How accurately can a machine learning algorithm make the distinction between city and nature sounds?

Dominic Armand Stamatoiu
u2051167
Group 22

Abstract

The goal of this study is to see how well a machine learning system can distinguish between city and nature sounds. It receives a large amount of audio data, which is evaluated and processed using three different machine learning methods. The accuracy of each algorithm is measured to evaluate which model is better at identifying the sounds.

Keywords: sound recognition; machine learning algorithms; accuracy

Introduction

During the last 30 years, mankind has achieved impressive results in the domain of machine learning by developing new technologies. One such achievement is the birth of image recognition which aids us today in various fields such as social services, engineering, communications and, many others. By using deep learning, which is a machine learning algorithm, image recognition has advanced a lot over the years. According to Fujiyoshi et al. (2019), image recognition is now being used in creating vehicles capable of autonomous driving, which is a very complex task.

On the other hand, while focusing on image recognition, other components of the human senses were disregarded, especially sound recognition. The machine learning algorithms that are being trained and that have to learn from sound data of musical instruments and human speech have issues with sound that comes from the environment, such as city and nature sounds which will be covered in this research study (Cowling & Sitte, 2003). According to Bountourakis et al. (2015), the best algorithms, which have the highest accuracy regarding the environmental sound recognition are: Support Vector Machines (SVM), K-Nearest Neighbour (KNN), and Artificial Neural Networks (ANN). The ANN classifier continues to outperform the K-NN classifier (Silva et al., 2019), while the SVM approach is considered to be the most accurate (Nolasco & Benetos, 2018).

Methods

The data set used for this research is “freefield 1010” (Stowell & Plumbley, 2014). It consists of 17807 sound recordings of a length of approximately 10 seconds. There are seven categories of sounds, however, for this study, the categories referring to the city (562 recordings) and the nature (905 recordings) sounds will suffice. In terms of methods, we compared the degree of accuracy for the binary classification composed of three machine learning algorithms: K-Nearest Neighbors model, Multilayer Perceptron (MLP) model and Convolutional Neural Network (CNN) model. We used the Python programming language and found a package called ”Librosa” which helped us retrieve the audio data and process it.

Results

This section’s results are divided into three sections, each detailing a model’s outcome. It should be mentioned that the data was split into train and test data in a 4:1 ratio using 10-fold cross-validation for each model. A validation set was included in both the Sequential model and the MLP Classifier, on which our team did hyper-parameter tuning. The displayed results represent the mean extracted from all the cross-validation folds.

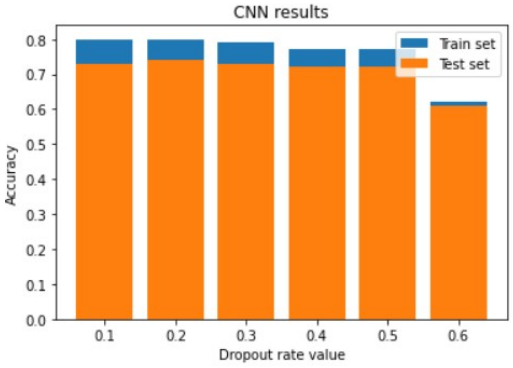

Figure 1: Accuracy results for CNN model

Changing the optimizer and activation function for our Convolutional Neural Network (CNN) had a negligible effect on prediction accuracy, similar to the MLP. The dropout rate, which is responsible for randomly deactivating neurons and their connections in order to prevent the network from depending on only a few nodes, was the sole parameter we modified during the iterations. The dropout rates on all hidden layers at the same time ranged from 0.1 to 0.6, increasing by 0.1 per iteration, while the input layers remained at 0.8 throughout. We can see that lower dropout rates resulted in the highest accuracies on both the test set and training set, with 0.2 being the top score for the test set, while larger dropout rates result in lower accuracies (Fig. 1). We noticed a sharp decline in accuracy from 0.5 to 0.6, which is why we decided to stop iterating at that point.

As previously stated, the only passable argument for the K-Nearest Neighbors model was the number of neighbors itself. The worst test-set accuracy (69%) was shown by the model that received 1 neighbor as a parameter while displaying the

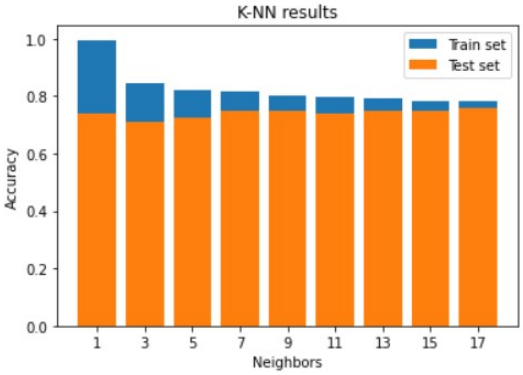

Figure 2: Accuracy results for K-NN model

best train set accuracy(99%). This is to be expected since the model tends to overfit the lower the number of neighbors is. As shown in (Fig. 2), the test set accuracy gradually increased until it peaked at 76% with 17 neighbors, while the training set began to decline while remaining at around an 80% mark. We terminated parameter testing at 17 neighbors because the test set accuracy did not improve after this peak.

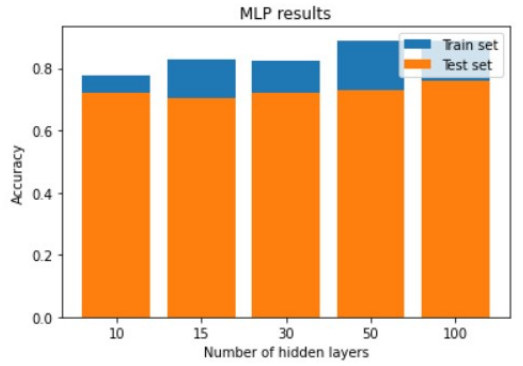

Figure 3: Accuracy results for MLP model

We decided not to present two of the three parameters in the Multilayer Perceptron model (optimizer and activation function) since they had little to no influence on the findings. The number of hidden layers passed while constructing the model was the only notable difference. Test set accuracy fluctuates between 70 and 76 percent (Fig. 3), whereas train set accuracy grew steadily from 78 percent to 89 percent. The 100-layer model had the best accuracy on the test set, and it also had the best accuracy for the train test. This is to be expected because this model underfits as the number of hidden layers is decreased. A higher number of hidden layers did not seem to affect the accuracy on the test set, therefore we stopped at 100.

Discussion

Our research yielded results ranging from 60% to 90%, which is a standard range of accuracy demonstrated by machine learning algorithms in most studies. We had to choose favorable data and determine what features we should utilize to enhance accuracy before we could focus on these algorithms, which took several minutes. The data download took around 5 minutes, and we ended up with numerous sound files that we had to analyze in order to extract the numerical data. The Librosa package assisted us in changing the data and provided arrays that could be used to train the machine learning models by reshaping them. In the end, all three models performed similarly in terms of accuracy, although K-NN and MLP outperformed the others.

Conclusion

Finally, the outcomes of the research study we gave were satisfactory, but they may be improved with additional future research, learning more about the relevant issue, and attempting to optimize our knowledge of sound recognition in relation to the surroundings. The study presented the training and testing of three machine learning algorithms using sound data collected from the city and nature, with the results indicating which model is the superior sound recognition solution. Because our research expertise is limited, this study may not be comparable to prior ones. We believe that this study will serve as a foundation for further research into sound recognition.

Acknowledgements

This research and paper would not have been possible without the help of my fellow group members, with who I managed on the last day to finish the paper. They did a really good job at explaining different concepts regarding artificial intelligence that I did not understood until know. I also want to thank my friends and family for their continuous support.

References

Bountourakis, V., Vrysis, L., & Papanikolaou, G. (2015). A computational model of language acquisition in the two-year old. *Proceedings of the Audio Mostly 2015 on Interaction with Sound*, *5*, 1–7.

Cowling, M., & Sitte, R. (2003). Comparison of techniques for environmental soundrecognition. *Pattern recognition letters*, *24*, 2895–2907.

Fujiyoshi, H., Hirakawa, T., & Yamashita, T. (2019). Deep learning-based image recognition for autonomous driving. *Cognition and Brain Theory*, *43*, 244–252.

Nolasco, I., & Benetos, E. (2018). To bee or not to bee: Investigating machine learning approaches for beehive sound recognition. *arXiv preprint arXiv:1811.06016.*.

Silva, B. D., Happi, A. W., Braeken, A., & Touhafi, A. (2019). Evaluation of classical machine learning techniques towards urban sound recognition on embedded systems. *Applied Sciences*, *9*, 3885.

Stowell, D., & Plumbley, M. (2014). An open dataset for research on audio field recording archives: freefield1010. *Journal of the audio engineering society*.