Title: GPT-1

URL: https://en.wikipedia.org/wiki/GPT-1

PageID: 68456032

Categories: Category:2018 in artificial intelligence, Category:2018 software, Category:Generative pre-trained transformers, Category:Large language models, Category:OpenAI, Category:Software using the MIT license

---

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

Generative Pre-trained Transformer 1 (GPT-1) was the first of OpenAI 's large language models following Google 's invention of the transformer architecture in 2017. [2] In June 2018, OpenAI released a paper entitled "Improving Language Understanding by Generative Pre-Training", [3] in which they introduced that initial model along with the general concept of a generative pre-trained transformer . [4]

Up to that point, the best-performing neural NLP models primarily employed supervised learning from large amounts of manually labeled data. This reliance on supervised learning limited their use of datasets that were not well-annotated, in addition to making it prohibitively expensive and time-consuming to train extremely large models; [3] [5] many languages (such as Swahili or

Haitian Creole) are difficult to translate and interpret using such models due to a lack of available text for corpus-building. [5] In contrast, a GPT's "semi-supervised" approach involved two stages: an unsupervised generative "pre-training" stage in which a language modeling objective was used to set initial parameters, and a supervised discriminative "fine-tuning" stage in which these parameters were adapted to a target task. [3]

The use of a transformer architecture, as opposed to previous techniques involving attention-augmented RNNs, provided GPT models with a more structured memory than could be achieved through recurrent mechanisms; this resulted in "robust transfer performance across diverse tasks". [3]

Architecture

The GPT-1 architecture is a twelve-layer decoder-only transformer , using twelve masked self-attention heads, with 64-dimensional states each (for a total of 768). Rather than simple stochastic gradient descent , the Adam optimization algorithm was used; the learning rate was increased linearly from zero over the first 2,000 updates to a maximum of 2.5×10^{-4} , and annealed to 0 using a cosine schedule. [3] GPT-1 has 117 million parameters. [4]

While the fine-tuning was adapted to specific tasks, its pre-training was not; to perform the various tasks, minimal changes were performed to its underlying task-agnostic model architecture. [3] Despite this, GPT-1 still improved on previous benchmarks in several language processing tasks, outperforming discriminatively-trained models with task-oriented architectures on several diverse tasks. [3]

Performance and evaluation

GPT-1 achieved a 5.8% and 1.5% improvement over previous best results [3] on natural language inference (also known as textual entailment) tasks, evaluating the ability to interpret pairs of sentences from various datasets and classify the relationship between them as "entailment", "contradiction" or "neutral". [3] Examples of such datasets include QNLI (Wikipedia articles) and MultiNLI (transcribed speech, popular fiction, and government reports, among other sources); [6] It similarly outperformed previous models on two tasks related to question answering and commonsense reasoning —by 5.7% on RACE, [7] a dataset of written question-answer pairs from middle and high school exams, and by 8.9% on the Story Cloze Test . [8]

GPT-1 improved on previous best-performing models by 4.2% on semantic similarity (or paraphrase detection), evaluating the ability to predict whether two sentences are paraphrases of one another, using the Quora Question Pairs (QQP) dataset. [3]

GPT-1 achieved a score of 45.4, versus a previous best of 35.0 [3] in a text classification task using the Corpus of Linguistic Acceptability (CoLA). Finally, GPT-1 achieved an overall score of 72.8 (compared to a previous record of 68.9) on GLUE, a multi-task test. [9]

See also

List of large language models

References

v

t

e

ChatGPT in education GPT Store DALL-E ChatGPT Search Sora Whisper

in education

GPT Store

DALL-E

ChatGPT Search

Sora

Whisper

GitHub Copilot

OpenAI Codex

Generative pre-trained transformer GPT-1 GPT-2 GPT-3 GPT-4 GPT-4o o1 o3 GPT-4.5 GPT-4.1 o4-mini GPT-OSS GPT-5

GPT-1

GPT-2

GPT-3

GPT-4

GPT-4o

o1

o3

GPT-4.5

GPT-4.1

o4-mini

GPT-OSS

GPT-5

ChatGPT Deep Research

Operator

Sam Altman removal

removal

Greg Brockman

Sarah Friar

Jakub Pachocki

Scott Schools

Mira Murati

Emmett Shear

Sam Altman

Adam D'Angelo

Sue Desmond-Hellmann

Zico Kolter

Paul Nakasone

Adebayo Ogunlesi

Nicole Seligman

Fidji Simo

Lawrence Summers

Bret Taylor (chair)

Greg Brockman (2017–2023)

Reid Hoffman (2019–2023)

Will Hurd (2021–2023)

Holden Karnofsky (2017–2021)

Elon Musk (2015–2018)

Ilya Sutskever (2017–2023)

Helen Toner (2021–2023)

Shivon Zilis (2019–2023)

Stargate LLC

Apple Intelligence

AI Dungeon

AutoGPT

Contrastive Language-Image Pre-training

" Deep Learning "

LangChain

Microsoft Copilot

OpenAI Five

Transformer

Category

v

t

e

Autoencoder

Deep learning

Fine-tuning

Foundation model

Generative adversarial network

Generative pre-trained transformer

Large language model

Model Context Protocol

Neural network

Prompt engineering

Reinforcement learning from human feedback

Retrieval-augmented generation

Self-supervised learning

Stochastic parrot

Synthetic data

Top-p sampling

Transformer

Variational autoencoder
Vibe coding
Vision transformer
Waluigi effect
Word embedding
Character.ai
ChatGPT
DeepSeek
Ernie
Gemini
Grok
Copilot
Claude
Gemini
Gemma
GPT 1 2 3 J 4 4o 4.5 4.1 OSS 5
1
2
3
J
4
4o
4.5
4.1
OSS
5
Llama
o1
o3
o4-mini
Qwen
Base44
Claude Code
Cursor
Devstral
GitHub Copilot
Kimi-Dev
Qwen3-Coder

Replit
Xcode
Aurora
Firefly
Flux
GPT Image 1
Ideogram
Imagen
Midjourney
Qwen-Image
Recraft
Seedream
Stable Diffusion
Dream Machine
Hailuo AI
Kling
Midjourney Video
Runway Gen
Seedance
Sora
Veo
Wan
15.ai
Eleven
MiniMax Speech 2.5
WaveNet
Eleven Music
Endel
Lyria
Riffusion
Suno AI
Udio
Agentforce
AutoGLM
AutoGPT
ChatGPT Agent
Devin AI
Manus

OpenAI Codex
Operator
Replit Agent
01.AI
Aleph Alpha
Anthropic
Baichuan
Canva
Cognition AI
Cohere
Contextual AI
DeepSeek
ElevenLabs
Google DeepMind
HeyGen
Hugging Face
Inflection AI
Krikey AI
Kuaishou
Luma Labs
Meta AI
MiniMax
Mistral AI
Moonshot AI
OpenAI
Perplexity AI
Runway
Safe Superintelligence
Salesforce
Scale AI
SoundHound
Stability AI
Synthesia
Thinking Machines Lab
Upstage
xAI
Z.ai
Category

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