Environmental Sound Classification Based on Vision Transformers

MISCHA RAUCH ZIYI WANG
mtrauch | ziyiwa @kth.se

December 27, 2024

Abstract

Passive acoustic monitoring (PAM) enables a wide range of classifying environmental sounds, such allows conservationists to sample much greater spatial scales with higher temporal resolution and explore the relationship between restoration interventions and biodiversity in depth. The state of the art in audio classification is limited by the conventional use of Convolutional Neural Networks (CNNs), necessitating a reevaluation of transformer-based approaches to address potential performance gaps capturing long-range dependencies and patterns. Here we study the efficacy of Vision Transformers and Audio Spectrogram Transformers in environmental sound classification, comparing their performance against CNNs. Our key findings demonstrate promising results, indicating that AST achieve competitive performance comparable to CNNs in accurately classifying environmental sounds, showcasing their potential in auditory analysis and paving the way for novel audio processing methodologies. The significance of these findings lies in the potential to revolutionize environmental sound classification, offering efficient and versatile models that can be applied beyond bird species classification, ultimately contributing to advancements in diverse domains such as wildlife monitoring, urban soundscape analysis, and public health, with far-reaching implications for our understanding of the acoustic world.

Contents

1 Introduction **3**
1.1 Background . . . 3
1.2 Problem discussion . . . 4

2 Research Method **5**
2.1 Choice of Research Method . . . 5
2.2 Application of Research Method . . . 5

3 Method(s) **6**
3.1 Ethical Considerations . . . 6
3.2 Data Preparation . . . 6
3.3 Training Setting . . . 7
3.3.1 ViT Based Method . . . 7
3.3.2 CNN Based Method . . . 7
3.3.3 AST Based Method . . . 7
3.4 Metrics . . . 8

4 Results and Analysis **8**
4.1 CNN and ViT Results . . . 8
4.2 AST Results . . . 10
4.3 Weltch's t-test Results . . . 11
4.3.1 CNN vs. ViT Model Comparison . . . 11
4.3.2 CNN vs. AST Model Comparison . . . 11
4.3.3 Implications of the T-test Results . . . 11

5 Discussion **12**
5.1 Vision Transformer and CNN . . . 12
5.2 AST Model's Superior Accuracy . . . 12

A Insensible Approximation **13**
A.1 CNN validation and training losses after 100 epochs . . . 13
A.2 Welch's t-test on the classification results of the CNN and ViT dataset . . . 13
A.3 Welch's t-test on the classification results of the CNN and AST dataset . . . 14

1 Introduction

Environmental sounds ranging from urban noise to natural phenomena like bird calls present a rich tapestry of audio signals that are inherently different from human speech or music. These sounds typically lack a stable temporal structure and are often polyphonic, posing unique challenges in classification[1]. The task becomes even more complex when considering the identification of specific species in bioacoustic studies, such as bird species recognition, where fine-grained audio analysis is crucial[2].

Recent research has shifted focus towards transformer models, like Vision Transformers (ViT) and Audio Spectrogram Transformers (AST), which have demonstrated promising results in image and audio processing tasks[3, 4]. These models leverage the power of self-attention mechanisms to capture long-range dependencies in data, a feature particularly useful in dissecting the complex nature of environmental sounds.

This paper aims to evaluate the efficacy of ViT and AST in environmental sound classification (ESC), particularly focusing on their suitability for general environmental sound scenarios. A comparative analysis with the widely-used Convolutional Neural Networks (CNN) will provide insights into the potential advantages and limitations of transformer models in this domain[5, 6].

With the growing importance of ESC in applications like urban planning, wildlife monitoring, and smart city implementations, understanding the capabilities of these advanced models is not only of academic interest but also of practical significance. This study is poised to contribute to this evolving field by providing a comprehensive analysis of transformer models, benchmarked against traditional CNN approaches, across various environmental sound datasets[7].

1.1 Background

The field of ESC has historically relied on traditional signal processing techniques, which have undergone a transformative shift with the advent of deep learning technologies. This background section delves into the evolution of ESC methodologies, highlighting the transition from traditional approaches to advanced deep learning models, primarily focusing on CNNs and the emerging use of transformer models.

Initially, ESC was primarily rooted in traditional signal processing techniques, involving feature extraction methods like Mel-frequency cepstral coefficients (MFCCs) and spectrogram analysis. These methods were combined with machine learning classifiers such as Support Vector Machines (SVM) and Gaussian Mixture Models (GMM) to classify environmental sounds. The introduction of deep learning, particularly CNNs, marked a significant advancement in ESC. CNNs, with their ability to automatically learn hierarchical feature representations from raw audio data, significantly improved the accuracy and efficiency of sound classification tasks[8]. Further, research explored the depths of CNN architectures, pushing the boundaries of what deep learning could achieve in the context of ESC[9].

The groundbreaking work of Salamon and Bello[8] established a benchmark in utilizing CNNs for ESC. However, the complex and unstructured nature of environmental sounds, especially in bioacoustic contexts like bird species recognition, demands more sophisticated approaches. In this regard, the research by Bird CLF23 delves into fine-grained audio analysis, highlighting the necessity for advanced techniques in ESC[2].

The background of ESC reflects a dynamic and evolving landscape, transitioning from traditional signal processing and machine learning to sophisticated deep learning approaches. This evolution, marked by the advent of CNNs and the recent introduction of transformer models, sets the stage for new breakthroughs in the field, heralding a new era in the accurate and efficient classification of environmental sounds.

The realm of ESC has witnessed significant advancements with the advent of deep learning techniques, notably CNNs. Historically, CNNs have dominated this field due to their remarkable ability to extract hierarchical features from sound spectrograms[8, 7]. However, the exploration of transformer models, primarily known for their success in natural language processing, has opened new avenues in audio signal processing[10].

With a more recent shifted focus towards transformer models, which are known for their exceptional performance in handling sequential data in natural language processing. The groundbreaking work from

Dosovitskiy et al.[3] introduced the concept of ViT, showcasing their capability to process image data effectively. This development sparked interest in applying transformer architectures to audio data, leading to the exploration of AST, as detailed by[4]. These models, employing self-attention mechanisms, provide an innovative approach to capture complex patterns in environmental sounds, offering a promising alternative to CNNs.

Yuan Gong's introduction of the AST marks a significant leap in this field. This model, being convolution-free and purely attention-based, showcases the power of transformer models in handling audio data. The AST model leverages the self-attention mechanism to analyze the intricacies of audio spectrograms, which is crucial for distinguishing subtle differences in environmental sounds[6].

Additionally, recent advancements have seen the integration of multimodal data, combining audio and visual information using transformer models. This approach has shown potential in enhancing the accuracy and robustness of ESC systems in complex and dynamic environments. The incorporation of AST and other transformer-based models illustrates the field's ongoing commitment to leveraging cutting-edge technology in pursuit of enhanced ESC capabilities.

In summary, the evolution of transformer-based models in ESC represents a paradigm shift from traditional CNN approaches. These developments not only offer more sophisticated tools for sound classification but also pave the way for innovative applications in environmental monitoring and bioacoustic research. The fusion of multi-modal data and the creation of domain-specific models further enhance the potential of ESC, promising significant advancements in the field.

1.2 Problem discussion

Correspondingly, the hypothesis of this study is that ViT and AST will exhibit superior or comparable performance to CNNs in terms of accuracy and efficiency in classifying environmental sounds. In the realm of ESC, recent years have witnessed a transformative shift from traditional signal processing techniques to the adoption of deep learning technologies. This evolution is highlighted by the transition from methods involving feature extraction like MFCCs combined with machine learning classifiers to the introduction of advanced deep learning models, especially CNNs and the emerging use of transformer models.

The field of ESC has historically leveraged traditional signal processing techniques, undergoing a significant transformation with the advent of deep learning. Initially rooted in traditional methods such as spectrogram analysis combined with classifiers like SVM and GMM, the introduction of deep learning has paved the way for more sophisticated tools for sound classification. Recent research has increasingly shifted focus towards transformer models, such as ViT [3]and AST[6], which have demonstrated promising results in both image and audio processing tasks. These models, leveraging self-attention mechanisms, are particularly adept at capturing long-range dependencies in data, a feature crucial in dissecting the complex nature of environmental sound.

However, there exists a gap in comprehensive analysis and benchmarking of these transformer models against traditional CNN approaches across various environmental sound datasets. This gap presents an opportunity to evaluate the efficacy of transformer models in ESC, particularly focusing on their suitability for general environmental sound scenarios compared to the widely-used CNNs. Given this background, this paper aims to address the following research question:

How does the performance of transformers-based models compare with traditional Convolutional Neural Networks (CNNs) in the context of environmental sound classification?

The hypothesis of this study is that ViT and AST will exhibit superior or comparable performance to CNNs in terms of accuracy and efficiency in classifying environmental sounds. This exploration is motivated by the need to understand the capabilities of these advanced models, which are not only of academic interest but also hold significant implications for applications in urban planning, wildlife monitoring, and smart city implementations.

2 Research Method

This section is structured to offer a comprehensive understanding of the methodologies employed, ensuring transparency and reproducibility of the research process. The methods adopted are tailored to address the research question effectively, ensuring a robust and scientifically valid approach to the investigation. The first part will be dealing with the general approach undertaken where the second part will deal with a more specific description of how the chosen research method will be applied.

2.1 Choice of Research Method

For this study, the primary focus was on quantitative data. Quantitative data, in the form of numerical values representing audio signal characteristics, were essential for applying and evaluating deep learning models. The audio signals, once converted into mel spectrograms, presented a quantifiable representation of sounds, enabling the application of advanced computational techniques for classification.

The collection of this data did not involve direct interactions or observations in natural settings. Instead, the data comprised pre-recorded and processed audio signals, ensuring that the research did not intrude upon or disturb natural habitats or wildlife. This approach aligns with ethical research practices in environmental studies, prioritizing the minimization of ecological impact.

The analysis of the collected data was also quantitative. Utilizing deep learning models, the mel spectrograms underwent a process of feature extraction, classification, and validation. The quantitative nature of this analysis facilitated precise, objective, and reproducible assessments of the model's performance.

To assess the performance of the deep learning models, the quantitative metric accuracy was employed. Accuracy, representing the overall effectiveness of the model in correctly classified sounds. The choice of this metric was motivated by its widespread acceptance and reliability in machine learning research. Furthermore, to test the significance of the different employed models a statistical test was used to underline the significance of the results achieved by different classifiers.

In conclusion, the quantitative approach in both data collection and analysis was pivotal in addressing the research question. This approach allowed for a precise evaluation of the model's capabilities in classifying sounds from mel spectrograms, ensuring that the research outcomes were grounded in objective and replicable methods.

2.2 Application of Research Method

In applying the chosen research method, a specific dataset containing bird sound recordings were employed[2]. This dataset was selected based on their diversity of bird species, sound quality, and the availability of labeled data for supervised learning. The dataset provided audio files along with corresponding labels indicating the bird species, facilitating a structured approach to supervised learning.

The raw audio files from this dataset was transformed into mel spectrograms, a visual representation of the sound spectrum over time. This conversion was achieved using the Librosa library in Python, which allowed for consistent and standardized spectrogram generation. Parameters such as the number of mel bands, hop length, and window size were carefully chosen to ensure that the spectrograms effectively captured the distinguishing features of different bird calls.

The training process involved dividing the dataset into training, validation, and testing subsets. The training set was used to train the model, while the validation set assisted in tuning hyperparameters and preventing overfitting.

Upon training, the model's performance was evaluated using the predefined metric accuracy. This evaluation was conducted on the test set, which consisted of unseen data, to assess the model's ability to generalize to new samples. The testing phase provided a critical evaluation of the model's effectiveness in classifying bird species from audio recordings.

Throughout the application of the research method, ethical considerations were upheld. The use of publicly available, non-invasive datasets ensured that the research did not negatively impact wildlife or their habitats.

Additionally, the research was conducted with the intention of contributing positively to ecological studies and biodiversity monitoring, aligning with broader environmental conservation goals.

3 Method(s)

3.1 Ethical Considerations

In conducting this research on environmental sound classification, we ensured adherence to ethical standards at every stage. The recordings were obtained from publicly available datasets or sources that explicitly permitted their use for research purposes, ensuring no infringement on proprietary rights or disturbance to natural ecosystems. Furthermore, the data did not contain any personally identifiable information, thereby eliminating concerns regarding privacy infringement[2].
Throughout the analysis, we maintained the integrity of the data, using it solely for the intended scientific inquiry without manipulation that could lead to misleading conclusions. In terms of the broader implications of our work, we recognized the potential impact of sound classification technologies on privacy and surveillance. Therefore, we emphasized the importance of using such technologies responsibly, advocating for their application primarily in non-invasive environmental monitoring and biodiversity research. We discouraged the use of our findings in any manner that could lead to the infringement of individual privacy or be detrimental to wildlife.
Our research complied with all relevant institutional and legal guidelines for ethical research, and we advocated for continued ethical vigilance in the field of environmental audio processing. By considering these ethical aspects, we aimed to contribute positively to the scientific community and society, promoting the use of technology for beneficial and non-intrusive applications.

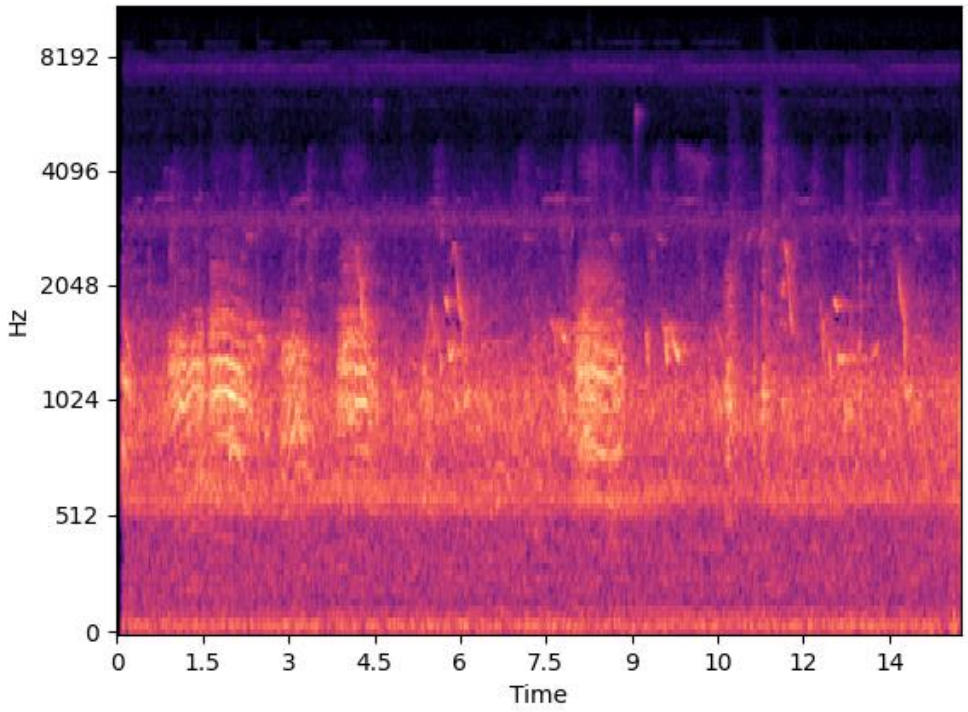

Figure 1: A Sample of Spectrograms Images

3.2 Data Preparation

To set the context for the model construction and data analysis, it was imperative to first establish a comprehensive understanding of the dataset used. Initially, the bird species dataset encompassed 264

distinct classes. However, a subset of these classes was notably underrepresented, each containing fewer than five samples. This scarcity of data posed significant challenges in creating balanced and effective training, validation, and testing splits. Consequently, to address this issue, these underrepresented classes were excluded from the analysis, resulting in a refined dataset comprising 247 classes. Furthermore, to manage computational demands effectively and streamline the classification process, the focus was narrowed to the first 20 classes that had the highest sample counts. This strategic reduction in dataset complexity was a critical step in ensuring the feasibility and robustness of the subsequent modeling and analysis phases.

The preprocessing of the audio data was executed through a multi-step procedure on a locally configured workstation, equipped with a 2.3 GHz 8-Core Intel Core i9 processor, an Intel UHD Graphics 630 1536 MB graphics card, and 32 GB of 2667 MHz DDR4 memory. This process involved transforming the sound recordings into spectrograms, utilizing the Python librosa library. Additionally, normalization of the input was performed by scaling it in accordance with its maximum and minimum values. The final step in the preprocessing phase entailed converting the power spectrum into decibels (dB), which facilitated improved visualization and analysis of the audio data.

3.3 Training Setting

The training of the models was conducted using Google Colab's complimentary infrastructure, which provided access to a T4 NVIDIA graphics card equipped with 2.560 NVIDIA CUDA cores. To facilitate an efficient workflow, the pre-processed mel-spectrograms were stored on Google Drive and seamlessly integrated with Google Colab's virtual environment. The models were trained using PyTorch, with specific adaptations made to tailor them to our research requirements.

3.3.1 ViT Based Method

For the Vision Transformer, the ViT_b_17 model from PyTorch's model repository was employed, incorporating its default weights. Crucially, the classifier head of this model was customized to align with our specific needs, including the correct number of output nodes and the integration of a softmax activation function for classification purposes.

3.3.2 CNN Based Method

In parallel, a standard ResNet-50 model was utilized for the CNN component of our study. This choice was driven by ResNet-50's optimal balance between computational efficiency and model complexity and success in image classification. Similar to the Vision Transformer, the ResNet model was initialized with its default weights, ensuring a consistent and reliable foundation for our comparative analysis. Both models used a learning rate of $5e^{-6}$.

3.3.3 AST Based Method

In the process of fine-tuning using the AST structure, we followed the above-mentioned experimental environment and combined the api rules provided on Huggingface for program design. For the processing of experimental data, in order to ensure that various models have the ability to compare with each other, we use the same method to process audio. That is, use the relevant functions of the librosa library to generate mel spectrogram. After generation, use Dataset to encapsulate classes and perform pad, feature extract and other operations. We set the basic training parameters of AST as learning-rate = $3e^{-5}$, warmup ratio=0.1, logging steps = 10, train epoch = 3. Select MIT/ast-finetuned-audioset-10-10-0.4593 as the pre-training model.

It is important to clarify our decision to run the model for only 3 epochs. This choice is primarily informed by the fact that the AST model utilized in our study has undergone extensive pretraining on an upstream task. Pretraining on a large dataset equips the model with a robust foundational understanding of audio features, thus significantly reducing the need for extensive training on the downstream task. Consequently,

the volume of data required for effective fine-tuning in the downstream task is considerably smaller. This factor, coupled with the efficient learning capabilities of the pretrained model, renders a prolonged training phase unnecessary. Therefore, a concise training duration of 3 epochs is deemed sufficient for the model to adapt and perform effectively on the downstream environmental sound classification task.

3.4 Metrics

In our study, the performance of the models was evaluated by calculating and plotting both the validation accuracy and the associated losses every 10 epochs. This approach provided a clear and continuous insight into the models' performance throughout the training process. To rigorously assess the significance of the models' accuracies, we employed a k-fold cross-validation technique on the training dataset. This method ensured a thorough and unbiased evaluation by dividing the dataset into k subsets and iteratively using one subset for validation while training on the remaining subsets.

Furthermore, to ascertain the statistical significance of the differences in performance between the models, a Welch t-test was conducted. This statistical test was particularly chosen for its robustness in handling potential disparities in sample variances and sizes, making it an appropriate tool for comparing the effectiveness of our models.

4 Results and Analysis

4.1 CNN and ViT Results

As in the Method section 3 described the experiments were run in google coolab using a Nvidia T4 GPU, the implemented code in pytorch was run for the CNN as well as the Vision Transformer until the GPU from coolab disconnected. Meanwhile every 10 epochs the model got saved and the corresponding validation accuracies and losses were plotted. For the first experiment, the training of a CNN model on the reduced bird dataset containing 20 species, a validation accuracy of around 60% was reached. Regarding the training loss it can be observed that after 60 epochs the loss seems to stagnate, while the validation loss starts to increase at around the same time. The second experiment, the training of a vision transformer on the reduced bird dataset containing 20 species, reached a validation accuracy of around 40%. Both losses in this experiment still decrease after 250 epochs while a stronger fluctuations can be seen in the training loss.

In accordance with the methodology outlined in Section 3, the experiments were conducted using Google Colab's Nvidia T4 GPU. The code, implemented in PyTorch, was executed for both the CNN and the Vision Transformer, continuing until the Colab GPU session ended. During this process, the models were saved and the corresponding validation accuracies and losses were recorded and visualized every 10 epochs.

The initial experiment involved training a CNN model on the reduced bird dataset containing 20 species. This experiment achieved a validation accuracy of approximately 60% . It was noted that the training loss began to plateau after 60 epochs, while the validation loss started to increase around the same epoch mark.

In the second experiment, a Vision Transformer was trained on the same reduced dataset of 20 bird species. This approach yielded a validation accuracy of around 40%. Notably, both the training and validation losses continued to decrease even after 250 epochs, although more pronounced fluctuations were observed in the training loss.

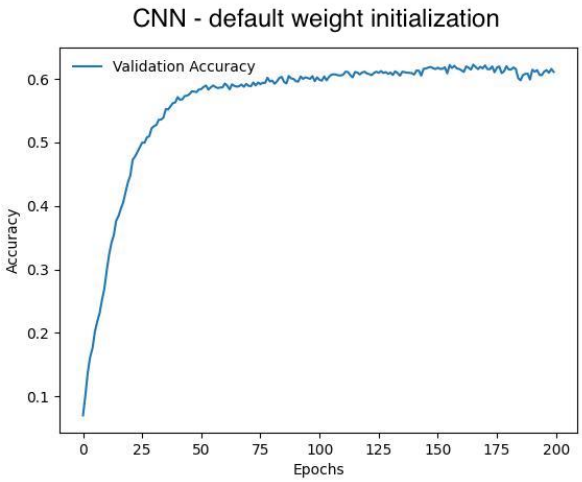

Figure 2: CNN: validation accuracy 200 epochs

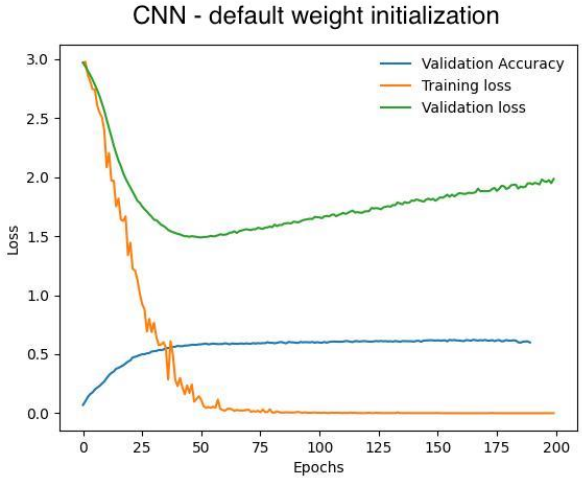

Figure 3: CNN: validation and training losses 200 epochs

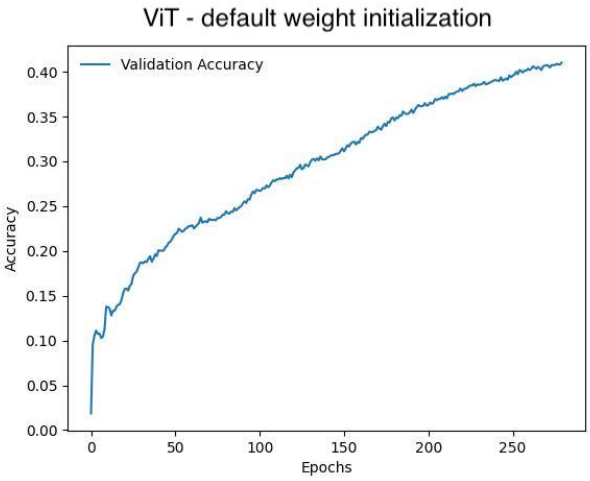

Figure 4: ViT: validation accuracy 200 epochs

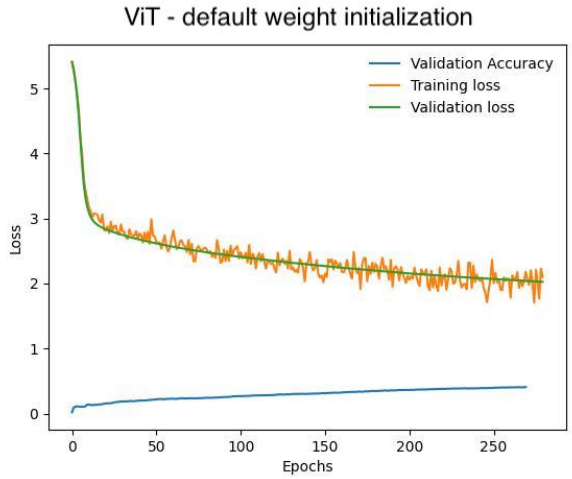

Figure 5: ViT: validation and training losses 200 epochs

4.2 AST Results

The AST model was also trained on the reduced bird dataset containing 20 species, following the same procedure as the CNN and Vision Transformer experiments. This training was conducted using Google Colab's Nvidia T4 GPU, and the experiment's progress was monitored with checkpoints every 10 epochs.

The AST model achieved a validation accuracy of around 88.35% by the end of the third epoch. This performance is significantly higher compared to the CNN (60%) and Vision Transformer (40%) models. The high accuracy of the AST model suggests its effectiveness in audio-based classification tasks, likely due to its design that specializes in handling spectrogram inputs. And the AST model showed a consistent decrease in training loss over the epochs, from 0.6984 in the first epoch to 0.1634 in the third epoch. This steady decline indicates effective learning and model optimization. The validation loss of the AST model decreased initially but showed a slight increase in the second epoch before decreasing again in the third epoch. This fluctuation might suggest some variability in the model's performance on the validation set, but overall, the loss trend is downward.

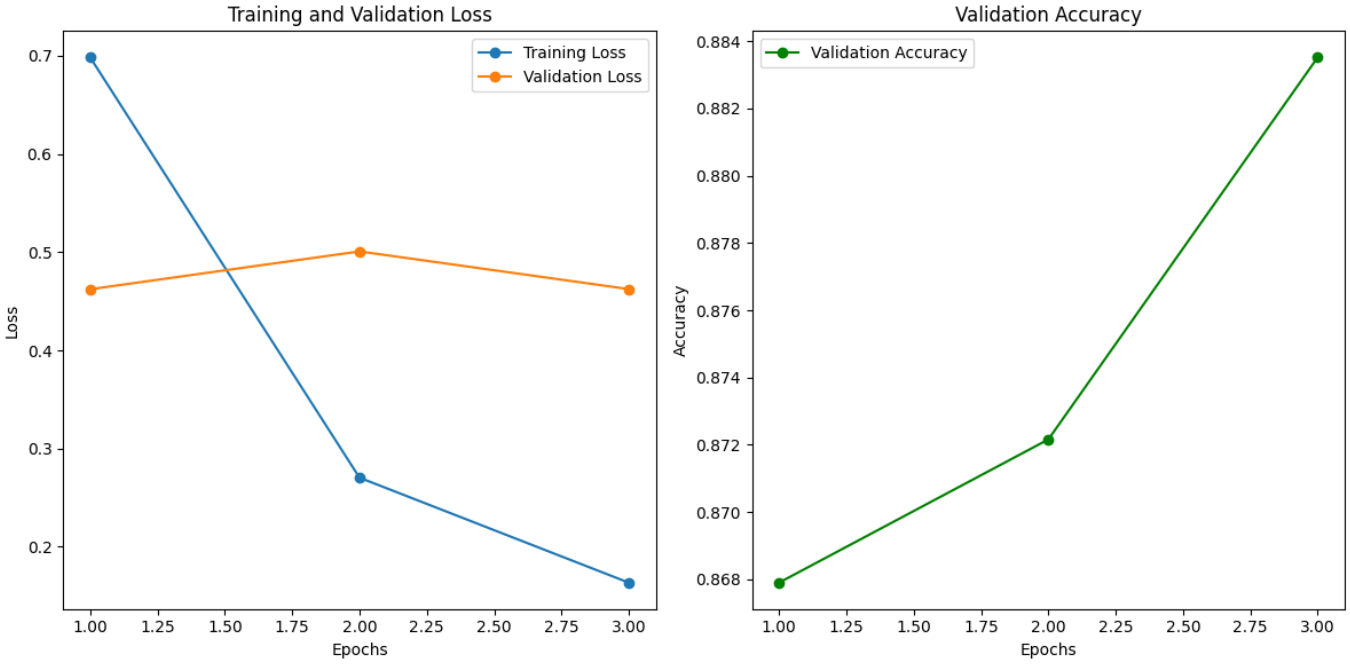

Figure 6: AST: validation and training losses 3 epochs & accuracy

4.3 Weltch's t-test Results

In this study, we conducted Welch's t-tests to evaluate the statistical significance of the differences in classification performance between different models: CNN, ViT, and AST. The Welch's t-test is particularly suited for our analysis as it does not assume equal variances between groups, which aligns with our dataset characteristics. We set our significance level (α) at 0.05 for these tests. We put the detailed results in our A.3.

4.3.1 CNN vs. ViT Model Comparison

Hypotheses:

- H_0: The mean performance score of the CNN model is equal to the mean performance score of the ViT model.
- H_1: The mean performance score of the CNN model is not equal to the mean performance score of the ViT model.

Results: The Welch's t-test yielded a t-statistic of 4.2459 and a p-value of 0.00307. Given that the p-value is less than the significance level of 0.05, we reject the null hypothesis (H_0). This result suggests a statistically significant difference in the mean performance scores between the CNN and ViT models, with the CNN model outperforming the ViT model.

4.3.2 CNN vs. AST Model Comparison

Hypotheses:

- H_0: The mean performance score of the CNN model is equal to the mean performance score of the AST model.
- H_1: The mean performance score of the CNN model is not equal to the mean performance score of the AST model.

Results: The t-test for this comparison reported a t-statistic of 20.4262 and a p-value of approximately 3.39e-05. The extremely low p-value, much less than our α level, leads us to reject the null hypothesis (H_0). This finding indicates a significant difference in the mean performance scores, with the CNN model showing a different level of performance compared to the AST model.

4.3.3 Implications of the T-test Results

These results demonstrate statistically significant differences in the performance of the models tested. The rejection of the null hypotheses in both cases indicates that the mean performance scores of the CNN model are significantly different from those of the ViT and AST models. This finding underscores the effectiveness of the CNN model in environmental sound classification, but it also highlights the need for further investigation into why the CNN model shows superior performance. Factors such as model architecture, data representation, and training methodologies might contribute to these differences and warrant detailed exploration in future studies.

In our study, we employed k-fold cross-validation for evaluating the performance of the CNN, ViT, and AST models in classifying environmental sounds. While k-fold cross-validation is a widely accepted method for model evaluation, it's important to consider the potential limitations of not using stratified k-fold cross-validation. The lack of stratification in our k-fold cross-validation approach might have led to higher variance and bias in model performance across different folds, especially if certain environmental sounds were unevenly distributed. This can obscure the true performance of the models, particularly in their ability to generalize across different types of sounds. To mitigate these limitations, future experiments should consider employing stratified k-fold cross-validation. This approach would ensure that each model

is evaluated on a more balanced and representative mix of sound classes, leading to more accurate and generalizable findings. Additionally, it would enhance the robustness of statistical tests like the Welch's t-test by providing a more uniform basis for comparison.

5 Discussion

5.1 Vision Transformer and CNN

- Answer: CNNS outperforms the Vision transformer and achieves a decent accuracy after a relatively short time.
- Result: The CNN though seems to overfit the training data since the loss of the validation goes towards a upward trend. High variance. The vision transformer has at the beginning a decent learning curve but than flatten outs, for further reaserch it would be interesting to investigate when the vision transformer reaches a plaetau, since right now it seems still to learn, although with a relative slow rate, not very efficient.
- Support: Vision Transformers (ViTs) can struggle to match the performance of Convolutional Neural Networks (CNNs) when trained from scratch on small-scale datasets. This issue primarily stems from the absence of inherent inductive biases, locality, and hierarchical structure in ViTs, which are typical in CNN architectures.[11]

5.2 AST Model's Superior Accuracy

- Answer: The AST model outperforms both CNN and Vision Transformer models in terms of validation accuracy.
- Result: The validation accuracy achieved by the AST model is approximately 88.35%, significantly higher than the CNN's 60% and the Vision Transformer's 40%. One key difference between ViT and AST is the use of overlapping patches in AST. This design choice is significant because overlapping patches can capture more nuanced relationships between different parts of the audio spectrogram. In a typical spectrogram, adjacent time-frequency areas often have correlated information, especially in environmental sounds where such correlations can signify important acoustic events. By allowing the patches to overlap, AST can better leverage these correlations, leading to improved performance in tasks like environmental sound classification.
- Support: This aligns with research by Yuan Gong [6], who demonstrate that models specifically designed for audio processing, like the AST, tend to perform better in audio classification tasks.
- Next Steps: Investigating the AST model's performance on a more extensive dataset and exploring its response to various hyperparameter adjustments could provide deeper insights into its capabilities.

References

[1] S. Chachada and C.-C. J. Kuo, "Environmental sound recognition: A survey," *APSIPA Transactions on Signal and Information Processing*, vol. 3, p. e14, 2014.

[2] S. Kahl, T. Denton, H. Klinck, H. Reers, F. Cherutich, H. Glotin, H. Goëau, W.-P. Vellinga, R. Planqué, and A. Joly, "Overview of birdclef 2023: Automated bird species identification in eastern africa," *Working Notes of CLEF*, 2023.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[4] A. F. R. Nogueira, H. S. Oliveira, J. J. Machado, and J. M. R. Tavares, "Transformers for urban sound classification—a comprehensive performance evaluation," *Sensors*, vol. 22, no. 22, p. 8874, 2022.

[5] D. Elliott, C. E. Otero, S. Wyatt, and E. Martino, "Tiny transformers for environmental sound classification at the edge," *arXiv preprint arXiv:2103.12157*, 2021.

[6] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[7] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[8] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal processing letters*, vol. 24, no. 3, pp. 279–283, 2017.

[9] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 421–425.

[10] A. Bansal and N. K. Garg, "Environmental sound classification: A descriptive review of the literature," *Intelligent Systems with Applications*, p. 200115, 2022.

[11] H. Gani, M. Naseer, and M. Yaqub, "How to train vision transformer on small-scale datasets?" *arXiv preprint arXiv:2210.07240*, 2022.

A Insensible Approximation

A.1 CNN validation and training losses after 100 epochs

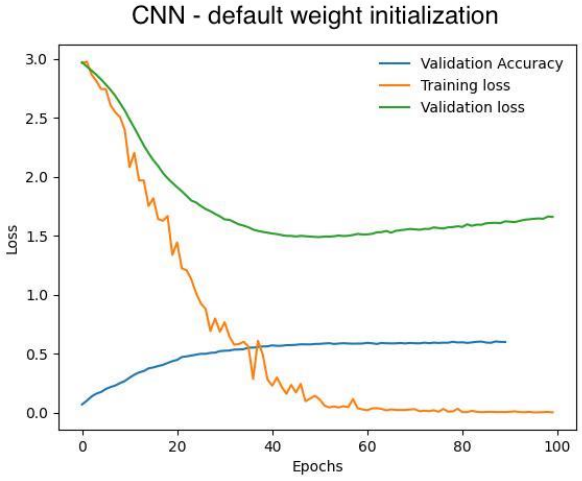

Figure 7: CNN: validation and training losses 100 epochs

A.2 Welch's t-test on the classification results of the CNN and ViT dataset

For testing the statistical significance of both classification results, with a significance value $\alpha = 0.05$, the following scores were used:

- k-fold from CNN: 64.0625, 69.375, 55.59322033898305, 60.15625, 53.125
- k-fold from ViT: 37.8125, 31.864406779661017, 48.125, 50.390625, 36.328125

Therefore the H_0 and H_1 hypothesis are:
H_0: *The mean performance score of the CNN model is equal to the mean performance score of the ViT model.*
H_1: *The mean performance score of the CNN model is not equal to the mean performance score of the ViT model.*
The following assumptions have to hold for a t-test:

1. The observations in one sample should be independent of the observations in the other sample.
2. The data in both samples was obtained using a random sampling method.
   (a) The two statements above are given for free in this case.
3. The data should be approximately normally distributed.
   (a) To test normality a Shapiro Wilk test was performed on both data series. They resulted in a $p-value$ of 0.8770557045936584 and 0.43037623167037964 respectively. Since both $p-values$ are larger than 0.05 we can assume normality.
4. The two samples should have approximately the same variance. If this assumption is not met, a Welch's t-test should be performed.
   (a) The variances are 34.00704718471703 and 50.86982893183533 respectively, since these are not approximately the same a Welch's t-test was performed.

The obtained result of the Welch's t-test was a statistic value of 4.245859096615145 and a *pvalue* of 0.003073694853434769. Since the *pvalue* is < 0.05 we have reason to reject the H_0 hypothesis.

A.3 Welch's t-test on the classification results of the CNN and AST dataset

For testing the statistical significance of both classification results, with a significance value $\alpha = 0.05$, the following scores were used:

- k-fold from CNN: 64.0625, 69.375, 55.59322033898305, 60.15625, 53.125
- k-fold from AST: 0.8680555555555556, 0.9357142857142857, 0.9142857142857143, 0.8785714285714286, 0.9214285714285714

Therefore the H_0 and H_1 hypothesis are:
H_0: *The mean performance score of the CNN model is equal to the mean performance score of the AST model.*
H_1: *The mean performance score of the CNN model is not equal to the mean performance score of the AST model.*
The following assumptions have to hold for a t-test:

1. The observations in one sample should be independent of the observations in the other sample.
2. The data in both samples was obtained using a random sampling method.
   (a) The two statements above are given for free in this case.
3. The data should be approximately normally distributed.
   (a) To test normality a Shapiro Wilk test was performed on both data series. They resulted in a $p-value$ of 0.8770557045936584 and 0.48218968510627747 respectively. Since both $p-values$ are larger than 0.05 we can assume normality.

4. The two samples should have approximately the same variance. If this assumption is not met, a Welch's t-test should be performed.

   (a) The variances are 34.00704718471703 and 0.0006706412194507425 respectively, since these are not approximately the same a Welch's t-test was performed.

The obtained result of the Welch's t.-teest was a statistic value of 20.426175832638407 and a *pvalue* of $3.39126914658412e-05$. Since the *pvalue* is < 0.05 we have reason to reject the H_0 hypothesis.