IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 6.078
(Volume 6, Issue 3)
Available online at: www.ijariit.com

Audio super resolution using Neural Networks

Gorli Harshini
gorliharshini98@gmail.com
Andhra University College of Engineering for Women, Visakhapatnam, Andhra Pradesh

Gayatri Yasaswini Pappala
gayatripy19@gmail.com
Andhra University College of Engineering for Women, Visakhapatnam, Andhra Pradesh

Gorle Manasa
manasagorle17@gmail.com
Andhra University College of Engineering for Women, Visakhapatnam, Andhra Pradesh

Gollamandala Sam Shalini
samshalini1998@gmail.com
Andhra University College of Engineering for Women, Visakhapatnam, Andhra Pradesh

M. Sion Kumari
msmondru@gmail.com
Andhra University College of Engineering for Women, Visakhapatnam, Andhra Pradesh

ABSTRACT

In this era of technological advancement, people tend to a demand for high quality videos, audios and images. So deep convolutional neural networks play an important role in learning low resolution data and obtaining high resolution data by performing interpolation method. This is similar to image super resolution. Here we introduce a signal processing technique to convert the low resolution into a high-resolution data with the help of signal processing methods such as up-sampling and down-sampling using subpixel convolution through Bottleneck architecture. Our model tests the missing values with in the low-resolution signals and forms high resolution signals. This technique is applied to a telephony, upscaling, text to speech conversion and also for investigations in many departments. We test the effect of convolutions used in the signal processing and measures the compatibility and scalability for generative model of audio.

Keywords— *Deep Convolutional Neural Networks, Interpolation Method, subpixel Convolution, Bottleneck Architecture..*

1. INTRODUCION

Convolutional neural networks have been one of the most influential innovations in the field of computer vision and it is most likely used in audio synthesis because audio samples are analogous to image pixels. Our model based on machine learning takes a step towards alleviating the difficulty of producing high quality audio signals in affordable cost by proposing a technique from input containing only a small fraction (15-50%) of the original signal's information. Our audio processing technique also raises basic idea on pertaining to time series and generative modelling The flexibility of the model depends on the computational power because we train thousands of data and test the complexity of the model. The flexibility and computational power is most important because it's functions are made by removing the leading and tailing phases time domain and frequency domain of raw data using neural networks.

2. EXISTING SYSTEM

In the existing system, the time series modelling is modelled with auto-regression models which has low rate of accuracy and even the bandwidth extension includes linear predictive models and gaussian mixture models which has low rate of scaling. Existing telephony, commercial efforts are underway to transmit voice at higher rates (typically 16 Khz) in specific handsets, This is also one of the drawback in existing system.

3. PROPOSED SYSTEM

Generally, the time series signals are not precise and altered well, our model generalizes conditional modelling ideas used in computer vision for tasks such as image super-resolution or colorization. We identify a huge class of conditional time series modelling problems that arise in signal processing, biomedicine, and other fields and that are characterized by a natural alignment among source/target series pairs and differences that are well-represented by local transformations. We propose a general architecture for such problems and show that it works well in different domains. Our work proposes the first convolutional architecture, which we find to scale better with dataset size and outperform recent, specialized methods. Moreover, while existing techniques involve many hand-crafted features. Our approach is fully domain-agnostic. our work demonstrates the effectiveness of feedforward convolutional architectures on an audio generation task.

4. WORKING PROCEDURE

4.1 Processing of audio signal

Audio signal can be represented in time domain as s(t) ie; s(t) = [0 T] in real world and T is time or duration of the audio raw data in (seconds) the scalar data s(t) is converted to vector x(t) which is R*T of time by the sampling rate R that is $\{\frac{1}{R},\frac{2}{R},\frac{3}{R},\frac{4}{R},\frac{5}{R},\ldots\frac{RT}{R}\} \rightarrow$ R. Sampling rate f_s is the number of samples obtained per second by an average or the rate at which data is sampled f_s = Samples/ second(Which ranges from Hzs to khzs). By giving various sample rates to the low resolution data x we predict the correct R of the resolution f x and increase the resolution of the audio from a fraction of $\frac{1}{R},\frac{2}{R},\frac{3}{R},\frac{4}{R},\frac{5}{R},\ldots\frac{RT}{R}$ so by the predicting the x value we can get the high and super resolution of the audio data. The low-resolution data can be of about 4Khz

4.2 Method

The method used here is up-sampling and down-sampling. we consider a low resolution signal of input-data with a sampling rate R1 and y be a high resolution signal with sampling rate R2, R2 >R1 where x = $\{x_1$/R1 ,….x_{R1T1}/R1 } and y = { y_1/ R2,…. y_{yR2T2}/R2} and the definition for the x,y can be given as y = f_θ(x) where θ is a parameter and f_θ. We obtain by running the model with convolutional neural network on large set of data.

4.2.1 Up-sampling: Up-sampling is the process of increasing the sampling rate of a discrete time audio file using discrete-time interpolation. This can be said as a new virtual sample which is placed between the two known adjacent samples and calculates the amplitude present between two sampling rates which gives a better sampling rate. This process is also called as bandwidth extension as the bandwidth of the discrete-time series is extended. our neural network contains (2-3 densely connected convolutional layers) where the hidden layers are present. The included learning bases algorithm is simple and easy to handle which can result in more accurate convolutional model. The band-width of the decimation data are at a low bit rate of 8bits/secs. In the reconstruction of high bandwidth, we use low pass filters. This interpolation takes about (L-1) Zero valued samples in between input samples and tries to minimize any disturbances. This sampling rate increases from fs -Lfs. The Nyquist rate helps to predict whether a sample rate of each component is within a certain range and takes the highest rate in that certain range and tries to fit every signals sample rate to that particular highest rate

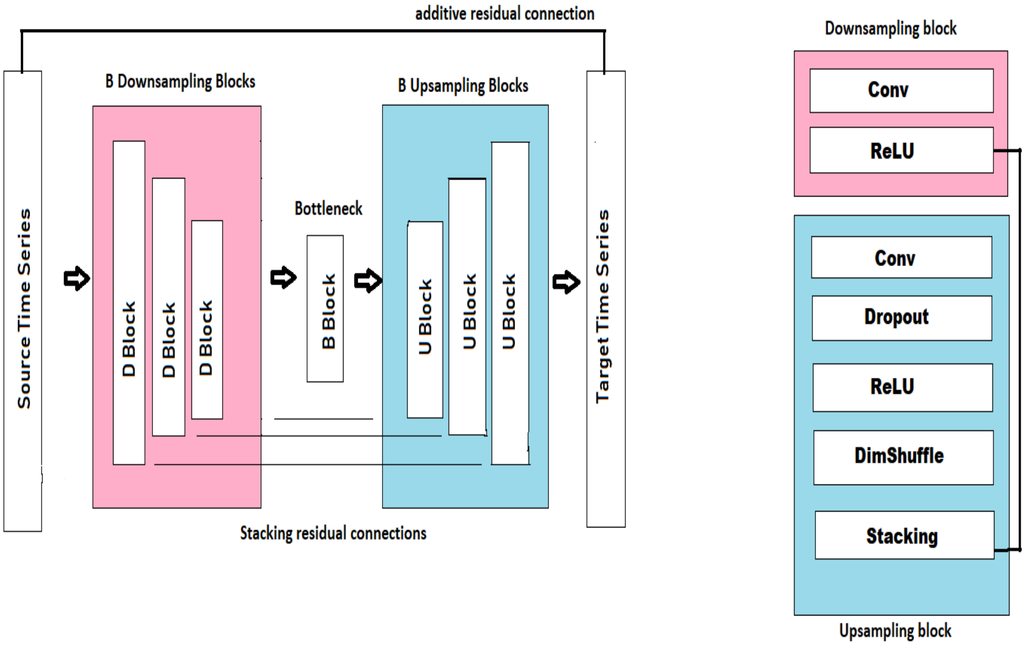

Fig. 1: Densely connected neural network implemented in Bottleneck method with input as time series and expected output as target time series. Up-sampling and down-sampling of data feed into CNN

4.2.2 Down-sampling: In down-sampling, signal can be down-sampled wherever the sample is oversampled. So we use decimation for this process and when the fraction values cannot be decimated, they are converted to float values and then the proper functionalities are added. Therefore the main purpose of this process is whenever a sample is oversampled that can be down-sampled using the low pass filter (the filter might be IIR or FIR)

$$\text{Stopband low frequency } = \frac{1}{output\ rate} - \text{ pass band upper frequency}$$

Aliasing is reduced by neglecting the values with decimating filters in between. This can be used for the up-sampling procedure to increase the bandwidth of the frequency or the signal.

The method used here is up-sampling and down-sampling. we consider a low resolution signal of input-data with a sampling rate R1 and y be a high resolution signal with sampling rate R2, R2 >R1 where x = $\{x_1$/R1 ,….x_{R1T1}/R1 } and y = { y_1/ R2,….

y_{yR2T2}/R2} and the definition for the x,y can be given as y = f_θ(x) where θ is a parameter and f_θ. We obtain by running the model with convolutional neural network on large set of data. For example, x_i , y_i.

4.3 Model Architecture

The model is the basic part of any deep learning project. As shown in diagram stacking residual models first the down-sampling files are let into the pipeline for computation of model for transforming and correlating data. We describe our architecture in figure 1. Since we use Generative model and needs deep convolutional neural networks which fill the gaps between the two frequencies rather than extracting features. Two common layers used are up-sampling and down-sampling blocks. The cublic up-sampling blocks B are given as inputs in between two samples with low resolution. Then we pass the following result into the B feed-forward Down-sampling convolutions with a stride of 2 and non-linear activation function Relu, so that each bock performs convolution and batch normalization by passing activation function. Here it increases the number of filters by 2 at each stage. All the operations and functions are stored in tensors. B of i^{th} down-sampling feature are concatenated or stacked with (B-i)th tensor of up-sampling feature for the model to improve cubic approximation by upscaling using subpixel shuffling layer. This subpixel is a transposed and subpixel convolution.

4.4 Setup

we evaluated out the data from the weights obtained by training vctk speaker dataset and also validating 105 new audio file recordings created at low and high resolutions. So from this audio set, we obtained an up-sampled high quality output. This output is connected to cubic up-sampling using additive residual connections image from Residual blocks- Building blocks of ResNet. If the network has more number of deep layers then the model which runs on simplex and complex functions may not understand very simple functions, this leads to a problem in running the model which encompasses both simplex and complex functions.

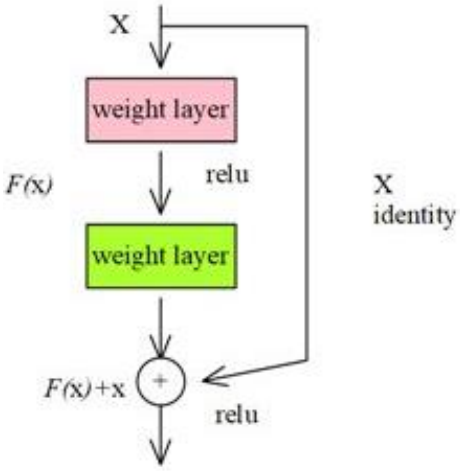

Fig. 2: shows how the residual connections are connected to the layers in CNN

4.5 Performance

The results are as we expected and gained a high and super resolution audio and also removed super distorted frequencies. The audios restore into their previous wave forms by checking the bit depth.

5. CONCLUSION

Machine learning techniques based on deep neural networks have been successful at solving underdefined problems in signal processing such as image super-resolution, colorization, in-painting, and many others. Learning-based methods often perform better in this context than general-purpose algorithms because they leverage sophisticated domain-specific models of the appearance of natural signals. In this work, we proposed new techniques that use this insight to up-sample audio signals. Our technique extends previous work on image super-resolution to the audio domain; it outperforms previous bandwidth extension approaches on both speech and non-vocal music. Our approach is fast and simple to implement, and has applications in telephony, compression, and text-to-speech generation. It also demonstrates the effectiveness of feedforward architectures on an important audio generation task, suggesting new directions for generative audio modelling

6. REFERENCES

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), Advances in Neural Information Processing Systems .

[2] J Balle´, V Laparra, and E P Simoncelli. End-to-end optimized image compression. In Int'l. Conf. on Learning Representations (ICLR2017), Toulon, France, April 2017.

[3] Yan Ming Cheng, Douglas O'Shaughnessy, and Paul Mermelstein. Statistical recovery of wideband speech from narrowband speech. IEEE Transactions on Speech and Audio Processing.

[4] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell.

[5] Per Ekstrand. Bandwidth extension of audio signals by spectral band replication. In in Proceedings of the 1st IEEE Benelux Workshop on Model Based Processing and Coding of Audio (MPCA02. Citeseer, 2002.

[6] Augustine Gray and John Markel. Distance measures for speech processing. IEEE Transactions on Acoustics, Speech, and Signal Processing, 24(5):380–391, 1976.

[7] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. arxiv, 2016.

[8] Erik Larsen and Ronald M Aarts. Audio bandwidth extension: application of psychoacoustics, signal processing and loudspeaker design. John Wiley & Sons, 2005.

[9] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. CoRR, abs/1609.04802, 2016.