

Main page Contents Current events Random article About Wikipedia Contact us

Contribute

Help Learn to edit Community portal Recent changes Upload file

Tools

What links here Related changes Special pages Permanent link Page information Cite this page Wikidata item

Print/export

Download as PDF

Languages

Español Euskara فارسی

Français 日本語 Українська

Tiếng Việt

中文

Edit links

Article Talk

Read

Edit View history

Not logged

Search Wikipedia

Talk Contributions Create account Log

BERT (language model)

From Wikipedia, the free encyclopedia

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique for natural language processing (NLP) pre-training developed by Google. BERT was created and published in 2018 by Jacob Devlin and his colleagues from Google. [1][2] In 2019, Google announced that it had begun leveraging BERT in its search engine, and by late 2020 it was using BERT in almost every English-language query. A 2020 literature survey concluded that "in a little over a year, BERT has become a ubiquitous baseline in NLP experiments", counting over 150 research publications analyzing and improving the model. [3]

The original English-language BERT has two models: $^{[1]}$ (1) the BERT_{BASE}: 12 Encoders with 12 bidirectional self-attention heads, and (2) the BERT_{LARGE}: 24 Encoders with 16 bidirectional self-attention heads. Both models are pre-trained from unlabeled data extracted from the BooksCorpus^[4] with 800M words and English Wikipedia with 2,500M words. $^{[5]}$

Contents [hide]

- 1 Architecture
- 2 Performance
- 3 Analysis
- 4 History
- 5 Recognition
- 6 See also
- 7 References
- 8 Further reading
- 9 External links

Architecture [edit]

BERT is at its core a Transformer language model with variable number of encoder layers and self-attention heads. The architecture is "almost identical" to the original Transformer implementation in Vaswani et al. (2017).^[6]

BERT was pretrained on two tasks: **language modelling** (15% of tokens were masked and BERT was trained to predict them from context) and **next sentence prediction** (BERT was trained to predict if a chosen next sentence was probable or not given the first sentence). As a result of the training process, BERT learns contextual embeddings for words. After pretraining, which is computationally expensive, BERT can be finetuned with less resources on smaller datasets to optimize its performance on specific tasks.^{[1][7]}

Performance [edit]

When BERT was published, it achieved state-of-the-art performance on a number of natural language understanding tasks:^[1]

- GLUE (General Language Understanding Evaluation) task set (consisting of 9 tasks)
- SQuAD (Stanford Question Answering Dataset) v1.1 and v2.0
- SWAG (Situations With Adversarial Generations)

Analysis [edit]

The reasons for BERT's state-of-the-art performance on these natural language understanding tasks are not yet well understood. [8][9] Current research has focused on investigating the relationship behind BERT's output as a result of carefully chosen input sequences, [10][11] analysis of internal vector representations through probing classifiers, [12][13] and the relationships represented by attention weights. [8][9]

History [edit]

BERT has its origins from pre-training contextual representations including Semi-supervised Sequence Learning, [14] Generative Pre-Training, ELMo, [15] and ULMFit. [16] Unlike previous models, BERT is a deeply

bidirectional, unsupervised language representation, pre-trained using only a plain text corpus. Context-free models such as word2vec or GloVe generate a single word embedding representation for each word in the vocabulary, where BERT takes into account the context for each occurrence of a given word. For instance, whereas the vector for "running" will have the same word2vec vector representation for both of its occurrences in the sentences "He is running a company" and "He is running a marathon", BERT will provide a contextualized embedding that will be different according to the sentence.

On October 25, 2019, Google Search announced that they had started applying BERT models for English language search queries within the US.^[17] On December 9, 2019, it was reported that BERT had been adopted by Google Search for over 70 languages.^[18] In October 2020, almost every single English-based query was processed by BERT.^[19]

Recognition [edit]

BERT won the Best Long Paper Award at the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).^[20]

See also [edit]

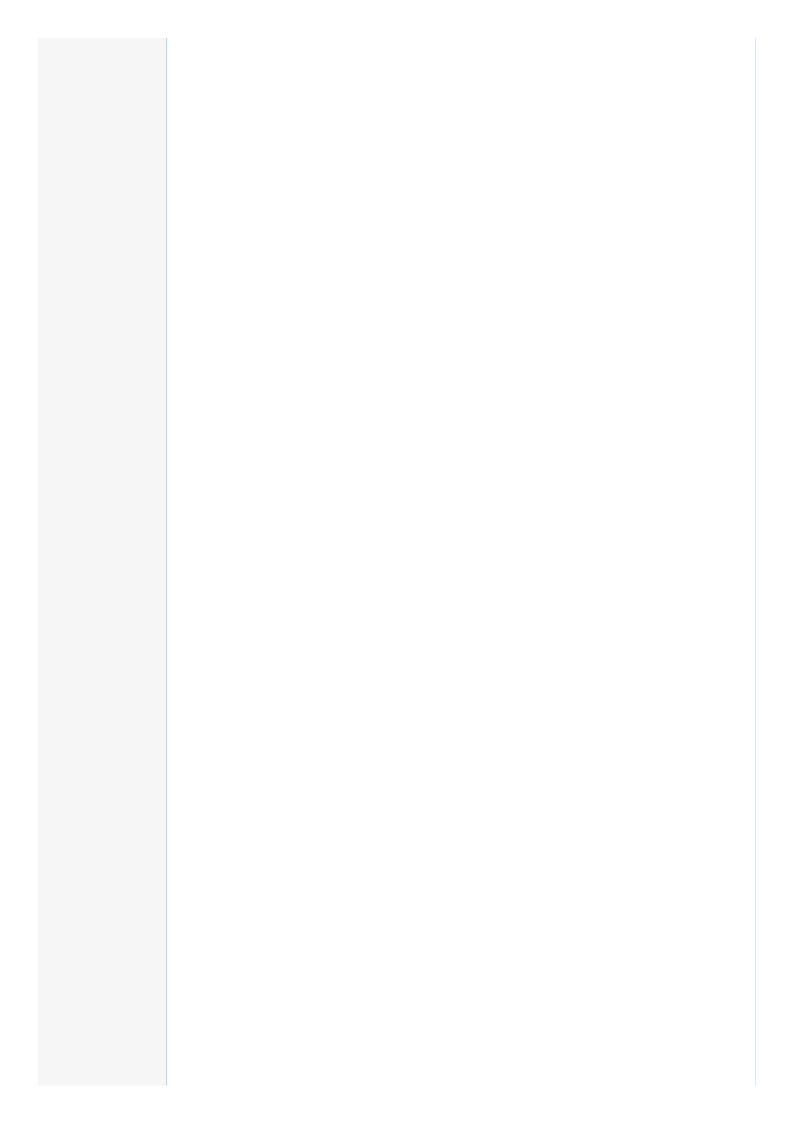
- Transformer (machine learning model)
- Word2vec
- Autoencoder
- Document-term matrix
- Feature extraction
- Feature learning
- Neural network language models
- Vector space model
- Thought vector
- fastText
- GloVe
- TensorFlow

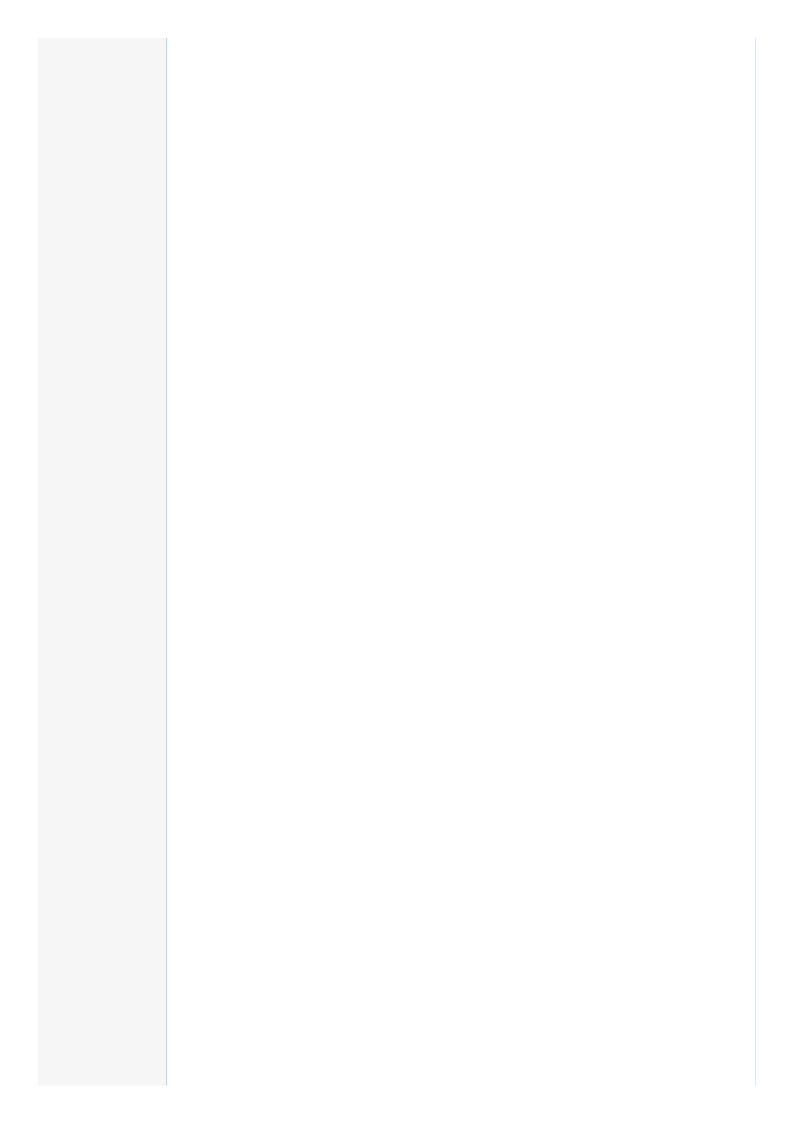
References [edit]

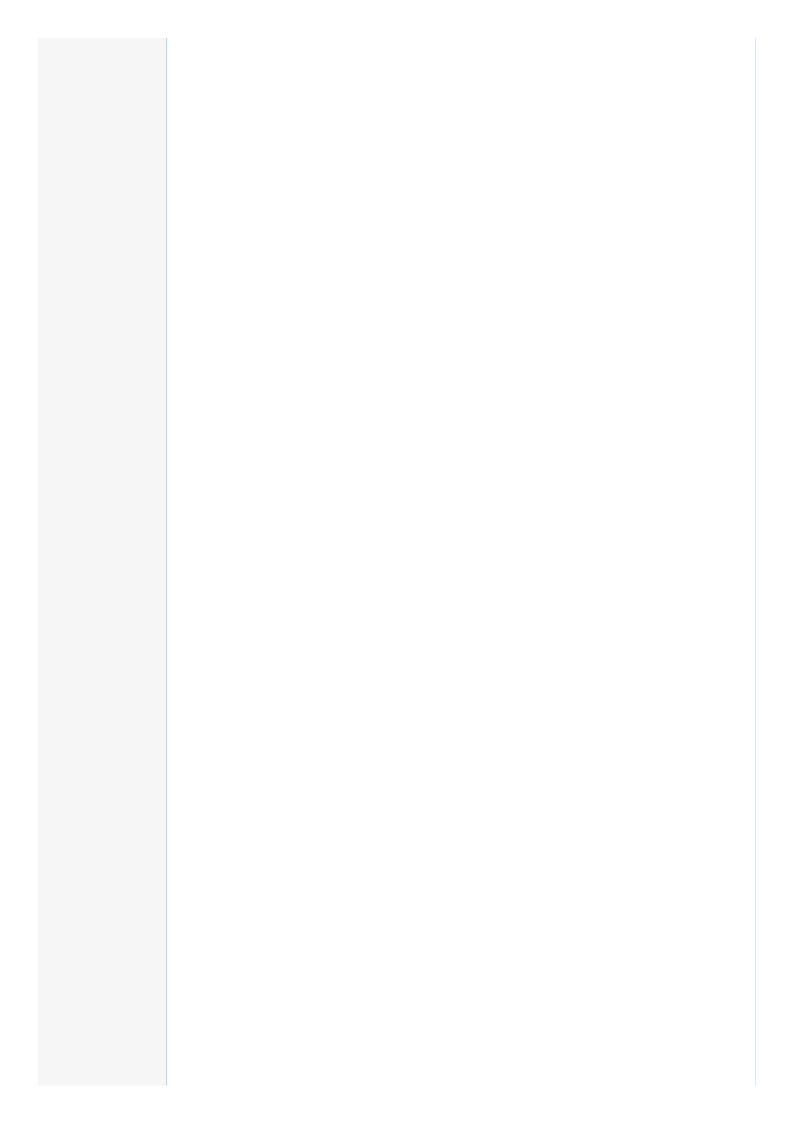
- 1. ^ a b c d Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina (11 October 2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". arXiv:1810.04805v2 ☑ [cs.CL☑].
- 2. A "Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing" 函. *Google Al Blog*. Retrieved 2019-11-27.
- 3. ^ Rogers, Anna; Kovaleva, Olga; Rumshisky, Anna (2020). "A Primer in BERTology: What We Know About How BERT Works" 다. Transactions of the Association for Computational Linguistics . 8: 842–866.
- 4. ^ Zhu, Yukun; Kiros, Ryan; Zemel, Rich; Salakhutdinov, Ruslan; Urtasun, Raquel; Torralba, Antonio; Fidler, Sanja (2015). "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books". pp. 19–27. arXiv:1506.06724 [cs.CV].
- 5. ^ Annamoradnejad, Issa (2020-04-27). "ColBERT: Using BERT Sentence Embedding for Humor Detection". arXiv:2004.12765 당 [cs.CL당].
- 7. ^ Horev, Rani (2018). "BERT Explained: State of the art language model for NLP" ☑. Towards Data Science.
 Retrieved 27 September 2021.
- 8. ^ a b Kovaleva, Olga; Romanov, Alexey; Rogers, Anna; Rumshisky, Anna (November 2019). "Revealing the Dark Secrets of BERT" 已. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 4364–4373. doi:10.18653/v1/D19-1445 已. S2CID 201645145 已.

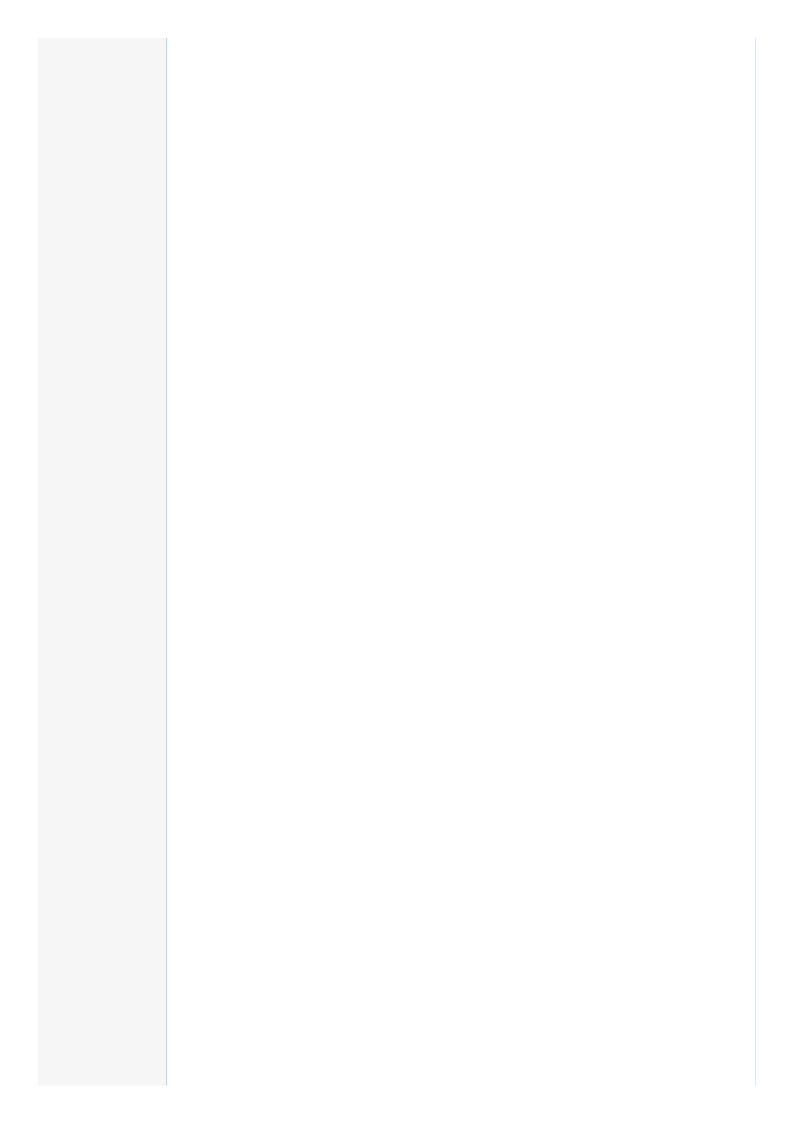
- 11. A Gulordava, Kristina; Bojanowski, Piotr; Grave, Edouard; Linzen, Tal; Baroni, Marco (2018). "Colorless Green Recurrent Networks Dream Hierarchically". *Proceedings of the 2018 Conference of the North American Chapter of*

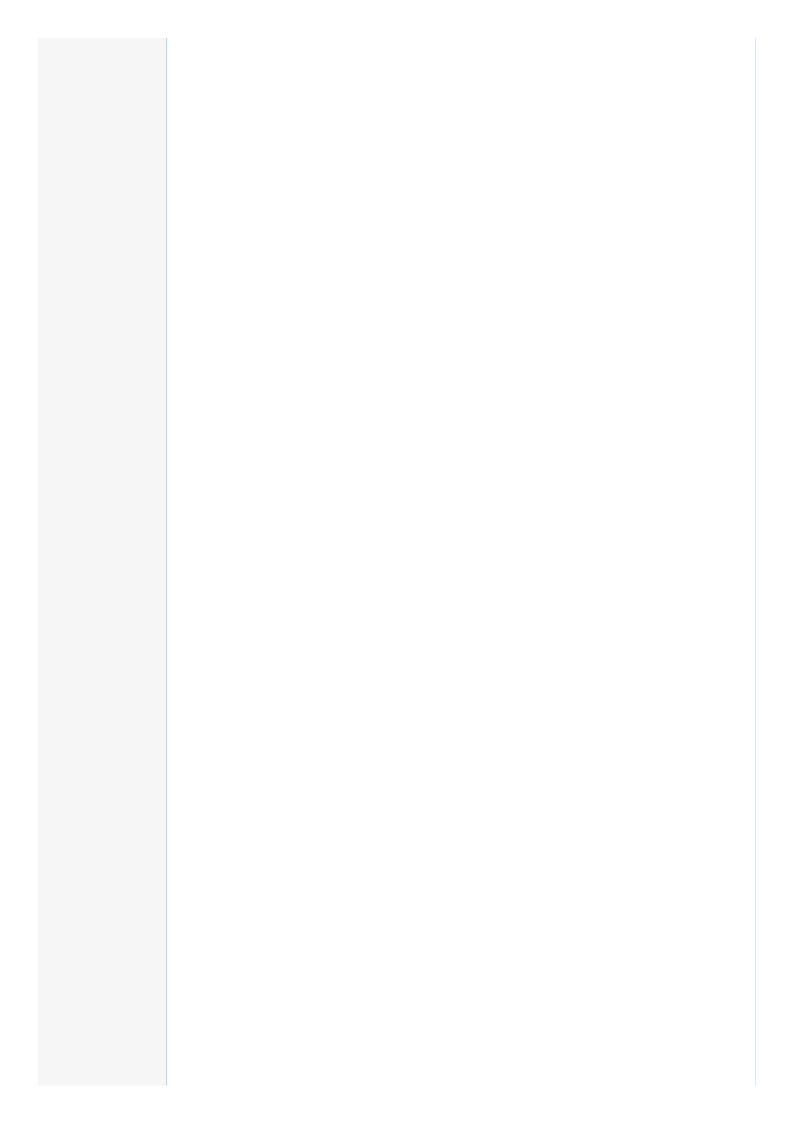
- 12. ^ Giulianelli, Mario; Harding, Jack; Mohnert, Florian; Hupkes, Dieuwke; Zuidema, Willem (2018). "Under the Hood: Using Diagnostic Classifiers to Investigate and Improve how Language Models Track Agreement Information".
 Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP.
 Stroudsburg, PA, USA: Association for Computational Linguistics: 240–248. arXiv:1808.08079 母.
 Bibcode:2018arXiv180808079G 母. doi:10.18653/v1/w18-5426母. S2CID 52090220母.
- 13. ^ Zhang, Kelly; Bowman, Samuel (2018). "Language Modeling Teaches You More than Translation Does: Lessons Learned Through Auxiliary Syntactic Task Analysis" 母. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Stroudsburg, PA, USA: Association for Computational Linguistics: 359–361. doi:10.18653/v1/w18-5448 母.
- 14. ^ Dai, Andrew; Le, Quoc (4 November 2015). "Semi-supervised Sequence Learning". arXiv:1511.01432 년 [cs.LG 답].
- 15. ^ Peters, Matthew; Neumann, Mark; Iyyer, Mohit; Gardner, Matt; Clark, Christopher; Lee, Kenton; Luke, Zettlemoyer (15 February 2018). "Deep contextualized word representations". arXiv:1802.05365v2 [[cs.CL].]
- 16. ^ Howard, Jeremy; Ruder, Sebastian (18 January 2018). "Universal Language Model Fine-tuning for Text Classification". arXiv:1801.06146v5 ☐ [cs.CL].
- 17. ^ Nayak, Pandu (25 October 2019). "Understanding searches better than ever before" ☑. Google Blog. Retrieved 10 December 2019.
- 18. ^ Montti, Roger (10 December 2019). "Google's BERT Rolls Out Worldwide" ②. Search Engine Journal. Search Engine Journal. Retrieved 10 December 2019.
- 19. ^ "Google: BERT now used on almost every English query" ☑. Search Engine Land. 2020-10-15. Retrieved 2020-11-24.
- 20. ^ "Best Paper Awards" 2. NAACL. 2019. Retrieved Mar 28, 2020.











Further reading [edit]

• Rogers, Anna; Kovaleva, Olga; Rumshisky, Anna (2020). "A Primer in BERTology: What we know about how BERT works". arXiv:2002.12327 ☑ [cs.CL ☑].

External links [edit]

v·t·e	Na	atural language proc	ess	ing
General terms		Al-complete · Bag-of-words · n-gram (Bigram · Trigram) · Computational linguistics · Natural-language understanding · Stopwords · Text processing		
Text analysis		Collocation extraction · Concept mining · Coreference resolution · Deep linguistic processing · Distant reading · Information extraction · Named-entity recognition · Ontology learning · Parsing · Part-of-speech tagging · Semantic role labeling · Semantic similarity · Sentiment analysis · Terminology extraction · Text mining · Textual entailment · Truecasing · Word-sense disambiguation · Word-sense induction Compound-term processing · Lemmatisation · Lexical analysis · Text chunking · Stemming · Sentence segmentation · Word segmentation		
Automati	c summarization	Multi-document summarization · Sentence extraction · Text simplification		
Machine translation		Computer-assisted • Example-based • Rule-based • Statistical • Transfer-based • Neural		
Distributional semantics models		BERT · Document-term matrix · Explicit semantic analysis · fastText · GloVe · Latent semantic analysis · Word embedding · Word2vec		
Language resources, datasets and corpora		Types and standards	L M F S	Corpus linguistics · Lexical resource · Linguistic Linked Open Data · Machine-readable dictionary · Parallel text · PropBank · Semantic network · Simple Knowledge Organization System · Speech corpus · Text corpus · Thesaurus (information retrieval) · Treebank · Jniversal Dependencies
		Data	(BabelNet · Bank of English · DBpedia · FrameNet · Google Ngram Viewer · ThoughtTreasure · UBY · WordNet
Automatic identification and data capture		Speech recognition · Speech segmentation · Speech synthesis · Natural language generation · Optical character recognition		
Topic model		Document classification · Latent Dirichlet allocation · Pachinko allocation		
Computer-assisted reviewing		Automated essay scoring · Concordancer · Grammar checker · Predictive text · Spell checker · Syntax guessing		
Natural language user interface		Chatbot · Interactive fiction · Question answering · Virtual assistant · Voice user interface		
Other software		Natural Language Toolkit · spaCy		
v·t·e		Differentiable comp	utin	g
General	Automatic differen	Differentiable programming · Neural Turing machine · Differentiable neural computer · Automatic differentiation · Neuromorphic engineering · Cable theory · Pattern recognition · Computational learning theory · Tensor calculus		
Concepts	Gradient descent (SGD) · Clustering · Regression (Overfitting) · Adversary · Attention · Convolution · Loss functions · Backpropagation · Normalization · Activation (Softmax · Sigmoid · Rectifier) · Regularization · Datasets (Augmentation)			
Programming languages	Python · Julia			

Application	Machine learning · Artificial neural network (Deep learning) · Scientific computing · Artificial Intelligence			
Hardware	IPU · TPU · VPU · Memristor · SpiNNaker			
Software library	TensorFlow · PyTorch · Keras · Theano			
	Audio-visual	AlexNet · WaveNet · Human image synthesis · HWR · OCR · Speech synthesis · Speech recognition · Facial recognition · AlphaFold · DALL-E		
Implementation	Verbal			
	Decisional	AlphaGo · AlphaZero · Q-learning · SARSA · OpenAl Five · Self-driving car · MuZero · Action selection · Robot control		
People	Alex Graves · Ian Goodfellow · Yoshua Bengio · Geoffrey Hinton · Yann LeCun · Andrew Ng · Demis Hassabis · David Silver · Fei-Fei Li			
Organizations	DeepMind · OpenAl · MIT CSAIL · Mila · Google Brain · FAIR			
⊕ Portals (Computer programming • Technology) • ⊕ Category (Artificial neural networks • Machine learning)				

Categories: Natural language processing | Computational linguistics | Speech recognition | Computational fields of study | Artificial intelligence

This page was last edited on 1 November 2021, at 03:20 (UTC).

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.

Privacy policy About Wikipedia Disclaimers Contact Wikipedia Mobile view Developers Statistics Cookie statement



