Workshop on Machine Learning on Low-Power Devices: Applications and Advanced Topics

AI and ML Overview

Prof. Jesús Alfonso López Sotelo
jalopez@uao.edu.co

UAO - Universidad Autónoma de Occidente, Cali, Colombia www.uao.edu.co

Jesús Alfonso López Sotelo

Born in Cali, Colombia. He is an Electrical Engineer, Master in Automation and Doctor in Engineering.

He has more than 25 years of experience in teaching and developing projects related to Artificial Intelligence. His areas of interest are artificial neural networks and deep learning (Deep Learning), Artificial Intelligence in edge devices, fuzzy systems, evolutionary computing, teaching artificial intelligence and the impact that this technology can have on our society.

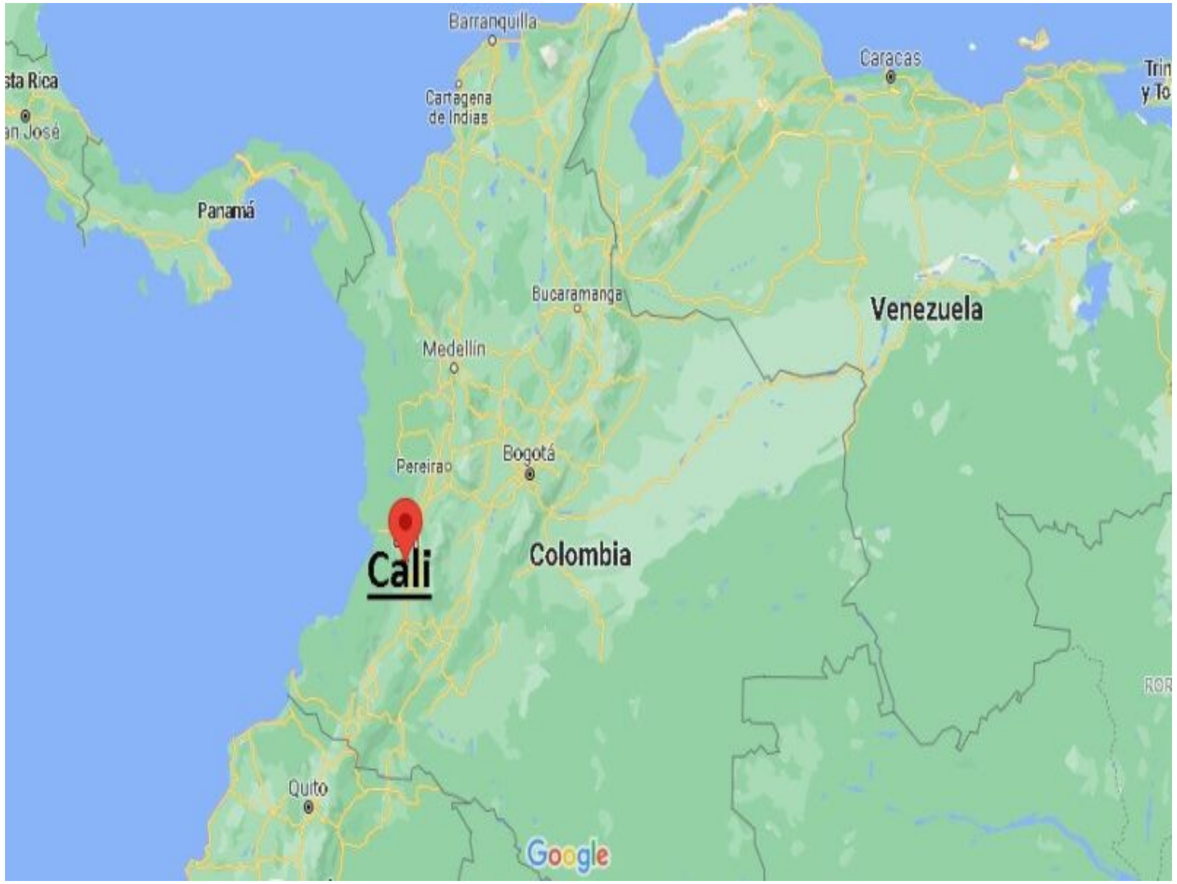
He is an Associate researcher of the national system of science, technology and innovation in Colombia of MinCiencias. He is a professional member of the IEEE where he belongs to the national chapter of the Computational Intelligence Society. He is currently linked to the Universidad Autónoma de Occidente in Cali and belongs to the Energy Research Group, GIEN. He has published various articles, book chapters and books on the topics of Artificial Neural Networks, Deep Learning and other artificial intelligence techniques.

Perfil Linkedin

https://www.linkedin.com/in/jesus-alfonso-lópez-sotelo-76100718/

Universidad Autónoma de Occidente https://www.uao.edu.co/

Cali Colombia



«If I have seen further than others, it is by standing upon the shoulders of giants».

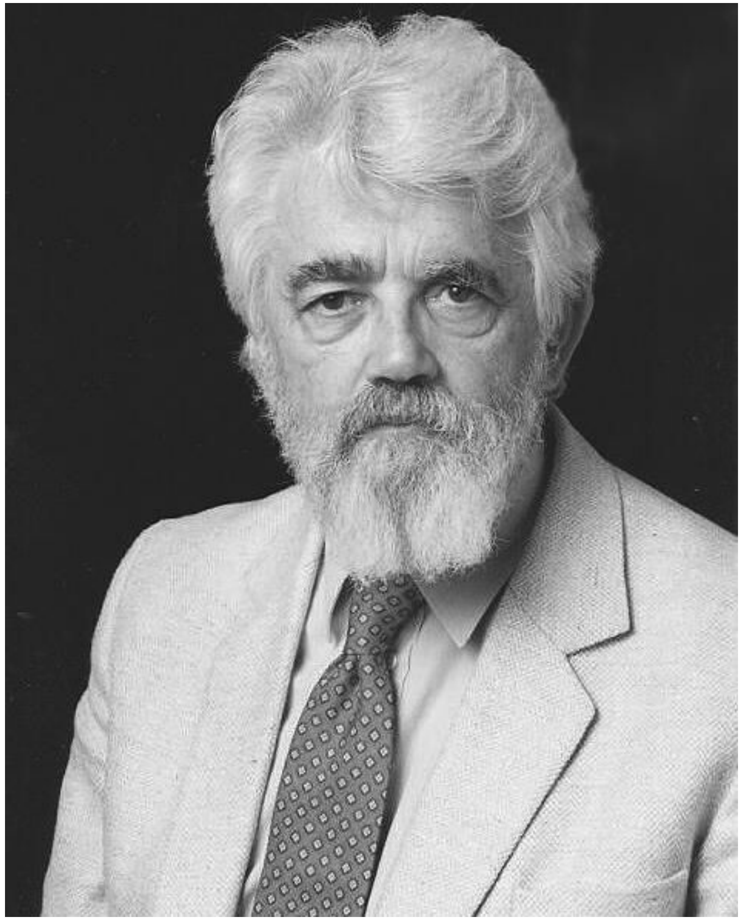
Isaac Newton

Artificial Intelligence (AI) and Machine Learning

Artificial Intelligence

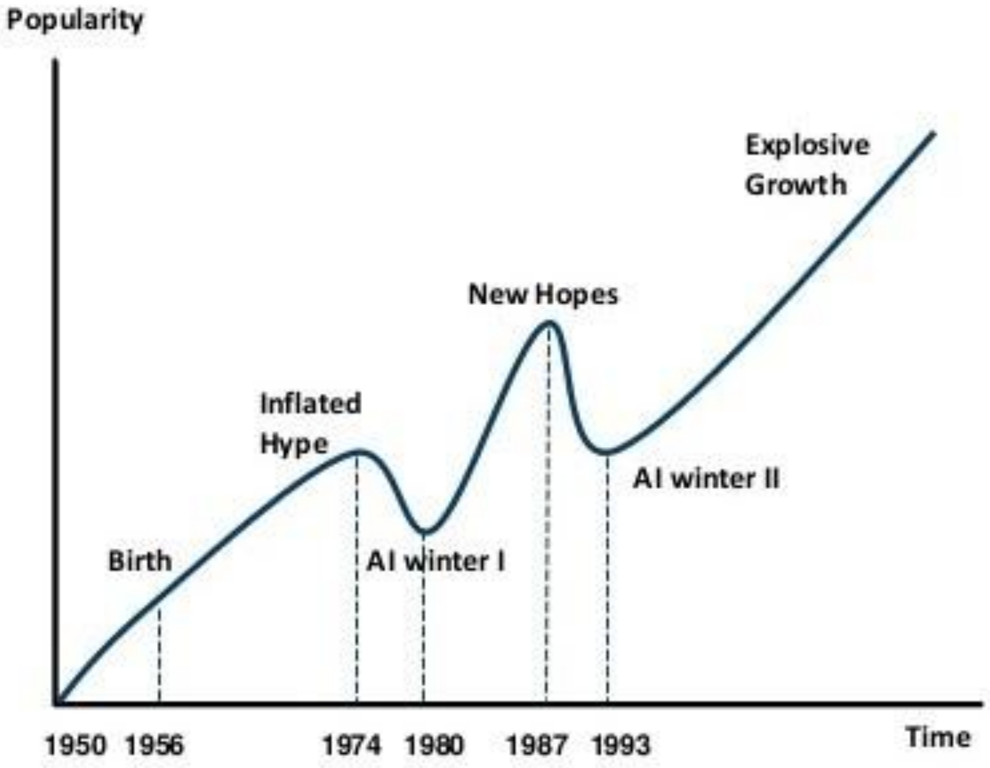
The Dartmouth Summer School on Artificial Intelligence (1956) is considered an important event in the history of AI and where the term artificial intelligence emerged, selected by computer scientist John McCarthy.

Artificial intelligence (AI) can be defined as the field of study and development of computer systems that can perform tasks that normally require human intelligence. These tasks include learning, perception, reasoning, problem solving, and natural language understanding.



https://francis.naukas.com/2011/10/25/john-mccarthy-1927-2011-el-padre-de-la-inteligencia-artificial/

Artificial Intelligence

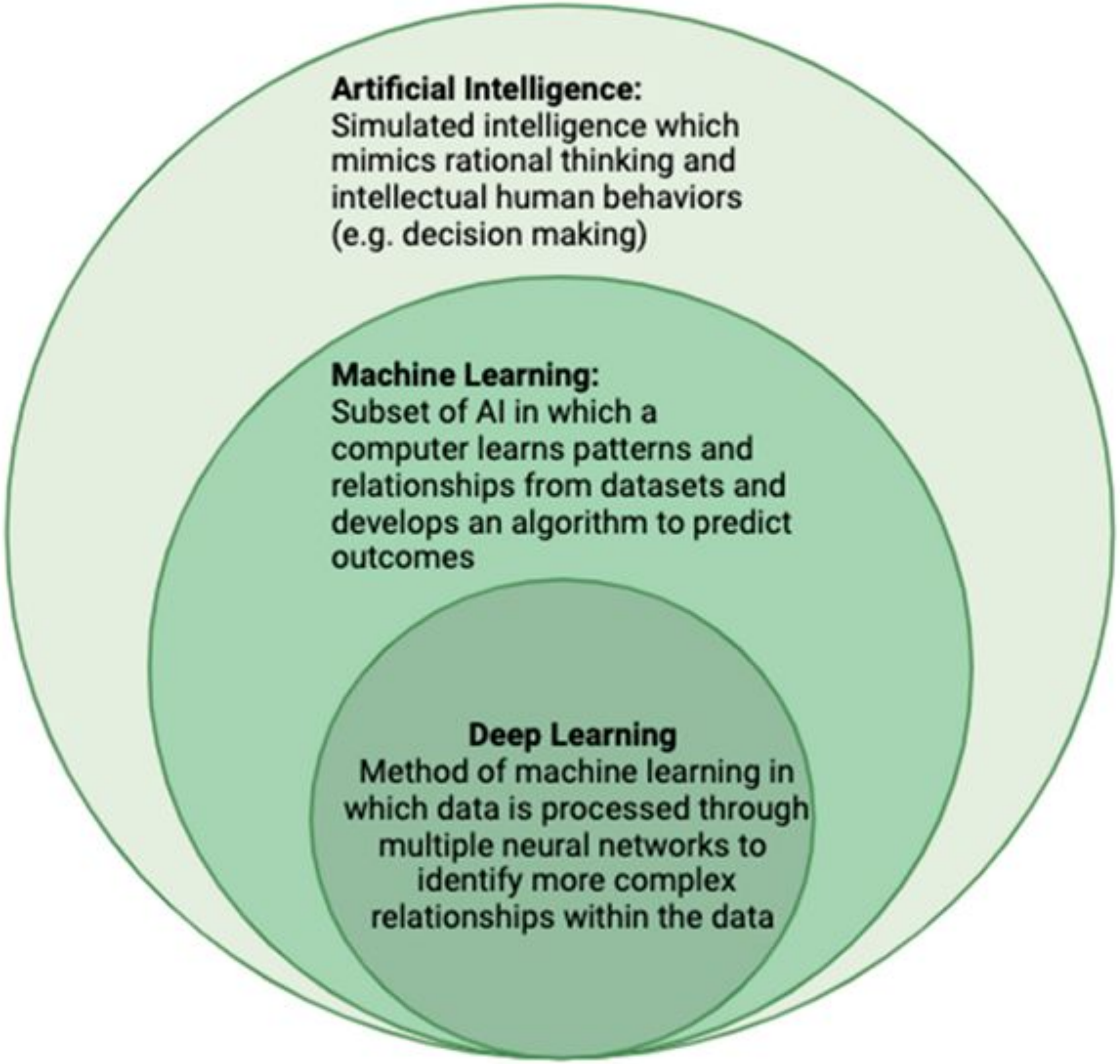


Timeline of AI Development

- **1950s-1960s**: First AI boom - the age of reasoning, prototype AI developed
- **1970s**: AI winter I
- **1980s-1990s**: Second AI boom: the age of Knowledge representation (appearance of expert systems capable of reproducing human decision-making)
- **1990s**: AI winter II
- **1997**: Deep Blue beats Gary Kasparov
- **2006**: University of Toronto develops Deep Learning
- **2011**: IBM's Watson won Jeopardy
- **2016**: Go software based on Deep Learning beats world's champions

https://www.actuaries.digital/2018/09/05/history-of-ai-winters/

Artificial Intelligence



Artificial Intelligence:
Simulated intelligence which mimics rational thinking and intellectual human behaviors (e.g. decision making)

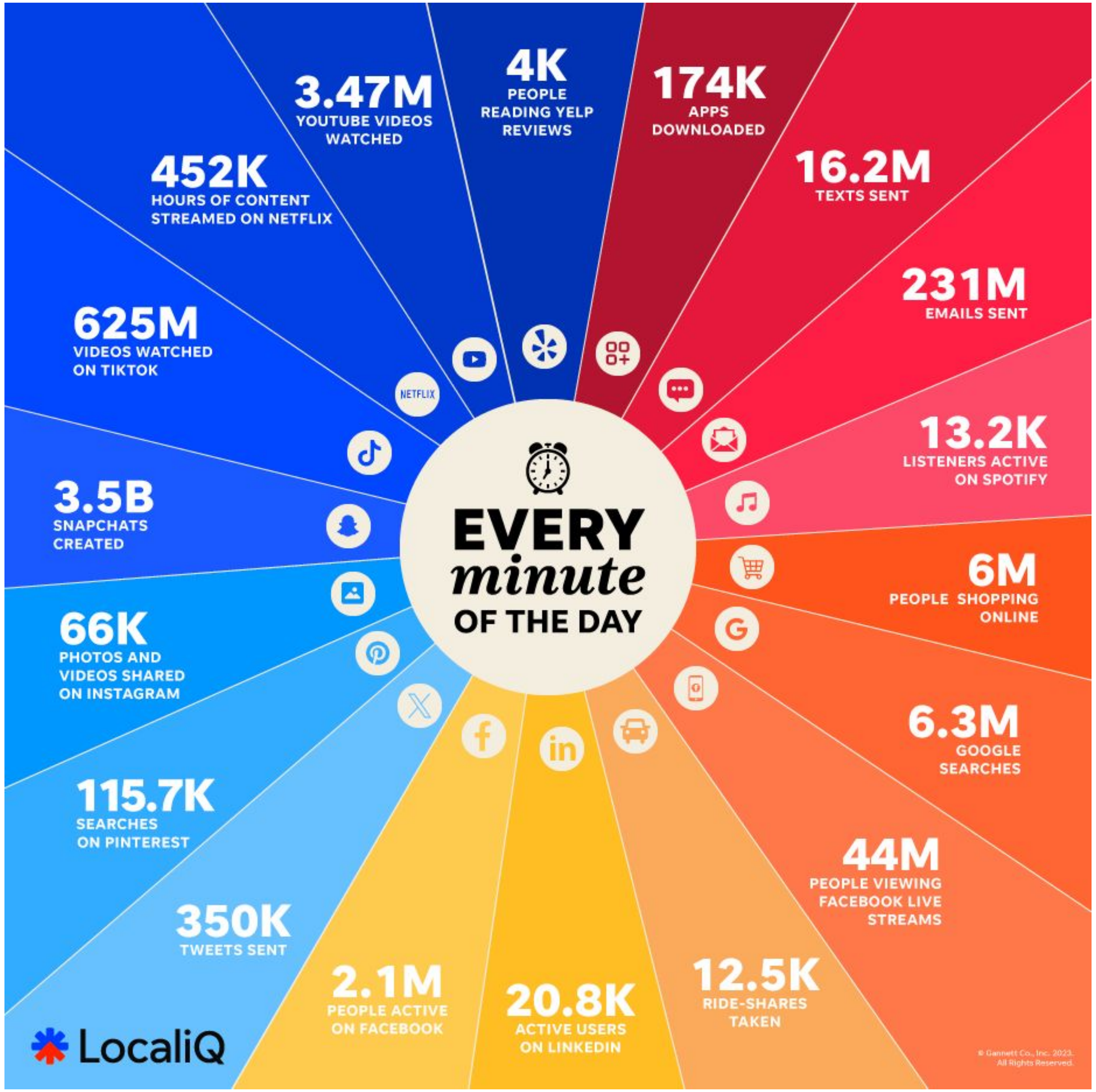
Machine Learning:
Subset of AI in which a computer learns patterns and relationships from datasets and develops an algorithm to predict outcomes

Deep Learning
Method of machine learning in which data is processed through multiple neural networks to identify more complex relationships within the data

Source: (Larrain, Torres-Hernandez, & Hewitt, 2024)

Machine Learning

Machine learning is a subset of artificial intelligence that has the ability to "learn" (i.e., progressively improve performance on a specific task) from data, without being explicitly programmed



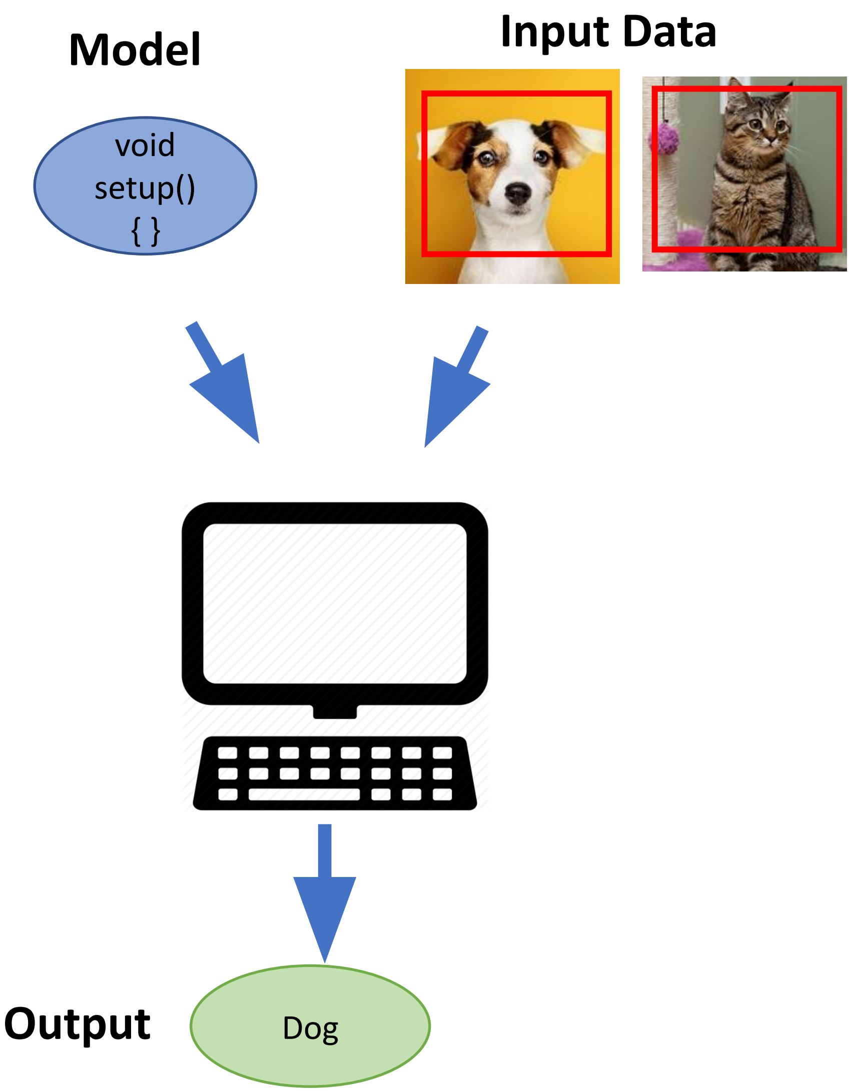
https://www.bondhighplus.com/2024/01/25/what-happens-in-an-internet-minute/

Machine Learning

People with no idea about AI, telling me my AI will destroy the world
Me wondering why my neural network is classifying a cat as a dog..
Dog

Dogs and cats recognizer

Machine Learning



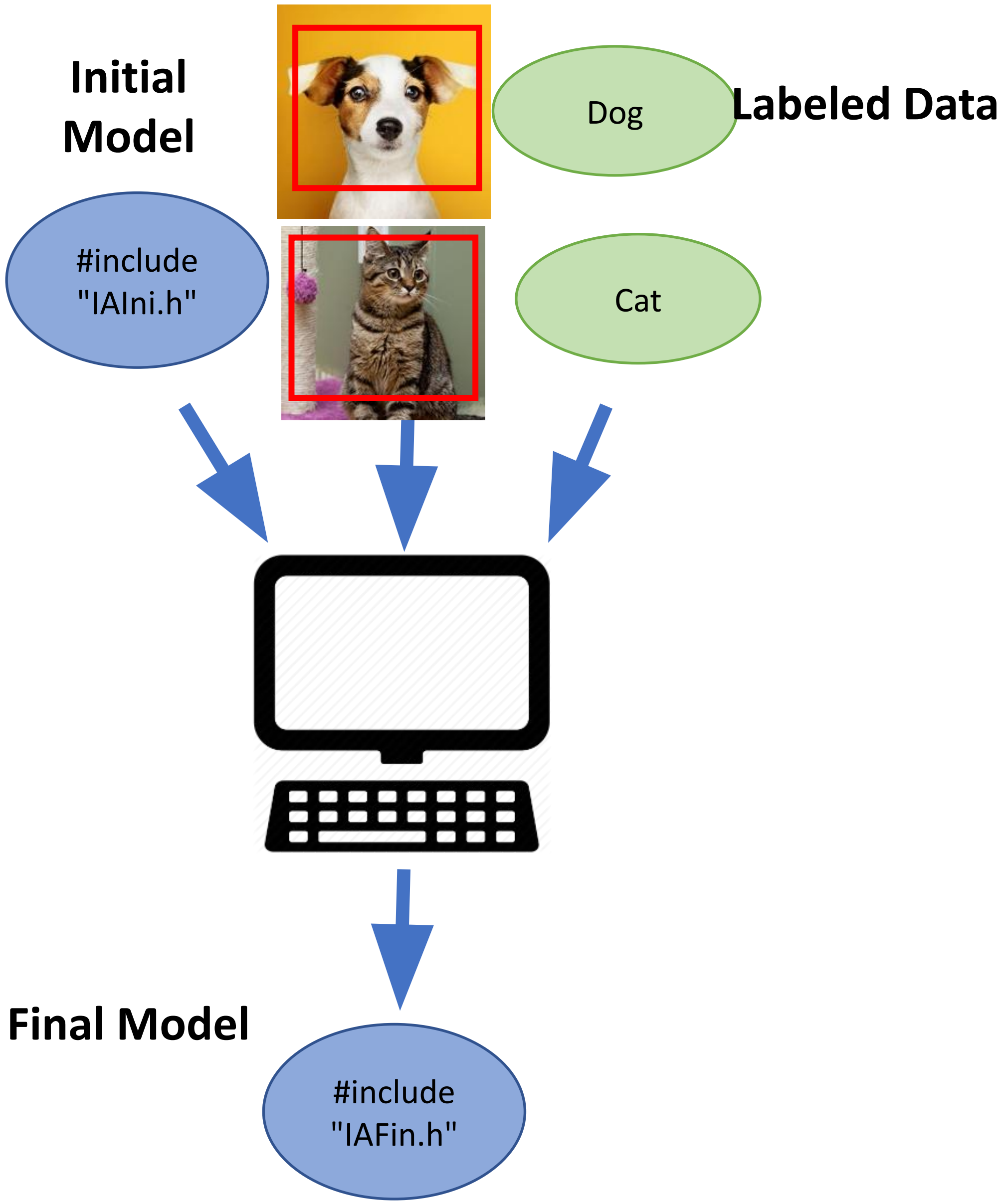
Model

```
void
setup()
{ }
```

Input Data

Output

Dog



Initial Model

```
#include
"IAIni.h"
```

Labeled Data

Dog

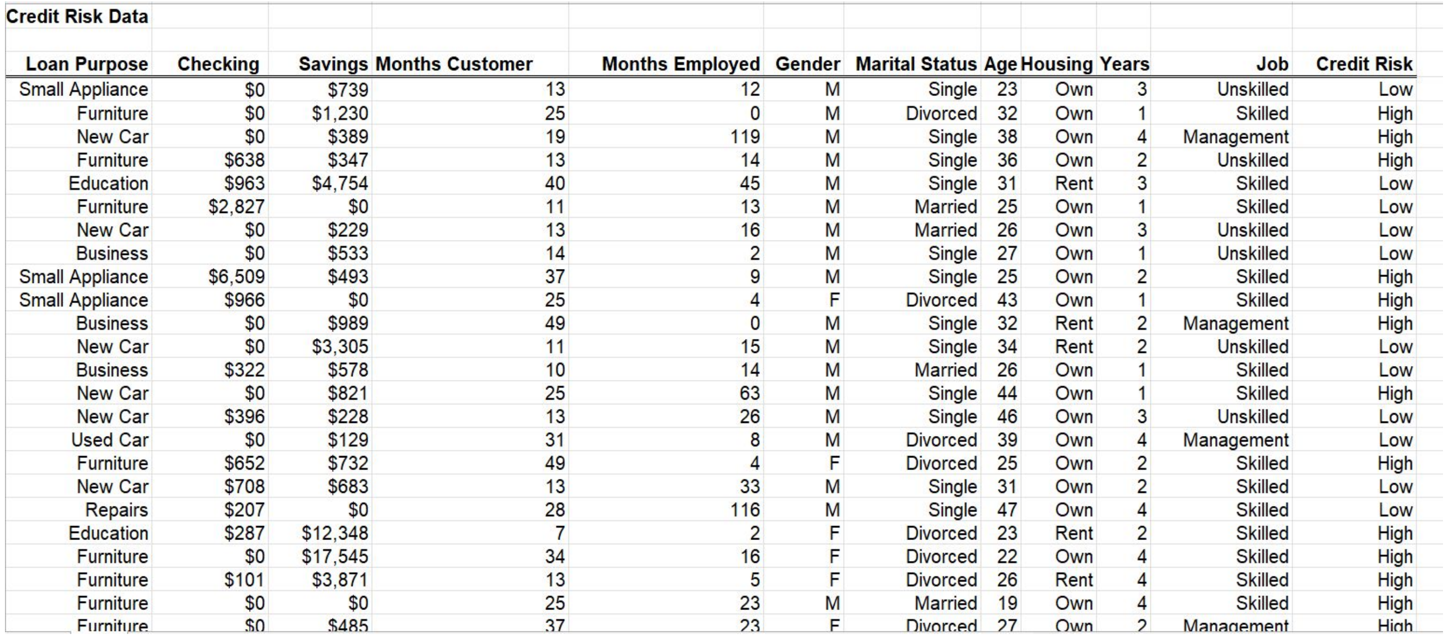
Cat

Final Model

```
#include
"IAFin.h"
```

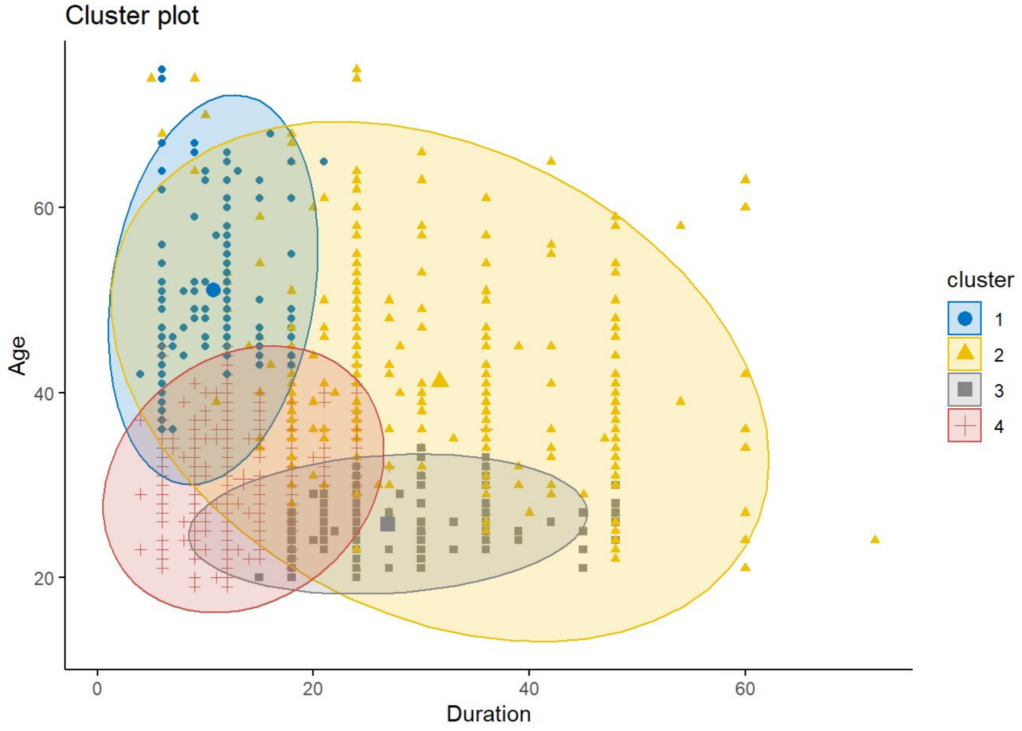
Data for AI Models

Tabulated or Structured Data



Credit Risk Data

Loan Purpose	Checking	Savings	Months Customer	Months Employed	Gender	Marital Status	Age	Housing	Years	Job	Credit Risk
Small Appliance	$0	$739	13	12	M	Single	23	Own	3	Unskilled	Low
Furniture	$0	$1,230	25	0	M	Divorced	32	Own	1	Skilled	High
New Car	$0	$389	19	119	M	Single	38	Own	4	Management	High
Furniture	$638	$347	13	14	M	Single	36	Own	2	Unskilled	High
Education	$963	$4,754	40	45	M	Single	31	Rent	3	Skilled	Low
Furniture	$2,827	$0	11	13	M	Married	25	Own	1	Skilled	Low
New Car	$0	$229	13	16	M	Married	26	Own	3	Unskilled	Low
Business	$0	$533	14	2	M	Single	27	Own	1	Unskilled	Low
Small Appliance	$6,509	$493	37	9	M	Single	25	Own	2	Skilled	High
Small Appliance	$966	$0	25	4	F	Divorced	43	Own	1	Skilled	High
Business	$0	$989	49	0	M	Single	32	Rent	2	Management	High
New Car	$0	$3,305	11	15	M	Single	34	Rent	2	Unskilled	Low
Business	$322	$578	10	14	M	Married	26	Own	1	Skilled	Low
New Car	$0	$821	25	63	M	Single	44	Own	1	Skilled	High
New Car	$396	$228	13	26	M	Single	46	Own	3	Unskilled	Low
Used Car	$0	$129	31	8	M	Divorced	39	Own	4	Management	Low
Furniture	$652	$732	49	4	F	Divorced	25	Own	2	Skilled	High
New Car	$708	$683	13	33	M	Single	31	Own	2	Skilled	Low
Repairs	$207	$0	28	116	M	Single	47	Own	4	Skilled	Low
Education	$287	$12,348	7	2	F	Divorced	23	Rent	2	Skilled	High
Furniture	$0	$17,545	34	16	F	Divorced	22	Own	4	Skilled	High
Furniture	$101	$3,871	13	5	F	Divorced	26	Rent	4	Skilled	High
Furniture	$0	$0	25	23	M	Married	19	Own	4	Skilled	High
Furniture	$0	$485	37	23	F	Divorced	27	Own	2	Management	High

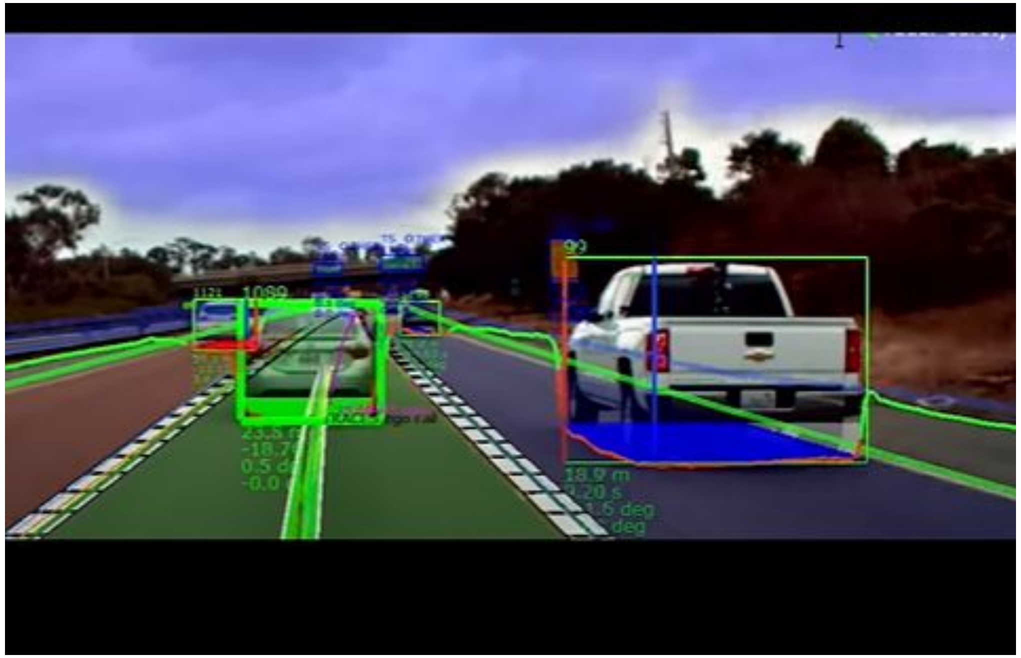


https://media.cheggcdn.com/media/d52/d52c60c8-60d4-4e55-882f-3ed24306f8cb/phpR8NHxM

https://rpubs.com/sid9715/580607

Data for AI Models

Images



https://www.youtube.com/watch?v=KS_4xjXNTxg&

https://viso.ai/applications/computer-vision-applications/

Data for AI Models

Language Data (spoken and written)

Hola, ¿cómo va mi pedido?

Hola Pilar, tu pedido #659548 ha sido enviado. Aquí tienes su código de seguimiento: SM1975MJA

https://www.grupoftp.com/noticias/el-futuro-de-los-chatbots/

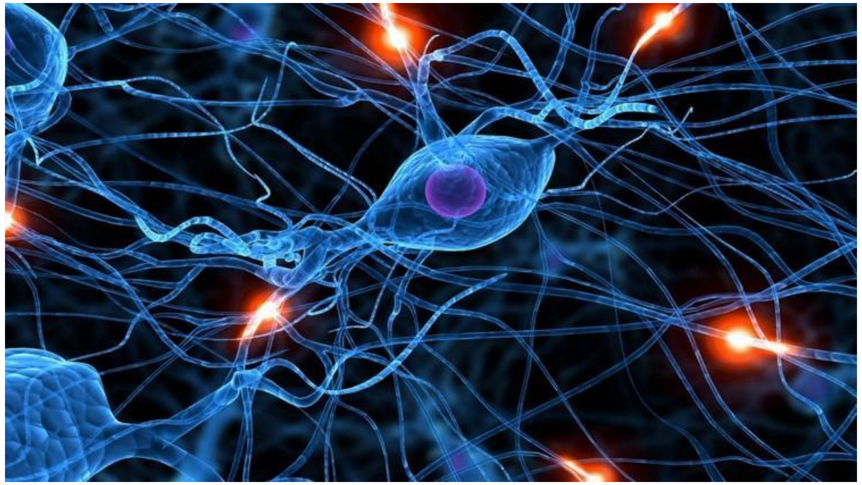
https://analyticsindiamag.com/google-translate-machine-learning/

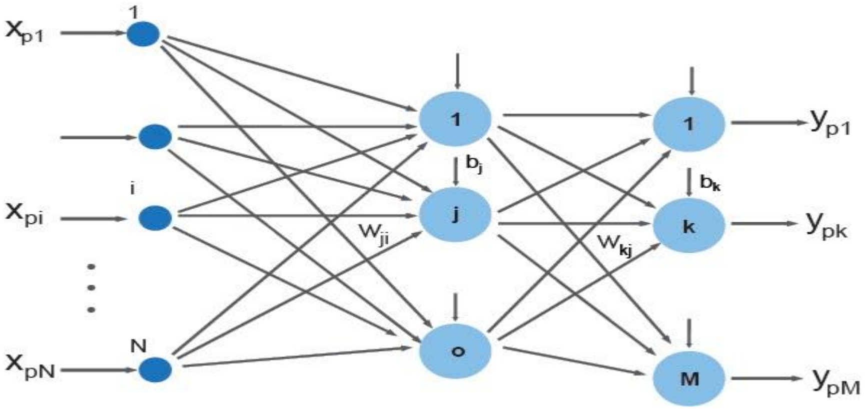
Deep Learning Concepts and Short History

Deep Learning

https://medicine.wustl.edu/news/slow-steady-waves-keep-brain-humming/

https://medium.com/espanol/avances-en-redes-neuronales-705c2efe53d2



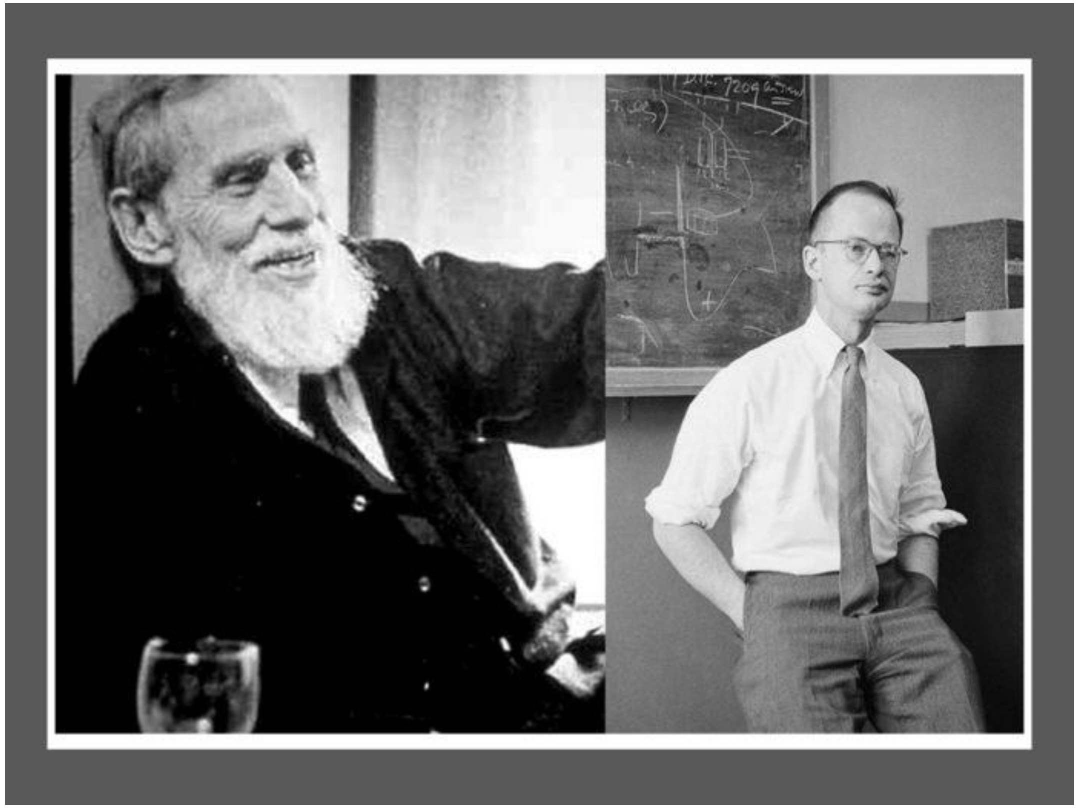


Fuente: Deep Learning. Teoría y Aplicaciones. Jesus Alfonso López. Alpha Editorial 2021

Deep Learning

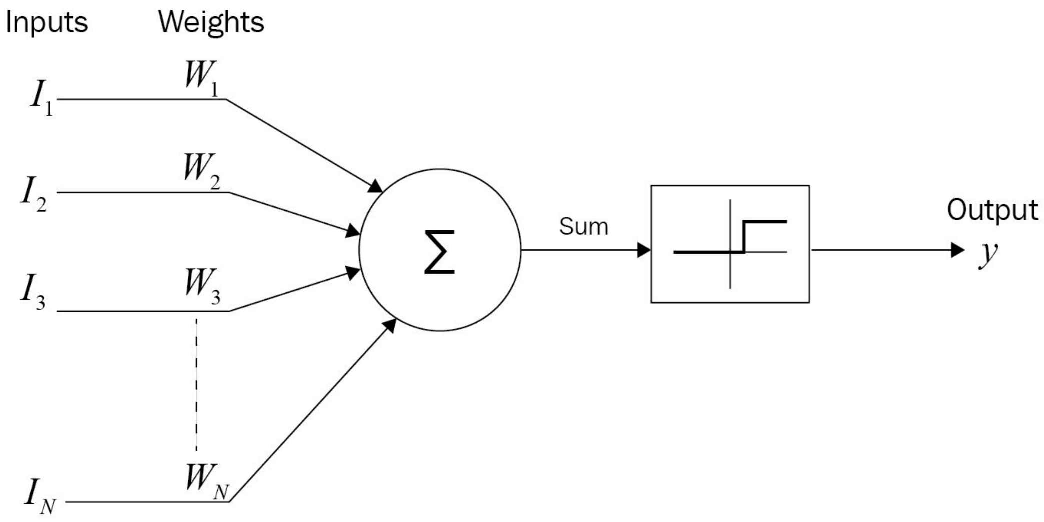
Dense Layers

McCulloch and Pitts



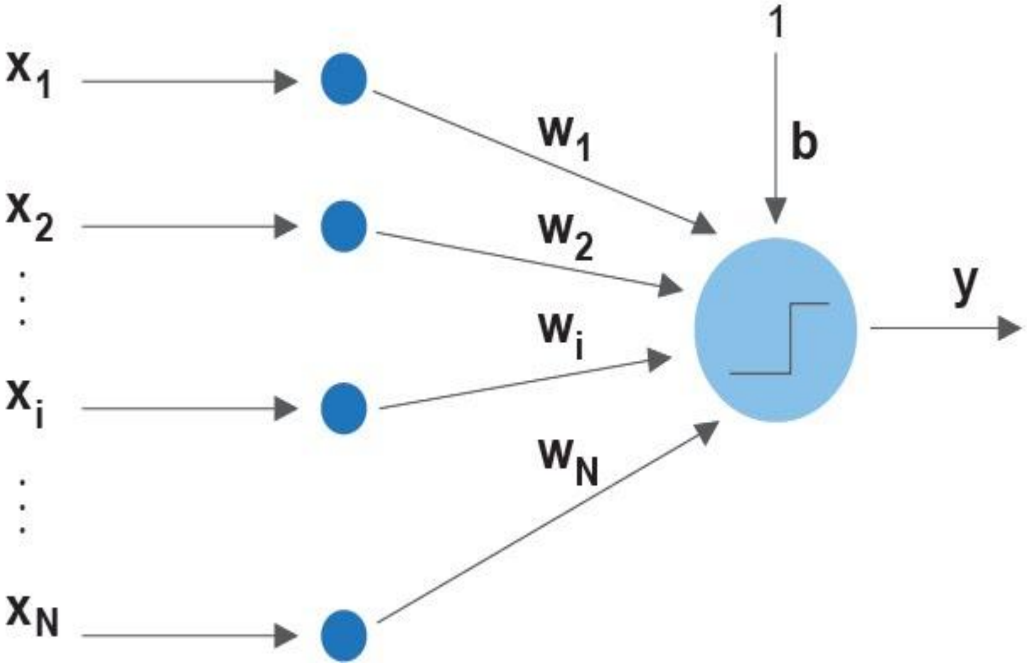
https://machinelearningknowledge.ai/timeline/mcculloch-pitts-neuron-the-beginning/

Artificial Neuron (1943)

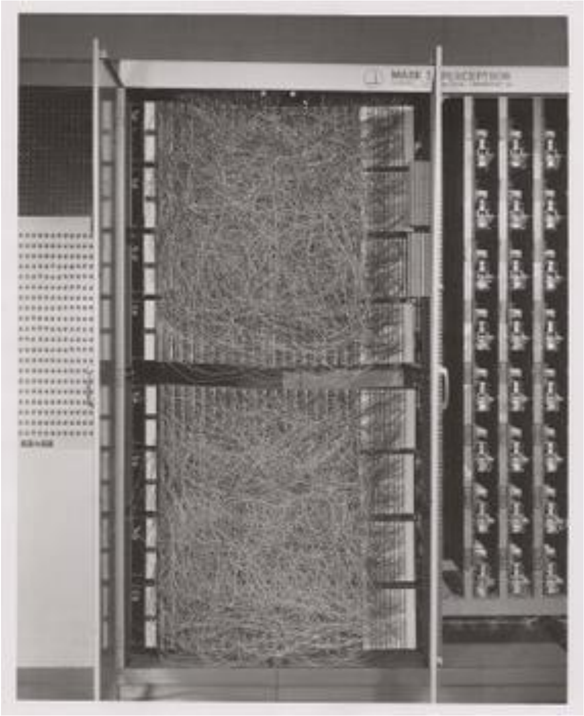


Deep Learning

Perceptron (1957)



The Mark I Perceptron machine was the first implementation of the perceptron algorithm

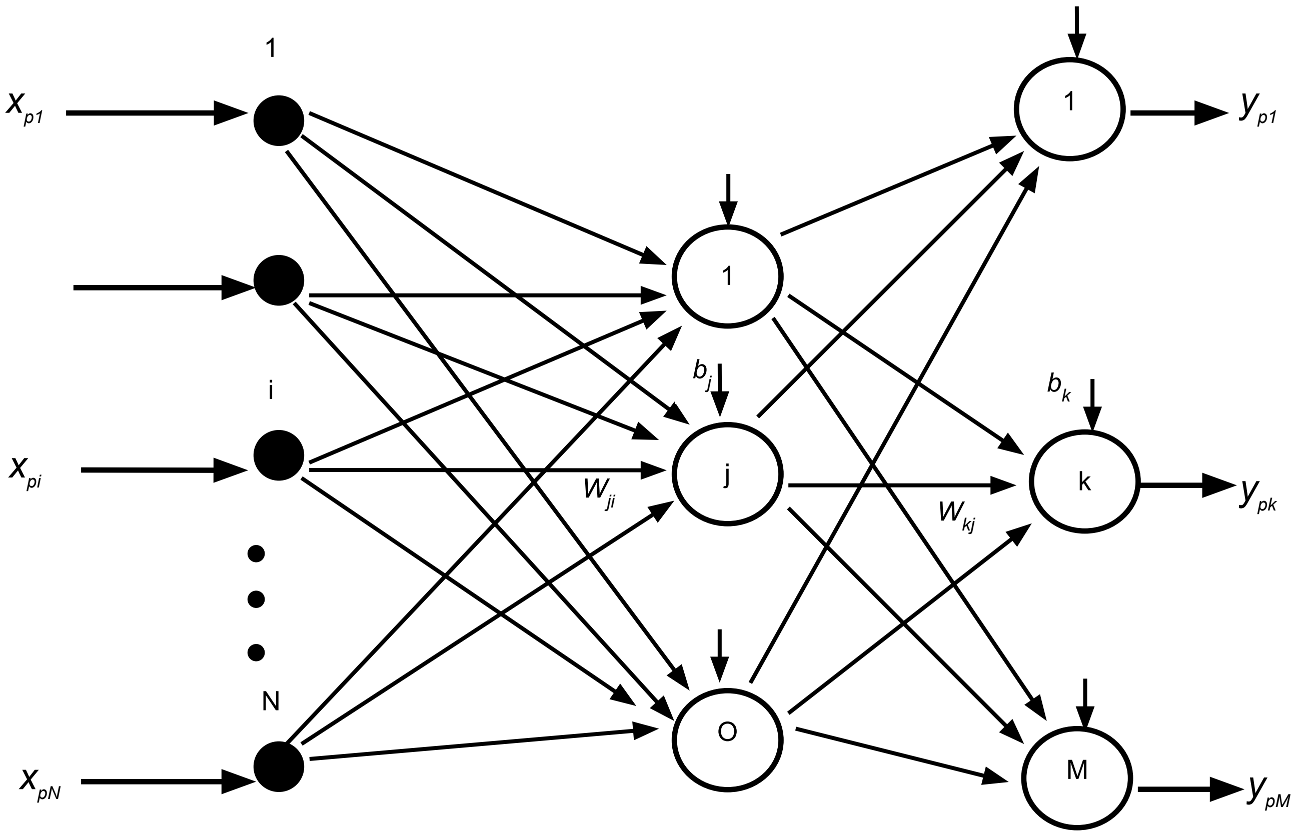


Frank Rossenblatt

https://en.wikipedia.org/wiki/Perceptron

https://blogs.umass.edu/comphon/2017/06/15/did-frank-rosenblatt-invent-de
-learning-in-1962/

Dense Layers

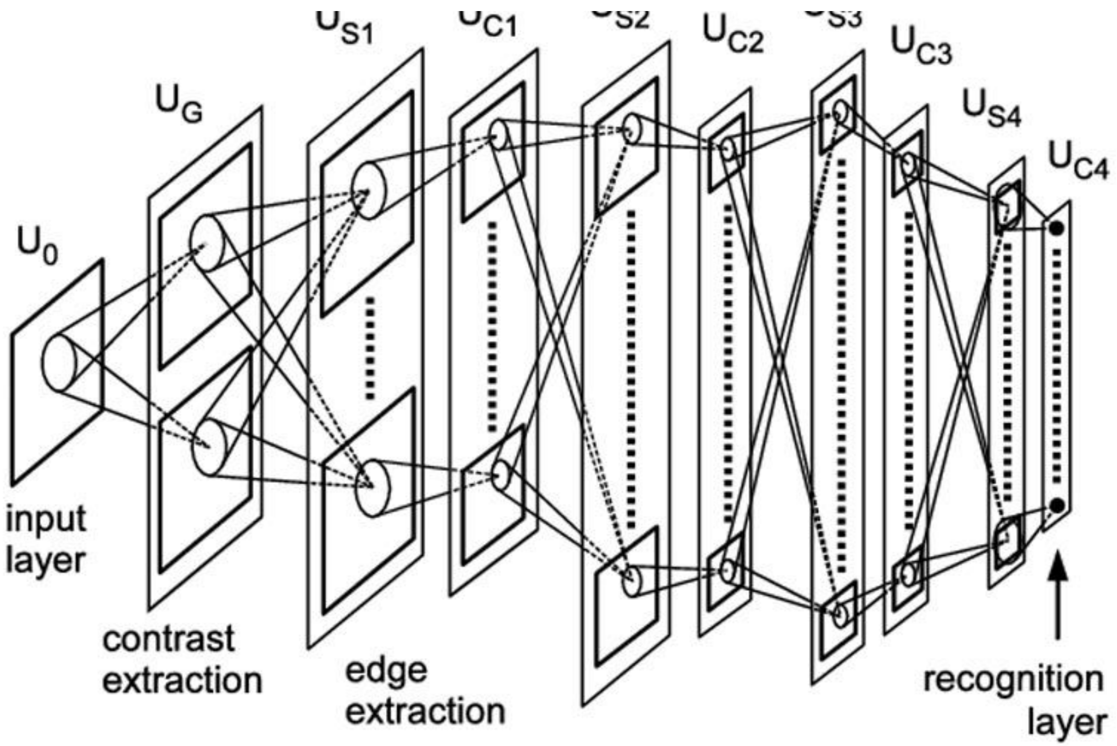


Fuente: Deep Learning. Teoría y Aplicaciones.. Jesus Alfonso López. 2021

Deep Learning

Convolutional Layers

Kunihiko Fukushima and the architecture of the Neocognitron (1979)



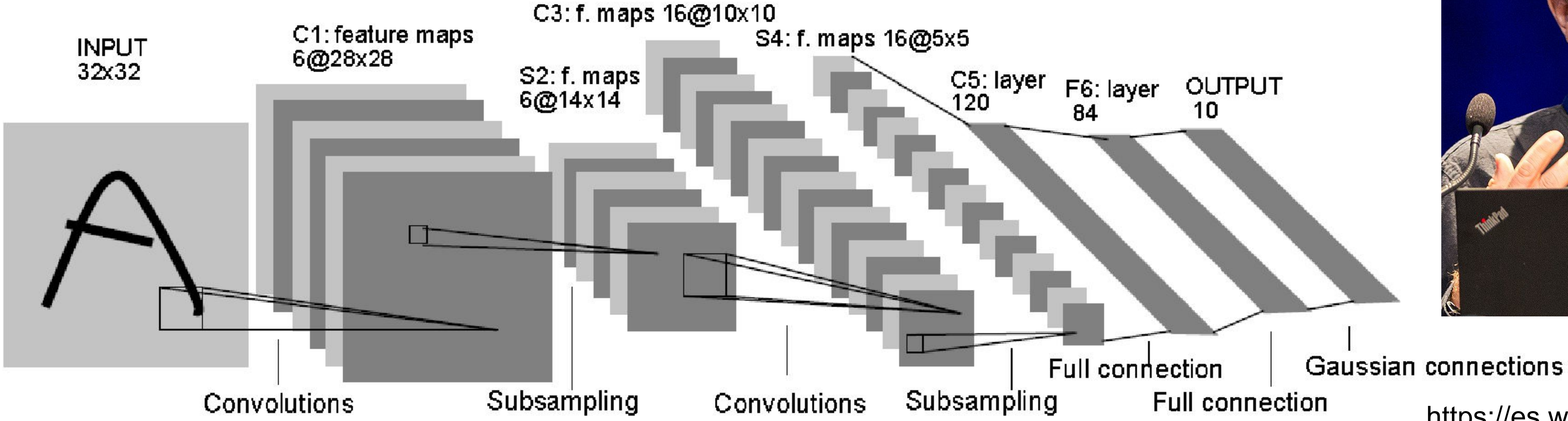
https://labelyourdata.com/articles/history-of-machine-learning-how-did-it-all-start#1979

Convolutional Layers

Yann LeCun

https://es.wikipedia.org/wiki/Yann_LeCun

LeNet-5 (1989)



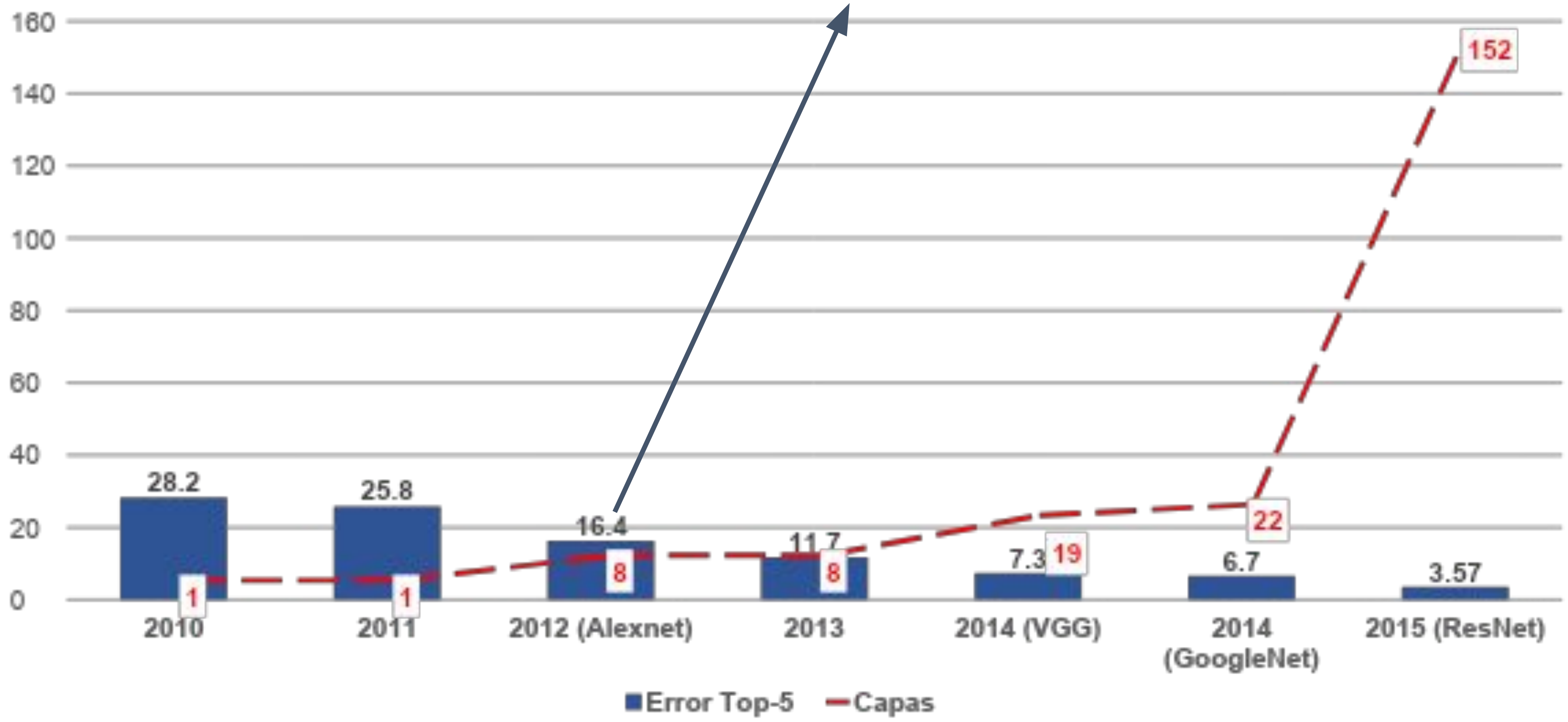
http://yann.lecun.com/exdb/publis/pdf/lecun-01a.pdf

LeNet- 1 Demo Video

https://www.youtube.com/watch?v=FwFduRA_L6Q

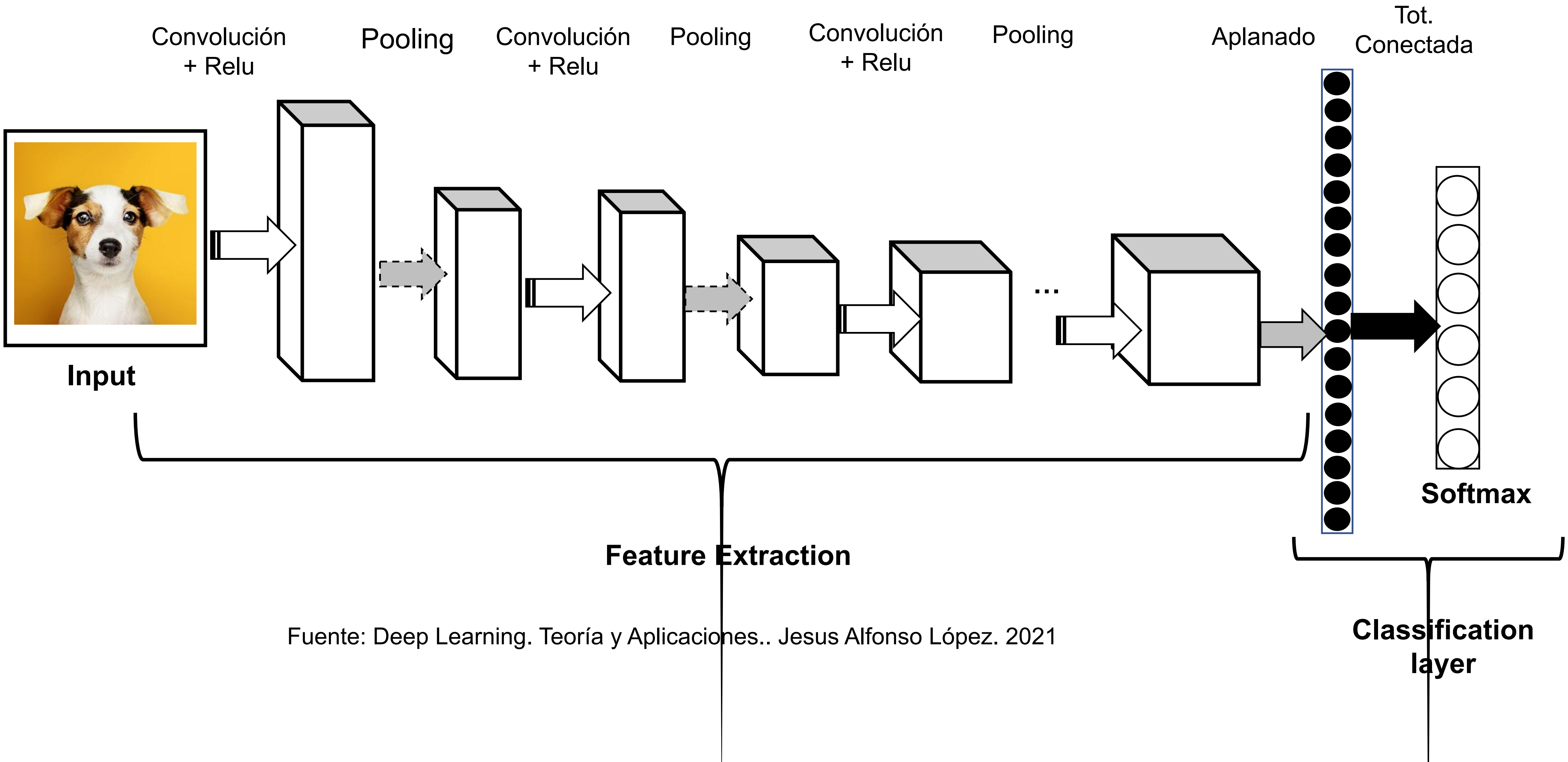
Convolutional Layers

Imagenet Moment



Fuente: Deep Learning. Teoría y Aplicaciones.. Jesus Alfonso López. 2021

Convolutional Layers



Convolución + Relu

Pooling

Convolución + Relu

Pooling

Convolución + Relu

Pooling

Aplanado

Tot. Conectada

Input

...

Softmax

Feature Extraction

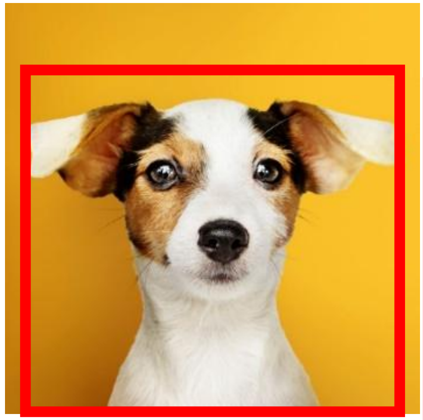
Classification layer

Fuente: Deep Learning. Teoría y Aplicaciones.. Jesus Alfonso López. 2021

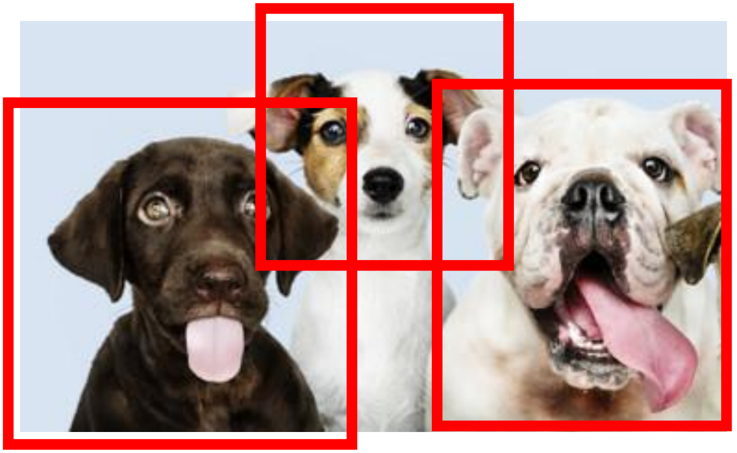
Convolutional Layers

Classification

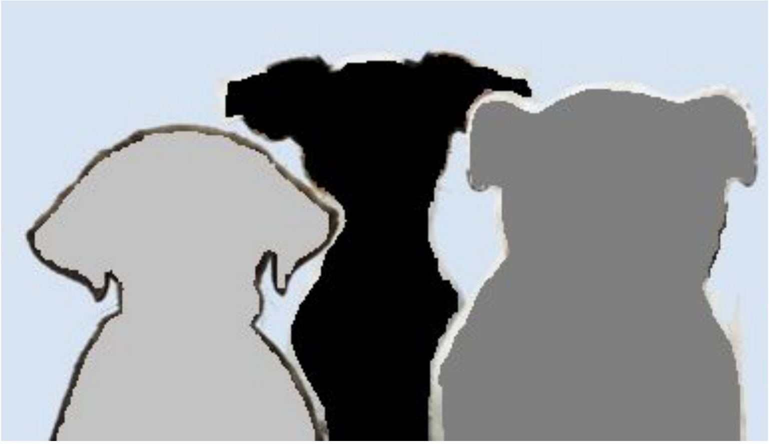
Classification and Localization



Detection



Segmentation



One Object

Several Objects

Fuente: Deep Learning. Teoría y Aplicaciones.. Jesus Alfonso López. 2021

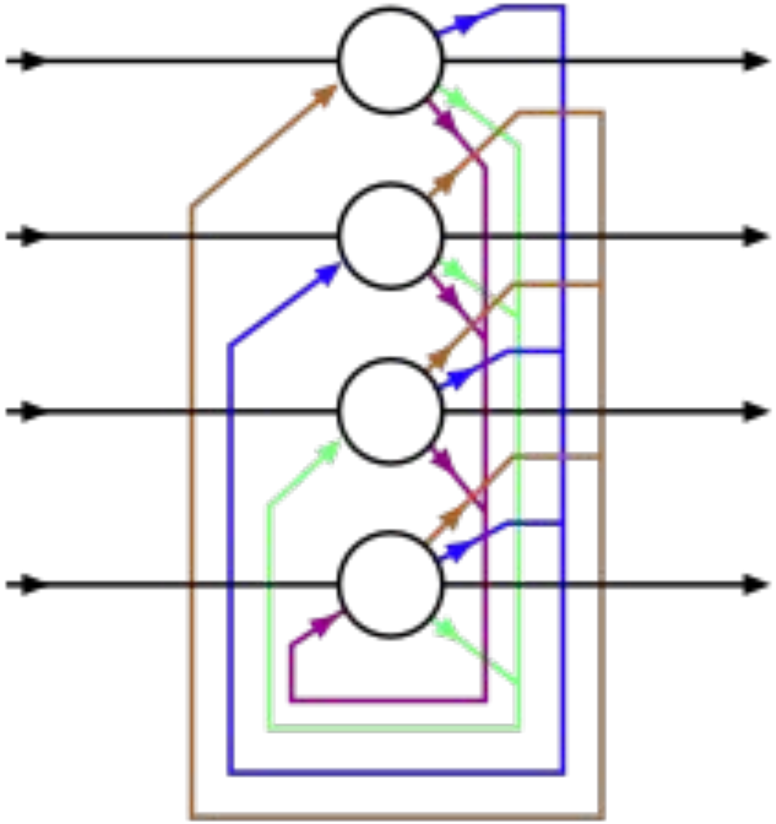
Deep Learning

Recurrent Layers

Hopfield neural network (1982)

John Joseph Hopfield

https://www.swarthmore.edu/bulletin/archive/wp/october-2009_john-hopfield-54.html

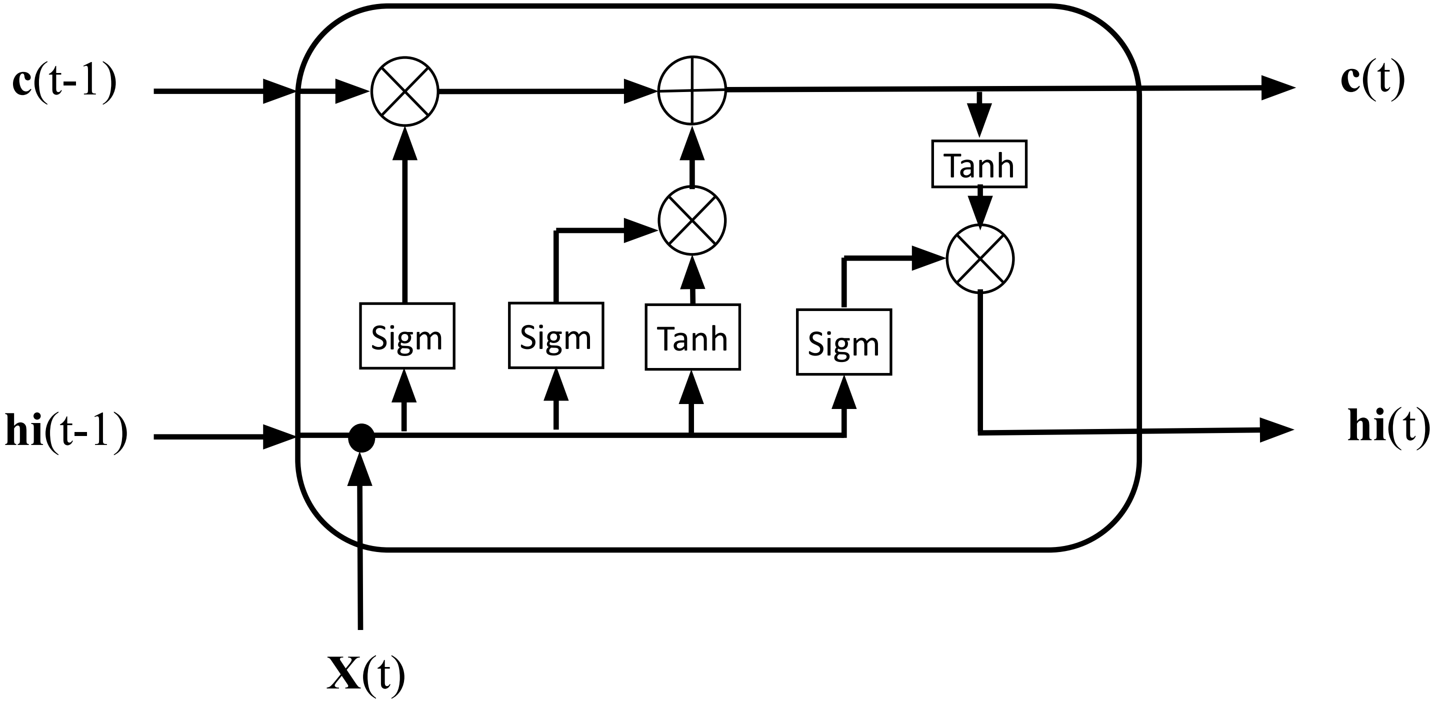


https://en.wikipedia.org/wiki/Hopfield_network

Recurrents Layers

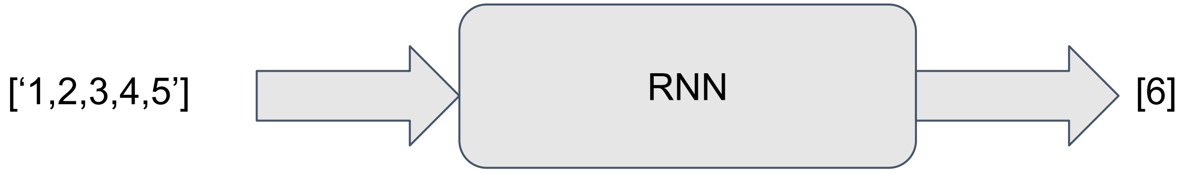
LSTM (1997)

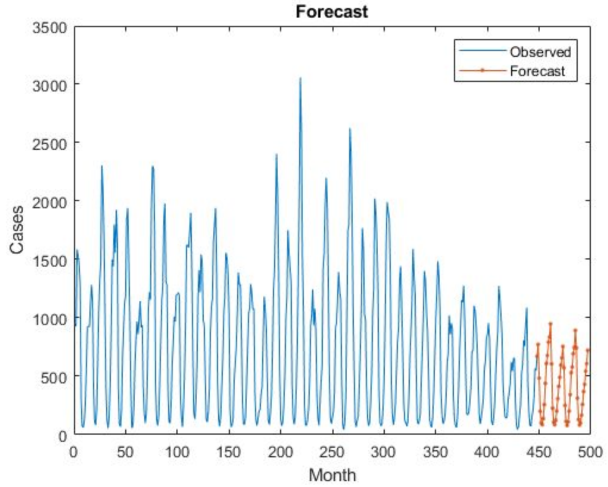
Jürgen Schmidhuber



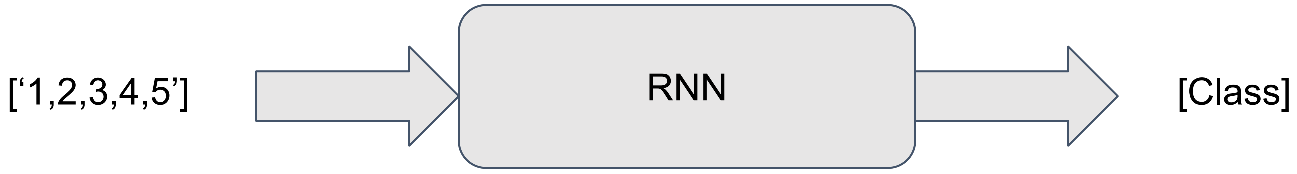
Fuente: Deep Learning. Teoría y Aplicaciones.. Jesus Alfonso López. 2021

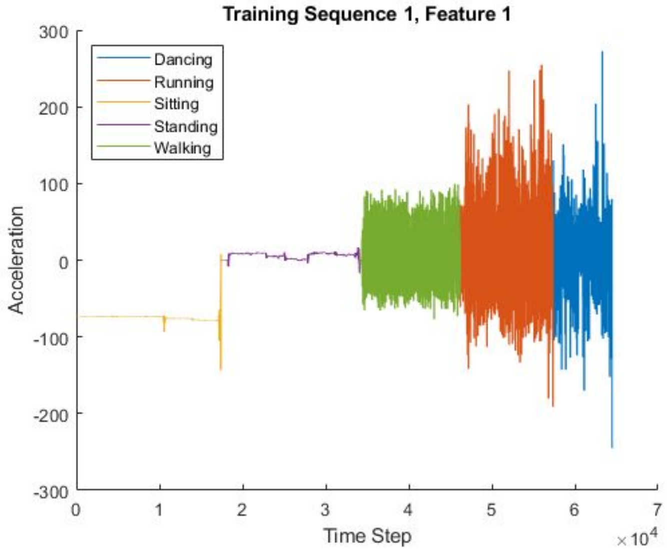
Recurrents Layers





https://la.mathworks.com/help/deeplearning/examples/time-series-forecasting-using-deep-learning.html





https://la.mathworks.com/help/deeplearning/examples/sequence-to-sequence-classification-using-deep-learning.html

Recurrents Layers

NLP (Natural Language Processing)

['I am sad']

RNN

['Sadness']

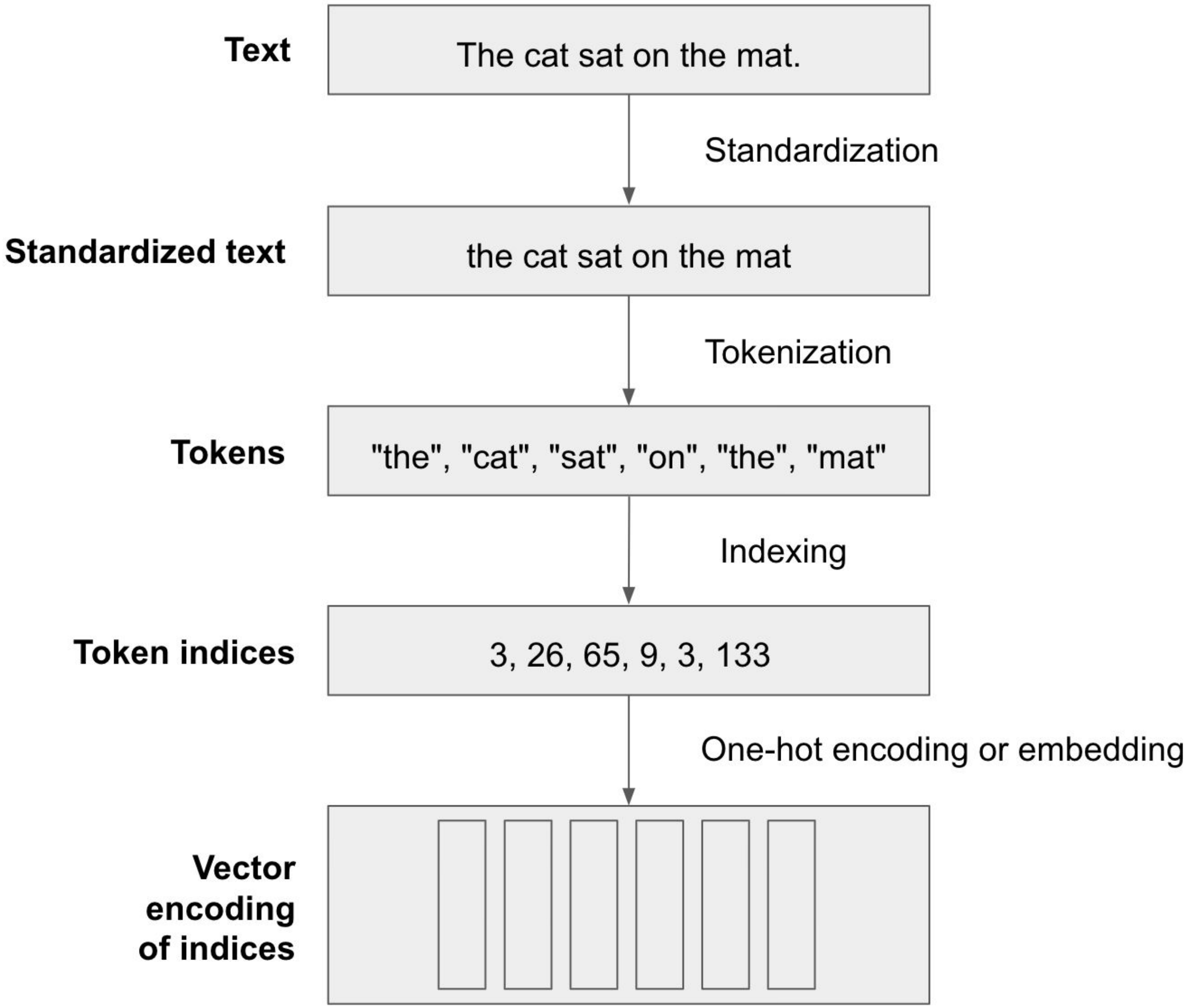
['La casa blanca']

RNN

['The white house']

Language Models

Embedding



Text: The cat sat on the mat.

Standardization

Standardized text: the cat sat on the mat

Tokenization

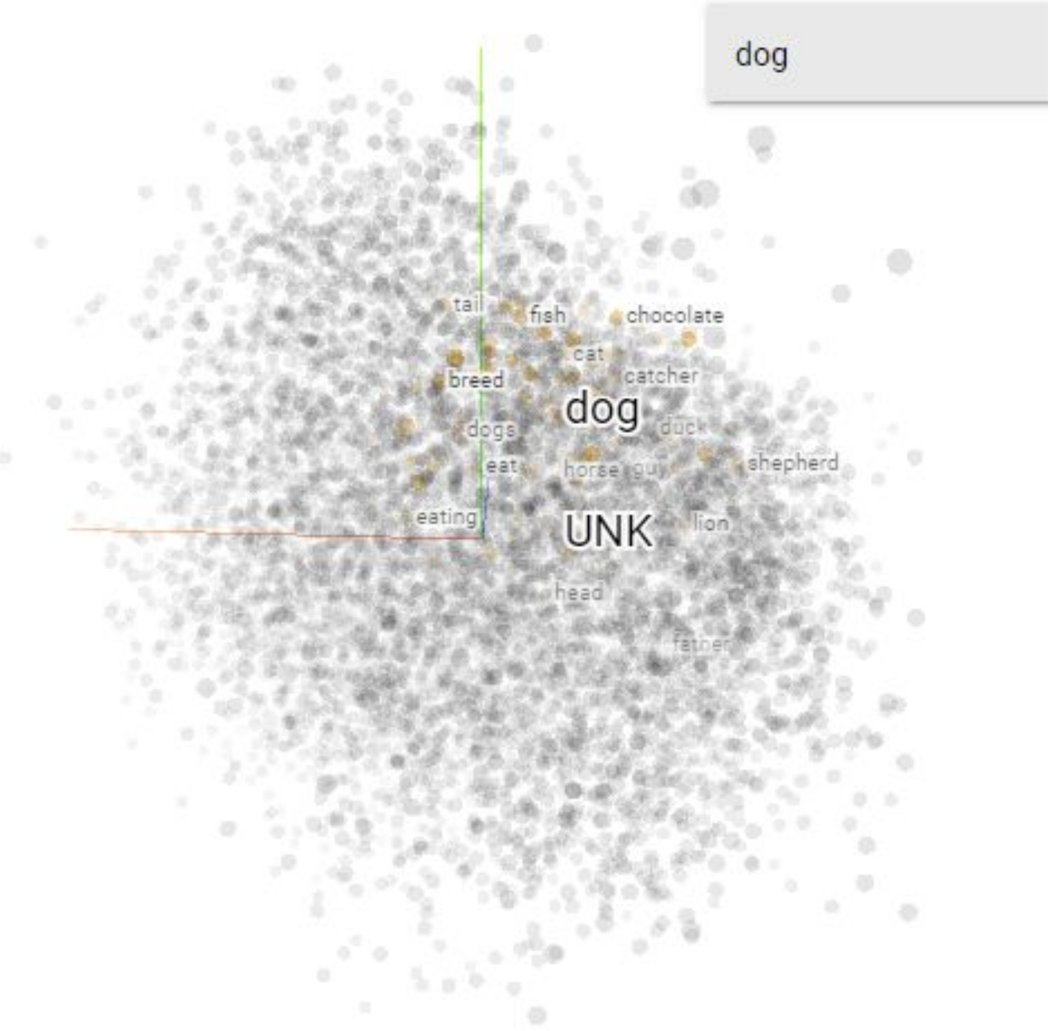
Tokens: "the", "cat", "sat", "on", "the", "mat"

Indexing

Token indices: 3, 26, 65, 9, 3, 133

One-hot encoding or embedding

Vector encoding of indices

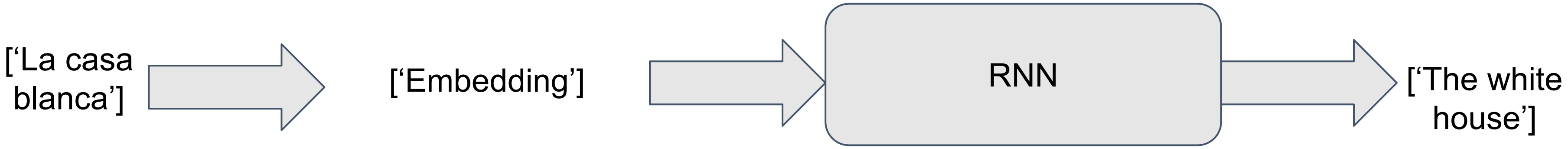


https://projector.tensorflow.org/

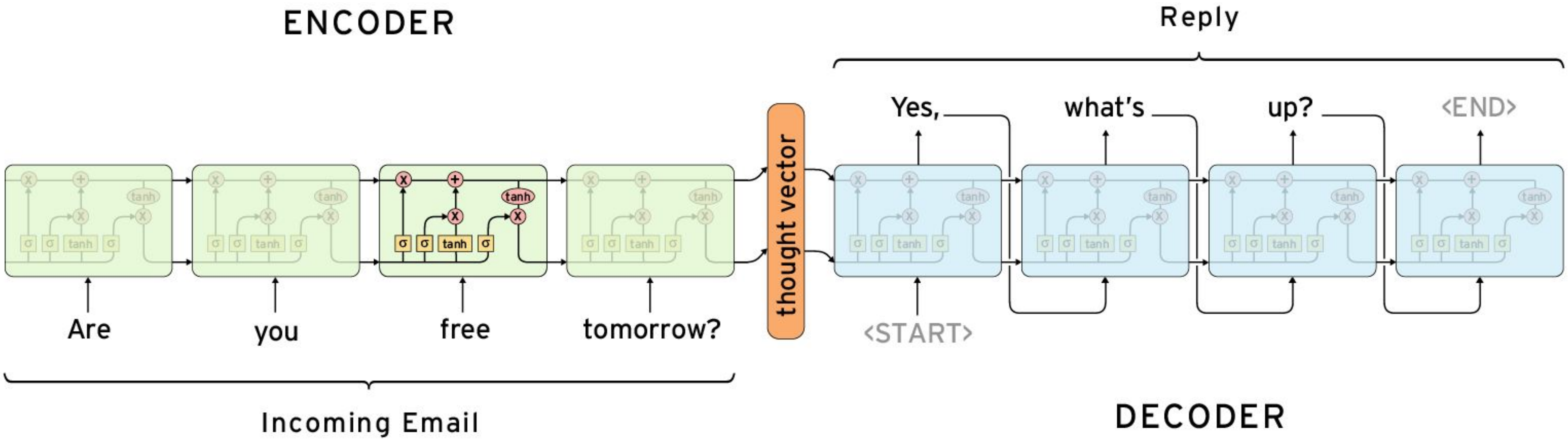
https://livebook.manning.com/book/deep-learning-with-python-second-edition/chapter-11/v-6/28

Language Models

Sequence to sequence



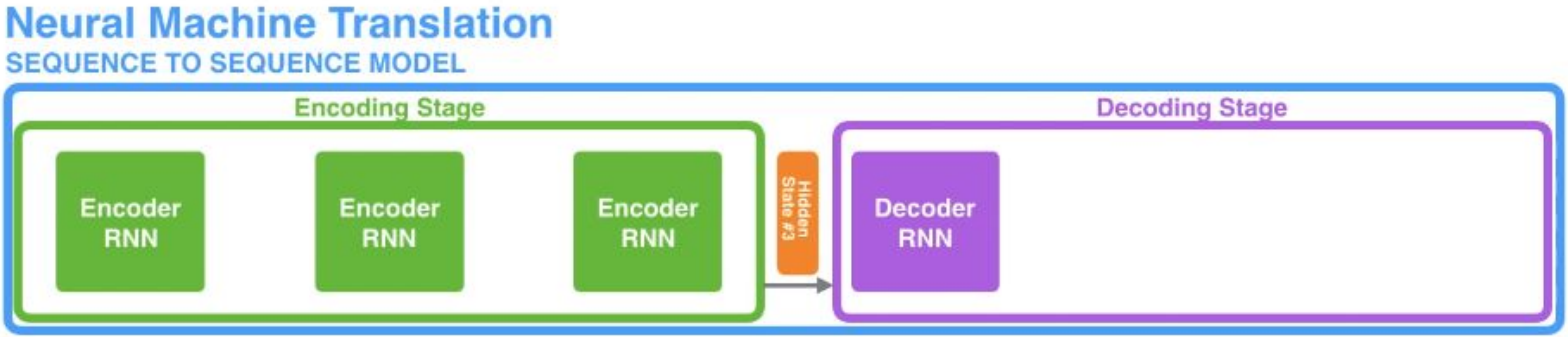
['La casa blanca'] → ['Embedding'] → RNN → ['The white house']



https://github.com/nicolas-ivanov/debug_seq2seq

Language Models

The last state of the encoder is used as input to the decoder.

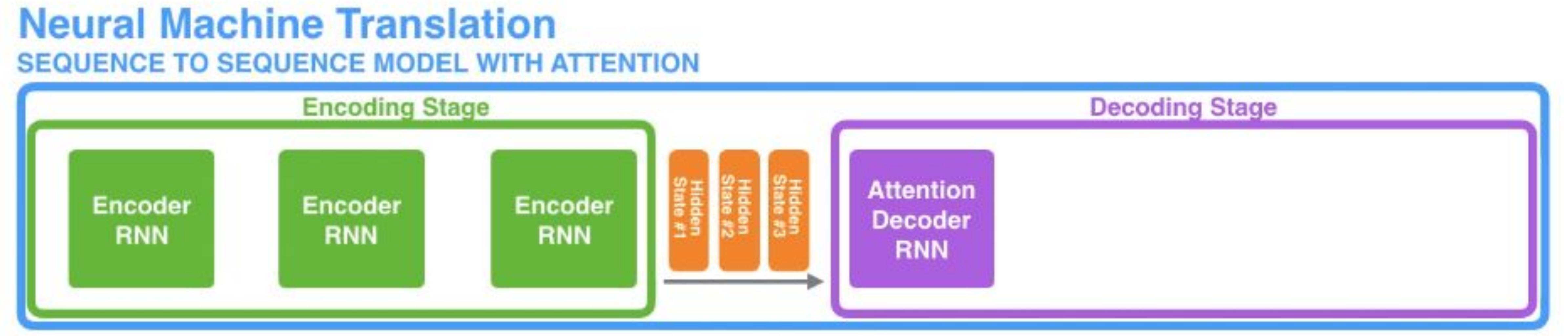


https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

Language Models

Attention

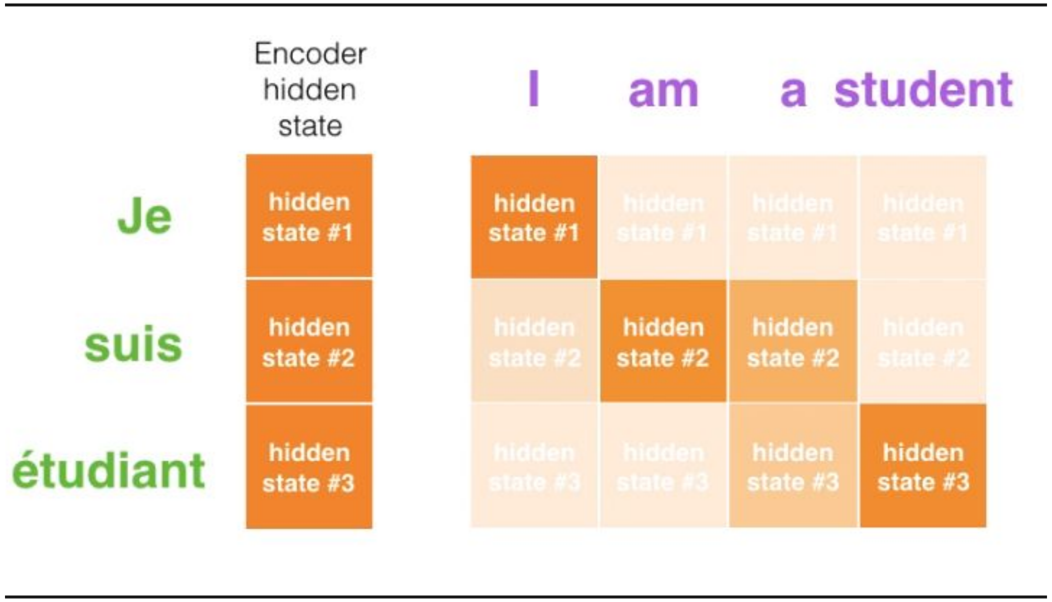
Considering attention all hidden states of the encoder are used.

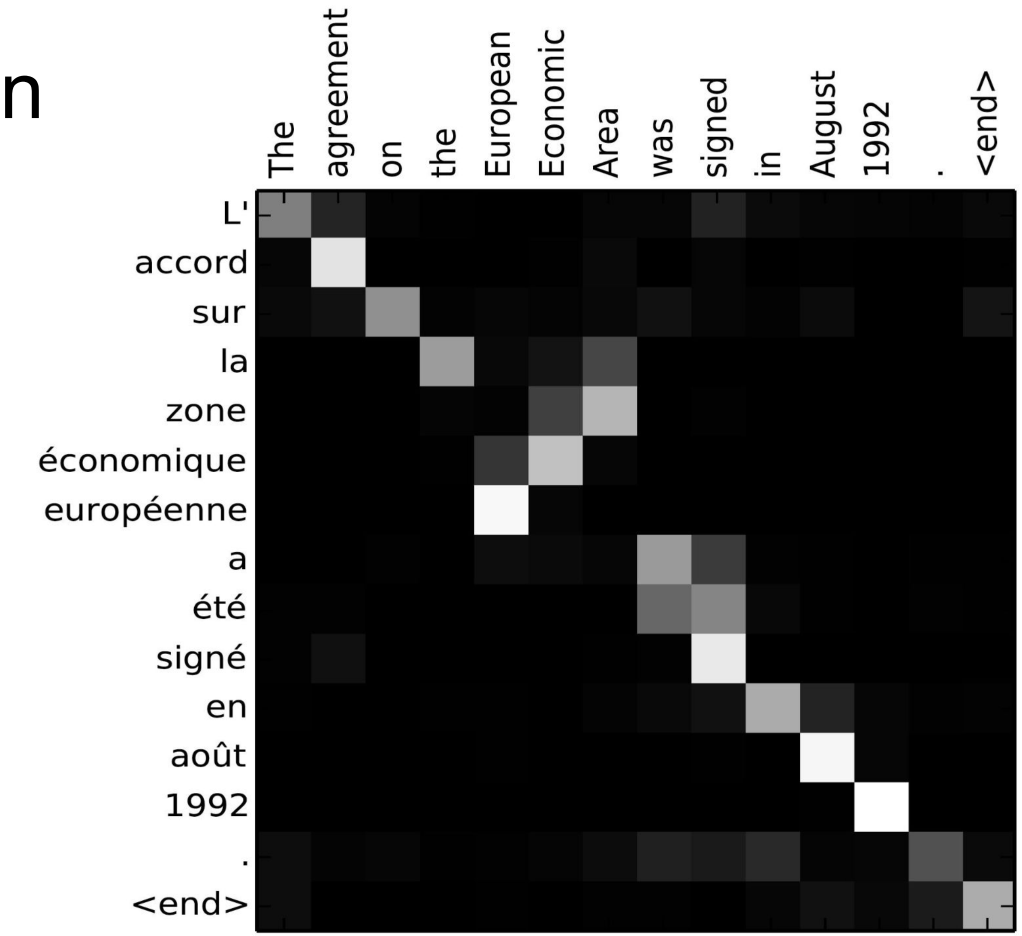


https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

Language Models

Attention



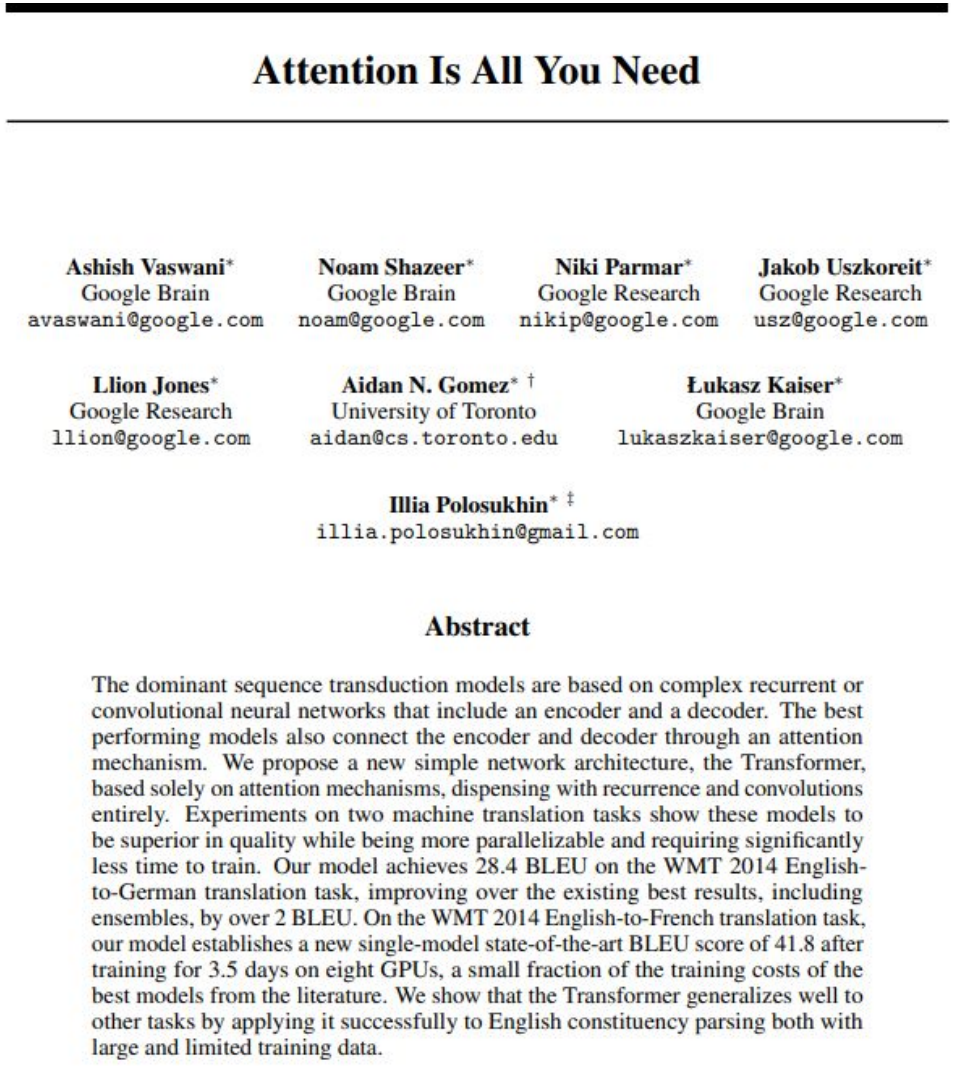


https://jalammar.github.io/visualizing-neural-machine-translation-mechanics-of-seq2seq-models-with-attention/

https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

Deep Learning

Self-Attention Layers



Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

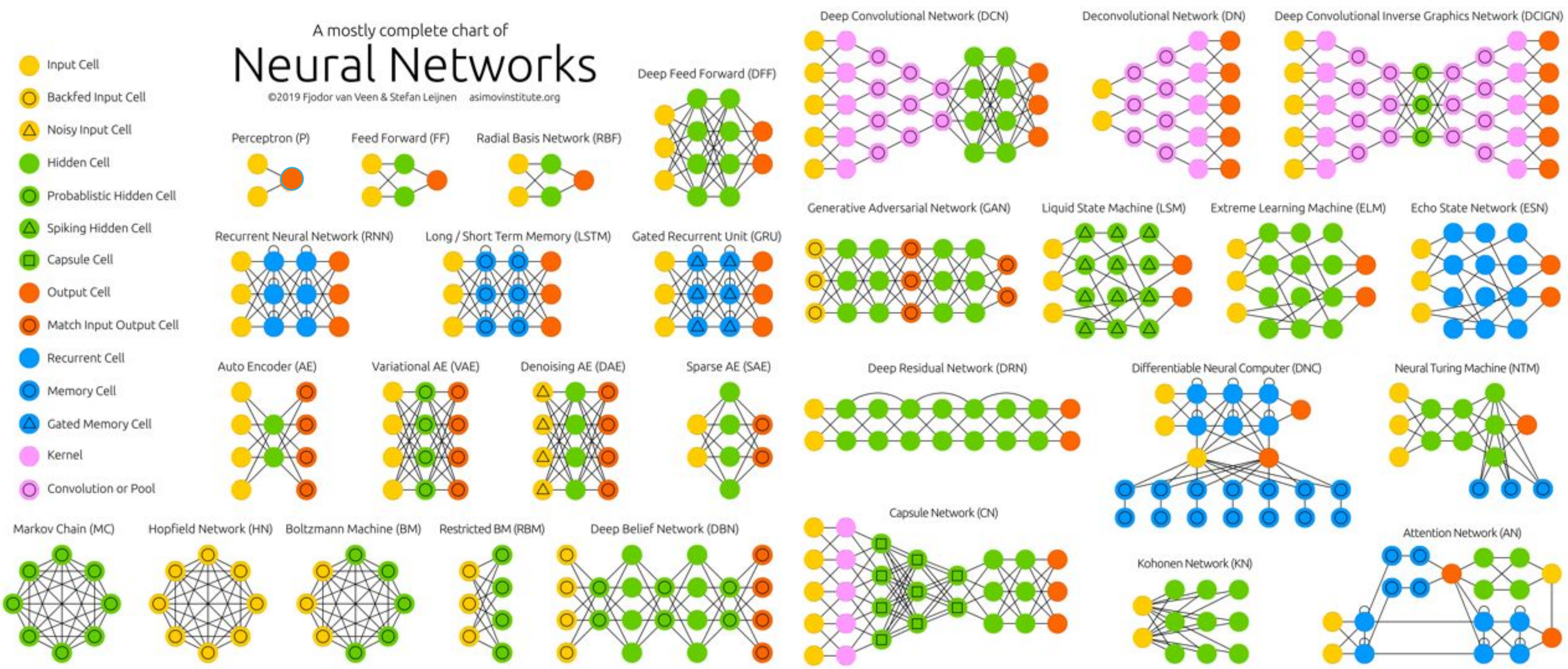
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Self-Attention equation

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\intercal}}{\sqrt{d_{keys}}}\right)\mathbf{V}$$

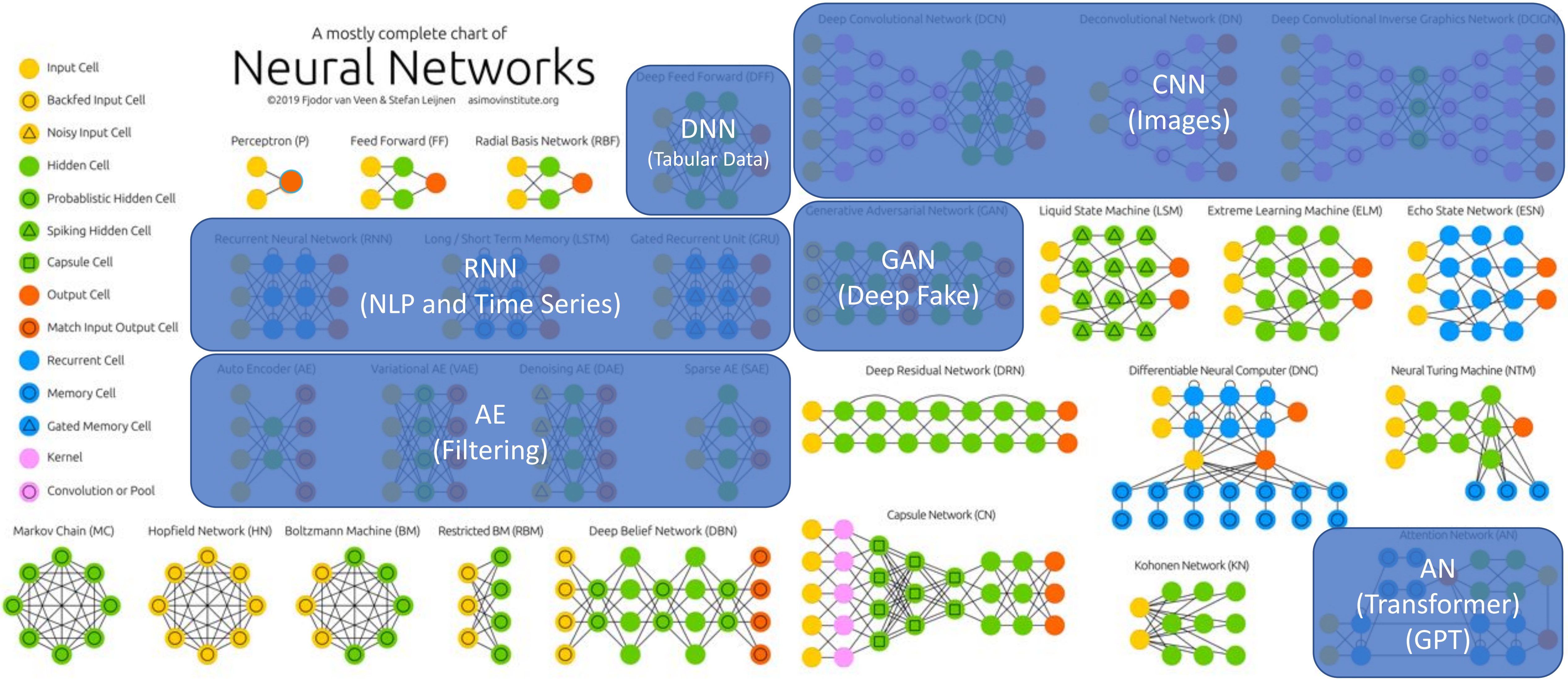
https://arxiv.org/pdf/1706.03762.pdf

Deep Learning ZOO



https://www.asimovinstitute.org/neural-network-zoo/

Deep Learning ZOO


A mostly complete chart of
Neural Networks
©2019 Fjodor van Veen & Stefan Leijnen asimovinstitute.org
Input Cell
Backfed Input Cell
Noisy Input Cell
Hidden Cell
Probablistic Hidden Cell
Spiking Hidden Cell
Capsule Cell
Output Cell
Match Input Output Cell
Recurrent Cell
Memory Cell
Gated Memory Cell
Kernel
Convolution or Pool
Perceptron (P)
Feed Forward (FF)
Radial Basis Network (RBF)
Deep Feed Forward (DFF)
Deep Convolutional Network (DCN)
Deconvolutional Network (DN)
Deep Convolutional Inverse Graphics Network (DCIGN)
Recurrent Neural Network (RNN)
Long / Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Generative Adversarial Network (GAN)
Liquid State Machine (LSM)
Extreme Learning Machine (ELM)
Echo State Network (ESN)
Auto Encoder (AE)
Variational AE (VAE)
Denoising AE (DAE)
Sparse AE (SAE)
Deep Residual Network (DRN)
Differentiable Neural Computer (DNC)
Neural Turing Machine (NTM)
Markov Chain (MC)
Hopfield Network (HN)
Boltzmann Machine (BM)
Restricted BM (RBM)
Deep Belief Network (DBN)
Capsule Network (CN)
Kohonen Network (KN)
Attention Network (AN)

DNN
(Tabular Data)

CNN
(Images)

RNN
(NLP and Time Series)

GAN
(Deep Fake)

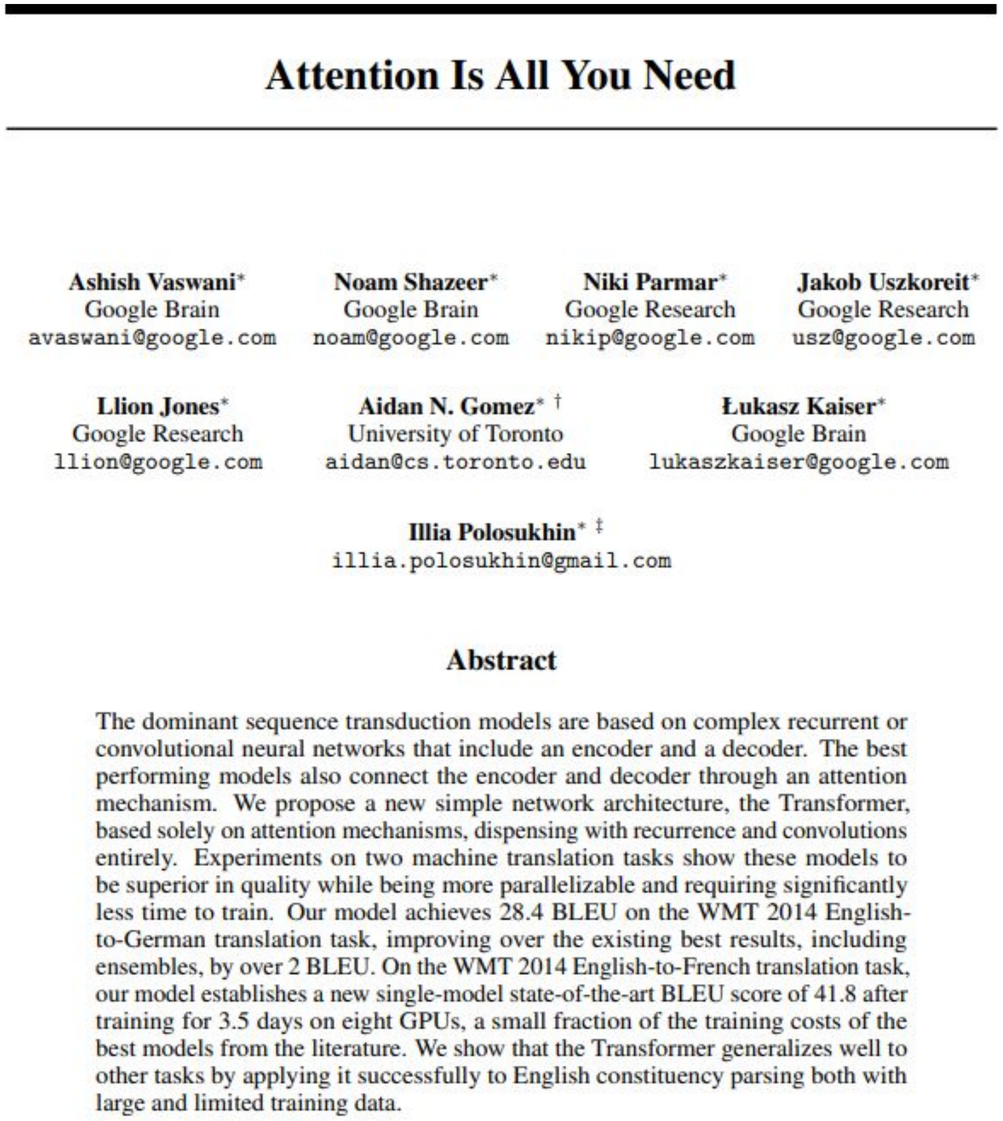
AE
(Filtering)

AN
(Transformer)
(GPT)

https://www.asimovinstitute.org/neural-network-zoo/

Current AI: Transformer World

Transformers



Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

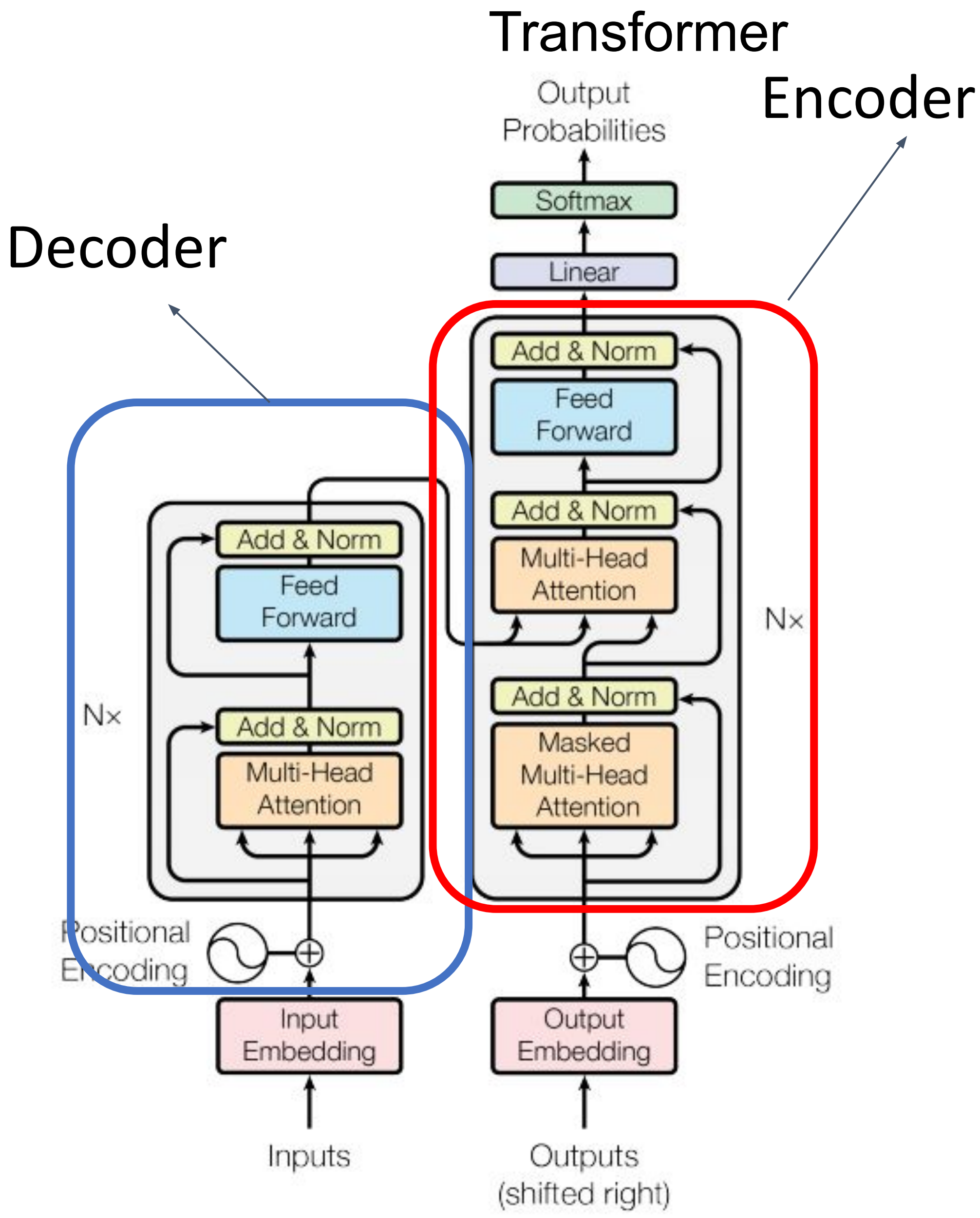
Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

https://arxiv.org/pdf/1706.03762.pdf

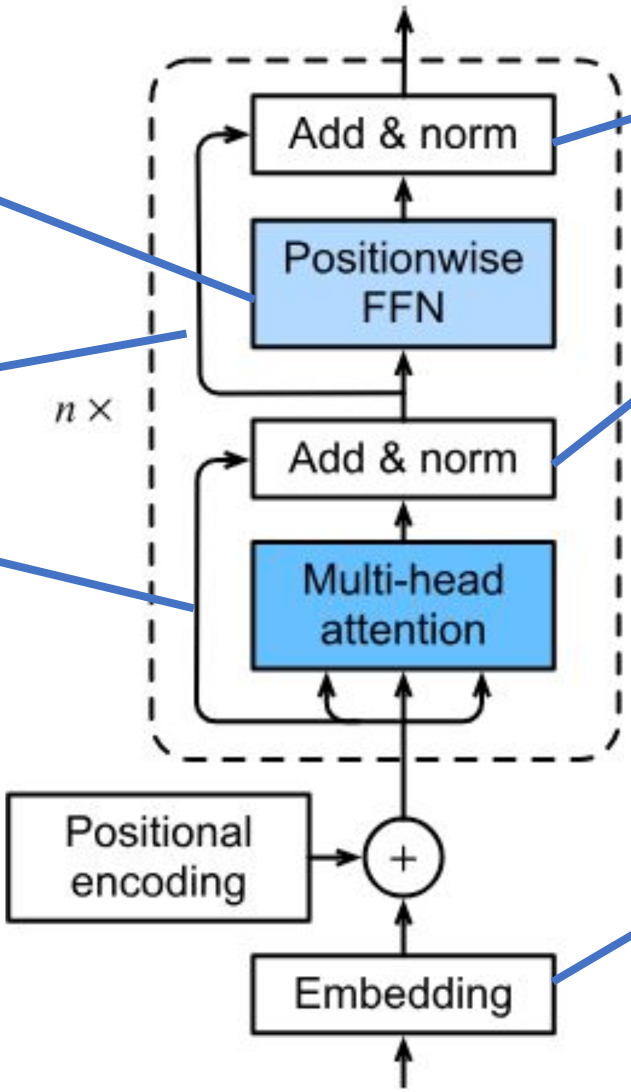


Transformer

Encoder

Decoder

Transformers



Dense Layer

Residuals Conexions
(CNN)

Normalization
(CNN)

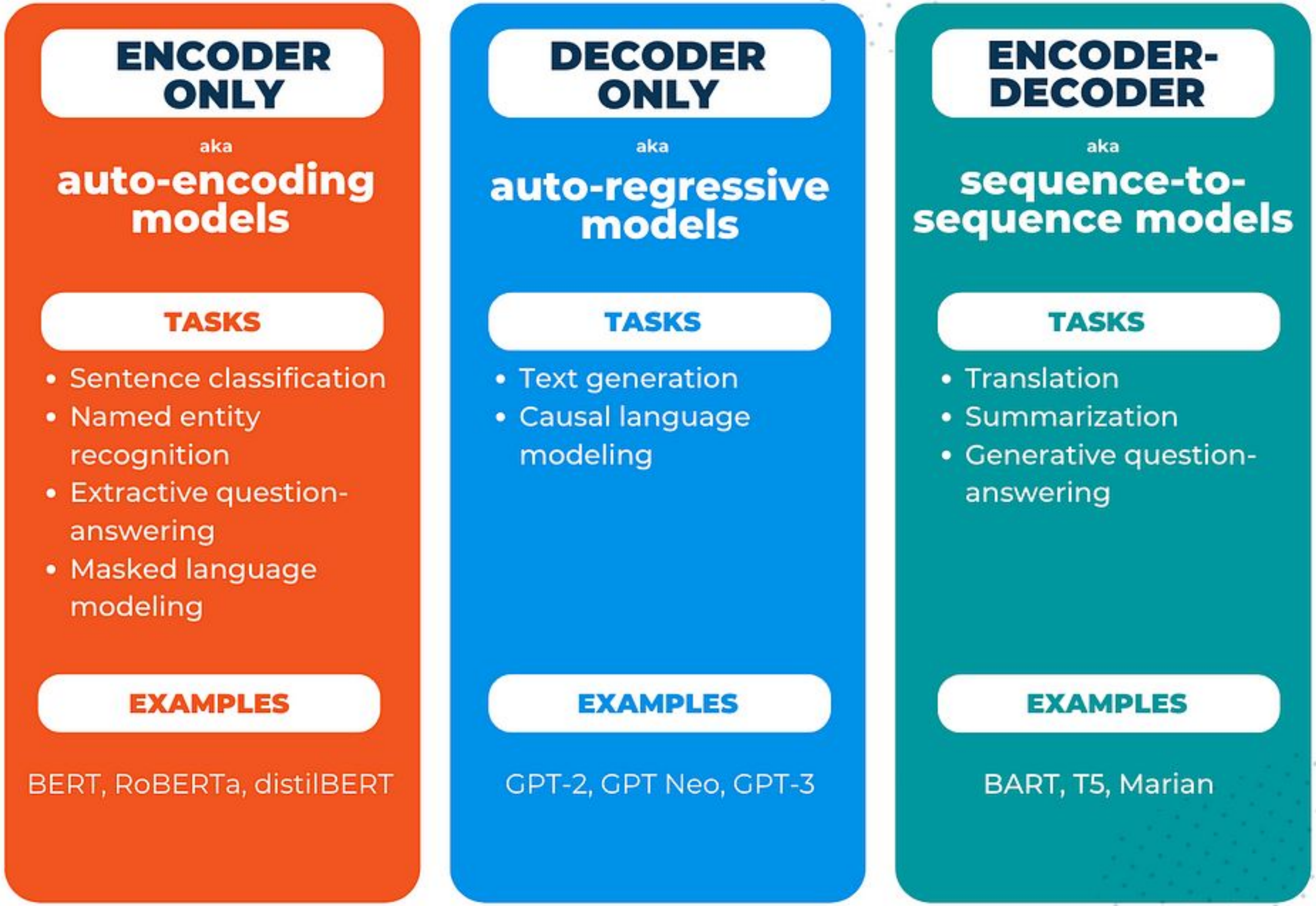
NLP based on RNN

Transformer language
model visualization

https://bbycroft.net/llm

https://d2l.ai/_images/transformer.svg

Transformers and Language Models

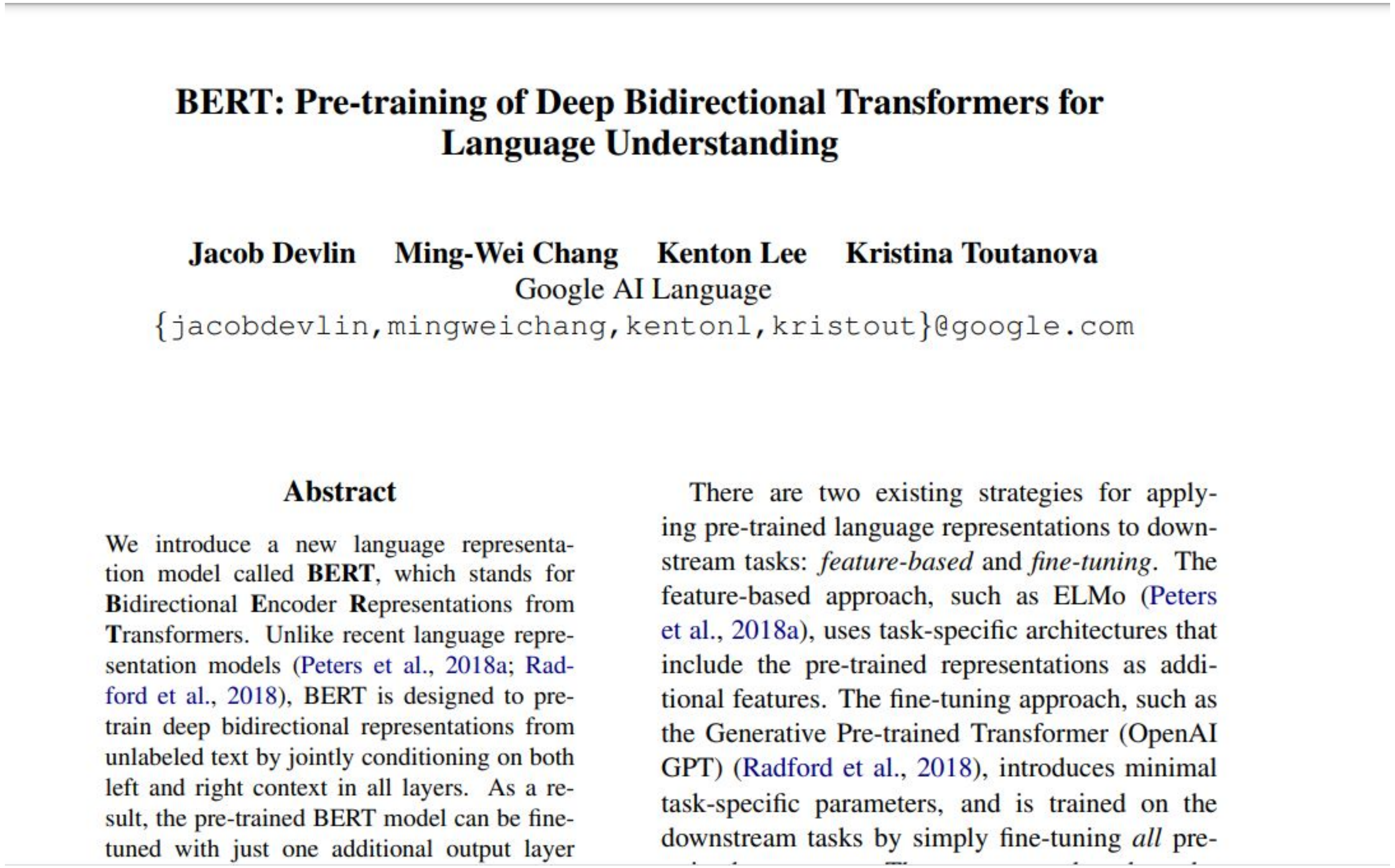


ENCODER ONLY	DECODER ONLY	ENCODER-DECODER
aka auto-encoding models	aka auto-regressive models	aka sequence-to-sequence models
TASKS	TASKS	TASKS
• Sentence classification • Named entity recognition • Extractive question-answering • Masked language modeling	• Text generation • Causal language modeling	• Translation • Summarization • Generative question-answering
EXAMPLES	EXAMPLES	EXAMPLES
BERT, RoBERTa, distilBERT	GPT-2, GPT Neo, GPT-3	BART, T5, Marian

https://mlops.community/explainable-ai-visualizing-attention-in-transformers/

Transformers and Language Models

BERT (Bidirectional Encoder Representations from Transformers)



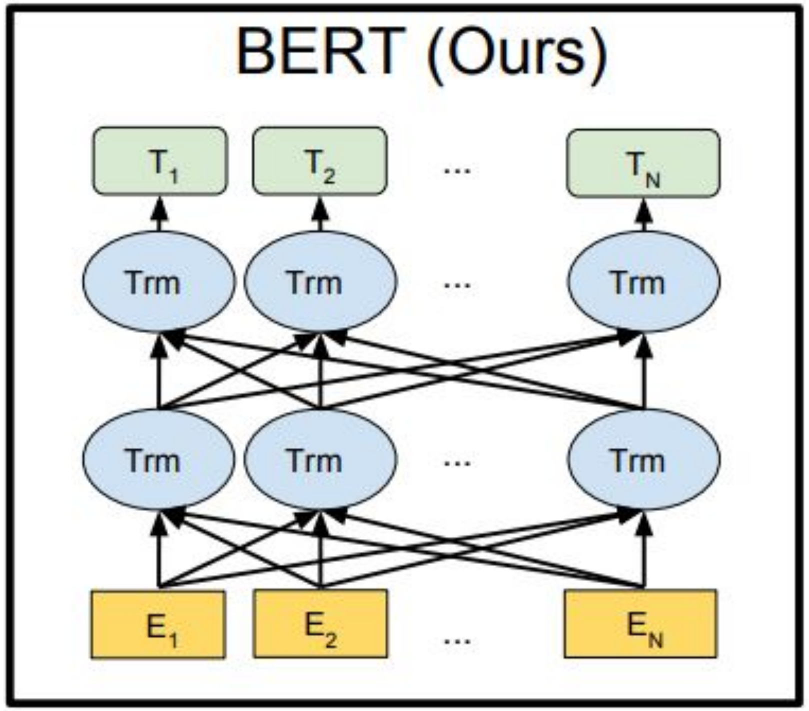
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
`{jacobdevlin,mingweichang,kentonl,kristout}@google.com`

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-

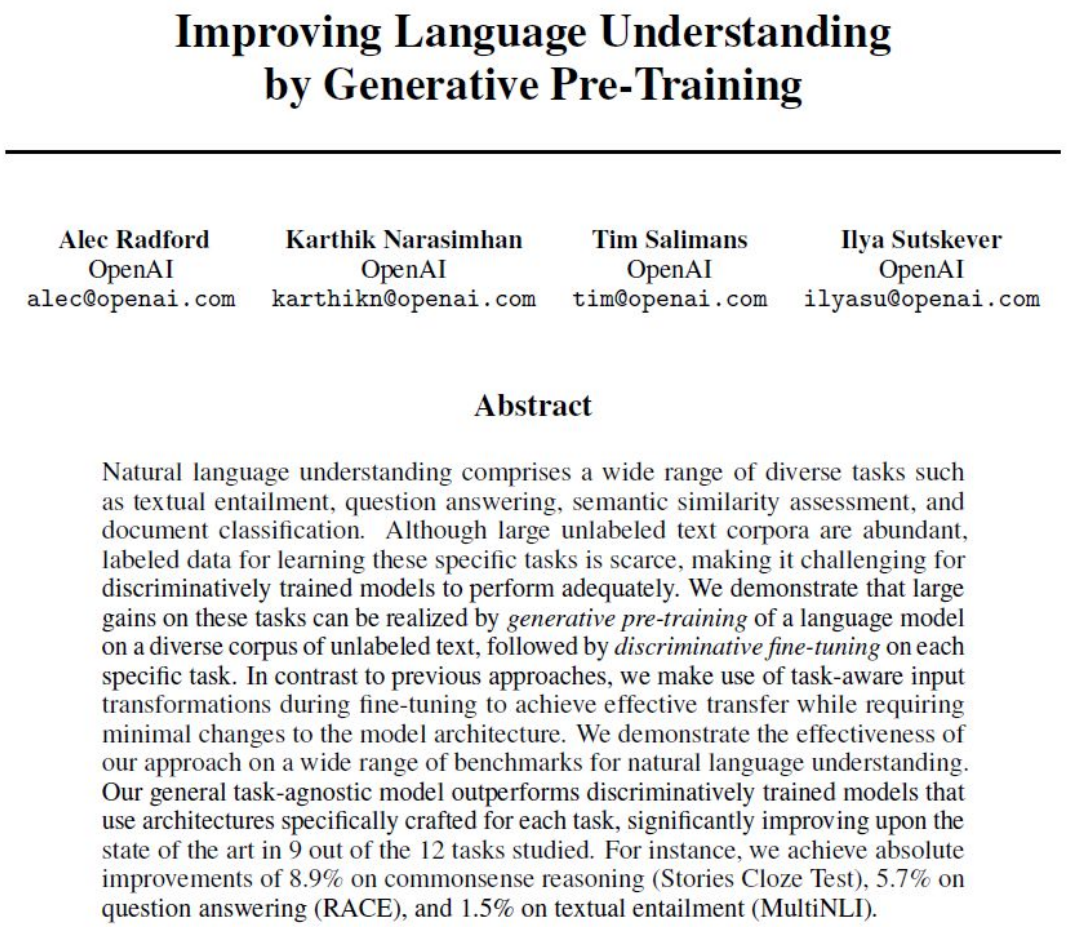


https://arxiv.org/pdf/1810.04805.pdf

Transformers and Language Models

GPT (Generative Pretrained Transformers**)**

12-layers, 768-hidden, 12-attention-heads, 117M parameters. Tamaño de secuencia 512



Improving Language Understanding by Generative Pre-Training

Alec Radford
OpenAI
alec@openai.com

Karthik Narasimhan
OpenAI
karthikn@openai.com

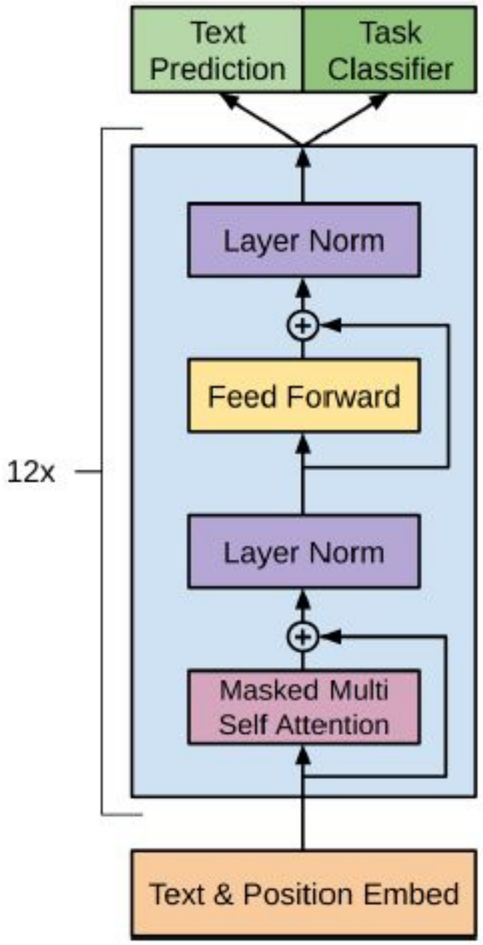
Tim Salimans
OpenAI
tim@openai.com

Ilya Sutskever
OpenAI
ilyasu@openai.com

Abstract

Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. We demonstrate that large gains on these tasks can be realized by *generative pre-training* of a language model on a diverse corpus of unlabeled text, followed by *discriminative fine-tuning* on each specific task. In contrast to previous approaches, we make use of task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture. We demonstrate the effectiveness of our approach on a wide range of benchmarks for natural language understanding. Our general task-agnostic model outperforms discriminatively trained models that use architectures specifically crafted for each task, significantly improving upon the state of the art in 9 out of the 12 tasks studied. For instance, we achieve absolute improvements of 8.9% on commonsense reasoning (Stories Cloze Test), 5.7% on question answering (RACE), and 1.5% on textual entailment (MultiNLI).

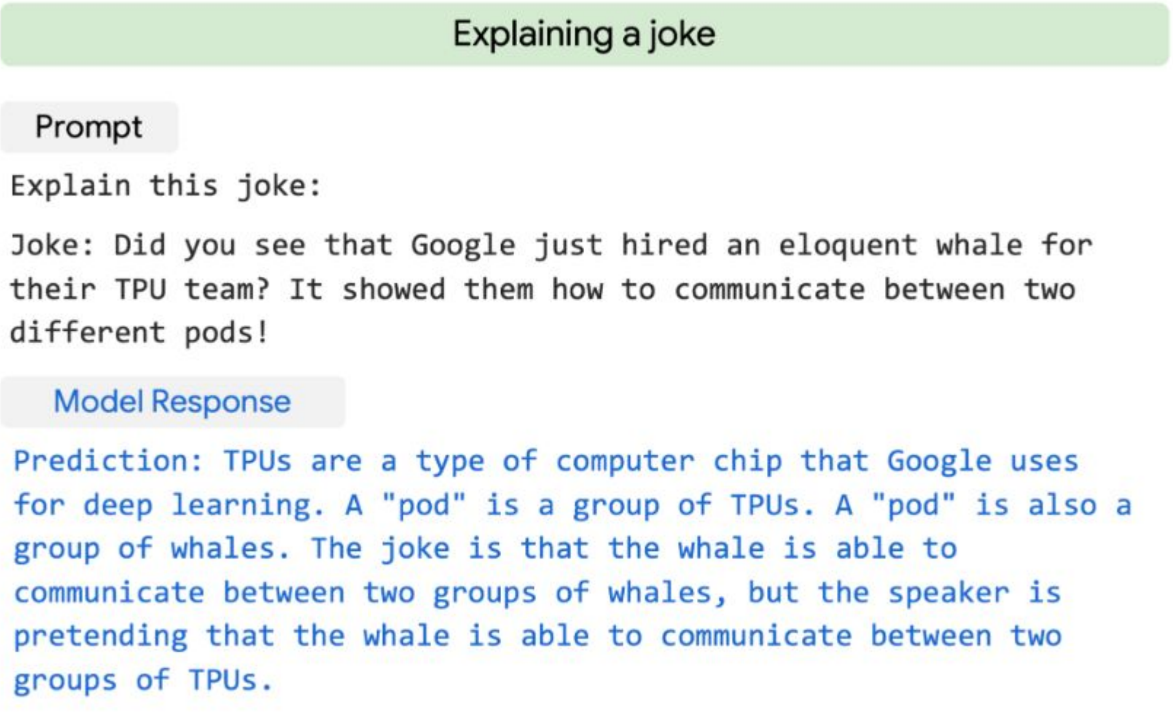
sed/language_understanding_paper.pdf



https://openai.com/blog/language-unsupervised/

Transformers and Language Models

Palm



Explaining a joke

Prompt

Explain this joke:

Joke: Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!

Model Response

Prediction: TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.

PaLM explains an original joke with two-shot prompts.

https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html

LaMDA

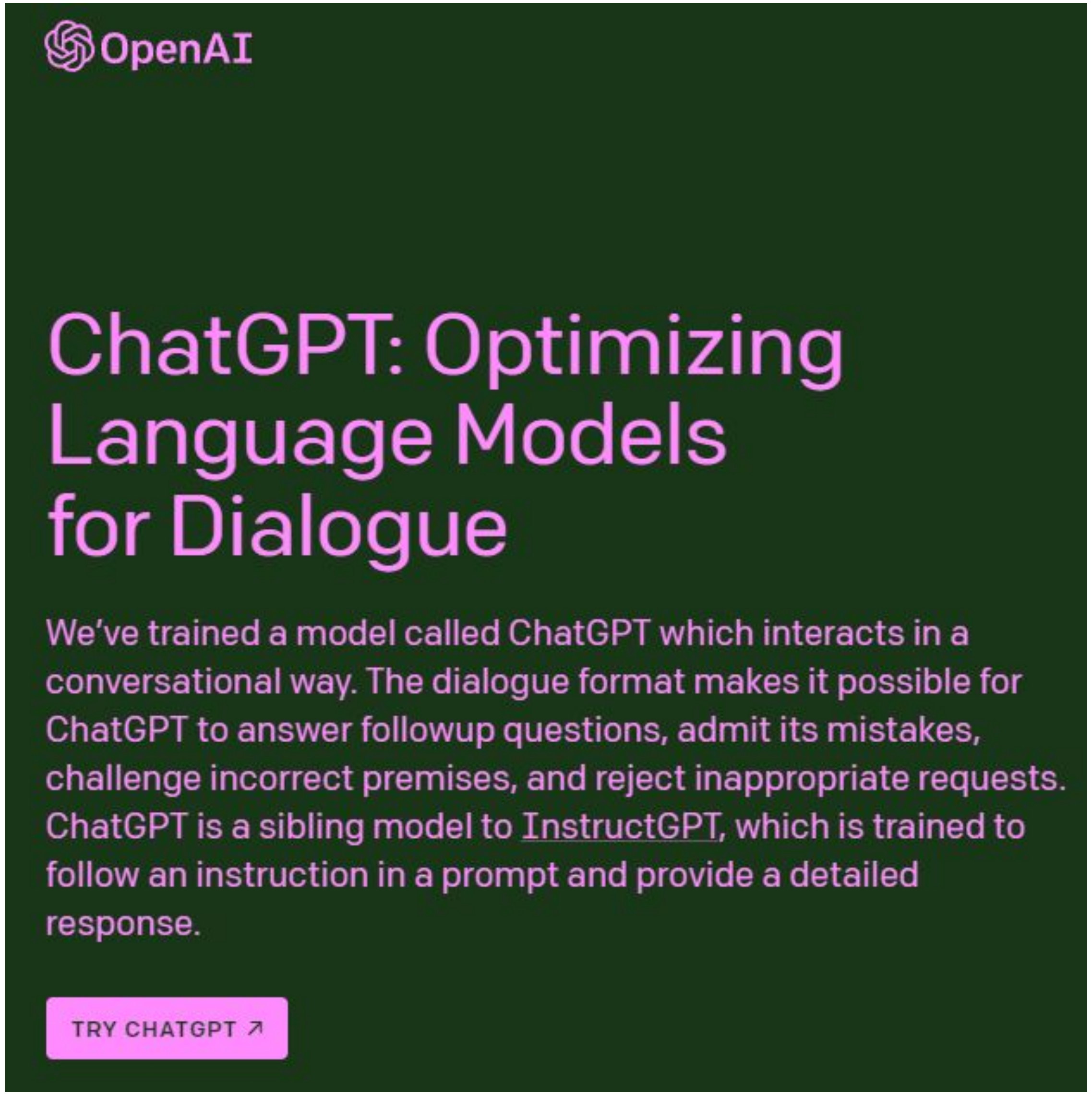
https://blog.google/technology/ai/lamda/

Transformers and Language Models

Gemini

https://gemini.google.com/app

Transformers and Language Models



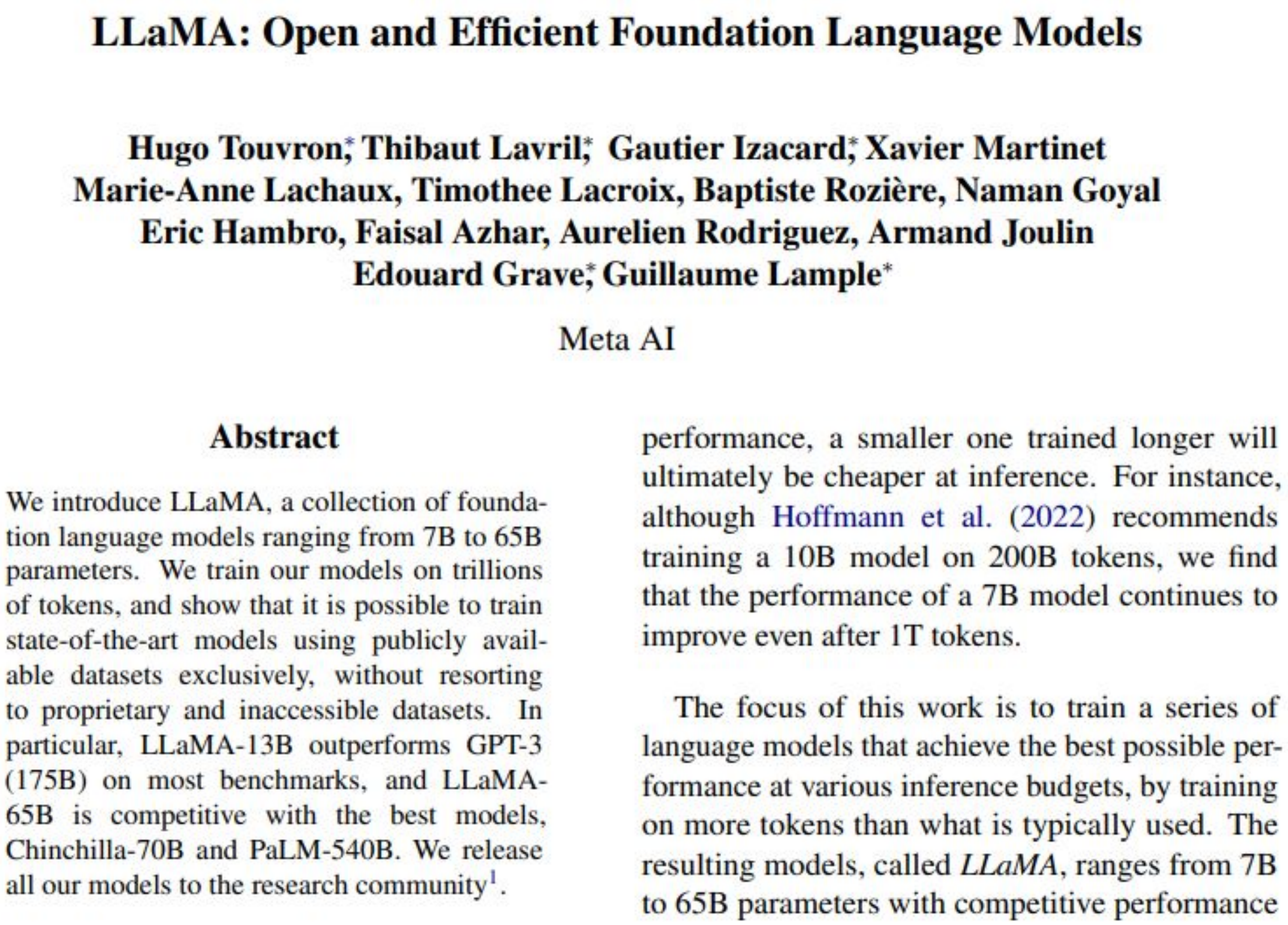
ChatGPT: Optimizing Language Models for Dialogue

We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.

ChatGPT

https://openai.com/blog/chatgpt/

LLaMA



LLaMA: Open and Efficient Foundation Language Models

Hugo Touvron*, Thibaut Lavril*, Gautier Izacard*, Xavier Martinet
Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal
Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin
Edouard Grave*, Guillaume Lample*

Meta AI

Abstract

We introduce LLaMA, a collection of foundation language models ranging from 7B to 65B parameters. We train our models on trillions of tokens, and show that it is possible to train state-of-the-art models using publicly available datasets exclusively, without resorting to proprietary and inaccessible datasets. In particular, LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla-70B and PaLM-540B. We release all our models to the research community[1].

performance, a smaller one trained longer will ultimately be cheaper at inference. For instance, although Hoffmann et al. (2022) recommends training a 10B model on 200B tokens, we find that the performance of a 7B model continues to improve even after 1T tokens.

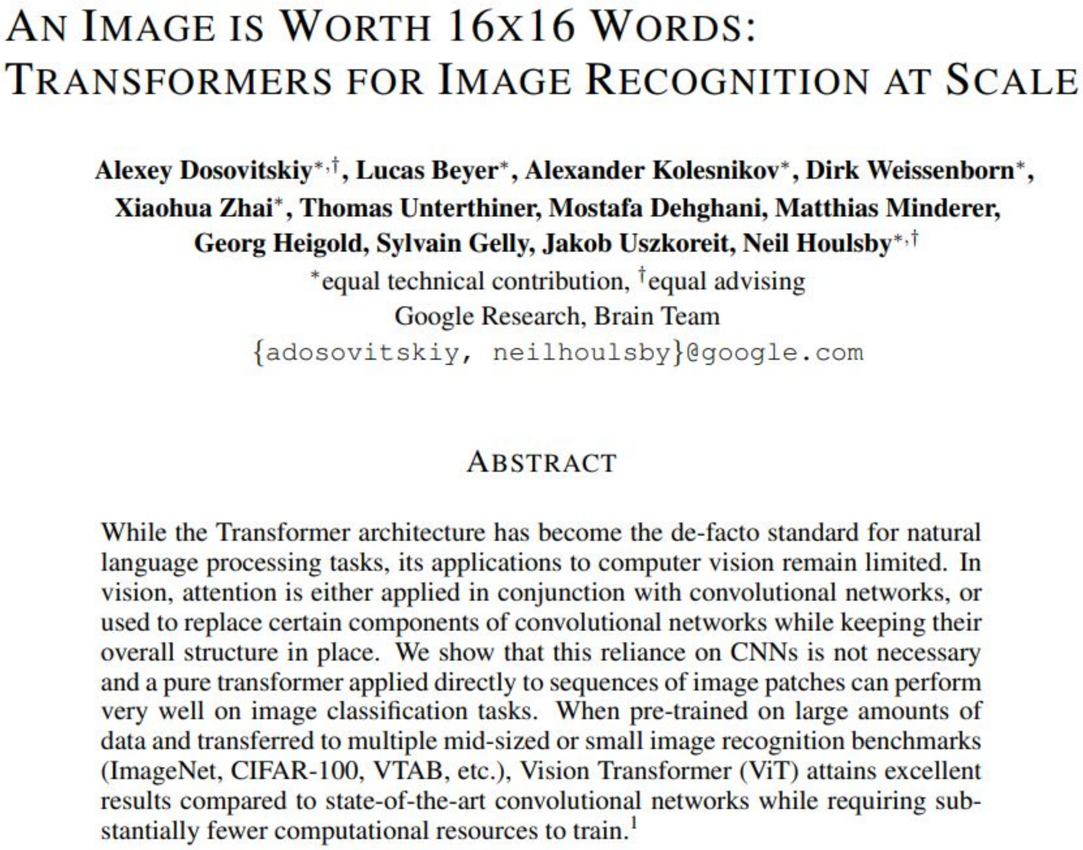
The focus of this work is to train a series of language models that achieve the best possible performance at various inference budgets, by training on more tokens than what is typically used. The resulting models, called *LLaMA*, ranges from 7B to 65B parameters with competitive performance

https://arxiv.org/pdf/2302.13971v1.pdf

- RMSNorm se utiliza para mejorar la estabilidad del entrenamiento, normalizando la entrada de cada subcapa del transformer, en lugar de normalizar la salida.
- La no linealidad de ReLU se reemplaza por la función de activación SwiGLU para mejorar el rendimiento.
- Se eliminan el embedding posicional absoluto y, en su lugar, se agregan el embedding posicional rotativo (RoPE) en cada capa de la red.

Transformers and Image Models

Vision Transformer

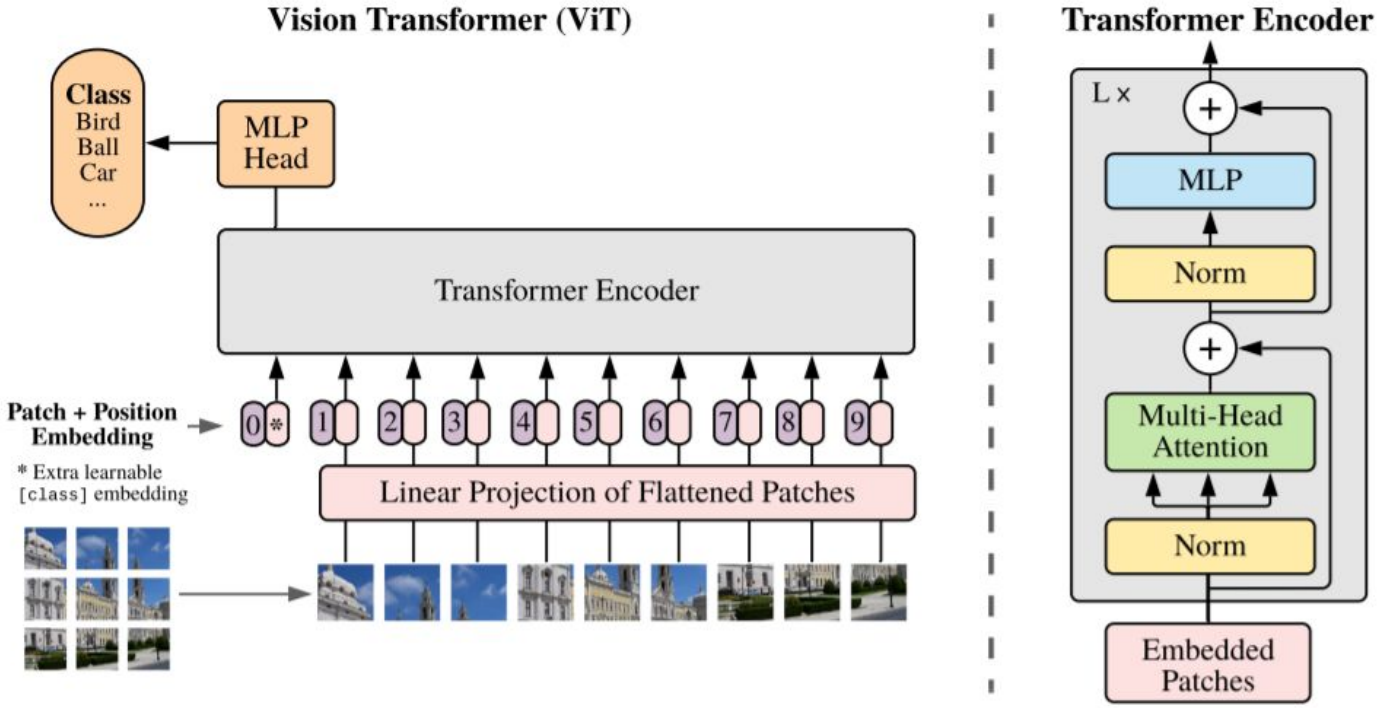


AN IMAGE IS WORTH 16X16 WORDS:
TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], **Lucas Beyer**[*], **Alexander Kolesnikov**[*], **Dirk Weissenborn**[*],
Xiaohua Zhai[*], **Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer,**
Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby[*,†]
[*]equal technical contribution, [†]equal advising
Google Research, Brain Team
`{adosovitskiy, neilhoulsby}@google.com`

ABSTRACT

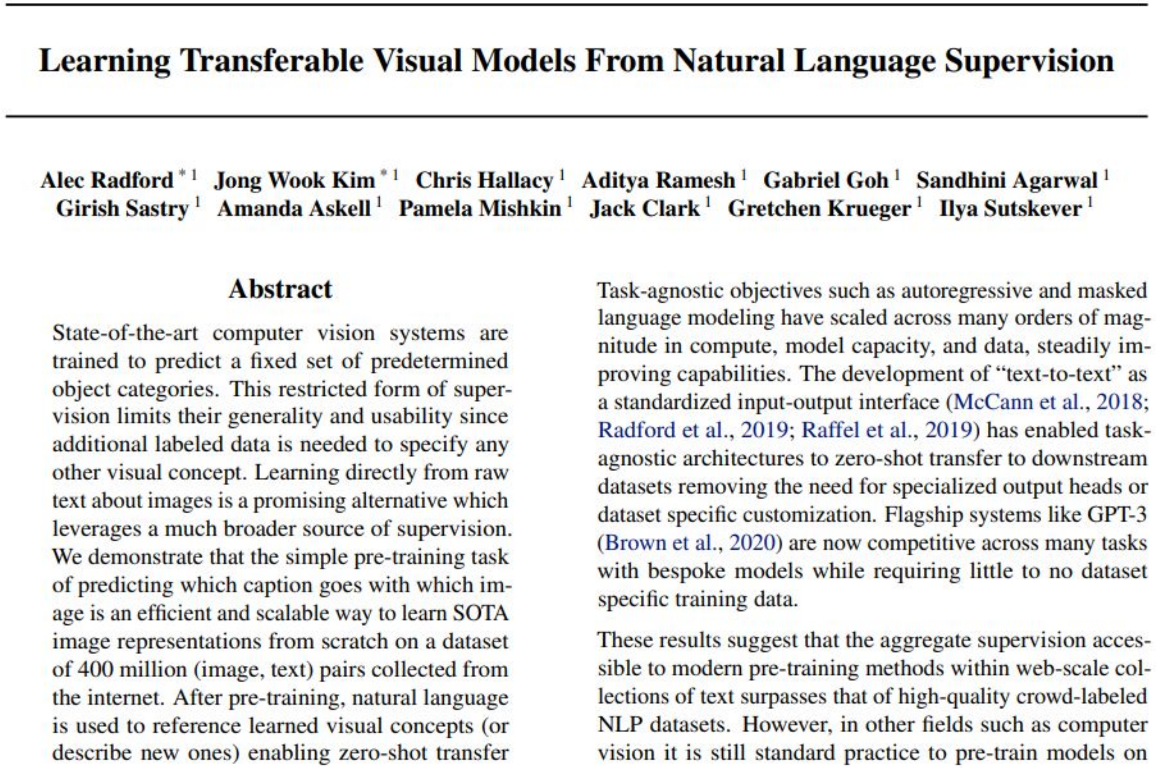
While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]



https://arxiv.org/pdf/2010.11929.pdf

Transformers and Image Generation

CLIP



Learning Transferable Visual Models From Natural Language Supervision

Alec Radford * 1 Jong Wook Kim * 1 Chris Hallacy 1 Aditya Ramesh 1 Gabriel Goh 1 Sandhini Agarwal 1
Girish Sastry 1 Amanda Askell 1 Pamela Mishkin 1 Jack Clark 1 Gretchen Krueger 1 Ilya Sutskever 1

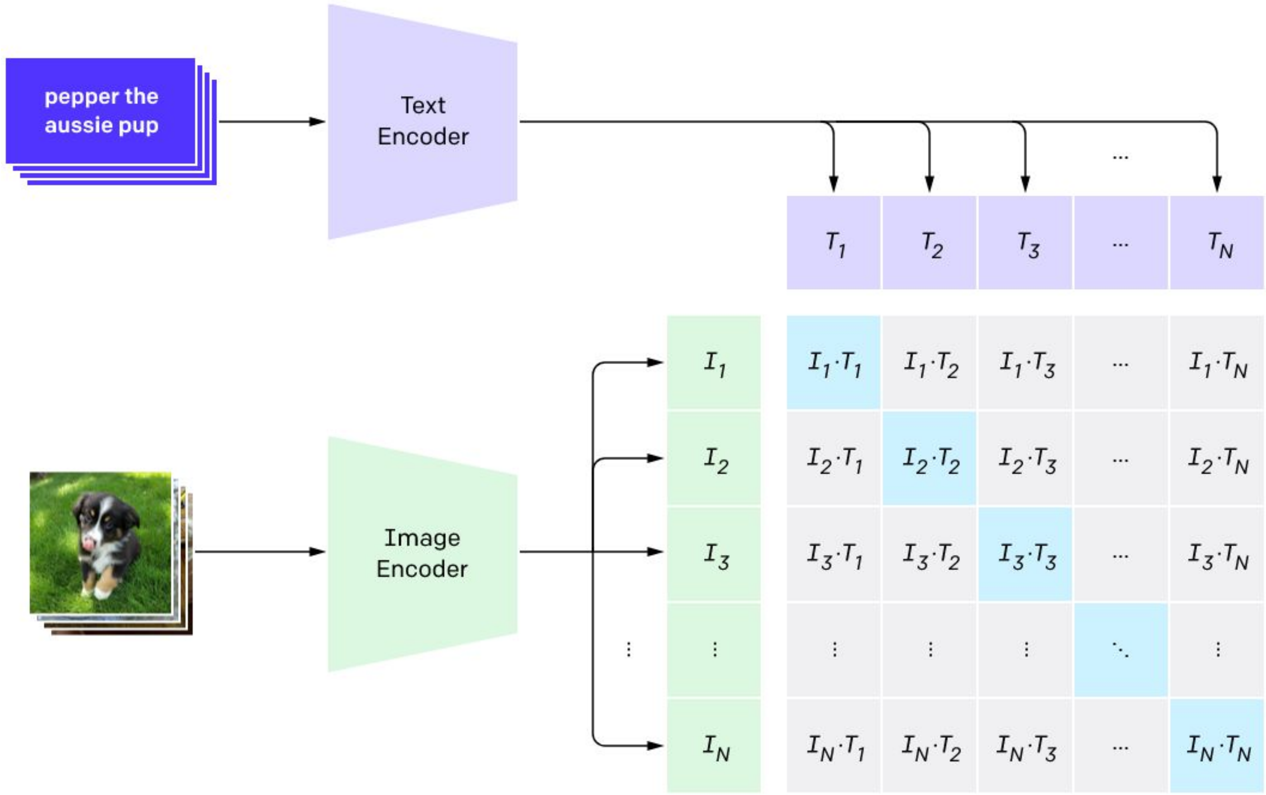
Abstract

State-of-the-art computer vision systems are trained to predict a fixed set of predetermined object categories. This restricted form of supervision limits their generality and usability since additional labeled data is needed to specify any other visual concept. Learning directly from raw text about images is a promising alternative which leverages a much broader source of supervision. We demonstrate that the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn SOTA image representations from scratch on a dataset of 400 million (image, text) pairs collected from the internet. After pre-training, natural language is used to reference learned visual concepts (or describe new ones) enabling zero-shot transfer

Task-agnostic objectives such as autoregressive and masked language modeling have scaled across many orders of magnitude in compute, model capacity, and data, steadily improving capabilities. The development of "text-to-text" as a standardized input-output interface (McCann et al., 2018; Radford et al., 2019; Raffel et al., 2019) has enabled task-agnostic architectures to zero-shot transfer to downstream datasets removing the need for specialized output heads or dataset specific customization. Flagship systems like GPT-3 (Brown et al., 2020) are now competitive across many tasks with bespoke models while requiring little to no dataset specific training data.

These results suggest that the aggregate supervision accessible to modern pre-training methods within web-scale collections of text surpasses that of high-quality crowd-labeled NLP datasets. However, in other fields such as computer vision it is still standard practice to pre-train models on

1. Contrastive pre-training



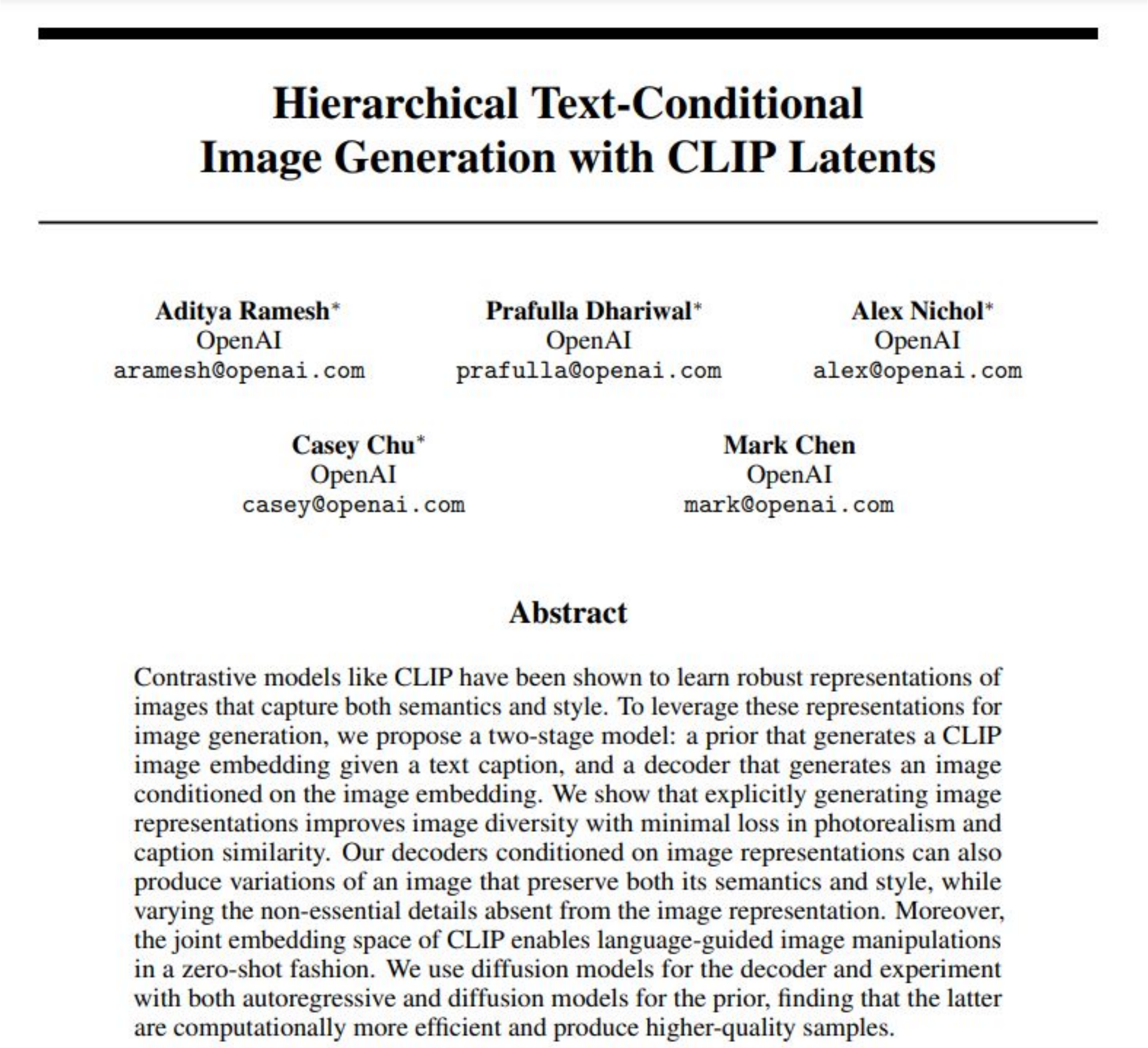
	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

https://openai.com/blog/clip/

https://arxiv.org/pdf/2010.11929.pdf

Transformers and Image Generation

Dalle-2



Hierarchical Text-Conditional Image Generation with CLIP Latents

Aditya Ramesh[*]
OpenAI
aramesh@openai.com

Prafulla Dhariwal[*]
OpenAI
prafulla@openai.com

Alex Nichol[*]
OpenAI
alex@openai.com

Casey Chu[*]
OpenAI
casey@openai.com

Mark Chen
OpenAI
mark@openai.com

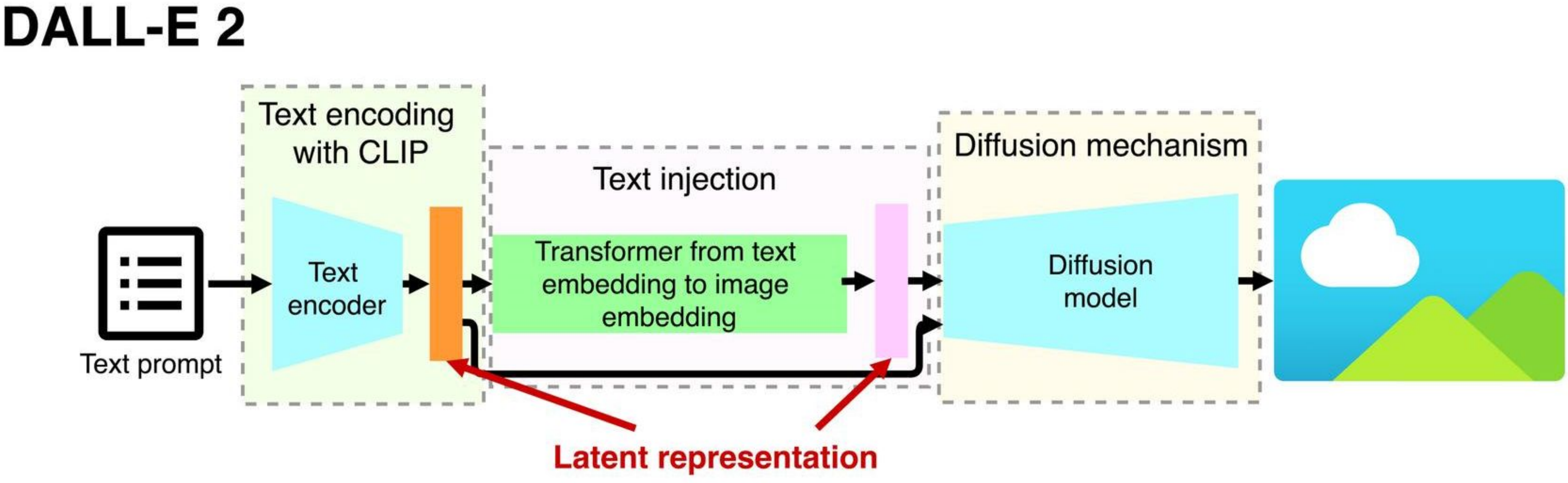
Abstract

Contrastive models like CLIP have been shown to learn robust representations of images that capture both semantics and style. To leverage these representations for image generation, we propose a two-stage model: a prior that generates a CLIP image embedding given a text caption, and a decoder that generates an image conditioned on the image embedding. We show that explicitly generating image representations improves image diversity with minimal loss in photorealism and caption similarity. Our decoders conditioned on image representations can also produce variations of an image that preserve both its semantics and style, while varying the non-essential details absent from the image representation. Moreover, the joint embedding space of CLIP enables language-guided image manipulations in a zero-shot fashion. We use diffusion models for the decoder and experiment with both autoregressive and diffusion models for the prior, finding that the latter are computationally more efficient and produce higher-quality samples.

https://openai.com/dall-e-2/

https://arxiv.org/pdf/2204.06125.pdf

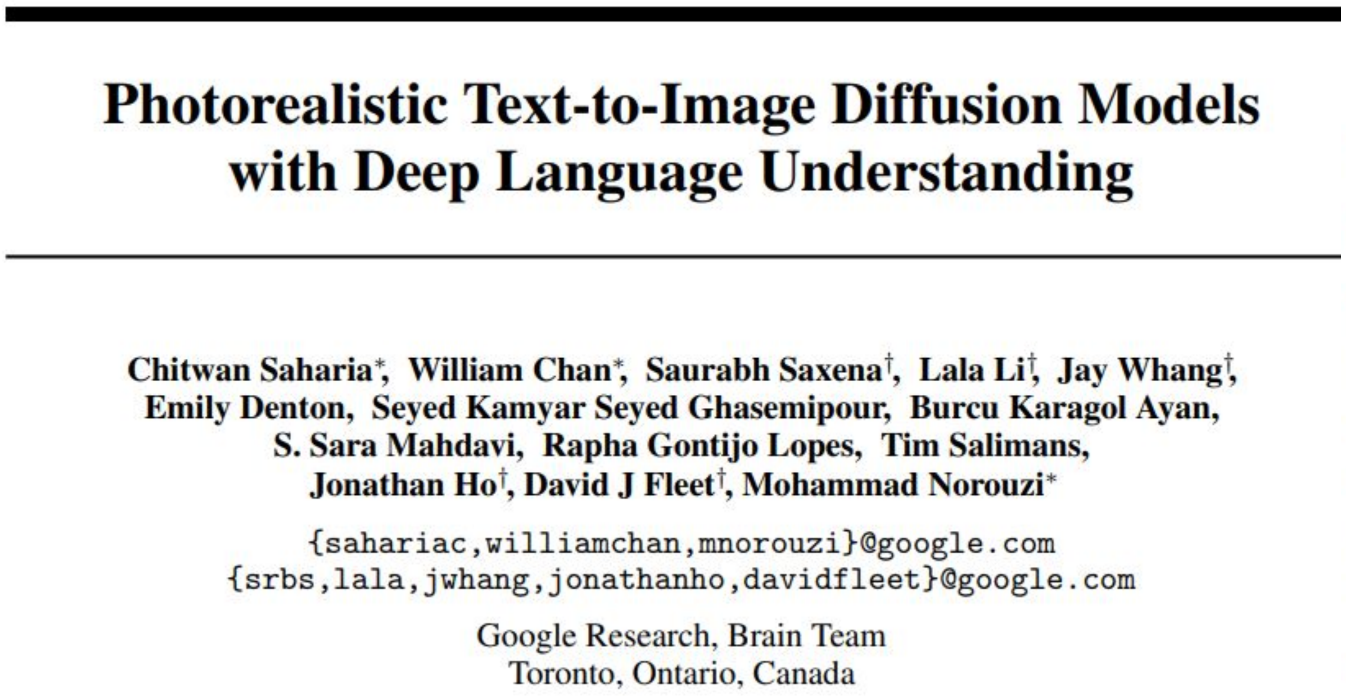
Transformers and Image Generation



https://newsletter.theaiedge.io/p/everything-you-need-to-know-about

Transformers and Image Generation

Imagen

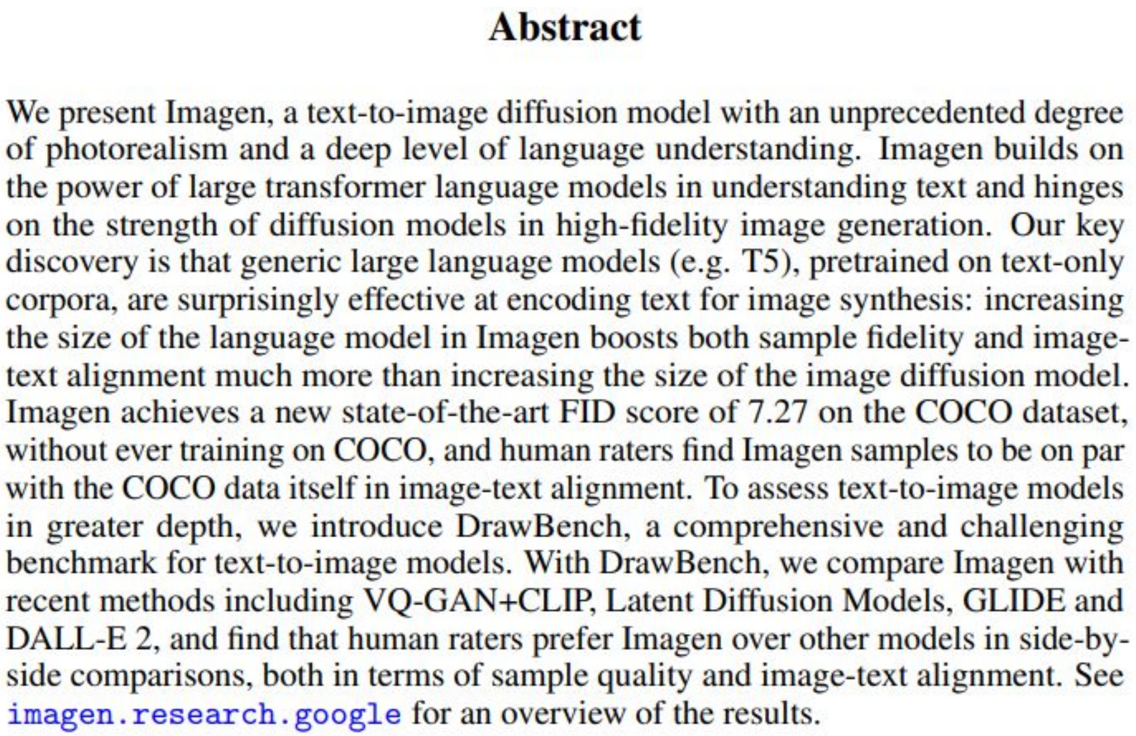


Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

Chitwan Saharia*, **William Chan***, **Saurabh Saxena**†, **Lala Li**†, **Jay Whang**†,
Emily Denton, **Seyed Kamyar Seyed Ghasemipour**, **Burcu Karagol Ayan**,
S. Sara Mahdavi, **Rapha Gontijo Lopes**, **Tim Salimans**,
Jonathan Ho†, **David J Fleet**†, **Mohammad Norouzi***

`{sahariac,williamchan,mnorouzi}@google.com`
`{srbs,lala,jwhang,jonathanho,davidfleet}@google.com`

Google Research, Brain Team
Toronto, Ontario, Canada

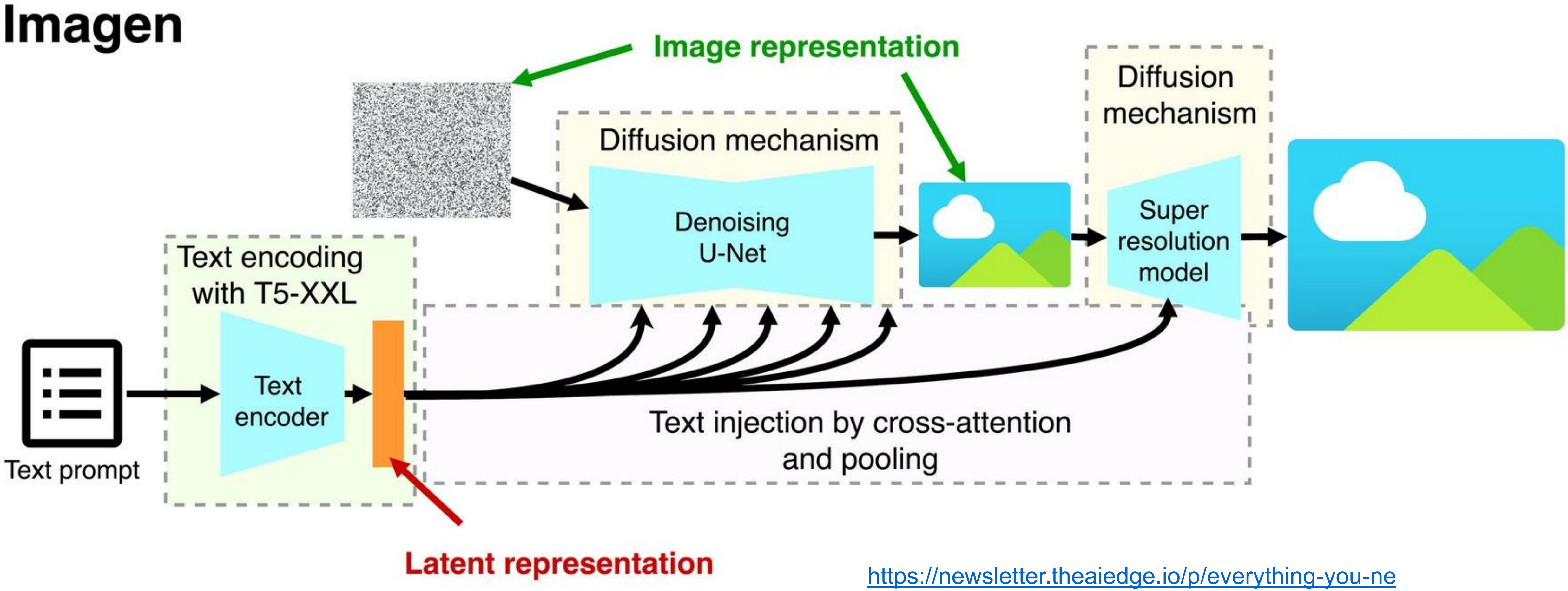


Abstract

We present Imagen, a text-to-image diffusion model with an unprecedented degree of photorealism and a deep level of language understanding. Imagen builds on the power of large transformer language models in understanding text and hinges on the strength of diffusion models in high-fidelity image generation. Our key discovery is that generic large language models (e.g. T5), pretrained on text-only corpora, are surprisingly effective at encoding text for image synthesis: increasing the size of the language model in Imagen boosts both sample fidelity and image-text alignment much more than increasing the size of the image diffusion model. Imagen achieves a new state-of-the-art FID score of 7.27 on the COCO dataset, without ever training on COCO, and human raters find Imagen samples to be on par with the COCO data itself in image-text alignment. To assess text-to-image models in greater depth, we introduce DrawBench, a comprehensive and challenging benchmark for text-to-image models. With DrawBench, we compare Imagen with recent methods including VQ-GAN+CLIP, Latent Diffusion Models, GLIDE and DALL-E 2, and find that human raters prefer Imagen over other models in side-by-side comparisons, both in terms of sample quality and image-text alignment. See `imagen.research.google` for an overview of the results.

https://arxiv.org/pdf/2205.11487.pdf

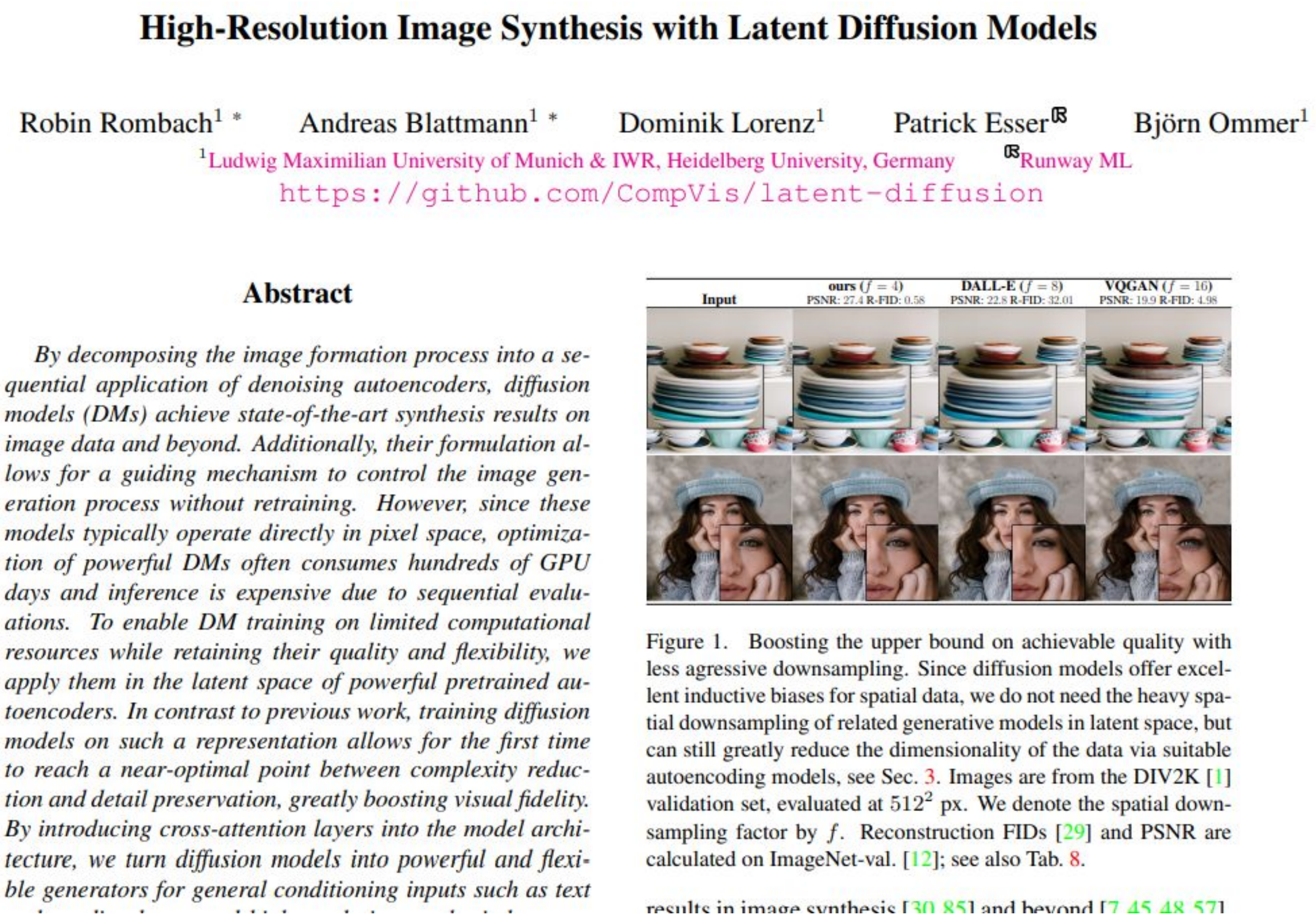
Transformers and Image Generation



https://newsletter.theaiedge.io/p/everything-you-need-to-know-about

Transformers and Image Generation

Stable Diffusion



High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach[1] * Andreas Blattmann[1] * Dominik Lorenz[1] Patrick Esser[R] Björn Ommer[1]
[1] Ludwig Maximilian University of Munich & IWR, Heidelberg University, Germany [R] Runway ML
https://github.com/CompVis/latent-diffusion

Abstract

By decomposing the image formation process into a sequential application of denoising autoencoders, diffusion models (DMs) achieve state-of-the-art synthesis results on image data and beyond. Additionally, their formulation allows for a guiding mechanism to control the image generation process without retraining. However, since these models typically operate directly in pixel space, optimization of powerful DMs often consumes hundreds of GPU days and inference is expensive due to sequential evaluations. To enable DM training on limited computational resources while retaining their quality and flexibility, we apply them in the latent space of powerful pretrained autoencoders. In contrast to previous work, training diffusion models on such a representation allows for the first time to reach a near-optimal point between complexity reduction and detail preservation, greatly boosting visual fidelity. By introducing cross-attention layers into the model architecture, we turn diffusion models into powerful and flexible generators for general conditioning inputs such as text

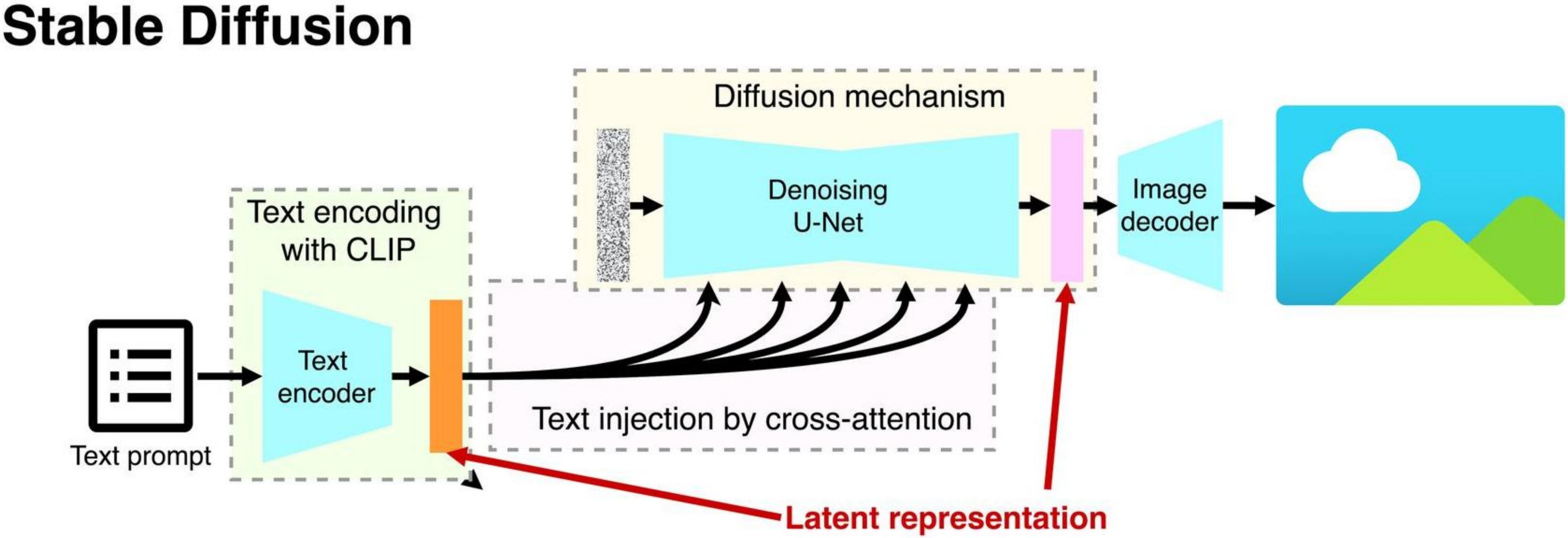
Figure 1. Boosting the upper bound on achievable quality with less agressive downsampling. Since diffusion models offer excellent inductive biases for spatial data, we do not need the heavy spatial downsampling of related generative models in latent space, but can still greatly reduce the dimensionality of the data via suitable autoencoding models, see Sec. 3. Images are from the DIV2K [1] validation set, evaluated at 512^2 px. We denote the spatial downsampling factor by f. Reconstruction FIDs [29] and PSNR are calculated on ImageNet-val. [12]; see also Tab. 8.

results in image synthesis [30, 85] and beyond [7, 45, 48, 57]

https://jalammar.github.io/illustrated-stable-diffusion/

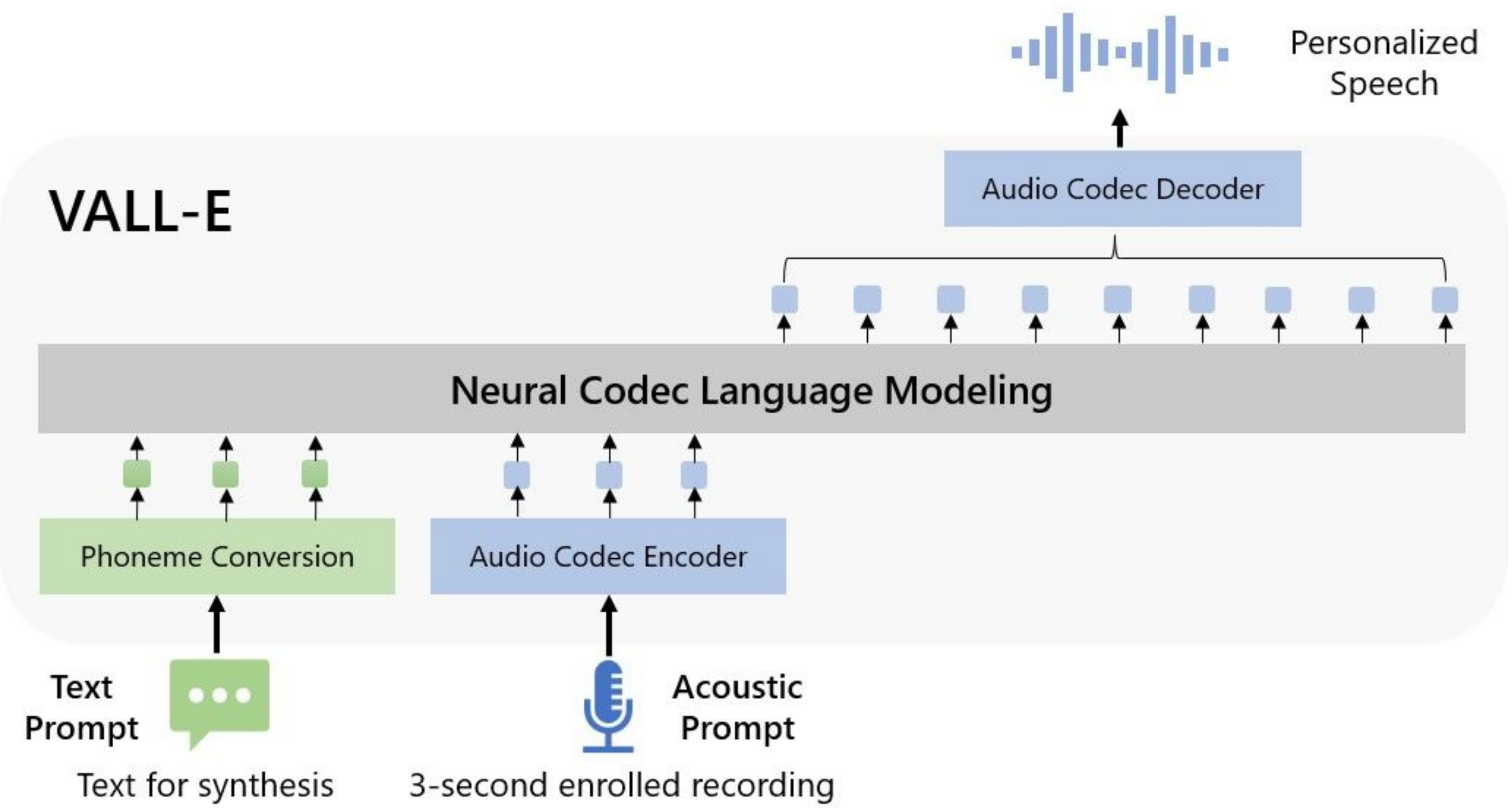
https://arxiv.org/pdf/2112.10752.pdf

Transformers and Image Generation



https://newsletter.theaiedge.io/p/everything-you-need-to-know-about

Transformers and Sound Generation

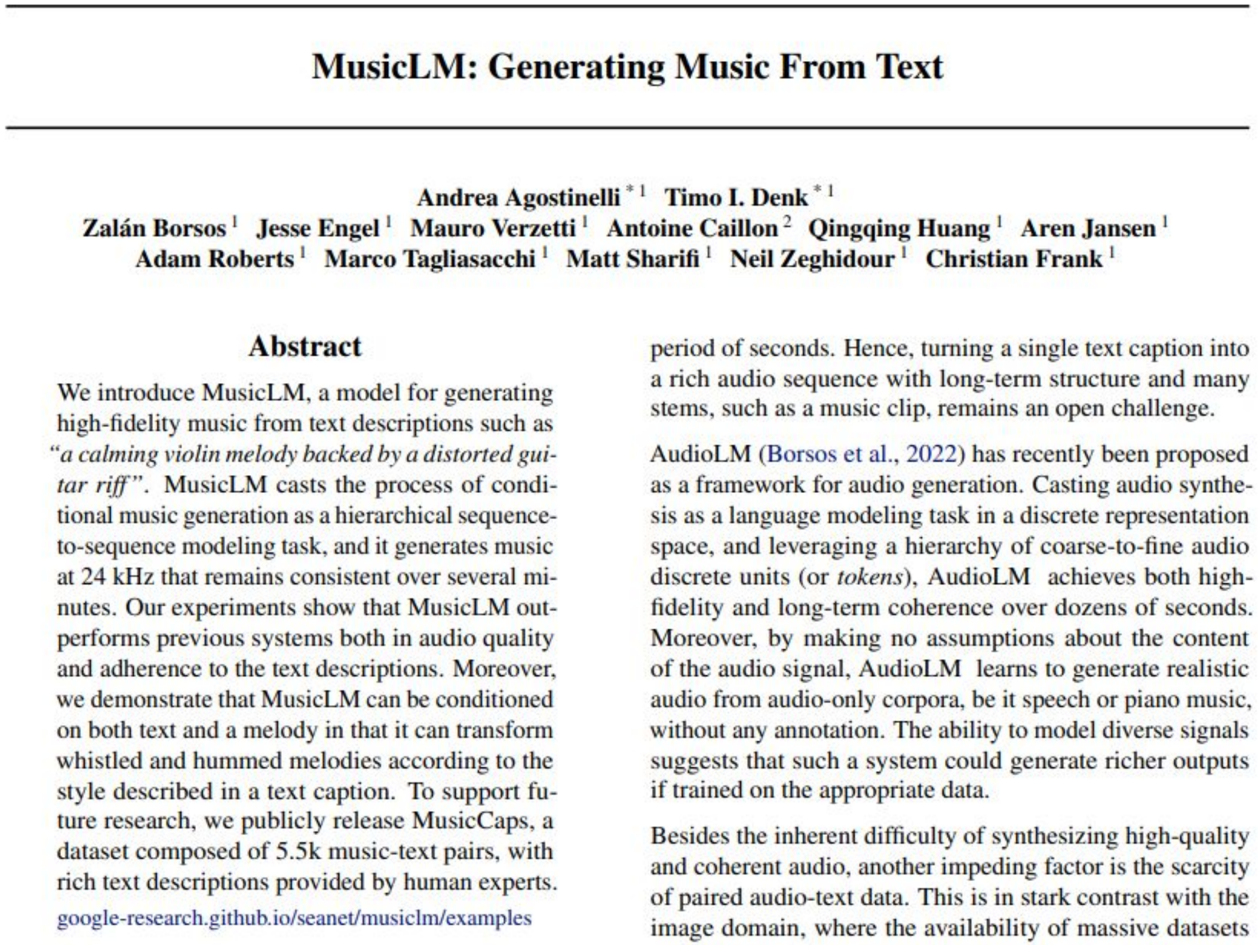


VALL-E

https://valle-demo.github.io/

Transformers and Sound Generation

Music LM



MusicLM: Generating Music From Text

Andrea Agostinelli *[1] Timo I. Denk *[1]
Zalán Borsos [1] Jesse Engel [1] Mauro Verzetti [1] Antoine Caillon [2] Qingqing Huang [1] Aren Jansen [1]
Adam Roberts [1] Marco Tagliasacchi [1] Matt Sharifi [1] Neil Zeghidour [1] Christian Frank [1]

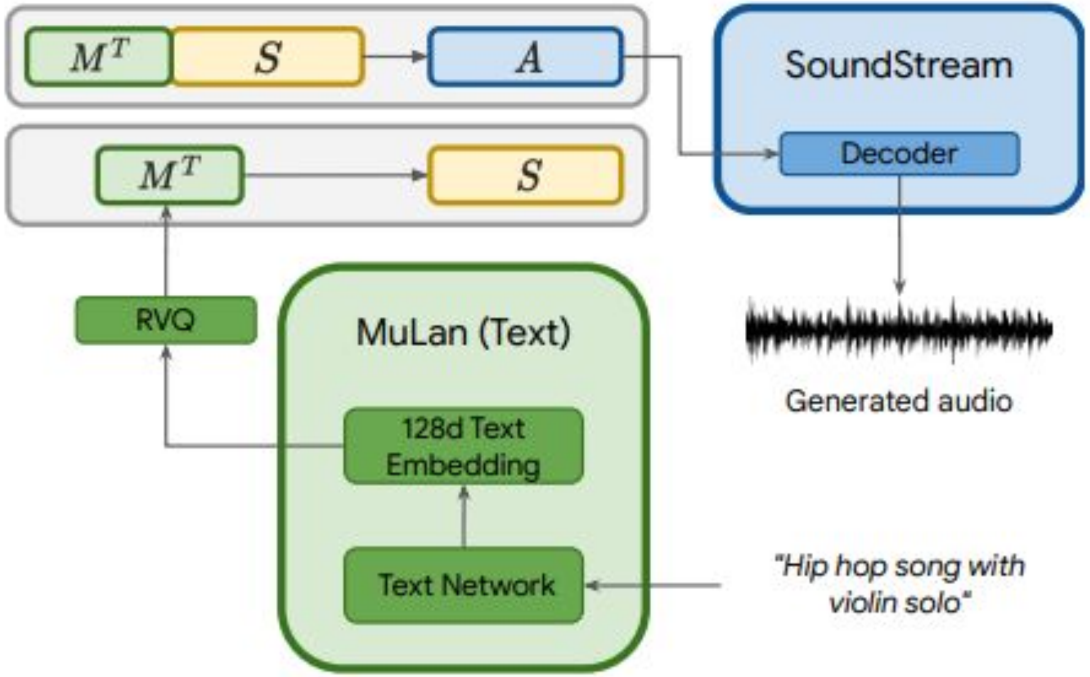
Abstract

We introduce MusicLM, a model for generating high-fidelity music from text descriptions such as *"a calming violin melody backed by a distorted guitar riff"*. MusicLM casts the process of conditional music generation as a hierarchical sequence-to-sequence modeling task, and it generates music at 24 kHz that remains consistent over several minutes. Our experiments show that MusicLM outperforms previous systems both in audio quality and adherence to the text descriptions. Moreover, we demonstrate that MusicLM can be conditioned on both text and a melody in that it can transform whistled and hummed melodies according to the style described in a text caption. To support future research, we publicly release MusicCaps, a dataset composed of 5.5k music-text pairs, with rich text descriptions provided by human experts.
google-research.github.io/seanet/musiclm/examples

period of seconds. Hence, turning a single text caption into a rich audio sequence with long-term structure and many stems, such as a music clip, remains an open challenge.

AudioLM (Borsos et al., 2022) has recently been proposed as a framework for audio generation. Casting audio synthesis as a language modeling task in a discrete representation space, and leveraging a hierarchy of coarse-to-fine audio discrete units (or *tokens*), AudioLM achieves both high-fidelity and long-term coherence over dozens of seconds. Moreover, by making no assumptions about the content of the audio signal, AudioLM learns to generate realistic audio from audio-only corpora, be it speech or piano music, without any annotation. The ability to model diverse signals suggests that such a system could generate richer outputs if trained on the appropriate data.

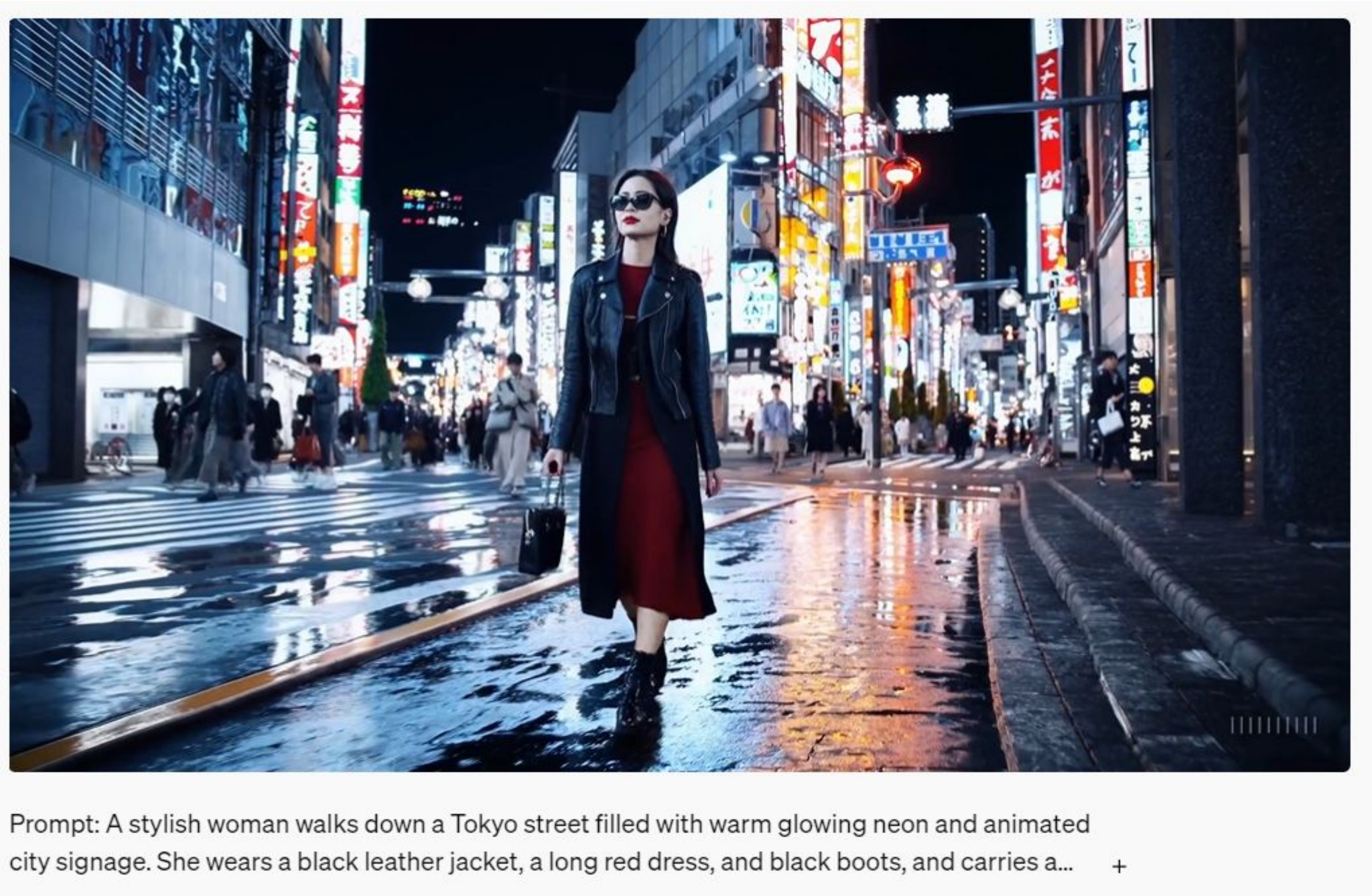
Besides the inherent difficulty of synthesizing high-quality and coherent audio, another impeding factor is the scarcity of paired audio-text data. This is in stark contrast with the image domain, where the availability of massive datasets



https://arxiv.org/pdf/2301.11325.pdf

Transformers and Video Generation

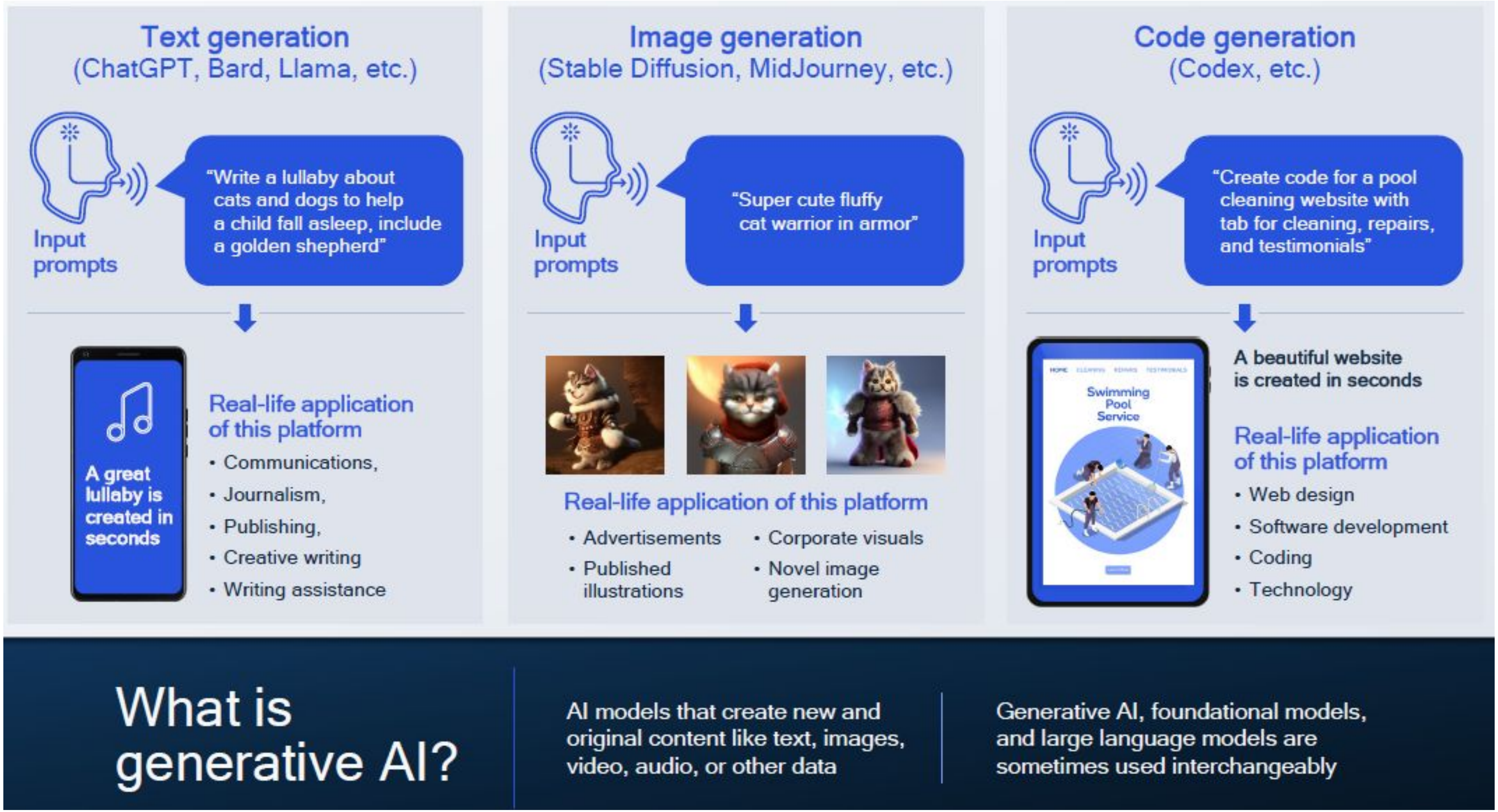
Sora



Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a... +

https://openai.com/index/sora

Generative Artificial Intelligence



Text generation
(ChatGPT, Bard, Llama, etc.)

Input prompts

"Write a lullaby about cats and dogs to help a child fall asleep, include a golden shepherd"

A great lullaby is created in seconds

Real-life application of this platform

- Communications,
- Journalism,
- Publishing,
- Creative writing
- Writing assistance

Image generation
(Stable Diffusion, MidJourney, etc.)

Input prompts

"Super cute fluffy cat warrior in armor"

Real-life application of this platform

- Advertisements
- Published illustrations
- Corporate visuals
- Novel image generation

Code generation
(Codex, etc.)

Input prompts

"Create code for a pool cleaning website with tab for cleaning, repairs, and testimonials"

Swimming Pool Service

A beautiful website is created in seconds

Real-life application of this platform

- Web design
- Software development
- Coding
- Technology

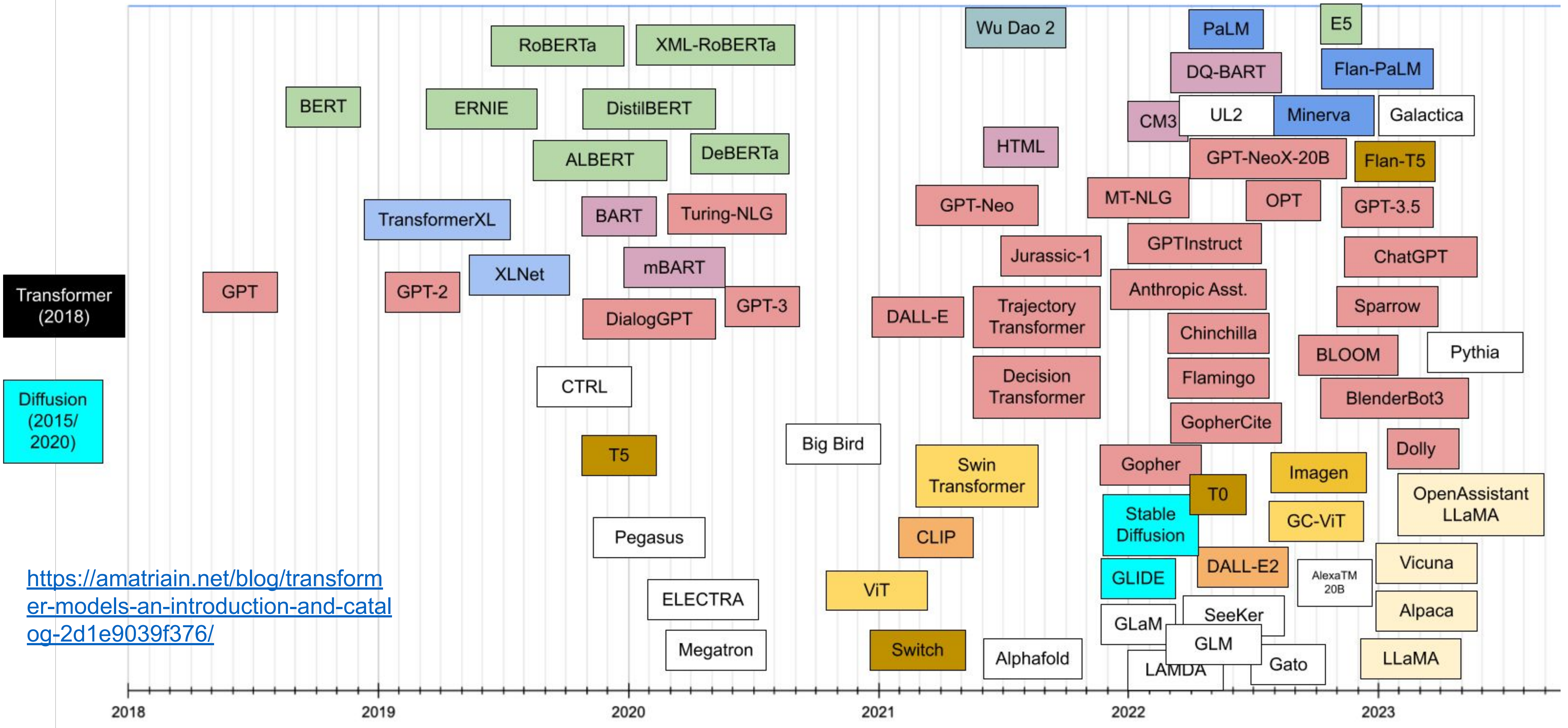
What is generative AI?

AI models that create new and original content like text, images, video, audio, or other data

Generative AI, foundational models, and large language models are sometimes used interchangeably

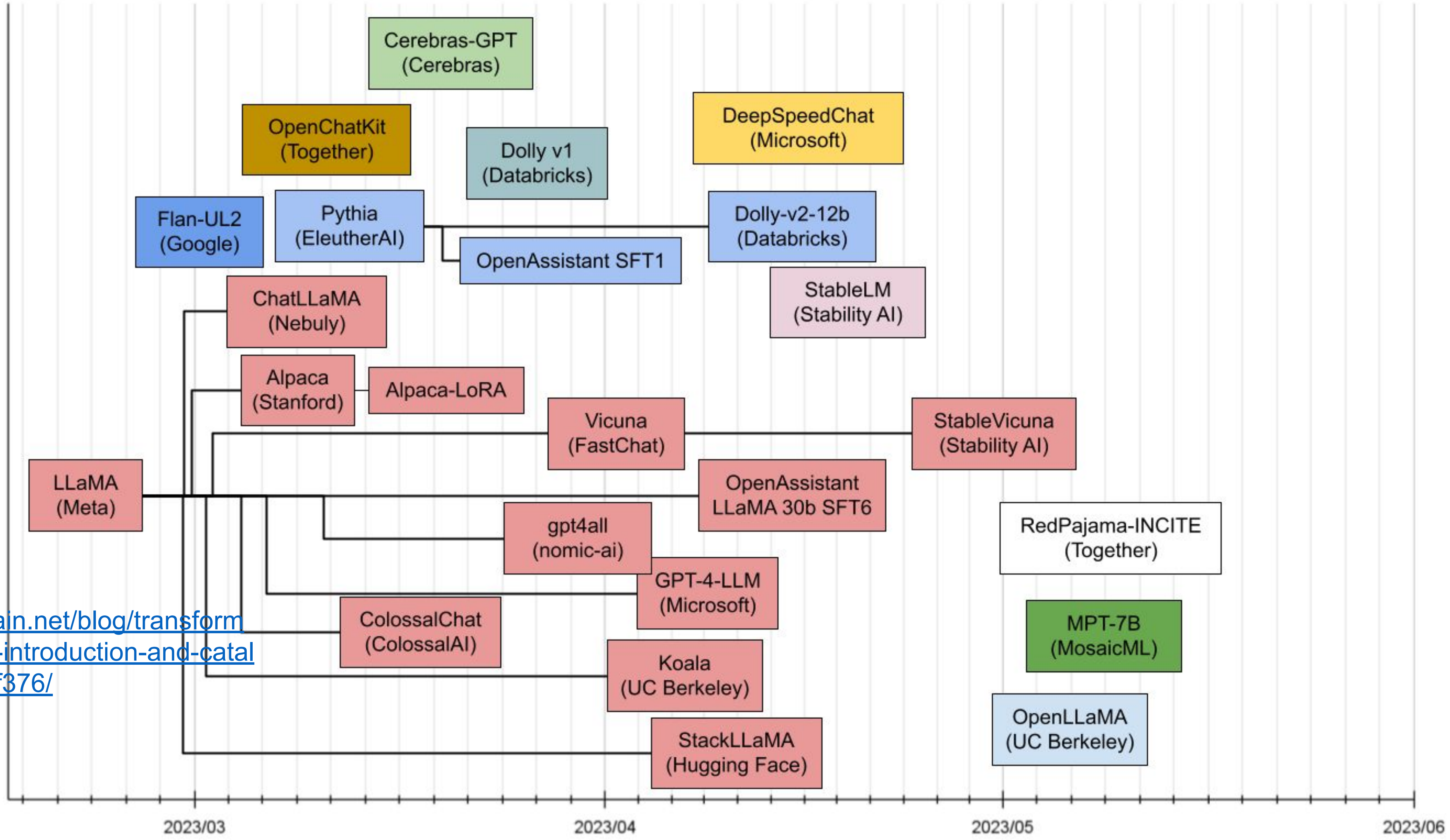
Generative AI at the edge. Joseph Soriaga Qualcomm Technologies, Inc. Nov. 8 2023

Transformers ZOO



https://amatriain.net/blog/transformer-models-an-introduction-and-catalog-2d1e9039f376/

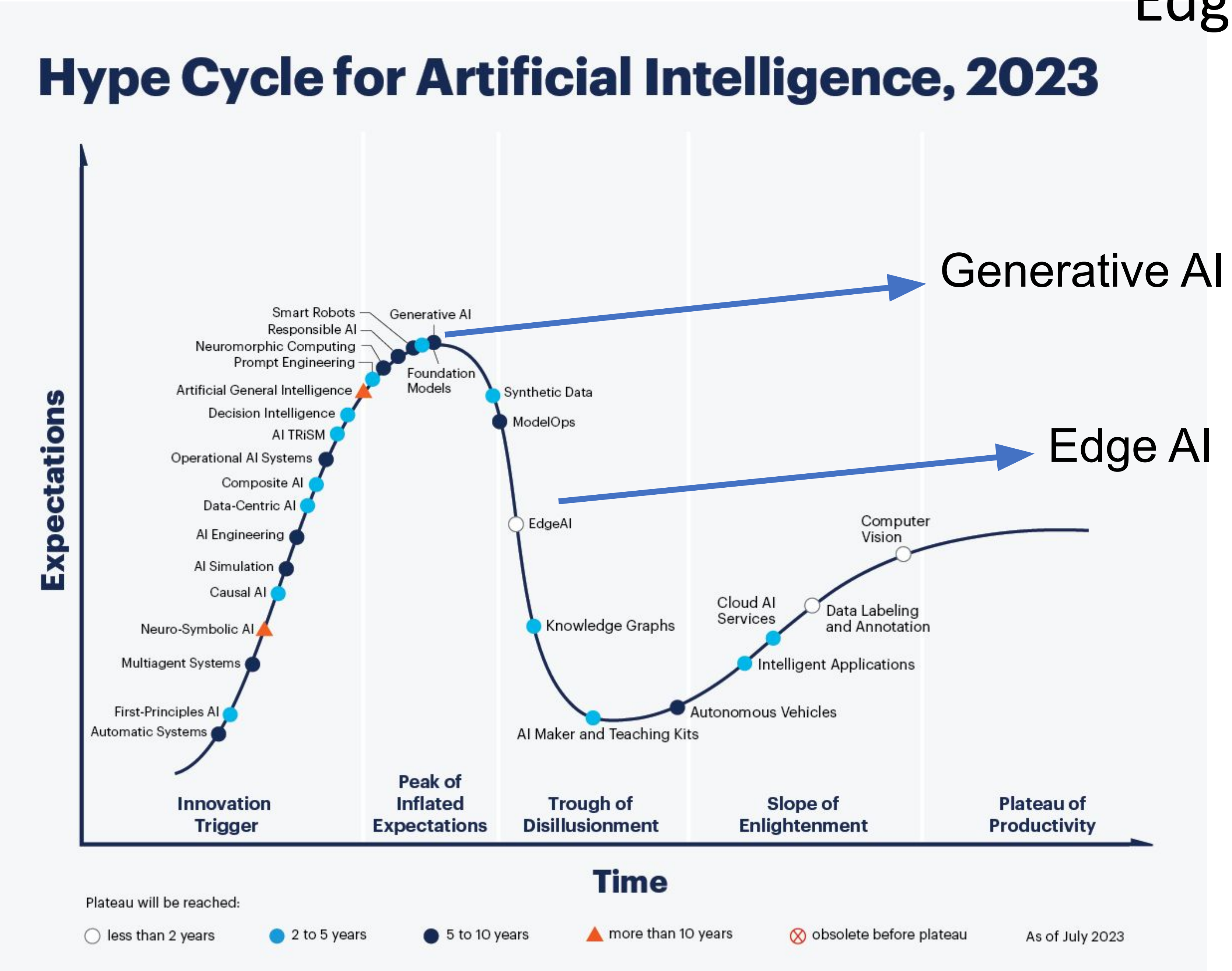
Transformers ZOO (Open Source Models)



https://amatriain.net/blog/transformer-models-an-introduction-and-catalog-2d1e9039f376/

“Tiny” and “large” meet together

Edge AI and Generative AI

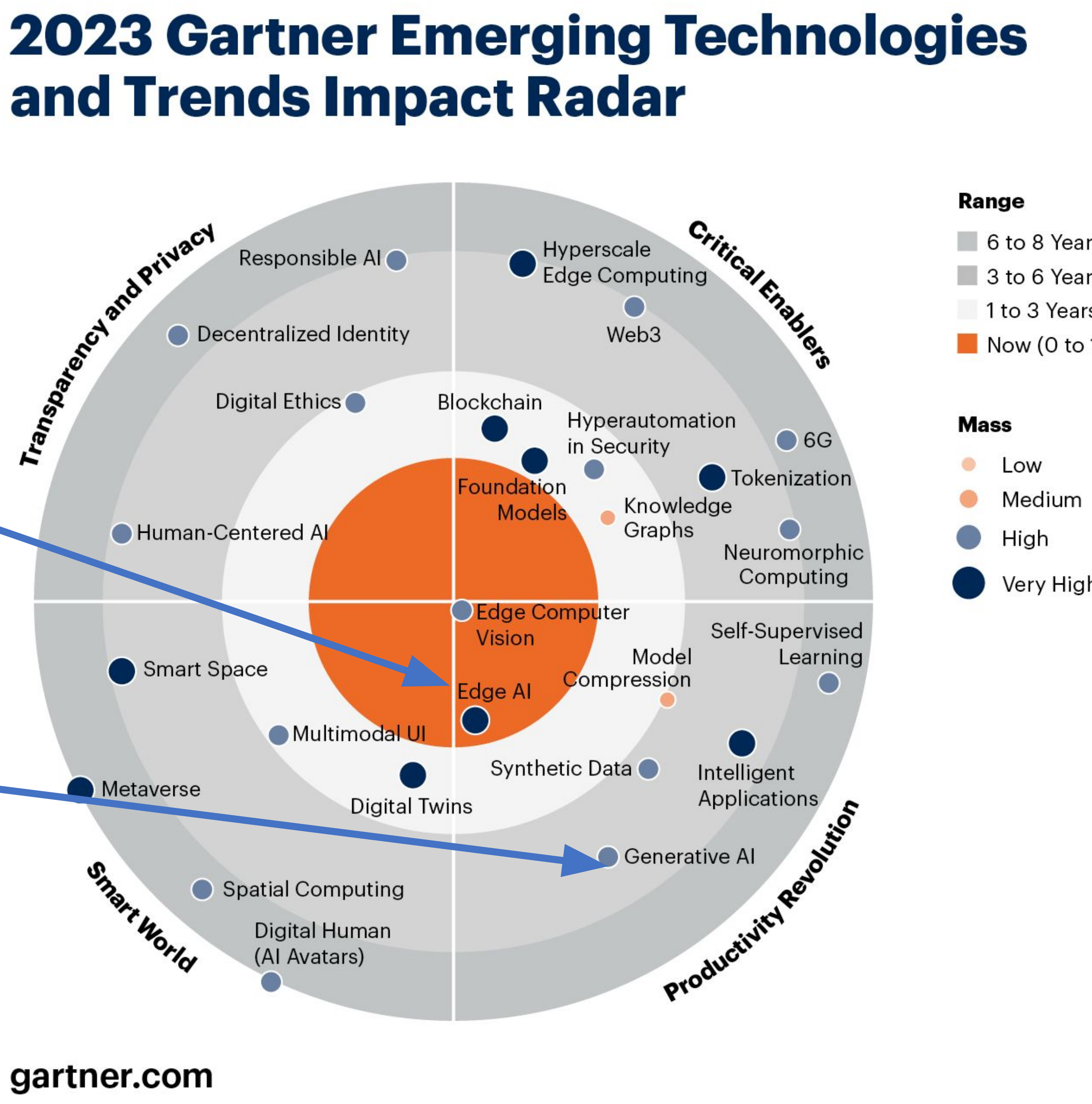


Generative AI

Edge AI

https://www.gartner.com/en/articles/what-s-new-in-artificial-intelligence-from-the-2023-gartner-hype-cycle

Edge AI and Generative AI


2023 Gartner Emerging Technologies and Trends Impact Radar

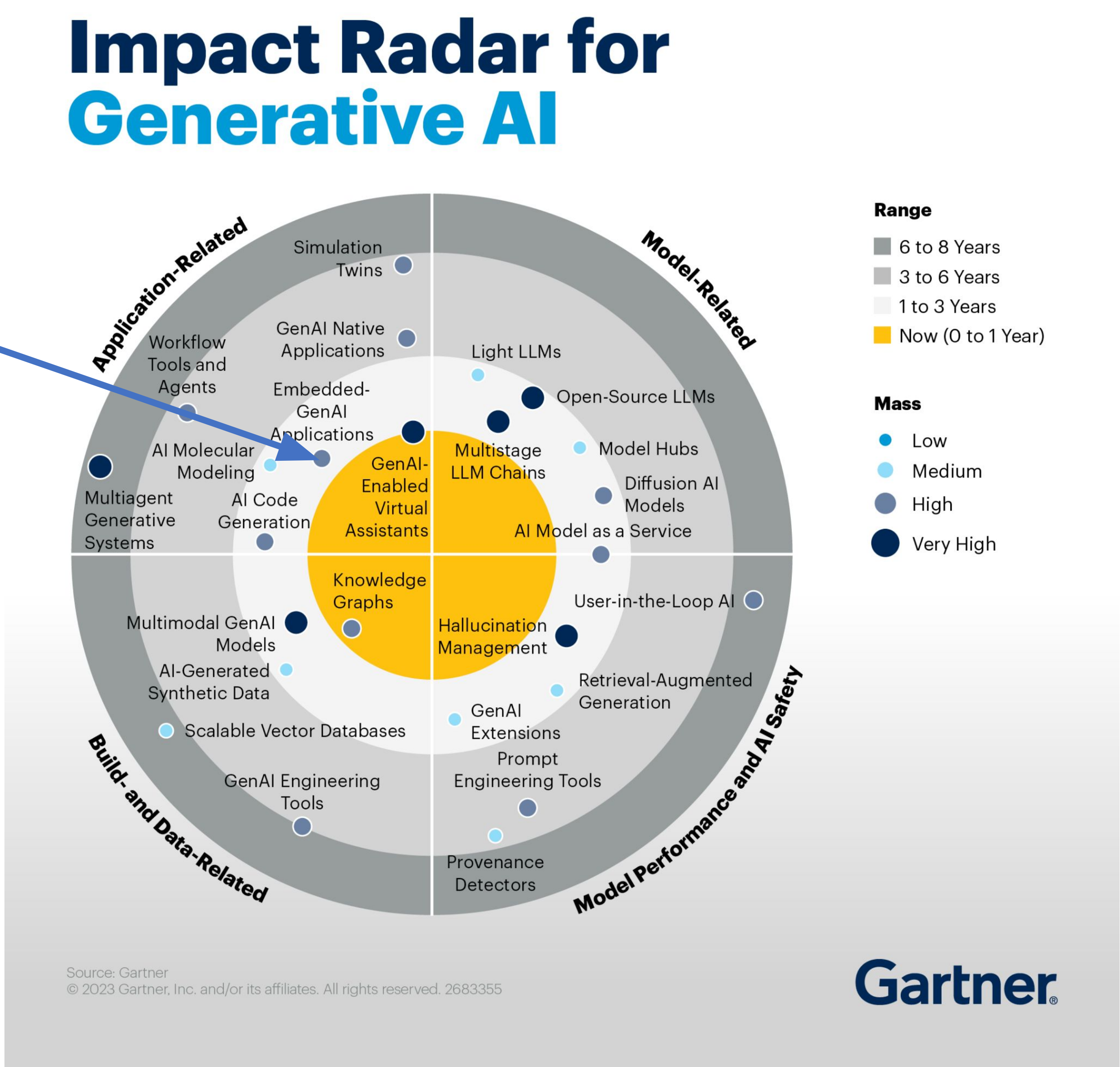
gartner.com

Edge AI

Generative AI

https://www.gartner.com/en/articles/4-emerging-technologies-you-need-to-know-about

Embedded GenAI



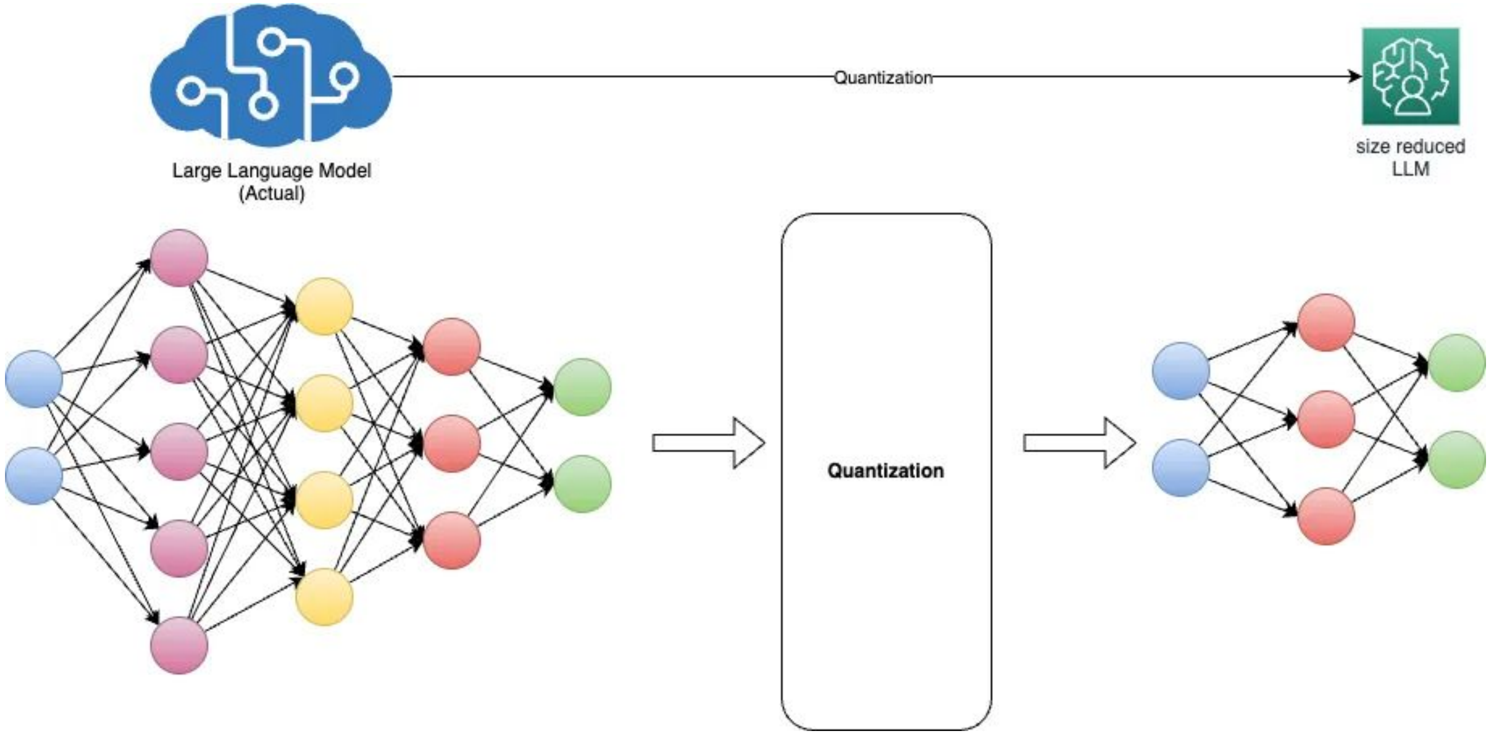
Source: Gartner

© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. 2683355

https://www.gartner.com/en/articles/understand-and-exploit-gen-ai-with-gartner-s-new-impact-radar

Edge AI and Generative AI

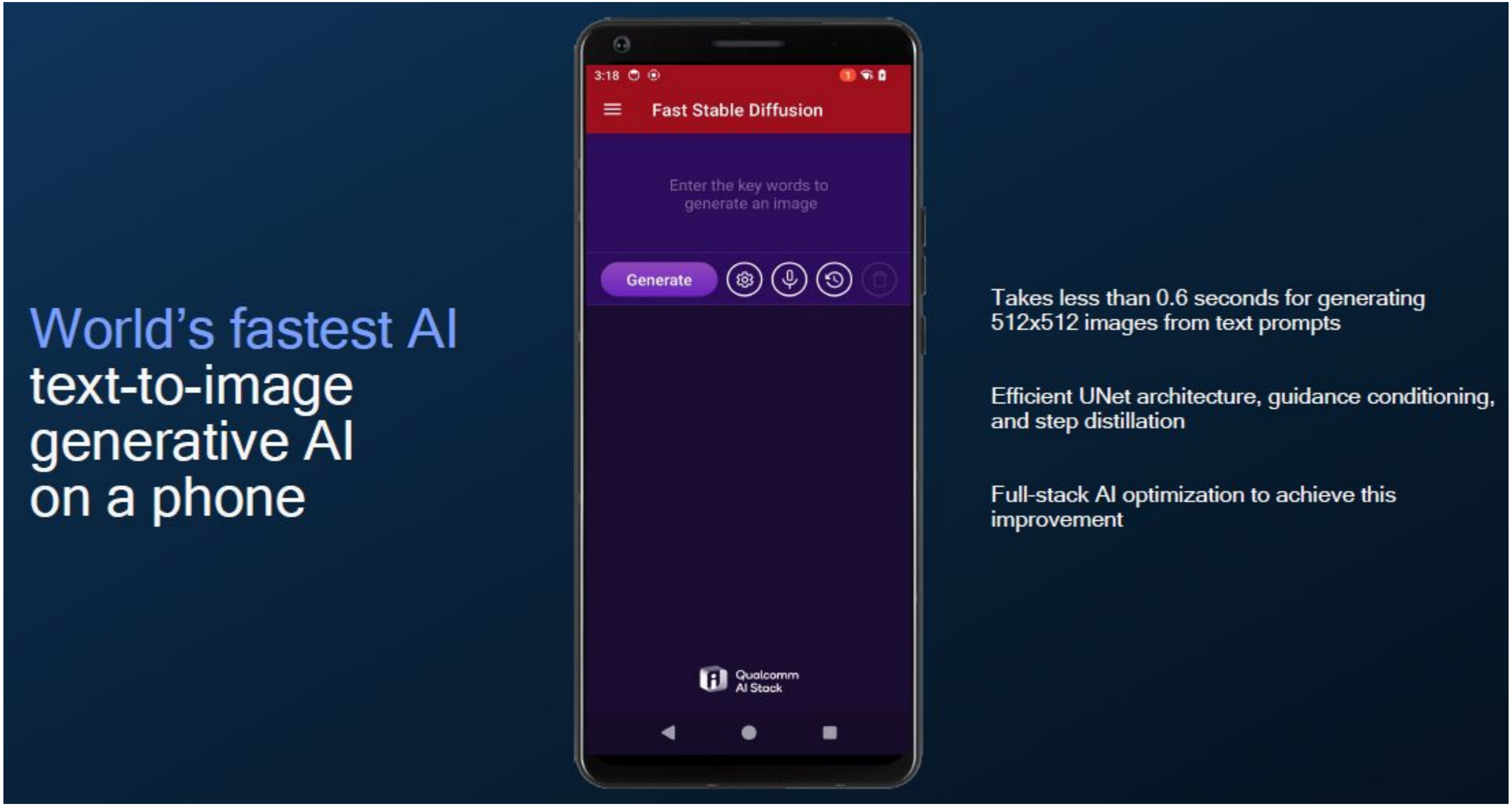
- AI models optimizations
- Quantization
- Pruning
- Knowledge distillation



https://int8.io/local-large-language-models-beginners-guide/

https://www.linkedin.com/pulse/quantization-what-you-should-understand-want-run-llms-pavan-mantha

Edge AI and Generative AI



World's fastest AI text-to-image generative AI on a phone

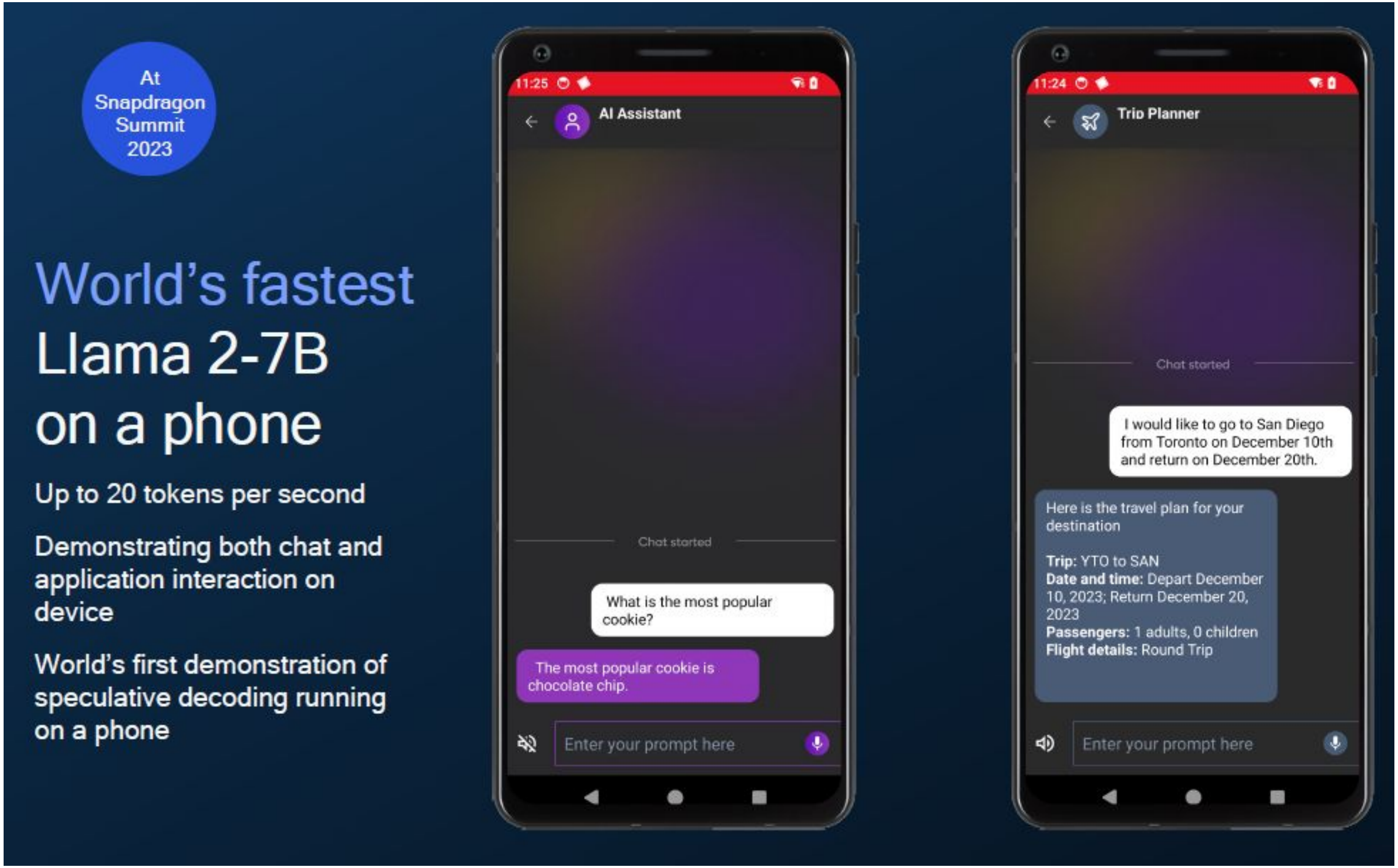
Takes less than 0.6 seconds for generating 512x512 images from text prompts

Efficient UNet architecture, guidance conditioning, and step distillation

Full-stack AI optimization to achieve this improvement

Generative AI at the edge. Joseph Soriaga Qualcomm Technologies, Inc. Nov. 8 2023

Edge AI and Generative AI



At Snapdragon Summit 2023

World's fastest Llama 2-7B on a phone

Up to 20 tokens per second

Demonstrating both chat and application interaction on device

World's first demonstration of speculative decoding running on a phone

What is the most popular cookie?

The most popular cookie is chocolate chip.

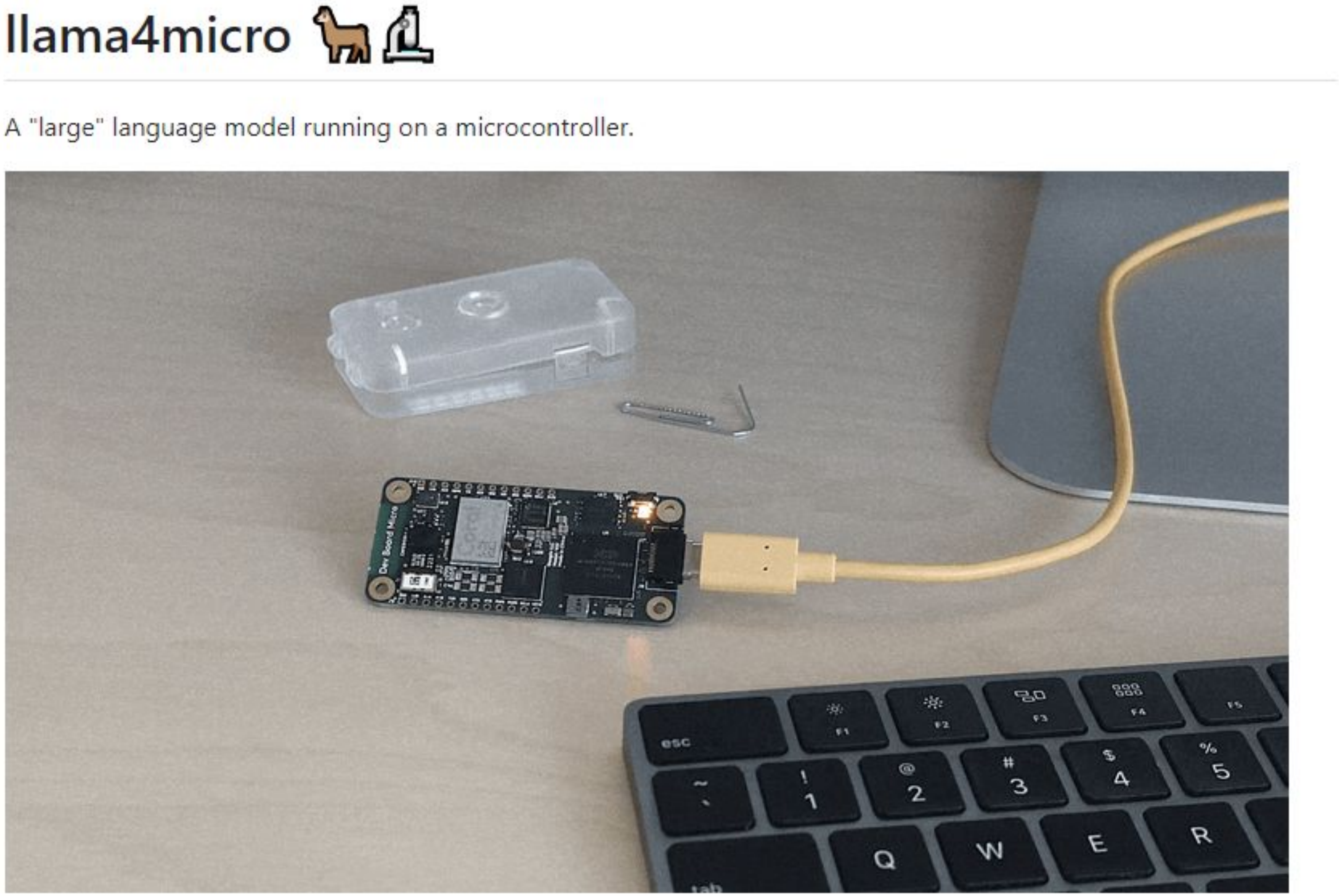
I would like to go to San Diego from Toronto on December 10th and return on December 20th.

Here is the travel plan for your destination

Trip: YTO to SAN
Date and time: Depart December 10, 2023; Return December 20, 2023
Passengers: 1 adults, 0 children
Flight details: Round Trip

Generative AI at the edge. Joseph Soriaga Qualcomm Technologies, Inc. Nov. 8 2023

Edge AI and Generative AI



llama4micro 🦙🔬

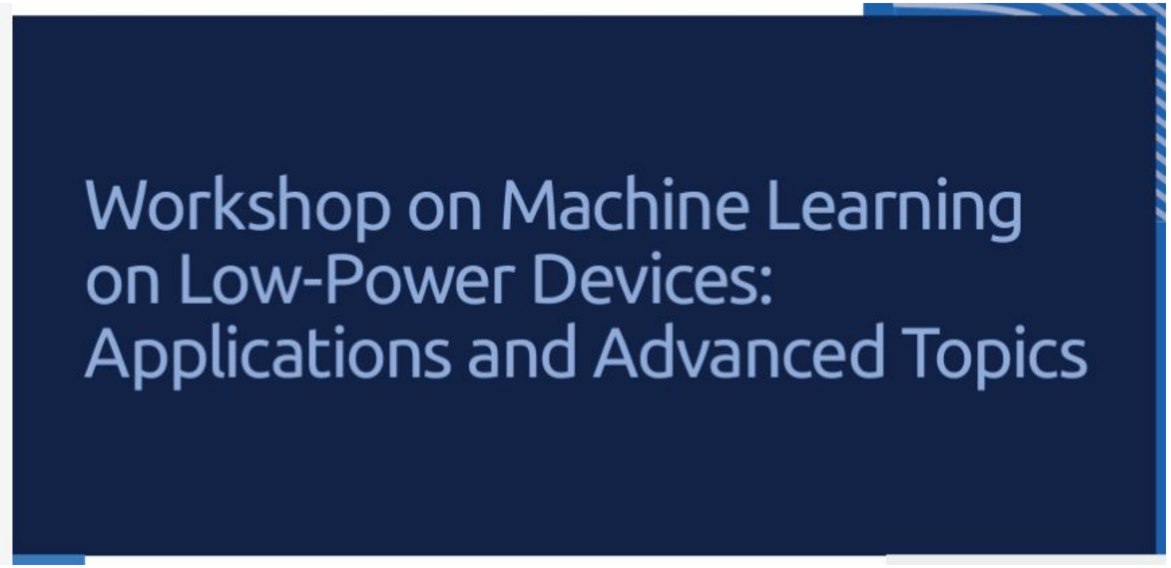
A "large" language model running on a microcontroller.

https://github.com/maxbbraun/llama4micro

Edge AI and Generative AI

https://www.dfrobot.com/blog-13412.html

Thanks!


Workshop on Machine Learning on Low-Power Devices: Applications and Advanced Topics

Prof. Jesús Alfonso López Sotelo
jalopez@uao.edu.co

UAO - Universidad Autónoma de Occidente, Cali, Colombia www.uao.edu.co