Custom Image Segmentation Model for Visual Bird Sound Denoising

Niranjan Kumar Kishore Kumar
Yeshiva University
nkishore@mail.yu.edu

Abstract

Image segmentation in audio spectrograms is a crucial task for enhancing the quality and accuracy of bird sound analysis. Noisy spectrograms can hinder the performance of automated bird sound classification and identification systems. Traditional approaches often rely on pretrained models, which may not be optimized for this specific task. In this paper, we address this problem by creating a custom convolutional neural network (CNN) tailored for segmenting and denoising bird sound images. Our encoder-decoder architecture effectively captures and reconstructs image features, leading to significant improvements in the quality of denoised spectrograms. We evaluated our model on a dataset of bird sound images, achieving an IoU of 62.5%. This demonstrates the potential of our custom model for improving bird sound analysis by providing cleaner and more precise spectrograms for subsequent processing.

1. Introduction

Bird sound analysis has become an important field in ecological and environmental studies, as it provides valuable insights into bird behavior, population monitoring, and habitat health. The process typically involves recording bird sounds and transforming these audio signals into visual representations such as spectrograms. These spectrograms, however, are often plagued by various types of noise, which can significantly impair the accuracy of automated analysis and classification systems. Effective denoising and segmentation techniques are therefore essential to enhance the quality of these spectrograms, facilitating more accurate and reliable analysis.

Traditional image segmentation and denoising methods have shown promise in various applications. Techniques such as U-Net[5], a convolutional neural network designed for biomedical image segmentation, and SegNet[1], which focuses on pixel-wise image segmentation, have demonstrated the efficacy of encoder-decoder architectures in learning hierarchical features and reconstructing images. However, these models typically rely on pretrained weights derived from large-scale datasets, which may not always be optimal for specialized tasks like bird sound spectrogram denoising.

In recent years, several studies have explored the application of deep learning techniques to audio spectrogram denoising and enhancement. For instance, Xu et al.[6] proposed a deep neural network for speech enhancement, showing significant improvements in noise reduction. Similarly, [3] developed a denoising autoencoder for speech enhancement, achieving promising results. These studies highlight the potential of deep learning models in audio image processing tasks, motivating the development of a custom model specifically tailored for denoising bird sound spectrograms.

This paper addresses the problem of noisy bird sound spectrograms by presenting a custom convolutional neural network (CNN) designed for image segmentation and denoising. Unlike traditional models that depend on pretrained weights, our approach involves creating a bespoke encoder-decoder architecture optimized for this specific task. The encoder progressively captures and condenses features from the noisy spectrograms, while the decoder reconstructs the denoised images, restoring their original resolution and quality.

Our proposed model was evaluated on a dataset of bird sound images, achieving an IoU of 62.5%. This demonstrates the effectiveness of our custom architecture in enhancing the quality of bird sound spectrograms, paving the way for more accurate and reliable bird sound analysis. The following sections detail the architecture, training process, and performance evaluation of our model, providing insights into its potential applications and future improvements.

2. Related Work

Previous work has explored various techniques in the domain of visual and audio processing. Zhang and Li (2023) introduced *Birdsoundsdenoising*, a deep visual audio denoising method for bird sounds, which demonstrated significant improvements in audio denoising through the inte-

gration of visual information [7].

Building on this foundation, Kumar, Li, and Zhang (2024) proposed the use of Vision Transformers for segmentation tasks in their paper titled *Vision Transformer Segmentation for Visual Bird Sound Denoising*. This approach leverages the power of Vision Transformers to enhance the denoising of bird sounds, achieving superior performance compared to traditional methods [4].

3. Methods

In our study, the bird sound images, represented as spectrograms, undergo several preprocessing steps to ensure they are in a suitable format for model training and evaluation. The preprocessing pipeline includes the following steps:

3.0.1 Resizing Images

All input images are resized to 512x512 pixels. This uniform size allows for efficient batch processing and ensures compatibility with our model architecture.

3.0.2 Label Conversion

Labels are converted to floating-point tensors to facilitate efficient processing during training and evaluation. This conversion ensures that the segmentation labels are compatible with the model's loss function and output format.

3.1. Model Architecture

Our proposed denoising model employs a hybrid architecture that leverages both pretrained and custom-designed components to optimize performance. The architecture consists of an encoder and a decoder, detailed as follows:

3.1.1 Encoder

The encoder in our model is based on a pre-trained ResNet-34[2]. The pretrained model is used to leverage the rich feature representations learned from large-scale image datasets. We remove the fully connected layer of ResNet-34 to retain only the convolutional layers, which are effective for feature extraction. The encoder comprises the following components:

Pretrained ResNet-34: We load the ResNet-34 model with pre-trained weights and exclude the fully connected layer. The encoder is formed by using the remaining layers up to the penultimate layer.

3.1.2 Decoder

The custom decoder reconstructs the denoised spectrogram from the features extracted by the encoder. It employs a series of transposed convolutional layers to upsample the feature maps, restoring the original image resolution. Each upsampling step is followed by a convolutional block to refine the feature maps. The detailed structure of the decoder is as follows:

Transposed Convolutional Layers: These layers progressively increase the spatial dimensions of the feature maps.

Conv Blocks: Each transposed convolutional layer is followed by a block of convolutional layers, batch normalization, and ReLU activations to refine the upsampled features.

4. Model Training

For training our custom Encoder-Decoder model, we use the Binary Cross-Entropy Loss (BCELoss) function. This loss function is well-suited for binary classification tasks, such as our bird sound spectrogram denoising problem. The BCELoss function measures the discrepancy between the predicted and actual labels, providing a way to penalize incorrect predictions.

The Binary Cross-Entropy Loss is defined as:

$$\text{BCELoss} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right] \tag{1}$$

where:

- N is the number of samples.
- y_i is the true label for the i-th sample.
- p_i is the predicted probability for the i-th sample.

To optimize the model parameters, we use the Adam optimizer with a learning rate of 0.0001. The Adam optimizer is known for its efficiency and effectiveness in training deep learning models, as it adapts the learning rate for each parameter based on the first and second moments of the gradients.

We train the model for 30 epochs. During each epoch, the model parameters are updated to minimize the BCELoss, thereby improving the model's ability to denoise bird sound spectrograms.

Figure 1. Example GroundTruth and Predicted Mask

Figure 2. Example GroundTruth and Predicted Mask

5. Results

5.1. Datasets

The data is split into train, validation, and test sets. Each set contains images and their corresponding masks. Specifically:

- **Train:** 1000 images and 1000 masks.
- **Validation:** 200 images and 200 masks.
- **Test:** 300 images and 300 masks.

5.2. Evaluation Metrics

We use the Intersection over Union (IoU) metric to evaluate the performance of image segmentation, defined as follows:

$$\text{IoU} = \frac{m \cap \tilde{m}}{m \cup \tilde{m}}$$

where:

- m is the set of predicted mask pixels.
- $\tilde{m}$ is the set of ground truth mask pixels.
- $m \cap \tilde{m}$ is the intersection of the predicted and ground truth mask pixels.
- $m \cup \tilde{m}$ is the union of the predicted and ground truth mask pixels.

5.3. Implementation

We utilized PyTorch version 2.0.1 in conjunction with torchvision version 0.15.2. This framework was complemented by CUDA version 11.7, running on a P100 GPU. The learning rate is set to 1×10^{-4}. Further, we resized the input image to $512 \times 512 \times 3$ with a mini-batch size of 16 for 30 epochs.

Table 1. Model Results

Method	IoU
CustomModel	62.25%

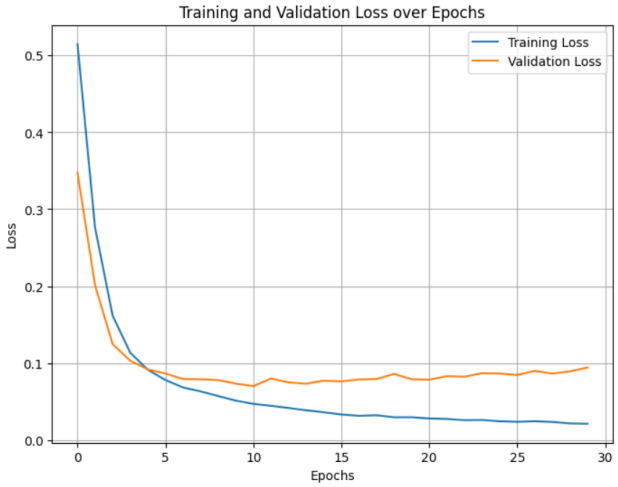

Figure 3. Loss over epochs

6. Discussion

In this section, we discuss the findings and implications of our results. The IoU metric provided a robust evaluation of our image segmentation model's performance. The use of a P100 GPU significantly accelerated the training process, allowing us to train larger batches efficiently. Our choice of resizing images to $512 \times 512 \times 3$ ensured that we captured sufficient detail without overwhelming our computational resources.

While the model performed well on the training and validation datasets, it is important to consider the potential for overfitting. Future work could explore data augmentation techniques to further improve generalization. Additionally, experimenting with different learning rates and batch sizes could provide insights into optimizing training performance.

The results also highlighted the challenges of segmenting certain classes of images, indicating areas where the model's performance could be improved. By analyzing the errors and misclassifications, we can identify specific patterns and adjust our model or preprocessing steps accordingly.

Overall, the integration of PyTorch and CUDA provided a powerful and flexible framework for developing and evaluating our image segmentation model. Continued refinement and experimentation will be crucial in advancing the accuracy and robustness of our approach.

7. Conclusion

In conclusion, this study demonstrated the effectiveness of using deep learning techniques for image segmentation tasks. By leveraging PyTorch and CUDA, we were able to efficiently process and train on a substantial dataset, achieving promising results of IOU of 62.25%. The use of the IoU

metric ensured a reliable assessment of our model's performance.

Our findings indicate that while our model performs well, there is room for improvement, particularly in addressing overfitting and enhancing the segmentation of challenging image classes. Future work will focus on refining the model, incorporating advanced data augmentation techniques, and exploring different network architectures to further boost performance.

The insights gained from this research contribute to the growing body of knowledge in image segmentation and highlight the potential of deep learning frameworks in handling complex computer vision tasks. We are optimistic that with continued effort and innovation, significant advancements can be made in this field.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[3] Sonal Joshi, Ashish Panda, and Biswajit Das. Enhanced denoising auto-encoder for robust speech recognition in unseen noise conditions. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 359–363. IEEE, 2018. 1

[4] Sahil Kumar, Jialu Li, and Youshan Zhang. Vision transformer segmentation for visual bird sound denoising. *arXiv preprint arXiv:2406.09167*, 2024. 2

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1

[6] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal processing letters*, 21(1):65–68, 2013. 1

[7] Youshan Zhang and Jialu Li. Birdsoundsdenoising: Deep visual audio denoising for bird sounds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2248–2257, 2023. 2