Title: WaveNet

URL: https://en.wikipedia.org/wiki/WaveNet

PageID: 54133326

Categories: Category:2016 in artificial intelligence, Category:2016 software, Category:Artificial neural networks, Category:Deep learning, Category:Google, Category:Google acquisitions, Category:Speech synthesis

---

Supervised learning

Unsupervised learning

Semi-supervised learning

Self-supervised learning

Reinforcement learning

Meta-learning

Online learning

Batch learning

Curriculum learning

Rule-based learning

Neuro-symbolic AI

Neuromorphic engineering

Quantum machine learning

Classification

Generative modeling

Regression

Clustering

Dimensionality reduction

Density estimation

Anomaly detection

Data cleaning

AutoML

Association rules

Semantic analysis

Structured prediction

Feature engineering

Feature learning

Learning to rank

Grammar induction

Ontology learning

Multimodal learning

Apprenticeship learning

Decision trees

Ensembles Bagging Boosting Random forest

Bagging

Boosting

Random forest

k -NN

Linear regression

Naive Bayes

Artificial neural networks

Logistic regression

Perceptron

Relevance vector machine (RVM)

Support vector machine (SVM)

BIRCH

CURE

Hierarchical

k -means

Fuzzy

Expectation–maximization (EM)

DBSCAN

OPTICS

Mean shift

Factor analysis

CCA

ICA

LDA

NMF

PCA

PGD

t-SNE

SDL

Graphical models Bayes net Conditional random field Hidden Markov

Bayes net

Conditional random field

Hidden Markov

RANSAC

k -NN

Local outlier factor

Isolation forest

Autoencoder

Deep learning

Feedforward neural network

Recurrent neural network LSTM GRU ESN reservoir computing

LSTM

GRU

ESN

reservoir computing

Boltzmann machine Restricted

Restricted

GAN

Diffusion model

SOM

Convolutional neural network U-Net LeNet AlexNet DeepDream

U-Net

LeNet

AlexNet

DeepDream

Neural field Neural radiance field Physics-informed neural networks

Neural radiance field

Physics-informed neural networks

Transformer Vision

Vision

Mamba

Spiking neural network

Memtransistor

Electrochemical RAM (ECRAM)

Q-learning

Policy gradient

SARSA

Temporal difference (TD)

Multi-agent Self-play

Self-play

Active learning

Crowdsourcing

Human-in-the-loop

Mechanistic interpretability

RLHF

Coefficient of determination

Confusion matrix

Learning curve

ROC curve

Kernel machines

Bias–variance tradeoff

Computational learning theory

Empirical risk minimization

Occam learning

PAC learning

Statistical learning

VC theory

Topological deep learning

AAAI

ECML PKDD

NeurIPS

ICML

ICLR

IJCAI

ML

JMLR

Glossary of artificial intelligence

List of datasets for machine-learning research List of datasets in computer vision and image processing

List of datasets in computer vision and image processing

Outline of machine learning

v

t

e

WaveNet is a deep neural network for generating raw audio. It was created by researchers at London-based AI firm DeepMind . The technique, outlined in a paper in September 2016, [1] is able to generate relatively realistic-sounding human-like voices by directly modelling waveforms using a neural network method trained with recordings of real speech. Tests with US English and Mandarin reportedly showed that the system outperforms Google's best existing text-to-speech (TTS) systems, although as of 2016 its text-to-speech synthesis still was less convincing than actual human speech. [2] WaveNet's ability to generate raw waveforms means that it can model any kind of audio, including music. [3]

History

Generating speech from text is an increasingly common task thanks to the popularity of software such as Apple's Siri , Microsoft's Cortana , Amazon Alexa and the Google Assistant . [4]

Most such systems use a variation of a technique that involves concatenated sound fragments together to form recognisable sounds and words. [5] The most common of these is called concatenative TTS. [6] It consists of large library of speech fragments, recorded from a single speaker that are then concatenated to produce complete words and sounds. The result sounds unnatural, with an odd cadence and tone. [7] The reliance on a recorded library also makes it difficult to modify or change the voice. [8]

Another technique, known as parametric TTS, [9] uses mathematical models to recreate sounds that are then assembled into words and sentences. The information required to generate the sounds is stored in the parameters of the model. The characteristics of the output speech are controlled via the inputs to the model, while the speech is typically created using a voice synthesiser known as a vocoder . This can also result in unnatural sounding audio.

Design and ongoing research

Background

WaveNet is a type of feedforward neural network known as a deep convolutional neural network (CNN). In WaveNet, the CNN takes a raw signal as an input and synthesises an output one sample at a time. It does so by sampling from a softmax (i.e. categorical) distribution of a signal value that is encoded using μ-law companding transformation and quantized to 256 possible values. [11]

Initial concept and results

According to the original September 2016 DeepMind research paper WaveNet: A Generative Model for Raw Audio , [12] the network was fed real waveforms of speech in English and Mandarin. As these pass through the network, it learns a set of rules to describe how the audio waveform evolves over time. The trained network can then be used to create new speech-like waveforms at 16,000 samples per second. These waveforms include realistic breaths and lip smacks – but do not conform to any language. [13]

WaveNet is able to accurately model different voices, with the accent and tone of the input correlating with the output. For example, if it is trained with German, it produces German speech. [14] The capability also means that if the WaveNet is fed other inputs – such as music – its output will be musical. At the time of its release, DeepMind showed that WaveNet could produce waveforms that sound like classical music . [15]

Content (voice) swapping

According to the June 2018 paper Disentangled Sequential Autoencoder , [16] DeepMind has successfully used WaveNet for audio and voice "content swapping": the network can swap the voice on an audio recording for another, pre-existing voice while maintaining the text and other features from the original recording. "We also experiment on audio sequence data. Our disentangled representation allows us to convert speaker identities into each other while conditioning on the content of the speech." (p. 5) "For audio, this allows us to convert a male speaker into a female speaker and vice versa [...] ." (p. 1) According to the paper, a two-digit minimum amount of hours (c. 50 hours) of pre-existing speech recordings of both source and target voice are required to be fed into WaveNet for the program to learn their individual features before it is able to perform the conversion from one voice to another at a satisfying quality. The authors stress that " [a] n advantage of the model is that it separates dynamical from static features [...] ." (p. 8), i. e. WaveNet is capable of distinguishing between the spoken text and modes of delivery (modulation, speed, pitch, mood, etc.) to maintain during the conversion from one voice to another on the one hand, and the basic features of both source and target voices that it is required to swap on the other.

The January 2019 follow-up paper Unsupervised speech representation learning using WaveNet autoencoders [17] details a method to successfully enhance the proper automatic recognition and discrimination between dynamical and static features for "content swapping", notably including swapping voices on existing audio recordings, in order to make it more reliable. Another follow-up

paper, Sample Efficient Adaptive Text-to-Speech , [18] dated September 2018 (latest revision January 2019), states that DeepMind has successfully reduced the minimum amount of real-life recordings required to sample an existing voice via WaveNet to "merely a few minutes of audio data" while maintaining high-quality results.

Its ability to clone voices has raised ethical concerns about WaveNet's ability to mimic the voices of living and dead persons. According to a 2016 BBC article, companies working on similar voice-cloning technologies (such as Adobe Voco) intend to insert watermarking inaudible to humans to prevent counterfeiting, while maintaining that voice cloning satisfying, for instance, the needs of entertainment-industry purposes would be of a far lower complexity and use different methods than required to fool forensic evidencing methods and electronic ID devices, so that natural voices and voices cloned for entertainment-industry purposes could still be easily told apart by technological analysis. [19]

Applications

At the time of its release, DeepMind said that WaveNet required too much computational processing power to be used in real world applications. [20] As of October 2017, Google announced a 1,000-fold performance improvement along with better voice quality. WaveNet was then used to generate Google Assistant voices for US English and Japanese across all Google platforms. [21] In November 2017, DeepMind researchers released a research paper detailing a proposed method of "generating high-fidelity speech samples at more than 20 times faster than real-time", called "Probability Density Distillation". [22] At the annual I/O developer conference in May 2018, it was announced that new Google Assistant voices were available and made possible by WaveNet; WaveNet greatly reduced the number of audio recordings that were required to create a voice model by modeling the raw audio of the voice actor samples. [23]

See also

15.ai

Deep learning speech synthesis

References

External links

WaveNet: A Generative Model for Raw Audio

v

t

e

Google

Google Brain

Google DeepMind

AlphaGo (2015)

Master (2016)

AlphaGo Zero (2017)

AlphaZero (2017)

MuZero (2019)

Fan Hui (2015)

Lee Sedol (2016)

Ke Jie (2017)

AlphaGo (2017)

The MANIAC (2023)
AlphaFold (2018)
AlphaStar (2019)
AlphaDev (2023)
AlphaGeometry (2024)
AlphaGenome (2025)
Inception (2014)
WaveNet (2016)
MobileNet (2017)
Transformer (2017)
EfficientNet (2019)
Gato (2022)
Quantum Artificial Intelligence Lab
TensorFlow
Tensor Processing Unit
Assistant (2016)
Sparrow (2022)
Gemini (2023)
BERT (2018)
XLNet (2019)
T5 (2019)
LaMDA (2021)
Chinchilla (2022)
PaLM (2022)
Imagen (2023)
Gemini (2023)
VideoPoet (2024)
Gemma (2024)
Veo (2024)
DreamBooth (2022)
NotebookLM (2023)
Vids (2024)
Gemini Robotics (2025)
" Attention Is All You Need "
Future of Go Summit
Generative pre-trained transformer
Google Labs
Google Pixel

Google Workspace
Robot Constitution
Category
Commons
v
t
e
History timeline
timeline
Companies
Projects
Parameter Hyperparameter
Hyperparameter
Loss functions
Regression Bias–variance tradeoff Double descent Overfitting
Bias–variance tradeoff
Double descent
Overfitting
Clustering
Gradient descent SGD Quasi-Newton method Conjugate gradient method
SGD
Quasi-Newton method
Conjugate gradient method
Backpropagation
Attention
Convolution
Normalization Batchnorm
Batchnorm
Activation Softmax Sigmoid Rectifier
Softmax
Sigmoid
Rectifier
Gating
Weight initialization
Regularization
Datasets Augmentation
Augmentation
Prompt engineering

Reinforcement learning Q-learning SARSA Imitation Policy gradient

Q-learning

SARSA

Imitation

Policy gradient

Diffusion

Latent diffusion model

Autoregression

Adversary

RAG

Uncanny valley

RLHF

Self-supervised learning

Reflection

Recursive self-improvement

Hallucination

Word embedding

Vibe coding

Machine learning In-context learning

In-context learning

Artificial neural network Deep learning

Deep learning

Language model Large language model NMT

Large language model

NMT

Reasoning language model

Model Context Protocol

Intelligent agent

Artificial human companion

Humanity's Last Exam

Artificial general intelligence (AGI)

AlexNet

WaveNet

Human image synthesis

HWR

OCR

Computer vision

Speech synthesis 15.ai ElevenLabs

15.ai

ElevenLabs

Speech recognition Whisper

Whisper

Facial recognition

AlphaFold

Text-to-image models Aurora DALL-E Firefly Flux Ideogram Imagen Midjourney Recraft Stable Diffusion

Aurora

DALL-E

Firefly

Flux

Ideogram

Imagen

Midjourney

Recraft

Stable Diffusion

Text-to-video models Dream Machine Runway Gen Hailuo AI Kling Sora Veo

Dream Machine

Runway Gen

Hailuo AI

Kling

Sora

Veo

Music generation Riffusion Suno AI Udio

Riffusion

Suno AI

Udio

Word2vec

Seq2seq

GloVe

BERT

T5

Llama

Chinchilla AI

PaLM

GPT 1 2 3 J ChatGPT 4 4o o1 o3 4.5 4.1 o4-mini 5

1

2

3

J

ChatGPT

4

4o

o1

o3

4.5

4.1

o4-mini

5

Claude

Gemini Gemini (language model) Gemma

Gemini (language model)

Gemma

Grok

LaMDA

BLOOM

DBRX

Project Debater

IBM Watson

IBM Watsonx

Granite

PanGu-Σ

DeepSeek

Qwen

AlphaGo

AlphaZero

OpenAI Five

Self-driving car

MuZero

Action selection AutoGPT

AutoGPT

Robot control

Alan Turing

Warren Sturgis McCulloch

Walter Pitts

John von Neumann
Claude Shannon
Shun'ichi Amari
Kunihiko Fukushima
Takeo Kanade
Marvin Minsky
John McCarthy
Nathaniel Rochester
Allen Newell
Cliff Shaw
Herbert A. Simon
Oliver Selfridge
Frank Rosenblatt
Bernard Widrow
Joseph Weizenbaum
Seymour Papert
Seppo Linnainmaa
Paul Werbos
Geoffrey Hinton
John Hopfield
Jürgen Schmidhuber
Yann LeCun
Yoshua Bengio
Lotfi A. Zadeh
Stephen Grossberg
Alex Graves
James Goodnight
Andrew Ng
Fei-Fei Li
Alex Krizhevsky
Ilya Sutskever
Oriol Vinyals
Quoc V. Le
Ian Goodfellow
Demis Hassabis
David Silver
Andrej Karpathy
Ashish Vaswani

Noam Shazeer

Aidan Gomez

John Schulman

Mustafa Suleyman

Jan Leike

Daniel Kokotajlo

François Chollet

Neural Turing machine

Differentiable neural computer

Transformer Vision transformer (ViT)

Vision transformer (ViT)

Recurrent neural network (RNN)

Long short-term memory (LSTM)

Gated recurrent unit (GRU)

Echo state network

Multilayer perceptron (MLP)

Convolutional neural network (CNN)

Residual neural network (RNN)

Highway network

Mamba

Autoencoder

Variational autoencoder (VAE)

Generative adversarial network (GAN)

Graph neural network (GNN)

Category