The audio used is private but the training code is publicly available and used for this work.

MuLan [19] experimented with ResNET-50 as well as Audio Spectrogram Transformers [25] for audio encoding and a pretrained BERT model for the text branch. Both were jointly optimized over the Contrastive Multi-view Coding loss [34], which is a cross-modal extension of InfoNCE. Neither the data nor the code is available.

LAION-CLAP tested 6 different combinations of audio and text embedding models, the best one of which was HTS-AT with RoBERTa [35]. The latter is the one that will be used in this paper. The LAION-Audio-630k dataset was formed by combining AudioCaps, CLotho and Audioset.

Generally, research for two-tower systems is limited to testing different combinations of audio and text encoders, optimized jointly over a form of contrastive loss and modality fusion. We believe that closer inspection of their embeddings and evaluation protocol is needed.

## 3. EVALUATION OF TWO-TOWER SYSTEMS

### 3.1 Dataset and models

We use the TinySOL dataset which contains 2913 audio clips with a single note played from a single instrument out of a set of 14 instrument classes. This dataset has been chosen as it has consistent recording settings without noise, it is a simple dataset for instrument recognition and finally, confounding factors (compression, sampling rate etc) are minimized. We consider 3 models in total:

1. **Music/Speech CLAP:** [20, 21] A CLAP-based model trained on music/speech data<sup>2</sup>

<sup>2</sup> music\_speech\_epoch\_15\_esc\_89.25.pt

2. **Music CLAP**: [21] A CLAP-based model trained on music data<sup>3</sup>
3. **MusCALL**: A version of [18], retrained on music data<sup>4</sup>

We use the two pretrained CLAP systems provided by LAION<sup>5</sup>. For this work, the original MusCALL implementation was retrained from scratch, as both the data and trained models used in the original paper are not publicly available. Instead, we train on the LPMusicCaps-MTT [36] dataset, which is built by leveraging the audio and 188 tags from Magna Tag A Tune [37] to artificially generate captions through a GPT-3.5 model. The audio is resampled to 44.1 and 16 KHz for CLAP and MusCALL respectively, and the pre-processing steps described in their respective code repositories are followed.

### 3.2 Zero-shot transfer for instrument classification

Given an unseen audio segment  $x^*$ , a text label  $l^*$  and a two-tower system  $f(x)$ , we want to model the likelihood  $P(l^*|x^*)$  based on the embeddings provided by  $f(x)$ . In the general case,  $f \mapsto \mathbb{R}^F$  is a function that represents a two-tower system and maps audio or text information to a joint audio-text space, where  $F$  is the dimension of the joint space. Also, let  $\delta : (\mathbb{R}^F \times \mathbb{R}^F) \rightarrow \mathbb{R}$  be a function that measures similarity between joint space embeddings. In this approach, we model the  $P(l^*|x^*)$  based on  $\delta$ , as in:

$$P(l^*|x^*) \propto \delta(f(x^*), f(l^*)) \quad (1)$$

Multiclass classification attributes the most probable class to each recording and equivalently, the one that has the maximum likelihood. Given our approximation, the output class for each recording is:

$$c^* = \operatorname{argmax}_{c \in C} \delta(f(x^*), f(c)) \quad (2)$$

for  $c \in C$ , where  $C = \{c_1, c_2, \dots, c_N\}$  is the set of classes that we are interested in.

In our work, the embedding similarity function  $\delta$  is the cosine similarity:

$$\delta(e_1, e_2) = \frac{e_1 \cdot e_2}{\|e_1\| \cdot \|e_2\|} \quad (3)$$

where  $\|\cdot\|$  is the  $L_2$  norm and  $e_i \in \mathbb{R}^F$  are embeddings.

### 3.3 Experiment 1: Are two-tower systems context dependent?

Two-tower systems are typically not trained on single words, but rather longer prompts. As a result, the embedding produced for a single-word text label (e.g. “guitar”) can be very different from the embedding for a longer prompt, with additional context (e.g. “a guitar track”). When using two-tower models for classification, the class label can be wrapped in a prompt such as “A <label>

track” [18], to better match the training distribution. Methods to introduce stochasticity in the prompts used during training have been empirically proven to lead to more robust results [38]. Retraining the systems and testing different ways of augmenting captions used for training is left for future work, but works in image-text [39] and video-text [40] two-tower systems provide some evidence for their usefulness.

The impact of different approaches for giving additional context to a single-word text label during inference has not been well-explored. We explore the prompt sensitivity of each system by slightly changing the text prompt used for zero-shot classification in order to better understand to which extent these systems leverage contextual information. As far as we are aware, we are the first to evaluate the use of different types of prompts for two-tower systems during inference. Specifically, we evaluate 3 systems against 6 different prompts:

1. MusCALL prompt: “A <label> track”
2. Generated definition: “The <label> is a ...”
3. Generated definitions without label words: “The <removed> is a ...”
4. Label word with random context: “<label> <randomly selected lorem ipsum segment>”
5. Musically informed #1: “This is a recording of a <label>”
6. Musically informed #2: “Solo musical instrument sound of a <label>”

The first prompt proposed is the prompt that was used in MusCALL. The second prompt is generated using GPT-3.5 [41]. The third prompt is the same as the second but we removed all instances of the label itself to evaluate the influence of the context on its own. To evaluate if the systems are sensitive to specific words and to further evaluate if the context is useful, the fourth prompt adds random words alongside the label. Lastly, we test 2 musically informed prompts.

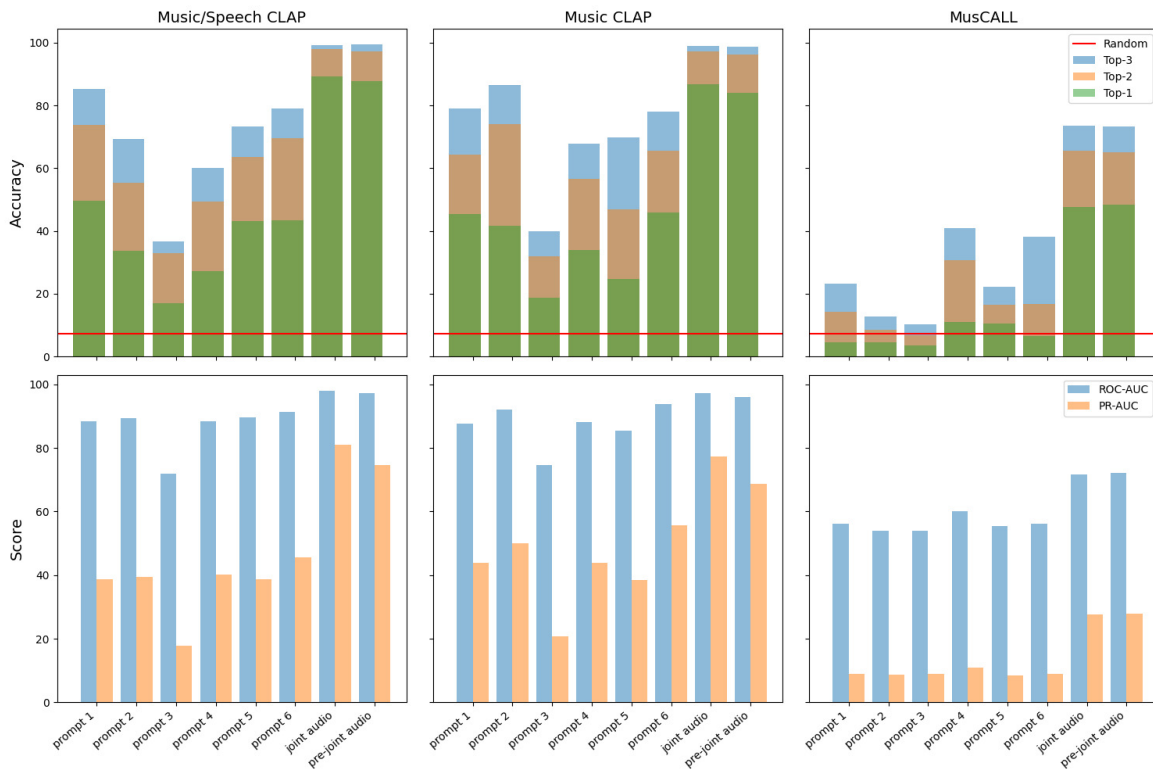
As a first metric, we consider Top-k accuracy with  $k = \{1, 2, 3\}$ . We calculate the cosine similarity between each recording and instrument prompt and sort them. We assign zero-shot class labels as described in Section 3.2 and check if the true label is present in the top-k assigned class labels. Furthermore, we calculate Receiver Operating Characteristic and Precision-Recall Area Under the Curve (ROC-AUC and PR-AUC respectively) following [42].

Figure 2 presents the zero-shot instrument recognition results for the three models across the 6 prompts, as well as the audio-only alternatives that will be described in Section 3.4. Despite the focus on music data, Music CLAP doesn’t display very different results from Music/Speech CLAP. While music-specific systems are generally expected to perform better, this is not the case for two-tower systems. This might be an indication that music requires special treatment, as the metrics approach the state of the art in audio-text [21] and image-text [43] two-tower systems.

<sup>3</sup> music\_audioset\_epoch\_15\_esc\_90.14.pt

<sup>4</sup> <https://github.com/ilaria-manco/muscall>

<sup>5</sup> <https://github.com/LAION-AI/CLAP>



**Figure 2:** Metrics for 6 textual prompts (See Section 3.3), 2 audio based label embeddings (See Section 3.4) and the 3 two-tower multimodal systems. The top row contains top-1 through top-3 accuracy and the bottom ROC-AUC and PR-AUC. The red line represents random choice.

Top-1 accuracy is worse than random for 4 out of 6 textual prompts for MusCALL. This might be caused by the small size of training data used, the absence of instrument-specific captions or their underrepresentation in the captions used, as well as the absence of single-note recordings in LPMusicCaps-MTT. While the metrics are low for MusCALL in most of the cases, a relatively large performance is still evident for the audio-only scenario. This implies that the problem lies in the audio-text alignment or the text branch.

The performance of CLAP models seems to be heavily correlated with the instrument labels themselves. Removing the label from definitions provides evidence that relevant context cannot be leveraged properly. Also, using musically informed prompts doesn't always result in greater or even comparable results. Specifically, top-1 accuracy drops when using the second musically informed prompt for CLAP models, despite the prompt being a more precise description of what is occurring in the audio.

These results suggest that CLAP models do not leverage extra context in the input prompt effectively. Both models performed worse when using relevant context without the instrument word, suggesting that the textual encoders put a lot more emphasis on the presence of specific words rather than the meaning of the prompts themselves. In addition, using a generic prompt provided better results than a musically informed one in most cases. Furthermore, any kind of context added at the prompts seems to harm the performance in most of the cases and provide more evidence that

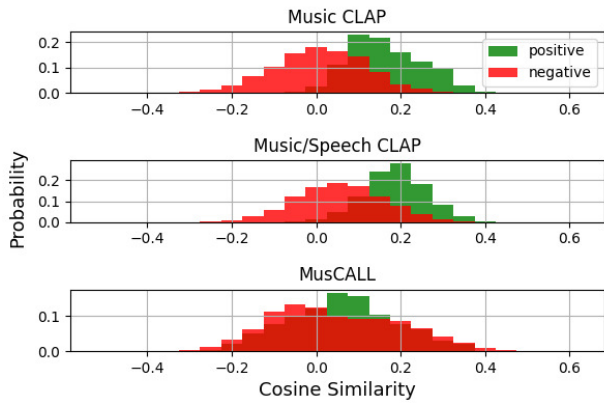
the model's text encoder cannot properly decompose the sentence to its constituents and use these semantically. Despite this observation, using definitions (prompt 2) seems promising for Music CLAP and for every metric apart from top-1 accuracy.

In the following experiments, we will consider only the "MusCALL prompt" as it leads to the highest top-1 accuracy, when model accuracy surpasses random choice.

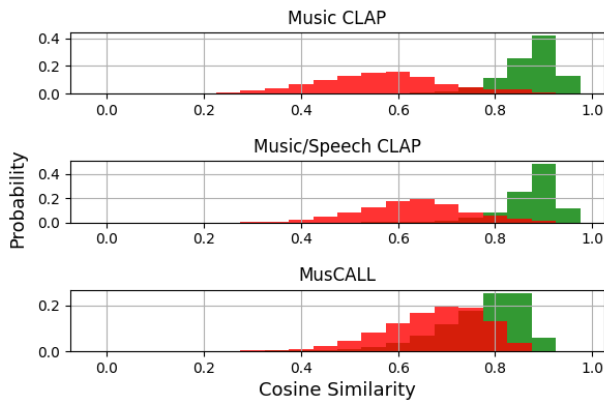
### 3.4 Experiment 2: Inspecting the cosine similarity distributions

As a next experiment, we calculate the cosine similarity between the joint space embeddings of each recording from TinySOL and the MusCALL prompt for each instrument, then compare the similarities of positive pairs vs negative pairs. We define a positive pair as an audio-label pair where the label corresponds to the instrument in the recording, and the negative pairs as all other pairs. Figure 3a presents histograms of similarities for positive and negative pairs when using text prompts. If the audio-text coherence is good, positive and negative histograms should be well separated.

Positive and negative similarity distributions overlap greatly, as can be seen in Figure 3a. As a result, retrieval is far from optimal. Fundamentally, a caption is a multi-faceted sentence. We suspect that treating a sentence as only one embedding point (mean of word embeddings) is fundamentally problematic and greatly hinders the semantic properties of the joint space. A hypothesis that needs



(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.

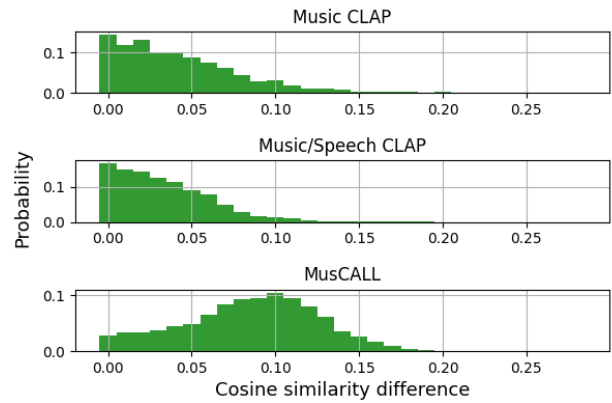
**Figure 3:** Histograms of audio and label embeddings for positive and negative pairs. When using textual prompts (a), the alignment is problematic, as can be seen from the overlap between positive and negative distributions.

testing is that by using composite sentences, a model cannot properly infer the relative embeddings of the sentence constituents.

To further evaluate if the audio encoder produces meaningful representations, we use the mean of intra-class song embeddings as the label embedding. This label embedding takes the role of the prompt embedding in the previous experiment. We generate the embeddings in joint audio-text space for each song. Then, we collect the songs that belong to  $k$ -th class  $c_k$  and estimate the mean of the embeddings. The latter serves as the optimal embedding that the text label would have to be mapped to in order to maximize performance and will be referred to as the “audio-only” label. Note that this embedding is only optimal in the case of TinySOL data.

The resulting histograms for positive and negative pairs are shown in Figure 3b. They are well separated, indicating that the audio-encoder itself produces meaningful, separable embeddings.

Audio embeddings seem to be of good quality before and after the projection to the joint audio-text space, as the metrics are almost equal before and after projection in Fig-



**Figure 4:** The histogram of top-2 class similarities for every song in TinySOL. The CLAP models tend to be not very confident while the metrics are greater than the over-confident MusCALL with the worst metrics.

ure 2. The metrics almost double when using any audio-only labels, which further provides evidence that the problem resides in the text branch, or joint-space projection and there remains a large performance gap to bridge.

### 3.5 Experiment 3: How confident are two-tower systems in their prediction?

We calculate the histogram of the difference between the top-2 candidate classes for each recording [44] to quantify the classification confidence. The similarity between each audio and instrument embeddings is estimated and they are sorted in descending order. The difference between top-2 similarities for each song is then calculated and a histogram of that difference is plotted in Figure 4.

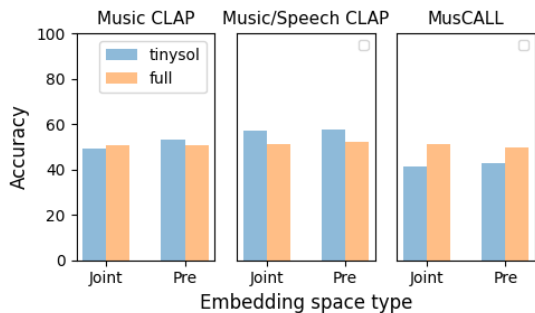
MusCALL seems to be overly confident in its prediction, which is unwarranted given the metrics reported. The opposite can be stated for CLAP models, where despite their better performance, the difference has a median value of 0.05-0.08.

### 3.6 Experiment 4: Quantitative evaluation of the text branch

While there are datasets that can be used to quantify the semantic properties and/or quality of a LM, there isn’t one that focuses on music. To overcome this lack of text data for the case of instrument recognition, we can utilize instrument ontologies, which encompass semantic similarity of instruments at multiple levels. We propose to leverage them to quantify the semantic similarity between different instruments and instrument families. In this experiment we use the instrument ontology by Henry Doktorski<sup>6</sup> (HDIO). As every instrument ontology has its limitations [45], repeating the same experiment with other ontologies is left for future work.

We extract the tree based on HDIO and form every possible triplet of TinySOL instrument labels in the tree for a total of  $364 = \binom{14}{3}$  combinations of positive word

<sup>6</sup> <https://free-reed.net/>



**Figure 5:** Semantic meaningfulness quantification leveraging Henry Doktorski’s instrument ontology. We evaluated the systems over valid triplets obtained through TinySOL labels, as well as every available triplet obtained from the ontology’s labels. Accuracy ranges from 49-59% which stresses that the models do not properly understand musical instruments in depth.

pairs without repetition. The triplets are of the form ( $\langle anchor \rangle$ ,  $\langle positive \rangle$ ,  $\langle negative \rangle$ ) where the  $\langle anchor \rangle$  label has to be more semantically similar according to HDIO to the  $\langle positive \rangle$  than the  $\langle negative \rangle$ , e.g. (“violin”, “violoncello”, “trumpet”). Subsequently, every ( $\langle anchor \rangle$ ,  $\langle positive \rangle$ ) pair that is linked through the root node of HDIO is excluded. The number of remaining triplets is 273. As a way to quantify semantic meaningfulness with respect to musical instruments, we calculate cosine similarity between the ( $\langle anchor \rangle$ ,  $\langle positive \rangle$ ) and ( $\langle anchor \rangle$ ,  $\langle negative \rangle$ ) pairs for each system. Triplets for which the similarity is higher for the first pair than the second are considered “correct”, and triplets where this is not the case are considered “incorrect”. We compute the accuracy score as the percentage of correct triples.

We repeat this procedure with every valid triplet from the full ontology, as opposed to just using the instrument labels appearing in TinySOL. This gives us  $\approx 443k$  triplets.

The accuracy for triplets from TinySOL and the full HDIO ontology are both presented in Figure 5. We see that half of the triplets are “incorrect” and this means that abstract semantic relations between instruments are not effectively captured in the textual branch, indicating a need for fine-tuning textual encoders on music related data. Note that the accuracy is roughly the same as we would obtain by creating arbitrary triplets, though it is important to highlight that several instruments and instrument categories are words that are not frequently used in English. Closer examination of the validity and usefulness of specific triplet cases (e.g. “stringed”, “plucked”, “violin”) is left for future work.

### 3.7 Experiment 5: Does joint space mapping introduce noise?

To further examine the origins of the problematic embedding alignment, we repeat zero-shot evaluation with audio-only labels described in Sections 3.4 and semantic mean-

ingfulness evaluation described in 3.6 with the embeddings in the audio space and text space before the joint audio-text space mapping.

A minor performance increment can be seen when using the joint embedding instead of the pre-joint audio embedding, as can be seen in the last two columns of Figure 2, apart from MusCALL where the metrics remain almost the same. We believe that the reduction in dimensionality of the joint space compared to the separate spaces is the underlying cause of these increments.

On the other hand, the accuracy based on HDIO remains the same, except for a decrement observed for Music CLAP and TinySOL subset of HDIO triplets, as can be seen in Figure 5. This could be an indication that the MLP can effectively map knowledge to the joint space. This is a further hint that potentially the problem lies in the LM used and fine-tuning might be needed to enforce musical semantics to be better represented.

## 4. CONCLUSIONS AND FUTURE WORK

In this paper, we evaluated 3 two-tower multimodal systems for instrument classification. We provided a zero-shot classification analysis and an elaborate evaluation of the audio and text embeddings in the pre-joint and joint audio-text spaces. We also proposed a novel way to quantify the semantic meaningfulness of text embeddings based on triplets derived from an instrument ontology.

Generally, experiments showed that audio encoders are of good quality and hence, the alignment issue might be traced back to the text branch and/or the joint audio-text space mapping. Therefore, a solution could be to freeze the audio encoder and map the text information to audio space. Also, further attention to modality imbalance [46] can be placed with weighing in negative and positive examples [18, 47–50]. Additionally, to avoid sensitivity towards instrument labels and the inability to leverage context, we propose to use text augmentation over captions or masking/removing the words from them. It is important to state that the relation between sentence and word embeddings is not as straightforward as with bag-of-words Language Models [51] and as a result, the way to utilize captions or put additional emphasis on their constituents have to be further tested.

As a result, using two-tower systems might not be very useful for multi-class scenarios, given the large overlap between positive and negative histograms of cosine similarities shown in our experiments. We believe that it is essential for a music terminology similarity corpus to be established. The benefits will be two-fold: (1) it will provide a useful way of quantifying the semantic meaningfulness of the textual branch for two-tower model and (2) it can serve as a baseline to quantify the need for music-informed fine-tuning. Last but not least, genre and emotion ontologies can be used to further evaluate the semantic meaningfulness of language models.

## 5. REFERENCES

- [1] N. Ndou, R. Ajoodha, and A. Jadhav, "Music genre classification: A review of deep-learning and traditional machine-learning approaches," in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2021, pp. 1–6.
- [2] D. Han, Y. Kong, J. Han, and G. Wang, "A survey of music emotion recognition," *Frontiers of Computer Science*, vol. 16, 2022.
- [3] M. Won, J. Spijkervet, and K. Choi, *Music Classification: Beyond Supervised Learning, Towards Real-world Applications*. <https://music-classification.github.io/tutorial>, November 2021.
- [4] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 303–319, 2011.
- [5] J. C. Lena and R. A. Peterson, "Classification as culture: Types and trajectories of music genres," *American Sociological Review*, vol. 73, no. 5, pp. 697–718, 2008.
- [6] A. van Venrooij and V. Schmutz, "Categorical ambiguity in cultural fields: The effects of genre fuzziness in popular music," *Poetics*, vol. 66, pp. 1–18, 2018.
- [7] R. Hennequin, J. Royo-Letelier, and M. Moussallam, "Audio based disambiguation of music genre tags," in *International Society for Music Information Retrieval Conference*, 2018.
- [8] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2019.
- [9] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2022.
- [10] L. Wang, P. Luc, Y. Wu, A. Recasens, L. Smaira, A. Brock, A. Jaegle, J.-B. Alayrac, S. Dieleman, J. Carreira, and A. van den Oord, "Towards learning universal audio representations," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4593–4597, 2021.
- [11] H. Xie, O. J. Räsänen, and T. Virtanen, "Zero-shot audio classification with factored linear and nonlinear acoustic-semantic projections," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 326–330, 2020.
- [12] J. Choi, J. Lee, J. Park, and J. Nam, "Zero-shot learning for audio-based music classification and tagging," in *International Society for Music Information Retrieval Conference*, 2019.
- [13] H. Xie and T. Virtanen, "Zero-shot audio classification via semantic embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1233–1242, 2020.
- [14] H. Xie and V. Tuomas, "Zero-shot audio classification based on class label embeddings," *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 264–267, 2019.
- [15] S. Pratt, I. Covert, R. Liu, and A. Farhadi, "What does a platypus look like? generating customized prompts for zero-shot image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 15 691–15 701.
- [16] Y. Ge, J. Ren, A. Gallagher, Y. Wang, M.-H. Yang, H. Adam, L. Itti, B. Lakshminarayanan, and J. Zhao, "Improving zero-shot generalization and robustness of multi-modal models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 11 093–11 101.
- [17] S. Goyal, A. Kumar, S. Garg, Z. Kolter, and A. Raghunathan, "Finetune like you pretrain: Improved finetuning of zero-shot vision models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 19 338–19 347.
- [18] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Contrastive audio-language learning for music," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022.
- [19] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. W. Ellis, "MuLan: A joint embedding of music audio and natural language," in *International Society for Music Information Retrieval Conference*, 2022.
- [20] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP: learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [21] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.
- [22] C. Emanuele, D. Ghisi, V. Lostanlen, F. Lévy, J. Fineberg, and Y. Maresz, "TinySOL: An audio dataset of isolated musical notes (5.0)," 2020.

- [23] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keysers, A. Kolesnikov, and L. Beyer, “LiT: Zero-shot transfer with locked-image text tuning,” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18 102–18 112, 2021.
- [24] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2014.
- [25] Y. Gong, Y.-A. Chung, and J. Glass, “AST: Audio Spectrogram Transformer,” in *Proc. Interspeech 2021*, 2021, pp. 571–575.
- [26] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543.
- [27] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *International Conference on Learning Representations*, 2013.
- [28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *North American Chapter of the Association for Computational Linguistics*, 2019.
- [29] D. Dogan, H. Xie, T. Heittola, and T. Virtanen, “Zero-shot audio classification using image embeddings,” *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 1–5, 2022.
- [30] X. Du, Z. Yu, J. Lin, B. Zhu, and Q. Kong, “Joint music and language attention models for zero-shot music tagging,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 1126–1130.
- [31] K. He, Y. Wang, and J. Hopcroft, “A powerful generative model using random weights for the deep image representation,” in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS’16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 631–639.
- [32] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Neural Information Processing Systems*, 2017.
- [33] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *CoRR*, vol. abs/1807.03748, 2018. [Online]. Available: <http://arxiv.org/abs/1807.03748>
- [34] Y. Tian, D. Krishnan, and P. Isola, “Contrastive multiview coding,” in *European Conference on Computer Vision*, 2019.
- [35] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A robustly optimized BERT pretraining approach,” *CoRR*, vol. abs/1907.11692, 2019.
- [36] S. Doh, K. Choi, J. Lee, and J. Nam, “LP-MusicCaps: LLM-based pseudo music captioning,” vol. abs/2307.16372, 2023.
- [37] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, “Evaluation of algorithms using games: The case of music tagging,” in *International Society for Music Information Retrieval Conference*, 2009.
- [38] S. Doh, M. Won, K. Choi, and J. Nam, “Toward universal text-to-music retrieval,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.
- [39] L. Zhen, P. Hu, X. Wang, and D. Peng, “Deep supervised cross-modal retrieval,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 386–10 395.
- [40] X. Wang, B. Ke, X. Li, F. Liu, M. Zhang, X. Liang, and Q. Xiao, “Modality-balanced embedding for video retrieval,” in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 2578–2582.
- [41] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.
- [42] M. Won, A. Ferraro, D. Bogdanov, and X. Serra, “Evaluation of cnn-based automatic music tagging models,” in *Proc. of 17th Sound and Music Computing*, 2020.
- [43] M. J. Mirza, L. Karlinsky, W. Lin, H. Possegger, M. Kozinski, R. Feris, and H. Bischof, “Lafter: Label-free tuning of zero-shot classifier using language and unlabeled image collections,” in *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 5765–5777.
- [44] C. C. Aggarwal, *Data Classification: Algorithms and Applications*, 1st ed. Chapman & Hall/CRC, 2014.



- [45] Şefki Kolozali, M. Barthez, G. Fazekas, and M. B. Sandler, “Knowledge representation issues in musical instrument ontology design,” in *International Society for Music Information Retrieval Conference*, 2011.
- [46] X. Wang, B. Ke, X. Li, F. Liu, M. Zhang, X. Liang, Q.-E. Xiao, and Y. Yu, “Modality-balanced embedding for video retrieval,” *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.
- [47] D. Oneață and H. Cucu, “Improving multimodal speech recognition by data augmentation and speech representations,” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4578–4587, 2022.
- [48] K. Margatina, G. Vernikos, L. Barrault, and N. Aletras, “Active learning by acquiring contrastive examples,” in *Conference on Empirical Methods in Natural Language Processing*, 2021.
- [49] J. Choi, S. Jang, H. Cho *et al.*, “Towards proper contrastive self-supervised learning strategies for music audio representation,” in *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2022, pp. 1–6.
- [50] S. Ma, Z. Zeng, D. McDuff, and Y. Song, “Learning audio-visual representations with active contrastive coding,” *CoRR*, vol. abs/2009.09805, 2020. [Online]. Available: <https://arxiv.org/abs/2009.09805>
- [51] M. Alian and A. Awajan, “Factors affecting sentence similarity and paraphrasing identification,” *Int. J. Speech Technol.*, vol. 23, no. 4, p. 851–859, dec 2020.