Interaction-Based Material Recognition with Multimodal Audio-Visual Datasets

Emily Liu
MIT
emizfliu@mit.edu

Sarah Zhao
MIT
sarahaz@mit.edu

Abstract

The human perceptual system integrates visual and audio input to learn a mental representation of real-world objects. When a human learns about a scene or object, they draw information from what they hear as well as what they see. In this paper, we ask if this sensory integration can be simulated using neural networks. The primary objective of this investigation is to assess the advantages provided by using multimodal data in scene annotation tasks. Specifically, we will be using both vision and sound to determine the materials (leather, wood, etc.) present in an image and comparing performance of the multimodal training setup to that of training purely on sound or image data.

Although it has been previously observed that incorporating audio information into a vision-based statistical summary task improves performance, the methods involved are complicated and difficult to tune. In this paper, we hope to propose a relatively simple training pipeline that can integrate visual and audio information. We explore two methods of aggregation, both of which involve a vision network and a sound network, one of which trains a network to predict classes from the concatenated image-sound embeddings and the other which predicts the degree of association between the image and sound embedding.

We found that the model using a fully connected layer converged fastest during training, however both models reach similar accuracy, demonstrating the ability to classify materials and a potential for image-based sound generation.

1. Introduction

In our daily life, we often come to associate certain sounds with our environment or objects, such as the crash of waves on a beach and the cry of seagulls overhead. These sounds could be directly indicated by physical interactions within the scene (visually indicated sounds [3], or implied off-image, such as the hum of electricity inside a building. We propose that combining these sounds with visual input will give us more information regarding a scene, allowing for higher accuracy or more efficient training on scene annotation tasks. The models we propose in this paper are based on standard visual CNN networks, and the goal is for the proposed methods to be easily implementable and replicable. If successful, the methods from our work can be used to aid in audio prediction from images to create a complete audio-visual scene.

In this paper, we propose two models of sound-vision aggregation. The first aggregation method concatenates the embeddings produced by the two networks to pass through a fully connected layer followed by a softmax layer, and the second aggregation method takes the cosine distance between the vision and sound embeddings, turning the problem into a binary classification task of whether the image and sound are associated or not.

2. Related Work

Several previous works have attempted to tackle the issue of multimodal audio-visual integration.

In [3], it was found that sound waveforms could be synthesized from silent visual input, by training on color and motion input. The model has two main tasks: predicting the type of sound (if any) is associated with the video, as well as detecting alignment of the sound with visual input. To achieve this, the video frames was represented with a convolution network and the time series with a recurrent neural network, mapping resulting sound features to waveforms using parametric or example-based synthesis as needed. The model was then trained on the Amazon Mechanical Turk dataset. It was found in the resulting psychophysical study that the rate participants mistook the model's result for the ground-truth sound outperformed state-of-the-art image-matching models, as well as a baseline random sound generator.

In [2], it was found that when an image with associated audio was trained to predict held-out sound from video frames, the learned features could be used to perform object and scene recognition tasks. To achieve this, a statistical summary of the sound data was first computed to match the time scale of the video. Then, the sounds were predicted from images. Although each video is associated

Figure 1. Example of image plus spectrogram pairing from the dataset. The spectrograms on the right are generated from the sound produced by the interaction of the drumstick with the table or the chair.

with multiple frames, sounds were predicted from a single frame, since the goal is to translate to object/scene recognition tasks on a still image. It was found that the neuron visualizations of the learned representations in the network match state-of-the-art models and the sound-trained network representation contained fewer unimportant elements of the image (such as grass).

In both these works, a neural network was used to predict sound data from an image in some capacity, and it was demonstrated that this sound data could be used in further downstream annotation tasks. However, the models used in these papers are parametrically complex and difficult to tune, meaning that from a practicality standpoint it may be infeasible to implement them. In this project, we propose two relatively simple aggregation modes. The first learns image and sound embeddings and makes predictions after passing through a fully connected layer. The second makes use of a two-tower neural network model to predict the likelihood of sound-image pairings.

The two tower model was originally used for ML-based recommender systems, as in [5]. The architecture itself consists of two neural networks that feed forward in parallel, and a final operation that computes the normalized dot product of the two network embeddings. In recommender systems, given a user (network 1) and a query (network 2) such as a movie or game, the algorithm will output a number close to 1 if the query is a good recommendation for the user, and close to -1 (or 0, depending on the normalization technique) if it is a bad recommendation. Using this architecture enables us to reframe the sound-image pairing likehihood problem as "recommending" a set of sounds for a given image. This model likewise can also be used for sound generation, by using nearest neighbors to find the top sound "recommendations" for an unseen image.

3. Methods

3.1. Dataset

In this investigation, we used the Greatest Hits dataset developed by Owens et. al. in [3]. The Greatest Hits dataset consists of 977 videos of drumsticks hitting various objects. Rather than capturing full scene information, each video is zoomed in close so that the material and fine-grained texture of the object is visible, to simulate the viewpoint of an observer interating with the system, such as a robot. 64% of the videos came from outdoor scenes and the remaining 36% came from indoor scenes. All sounds in the video consisted of interaction with 1 of 17 different materials (Table 1). A denoising algorithm was applied to the audio from all of the videos.

Since the image in the video is steady (save for the movement of the drumstick), we selected one frame of the video to use as visual input. We selected ten sound interactions from each video. For each selected sound interaction, we took a one-second long clip centered around the interaction which was then converted into a spectrogram, as demonstrated in Figure 1. The spectrogram is an image representation of sound intensity over time, meaning we can process the sound data using the same network architectures we use for the visual data. This is standard practice when working

Material	ID
Plastic	0
Drywall	1
Rock	2
Metal	3
Leaf	4
Grass	5
Paper	6
Water	7
Gravel	8
Glass	9
Tile	10
Ceramic	11
Plastic Bag	12
Dirt	13
Cloth	14
Wood	15
Carpet	16

Table 1. Materials and Material ID

with audio data in deep learning.

The Greatest Hits dataset itself was split into train and test data. To generate a validation set, we held out 20% of the training data. In total, out of the 977 videos, 586 were used as training data, 147 as validation data, and 244 as test data. As seen in Figure 2, the class distribution of Material IDs from the train, validation, and test datasets are approximately equivalent.

3.2. Models

We used the following neural network models as building blocks for our sound-image aggregation models (Described in greater detail in Section 3.4). In the rest of the paper, these networks will be generally referred to as convolutional networks, as they can be interchanged in our visual and sound models.

3.2.1 ResNet-18

The first architecture we used was ResNet [1]. The ResNet is a deep convolutional neural network that addresses the issue of vanishing gradients in deep networks through the introduction of skip connections that apply an identity transform to layer outputs. The ResNet won the 1st place on the ILSVRC 2015 classification task and is one of the most-often cited visual network architectures. In this paper, we try a ResNet-17 (17-layer deep ResNet) architecture in the image and sound recognition networks.

3.2.2 VGG

The second network architecture we used was VGG [4], short for Visual Geometry Group. The VGG architecture is a CNN based architecture that makes use of small kernels (3×3, that captures the minimum possible receptive field) followed by ReLU activations. This modification leads to higher performance; VGG was the top submission for the ILSVRC in 2014.

3.3. Baselines

In the following sections, we detail the different models that we trained in this investigation. All models were trained in PyTorch using a SGD optimizer with a learning rate of 0.01 for 30 epochs. Validation accuracy was computed at each epoch. Models were trained on Google Colab GPU.

3.3.1 Sound Network

To obtain a baseline for performance, we tried training a network based solely on sound. Because sound data is represented visually as a square spectrogram, we directly used the convolutional networks (ResNet, VGG) to learn the material IDs from the spectrogram as a multiclass classification task with 17 classes using the Pytorch cross entropy loss.

3.3.2 Visual Network

We similarly directly applied the ResNet and VGG convolutional networks to the purely visual classification task. Since the convolutional models take square images only, all images were uniformly resized to size 224×224. Since the original images were all of the same dimension before resizing, compression along the vertical axis is likewise uniform and ideally does not impact material classification. A random horizontal flip was applied with probability 0.5 for data augmentation purposes.

Because images potentially contain more than one material, we performed multilabel classification with 17 outputs, one for each material. We used binary cross entropy loss from Pytorch. For a given image, each label was assigned 1 or 0 depending on whether or not the corresponding material is present in the image and accuracy was computed by counting the percentage of correctly predicted outputs. Since each image contains a small number of distinct labelled materials, the classification task is sparse and the quantities of 1 and 0 labels are unbalanced. Specifically, there are around 9 times as many 0 labels as there are 1 labels. To address this, we compute precision, recall, and F1 score alongside accuracy. We additionally train a separate run in which we reweight the loss function to place more emphasis on predicting positive labels and discourage false

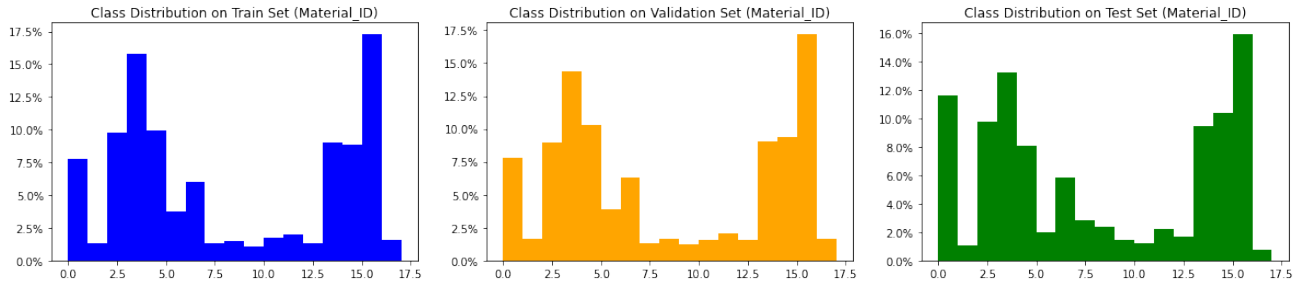

Figure 2. Distribution of Material IDs on train, validation, and test sets. Specific descriptions of each class can be found in Table 1.

negatives (using the pweight parameter in the torch binary cross entropy function).

3.4. Multimodal Aggregation

In this section, we cover the architectures of the two vision-sound aggregation models that we used.

3.4.1 MLP-based aggregation (FC_AGG)

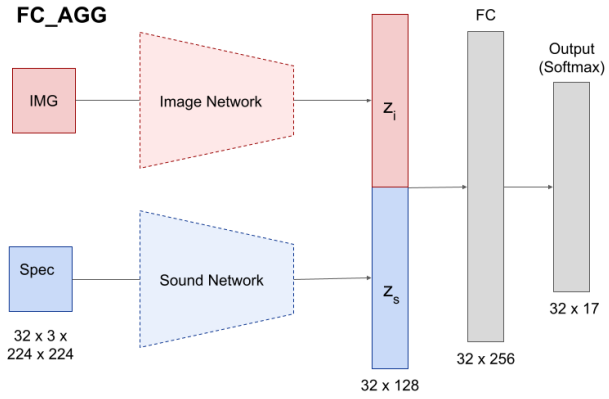

Figure 3. FC_AGG

The first aggregation model we considered is the FC_AGG model, which makes use of a fully connected layer. Given a pair of vision and sound input (one image and one spectrogram), we pass the inputs through the vision and sound networks respectively as detailed in Section 3.3. For the sake of simplicity and uniform backpropagation, we used the same convolutional network (ResNet or VGG) for both vision and sound networks. We configure both convolutional networks to output a 128-dimension embedding, which we concatenate and pass through a fully connected layer before predicting 1 out of 17 classes that corresponds to the material ID of the sound.

3.4.2 Two Tower Approach (TT_AGG)

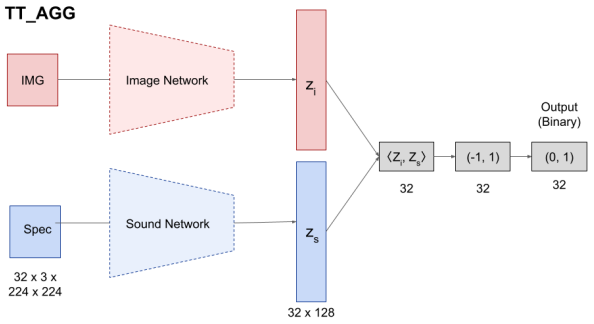

Figure 4. TT_AGG

The second aggregation model is the TT_AGG model, which is based on the two-tower architecture described in [5]. Like in the FC_AGG model, we pass the (vision, sound) input through convolutional networks to obtain 128-dimension embeddings. However, in addition to passing positive image-spectrogram pairings (where the spectrogram is directly associated with the image), we introduce negative pairings where the image is paired with a spectrogram whose associated material is not present in the image. This turns the prediction task into a contrastive classification task wherein we train our model to predict the likelihood that the image and sound are from the same video.

After embeddings are computed via convolutional network, they are normalized and a dot product is taken between them (range from [-1, 1]). The dot product is scaled to have a range of [0, 1] to match the binary classification task. The loss function used for this is the torch binary cross entropy loss for a single label. To evaluate the performance of this model, we examine the accuracy, recall, precision, and F1 score of the model predictions, but we also look at how many of the model's top sound predictions belong to classes of materials present in the image. Specifics for generating/predicting sound from image input are detailed in Section 3.5. Because we configure the number of positive and negative samples in our dataset to be equal, we do not need to reweight the loss function.

Model	Validation Accuracy	Testing Accuracy
Resnet	0.9449	0.5501
VGG	0.7978	0.5450

Table 2. Validation and Test Accuracy on Sound Network.

3.5. Sound Generation via Two-Tower Network

Because the two-tower model assigns a positive prediction to embedding pairs whose cosine distance is close to 1, it trains the image and sound networks to produce similar embeddings for instances of the same class. We can then apply this property to generate "recommendations" for the sound associated with a given image. First, we pass the image through the trained visual network on the TT_AGG model, and pass all sounds through the trained sound network. We then take the normalized dot product of the vision embedding (from the test set) with all of the sound embeddings (from the training set). The closer the dot product is to 1, the more likely a sound will be a good "fit" for an image. By sorting in descending order, we can thus pick the k "best" sounds for any given image, thereby creating a procedure for sound generation. Another way of thinking about this procedure is that given an image embedding, it is performing nearest neighbors on embeddings (of sounds from a previously existing sound library with cosine distance as the distance metric.

4. Results

4.1. Sound Network

Table 2 shows the validation and test set accuracy of the ResNet and VGG networks after being trained for the material classification task on spectrogram data only. ResNet achieves very high accuracy on the validation set, at around 95%, and VGG also achieves fairly high accuracy at 79.8%. However, both models overfit drastically as their test accuracies are both around 55% accuracy. This could likely be due to our small dataset size, as we only trained on ten sounds per video. Figure 5 shows the confusion matrices for class prediction on the test set for both models. It is observed that classes that are less prevalent are often misclassified as a similar, more prevalent class. For example, the sound of dirt (ID=13), which comprises less than 2% of samples in training, testing, and validation sets, is often misclassified as leaves (ID=4) which comprises approximately 10% of training, validation, and test sets. Leaves and dirt sound similar when hit due to both being relatively soft organic material, but the small percentage of true dirt samples means the model overfits to those that are in the training set and mistakes new examples as belonging to a more abundant class.

4.2. Visual Network

Table 3 shows the validation and testing results on the purely visual network. We note that both ResNet and VGG attain training accuracy of around 80% regardless of how the loss was weighted, although there is some mild overfitting as the test accuracy is around 75% for VGG and 76-77% for ResNet. Surprisingly, the unweighted models attain higher recall, which we would have expected to be the other way around as having a drastically uneven dataset may lead to the appearance of many false negatives since 0 is the majority label. Recall for all models is high (around 0.90-0.95) on the validation set, and reasonably high (around 0.70 for ResNet and around 0.75-0.76 for VGG) on the test set. However, it is worth noting that all models achieved relatively low precision, meaning that they predicted too many false positives. There is no clear trend on the increase or decrease in precision across weighted or unweighted loss functions or across convolutional architectures, indicating that it is inherently difficult to achieve high precision when training a sparse multilabel classification task due to the drastic class frequency imbalance.

4.3. MLP-based Aggregation (FC Layer)

Table 4 shows the validation and testing accuracy when we aggregate sound and image data during training using a fully connected layer. A marginal improvement (95% vs 94%) is observed in validation accuracy on ResNet and a larger improvement (90.7% vs 79.8%) is observed on VGG. This indicates that including image data helps with training. In fact, this observation is further supported by the training progress as seen in Figure 7. In the ResNet training, both sound and FC_AGG converge at the same approximate final value, but using both image and sound allows for faster convergence in earlier epochs. This phenomenon is more pronounced in the VGG training, where the image-sound combined model learns higher accuracy much earlier on than the pure sound model - notably, the VGG model could have been trained for a longer time until convergence, although we standardized all runs to 30 epochs due to resource constraints, but the observed trend would be the same (the only difference being higher end accuracy).

Model	Validation Accuracy	Testing Accuracy
Resnet	0.9526	0.5425
VGG	0.9066	0.5480

Table 4. Validation and Test Accuracy on FC_AGG Network.

When looking at the test set, we observe that FC_AGG overfits and achieves around the same test performance as pure sound data at around 54-55% accuracy. This indicates that the images used in the training set are perhaps not indicative of the images found in the test set, or it may indicate that the training set needs to be larger. Interestingly, a look

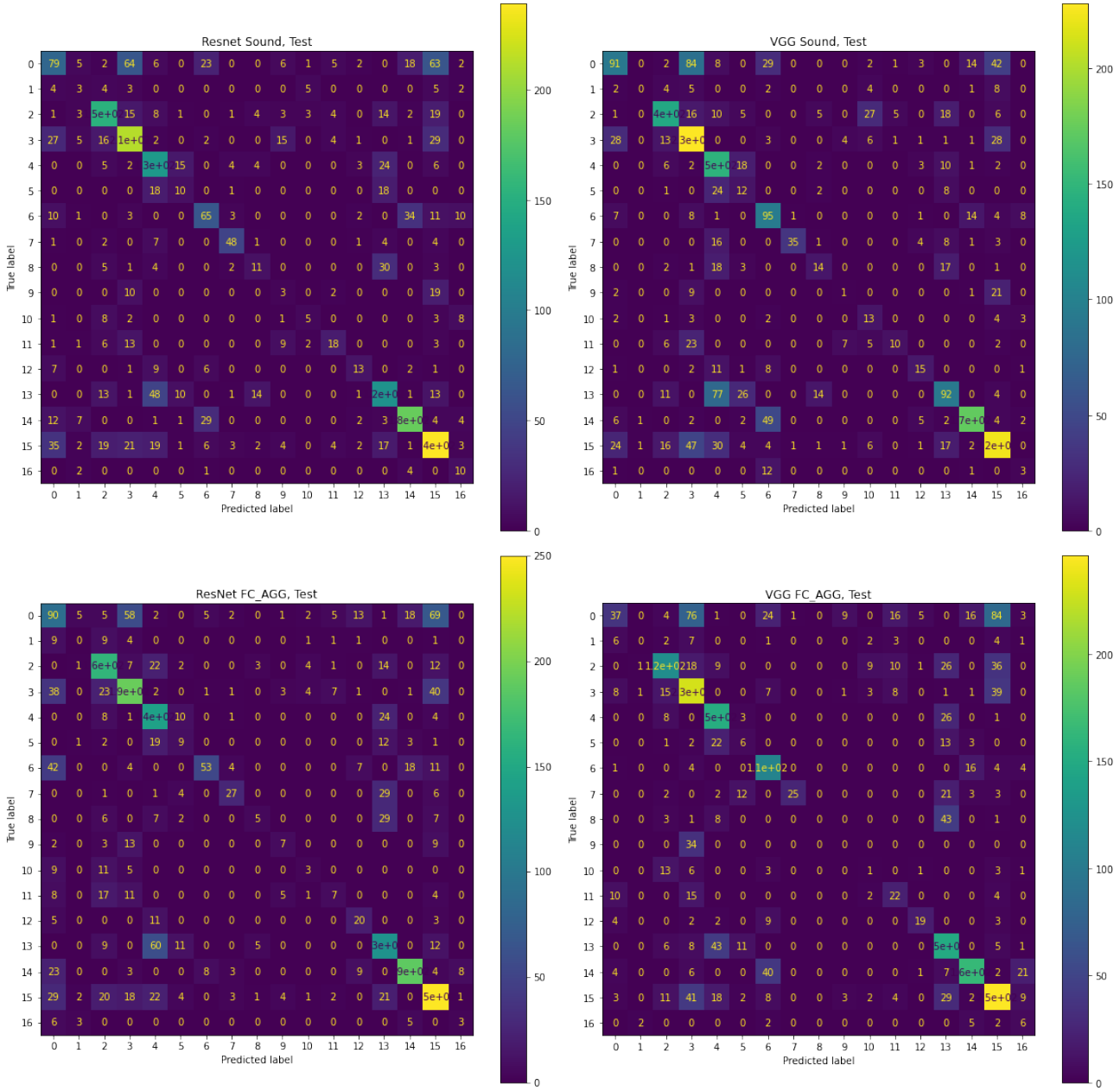

Figure 5. Confusion matrices for material ID classification, Sound only (top row) and FC_AGG (bottom row). Label IDs are as in Table 1.

		Validation Set				Test Set			
Network	Loss	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
ResNet	Unweighted	0.8148	0.4081	0.9665	0.5739	0.7623	0.2899	0.7107	0.4118
ResNet	Weighted	0.8148	0.4026	0.8996	0.5563	0.7728	0.3018	0.6954	0.4209
VGG	Unweighted	0.7918	0.3557	0.9518	0.5179	0.7580	0.2898	0.7437	0.4171
VGG	Weighted	0.7661	0.3265	0.9407	0.4847	0.7363	0.2707	0.7792	0.4018

Table 3. Accuracy, Precision, Recall, and F1 scores on the vision network, validation and test sets.

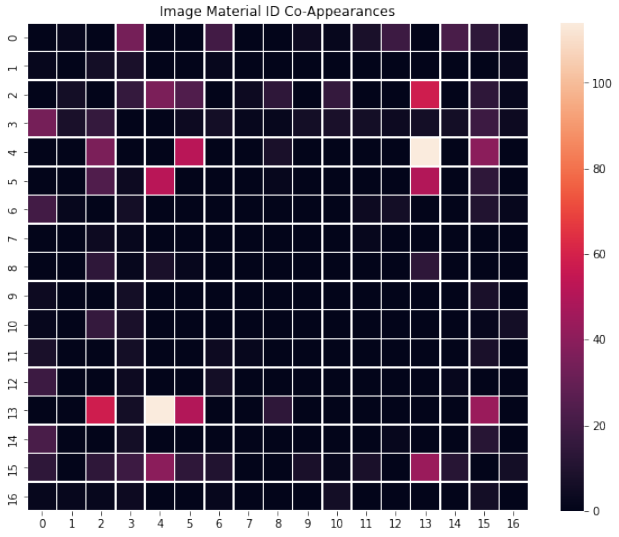

Figure 6. "Co-appearances" of material classes from the training set. A higher value means the two classes are more likely to be found together in the same image.

at the confusion matrix in Figure 5 indicates that although both sound-only and sound-image FC_AGG models tend to misclassify the same classes, the FC_AGG network is less likely to correctly classify the rare classes than the sound-only network. A look at Figure 6 tells us that the FC_AGG models make more mistakes between classes that are more likely to appear in the same image (as one image can contain multiple materials). The more "co-appearances" a pair of material classes have, the more likely they are to be confused for each other. This means that if a rare class and a common class appear in the same image, the rare class will likely be misclassified as the common class.

4.4. Two Tower Aggregation (Contrastive)

Table 5 shows the accuracy, precision, recall, and F1 scores of the two-tower aggregation model on the validation and test sets. The ResNet-based TT_AGG model attains similar validation accuracy to the image baselines described in Section 4.2 (around 80%) and slightly lower test accuracy (71%, as opposed to 73 - 77%). On the other hand, the VGG-based TT_AGG model achieves much lower accuracy, at around 64% for validation and 53% for test. This may indicate that the VGG architecture may be unsuitable for the TT_AGG task, or that it needs to be trained for longer, or that the hyperparameters need to be adjusted. Further investigation is required before a VGG-based two tower image-sound model is completely disregarded.

Both models attain higher precision on both the validation set and test set than the pure visual baselines (64-75% vs 32 - 40% on validation set, 53-69% vs 27-30% on train set). This means that the false positive rate is far lower with the TT_AGG model than with both the unweighted and weighted vision networks. Although on average the recall for the TT_AGG model is lower than that of the vision networks on the validation set (75-89% vs 89-97%), they are higher on the test set (80% vs 69-78%). This means that the false negative rate of the visual networks was artificially low, likely because the model is overenthusiastic in assigning positive cases, resulting in a higher false negative rate in the test set. Overall, the TT_AGG model attains a higher F1 score than the pure visual network model. In Figure 9, we can see that this is learned fairly early on in training. Although the formulation of the TT_AGG model may be a key contributor to these results (as we construct negative examples to appear as a 1 on 1 ratio to positive examples), this demonstrates that the contrastive training technique of the TT_AGG model allows us to circumvent the issues that come with training a sparse multilabel dataset while attaining similar accuracy in the case of the ResNet.

4.4.1 Nearest-Neighbor Sound Generation

Because the ResNet achieved higher validation and test accuracy, we will use the trained ResNet TT_AGG model for nearest-neighbor sound generation (recommendation).

We observed the K best predictions for sounds (from combined train, validation, and test sets) for a random image from the test set, where we let K go from 1 to 50. We judged a prediction to be "good" if the material ID of the recommended sound is also included in the image label, meaning that the sound could be plausibly generated from an interaction with the relevant material in the image. Figure 8 shows that the top 25 sound recommendations are all good sounds. The percentage decreases as K increases, which is to be expected, but the lowest the percentage gets is around 75%, meaning around 3 out of 4 recommendations for a sound are good. This demonstrates that a trained TT_AGG model can be feasibly used for sound generation.

5. Conclusion

In this paper, we investigated training on a vision-sound combined multimodal dataset. Our specific task was predicting labelled material IDs from videos as a way of simulating interaction with the environment, such as with a robot. To assess the performance of combined sound-image training, we computed baselines using purely spectrogram or visual datasets.

We performed multimodal training in two ways. In the first way, FC_AGG, we concatenated embeddings before passing through a fully connected layer to predict one of 17 classes. We compared this method to the sound baseline for multiclass classification. In the second method, TT_AGG, we computed the likelihood that a given image and spectrogram came from the same video by finding the similar-

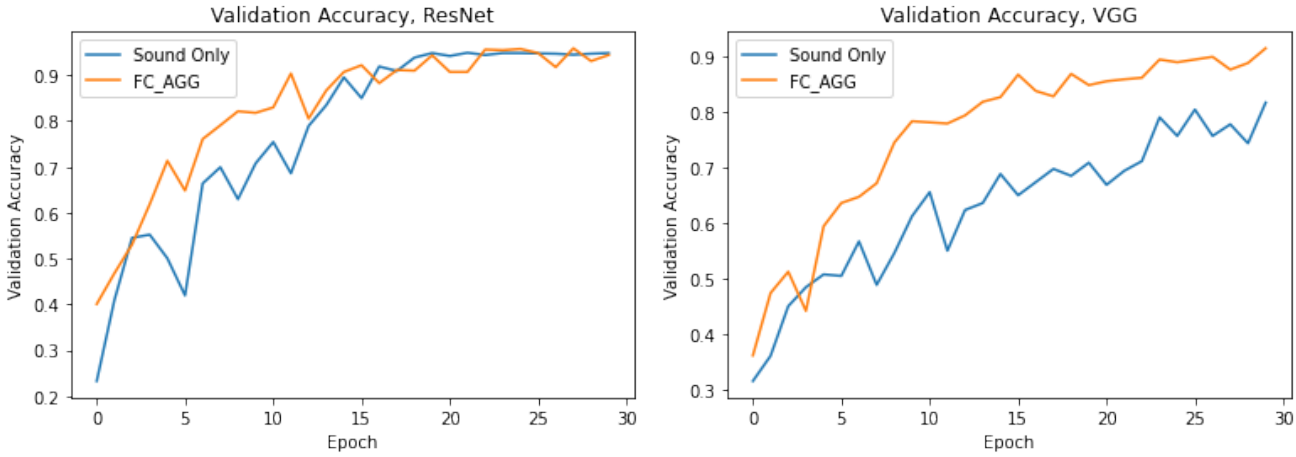

Figure 7. Validation accuracy over training, sound only vs FC_AGG

	Validation Set				Test Set			
Network	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
ResNet	0.8040	0.7563	0.8959	0.8202	0.7124	0.6897	0.8048	0.7428
VGG	0.6634	0.6424	0.7536	0.6936	0.546	0.5305	0.8019	0.6385

Table 5. Accuracy, Precision, Recall, and F1 scores on the TT_AGG network, validation and test sets.

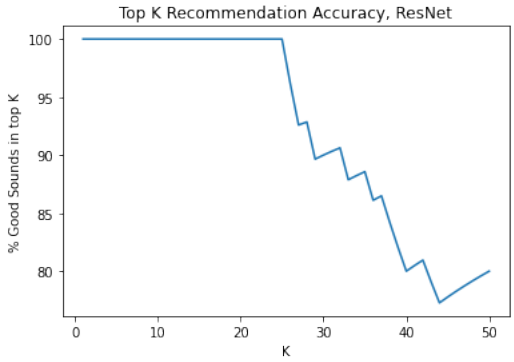

Figure 8. Percentage of "good" sounds in the top K predictions.

ity between their embeddings. This method was compared to the visual baseline for multilabel classification, and was also used for sound generation.

It was found that combining images and sounds in FC_AGG allowed for faster convergence during training when compared to training only on sound data. However, the validation accuracies were similar after training to convergence, and both models severely overfit to the data, resulting in much lower test accuracy. This can be mitigated through use of a larger training set and additional hyperparameter tuning to prevent overfitting (such as regularization, dropout, etc which were not explored in this paper due to time and resource constraints). The sound network and the FC_AGG model both misclassify data points that belong to less frequently occuring classes as similar but more frequently occurring classes. In the FC_AGG network, this can be explained by co-appearances, since the network is likely to misclassify when two or more materials are present in the same image, in which case the model will wrongly predict one of the other classes present in the image. To further investigate this phenomenon, it is worthwhile to create another dataset with less complex images where there is only one material or object per image, and repeat the same experiment comparing with a pure sound baseline. It would also be beneficial to investigate the sound to image ratio. In this investigation, we used 10 sounds per 1 image to capture the fact that images are more complex and contain more information than the spectrograms do. However, if we were to include simpler images in our dataset where only 1 class was present for each image, this may eliminate the need for multiple sounds per image.

Training on the pure visual model introduced the issue of sparse multilabel datasets, where the number of negative samples far outweighed the number of positive samples. To address this, we tried training separately using the original binary cross entropy loss function and using a weighted loss function to discourage predicting 0 to naively achieve base accuracy using the majority class. However, we observed the opposite result in both weighted and unweighted classes in which the model did a good job at avoiding false negatives but began predicting many false positives as well.

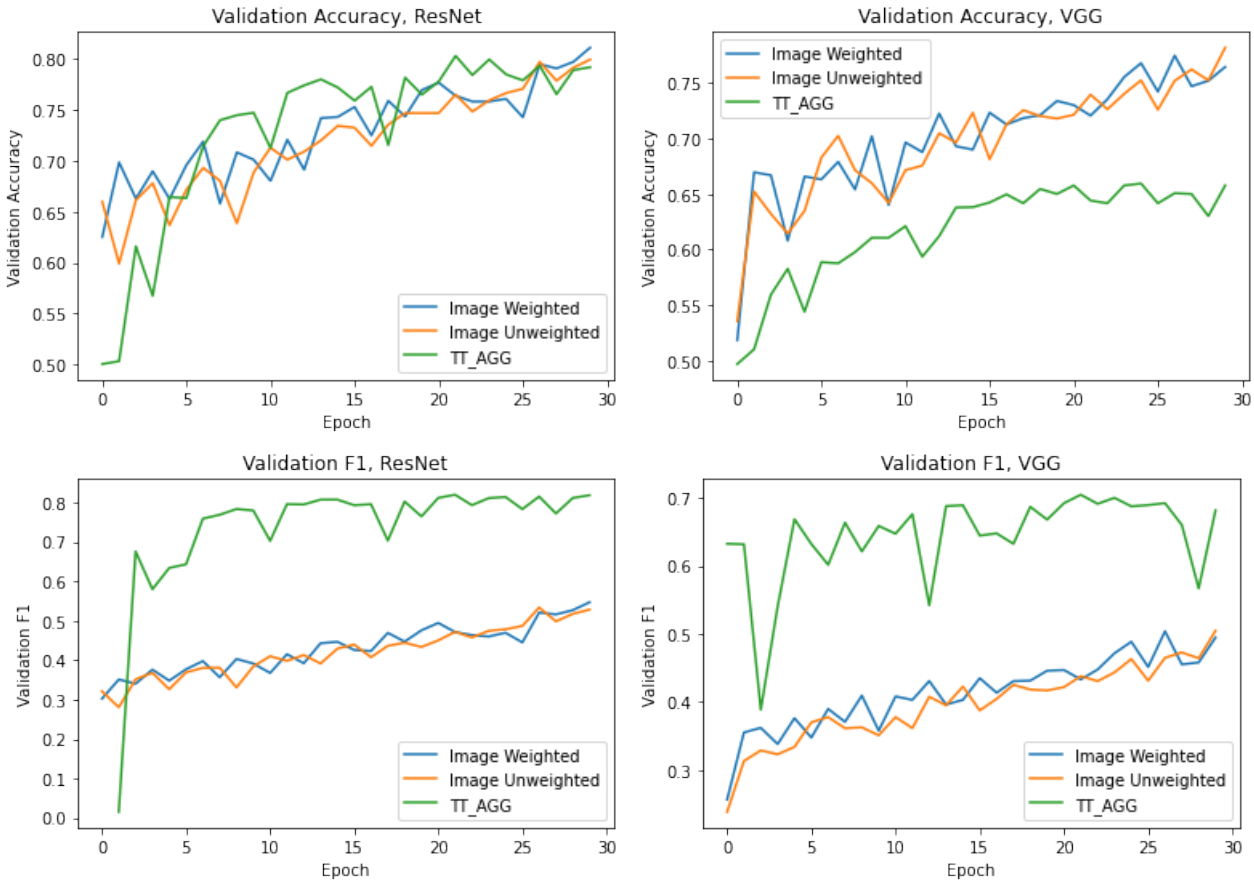

Figure 9. Validation accuracy and F1 score over training, TT_AGG vs vision networks trained on weighted and unweighted loss functions

Using the TT_AGG model formulation where we created a balanced dataset helped address this issue. Even though accuracy was constant between the ResNet TT_AGG model and the baselines, the F1 score was improved by switching to the TT_AGG model. The VGG TT_AGG model had far lower accuracy than the baseline, but the reason for this is unclear. It may have to do with needing to further tune training parameters, or the VGG architecture may have inadvertently caused some embeddings to converge meaninglessly due to having too many layers (the ResNet, with skip connections, would be able to sidestep this issue). However, at this point these are only hypotheses.

In this paper, we also demonstrated the feasibility of using a trained TT_AGG model for sound generation. Out of the 50 closest embeddings, around 80% were feasible sounds given the image, which was determined by matching material labels. However, this trend is not monotonically decreasing, which means that some tuning is needed to find the ideal K closest points to search for good sounds. We have demonstrated on a toy dataset that TT_AGG has the potential for sound generation, but to actually implement it in practice, further training (on larger and more complex datasets, for more epochs, and with a set of optimized hyperparameters) is needed to develop a useable model. However, in this paper we have demonstrated a good set of metrics that can be used to evaluate the performance of a two tower based sound generation scheme, and we have demonstrated that the idea shows promise. If done well, image-based sound generation can be used in many applications, from accessibility tools to dataset augmentation for multimodal datasets such as the one used in this paper.

References

[1] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. DOI: 10.48550/ARXIV.1512.03385. URL: https://arxiv.org/abs/1512.03385.

[2] Andrew Owens et al. "Ambient Sound Provides Supervision for Visual Learning". In: *CoRR* abs/1608.07017 (2016). arXiv: 1608.07017. URL: http://arxiv.org/abs/1608.07017.

[3] Andrew Owens et al. “Visually Indicated Sounds”. In: *CoRR* abs/1512.08512 (2015). arXiv: 1512.08512. URL: http://arxiv.org/abs/1512.08512.

[4] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2014. DOI: 10.48550/ARXIV.1409.1556. URL: https://arxiv.org/abs/1409.1556.

[5] Xinyang Yi et al., eds. *Sampling-Bias-Corrected Neural Modeling for Large Corpus Item Recommendations*. 2019.