Multimodal Material Classification and Sound Prediction from Visual Scenes

Catherine Mei
MIT
meic1212@mit.edu

Linette Kunin
MIT
linette@mit.edu

Abstract

The application of predicting audio from visual information has become increasingly relevant in various fields, including accessibility technology, virtual reality, and robotics. In this paper, we investigate the use of visually indicated sounds by adding sound to silent video frames of drumsticks hitting different materials from the Greatest Hits dataset. Our first goal is to determine if combining both image and sound information improves our ability to predict the material being hit compared to using a single modality. To achieve this, we explore a range of neural network architectures and parameters. Our second objective is to create a model that takes in video frames of hits and returns a one-second audio clip corresponding to the hit using a state-of-the-art convolutional neural network. To prepare the data, we take singular frames at the exact moment of contact between the drumstick and the object and crop the one-second audio file centered around the hit. Our findings indicate that combining both image and sound information leads to more accurate predictions of the material in new images and sounds. We were also able to create 1-second audio clips that are louder at the time of the hit. However, our audio clips are noisy and do not change depending on the material being hit, which we attribute to limited computational power and the temporal dependencies inherent in spectrograms that our model does not consider.

1. Introduction

Visual information and auditory information are two crucial components of human perception that work in tandem to provide us with a comprehensive understanding of the world around us. While these two senses are often treated as separate modalities, recent research has shown that they are highly interconnected and can influence each other in fascinating ways [6]. Sound can be very important in determining materials especially when the texture of the material is hard to see. For example, by hitting a cup that looks like plastic or glass, the sound can help you to distinguish between the two materials. Thus, it is likely that combining information from two modalities (vision and sound) provides more information than either of the two modalities separately, and could be very useful in scene annotation tasks. Additionally, it can also be used to detect anomalies in manufacturing processes based on the sounds produced.

Another area where the interaction between vision and sound has been studied extensively is the use of visually indicated sounds, where visual cues are used to inform the creation or perception of auditory signals. The ability to add sounds to silent videos is an exciting development that has numerous practical applications. For example, it can enhance accessibility for people with hearing impairments and improve sound localization in noisy environments. In robotics, the ability to predict sounds made by objects can aid in improving object recognition and understanding the physical properties of objects. The predicted sounds can also help in identifying the action performed by a robot on an object and provide feedback on the action's effectiveness. In virtual and augmented reality, the ability to predict sounds can enhance the user's immersive experience by adding realistic sounds to the visual scene. Overall, the ability to predict visually indicated sounds can offer a more comprehensive understanding of physical interactions within a visual scene and their practical applications in various domains.

In this paper, we first use video still frames and the one second audio clip of drumsticks hitting different materials from the Greatest Hits dataset to classify the material being hit. Second, we create a model using a state-of-the-art convolutional neural network that takes in video frames of hits and returns the predicted one second audio clip corresponding to the hit. These two objectives are intrinsically linked since creating the one-second audio clip necessitates that both the visual and auditory cues convey information about the material being hit.

2. Related Work

The relationship between visual and auditory information is deeply rooted in psychological research, and we can gain valuable insights into model design principles by examining human cognition, development, and learning. In

particular, Smith and Gasser provided a theoretical framework for embodied cognition, which suggests that cognitive abilities are closely tied to the development of the body and the sensory-motor system [6]. The body plays a crucial role in shaping cognitive development by providing sensory feedback that helps to ground abstract concepts in concrete experience. Therefore, integrating multiple sensory modalities like sound and vision can enable effective learning and perception [6].

Numerous attempts have been made to predict sounds from images and videos, including the PixelPlayer system designed by Zhao et al., which generates sounds from images by mapping each pixel to a unique sound waveform [7]. The authors demonstrated that PixelPlayer achieved higher accuracy on tasks like image classification, object detection, and semantic segmentation, indicating that the sound predictions provided additional cues for visual recognition tasks. Furthermore, the predicted sounds captured some of the visual characteristics of the images, such as texture and shape [7]. Zhao et al.'s work has significant implications for areas such as accessibility and multimedia content creation, where the inclusion of sound information can enhance the experience for users with visual impairments.

With the same objective in mind, Owens et al. proposed a method to predict sound from silent videos. The authors used a convolutional neural network (CNN) to extract visual features and a recurrent neural network (RNN) to model temporal dependencies in the sound signal. They trained these models on a dataset containing videos of humans using a drumstick to interact with the environment [4]. This approach significantly outperformed state-of-the-art image matching methods. Finally, the authors showed that while information about the material being hit by the drumstick was helpful, it was not necessary to generate representative sounds for the silent videos [4]. The authors proposed a method for using ambient sound as a form of supervision for visual learning, where the model learns to associate sound with visual scenes without explicit sound labels [5]. They demonstrated that the model's performance on recognition tasks is comparable to that of state-of-the-art unsupervised learning methods [5].

3. Methods

This paper aims to address two primary research questions. Firstly, we investigated the effectiveness of combining visual and auditory modalities to enhance material labeling. Our hypothesis is that incorporating both visual and audio inputs would provide our model with more comprehensive information and consequently increase the model's accuracy. To test this hypothesis, we designed a model that integrates both image and spectrogram inputs, and outputs the material label for the object struck by the drumstick. We then compared the performance of this multimodal model with that of a model that exclusively employs sound or image inputs to label materials.

Secondly, we examined whether video frames could be labeled with sound without relying on recurrent neural net architectures. To achieve this, we developed a model that utilizes a convolutional neural network (CNN). In this model, video frames are used as inputs, while spectrograms serve as labels for the video frames. We compared the model's generated outputs to the actual sounds produced by a drumstick hitting various materials, with the aim of assessing the qualities of the hit (e.g. frequencies or amplitude) captured by the CNN.

3.1. Data Source

We used the Greatest Hits dataset, which features videos of humans striking or scratching various objects made from diverse materials using a drumstick. We selected this dataset because the use of a drumstick ensures consistency in the type of sound produced when an object or material is hit or scratched. Additionally, it has the advantage of not obstructing much of the scene, making the deformation visible when an object is struck, thereby providing essential information about the material type and the sound of the impact. The dataset also includes information on the time of each hit, the material struck by the drumstick, and the type of contact made (hit or scratch). We considered these labels as valuable information that could aid in determining the sound to be added to the video.

3.2. Data Pre-processing

Due to limitations in computational power, we recognized that it would be impossible to use all frames in a video as inputs to our model that adds sound. Therefore, we reviewed the videos of the drumstick striking objects, as well as the files that detailed the exact moments of each hit, extracting single video frames at the precise contact instance between the drumstick and the object. Afterward, we cropped the one-second audio file centered on the hit, with the consideration that this duration was sufficient to capture any material reverberations, without clustering too many hits into a single audio file. This audio file was then converted into a spectrogram to explicitly encode frequency information, which would aid in creating a more realistic audio output. Thus, for each video in the dataset, we saved an image of all the hits, an audio clip of the hit sound, the material being hit, and the spectrogram of the hit sound.

3.3. Multimodal Material Labeling

In this section, we will discuss how we combined visual and auditory modalities to enhance material label predictions as well as the implementations of the baseline sound-only and image-only networks.

3.3.1. Model

In this study, we employed ResNet [3] for our model. ResNet is a deep convolutional neural network that features skip connections, which effectively address the problem of vanishing gradients. Specifically, we utilized the pre-trained ResNet18 model as the baseline for our training conditions. This baseline was applied to two separate scenarios, namely, training the model to output material labels using either image inputs or sound inputs exclusively. Additionally, the same ResNet18 model was incorporated into our multimodal approach, which combined image and sound data to enhance the model's performance at predicting material labels.

3.3.2. Baseline: Single Modality Networks

We first attempted to classify materials based on either only sound or image properties. For both our sound-only and image-only networks, we employed a ResNet18 architecture. For the sound-only model, the model inputs were 2D arrays of size 1025 x 44 representing the spectrogram of the 1 second audio clip of the drumstick hit, while the output was the material label. For the image-only model, the model inputs were 3D array of dimensions 3 x 256 x 454, representing a static image of the drumstick and the struck material at the precise moment of impact, and the output was the material label. During training, both models used the Adam optimizer and Cross Entropy loss, with a batch size of 32, learning rate of 0.001, and 25 epochs of training. The dataset used in this experiment comprised of 733 training videos and 244 test videos, each of which had 15-25 instances of material hits by the drumstick. However, due to computational limitations, only 1/20th of the dataset was sampled for training.

3.3.3. Multimodal Model

We combined the image and sound inputs into a multimodal model to see if this better predicted the material being hit. Given a pair of video frame and audio spectrogram input, we passed the inputs through the image-only and sound-only networks detailed above. Both models returned a 128-dimensional embeddings, which we then concatenated and passed through a fully connected layer that predicted one of the 17 classes of materials from the combined embedding. Figure 1 is a diagram summarizing our model architecture.

3.3.4. Performance Metrics

The sound-only, image-only, and multimodal models were evaluated based on their respective validation and testing accuracy. Accuracy was determined by comparing the predicted label with the actual material label. Furthermore, we plotted the validation accuracy over epochs for each of the three methods in Figure 2. For our multimodal model, we constructed a confusion matrix to assess the distribution of accurate and inaccurate classifications in Figure 3.

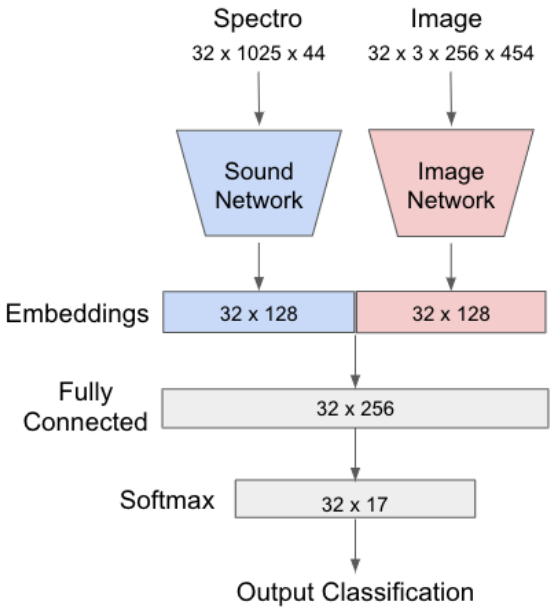

Figure 1. Multimodal Model architecture. The network combines the output embeddings of the image-only and sound-only networks to classify material labels.

3.4. Sound Prediction Using CNNs

In this section, we present our investigation into adding audio to video frames using a convolutional neural network (CNN) architecture, instead of relying on recurrent neural network (RNN) approaches. The current state-of-the-art methods for audio processing and prediction employ RNNs with long short-term memory units (LSTM) [4] or generative adversarial networks (GAN) [1]. However, we aim to examine the feasibility of utilizing a CNN architecture, such as ResNet, to capture details sound signals, and determine the type of information the model could identify.

3.4.1. Model

For our model, we used the ResNet18 architecture with a batch size of 32, learning rate of 0.001, and 25 epochs of training. We also utilized the Adam optimizer and cross entropy loss to optimize the model's performance. The input to our network was a tensor of size 256 x 454 x 3 representing images of the drumstick hit. Furthermore, the output label of the network was a spectrogram of size 1025 by 44. We trained out model to predict the output spectrogram from the input image tensor.

3.4.2. Approach & Qualitative Evaluation

Our dataset contained 733 training videos, each with between 15-25 instances of material hits by a drumstick. Due to computational limitations, we only used 1/20th of the dataset for our study. During training, we flattened the 1025 x 44 spectrogram array so that our model predicted 45100 outputs from the input image. For testing, we reshaped the predicted spectrogram array created by our model back into dimensions of 1025 x 44. Then, to assess the qualitative differences between the expected and predicted sounds, we converted the reshaped spectrogram output into a sound file and played it. We also visually compared the predicted output's spectrogram with the spectrogram of the expected audio output. By performing these analyses, we were able to evaluate the accuracy of our model and gain insights into the effectiveness of our approach.

4. Results & Discussion

4.1. Multimodal Material Labeling

We will begin by discussing our findings for our first research question. In this section, we examined whether the combination of information from multiple modalities could improve the classification of material labels.

4.1.1. Comparing Baseline with Multimodal Model

Table 1 shows the validation accuracy and testing accuracy for the sound-only, image-only, and multimodal networks.

	Validation	Test
Sound-only	0.9986	0.4440
Image-only	0.7669	0.5201
Multimodal	0.9535	0.5412

Table 1. Validation accuracy at the last epoch and testing accuracy for sound, image, and multimodal models

Our results show that the validation accuracy for both the sound-only and multimodal models is more than 90%. The validation accuracy for the image-only network is slightly lower. Additionally, the test accuracy is slightly higher for the multimodal model compared to the image-only model and significantly better than the sound-only model. This could imply that the multimodal approach helps to improve model accuracy.

4.1.2. Interpretations & Limitations

We saw that the validation accuracy increased across epochs for all models, as shown in Figure 2. We believe that the sound model's validation accuracy was larger than that of image or multimodal because the spectrogram was smaller and thus could have trained faster. Additionally, all of the models seem to overfit as test accuracy is much worse than validation accuracy.

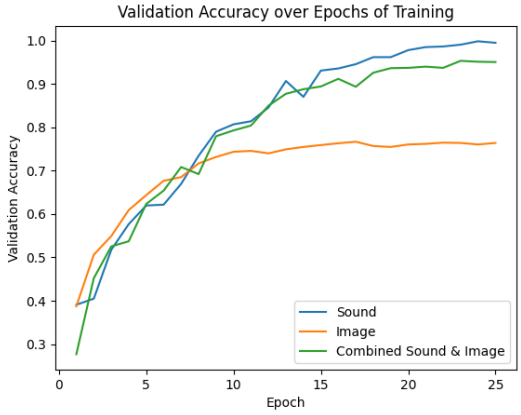

Figure 2. Validation accuracies across epochs of training for image, sound, and multimodal models.

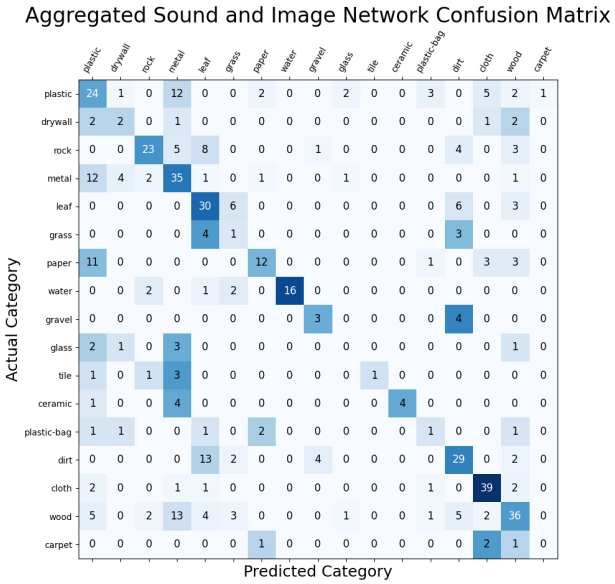

Figure 3. Confusion Matrix of material label classifications by multimodal model.

We saw that the test accuracy was better for the multimodal model than for the model with image or sound alone. This aligns with our intuition that adding more information improves model performance. We also notice that when materials are classified wrong, they tend to be materials that were also likely to be present in the image or sound similar when hit. We can see from Figure 3, that grass and dirt are often incorrectly classified as leaves (both often found outside) and glass, tile, and ceramic are often classified as metal (often all found in kitchen sinks).

However, our best test accuracy, although better than chance, was less than 55 percent. We believe this is due

to the lack of computational resources necessary to run a more significant number of training examples, and believe that our accuracy would significantly improve given more training examples. Additionally, the number of examples in each class was not evenly distributed. Although this distribution was fairly similar across training, validation, and test, looking at our confusion matrix in Figure 3, we notice that classes for which we have less examples generally are classified worse (e.g. glass), and that they are generally classified as materials for which there are more examples (e.g. metal).

4.2. Sound Prediction Using CNNs

This section presents the findings of our investigation into our second research question. Specifically, we address the feasibility of incorporating audio into video frames without relying on a recurrent neural network (RNN) approach. Additionally, we examine which characteristics of audio is captured by the CNN approach.

4.2.1. Performance & Evaluation

To evaluate the performance of our model, we initially used mean squared error (MSE) as a metric. Specifically, we computed the MSE between the predicted spectrogram and the actual spectrogram. However, due to the limited amount of training data, the resulting MSE was very large. Additionally, although the MSE decreased over epochs of training, analyzing just the MSE failed to provide insights into the underlying structure of the audio spectrogram. Therefore, we opted to conduct a qualitative analysis of our results instead. Specifically, we generated audio wave files corresponding to the spectrogram produced by our model and compared it with the expected audio clip. This allowed us to gain insights into the performance of our model.

Upon conducting a qualitative analysis of the audio clips generated from our model, we observed that all of the audio clips contained a distinct strike sound in the middle of the clip. While the model successfully captured the presence of a hit, almost all of the generated clips sounded the same and lacked differentiation in terms of the qualities of the material being struck. Furthermore, the audio contained some static noise, which was expected due to the limited amount of training data and the model's inability to fully capture the intricacies of the audio clip. Consequently, we concluded that the CNN can effectively capture the existence of a hit and the amplitude or loudness of the hit. However, the model was unable to capture any additional qualities of the material being struck.

Finally, we also assessed the images of the predicted spectrogram and actual spectrogram corresponding to the audio clip. Consistent with our previous findings, we observed a clear strike in the middle of both spectrograms. However, we noted that the range of frequencies associated with the strike was smaller in the predicted spectrogram compared to the actual spectrogram. This supports our earlier conclusion that the CNN is capable of capturing the presence of a hit but is unable to capture certain qualities of the audio, such as the frequency range associated with the strike.

4.2.2. Interpretations & Limitations

Our study has two primary limitations. First, we were limited in computational power, which may have hindered our ability to generate higher-quality audio. It is possible that with greater computational resources, we could have obtained more accurate and detailed audio representations. Second, audio data inherently exhibits temporal dependencies, which has led previous research to use approaches with better temporal resolution, such as LSTMs and GANs [2]. Despite this limitation, our study demonstrates that CNNs can still effectively capture the presence of hits and their amplitude. As such, future research may aim to improve CNNs to resolve temporal dependency issues, potentially leading to CNNs that perform as well as LSTMs and GANs on audio prediction tasks.

5. Conclusion

In this paper, we first investigated whether combining data from two modalities (vision and sound) improves material classification of materials being hit over vision and sound separately. We found the multimodal model had a better accuracy than either of the single-modal models. Secondly, we investigated whether we can use a CNN like ResNet to predict a one second audio clip of the sound the material would make when hit by a drumstick, given a single frame of a drumstick hitting the object. We found that although the model successfully captured the presence of a hit, specifically the amplitude or loudness (as shown by the distinct strike sound exactly in the middle of each clip), the sound generated is not notably different depending on the material that is being hit, and also contains a lot of static noise. We think this is due to the simplicity of our model, which does not capture temporal dependencies. Additionally, both tasks (material classification and sound prediction) would benefit from additional training examples which unfortunately, we were unable to include due to limitations in computational power.

In the future, we could improve the material classification task by artificially balancing the dataset, so it classifies materials for which there are less training examples better. For the sound prediction task, we could explore different models like LSTMs and GANs that capture more temporal dependencies, and see if these have better sound prediction.

6. Author Contributions

For this project, most of the work was done synchronously together. Therefore, the author contributions are very similar. We've provided a comprehensive list below.

Catherine Mei: Scripts to perform data processing, which included extracting video frames, audio clips, and spectrograms of the hits. Model development for the sound-only, image-only, and multimodal network. Model development for CNN sound prediction model. Paper writeup sections Related Works, Methods and Results for CNN sound prediction, Conclusion, and helped proofread paper.

Linette Kunin: Scripts to perform data processing, which included extracting video frames, audio clips, and spectrograms of the hits. Model development for the sound-only, image-only, and multimodal network. Model development for CNN sound prediction model. Paper writeup sections: Introduction, Methods and Results for material label prediction, Conclusion, and helped proofread paper.

7. Supplemental Materials

The code and data for our paper can be found below.

1. Greatest Hits Dataset
2. Processed Dataset (video frames, 1 second audio clips, and spectrograms)
3. Data Pre-processing script
4. Code for Material Labeling
5. Code for Sound Prediction CNN
6. Example outputs of Sound Prediction CNN

References

[1] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. In *International Conference on Learning Representations*, 2019. 3

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 5

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 3

[4] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. Visually indicated sounds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 3

[5] Andrew Owens, Jiajun Wu, Josh McDermott, William Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. volume 9905, pages 801–816, 10 2016. 2

[6] Linda Smith and Michael Gasser. The development of embodied cognition: six lessons from babies. *Artificial life*, 11:13–29, 2005. 1, 2

[7] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2