Title: Highway network

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

PageID: 55375136

Categories: Category:2015 in artificial intelligence, Category:Machine learning, Category:Neural network architectures

---

In machine learning , the Highway Network was the first working very deep feedforward neural network with hundreds of layers, much deeper than previous neural networks . [1] [2] [3] It uses skip connections modulated by learned gating mechanisms to regulate information flow, inspired by long short-term memory (LSTM) recurrent neural networks . [4] [5] The advantage of the Highway Network over other deep learning architectures is its ability to overcome or partially prevent the vanishing gradient problem , [6] thus improving its optimization. Gating mechanisms are used to facilitate information flow across the many layers ("information highways"). [1] [2]

Highway Networks have found use in text sequence labeling and speech recognition tasks. [7] [8]

In 2014, the state of the art was training deep neural networks with 20 to 30 layers. [9] Stacking too many layers led to a steep reduction in training accuracy, [10] known as the "degradation" problem. [11] In 2015, two techniques were developed to train such networks: the Highway Network (published in May), and the residual neural network , or ResNet [12] (December). ResNet behaves like an open-gated Highway Net.

Model

The model has two gates in addition to the H (W H , x) {\displaystyle H(W_{H},x)} gate: the transform gate T (W T , x) {\displaystyle T(W_{T},x)} and the carry gate C (W C , x) {\displaystyle C(W_{C},x)} . The latter two gates are non-linear transfer functions (specifically sigmoid by convention). The function H {\displaystyle H} can be any desired transfer function.

The carry gate is defined as:

C (W C , x) = 1 – T (W T , x) {\displaystyle C(W_{C},x)=1-T(W_{T},x)}

while the transform gate is just a gate with a sigmoid transfer function.

Structure

The structure of a hidden layer in the Highway Network follows the equation:

y = H (x , W H) ⋅ T (x , W T) + x ⋅ C (x , W C) = H (x , W H) ⋅ T (x , W T) + x ⋅ (1 – T (x , W T)) {\displaystyle {\begin{aligned}y=H(x,W_{H})\cdot T(x,W_{T})+x\cdot C(x,W_{C})\\=H(x,W_{H})\cdot T(x,W_{T})+x\cdot (1-T(x,W_{T}))\end{aligned}}}

Related work

Sepp Hochreiter analyzed the vanishing gradient problem in 1991 and attributed to it the reason why deep learning did not work well. [6] To overcome this problem, Long Short-Term Memory (LSTM) recurrent neural networks [4] have residual connections with a weight of 1.0 in every LSTM cell (called the constant error carrousel) to compute y t + 1 = F (x t) + x t {\textstyle y_{t+1}=F(x_{t})+x_{t}} . During backpropagation through time , this becomes the residual formula y = F (x) + x {\textstyle y=F(x)+x} for feedforward neural networks. This enables training very deep recurrent neural networks with a very long time span t. A later LSTM version published in 2000 [5] modulates the identity LSTM connections by so-called "forget gates" such that their weights are not fixed to 1.0 but can be learned. In experiments, the forget gates were initialized with positive bias weights, [5] thus being opened, addressing the vanishing gradient problem.

As long as the forget gates of the 2000 LSTM are open, it behaves like the 1997 LSTM.

The Highway Network of May 2015 [1] applies these principles to feedforward neural networks .

It was reported to be "the first very deep feedforward network with hundreds of layers". [13] It is like a 2000 LSTM with forget gates unfolded in time , [5] while the later Residual Nets have no equivalent of forget gates and are like the unfolded original 1997 LSTM. [4] If the skip connections in Highway Networks are "without gates," or if their gates are kept open (activation 1.0), they become Residual Networks.

The residual connection is a special case of the "short-cut connection" or "skip connection" by Rosenblatt (1961) [14] and Lang & Witbrock (1988) [15] which has the form x ■ F (x) + A x {\displaystyle x\mapsto F(x)+Ax} . Here the randomly initialized weight matrix A does not have to be the identity mapping. Every residual connection is a skip connection, but almost all skip connections are not residual connections.

The original Highway Network paper [16] not only introduced the basic principle for very deep feedforward networks, but also included experimental results with 20, 50, and 100 layers networks, and mentioned ongoing experiments with up to 900 layers. Networks with 50 or 100 layers had lower training error than their plain network counterparts, but no lower training error than their 20 layers counterpart (on the MNIST dataset, Figure 1 in [16]). No improvement on test accuracy was reported with networks deeper than 19 layers (on the CIFAR-10 dataset; Table 1 in [16]). The ResNet paper, [17] however, provided strong experimental evidence of the benefits of going deeper than 20 layers. It argued that the identity mapping without modulation is crucial and mentioned that modulation in the skip connection can still lead to vanishing signals in forward and backward propagation (Section 3 in [17]). This is also why the forget gates of the 2000 LSTM [18] were initially opened through positive bias weights: as long as the gates are open, it behaves like the 1997 LSTM. Similarly, a Highway Net whose gates are opened through strongly positive bias weights behaves like a ResNet. The skip connections used in modern neural networks (e.g., Transformers) are dominantly identity mappings.

References

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