Title: Mamba (deep learning architecture)

URL: https://en.wikipedia.org/wiki/Mamba_(deep_learning_architecture)

PageID: 75795581

Categories: Category:2023 in artificial intelligence, Category:Language modeling, Category:Neural network architectures

-----

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

Mamba [a] is a deep learning architecture focused on sequence modeling. It was developed by researchers from Carnegie Mellon University and Princeton University to address some limitations of transformer models , especially in processing long sequences. It is based on the Structured State Space sequence (S4) model. [2] [3] [4]

Architecture

To enable handling long data sequences, Mamba incorporates the Structured State Space Sequence model (S4). [2] S4 can effectively and efficiently model long dependencies by combining continuous-time, recurrent , and convolutional models. These enable it to handle irregularly sampled data, unbounded context, and remain computationally efficient during training and inferencing. [5]

Mamba introduces significant enhancements to S4, particularly in its treatment of time-variant operations. It adopts a unique selection mechanism that adapts structured state space model (SSM) parameters based on the input. [6] [2] This enables Mamba to selectively focus on relevant information within sequences, effectively filtering out less pertinent data. The model transitions from a time-invariant to a time-varying framework, which impacts both computation and efficiency. [2] [7]

Mamba employs a hardware-aware algorithm that exploits GPUs , by using kernel fusion, parallel scan , and recomputation. [2] The implementation avoids materializing expanded states in memory-intensive layers, thereby improving performance and memory usage. The result is significantly more efficient in processing long sequences compared to transformers . [2] [7]

Additionally, Mamba simplifies its architecture by integrating the SSM design with MLP blocks, resulting in a homogeneous and streamlined structure, furthering the model's capability for general sequence modeling across data types that include language, audio, and genomics, while maintaining efficiency in both training and inference. [2]

Key components

Selective-State-Spaces (SSM): The core of Mamba, SSMs are recurrent models that selectively process information based on the current input. This allows them to focus on relevant information and discard irrelevant data. [2]

Simplified Architecture: Mamba replaces the complex attention and MLP blocks of Transformers with a single, unified SSM block. This aims to reduce computational complexity and improve inference speed. [2]

Hardware-Aware Parallelism: Mamba utilizes a recurrent mode with a parallel algorithm specifically designed for hardware efficiency, potentially further enhancing its performance. [2]

Variants

Token-free language models: MambaByte

Operating on byte-sized tokens, transformers scale poorly as every token must "attend" to every other token leading to $O(n^2)$ scaling laws, as a result, Transformers opt to use subword tokenization to reduce the number of tokens in text, however, this leads to very large vocabulary tables and word embeddings .

This research investigates a novel approach to language modeling, MambaByte, which departs from the standard token-based methods. Unlike traditional models that rely on breaking text into discrete units, MambaByte directly processes raw byte sequences. This eliminates the need for tokenization, potentially offering several advantages: [8]

Language Independence: Tokenization often relies on language-specific rules and vocabulary, limiting applicability across diverse languages. MambaByte's byte-level representation allows it to handle different languages without language-specific adaptations.

Removes the bias of subword tokenisation: where common subwords are overrepresented and rare or new words are underrepresented or split into less meaningful units. This can affect the model's understanding and generation capabilities, particularly for languages with rich morphology or tokens not well-represented in the training data.

Simplicity in Preprocessing : It simplifies the preprocessing pipeline by eliminating the need for complex tokenization and vocabulary management, reducing the preprocessing steps and potential errors.

Subword tokenisation introduces a number of quirks in LLMs, such as failure modes where LLMs can't spell words, reverse certain words, handle rare tokens, which are not present in byte-level tokenisation. [9]

Mamba Mixture of Experts (MOE)

MoE Mamba represents a pioneering integration of the Mixture of Experts (MoE) technique with the Mamba architecture, enhancing the efficiency and scalability of State Space Models (SSMs) in language modeling. This model leverages the strengths of both MoE and SSMs, achieving significant gains in training efficiency—requiring 2.2 times fewer training steps than its predecessor, Mamba, while maintaining competitive performance. MoE Mamba showcases improved efficiency and effectiveness by combining selective state space modeling with expert-based processing, offering a promising avenue for future research in scaling SSMs to handle tens of billions of parameters. The model's design involves alternating Mamba and MoE layers, allowing it to efficiently integrate the entire sequence context and apply the most relevant expert for each token. [10] [11]

Vision Mamba

Vision Mamba (Vim) integrates SSMs with visual data processing, employing bidirectional Mamba blocks for visual sequence encoding. This method reduces the computational demands typically associated with self-attention in visual tasks. Tested on ImageNet classification, COCO object detection, and ADE20k semantic segmentation, Vim showcases enhanced performance and efficiency and is capable of handling high-resolution images with lower computational resources. This positions Vim as a scalable model for future advancements in visual representation learning. [12]

Jamba

Jamba is a novel architecture built on a hybrid transformer and mamba SSM architecture developed by AI21 Labs with 52 billion parameters, making it the largest Mamba-variant created so far. It has a context window of 256k tokens. [13]

Impact and Future Directions

Mamba LLM represents a significant potential shift in large language model architecture, offering faster, more efficient, and scalable models [citation needed] .

Applications include language translation, content generation, long-form text analysis, audio, and speech processing [citation needed] .

See also

Language modeling

Transformer (machine learning model)

State-space model

Recurrent neural network

Notes

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