25·29 August 2024
inter.noise
NANTES FRANCE

Comparative analysis of CNN and TNN models in environmental noise source identification

Antonio Topalović[1]
IMS Merilni Sistemi
Cesta Ljubljanske brigade 23A, 1000 Ljubljana, Slovenija

Metod Zupančič[2]
IMS Merilni Sistemi
Cesta Ljubljanske brigade 23A, 1000 Ljubljana, Slovenija

Jure Bovha[3]
IMS Merilni Sistemi
Cesta Ljubljanske brigade 23A, 1000 Ljubljana, Slovenija

Thomas King[4]
Hottinger Brüel & Kjaer
Teknikerbyen 28, DK-2830 Virum, Denmark

Abstract

This paper presents a focused examination of environmental noise, an issue of relevance due to its implications for human health and well-being, compliant with ISO 1996-2:2017 standards. Moving beyond the traditional methodologies that primarily rely on level exceedance, our study integrates advanced machine learning techniques to address the challenges in noise source identification. With the advent of IoT Noise Monitoring Terminals, the conventional manual auditory analysis methods have become less feasible. Our research explores the application of convolutional neural networks (CNNs), a standard in environmental noise analysis, and assesses the emerging utility of Transformer Neural Network (TNN) models in this domain. The aim is to conduct an objective comparison of these models, applying them to identical datasets to determine their effectiveness in identifying noise sources. Through this analysis, the study seeks to contribute to the field of environmental acoustics by offering insights into the comparative strengths and limitations of CNN and TNN models.

[1]antonio.topalovic@ims.si

[2]metod.zupancic@ea-lab.eu

[3]jure.bovha@ea-lab.eu

[4]thomas.king@hbkworld.com

1. INTRODUCTION

In the contemporary landscape of technological advancement, the quality of human life is intricately linked with our ability to understand and manage environmental factors. Among these, noise pollution stands out as a critical concern, given its significant implications on health and well-being. The importance of environmental noise monitoring cannot be overstated, as recognized by ISO 1996-2:2017 standards, which underscore the necessity for accurate noise source identification and management. This paper presents a focused examination of the challenges associated with environmental noise, particularly in urban settings where the amalgamation of various sound sources complicates the task of noise classification. Traditional methodologies, while effective in capturing level exceedance, fall short in addressing the nuanced requirements of noise source identification. In this context, the advent of Internet of Things (IoT) Noise Monitoring Terminals marks a pivotal shift from conventional manual auditory analysis methods to more sophisticated, data-driven approaches. Given the current state-of-the-art (SOTA) Transformer models' success across various fields, including text-based applications like ChatGPT, exploring these models for audio classification represents a promising research direction. This paper aims to compare the Transformer model, specifically the AST model [1], for its potential in handling the complexities of urban sound classification, against the previously favored Convolutional Neural Network (CNN) architecture. By doing so, it seeks to align with the emerging trends in environmental noise analysis, leveraging advanced machine learning techniques to overcome the limitations of traditional methods.

2. BACKGROUND

The advancement of machine learning techniques has significantly impacted the field of audio classification, particularly in the context of environmental noise monitoring. The selection of appropriate models for the classification of urban sounds is pivotal, guiding the transition from traditional methodologies to innovative, data-driven approaches. This section explores the evolution of model selection for audio classification, from the dominance of Convolutional Neural Networks (CNNs) to the emerging interest in Transformer-based models.

2.1. Previous Model Selection (CNN) for Audio Classification

Convolutional Neural Networks (CNNs) have long been the preferred choice for audio classification tasks. The theoretical foundation of CNNs lies in their ability to automatically and adaptively learn spatial hierarchies of features from input data. CNNs excel in handling data with a grid-like topology, such as images, which extend to audio data when represented as spectrograms or mel-frequency cepstral coefficients (MFCCs). The use of convolution layers allows CNNs to capture local dependencies and scale invariance in the data, making them particularly suitable for extracting features from complex audio signals. This capability, coupled with their efficiency and generalization power, has made CNNs a go-to model for a variety of audio classification tasks, including environmental noise classification.

2.2. Transformer-Based Models in Audio Classification

The introduction of Transformer models [2] marked a paradigm shift in the handling of sequential data, initially revolutionizing the field of natural language processing (NLP). Unlike their predecessors, Transformers do not require sequential data to be processed in order. This is achieved through the mechanism of self-attention, which weights the significance of different parts of the input data independently of their sequential order. This novel approach presents a promising avenue for audio classification, where the complex structure of audio signals can benefit from the Transformer's ability to capture long-range dependencies without the constraints of sequential processing.

Recent works exploring the application of Transformer models in audio classification have demonstrated their potential. For instance, the AST (Audio Spectrogram Transformer) [1] model has shown promise in capturing the rich, hierarchical structures within audio data, offering a new perspective on urban sound classification. Although research on Transformers in audio classification is still emerging, preliminary findings suggest their capability to outperform traditional models in certain contexts.

3. LITERATURE REVIEW

Combining the insights from both CNN and Transformer model research in audio classification provides a comprehensive overview of the current landscape. While CNNs have established a strong foothold due to their proven efficacy and robustness in handling audio data, Transformers are gaining traction for their innovative approach to sequential data analysis. The literature reveals a growing interest in leveraging the strengths of both model types to address the complexities of audio classification tasks. Studies comparing the performance of CNNs and Transformers on audio datasets, such as Audio Set [3], FreesoundDataset - FSD50K [4], UrbanSound8K [5], offer valuable insights into the advantages and limitations of each model type. By synthesizing these findings, this review aims to foster a nuanced understanding of how contemporary machine learning models can be harnessed to advance the field of environmental noise classification, aligning with the evolving standards and expectations of noise monitoring and management. This exploration underscores the importance of continuous innovation in model selection and development, highlighting the dynamic interplay between established methodologies and cutting-edge research in shaping the future of audio classification.

3.1. Convolutional Neural Network

- Piczak, K. J. (2015) [6] introduced the early use of CNNs for environmental sound classification through spectrogram analysis, setting a foundational approach for audio recognition tasks.

- Hershey et al. (2017) [7] developed the VGGish model, adapting VGG image recognition architectures to audio, pioneering the application of image recognition techniques to audio classification.

- Takahashi, Gygli, & Pfister (2016) [8] highlighted the efficacy of deep CNNs combined with data augmentation for acoustic event recognition, emphasizing the model's adaptability to augmented audio data.

- Salamon & Bello (2017) [9] applied deep CNNs to the UrbanSound8K dataset, demonstrating significant improvements in urban sound classification and the critical role of data augmentation.

- The thesis of Metod Zupančič (2021) [10] provides a detailed account of leveraging MFCC and MEL features in conjunction with convolutional neural networks for environmental sound classification. Notably, it offers a set of model checkpoints that facilitated the direct comparison and evaluation of CNNs against Transformer models in this study. While focusing on the improvement of noise detection solutions, the thesis highlights the utility of specific audio features and augmentation techniques in achieving high classification accuracy. This resource proved instrumental in establishing a baseline for performance comparison within our research.

3.2. Transformer Neural Network

The evolution of machine learning models, particularly in the context of audio classification, has been marked by significant advancements from Convolutional Neural Networks (CNNs) to Transformer models. While CNNs have been the cornerstone in the analysis of audio data, offering robustness and efficacy, the landscape began to shift with the seminal work "Attention is All You Need" by Vaswani et al. [2]. This paper introduced the Transformer model, which revolutionized the way sequential data is processed, emphasizing the use of attention mechanisms over traditional sequence-based processing methods like RNNs and CNNs. The Transformer model's ability to handle sequences in parallel, without the need for recurrent connections, paved the way for its adaptation beyond natural language processing (NLP) into other domains, including audio classification.

The subsequent publication "An Image is Worth 16x16 Words" by Dosovitskiy et al. [11] extended the application of Transformers to the visual domain, proposing the Vision Transformer (ViT) model. By treating image patches as tokens similar to words in NLP, this work demonstrated that Transformers could achieve state-of-the-art results on image classification tasks, challenging the dominance of CNNs in computer vision. This innovative approach underscored the model's versatility and its potential for cross-domain applicability, setting the stage for further explorations into how Transformers could be leveraged for audio and environmental noise classification.

Building on these foundational works, the Audio Spectrogram Transformer (AST) [1] model paper further adapted the Transformer architecture to the task of audio classification. By applying the principles of ViT to spectrogram images of sound, the AST model offered a novel method for capturing the complex patterns within audio data. This adaptation highlighted the Transformer's capability to understand and classify diverse sound environments, including the challenging nuances of urban soundscapes as represented in the UrbanSound8K dataset.

In synthesizing the insights from these pivotal studies, it becomes evident that the journey from "Attention is All You Need" [2] through "An Image is Worth 16x16 Words" [11] to the AST model [1] marks a significant evolution in our approach to audio classification. This literature review underscores the transformative impact of Transformer models in the field, showcasing their potential to redefine the boundaries of environmental noise classification. The exploration of Transformers in audio classification not only aligns with the evolving standards and expectations of noise monitoring and management but also highlights the importance of continuous innovation in model selection and development. The dynamic interplay between established methodologies and cutting-edge research is shaping the future of audio classification, promising new avenues for understanding and managing the complex world of urban sounds.

4. METHODOLOGY

4.1. Dataset Description (UrbanSound8K)

The UrbanSound8K dataset is a compilation of 8,732 labeled sound clips of urban noises, each lasting up to 4 seconds. These clips are drawn from field recordings and are classified into 10 categories that are common in urban environments: air conditioners, car horns, children playing, dogs barking, drilling, idling engines, gunshots, jackhammers, sirens, and street music. This categorization is designed to encapsulate a wide range of urban soundscapes, making the dataset a comprehensive resource for training and evaluating models for urban sound classification. The diversity and real-world complexity of the sounds in UrbanSound8K render it particularly suitable for this study, offering a robust framework for assessing the efficacy of different audio classification models.

4.2. Audio Features

Mel-Frequency Cepstral Coefficients (MFCCs) are a cornerstone in audio signal processing for sound classification and speech recognition. The derivation of MFCCs mimics the human auditory system, beginning with the division of the sound signal into short frames and converting these through a Fourier transform to analyze frequency components. Critical to this process is the application of the Mel scale through a series of filters, reflecting the human ear's logarithmic sensitivity to pitch. The culmination of this process involves the Discrete Cosine Transform (DCT) on the log Mel spectra, resulting in MFCCs that encapsulate the auditory-relevant characteristics of sound.

Mel Spectrograms provide a complementary analysis, visually representing the sound's frequency content over time on the Mel scale. This visual adjustment is essential for aligning frequency representation with human auditory perception, especially for distinguishing between diverse sound types in urban environments. The generation involves dividing the sound signal into frames and mapping the frequency spectrum onto the Mel scale, creating a spectrogram that intuitively displays sound energy and amplitude.

In urban sound classification, the nuanced capabilities of MFCCs and Mel spectrograms are invaluable. MFCCs excel in distilling sound patterns conducive to classification amidst urban noise, while Mel spectrograms offer a visual, intuitive grasp of sound events over time. Their integration into sound classification models capitalizes on their respective strengths, enhancing model performance through a sophisticated understanding of sound as perceived by human auditory mechanisms.

4.3. Test Methodology

The experimental setup acknowledges the substantial computational resources required for training advanced machine learning models, such as CNNs and the untrained TNN model. To accommodate the computational demands of our experiments, an Nvidia A6000 GPU was employed. This choice of hardware reflects the resources that were actually utilized for the research process, enabling the training and evaluation of sophisticated models to proceed without being hindered by immediate hardware constraints.

This methodological decision underscores the practical aspect of conducting machine learning research, where the availability of computational resources can directly impact the scope and depth of possible investigations. By specifying the hardware used, we aim to provide a transparent overview of the experimental conditions, allowing for a clearer interpretation of the results and their reproducibility in similar or differently resourced environments.

5. RESULTS

After delving into the comparative performance of CNN and TNN models in the context of shared augmentation techniques, it's pivotal to acknowledge the broader scope of our study and the limitations encountered during the experimentation phase.

5.1. Model Configuration Adjustments

Recognizing the need for efficiency and mindful of the extensive training times associated with larger models, we implemented the following adjustments:

– Reduction in Encoder Layers: We reduced the number of encoder layers from the original 12 to a single layer, based on preliminary tests that showed no substantial difference in performance for our specific preprocessed data.

– Adjustment in MLP Nodes: The model's multilayer perceptron (MLP) component was adjusted from 3072 nodes to 1536 nodes, as the larger configuration did not yield significant improvements in performance for our data.

These adjustments to the TNN model were driven by a pragmatic approach to experimentation, seeking to balance the model's complexity with our resource constraints and the unique demands of our dataset.

5.2. Results of 10-fold Cross-validation

Table 1: Results of 10-fold cross-validation using augmented data for MFCC and MEL features

	MFCC			MEL		
	Mean Acc	**Mean Loss**	**Mean Epochs**	**Mean Acc**	**Mean Loss**	**Mean Epochs**
CNN	0.984	0.187	384.2	0.974	0.253	231.3
TNN	0.635	1.582	6.2	0.653	1.450	15.5
	SD Acc	**SD Loss**	**SD Epochs**	**SD Acc**	**SD Loss**	**SD Epochs**
CNN	0.007	0.034	225.5	0.010	0.038	78.7
TNN	0.039	0.240	3.7	0.045	0.344	7.1

5.3. Augmentation Techniques and Model Performance

While both models benefited from the augmentation strategies implemented, it's important to note that the TNN model's performance, though not reaching the same level as the CNN, did demonstrate improvement from these techniques. This improvement, however, was constrained by the trial-and-error approach in hyperparameter optimization, a process inherently limited by the available computational resources and the project's timeframe.

5.4. Considerations on the TNN Model's Potential

The disparity observed in the TNN model's performance compared to its documented success in handling larger datasets (in the original AST model paper [1]), such as Audio Set [3], suggests that our implementation has not fully tapped into its capabilities. This is attributed to the iterative nature of finding optimal hyperparameters within the constraints of our hardware and the preliminary understanding of the model's architecture. It's reasonable to infer that with extended experimentation and access to more substantial computational resources, the TNN model's true potential could be realized, possibly surpassing CNN models in similar tasks.

The table above presents a stark contrast in performance between the CNN and TNN models, highlighting the effectiveness of our tailored approach for the CNN and the potential areas for improvement in the TNN model.

6. MOVING FORWARD

In reflecting on our findings and considering the pathways for future exploration, it's essential to highlight the evolving landscape of machine learning models in audio classification. Notably, the AST model adaptation [1], grounded in the architecture of Vision Transformers (ViT) [11], presents promising avenues for enhancement through the utilization of pre-trained models. The potential for significant improvement by adopting models pre-trained on vast image datasets, such as ImageNet, has been well documented in the AST paper [1]. Preliminary experiments adapting the `vit_large_patch16_224` model (gathered from the PyTorch Image Models [12]) with a slow learning rate have already shown encouraging results surpassing those reported in this study, albeit without the comprehensive validation of a 10-fold cross-validation due to time constraints.

6.1. Exploring Pre-trained Models and Hybrid Architectures

Employing pre-trained Vision Transformer (ViT) models as a foundation for audio classification tasks utilizes a substantial base of learned features and representations, which can significantly enhance the performance of Transformer Neural Network (TNN) models. This approach is currently under investigation in our ongoing tests and indicates a viable method for repurposing the extensive array of visual knowledge contained within these models for auditory applications. Nevertheless, the comprehensive effectiveness and applicability of this strategy require further validation in subsequent studies.

Additionally, investigating hybrid models that integrate the precision of convolutional layers with the advanced attention mechanisms of Transformers presents a promising avenue for enhancing audio classification. This interest is sparked by research such as "Conformer: Convolution-augmented Transformer for Speech Recognition" [13], which demonstrates the synergy of combining CNNs' detailed feature extraction with the contextual awareness provided by Transformers. This combination could lead to models offering exceptional accuracy and efficiency in environmental sound classification.

The comparative data presented in the table above clearly delineates the performance differences between CNN and TNN models. It underscores the effectiveness of our customized approach for CNNs and identifies potential areas for improvement within the TNN framework.

7. CONCLUSION AND FUTURE WORK

This study has advanced the exploration of Convolutional Neural Networks (CNN) and Transformer Neural Networks (TNN) for urban sound classification. Despite the constraints of limited computational resources and time, our research highlights the potential of the Audio Spectrogram Transformer (AST) model to address the intricacies of sound classification effectively. The AST model demonstrated notable adaptability and performance, even with a restricted dataset. Nonetheless, the limited scope of our training data and the lack of an extensive validation phase point to the preliminary nature of our results.

The outcomes achieved within these constraints reveal the potential of TNN models to process and classify complex urban soundscapes efficiently. However, the limited validation of our model underscores the need for further research. Future work should focus on enlarging the dataset to improve the training robustness of these models. There is also a significant opportunity to refine AST model configurations and explore the use of pre-trained models and hybrid architectures. Such efforts could lead to improvements in efficiency and accuracy in urban sound classification.

REFERENCES

1. Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021.
2. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.
3. Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.
4. Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. Fsd50k, Oct 2020.
5. J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *22nd ACM International Conference on Multimedia (ACM-MM'14)*, pages 1041–1044, Orlando, FL, USA, Nov. 2014.
6. Karol J. Piczak. Environmental sound classification with convolutional neural networks. *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2015.
7. Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. Cnn architectures for large-scale audio classification. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2016.
8. Naoya Takahashi, Michael Gygli, Beat Pfister, and Luc Van Gool. Deep Convolutional Neural Networks and Data Augmentation for Acoustic Event Recognition. In *Proc. Interspeech 2016*, pages 2982–2986, 2016.
9. Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283, 2017.
10. Metod Zupančič. Implementacija modela za klasificiranje urbanih zvokov, 2021.
11. Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.
12. Ross Wightman. Pytorch image models, 2019.
13. Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. pages 5036–5040, 10 2020.