Denoising Autoencoders

Chibuzor John Amadi[502623] & Kristian Perriu[505571]

[1] University of Pavia, Pavia
[2] University of Milano-Biccoca, Milan
[3] University of Milano-Statale, Milan

1 Introduction

With regards to audio data, Denoising Autoencoders is a type of neural network that is built to remove noise from audio signals. The basic structure of an autoencoder includes an encoder and a decoder. The encoder compresses the input audio signals into a lower dimensional representation, a latent space. This latent space captures the essential features of the input and is passed to the decoder. The decoder reconstructs the signal from the compressed representation. Denoising autoencoders are trained to remove noise from the input data, the neural network here is trained with pairs of noisy and clean audio signals, and its objective is to minimize the difference between the clean audio signals and the reconstructed output from the noisy input. The loss function used is the Mean Squared Error (MSE) measuring the difference between the denoised output and the clean target. In this paper we will study the denoising autoencoders with a UNet model architecture to understand the concepts of denoising auntoencoders.

2 Model Architecture

U-Net Architecture The U-Net[1] architecture is known for its symmetric encoder-decoder design and the use of skip connections, which help in preserving spatial information and avoid the vanishing gradient problem. To understand the model we will break down the different components that make up the model. ***Encoder*** The encoder part of the U-Net architecture is responsible for compressing the input noisy audio signal into a lower-dimensional representation. The convolutional layers helps in extracting features from the audio signal. ReLu activation function is applied after each convolutional layer, It introduces non-linearity into the model, enabling it to learn complex patterns and relationships within the data. This process ensures that the essential features of the noisy input are captured in a compact representation.

Decoder The decoder part of the U-Net architecture is responsible for reconstructing the clean audio signal from the compressed representation obtained from the encoder. The decoder part of the U-Net architecture is responsible for reconstructing the clean audio signal from the compressed representation obtained from the encoder. Similar to the encoder, ReLU activations are applied

after each transposed convolutional layer. They introduce non-linearity, enabling the decoder to learn and reconstruct complex patterns in the clean audio signal.

Skip Connections Skip connections[2] are a key feature of the U-Net architecture, providing direct connections between corresponding layers in the encoder and decoder. Skip connections bypass certain layers in the network by directly connecting the output of an encoder layer to the input of the corresponding decoder layer. Skip connections help retain detailed spatial information, which is crucial for accurate reconstruction of the clean audio signal. By providing alternate paths for the gradient to flow during back-propagation, skip connections mitigate the vanishing gradient problem, ensuring more effective and stable training.

3 Evaluation

Our Dataset is gotten from *datashare.uk* [3] and it is divided into a train, test and validation set. For this project we will make use of the Mel Spectogram [4]. MEL scale is realated to human-audible frequencies and to convert audio signals to Mel spectrograms we make use of the Fast Fourier Transform (FFT)[5] . Mel spectrograms are a representation of the short-time Fourier transform (STFT) with a Mel scale frequency axis. To initiate this process the raw audio signal is divided into overlapping windows. This helps in capturing the temporal characteristics of the audio. Each window is multiplied by a window function to reduce spectral leakage. The windows overlap to ensure smooth transitions and continuity in the frequency domain. The amount of overlap is determined by the hop length. Apply FFT to each windowed segment to convert the time-domain signal into the frequency domain, resulting in a series of FFT outputs. The FFT outputs are mapped onto the Mel scale, which is a perceptually motivated scale of pitches. Combine the Mel scale frequency components to form the Mel spectrogram, representing the audio signal in both time and frequency domains. Figure 1 illustrates the steps to carry out FFT.

The training of the denoising autoencoder is carried out with the Mean Squared Error (MSE) loss function with the formula 1 1. We optimize the MSE loss function to minimize the difference between the denoised output and the clean target audio. Fine-tune hyperparameters such as learning rate, batch size, and the number of layers/filters in the network using the validation set. The grid search and random search methods to explore different hyperparameter combinations. After training, the model is evaluated on the test set to measure its generalization capability. Compute the MSE loss on the test set to quantify the reconstruction error. The Mel spectrograms then visualized for both the noisy input, clean target, and denoised output to qualitatively assess the denoising performance.

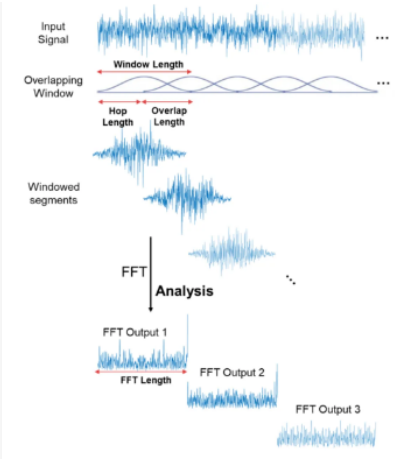

Fig. 1. Fast Fourier Transform Process

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

Fig. 2. Enter Caption

4 Results

The Mel spectogram is then analysed along with the audio samples that are audibly in order to evaluate the quality of the denoised output. we visualize and compare the Mel spectrograms of the noisy input, clean target, and denoised output. we also assess how closely the denoised output matches the clean target, indicating the effectiveness of noise removal.

5 Conclusion

The evaluation process demonstrates the effectiveness of the denoising autoencoder in removing noise from audio signals. By combining visual Mel spectrogram analysis, audio playback, and quantitative metrics, we gain a comprehensive understanding of the model's performance and its potential for real-world applications. The U-Net architecture's use of skip connections and instance normalization proves to be highly effective in preserving high-resolution features and stabilizing the training process. Future work could explore the integration of more advanced loss functions or hybrid architectures to further enhance denoising capabilities and generalize to a wider range of noise types.

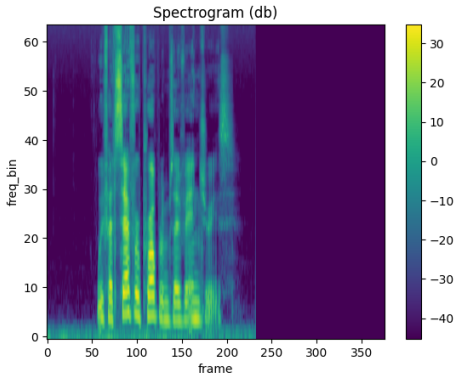

Fig. 3. clean spectogram

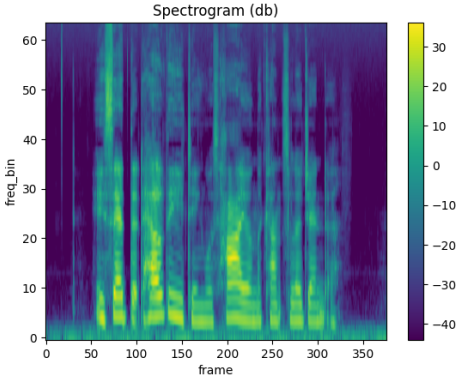

Fig. 4. noisy spectogram

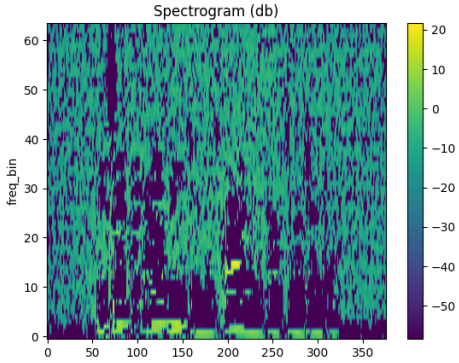

Fig. 5. cleaned spectogram

6 References

1 Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In International Conference on Medical image computing and computer-assisted intervention, pp. 234-241. Springer, Cham, 2015.

2 He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778. 2016.

3 Veaux, Christophe, Junichi Yamagishi, and Kirsten MacDonald. "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)." https://datashare.ed.ac.uk/handle/10283/3443 (2017).

4 Stevens, Stanley S., John Volkmann, and Edwin B. Newman. "A scale for the measurement of the psychological magnitude pitch." The Journal of the Acoustical Society of America 8, no. 3 (1937): 185-190.

5 Oppenheim, Alan V., and Ronald W. Schafer. "Discrete-time signal processing." Pearson Higher Education, 2009.