V. MODERN INFORMATION TECHNOLOGIES AND THEIR APPLICATION
V. СОВРЕМЕННЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И ИХ ПРИМЕНЕНИЕ

Abramova Alexandra
Student
Ural Federal University
Russia, Yekaterinburg
Research advisor: Kovaleva Alexandra

IMAGE PROCESSING AND GENERATION METHODS IN AUDIO PROCESSING AND GENERATION

Abstract: *The article provides an overview of modern methods in audio processing, including the generation of music and speech. The paper considers 3 approaches that have better parameters in comparison with existing implementations. Models are built on different types of neural networks. A comparison of the constructed implementations is given.*

Keywords: *recurrent (RNN), convolutional (CNN) and generative adversarial (GAN) neural networks, sound processing, sound sequences generation, music generation, speech generation.*

Абрамова Александра Егоровна
Студент
Уральский Федеральный Университет
Россия, г. Екатеринбург
Научный руководитель: Ковалева Александра Георгиевна

МЕТОДЫ ОБРАБОТКИ И СИНТЕЗА ИЗОБРАЖЕНИЙ В ОБРАБОТКЕ И ГЕНЕРАЦИИ АУДИО

Аннотация: *Статья содержит обзор современных методов в обработке звуковых последовательностей, в том числе генерации музыки и речи. В работе рассмотрены 3 подхода, имеющих хорошие показатели в сравнении с имеющимися реализациями. Модели построены на разных типах нейронных сетей. Приводится сравнение построенных реализаций.*

Ключевые слова: *рекуррентные, сверточные и генеративно-состязательные нейронные сети, обработка звука, генерация звуковых последовательностей, синтез музыки, синтез речи.*

Introduction

Machine learning has become popular through the automation of a wide range of tasks and simplification of human life. Various training methods are used in banking, telecommunications companies, consulting, as well as in science, medicine, etc. But they can help people not only in such familiar and understandable tasks, but also, for example, in creative activity. Artificial intelligence already knows how to talk, draw pictures, edit photos and videos. It can also create music, including for commercial purposes.

For example, the AIVA (Artificial Intelligence Virtual Artist) [1] generates music in several styles for films, video games, etc. Anyone can feel like a composer with the help of this project. The resulting music tracks can be downloaded and edited.

This product helped to complete the unfinished music piece by composer Antonin Dvořák 115 years after his death. The algorithm was trained on a data set that contains 30 thousand works, including all the works of the composer. Then people selected the most suitable passages for the composer's style and combined them.

Thus, machine learning methods can develop and complement various spheres of human creativity.

WaveNet

WaveNet [2] is a deep neural network designed to generate audio signals in the raw format (the format of raw or minimally processed data, for audio files – without compression and headers). The network is capable of generating musical works, as well as working with speech: generate (in TTS format – text-to-speech), transform and improve it. The base of the work was PixelCNN [3] – a convolutional neural network for generating images adapted one-dimensional WaveNet from 2-dimensional networks for images. The main ingredient of WaveNet is dilated convolution. It is used to increase the receptive field without a significant gain of compute capacity and allows the network to work efficiently on a larger scale than with vanilla convolution. Experiments include multi-speaker speech generation, text-to-speech, music and speech recognition. In this review we will focus only on the generation of human speech and musical samples. Multi-speaker aimed at free-form speech generation. Because the model is not conditioned on text, it generates non-existent but human language-like words with realistic sounding intonations.

Text-To-Speech approach learn on single-speaker speech databases from which Google's TTS systems are built. WaveNet is compared with another speech synthesizers. As result, WaveNet outperformed the baseline statistical parametric and concatenative speech synthesizers in both languages: English and Chinese. But the network sometimes generates samples that had unnatural prosody by stressing wrong words in a sentence.

For music generation WaveNets learned on two music datasets. It was defined that the size of the receptive field was crucial to obtain samples that were produced by music. Even with a receptive field of several seconds, the models did not enforce long-range consistency which resulted in variations in genre, instrumentation, volume and sound quality. Nevertheless, the samples were often harmonic and aesthetically pleasing.

Estimation of quality of the received results is quite difficult, therefore, many researchers in this field in their work use the method of "blind assessment" by independent listeners who put points to different samples on a certain scale. An average

rating is compiled. You can listen and evaluate the samples obtained using WaveNet using the link [2.2].

MelNet

As in the previous instance, MelNet [4] works similarly to PixelCNN, it has the same functions as WaveNet such as the generation of speech and music, but is based on Recurrent neural networks.

The authors of the article [4.1] assert WaveNet model is well adapted for modeling local dependencies, but does not capture a higher-level structure. In MelNet managed to take this aspect into account by modeling spectrograms, instead of one-dimensional signals (Fig. 1).

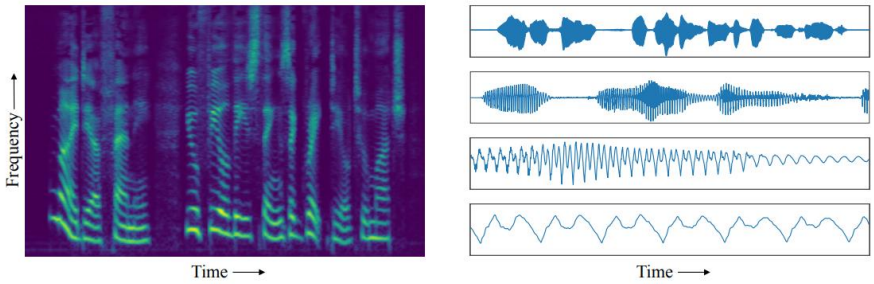

Figure 1. Spectrogram (on the left) and signal waveform

The usage of spectrograms in generative models also has minus the loss of the local structure of audio samples. To reduce the information loss at the lower level, it is necessary to simulate high-resolution spectrograms that have the same dimension as the corresponding signals in the time domain.

The independent experts' assessment of samples generated by the MelNet and WaveNet is shown in Figure 2. MelNet illustrates that its samples have a deeper structure for all types of tasks: speech generation in various ways and music. To rate MelNet samples quality at [4.2].

	WaveNet	MelNet
Blizzard	0.0 %	100.0 %
VoxCeleb2	0.0 %	100.0 %
MAESTRO	4.2 %	95.8 %

Рисунок 2 – Сравнительная таблица WaveNet и MelNet

GANSynth

GANSynth [5] is the method for generating high-quality sound using Generative Adversarial Networks.

As opposed to standard audio generation models, GANSynth generates the entire sequence in parallel, synthesizing samples much faster than in real time on a modern GPU and ~ 50,000 times faster than the standard WaveNet. Using a data set of music sounds from various NSynth musical instruments, the network can control the pitch sound and timbre independently.

Figure 3 shows a comparison of the generation methods and comparison with other models using the following metrics: Number of Statistically-Different Bins (NDB), Inception Score (IS), Pitch Accuracy (PA) и Pitch Entropy (PE), (Frechet Inception Distance – FID). «+ H» means higher frequency resolution, and «Real Data» means that the sample is taken from a test set.

Examples	Human Eval (wins)	NDB	FID	IS	PA	PE
Real Data	549	2.2	13	47.1	98.2	0.22
IF-Mel + H	**485**	**29.3**	167	38.1	97.9	0.40
IF + H	308	36. 0	**104**	**41.6**	**98.3**	**0.32**
Phase + H	225	37.6	592	36.2	97.6	0.44
IF-Mel	479	37.0	600	29.6	94.1	0.63
IF	283	37.0	708	36.3	96.8	0.44
Phase	203	41.4	687	24.4	94.4	0.77
WaveNet	359	45.9	320	29.1	92.7	0.70
WaveGAN	216	43.0	461	13.7	82.7	1.40

Figure 3. Comparison of various methods and models

The GANs generating instantaneous frequency (IF) for the phase component outperform other representations and strong base models, including signal-generating networks such as WaveNet. Progressive training (P) and an increase in STFT frequency resolution (H) increase productivity by helping to separate closely spaced harmonics. The obtained samples can be heard on the link [5.2].

Conclusion

WaveNet creates sound samples based on the previous set of sounds sequentially. MelNet and GANSynth models generate sound sequences in parallel, so they work much faster than WaveNet.

Also in the 2 and 3 implementations the spectrogram design method is used, which in addition to visualizing the sound allows working with audio samples as images. Spectrograms also capture the structure of a higher level of signals than using oscillograms and one-dimensional visualization of audio samples as in WaveNet. In addition, spectrograms are more compact than waveforms, which allow models to work with a large time interval of audio samples.

The disadvantage of spectrograms is in inability to maintain a local structure as sequential methods for generating sound. Therefore, they may be generated with high resolution, which also requires certain compute capacity.

All implementations are based on image processing methods adapted to the task of generating music and speech, but in the case of WaveNet, a transition was made from 2-dimensional to 1-dimensional. WaveNet is the basis for the development of the last two models, as the authors are guided by the results obtained by this model. MelNet and GANSynth models contain all the features that WaveNet has, and in many cases even surpass it.

However, the proposed models have disadvantages. While listening to synthesized music it was noticed that musical samples have fuzzy rhythms - sometimes a rhythm dimension changes. In the generated speech this problem was not found. It is necessary to underline, the algorithms imitate speech generation better than music because it has a more complex structure and more restrictions.

REFERENCES

1. AIVA - Artificial Intelligence Virtual Artist. URL: https://www.aiva.ai/about

2. van den Oord, Aaron, Dieleman, Sander, Zen, Heiga, Simonyan, Karen, Vinyals, Oriol, Graves, Alex, Kalchbrenner, Nal, Senior, Andrew, and Kavukcuoglu, Koray. WaveNet: A Generative Model for Raw Audio. 2.1 URL: https://arxiv.org/pdf/1609.03499

2.2 URL: https://deepmind.com/blog/wavenet-generative-model-raw-audio

3. van den Oord, Aaron, Kalchbrenner, Nal, Vinyals, Oriol, Espeholt, Lasse, Graves, Alex, and Kavukcuoglu, Koray. Conditional image generation with PixelCNN decoders. CoRR, abs/1606.05328, 2016b. URL: http://arxiv.org/abs/1606.05328.

4. Vasquez, Sean, Lewis, Mike. MelNet: A Generative Model for Audio in the Frequency Domain.

4.1 URL: https://arxiv.org/pdf/1906.01083v1.pdf

4.2 URL: https://audio-samples.github.io

5. Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, & Adam Roberts. GANSynth: Adversarial Neural Audio Synthesis. ICLR 2019.

5.1 URL: https://openreview.net/pdf?id=H1xQVn09FX

5.2 URL:

https://storage.googleapis.com/magentadata/papers/gansynth/index.html