Rapport Projet MLA : WaveNet

Lounes ALLIOUI LOUNES.ALLIOUI@ETU.SORBONNE-UNIVERISTE.FR
Branis GHOUL BRANIS.GHOUL@ETU.SORBONNE-UNIVERISTE.FR
Amel MEZEMATE AMEL.MEZEMATE@ETU.SORBONNE-UNIVERISTE.FR
Sara BRAHAMI SARA.BRAHAMI@ETU.SORBONNE-UNIVERISTE.FR
Master Ingénierie des Systemes Intelligents
Sorbonne Université
Paris, France

Editor: Machine Learning Avancé (2022-2023)

Abstract

In this paper, we introduce a deep learning model based on Wavenet, for generating raw audio waveforms. First, we will present its architecture, the utility of each block, as well as its functioning. In order to do the supervised learning of our neural network we have downloaded an open source database under the name VTCK-Corpus which contains audio files under the wav extension and their text files. We have done a pre-processing of the data, once they are ready to be used, we will train the model, to finally evaluate its performance using the chosen metrics. The main challenge is to implement the speech to text application.

Keywords: Wavenet, VTCK-Corpus, Speech to text, Supervised learning.

The code for this work is available via this link on Github

1. Introduction

Generating speech, music or texts are increasingly common in recent years thanks to the development of deep learning, several companies are involved, including Apple with Siri or Google Assitant. Indeed, the challenge is to achieve realistic voice synthesis and close to the human voice. Several models have been developed, we cite examples such as HMM, LSTM-RNN (1), but it turned out that the results obtained from these models are not that satisfactory, which pushed the researchers of the DeepMind company to create an architecture more powerful than these called WaveNet. The WaveNet is mainly based on convolutional neural networks, inspired by PiwelCNN having almost the same principal (1). During this project we will re-implement this model in order to evaluate its performance in different applications such as speech to text.

2. Presentation of the algorithm

2.1 Approach

WaveNet is a convolutional and autoregressive generative model. It generates human voices more realistic and natural than what the speech synthesis can generate and it models any type of audio (music, speech..) (2), this is due to the fact that it models sounds sample by sample, that is to say raw waveforms of the audio signal. Its architecture was inspired by PixelCNN (1) which models images pixel by pixel (3).

Inputs :

The input data is in the form of raw signals, which means that we can model any type of audio. The data is recovered in amplitudes, which means that it is in digital form that varies over time. At the time of training, the input sequences are real waveforms (4).

Causal dilated convolution

Wavenet is structured mainly around convolutional neural networks. We have two types of convolution : causal and dilated convolution. Let's focus on causal convolution first.

Temporal data often have superposition problems, i.e. future and past data get mixed up. The output data obviously cannot depend on the future data (since it has not been generated yet), the solution applied to this kind of problem is causal convolution.

Causality is only the consequences of some past actions. This convolution avoids the leakage of information from the future to the past, thus respecting the order and the meaning

$$p(x) = \prod p(xt|x1, ..., x_{t-1}) \tag{1}$$

will not depend on future times (5). There are two ways to implement a causal filter, the first one is to set to 0 the parts of the filter kernel that are concerned by the future input i.e. to mask them which corresponds to a masked convolution for images, where the principle is to mask some pixels so that the model can predict only according to the pixels already seen PixelRecurrentNeuralNetworks. The second method is based on the translation-equivariance property of the convolution, it consists in shifting and padding the signal by the size of the kernel, then canceling the shift.

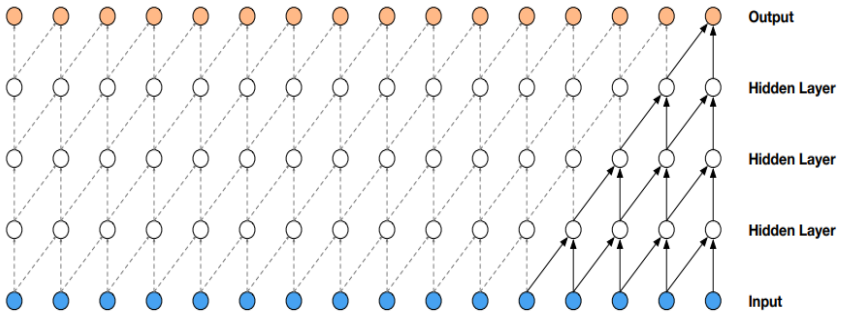

FIGURE 1 – causal convolution (1)

In second place we have dilated convolution. The main shortcoming of Wavenet is that the data is in the form of a high frequency waveform, with several samples of size greater than or equal to 16,000 in order to have a better result even if the sound is of short duration (we can also have less than 16,000 samples but the ideal for better quality is to take 16,000 samples or more).

So to be able to take into account as many samples as possible we proceed to a dilated convolution which will increase the reception field in a progressive way. The principle is simple, we increase the size of the filter by inserting zeros between the non-zero elements (increase the size of the filter) that is to say that it is applied on an area larger than its length so we skip input data with a certain step (voir la figure 5), here we have taken a step that is equal to 12.

$$receptivefield = 5 = (layers(= 4) + filterlength(= 2) - 1) \tag{2}$$

Softamax distribution

We convert continuous variables to discrete variables, one option is to look at 16-bit integers, one per time step, in this case we can model the probabilities from 2^{16} to $65,536$, if we put softmax on

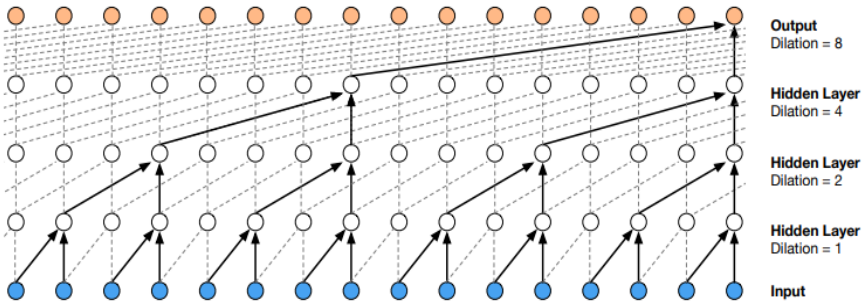

FIGURE 2 – stack of dilated causal convolutional layer (1)

it, it will be slow, there are many numbers, another option is to use the function (U law companding transformation) such as the equation (3).

$$f(x_t) = sign(x_t)\frac{ln(1+u|x_t|)}{1+u} \tag{3}$$

We assume that signal is already normalized to -1$\prec xt \prec$1 and u=255, then we can quantize the transformed output to 256 possible values. 256 values is much better than 65.536 probabilities, as mentioned in Wavenet.

Gated Activation Units(GAU)

GAUs allow the network to control the information to be propagated through the hierarchy of layers. We show that this mechanism is useful for language modeling because it can select the words or features that are relevant for predicting the next word. This works better than using the rectified linear activation function.

$$z = tanh(W_f k * x) \circledast \sigma(W_g k * x) \tag{4}$$

where * denotes a convolution operator, $\circledast$ denotes an element-wise multiplication operator, σ is a sigmoid function, k is the layer index, f and g denote filter and gate, respectively, and W is a learnable convolution filter,as in equation4 and as described in (1).

Residual and skip connections

Using residual and skip connections allows the network to be deeper , by adding skip conncetions we don't add parameters to learns. as described in (6).

3. Speech to text

the architecture used in speech to text is the same as explaned before,just in the output we will have the ground truth and the prediction, the will calculate the error to minimize it by using the loss.

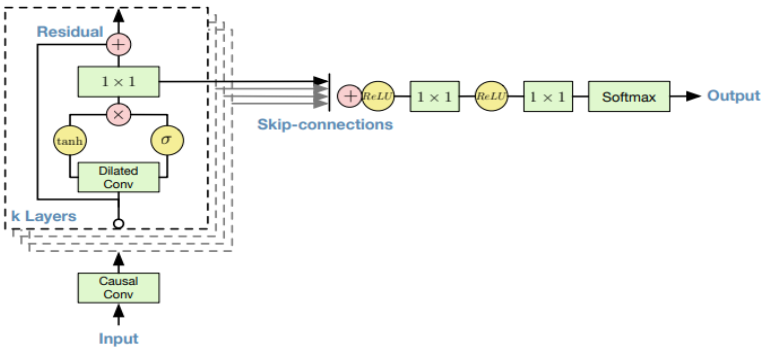

FIGURE 3 – wavenet neural network architecture (1)

4. Database

For the implementation of our application, we need a dataset that includes text, speech and the ID of each speaker. We used the Corpus CSTR VCTK : English Multi-Speaker Corpus for the CSTR Voice Cloning Toolkit. This includes speech data from 110 English speakers with different accents. Each speaker reads about 400 sentences, which were selected from a newspaper, the rainbow passage and an elicitation paragraph used for the speech accent archive. The newspaper texts were taken from the Glasgow Herald, with permission from the Herald Times Group. Each speaker has a different set of newspaper texts selected based on a greedy algorithm that increases contextual and phonetic coverage. WaveNet can be conditioned on the English alphabet for text generation by providing a vocabulary describing each letter of the English alphabet and on the speaker ID for speech recognition by providing the speaker ID to the model as a one-hot vector. The dataset consisted of 44 hours of data from 109 different speakers.

5. Evaluation expérimentale

5.1 Description of the experiment carried out, methodology and evaluation metrics.

Speech to Text is the transcription of speech into text. Today, Speech to Text is becoming more and more widespread, and in a wide variety of fields. It is therefore used in connected speakers, but can also be used to help people with disabilities or illiterate people by automatically transcribing what they say on a computer, without them having to type it. More traditionally, Speech-to-Text can be applied to do scripting work, particularly in call centres to keep a written record of all calls received without spending hours doing it by hand (7).

Deep Learning algorithms are based on mathematical calculations, in particular the propagation and back propagation of the gradient (8). For this purpose, the input and output data of the neural networks are qualitative (discrete or continuous) values representative for each application. However, our application "Speech-To-Text" processes both raw audio and text as inputs. Therefore, it is essential to pre-process this information to best represent our data. Then, extract the key features for the proper functioning of the model.

5.1.1 Preprossecing

The most common way to perform text or speech recognition is to slice the recording at each silence and then find out what was said there. To do this, coefficients representing few millise-

conds slice of the audio are calculated and placed in vectors : the MFCCs (Mel Frequency Cepstral Coefficients). These are specifically designed for sound analysis and are frequently used as input parameters for Speech to Text algorithms. These parameters are then sent to a model that will perform speech recognition. WaveNet has a particular architecture that consists of directly modelling the raw waveform of the audio signal, one sample at a time (letter by letter).

5.1.2 Pre-processing of audio files by MFCC

Cepstral Transactions (MFCC) is a method of extracting features from audio. MFCC uses the MEL scale to subdivide a band of frequencies, and then extracts Cepstral coefficients using the Discrete Cosine Transform (DCT). The MEL scale is based on the way humans distinguish frequencies, which makes it very practical for sound processing. However, the fundamental vocal frequency range of adult humans is 85 Hz to 255 Hz. In addition to the fundamental frequency, there are harmonics of fundamental frequencies. The harmonics are integer multiplications of the fundamental frequency. An example can be seen in the image below which shows the frequency versus time of several spoken words and the colour represents the frequency power at that point (the stronger yellow) :

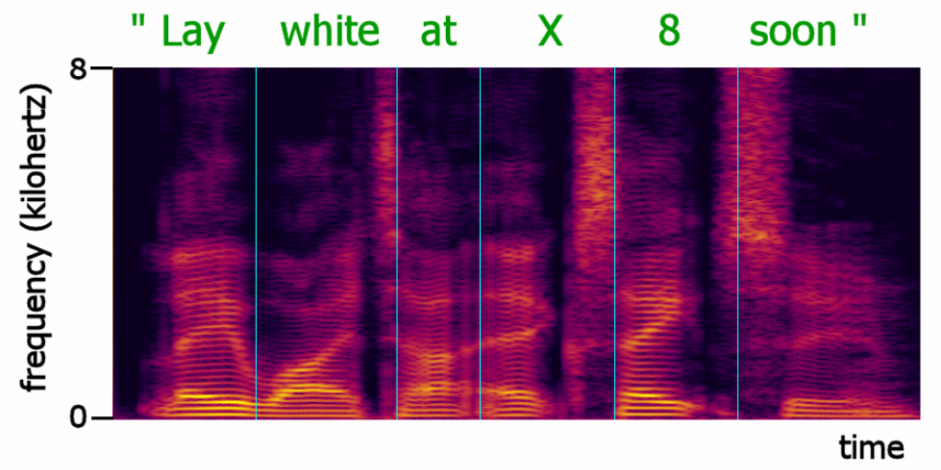

Figure 4 – Spectogram of an audio recording

Notice the first horizontal yellow line at the bottom of each segment. This is the fundamental frequency and is the strongest. Above it, there are harmonics with the same frequency distance from each other. Humans can hear approximately between 20Hz and 20KHz. Sound perception is non-linear and we can distinguish low frequency sounds better than high frequency sounds. And so the MFCC function in the code gives us ...

5.1.3 Pre-processing of text files

As far as the text files are concerned, the processing is quite simple. For each audio recording, we have a text file describing its content. The pre-processing consists in extracting the characters constituting the sentence without taking into account the special characters " !, ?, *, ect." and the numbers in order to improve the quality of the data. This increases the performance of the machine learning model by focusing only on the actual text data. However, for certain types of applications it may be interesting to preserve certain character types. Each character is represented by its index (its position) in a predefined vocabulary which will be given as input to the neural network as labels.c

5.1.4 Result of the pre-treatment

We have created files of type ***(.record)*** where we save the results of the pre-processing in the form of a dictionary containing the encoding, the dimensions, the audio data and the text data (labels) which are compatible with our WaveNet model. This allows us to create a local dataset that will be called at each task without the need for pre-processing at each stage of training and validation, thus accelerating learning.

5.1.5 Training

First, while the Paper used the TIMIT dataset for the speech recognition experiment, we used the free VTCK dataset.

Second, the Paper added a mean-pooling layer after the dilated convolution layer for down-sampling. We extracted MFCC from wav files and removed the final mean-pooling layer because the original setting was impossible to run on our AMD GPU.

Third, since the TIMIT dataset has phoneme labels, the Paper trained the model with two loss terms, phoneme classification and next phoneme prediction. We, instead, used a single CTC loss because VCTK provides sentence-level labels. As a result, we used only dilated conv1d layers without any maxpooling layer.

As an optimizer, we employed Adam, with an initial learning rate of 0.001 and batch size of 8. We devided our dataset to 80% for training and 20% for test.

Loss Function :

CTC (Connectionist Temporal Classification) calculates loss between a continuous (unsegmented) time series and a target sequences, and they sums over the probability of possible alignments of input to target. (9).

as as presented in (Graves et al., 2006)

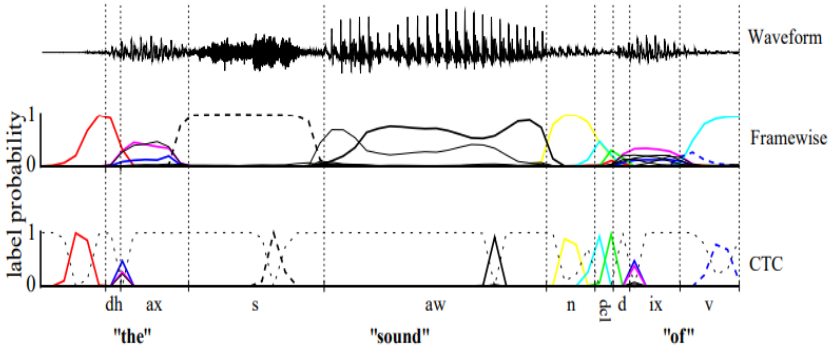

Figure 5 – Connectionist Temporal Classification

5.1.6 evaluation metrics

As the evaluation metric, we used the **F1-Score**(combining Precision and Recall). The goal of the F1 score is to combine the precision and recall metrics into a single metric. At the same time, the F1 score has been designed to work well on imbalanced data.the formula for the F1 score is following :

$$F1Score = 2 \times \frac{Presioncion \times Recall}{Precision + Recall} \tag{5}$$

5.2 Results

Presentation of the experimental results obtained and comparison with those of the reference article.

We trained our WaveNet model with the CSTR VCTK database with 450 minutes of raw audio recording for 24 epochs. We get F1-Score of 82%.

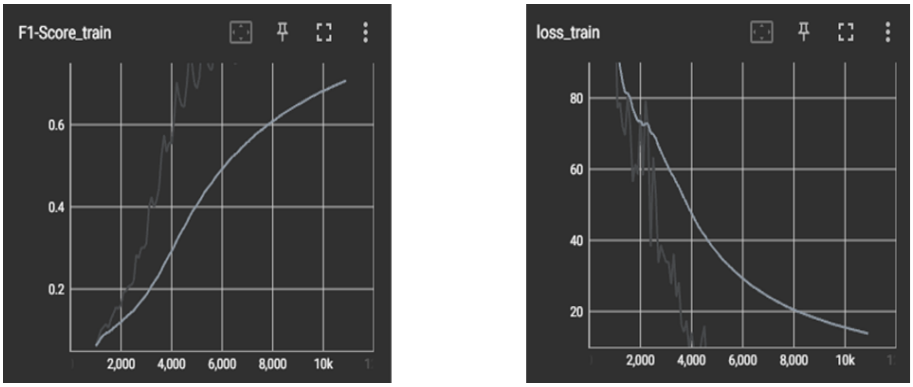

FIGURE 6 – The graph represents the training loss and the f1 score of the model, visualized on tensorboard.The abscissa axis represents the step, and the coordinate axis represents the value of the loss and the F1 score

As we can see from the figures above, our model is able to learn the raw audio samples. This is represented by the decrease in the Loss errors during training. The curve on the figure also shows that the model can further improve its accuracy by training on more data and for longer.

Results of text generation

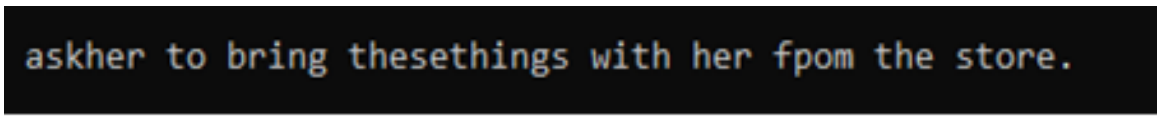

FIGURE 7 – Ask her to bring these things with her from the store

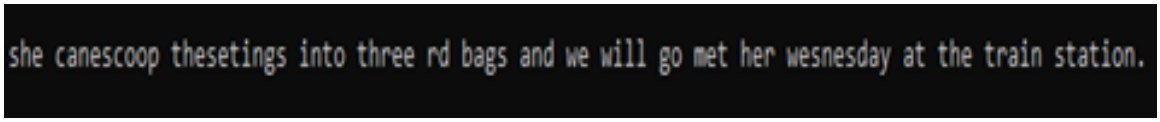

FIGURE 8 – She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

Nevertheless, our model fails to generate a correct sentence, it is able to initially respect the spaces between words (knowing that the prediction is done letter by letter) so the prediction of spaces is true and also predicts some short words, this proves that our model manages to learn.

However, we achieve F1-Score of ...

6. Conclusion

A summary of the main results obtained and presentation of the main areas for improvement.

This work presented a simplified architecture of the WaveNet model, a deep generative model of audio data that operates directly at the waveform level. WaveNets are autoregressive and combine causal filters with dilated convolutions to allow their receptive fields to grow exponentially with depth, which is important for modelling long-range temporal dependencies in audio signals. We have shown how WaveNets can be conditioned by other inputs either globally (e.g. speaker identity) or locally (e.g. linguistic features). This project allowed us to apply the different notions seen in the Sound Processing and Advanced Machine Learning course to the creation of deep learning voice recognition models.

We would like to thank all the people who contributed to this project.

Références

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet : A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016. [Online]. Available : http://arxiv.org/abs/1609.03499

[2] wavenet-a-generative-model-for-raw-audio. [Online]. Available : https://www.deepmind.com/blog/wavenet-a-generative-model-for-raw-audio/

[3] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," *CoRR*, vol. abs/1601.06759, 2016. [Online]. Available : http://arxiv.org/abs/1601.06759

[4] Neural networks for real-time audio : Wavenet. [Online]. Available : https://towardsdatascience.com/neural-networks-for-real-time-audio-wavenet-2b5cdf791c4f

[5] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet : A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016. [Online]. Available : http://arxiv.org/abs/1609.03499

[6] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph wavenet for deep spatial-temporal graph modeling." [Online]. Available : https://arxiv.org/abs/1906.00121

[7] R. Shadiev, W.-Y. Hwang, N.-S. Chen, and Y.-M. Huang, "Review of speech-to-text recognition technology for enhancing learning," *Educational Technology Society*, vol. 17, p. 65–84, 11 2014.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[9] A. Hannun, "Sequence modeling with ctc," *Distill*, 2017, https ://distill.pub/2017/ctc.