1 Feature Selections

Even best minds in the world will perform poorly on test which they have studied the wrong material- neural networks are no different. In order to properly train a neural network, the model must be presented with an appropriate set of training input x and a complementary set of training labels y [1, 3, 4]. It becomes quickly apparent that the nature of information contained in the input object x, called *features*, is *extremely important* to the performance of the classifier. Consider if you were tasked to identify cats and dogs from images, but instead were presented with only the top-most row of pixels- The task would be nearly impossible because of an inappropriate or in complete set of information.

Tuomas Virtanen, machine learning and audio engineer writes in his book, "Computational Analysis of Sound Scene and Events" [16]:

> For recognition algorithms, the necessary property of the acoustic features is low variability among features extracted from examples assigned to the same class, and at the same time high variability allowing distinction between features extracted from examples assigned to different classes.

In constructing a neural network classifier, the development of appropriate features is of the utmost importance. To ensure the construction of a suitable model, we derive features based from three sources (i) a spectrogram matrix of the waveform, (ii) the time-space representation of the waveform, and (iii) the frequency space representation of the waveform. It is important to note that although this algorithm will classify sound waves to instruments, the model will never actually be presented with a waveform directly, instead it will rely on these features.

Once we produce an sufficient set of features, we concatenate them into a single object, $\hat{x}$, called the *feature-vector* [3]. In the training process, this object, along with the appropriate classification label, y is presented to the neural network for processing. This process of constructing a feature vector from any data set is vital and is used to represent the data set in a far more compact and non-redundant format [16, 7]

To ensure suitable performance of this sound wave classification neural network, a great deal of time has been devoted to the construction of the elements of the feature vector. These features are derived are principles of music, digital signal processing, previous work success, and most importantly physics. In the following sections, we outline the set of 24 features used in the classification process

1.0.1 Audio Preprocessing

1.1 Spectrogram Feature

The field of neural classification is well studied in the application of image-processing. Many large-scale, and introductory neural network projects find themselves under the um-

brella of image classification [1, 3, 8, 10]. As a result, model architectures for image related tasks are well-explored and have experimentally shown successful behavior. Following in those footsteps, it make senses to provide an image-like representation of a sound wave as a feature. We do this in the form of a spectrogram.

A spectrogram is a representation of the energy distribution of a sound wave as a function of both space and time. In a conventional spectrogram, the passing of time is shown along the x-axis, and the frequency spectrum is shown on the y-axis. Thus each point in the 2-Dimensional space is an energy at a given time and frequency. Examples spectrograms from the wave form data set are shown in Fig. (1.1).

Insert spectrograms here

Figure 1: Spectrogram representations of various waveforms

A spectrogram is produced by the method of *frame-blocking*, which is very prevalent in audio signal classification. Frame-blocking creates a set of analysis frames, each of which is N samples in length, and has a fixed overlap with the next adjacent frame. Each of the k frames then allows for a section of the signal in somewhat stationary state [7, **?**, 5, 16]. For this project, we have chosen to use frames of size $N = 4096$ with a 75% or 3072 sample overlap. At a sample rate of $f_s = 44100$ samples/second, each frame represents a slice of time that is about 0.1 seconds long.

After frame-blocking, we apply a *windowing function* to each frame. The analysis frame are concatenated into a $k \times N$ sized matrix, A, where each row is a frame. A standard *Hann Window* of N samples is generated and reshaped into a $N \times 1$ array,H. The n-th index in a Hann window with N samples is defined:

$$H[n] = \frac{1}{2}\left[1 - \cos\left(\frac{2\pi n}{N-1}\right)\right] \tag{1}$$

The window function is applied to each analysis frame by computing the matrix-product of the $k \times N$ block-frame matrix and the $N \times 1$ window function array. Math is wrong here! Check this, check the code too!

1.2 Time-Space Features

1.3 Frequency-Space Features

References

[1] Geron, Aurelien. *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly, 2017.

[2] Geron, Aurelien. *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2nd ed., O'Reilly, 2019.

[3] Goodfellow, Ian, et al.*Deep Learning*. MIT Press, 2017.

[4] James, Gareth, et al. *An Introduction to Statistical Learning with Applications in R*. Springer, 2017.

[5] Khan, M. Kashif Saeed, and Wasfi G. Al-Khatib. "Machine-Learning Based Classification of Speech and Music." Multimedia Systems, vol. 12, no. 1, 2006, pp. 55–67., doi:10.1007/s00530-006-0034-0.

[6] Levine, Daniel S. *Introduction to Neural and Cognitive Modeling*. 3rd ed., Routledge, 2019.

[7] Liu, Zhu, et al. "Audio Feature Extraction and Analysis for Scene Segmentation and Classification." Journal of VLSI Signal Processing, vol. 20, 1998, pp. 61–79.

[8] Loy, James , *Neural Network Projects with Python*. Packt Publishing, 2019

[9] McCulloch, Warren S., and Walter Pitts. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *The Bulletin of Mathematical Biophysics*, vol. 5, no. 4, 1943, pp. 115–133.

[10] Mierswa, Ingo, and Katharina Morik. "Automatic Feature Extraction for Classifying Audio Data." *Machine Learning*, vol. 58, no. 2-3, 2005, pp. 127–149., doi:10.1007/s10994-005-5824-7.

[11] Mitchell, Tom Michael. Machine Learning. 1st ed., McGraw-Hill, 1997.

[12] Olson, Harry E. *Music, Physics and Engineering*. 2nd ed., Dover Publications, 1967.

[13] Peatross, Justin, and Michael Ware. *Physics of Light and Optics*. Brigham Young University, Department of Physics, 2015.

[14] Petrik, Marek. "Introduction to Deep Learning." Machine Learning. 20 April. 2020, Durham, New Hampshire.

[15] Short, K. and Garcia R.A. 2006. "Signal Analysis Using the Complex Spectral Phase Evolution (CSPE) Method." AES: *Audio Engineering Society Convention Paper*.

[16] Virtanen, Tuomas, et al. *Computational Analysis of Sound Scenes and Events*. Springer, 2018.

[17] White, Harvey Elliott, and Donald H. White. *Physics and Music: the Science of Musical Sound*. Dover Publications, Inc., 2019.

[18] Zhang, Tong, and C.-C. Jay Kuo. "Content-Based Classification and Retrieval of Audio." *Advanced Signal Processing Algorithms, Architectures, and Implementations VIII*, 2 Oct. 1998, pp. 432–443., doi:10.1117/12.325703.