EXPLORING MODALITY-AGNOSTIC REPRESENTATIONS FOR MUSIC CLASSIFICATION

Ho-Hsiang WU[1], **Magdalena FUENTES**[1,2], and **Juan P. BELLO**[1,2]

[1]*Music and Audio Research Laboratory*, **New York University**, New York, NY USA
[2]*Center for Urban Science and Progress*, **New York University**, New York, NY USA

ABSTRACT

Music information is often conveyed or recorded across multiple data modalities including but not limited to audio, images, text and scores. However, music information retrieval research has almost exclusively focused on single modality recognition, requiring development of separate models for each modality. Some multi-modal works require multiple coexisting modalities given to the model as inputs, constraining the use of these models to the few cases where data from all modalities are available. To the best of our knowledge, no existing model has the ability to take inputs from varying modalities, e.g. images or sounds, and classify them into unified music categories. We explore the use of cross-modal retrieval as a pretext task to learn modality-agnostic representations, which can then be used as inputs to classifiers that are independent of modality. We select instrument classification as an example task for our study as both visual and audio components provide relevant semantic information. We train music instrument classifiers that can take both images or sounds as input, and perform comparably to sound-only or image-only classifiers. Furthermore, we explore the case when there is limited labeled data for a given modality, and the impact in performance by using labeled data from other modalities. We are able to achieve almost 70% of best performing system in a zero-shot setting. We provide a detailed analysis of experimental results to understand the potential and limitations of the approach, and discuss future steps towards modality-agnostic classifiers.

1. INTRODUCTION

Musical objects and concepts appear in different heterogeneous data modalities, including but not limited to audio, images, text and scores, where sonic, visual and tactile modalities contribute to the overall experience. However, most music information retrieval (MIR) research has largely focused on developing systems that interact with a single modality, requiring development of separate models for audio, image or text, and over simplifying the musical modeling. There are approaches that exploit multiple modalities [1–4], but existing multi-modal systems in the context of MIR require *coexisting* modalities as inputs [5–10], which is a big constrain for their deployment since it limits the scope of systems to only work when the modality they have been design for is at hand.

In a context of rapidly increasing availability of information in all forms (video, audio, text, etc) it is desirable that models are able to overcome this single-modality limitation and can interact with information in any common form, for instance, a system able to classify musical instruments by the way they look and sound. To the best of our knowledge, no existing model in the context of MIR can be used if one of those modalities is missing (e.g. if it was trained with audio and text, can not be used in a dataset with only audio).

Based on recent work [11] we hypothesize that *modality-agnostic* systems can be developed by learning joint representations from different modalities when they represent the same concepts. If the embedding of an image of a guitar and the sound of a guitar are similar to each other (i.e. grouped closely in the embedding space) but different from those of a piano, we can build classification systems that would work with either image or audio. This would allow to train models in settings where there is big amounts of data from one modality but not from the other, but still be able to work in both cases.

This type of approach, often called *translation* since it implies "translating" one modality to another (e.g. being able to retrieve an image with a description of it) has received renewed attention recently given the combined efforts of the computer vision and natural language processing communities, and has been gaining more interests in the MIR community [12–18]. Recently, it has been proposed to learn translated representations using self-supervision [11] which is very promising since it doesn't rely on human-annotated data, but has the drawback of requiring millions pairs of raw data to train embedding models from scratch. We propose an intermediate solution, to use pre-trained embeddings and only learn the translation between them in a self-supervised manner, as a way of relaxing the amount of computation time and data needed for training the system.

In this paper, we take the first steps towards modality-agnostic music classification. We focus on the problem of classifying musical instruments using audio and/or image. We investigate the use of pre-trained audio and image embeddings in combination with training translation models to obtain a joint representation, in a self-supervised setting. We use the learned representations to train modality-

agnostic classifiers in a supervised manner, and we investigate the performance of the classifier compared to its single-modality counterpart in different scenarios, including one with varying amount of data available from either modality. Our implementation is available in `https://github.com/hohsiangwu/crossmodal`.

2. METHOD

Our method is summarized in Figure 1. It consists of three different stages: 1) First, we select a set of pre-trained embeddings from both audio and image, and translate or project the pre-trained embeddings into a common space, either by training a translation model, or simply using principal component analysis (PCA) to convert both embeddings to the same dimension; 2) We then conduct a study to find the best combination by comparing configuration performances in cross-modal retrieval; and 3) We use the resulting joint embeddings to train a classifier in a supervised setting and study the performance of the different configurations (i.e. translation vs. PCA vs. single modality) with different amount of data from each modality. We explain the different stages of the method in the following.

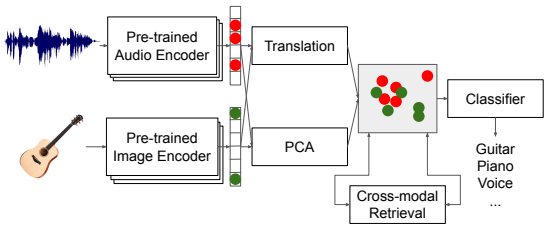

Figure 1. Method overview. The pre-trained audio and image embeddings are projected to a joint space by either the translation model or PCA, and the obtained embeddings are used to train the classifier in the downstream task: musical instrument classification. Cross-modal retrieval is used to select best configuration of pre-trained embeddings.

2.1 Pre-trained Image and Audio Embeddings

We select a set of state-of-the-art embeddings for both image and audio. For image embeddings, we use two pre-trained embedding models provided by *keras* library [1], trained using the ImageNet [19] dataset on a classification task. In particular, we use VGG Net [20] and ResNet [21]. These are both deep convolutional neural networks based architectures. We refer to [20, 21] for further details. For both models, we remove the last layer and apply average pooling to get the final image embeddings.

We also use pre-trained models to obtain the audio embeddings, particularly VGGish [22] , and YamNet. Finally, we use the open source implementation of OpenL3 [2] [4] trained with music data from AudioSet [23] to obtain another pair of image and audio embeddings.

[1] `https://keras.io/api/applications/`
[2] `https://github.com/marl/openl3`

Embedding model	# Parameters	Output dimension
OpenL3 (Image)	4.7M	8192
VGG16	15M	512
ResNet50	23.6M	2048
OpenL3 (Audio)	9M	6144
VGGish	62M	128
YamNet	3.2M	1024

Table 1. Overview of pre-trained image and audio embedding models.

In Table 1 we summarize the characteristics of each image and audio embedding model. Pre-trained VGG and ResNet image embeddings, VGGish and YamNet audio embeddings are trained on classification tasks, while OpenL3 is trained with audio-visual correspondence without labeled data.

To select the best combination, we evaluate how good the different pairs of embeddings blend together in a common space using a translation model. To quantify the success of this translation, we perform cross-modal retrieval (i.e. retrieve the image of an instrument using it's respective sound and vice versa) as further explained in Section 3.3. Our reasoning behind this is that for the modality-agnostic classifier to be successful, the embeddings should be very close to each other in the joint embedding space, and so they should be accurate in a cross-modal retrieval task when retrieving examples by distance. We select the best performing pair of audio and image embeddings and use it for the following stages.

2.2 Translation and Dimensional Reduction

We explore two ways of relating the audio and image embeddings: translation and a simple dimensional matching with PCA. For translation, in the self-supervised learning literature, various metric learning losses are used to learn a shared embedding space [24–26]. In particular, the *Contrastive loss* [27, 28] works well empirically with a careful selection of negative samples. It aims to minimize the distance of a given sample to positive examples (i.e. samples semantically related) while increasing the distance to negative examples concurrently. For the translation layer, we implement a 2 layer multi-layer perceptron (MLP) network, with pre-trained image and audio embeddings as inputs and train using contrastive loss, with cosine distance. We train the translation model using sample pairs from both modalities *without labels*, in particular we use the Musical Instruments AudioSet subset, as explained in Section 3. The output dimension is 128 for all of our embeddings.

As baseline, we apply PCA to each pre-trained embedding model to reduce their dimension to 128, and we train a single-modality classifier as well as a multi-modality classifier with such embeddings. The idea is to understand whether a simple solution is enough to build a modality-agnostic classification system, where the classifier is mainly responsible for the work of translating the modalities and learning the mapping to the labels.

2.3 Classification

We work with random forest classifiers. We train multi-modal (MM) classifiers either with translation (MMT) or PCA (MMP) and single-modality (SM) classifiers using dimension-reduced embeddings from audio (SMA) or image (SMI). We study how translation affects performance in scenarios with different amount of data used to train the classifiers. We do so by training the MMT and MMP with data from one modality and testing in the other, which we call *target modality*. We incorporate data from the target modality to the training of the classifiers by batches and see the impact in performance.

3. EXPERIMENTAL DESIGN

3.1 Dataset

Subset	Stage	# Samples
Translation	Train translation	130k
Cross-modal	Evaluate translation	10k
Classification	Train classifier	16.2k
	Evaluate classifier	1.8k

Table 2. Overview of subsets used for training and evaluation.

We use non-overlapping subsets of AudioSet [23] for the cross-modal experiments, the training of the translation model and the classifier. AudioSet is a multi-modal dataset containing YouTube videos with weak audio labels for a diverse set of real-world situations. We follow [11] and [4] by getting samples labeled at least with a descendant of "Musical instrument", "Singing" and "Tools". We then carefully split the dataset into three subsets, one for evaluating the pre-trained embedding combinations and select the best pair, another for training the translation model, and the last one for the downstream musical instrument classification task. From all qualified videos, we sample 1 second audio and one video frame as image within the second period. The assumption is that image and audio from roughly the same timestamp contain highly related semantic content. For evaluating the cross-modal retrieval experiments, we use a total 10k image/audio pairs. We call this subset the *cross-modal-subset*. We use 130k pairs to train the translation model, which is roughly half the amount of data used to train end-to-end models in [11]. We call this the *translation-subset* as shown in Table 2.

For the classification task, we carefully curated samples from 18 classes. Our categories include mapping from "Violin, fiddle" to "violin", "Choir" to "voice" and both "Drum" and "Drum kit" to "drums", and the remaining are "accordion", "banjo", "cello", "clarinet", "flute", "guitar", "mandolin", "organ", "piano", "saxophone", "synthesizer", "trombone", "trumpet", "ukulele", "cymbals". We manually audited the quality of the test set removing irrelevant samples (e.g. those labeled by piano but with image of an album cover with no piano on it) until we had 1,000 samples per instrument. Having a balanced dataset for the training of the classifier is important to prevent issues at this stage interfering with the assessment of the embeddings performance. We formulate the classification problem as a multi-class problem, where samples labeled only once from the above categories are selected. The result classification-subset consists of a balanced dataset with 16200 training samples and 1800 testing samples (10% split). We call it the *classification-subset*.

3.2 Model implementation

For ResNet and VGG image embeddings, we use the *preprocess* function provided from keras to normalize the pixel values. And we follow the pre- and post-processing steps of VGGish and YamNet [3]. For training the translation model, we do not use labels. Instead we randomly sample batches of size 4096 from translation-subset, extract pre-trained embeddings from both modalities, train a 2 layer MLP with both input dimensions as original pre-trained embeddings, 256 middle dimension, and 128 output dimension, implemented with PyTorch [4]. We use pairs of both modalities sampled from the same clip as positive examples, and other samples in the same batch as negative examples, with margin value as 1.0. We optimize using Adam optimizer with learning rate as 0.001, and we apply early stopping criteria on validation loss with patience as 5 epochs. For cross-modal retrieval, we take the outputs of translation model with corresponding pre-trained embeddings of the 10k image/audio pairs from cross-modal-subset, and use all embeddings from one modality as queries to fetch top 30 closest embeddings from another modality. For training the classification model, we use a random forest classifier from scikit-learn [5], with maximum depth set to 32, and 100 estimators.

3.3 Evaluation metrics

For the evaluation of the cross-modal retrieval results we follow the setup from [11]. We use *normalized discounted cumulative gain* (NDCG) score considering 30 elements. This score is a measure of ranking quality between 0 and 1 (from low to high quality), which assesses the gain of an element based on a relevance score and its position in the result list. Following [11], we use the relevance $r = C - d$, where d is the distance in the taxonomy graph between two labels in the AudioSet ontology, $C = 21$ being the maximum distance. As the AudioSet ontology is defined, the top labels (e.g. "Music", "Musical instrument", "Tools", "Singing") are included in computing the relevance, which make most of the example relevant since most of them convey one of those labels. [6] Therefore, we removed those top labels while computing the NDCG. We report the results of audio-to-image and image-to-audio retrieval.

For the evaluation of the classification results we use the macro F-measure or F1 score. Finally to assess the structural properties of the embedding spaces generated by the translation or the PCA projections we compute

[3] https://github.com/tensorflow/models/tree/master/research/audioset

[4] https://pytorch.org/

[5] https://scikit-learn.org/

[6] See https://research.google.com/audioset/ontology/index.html for further details.

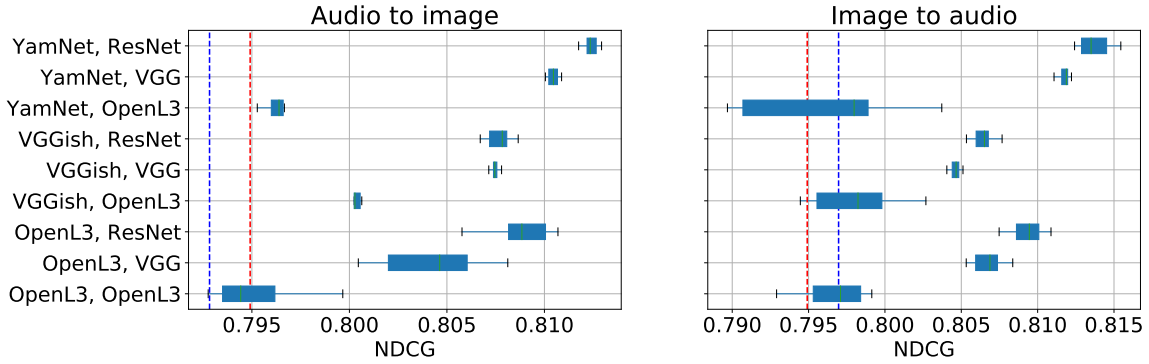

Figure 2. Cross-modal retrieval results with NDCG scores on the x-axis. On the y-axis we have different combinations of *audio, image* pre-trained embeddings used to train the translation model. The red dotted line is a random baseline, and the blue dotted line is OpenL3 (512 dimensions) for both image and audio without translation. We show cross-modal retrieval results of audio to image (left), and image to audio (right).

inter-cluster distances from the clusters of the different instrument classes and modality pairs before classification. For that, we take the test split of the classification-subset, compute average (centroids) of all the projected embeddings (PCA or translated) with the same modality and same instrument labels, and then compute pair-wise distance among modality/instrument clusters as the inter-cluster distance.

4. RESULTS AND DISCUSSIONS

4.1 What combination of pre-trained embeddings?

In this experiment we would like to determine which is the best audio-visual embeddings combination. We will have to simultaneously answer whether we are able to learn meaningful joint-embeddings from this data using translation, or if translation of pre-trained embeddings does not work at all.

To do so, we take all combination of audio and image embeddings and train the translation model with them, obtaining a total of nine separated translation models, i.e. nine mappings to joint embedding spaces. We then evaluate them using cross-modal retrieval in the cross-modal-subset explained in Section 3.1. The NDCG scores of different configurations are depicted in Figure 2, where both audio to image, and image to audio are shown. Following the ideas in [11], we use two baselines: random (red dotted line in Figure 2) which means randomly ordering the embeddings and get the first 30, and the OpenL3 (blue dotted line) image and audio embeddings both with 512 dimension [7], used for retrieval directly without translation.

The first observation is that the relative difference in NDCG scores between the baselines and our best performing model are comparable to those shown in [11], which is promising because it means that the translation model is effectively learning to relate the embeddings. Also unlike the systems in [11] which were trained from scratch, we obtained our joint embedding by translating pre-trained embeddings, and obtained competitive results. The random baseline performs better than OpenL3 for audio to image retrieval, which is consistent with the results reported in [11].

We observe a big gap in performance in all combinations that include the OpenL3 image embedding, which can be partially explained by the fact that VGG and ResNet greatly outperform that embedding in image classification downstream tasks, and thus more expressive embeddings would be better candidates for translation.

Overall, the combination of YamNet and ResNet performs the best across all configurations. We checked that is the case for the classification performance as well, therefore, we discuss only YamNet and ResNet results in the rest of the experiments.

4.2 How does translation affect performance?

We want to understand how translation affects the performance of a classifier in comparison to its multi-modal non-translated and single-modality counterparts. For that we compare their performance in the classification task, by training the classifier using embeddings from one modality and testing with embeddings from the other, and by adding batches of the training modality with balanced number of instrument classes by bits. We use the classification-subset for this experiment. The results of this process are shown in Figure 3, where the macro F1 scores are reported for a test set of only images (left) and only audio (right).

No data from target modality. First, we discuss the results in the 0 point of the x-axis, corresponding to the performance of the classifiers without any data from the target modality (e.g. when testing in image, only training the MM classifiers with audio). We see a similar and expected behaviour in both modalities: the MMP classifier is guessing some classes right (very little), probably exploiting some unintended relations between YamNet and ResNet embeddings after PCA, and the MMT classifier clearly outperforms the others, being able to achieve al-

[7] Note that the OpenL3 implementation allows for multiple output dimensions, and we choose 512 here for both embeddings to be comparable.

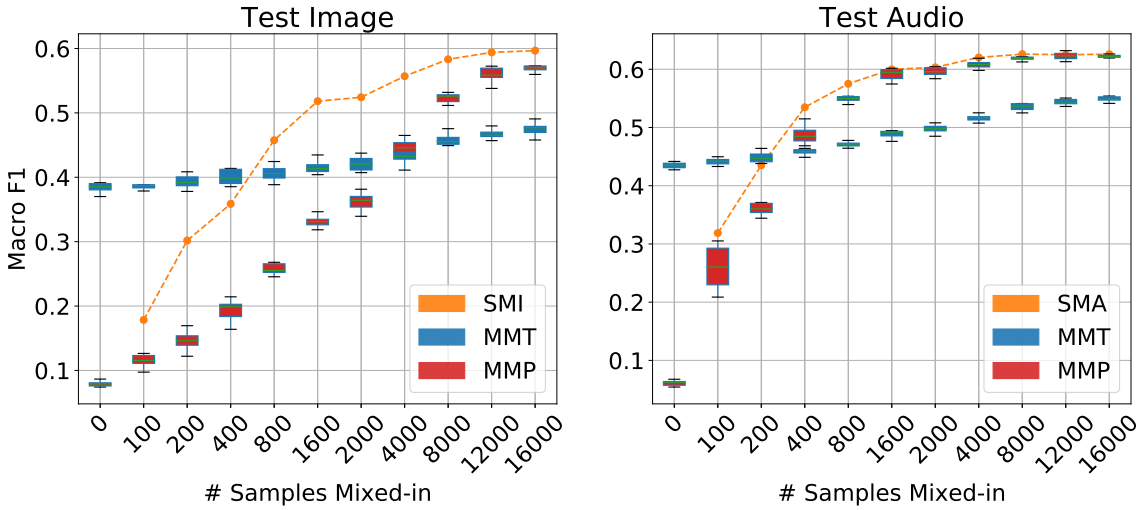

Figure 3. Instrument classification results in different modalities: image (left) and audio (right). The orange dashed line is a SM classifier trained and tested with same modality, image (I) or audio (A). Blue (translated) and red (PCA) boxes are MM classifiers trained with data from different modality to the test set, with the x-axis indicating the number of samples from the target modality mixed-in in during training (e.g. when testing on image, MMT and MMP are trained with audio, and image data is mixed in by bits).

most 70% of the best performance already. This confirms what we saw from the cross-modal retrieval examples, that the translation is doing a meaningful mapping, and further this allows to learn from one modality and test in another in a zero-shot fashion.

Adding data from target modality. However, for an ideal translation, image and audio embeddings would be interchangeable. That means that the performance of MMT without seeing any embedding from the target modality or after seeing all of them should be the same (since no *new* data would be added to the classifier). And so, the blue curve we see in Figure 3 with a small slope should flat at the maximum performance independently of the test data we add in. The fact that the performance of MMT increases by adding this data is showing that the translation failed in combining *some* meaningful information. And this makes the MMT classifier's performance to fall behind when all the data from both modalities are available (point 16000 in the x-axis of Figure 3).

Observing the performance of MMP, we also see that with the right amount of data and without translation, the classifier is able to learn the mapping between embeddings and classify the instruments correctly. This is an interesting result since it implies that for specific tasks with available labeled data from both (or multiple) modalities, it is enough to train a classifier to learn to deal with different modalities all together and be able to work with whatever modality is available at inference time with almost the same performance than a classifier fully dedicated to one modality.

4.3 What is translation doing?

We want to understand what is happening during translation that the MMT classifier is struggling to keep up with the others when enough data from both modalities is available, and why it does not reach best performance starting from the zero-shot setting. We compare the structure of the non-translated and translated embeddings before feeding them into the classifier. In particular we measure the distance between the cluster centroids of the different clusters of classes in each setting. Figure 4 shows the pairwise distance for the different modalities and instrument classes. On the left we see the non-translated embeddings sorted by modality, and on the right we see translated embeddings sorted by instrument class. The two figures show that the embedding spaces are indeed different, and that the translation is structuring and bringing together the audio and image embeddings of the same class (shown as small 2x2 squares on the diagonal), i.e. grouping the embeddings by concepts. This makes sense and explains why the translation works in the zero-shot setting, and is interesting considering that the translation layer is trained in an unsupervised manner.

However, there are noticeable small distance square blocks in both images: the one in the left on the top left, shows that the audio embeddings are closer to each other, which is an artifact of YamNet embeddings. This is not what happens with the ResNet embeddings, which after the PCA projection are sometimes closer to other image embeddings but also sometimes closer to audio embeddings. An exception to this are the image embeddings for voice, cymbals and synthesizer, which are very different from all the other embeddings.

The other big block where embeddings are grouped to-

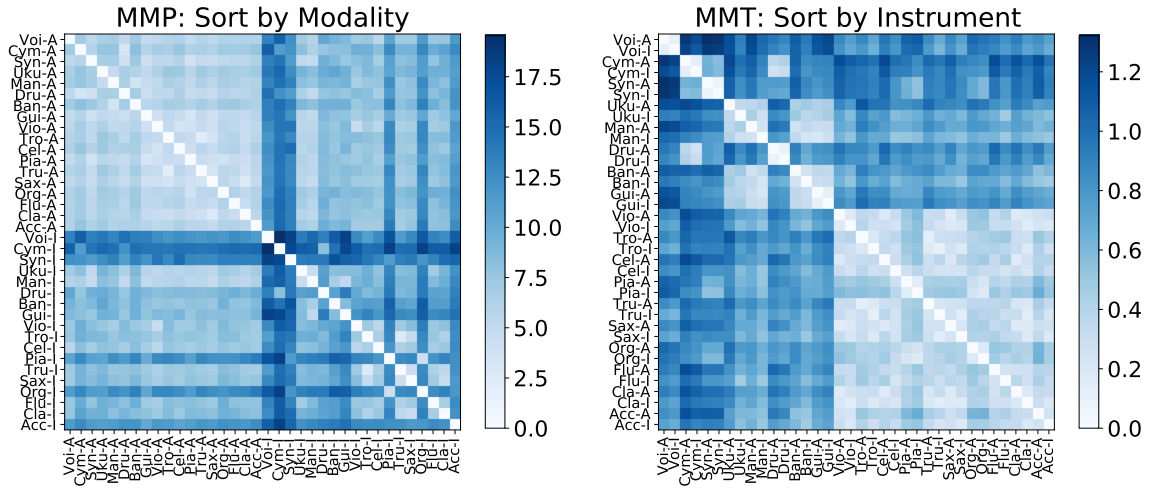

Figure 4. Pair-wise distance of inter-cluster centroid for modality and instrument classes. On the left we have PCA sorted first by modality. On the right we have translated sorted first by instrument class. Note: This is results of classification-subset but before training the classifier.

gether, in the right image of Figure 4, shows that the translation is bringing together some classes (e.g. accordian, clarinet, flute, organ, saxophone, etc.) that should not be blend together, and the overall distances in the translated embedding space are smaller than in the non-translated one. Observing the class distribution of the data used for training the translation model in Figure 6 (note the labels were not use in the training, only here for the analysis), we observe it is skewed and the classes with fewer number of instances correlate with most of the confused ones in the translated embeddings. We think that this is probably causing the classification performance of MMT to drop with respect to MMP and SM. The exception is voice, which we speculate has to do with an effect from the ResNet embedding which is very different from all other embeddings and we suspect that helped the translated cluster to be sufficiently different as well.

To see the correlation between our observations in the embedding space and the classification performance, we look at the per instrument F1 and confusion matrices of the MMT classifier, using all data from both modalities as shown in Figure 5. In each figure we show the per instrument F1 on the left, and the confusion matrix of MMT on the right. Looking at the per instrument F1 on the left, we can also observe the correlations between less performant instrument classes with smaller distances in Figure 4 and number of fewer samples in Figure 6. Looking at the confusion matrix of MMT on the right, we see that there are common mistakes of cymbal vs drum (both modalities), trombone vs trumpet (both modalities), guitar vs mandolin (mostly image), and organ vs piano (mostly image), which make sense because of the acoustic or visual similarity of those instruments in each modality. We observe a similar trend in the confusions made by the MMP classifier, but in a lesser extent (which explains the better performance).

To sum up, we believe that the bias of label distribution

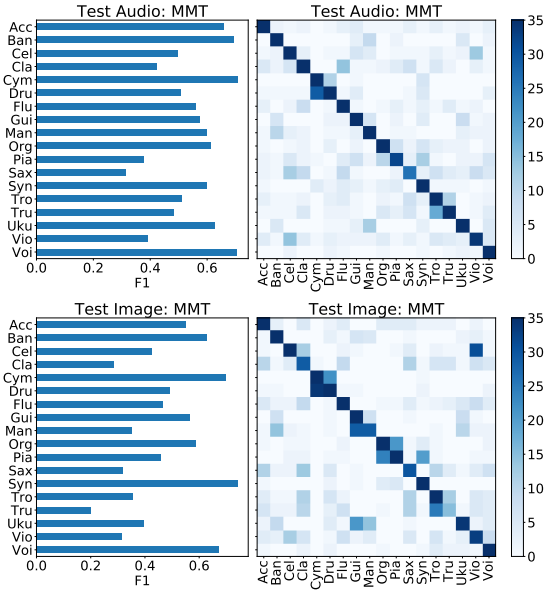

Figure 5. Classification results: training with all data using both modalities and testing in audio (top) and image (bottom). Per instrument F1 on the left, confusion matrix on the right.

in the data we used for training translation is the main cause of the performance drop in classification, and this is a trade-off of self-supervised learning without using the labels. We plan to explore in the future unsupervised methods for sample selection to balance the training set used for the translation model, such as determinant point processes [29].

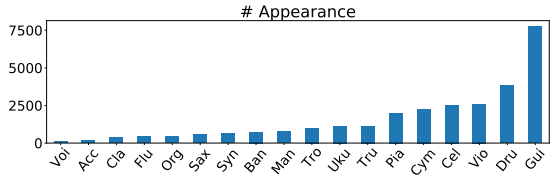

Figure 6. Number of samples containing classification instrument labels in translation-subset. We show here only the classes under study in the classification task.

5. CONCLUSIONS AND FUTURE WORK

In this work we propose and investigate modality-agnostic representations for music classification. We first present a study on different combinations of pre-trained audio and image embeddings to determine the best configuration to obtain modality-agnostic representations via cross-modal evaluation. We then use this representation to train instrument classifiers, comparing with non-translated and single-modality baselines. We show promising results as well as interesting potential applications using data from one modality to train and another modality to test with reasonable performance (almost 70% of best performing system in a zero-shot setting). We also investigate how biases in the training data used for the translation affect the classification performance.

For future work, we are interested in exploring sampling methods [30, 31] that could help balance the training set to obtain a more unbiased translation model, which from our analysis could lead to better performance. Also, we are interested in exploring the joint training of the translation and classification models, instead of the sequential method proposed in this paper. Furthermore, we think that exploring novel loss functions specifically for multi-modal data will also be an interesting direction as most of the current contrastive methods are applied to single modality. This work presented first steps and analysis towards the use of modality-agnostic representations in music, which we consider to be a promising idea in the context of MIR since it allows the use of data from different datasets and modalities in a flexible way, relaxing concerns about data scarcity and other data-availability related issues.

Acknowledgments

This work is partially supported by the National Science Foundation award #1544753. Magdalena Fuentes is a faculty fellow in the NYU Provost's Postdoctoral Fellowship Program at the NYU Center for Urban Science and Progress and Music and Audio Research Laboratory.

6. REFERENCES

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[2] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[3] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt *et al.*, "From captions to visual concepts and back," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1473–1482.

[4] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3852–3856.

[5] A. Yazdani, K. Kappeler, and T. Ebrahimi, "Affective content analysis of music video clips," in *Proceedings of the 1st international ACM workshop on Music information retrieval with user-centered and multimodal strategies*, 2011, pp. 7–12.

[6] A. Schindler and A. Rauber, "An audio-visual approach to music genre classification through affective color features," in *European Conference on Information Retrieval*. Springer, 2015, pp. 61–67.

[7] O. Slizovskaia, E. Gómez, and G. Haro, "Musical instrument recognition in user-generated videos using a multimodal convolutional neural network architecture," in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, 2017, pp. 226–232.

[8] S. Oramas, F. Barbieri, O. Nieto, and X. Serra, "Multimodal deep learning for music genre classification," *Transactions of the International Society for Music Information Retrieval. 2018; 1 (1): 4-21.*, 2018.

[9] Z. Duan, S. Essid, C. C. Liem, G. Richard, and G. Sharma, "Audiovisual analysis of music performances: Overview of an emerging field," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 63–73, 2018.

[10] J. Choi, J. Lee, J. Park, and J. Nam, "Zero-shot learning for audio-based music classification and tagging," in *The 20th International Society for Music Information Retrieval Conference (ISMIR)*. International Society for Music Information Retrieval Conference (ISMIR), 2019, pp. 67–74.

[11] R. Arandjelovic and A. Zisserman, "Objects that sound," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 435–451.

[12] M. Dorfer, J. Hajič Jr, A. Arzt, H. Frostel, and G. Widmer, "Learning audio–sheet music correspondences for cross-modal retrieval and piece identification," *Transactions of the International Society for Music Information Retrieval*, vol. 1, no. 1, 2018.

[13] Y. Zhang, B. Pardo, and Z. Duan, "Siamese style convolutional neural networks for sound search by vocal imitation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 429–441, 2018.

[14] K. Lee and J. Nam, "Learning a joint embedding space of monophonic and mixed music signals for singing voice," in *The 20th International Society for Music Information Retrieval Conference (ISMIR)*. International Society for Music Information Retrieval Conference (ISMIR), 2019, pp. 295–302.

[15] B. Li and A. Kumar, "Query by video: Cross-modal music retrieval." in *ISMIR*, 2019, pp. 604–611.

[16] K. Watanabe and M. Goto, "Query-by-blending: A music exploration system blending latent vector representations of lyric word, song audio, and artist." in *ISMIR*, 2019, pp. 144–151.

[17] F. Zalkow and M. Müller, "Learning low-dimensional embeddings of audio shingles for cross-version retrieval of classical music," *Applied Sciences*, vol. 10, no. 1, p. 19, 2020.

[18] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 609–617.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[23] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[24] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 539–546.

[25] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[26] J. Wang, F. Zhou, S. Wen, X. Liu, and Y. Lin, "Deep metric learning with angular loss," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2593–2601.

[27] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[28] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[29] A. Kulesza and B. Taskar, "Determinantal point processes for machine learning," *Machine Learning*, vol. 5, no. 2-3, pp. 123–286, 2012.

[30] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, "Sampling matters in deep embedding learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2840–2848.

[31] M. Won, S. Oramas, O. Nieto, F. Gouyon, and X. Serra, "Multimodal metric learning for tag-based music retrieval," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 591–595.