Title: Gating mechanism

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

PageID: 78052490

Categories: Category:Deep learning, Category:Neural network architectures

-----

In neural networks , the gating mechanism is an architectural motif for controlling the flow of activation and gradient signals . They are most prominently used in recurrent neural networks (RNNs), but have also found applications in other architectures.

RNNs

Gating mechanisms are the centerpiece of long short-term memory (LSTM). [1] They were proposed to mitigate the vanishing gradient problem often encountered by regular RNNs.

An LSTM unit contains three gates:

An input gate , which controls the flow of new information into the memory cell

A forget gate , which controls how much information is retained from the previous time step

An output gate , which controls how much information is passed to the next layer.

The equations for LSTM are: [2]

I t = σ (X t W x i + H t – 1 W h i + b i) F t = σ (X t W x f + H t – 1 W h f + b f) O t = σ (X t W x o + H t – 1 W h o + b o) C ~ t = tanh ■ (X t W x c + H t – 1 W h c + b c) C t = F t ■ C t – 1 + I t ■ C ~ t H t = O t ■ tanh ■ (C t) {\displaystyle {\begin{aligned}\mathbf {I} _{t}&=\sigma (\mathbf {X} _{t}\mathbf {W} _{xi}+\mathbf {H} _{t-1}\mathbf {W} _{hi}+\mathbf {b} _{i})\\\mathbf {F} _{t}&=\sigma (\mathbf {X} _{t}\mathbf {W} _{xf}+\mathbf {H} _{t-1}\mathbf {W} _{hf}+\mathbf {b} _{f})\\\mathbf {O} _{t}&=\sigma (\mathbf {X} _{t}\mathbf {W} _{xo}+\mathbf {H} _{t-1}\mathbf {W} _{ho}+\mathbf {b} _{o})\\{\tilde {\mathbf {C} }}_{t}&=\tanh(\mathbf {X} _{t}\mathbf {W} _{xc}+\mathbf {H} _{t-1}\mathbf {W} _{hc}+\mathbf {b} _{c})\\\mathbf {C} _{t}&=\mathbf {F} _{t}\odot \mathbf {C} _{t-1}+\mathbf {I} _{t}\odot {\tilde {\mathbf {C} }}_{t}\\\mathbf {H} _{t}&=\mathbf {O} _{t}\odot \tanh(\mathbf {C} _{t})\end{aligned}}}

Here, ■ {\displaystyle \odot } represents elementwise multiplication .

LSTM architecture, with gates

The gated recurrent unit (GRU) simplifies the LSTM. [3] Compared to the LSTM, the GRU has just two gates: a reset gate and an update gate . GRU also merges the cell state and hidden state. The reset gate roughly corresponds to the forget gate, and the update gate roughly corresponds to the input gate. The output gate is removed.

There are several variants of GRU. One particular variant has these equations: [4]

R t = σ (X t W x r + H t – 1 W h r + b r) Z t = σ (X t W x z + H t – 1 W h z + b z) H ~ t = tanh ■ (X t W x h + (R t ■ H t – 1) W h h + b h) H t = Z t ■ H t – 1 + (1 – Z t) ■ H ~ t {\displaystyle {\begin{aligned}\mathbf {R} _{t}&=\sigma (\mathbf {X} _{t}\mathbf {W} _{xr}+\mathbf {H} _{t-1}\mathbf {W} _{hr}+\mathbf {b} _{r})\\\mathbf {Z} _{t}&=\sigma (\mathbf {X} _{t}\mathbf {W} _{xz}+\mathbf {H} _{t-1}\mathbf {W} _{hz}+\mathbf {b} _{z})\\{\tilde {\mathbf {H} }}_{t}&=\tanh(\mathbf {X} _{t}\mathbf {W} _{xh}+(\mathbf {R} _{t}\odot \mathbf {H} _{t-1})\mathbf {W} _{hh}+\mathbf {b} _{h})\\\mathbf {H} _{t}&=\mathbf {Z} _{t}\odot \mathbf {H} _{t-1}+(1-\mathbf {Z} _{t})\odot {\tilde {\mathbf {H} }}_{t}\end{aligned}}}

Gated Recurrent Unit architecture, with gates

Gated Linear Unit

Gated Linear Units (GLUs) [5] adapt the gating mechanism for use in feedforward neural networks, often within transformer-based architectures. They are defined as:

$$\mathrm{GLU}(a,b)=a\odot \sigma (b)$$

where a,b are the first and second inputs, respectively. σ represents the sigmoid activation function.

Replacing σ with other activation functions leads to variants of GLU:

$$\begin{aligned}\mathrm{ReGLU}(a,b)&=a\odot \text{ReLU}(b)\\\mathrm{GEGLU}(a,b)&=a\odot \text{GELU}(b)\\\mathrm{SwiGLU}(a,b,\beta)&=a\odot \text{Swish}_{\beta}(b)\end{aligned}$$

where ReLU, GELU, and Swish are different activation functions.

In transformer models, such gating units are often used in the feedforward modules. For a single vector input, this results in: [6]

$$\begin{aligned}\operatorname{GLU}(x,W,V,b,c)&=\sigma (xW+b)\odot (xV+c)\\\operatorname{Bilinear}(x,W,V,b,c)&=(xW+b)\odot (xV+c)\\\operatorname{ReGLU}(x,W,V,b,c)&=\max(0,xW+b)\odot (xV+c)\\\operatorname{GEGLU}(x,W,V,b,c)&=\operatorname{GELU}(xW+b)\odot (xV+c)\\\operatorname{SwiGLU}(x,W,V,b,c,\beta)&=\operatorname{Swish}_{\beta}(xW+b)\odot (xV+c)\end{aligned}$$

Other architectures

Gating mechanism is used in highway networks, which were designed by unrolling an LSTM.

Channel gating [7] uses a gate to control the flow of information through different channels inside a convolutional neural network (CNN).

See also

- Recurrent neural network
- Long short-term memory
- Gated recurrent unit
- Transformer
- Activation function

References

Further reading

Zhang, Aston; Lipton, Zachary; Li, Mu; Smola, Alexander J. (2024). "10.1. Long Short-Term Memory (LSTM)". Dive into deep learning. Cambridge New York Port Melbourne New Delhi Singapore: Cambridge University Press. ISBN 978-1-009-38943-3.

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