CentraleSupélec | université PARIS-SACLAY

Detecting forest fires with sound

Filière Métiers de la Recherche – Sujet 17

Group members :
Adil EL YAALAOUI
Aya KHAZRI
Joel CARLES-GONZALEZ
Gabrielle CAILLAUD
Ibrahim Al Khalil RIDENE

Supervisor :
Frédéric MAGOULÈS

September 17, 2024

Table des matières

1 Introduction

Forest fires have always been a naturally occurring event, an integral part of the balanced life of certain ecosystems whose biodiversity relies on the occurrence of such events. Forest fires have, however, become the subject of growing concern only in recent years, as the frequency of catastrophic instances in western countries has seemingly risen to new heights. Some precise statistics and corresponding analyses are available on the European Forest Fire Information System `https://forest-fire.emergency.copernicus.eu`. It is widely understood that this rise is due to the growing consequences of climate change. The predictions therefore indicate that the forest fire hazard will continue to grow in the foreseeable future. All countries with dry-climate areas particularly prone to wildfires must anticipate and implement prevention measures.

The early detection of forest fires is a key challenge in this fight. Among the techniques that have been explored to tackle this challenge, we notably find visual, sound or thermal sensors. The main advantages of using sound detection are that it isn't subject to visibility issues and the technology is cheaper. The main difficulty resides in the algorithm that must be put in place to recognize fire sounds in the landscape of environmental sounds that exist in the forest.

The aim of this project was to implement an AI algorithm capable of detecting forest fire sounds through the classification of environmental sounds. With the guidance of our tutor Mr Magoulès and with the help of the same work made by previous students, we explored an innovative solution to the problem.

2 Literature review and related work

We analyzed several relevant articles on the detection of forest fires and on the classification of audio using machine learning. The articles [1, 2, 3, 4], develop solutions specifically applied to the recognition of forest fires, whereas the other articles explore subjects related to machine learning related problems. A summary of each article, along with a short synthesis of main leads resulting from the aggregation of all articles, is provided in our Bibliography review document.

The key takeaways from the literature regarding our topic are summarized hereafter.

2.1 Gong et al. 2021

The Audio Spectrogram Transformer model (AST) was proposed by Yuan Gong, Yu-An Chung, James Glass in 2021 [5]. It is is the first convolution-free, purely attention-based model for audio classification. It builds upon the transformer architecture, originally developed for natural language processing, and applies it directly to audio spectrograms. It applies a Vision Transformer to audio, by turning audio into an image (spectrogram).

2.2 Alternative approaches

2.2.1 Bardou et al. 2018

The authors of this article compares traditional machine learning approaches to Convolutional Neural Networks (CNNs) for lung sound classification [6]. The authors employed handcrafted features such as Mel-frequency cepstral coefficients (MFCCs) and Local Binary Patterns (LBPs) extracted from audio spectrograms. These features were then fed into classifiers like Support Vector Machines (SVM), k-nearest neighbors (KNN), and Gaussian Mixture Models (GMM). These handcrafted models yielded decent performance; however, their reliance on manual feature extraction introduced limitations. To overcome this, the authors experimented with CNNs that automatically learned relevant features from raw input data, specifically spectrograms of lung sounds. The CNN approach outperformed the handcrafted feature-based models.

2.2.2 Z. Zhang et al. 2019

(Z. Zhang et al. 2019)[7] implements a Convolutional Recurrent Neural Network (CRNN) to classify several environmental sounds. TThe CRNN architecture combined convolutional layers for feature extraction from Log-GammaTone spectrograms with a recurrent layer for capturing temporal dynamics. The authors reported impressive results, achieving an accuracy of 93.7%} on the ESC-10 dataset and 86%} on the ESC-50 dataset using 5-fold cross-validation.

3 Dataset

We used the ESC-50 dataset which is a labeled collection of 2000 environmental audio recordings suitable for benchmarking methods of environmental sound classification. The dataset consists of 5-second-long recordings organized into 50 semantical classes.

To improve fire detection, we applied a data augmentation technique. We created a new dataset by combining the "fire" audio recordings with other environmental sounds from the ESC-50 dataset (like birds chirping, rain, or wind) as background noise. In this new dataset, "fire" remains the primary audio signal, while the other classes act as secondary background sounds, simulating real-world scenarios where fire sounds might be mixed with other noises.

We found two other sources of forest fire sounds with open access, that we didn't use for the training but that we kept on the side just in case. They can be found at the following internet addresses :
- `https://freesound.org/people/tim.kahn/sounds/253770/`
- `https://www.nps.gov/yell/learn/photosmultimedia/sounds-fire.htm`

4 Methodology

Our proposed sound classification pipeline follows the general structure of pre-processing, features extraction and classification as illustrated on figure 2.

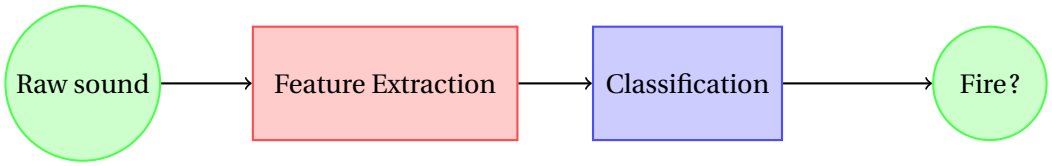

FIGURE 2 – General form of sound detection

4.1 AST : Audio Spectrogram Transformer

This section outlines the architecture and implementation details of the Audio Spectrogram Transformer (AST), which effectively captures temporal relationships and patterns in audio data using a self-attention mechanism.

4.1.1 Overview of AST Architecture

The Audio Spectrogram Transformer (AST) employs a transformer-based approach to process audio spectrograms, focusing entirely on attention mechanisms to model audio features. Below are the key steps involved in the architecture :

1. **Data Preprocessing and Feature Extraction :** The raw audio signal undergoes preprocessing, where it is transformed into a spectrogram representation, which serves as the input to the model.
2. **Patch Generation :** The spectrogram is divided into small, non-overlapping patches, similar to how image patches are processed in vision transformers.
3. **Patch Embedding :** Each spectrogram patch is flattened and embedded into a 1D vector, capturing essential frequency and time-based features. Positional embeddings are then added to maintain the sequential order of the patches.
4. **Transformer Encoder :** The embedded patches, now containing both feature and positional information, are fed into a Transformer encoder. This encoder applies a series of self-attention layers to capture long-range dependencies and relationships across the audio sequence.
5. **Classification :** After processing through the Transformer layers, the output representations are aggregated, and the final classification is performed to predict the target labels.

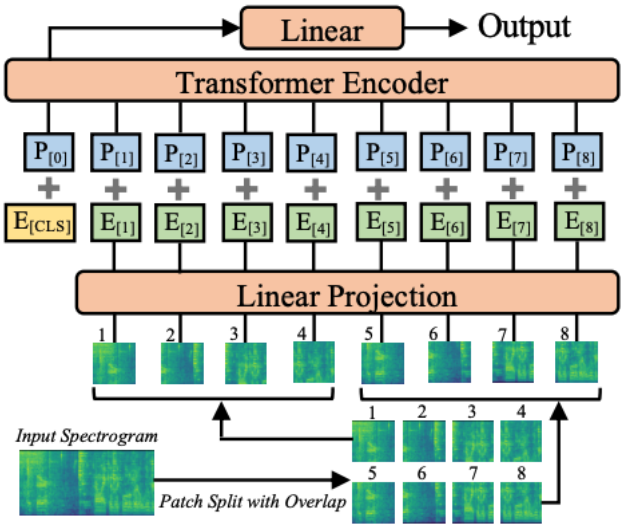

FIGURE 3 – Audio Spectrogram Transformer architecture

4.1.2 Application of AST

In our case, we use the AST model as an input feature to a multi layer perceptron (MLP). The full architecture is available on the schema 4.

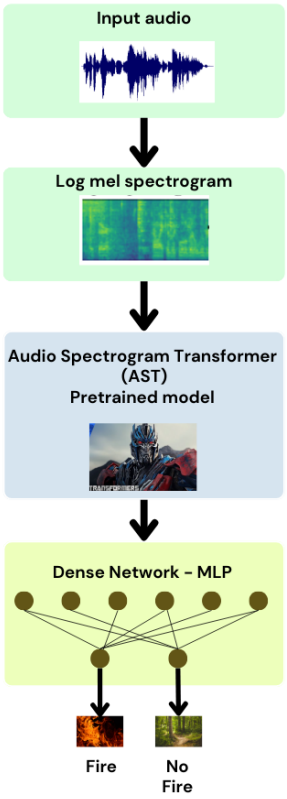

FIGURE 4 – Audio Spectrogram Transformer architecture

We have two approaches with the AST :

1. **Using a pretrained AST** An already trained AST is available with an open source license via Hugging face hub. This model is easily downloaded in Python via the transformers library. It was trained on a large dataset called Audioset, and therefore has learnt to analze audio and encode them in a meaningful way. Therefore, in the first approach, we use the output of the pretrained AST encoder, and pass that through a MLP that we train ourselves. During training, we freeze AST's weights and train only the MLP. This approach is called transfer learning.
2. **Training the full model from scratch** In this second approach, we don't use a pretrained AST model; we train the whole AST + MLP model.

5 Results

We ran several trainings of the different models and we chose to illustrate our results with 3 graphs. We compared pre-trained models and models we trained from scratch, as well as binary classification models and 50-class classification models.

Starting with binary classification, we compared three models, building on the work of the previous group : the Zhang model using a spectrogram as input, the Zhang model using a mel-spectrogram, and

the Bardou model. We trained these models ourselves, alongside our AST model, using the ESC-50 dataset with our data augmentation procedure.

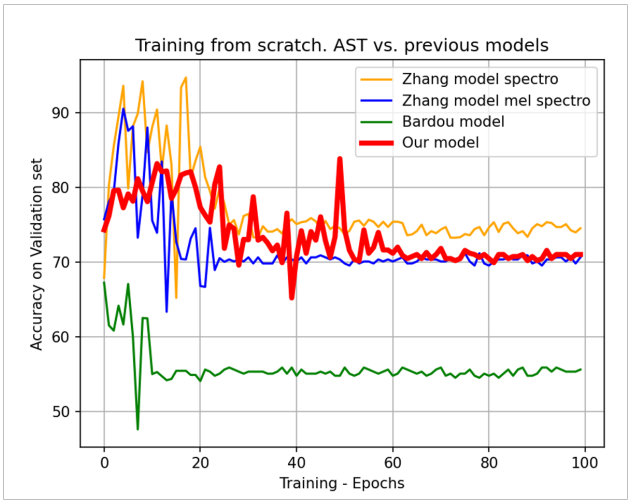

FIGURE 5 – Binary classifications with models trained from scratch

In figure 5 we can see that the model that reached the highest level of accuracy (about 95%) was the Zhang CRNN model with spectrogram, overperforming our proposed AST model by about 10%.

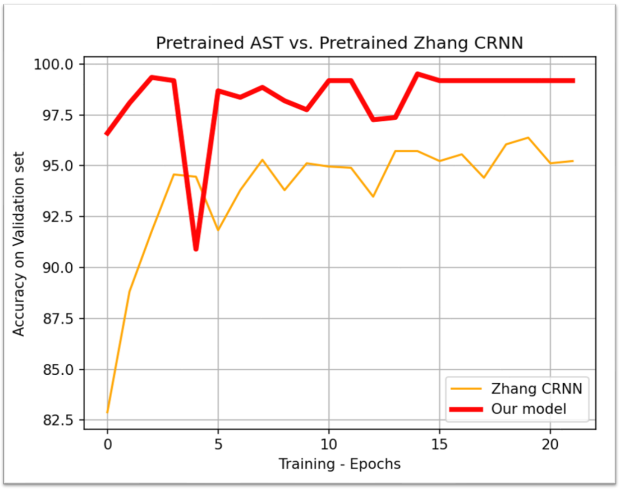

FIGURE 6 – Pretrained models for Binary classification

In figure 6 we observe that our proposed model, the AST model, reached a top accuracy of 99,51%. In comparison, the CRNN model proposed by Zhang reached a top accuracy of 96,50%, which is a similar score to the one this same algorithm obtained when it was trained from scratch.

On the other side, we compared our models on a multiple-classes scenario, where all 50 classes from the ESC50 dataset were kept. Starting by models from scratch, We can see in figure 7, the Bardou model has the best validation accuracy score of 27%. The AST model has low performance here.

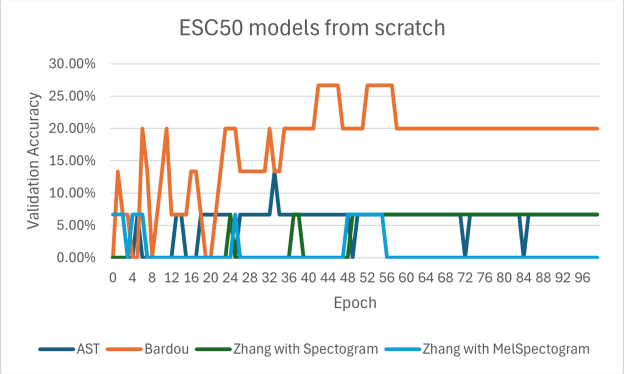

FIGURE 7 – ESC50-models trained from scratch

On the other hand, comparing pretrained models results are shown in figure 8. Again we observe that the AST model overperforms the Zhang CRNN model, by 20% percent.

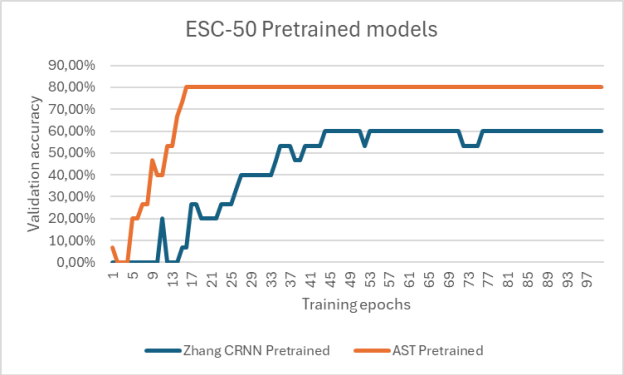

FIGURE 8 – ESC50-pretrained models

6 Discussion

The results of comparison between different models for binary classification, as well as 'FIFTY' classification match our expectations.

In the binary classification problem, we implemented two approaches. The first one deals with models trained from scratch (weights are initialized randomly at the start of the training process). The results reveal that Zhang model outperforms AST model, which can be explained by the fact that the architecture of the AST is more complex (based on transformers), thus needs more data to reach outstanding performance and it performs poorly with small datasets. The second approach is based on pre-trained Zhang and AST models, both are pre-trained before on the same AudioSet data. In this approach, we are performing finetuning step on our binary classification dataset. As expected, the results shows that the AST outperforms Zhang model with an outstanding validation accuracy of 99.51%. Thanks to the power of transfer learning, the AST model reached the highest performance and beated Zhang by far.

Concerning the ESC50 classification (50 class classification), we also implemented both approaches.

The first one (based on models trained from scratch) reveals that Bardou model is the best one with 27% validation accuracy. We can also explain this result based on models architecture : Bardou architecture is less complex (based only on CNNs) compared to Zhang architecure (CNNS and RNNs) and AST (Transformers). Thus, with small data as we have, Bardou with give better results. On the other approach based on pre-trained models, the results match our expectation. Tha AST model outperforms Zhang model with a validation accuracy of 80%. We can explain such a result with the same reason that transformers reach good performance when trained on large datasets.

Finally, the results we obtained with our implementation of AST are promising. Our results coincide with our expectations, as the AST algorithm has been proven to out perform other algorithms before.

For future research we recommend exploring the idea of characterizing the fire with its intensity and direction. Such knowledge could prove to be extremely useful for the firefighters to have well adapted responses to each forest fire. The article cite 1 offers an insightful lead with their investigation of crown fire versus surface fire identification. The identification of the direction of the forest fire could be pursued with a tight enough network of audio sensors in the forest. If the data from multiple sensors could be processed in real time at the same time, the comparison of the different intensities in different locations could lead to the determination of the movement of the fire.

7 Conclusion

In this study, we used the Audio Spectrogram Transformer (AST) model and architecture to classify environmental sounds that contains fire. The dataset ESC50 is used for binary classification after being augmented for a balanced class representation. We then compare this model with other existing approaches, first by training the AST on the augmented ESC50. This gives us an accuracy of 83.22% for classification which is higher than the "Bardou" model but remains lower than the "Zhang" model with both spectrogram and mel-spectrogram that go up to 94.21% . We then use the pre-trained AST model and fine tune it with the augmented dataset. This model reaches an accuracy of 99.51% and it is the best among the other pre-trained model for environmental sound classification.

8 Acknowledgements

We thank the group from last year's project for their work, which greatly helped us to jumpstart our own project. We warmly thank our supervisor and tutor Mr Magoulès for proposing this interesting project and for guiding us through it.

Références

[1] Shuo ZHANG et al. "Wildfire Detection Using Sound Spectrum Analysis Based on the Internet of Things". In : *Sensors* 19.23 (2019). ISSN : 1424-8220. DOI : 10.3390/s19235093. URL : https://www.mdpi.com/1424-8220/19/23/5093.

[2] Hung-Tien HUANG, Austin R. J. DOWNEY et Jason D. BAKOS. "Audio-Based Wildfire Detection on Embedded Systems". In : *Electronics* 11.9 (2022). ISSN : 2079-9292. DOI : 10.3390/electronics11091417. URL : https://www.mdpi.com/2079-9292/11/9/1417.

[3] Mounir GRARI et al. "FOREST FIRE DETECTION AND MONITORING THROUGH ENVIRONMENT SOUND SPECTRUM USING DEEP LEARNING". In : (oct. 2023).

[4] Giacomo PERUZZI, Alessandro POZZEBON et Mattia VAN DER MEER. “Fight Fire with Fire : Detecting Forest Fires with Embedded Machine Learning Models Dealing with Audio and Images on Low Power IoT Devices”. In : *Sensors* 23.2 (2023). ISSN : 1424-8220. DOI : 10 . 3390 / s23020783. URL : https://www.mdpi.com/1424-8220/23/2/783.

[5] Yuan GONG, Yu-An CHUNG et James GLASS. *AST : Audio Spectrogram Transformer*. 8 juill. 2021. arXiv : 2104.01778[cs]. URL : http://arxiv.org/abs/2104.01778 (visité le 06/09/2024).

[6] Dalal BARDOU, Kun ZHANG et Sayed Mohammad AHMAD. “Lung sounds classification using convolutional neural networks”. In : *Artificial Intelligence in Medicine* 88 (2018), p. 58-69. ISSN : 0933-3657. DOI : https : / / doi . org / 10 . 1016 / j . artmed . 2018 . 04 . 008. URL : https : / / www . sciencedirect.com/science/article/pii/S0933365717302051.

[7] Guan-Bo WANG et Wei-Qiang ZHANG. “An RNN and CRNN Based Approach to Robust Voice Activity Detection”. In : *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 2019, p. 1347-1350. DOI : 10.1109/APSIPAASC47483.2019.9023320.