Frontiers in Foundation Models

April 24

Zhenzhong Lan, Westlake University

Website:
https://ai4s.lab.westlake.edu.cn/course/frontiers-2025/

Class 1: A bird-eye view of deep learning

Tasks

- Classification/ regression
- Simulation
- Inverse design/ inverse problem
- Control/planning

×

Neural architecture

- Multilayer perceptron
- Graph Neural Networks
- Convolutional Neural Networks
- Transformers

×

Learning paradigm

- Supervised learning
- Generative modeling
- Foundation models
- Reinforcement learning
- Evolutionary and multi-objective optimization

Application (AI & Science)

- Robotics
- Games (e.g., Go, atari)
- Autonomous Driving
- PDEs
- Life science
- Materials science

Class 2: Deep learning fundamentals

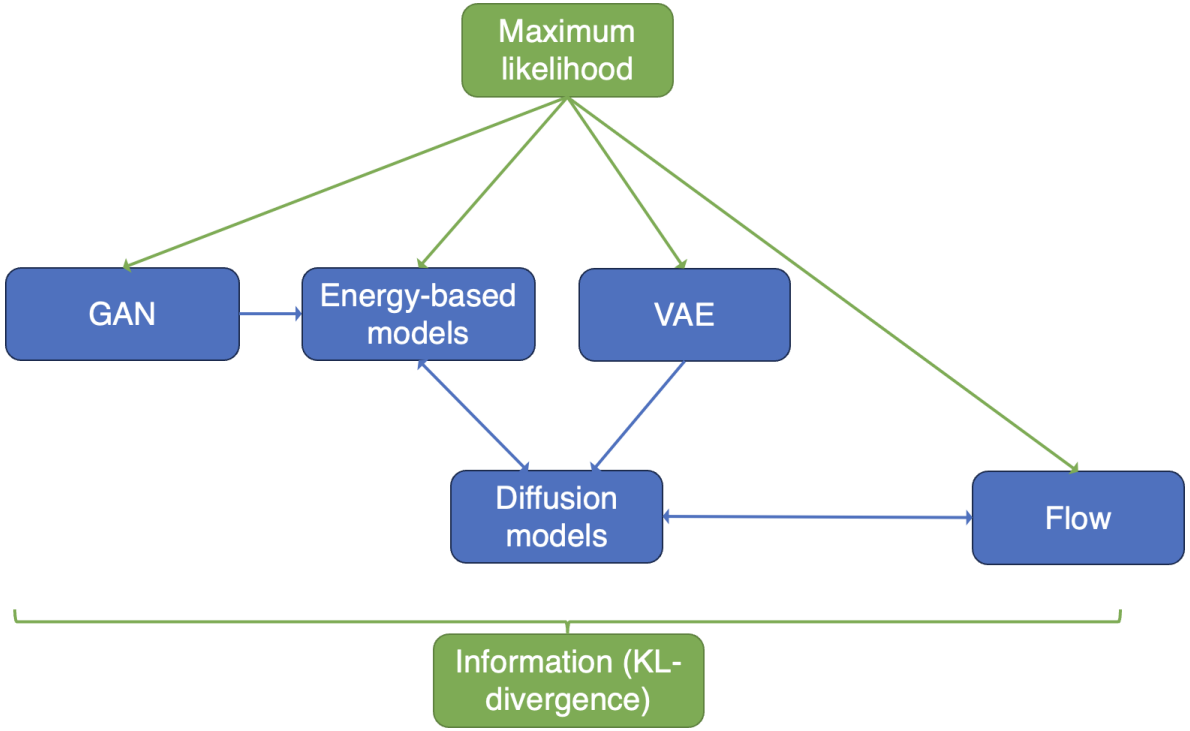
1. Principle 1: Model a hard transformation by composing simple transformations:
   - Multilayer Perceptron (MLP)
   - Backpropagation

2. Principle 2: Directly optimizing the final objective using maximum likelihood and information theory:
   - Maximum likelihood: MSE, uncertainty estimation
   - Information: cross-entropy, Information Bottleneck

3. Optimization
   - Adam: combining momentum and per-dimension magnitude
   - SAM (sharpness-aware minimization): $\max_{\epsilon \in N_{\theta}} \ell(\theta + \epsilon)$ finds flat and robust minima
   - Federative learning: improves the data privacy by only sharing client models

Class 3: Generative Models

- Generative models
  - VAE
  - GAN
  - Energy-based models
  - Diffusion models
  - Flows
- Application of diffusion models
  - Image, video, and shape generation
  - Simulation
  - Inverse design/inverse problem
  - Control/planning



Class 4: Foundation Models

Shift in learning paradigm through time:

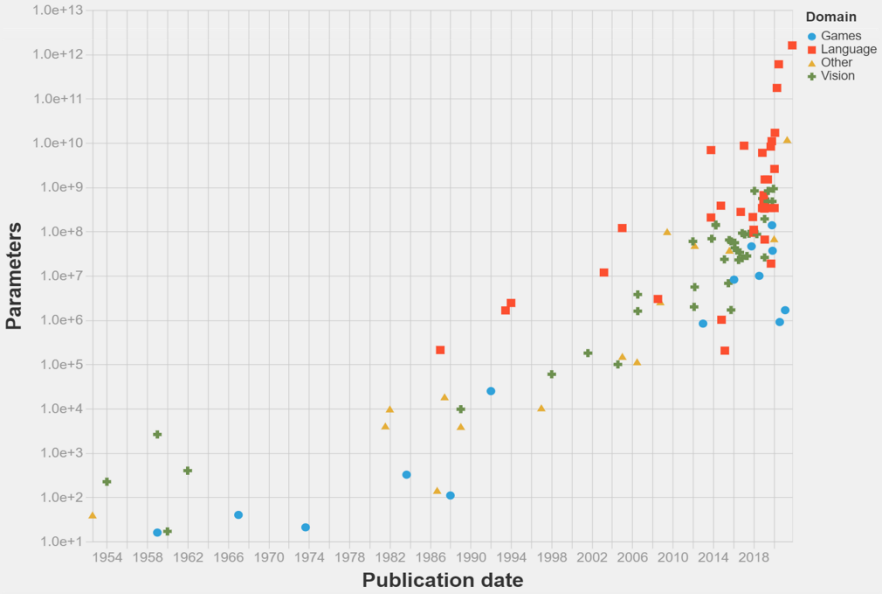
Shallow-Network (e.g. SVM) → Supervised (e.g. Alexnet)→ Partially supervised (e.g. word2vec) → Self-supervised + Finetunning (e.g. BERT) → Self-supervised + Prompting with examples (e.g. GPT3.5) → Self-supervised + Prompting (e.g. InstructGPT or ChatGPT)

Foundation model ≈ Pretrained networks

Foundation for foundation models

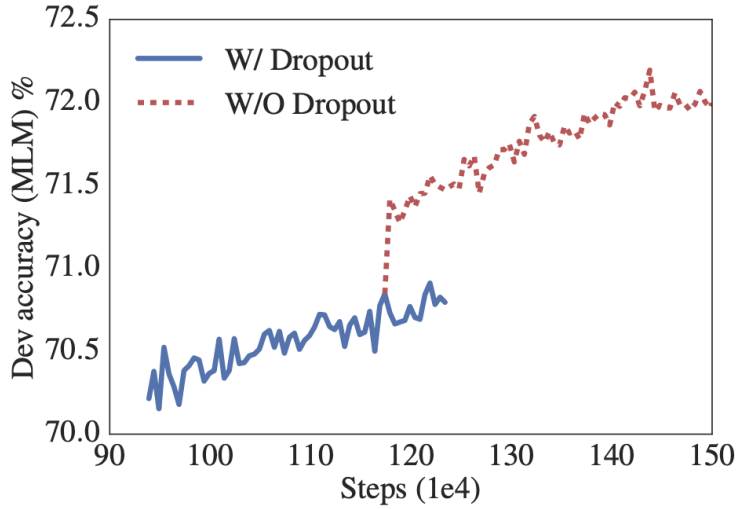
Principle 3 (the scaling law): AI methods that leverage computation are ultimately the most effective way of improvements (from "The bitter lesson" by Rich Sutton)

Principle 4 (the data law): Data is the ultimate way of regularization



Parameter count of ML systems through time

Image from: lesswrong.com



(b) Removing dropout

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. . Albert: A lite bert for self-supervised learning of language representations, ICLR 2020

Foundation for foundation models

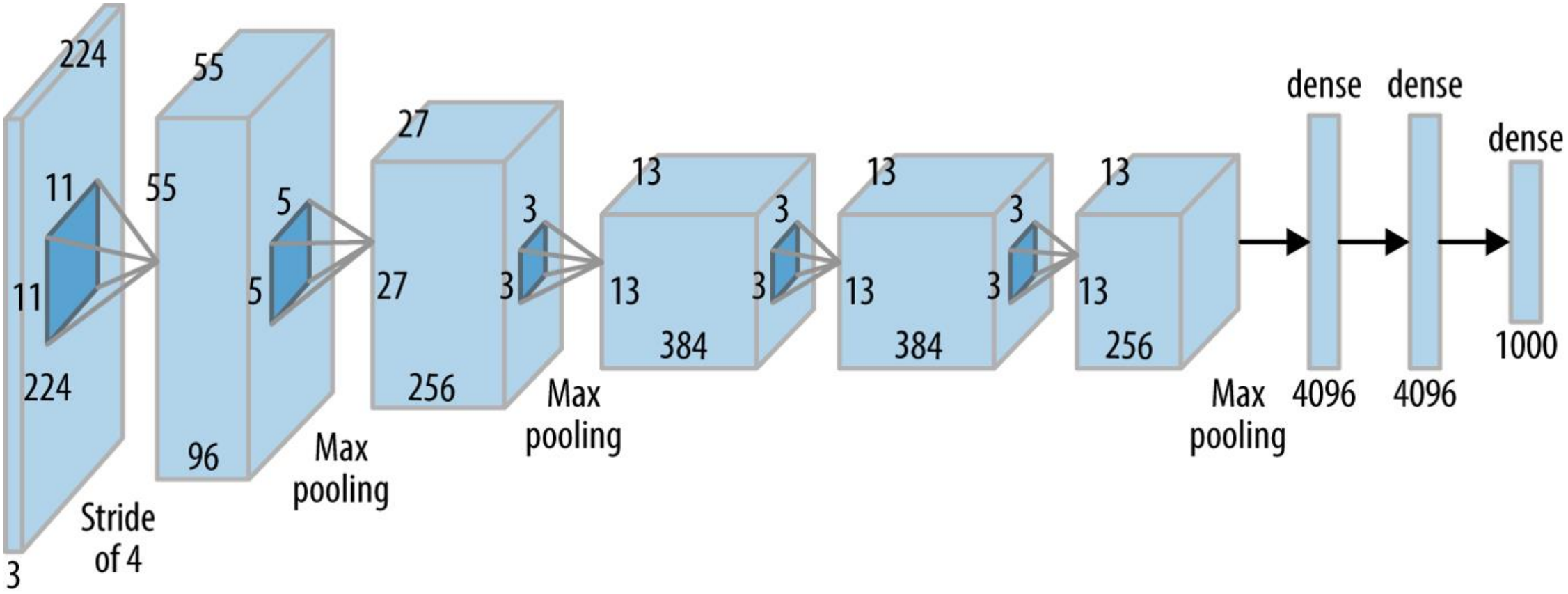
Principle 3 (the scaling law): AI methods that leverage computation are ultimately the most effective way of improvements (from "The bitter lesson" by Rich Sutton)

What is the most effective network architecture to leverage computation?
(Comparison between CNNs and Transformer)

Principle 4 (the data law): Data is the ultimate way of regularization

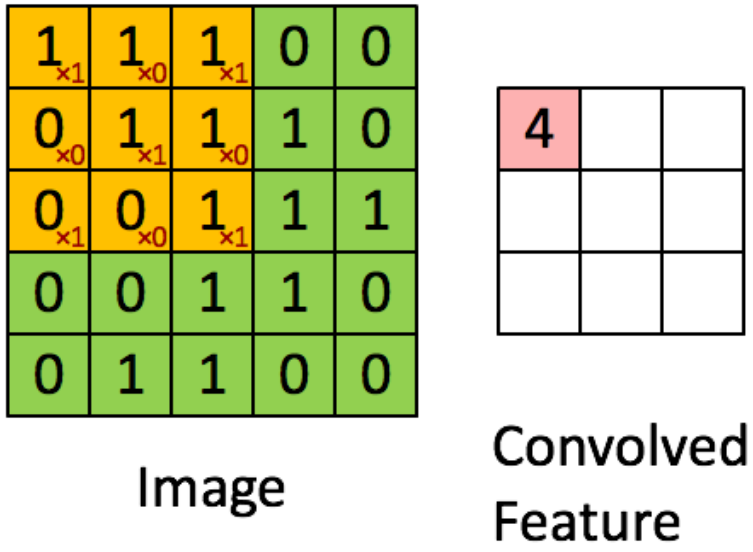
What is the most effective way of (pre-)training the network?

Convolutional Neural Networks for Image Classification

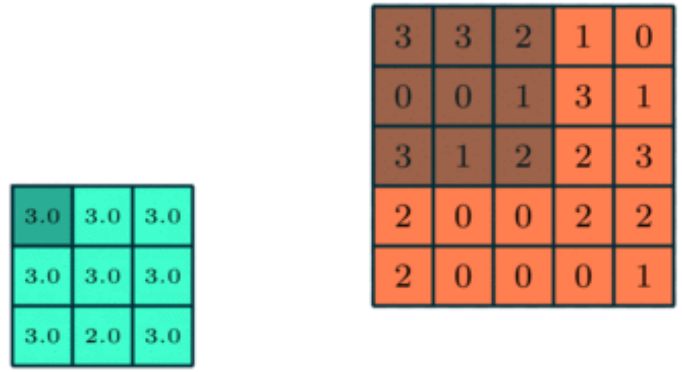


Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *NIPS 2012*

Convolutional and Pooling operation for CNNs



Convolution Operation

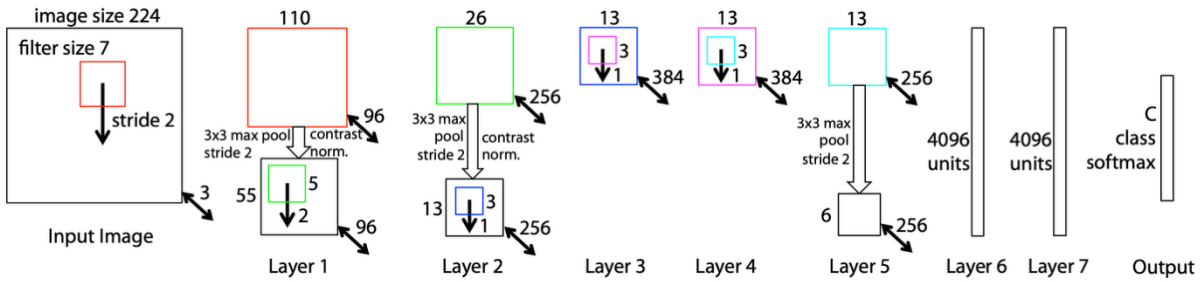


Pooling Operation

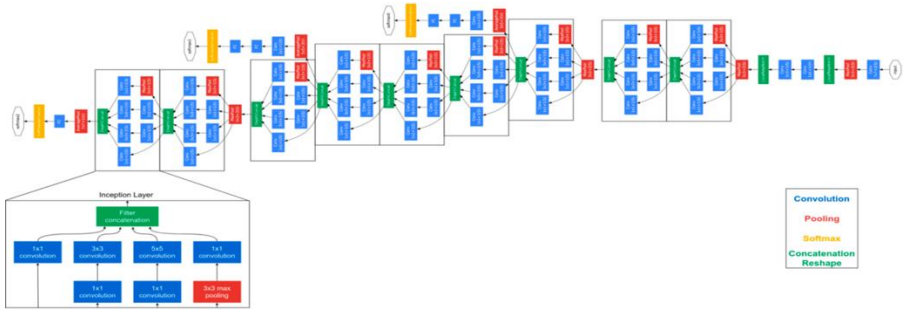
Both convolution and pooling operations are local operations and compress represetations from high dimensions to low dimensions

Gif credit Sumit Saha

Improvements on CNNs: deeper networks

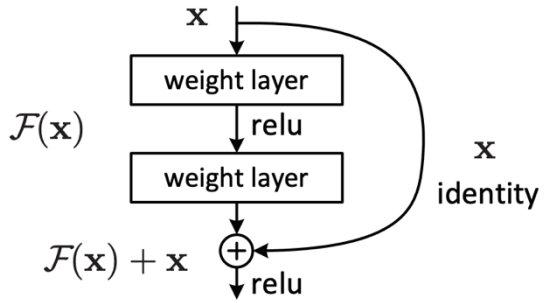


Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." *ECCV* 2014.



Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. CVPR 2015

Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." 2014.

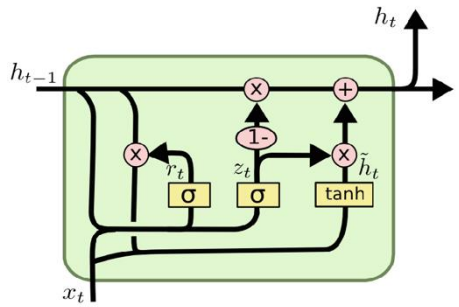


> 1000 layers

He, Kaiming, et al. "Deep residual learning for image recognition." *CVPR* 2016.

Sequence modeling using CNNs

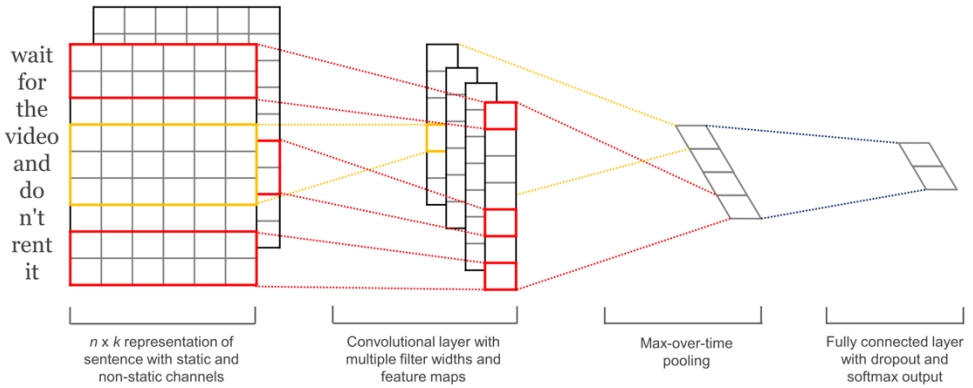
Gated Recurrent Networks



$$\mu_i = \sigma(U_\mu x_i + W_\mu h_{i\_1})$$
$$\widehat{h}_i = \tanh(U x_i + W \mu_i \odot h_{i\_1})$$
$$\lambda_i = \sigma(U_\lambda x_i + W_\lambda h_{i\_1})$$
$$h_i = (1 - \lambda_i) \odot h_{i\_1} + \lambda_i \odot \widehat{h}_i. \textit{ (shortcut)}$$

Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." EMNLP 2014.

CNNs



https://dennybritz.com/posts/wildml/understanding-convolutional-neural-networks-for-nlp/

Sequence modeling using parallel processing

Transformer: a sequence modeling architecture with parallel encoding and global view

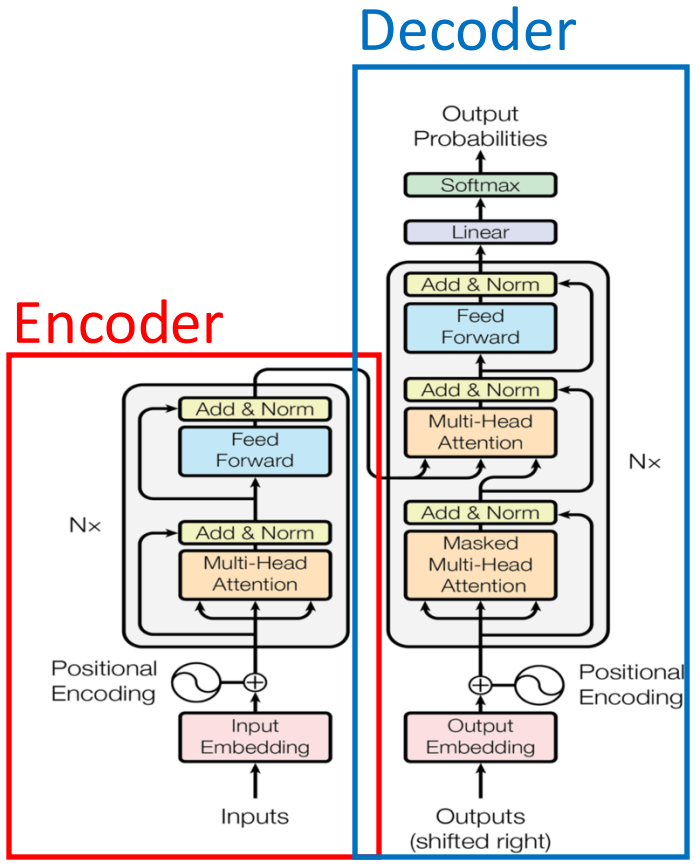
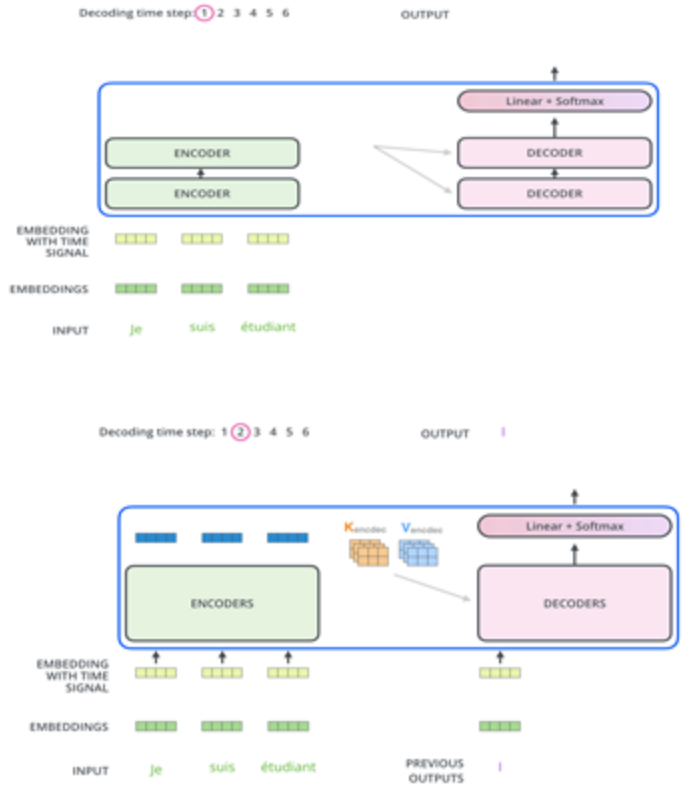


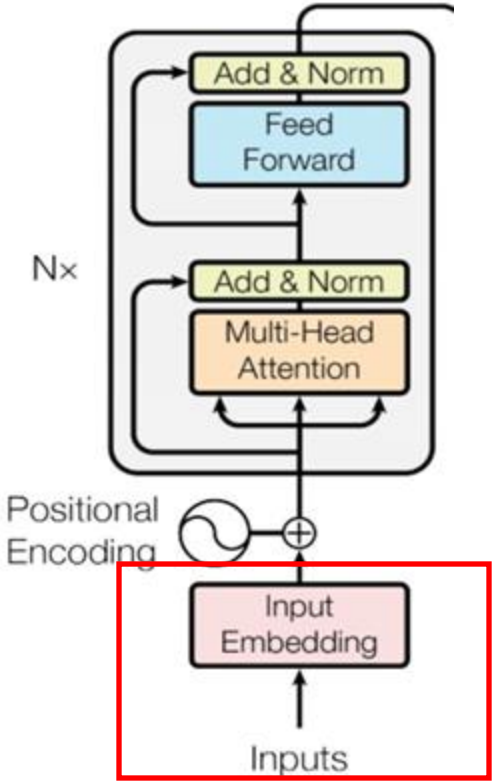
Figure 1: The Transformer - model architecture.

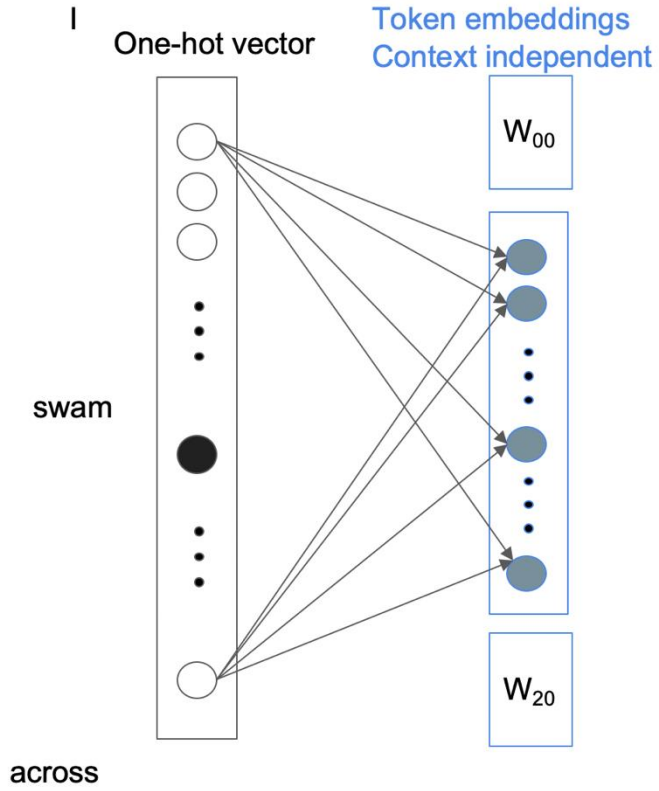


gif credit: Jay Alammar

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In *Advances in neural information processing systems*, pp. 5998-6008. 2017.

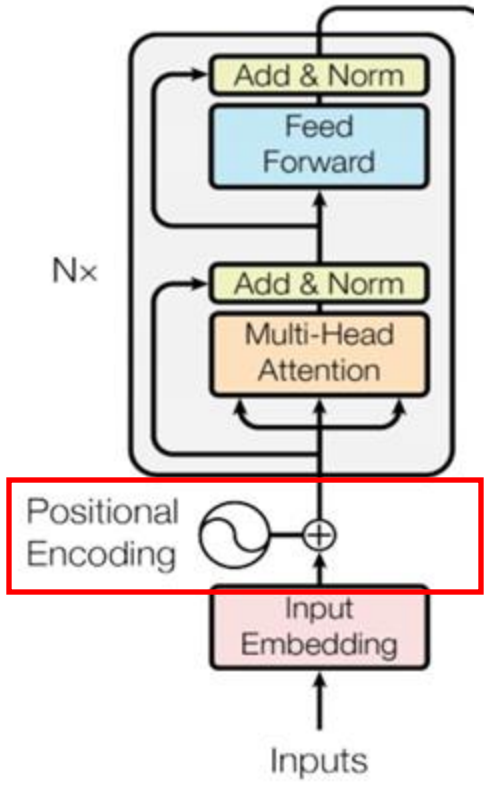
Token embedding map words into representations

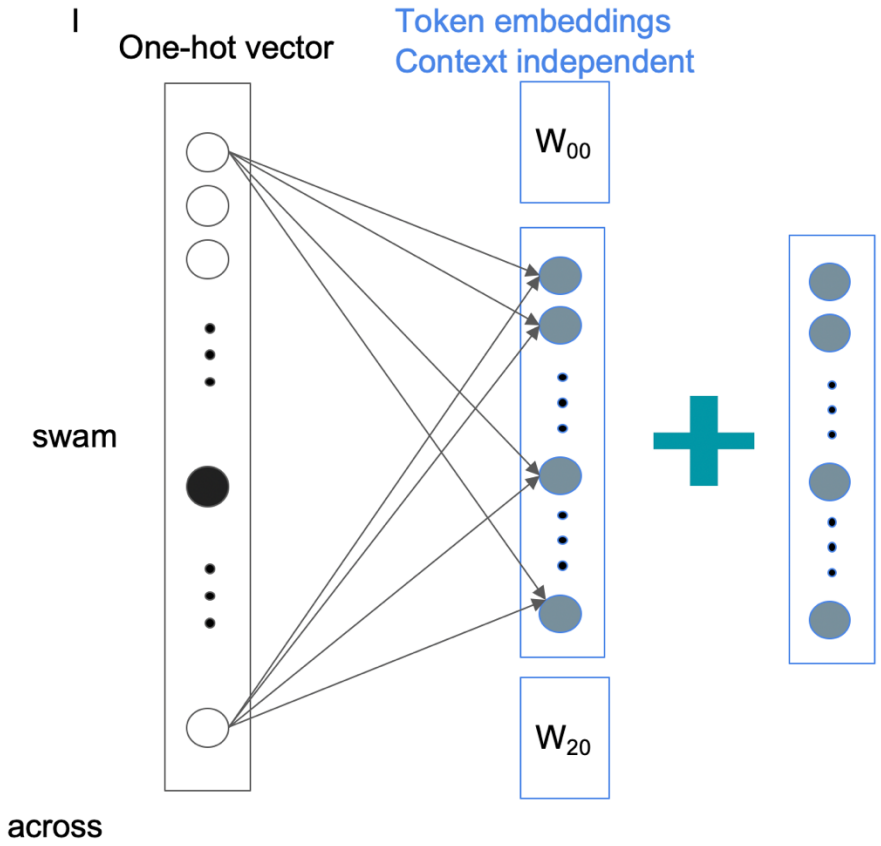




I	swam	across	the	river	to	get	to	the	other	bank

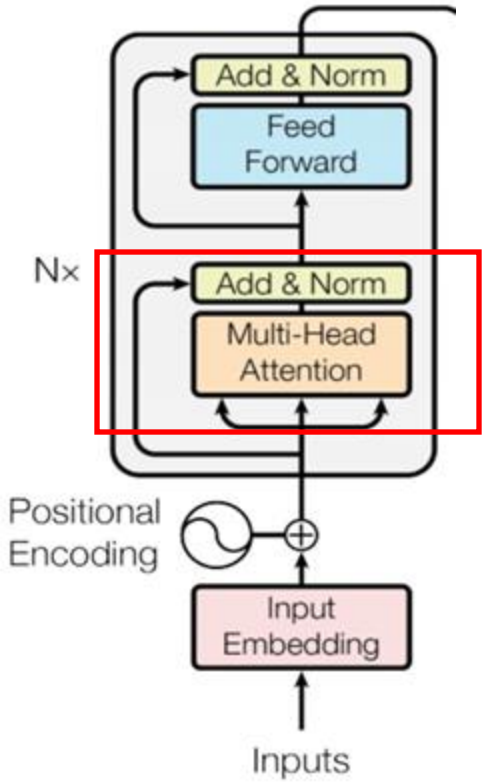
Positional embedding differentiates words in different positions

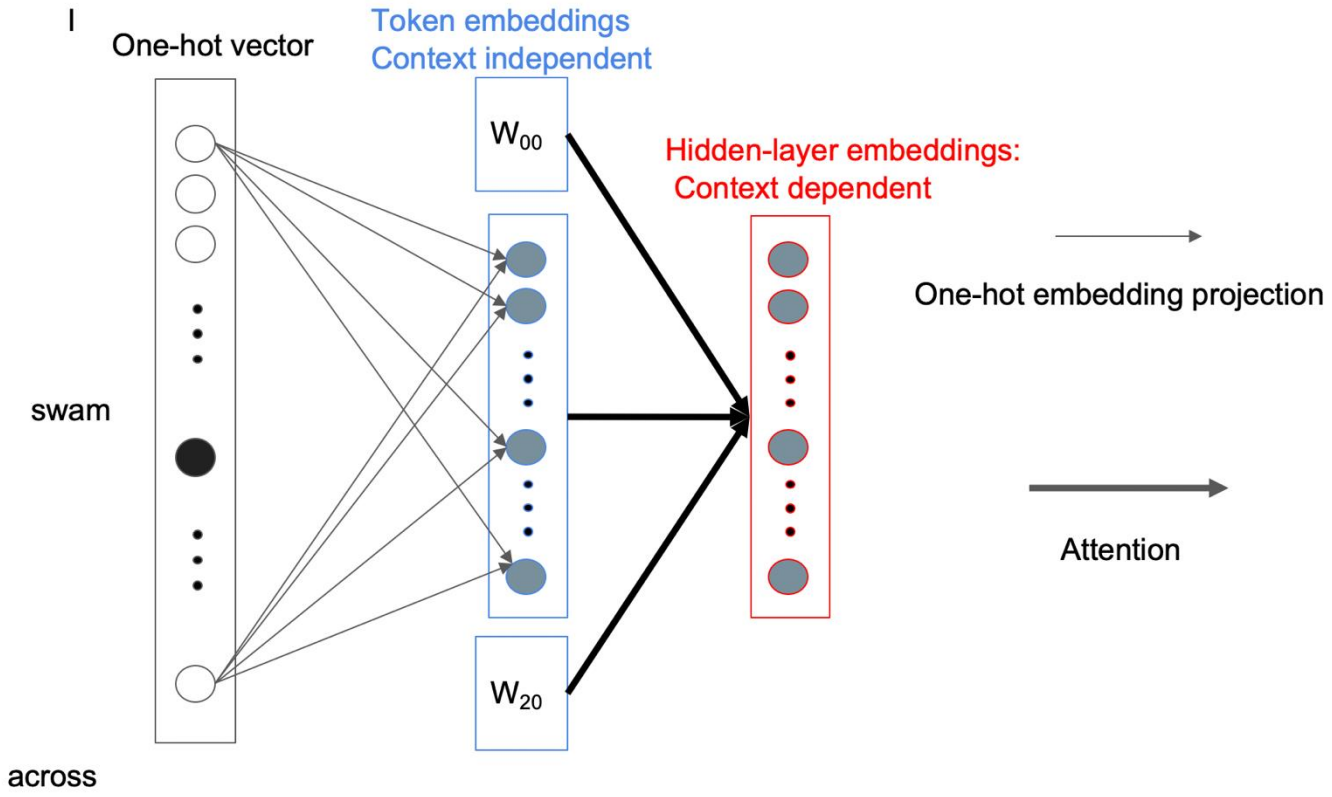




I	swam	across	the	river	to	get	to	the	other	bank

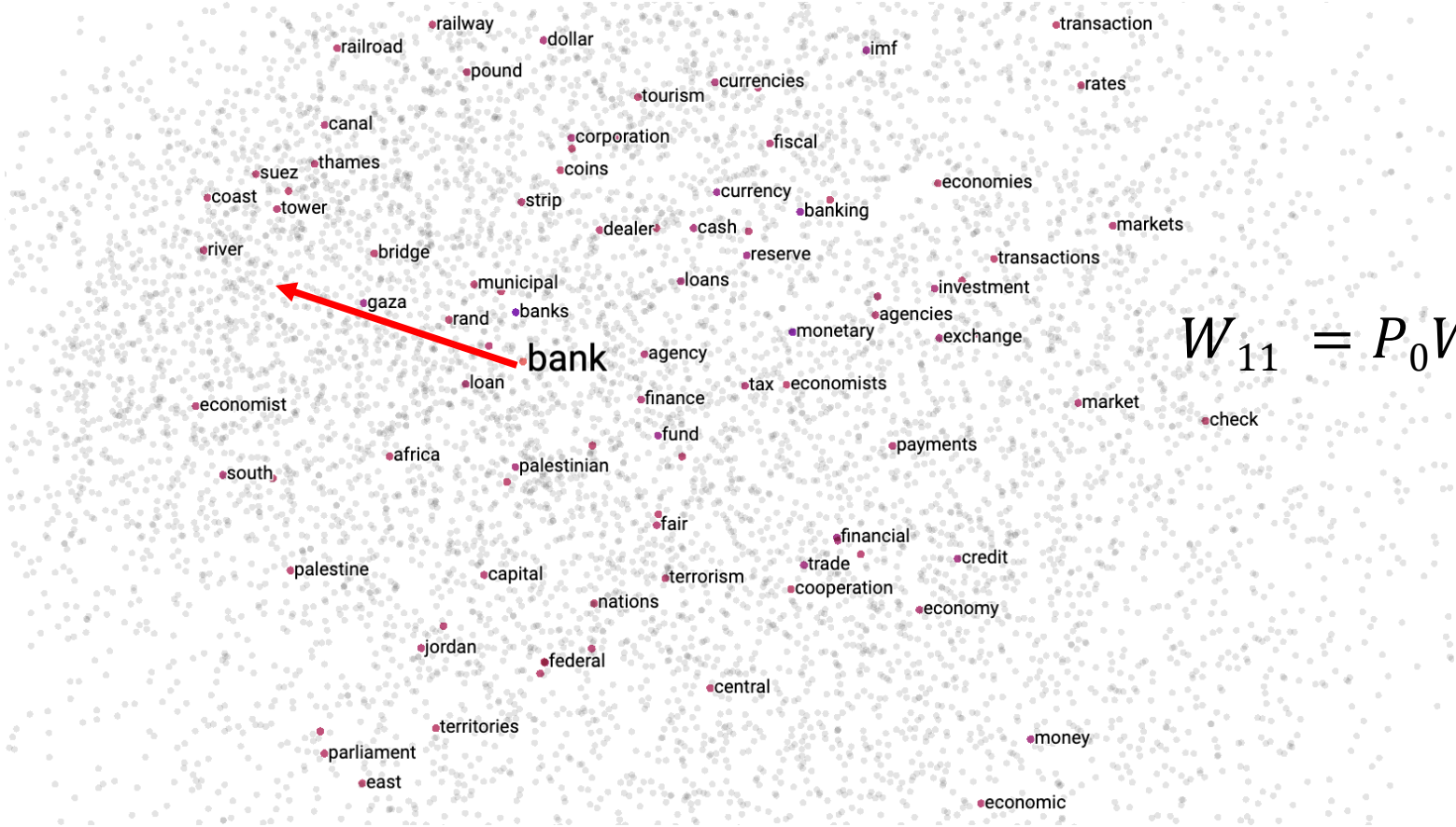
Multi-head attention gives words contextual meaning





I	swam	across	the	river	to	get	to	the	other	bank

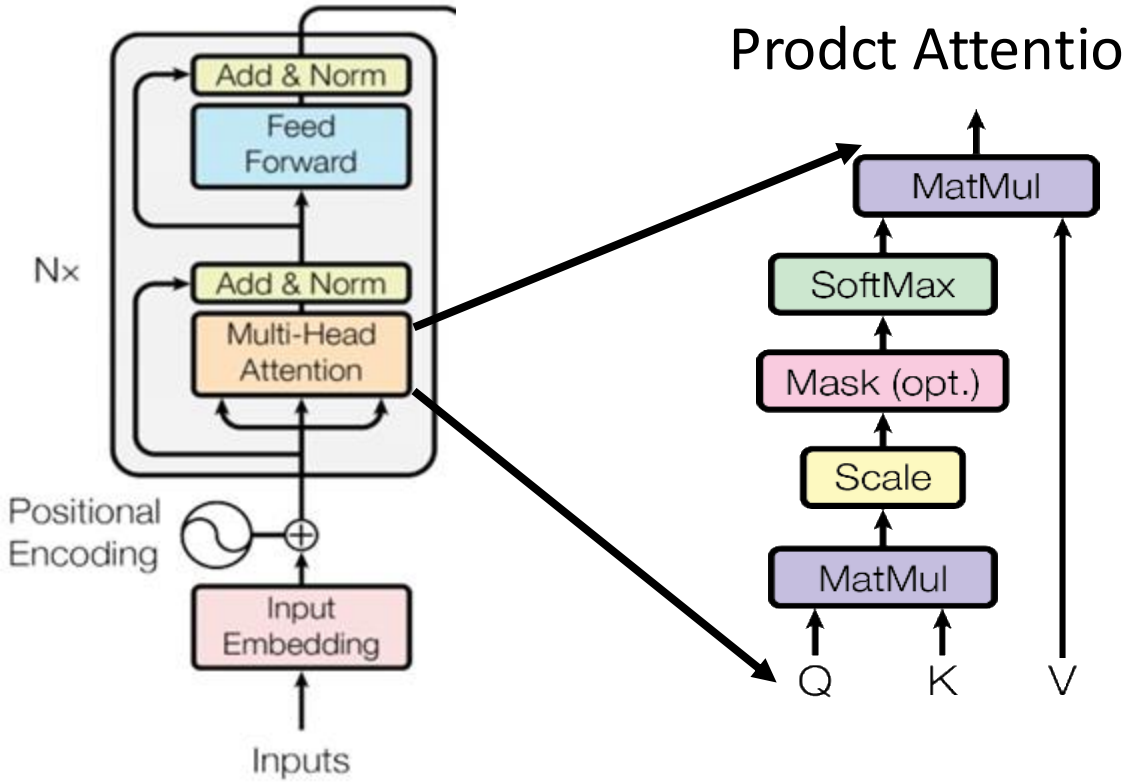
Multi-head attention makes similar words closer through weighted average operation



$$W_{11} = P_0 W_{00} + P_1 W_{10} + P_2 W_{20} + \cdots$$

https://projector.tensorflow.org/

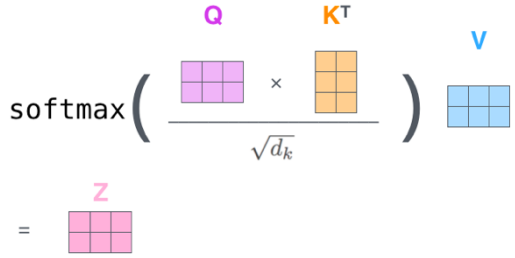
Multi-head attention makes similar words closer



Similar words get higher weights

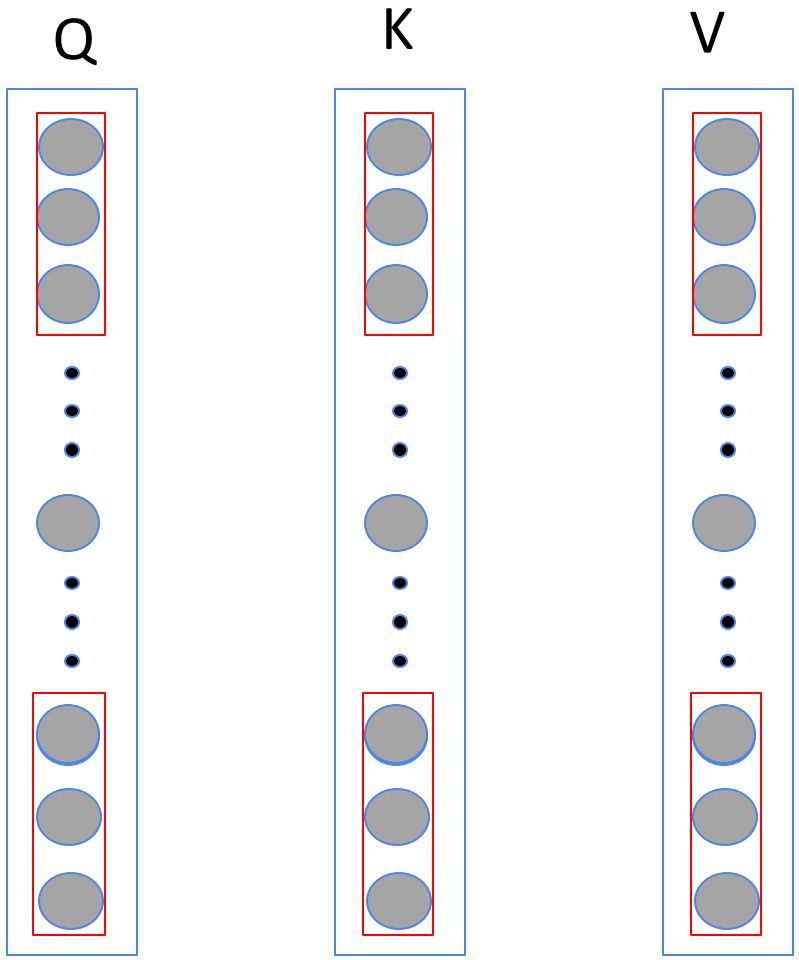
$$W_{11} = P_0 W_{00} + P_1 W_{10} + P_2 W_{20} + \cdots$$

$$P_0 : W_{10} * W_{00}$$
$$P_1 : W_{10} * W_{10}$$
$$P_2 : W_{10} * W_{20}$$



$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Multi-head attention: why we needs multi-heads

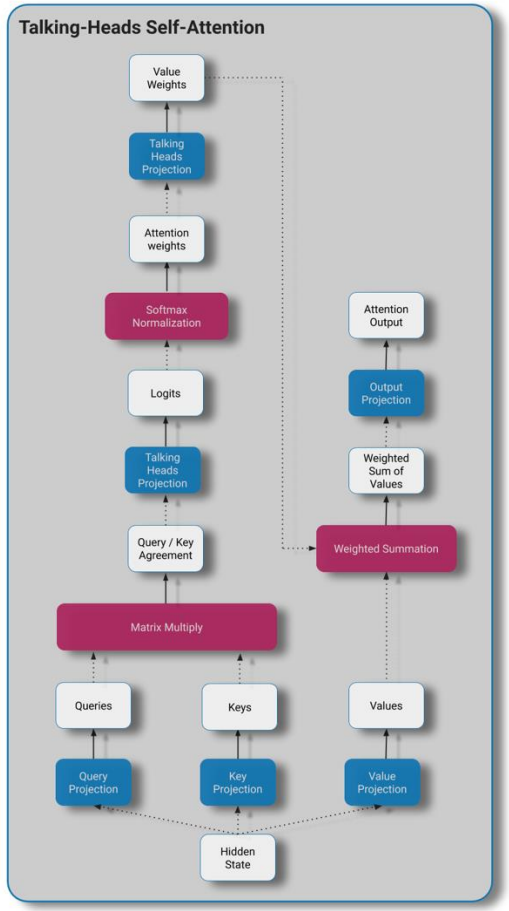


$$L \times D \rightarrow L \times n \times \mathrm{d}$$

Talking-head attention: does more heads leads better results

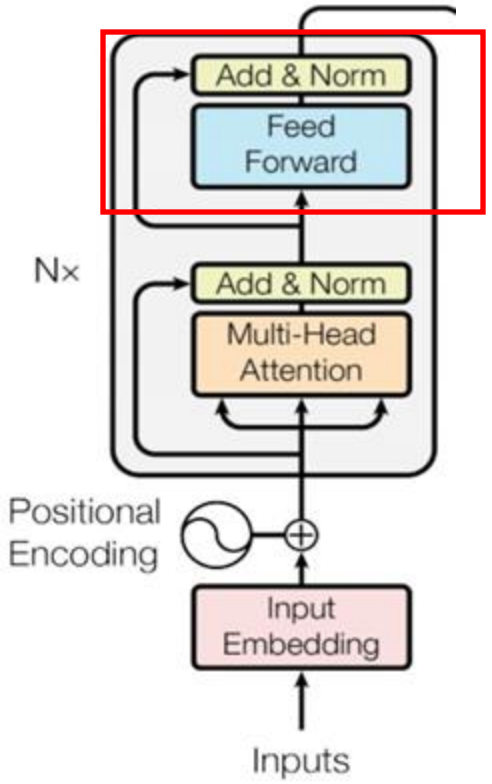
	N	d_{model}	d_{ff}	h	d_k	d_v	P_{drop}	ϵ_{ls}	train steps	PPL (dev)	BLEU (dev)	params $\times 10^6$
base	6	512	2048	8	64	64	0.1	0.1	100K	4.92	25.8	65
(A)				1	512	512				5.29	24.9	
				4	128	128				5.00	25.5	
				16	32	32				4.91	25.8	
				32	16	16				5.01	25.4	
(B)					16					5.16	25.1	58
					32					5.01	25.4	60
(C)	2									6.11	23.7	36
	4									5.19	25.3	50
	8									4.88	25.5	80
		256			32	32				5.75	24.5	28
		1024			128	128				4.66	26.0	168
			1024							5.12	25.4	53
			4096							4.75	26.2	90
(D)							0.0			5.77	24.6	
							0.2			4.95	25.5	
								0.0		4.67	25.3	
								0.2		5.47	25.7	
(E)	positional embedding instead of sinusoids									4.92	25.7	
big	6	1024	4096	16			0.3		300K	**4.33**	**26.4**	213

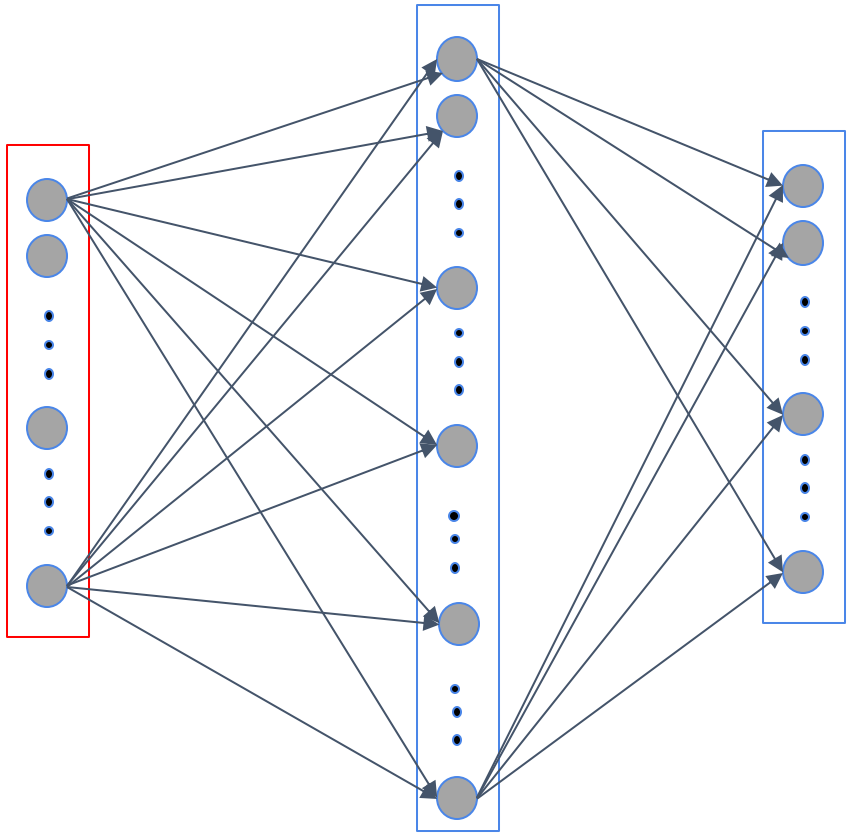
Ashish Vaswani, Noam Shazeer, etc, attention is all you need, 2017

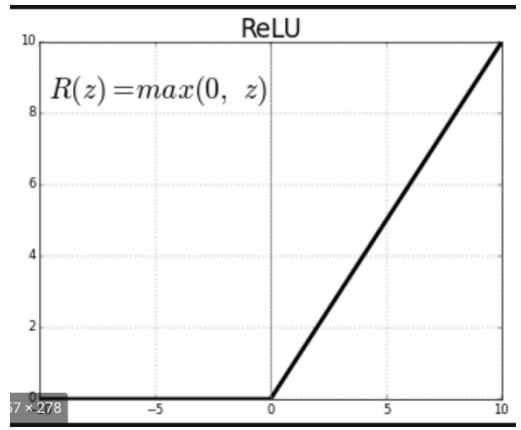


Noam Shazeer, Zhenzhong Lan, Youlong Cheng, Nan Ding, Le Hou, talking-head attention, 2020

Feed forward network: within token transformation







Sandler, Mark, et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." *CVPR* 2018.

I	swam	across	the	river	to	get	to	the	other	bank

Decoder: sequential decoding

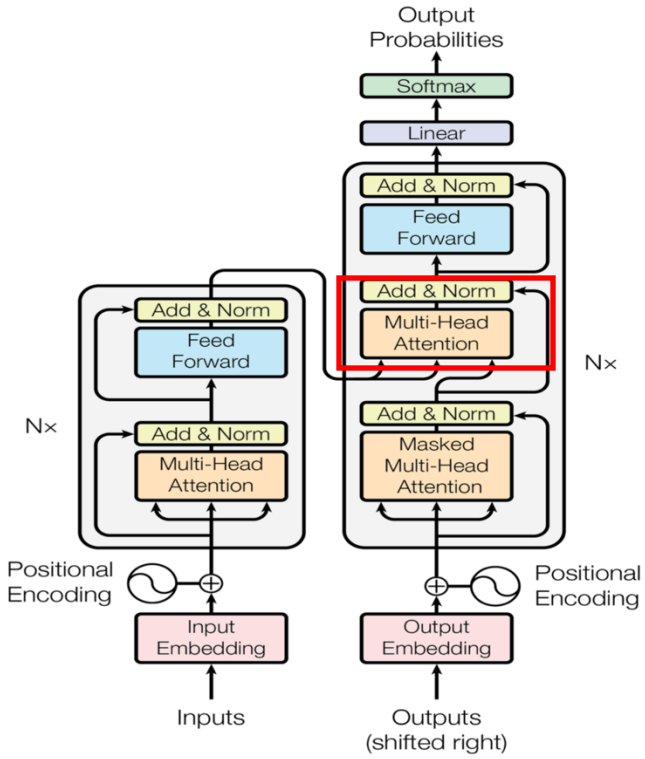
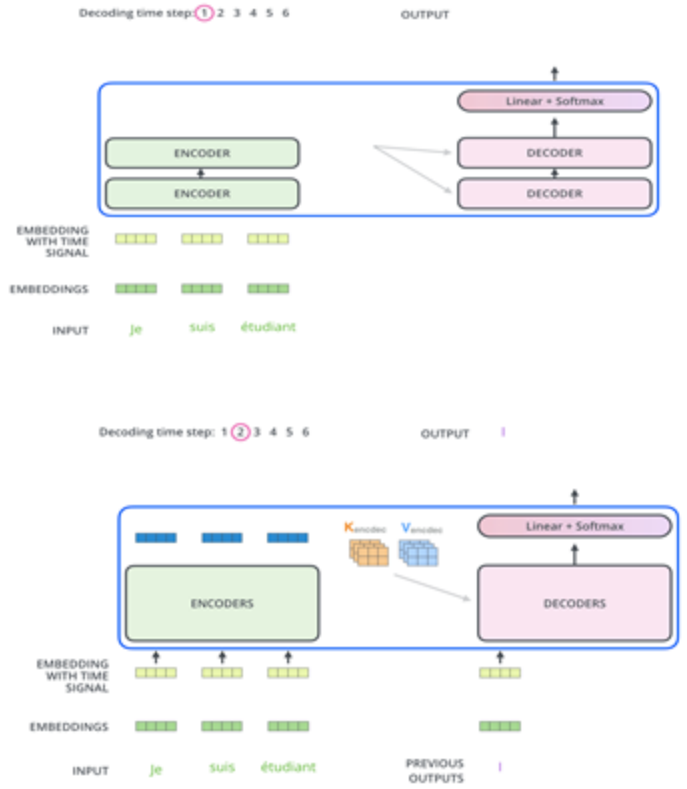


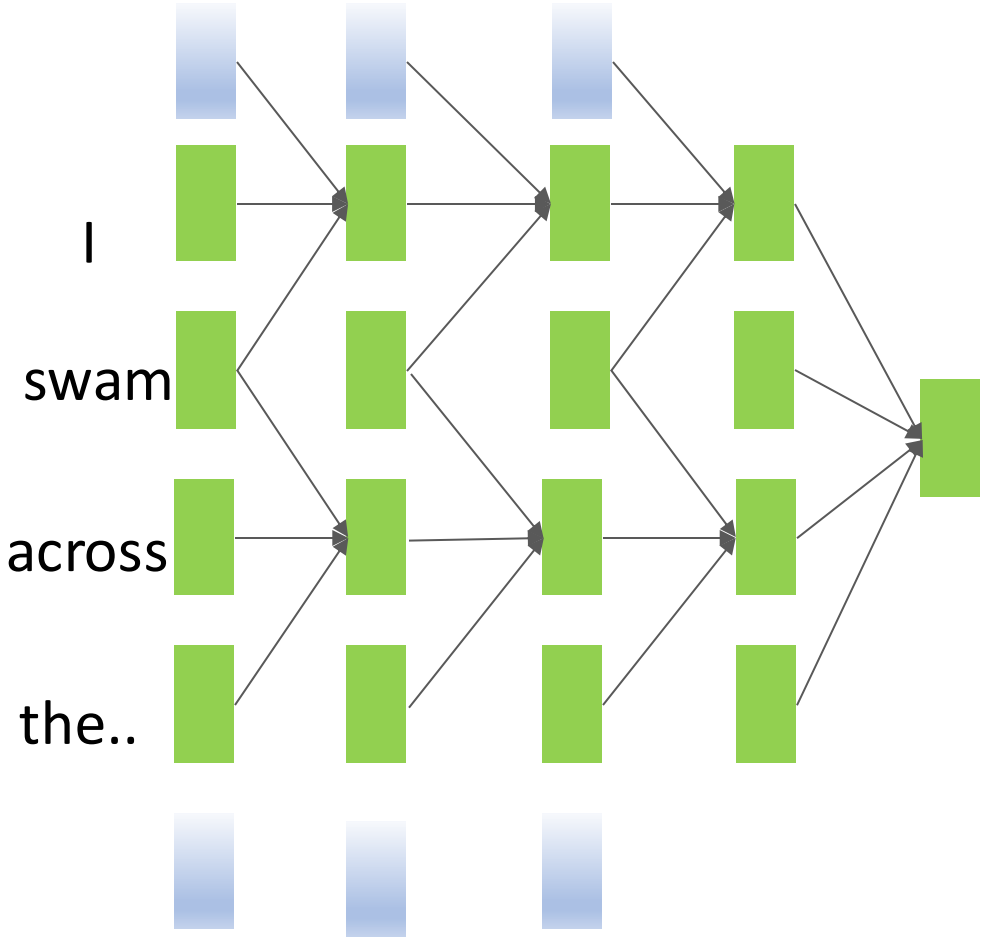
Figure 1: The Transformer - model architecture.



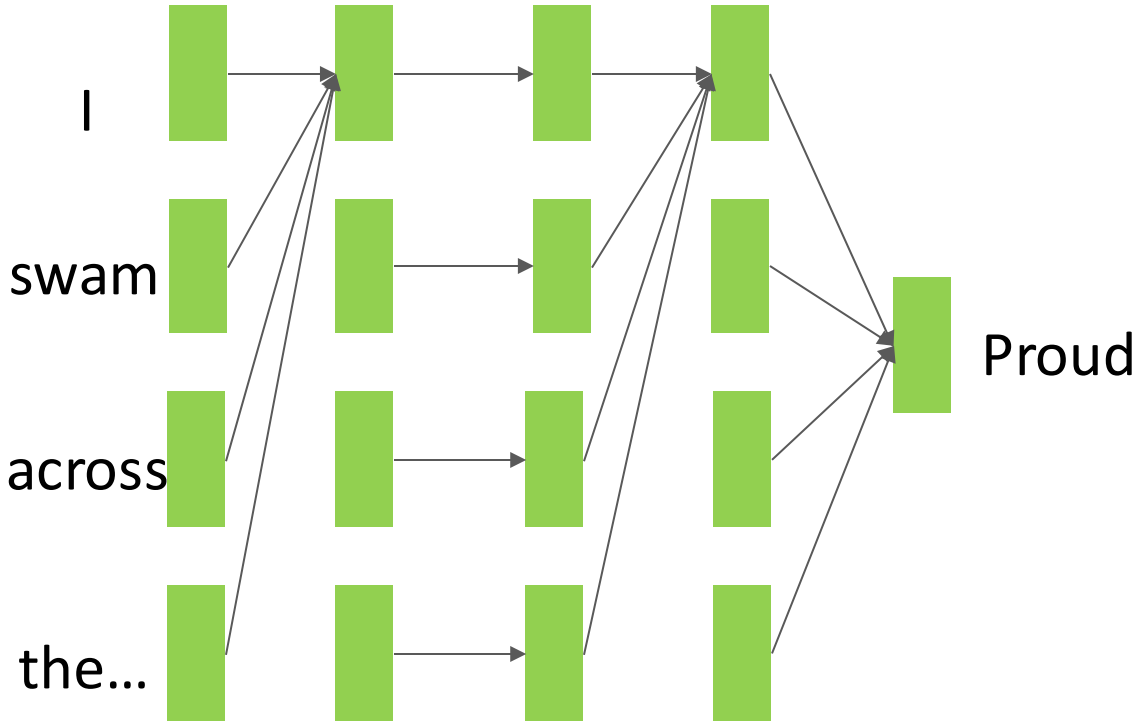
gif credit: Jay Alammar

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In *Advances in neural information processing systems*, pp. 5998-6008. 2017.

CNNs exploits local constraints while Transformers has a more global view

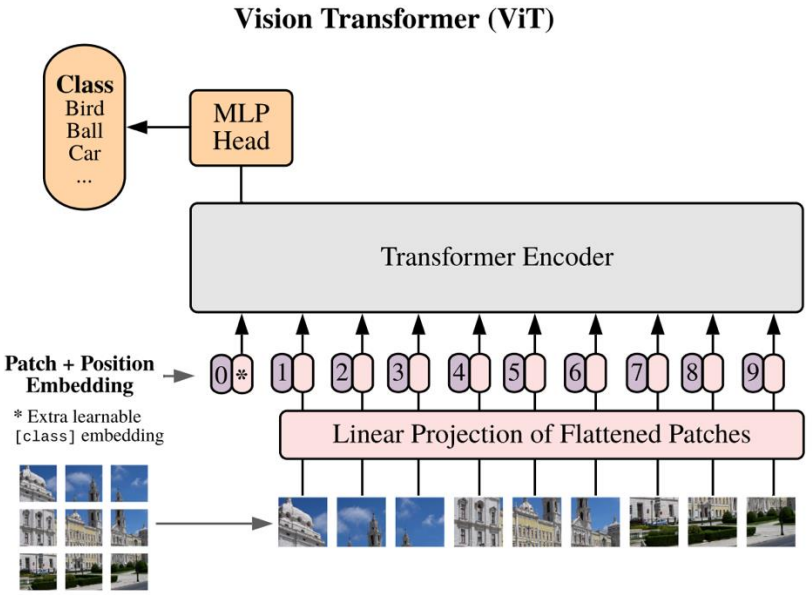


CNNs

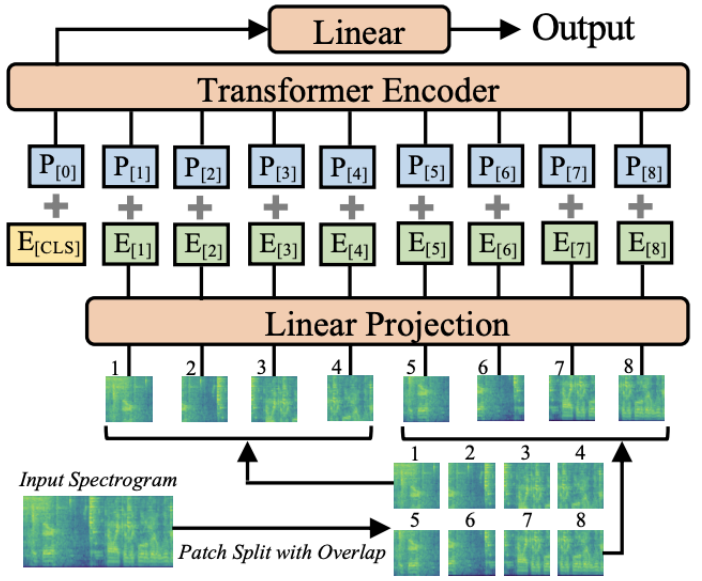


Transformers

Transformer is everywhere nowadays

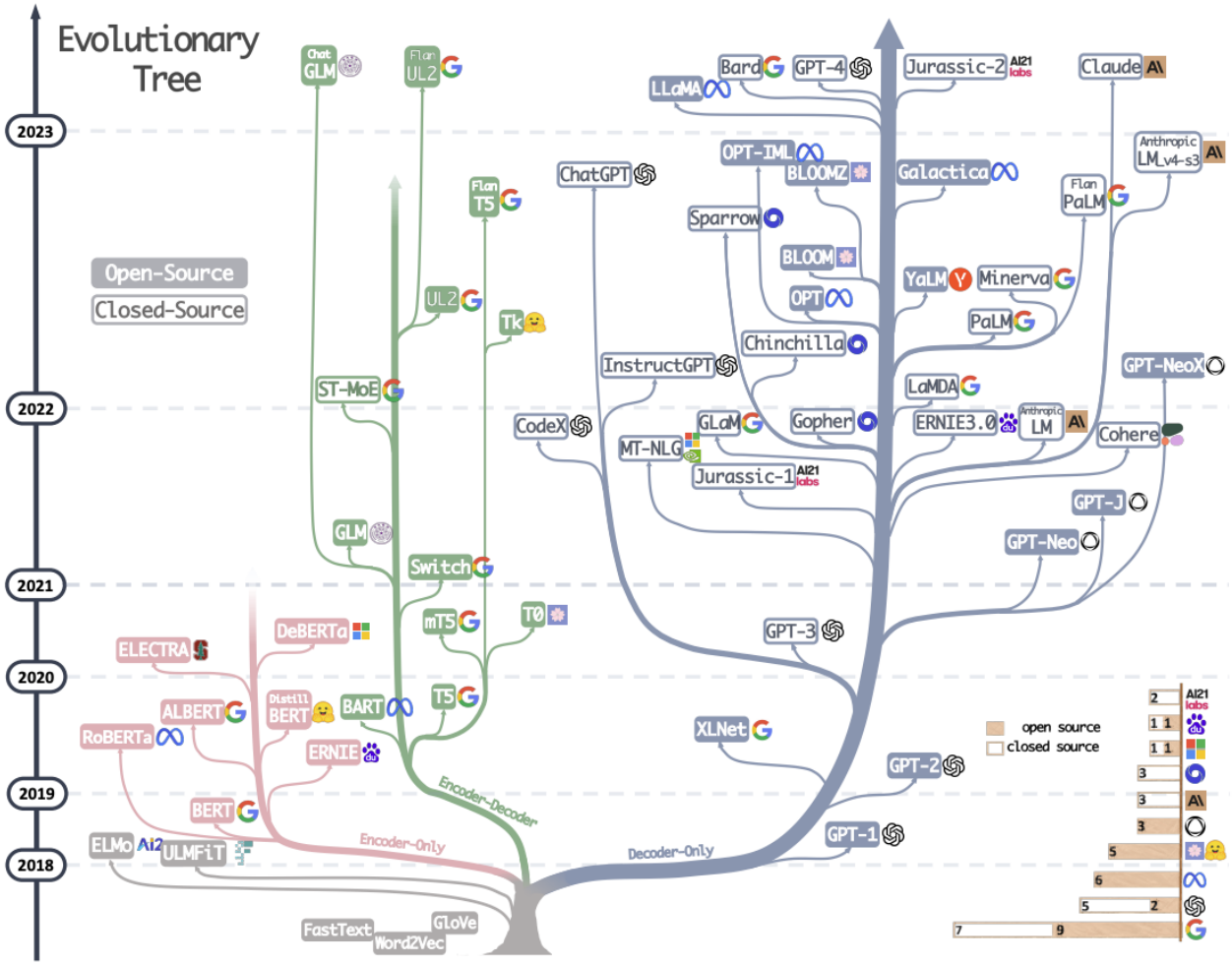


Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

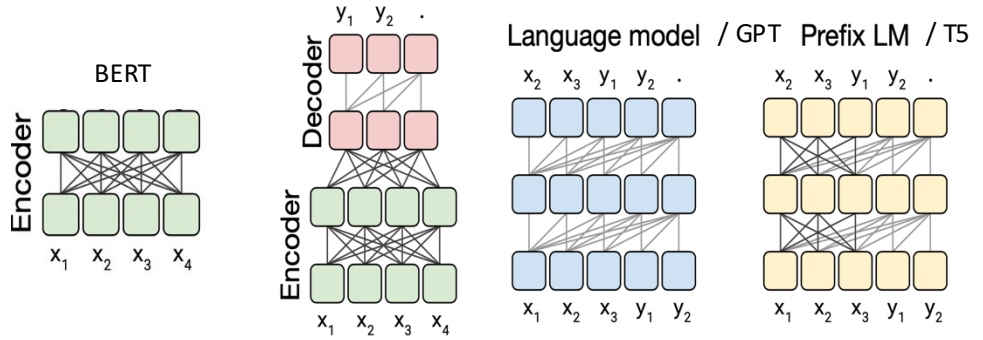


Gong, Yuan, Yu-An Chung, and James Glass. "Ast: Audio spectrogram transformer." *arXiv preprint arXiv:2104.01778* (2021).

Decoder-only Transformer is dominant



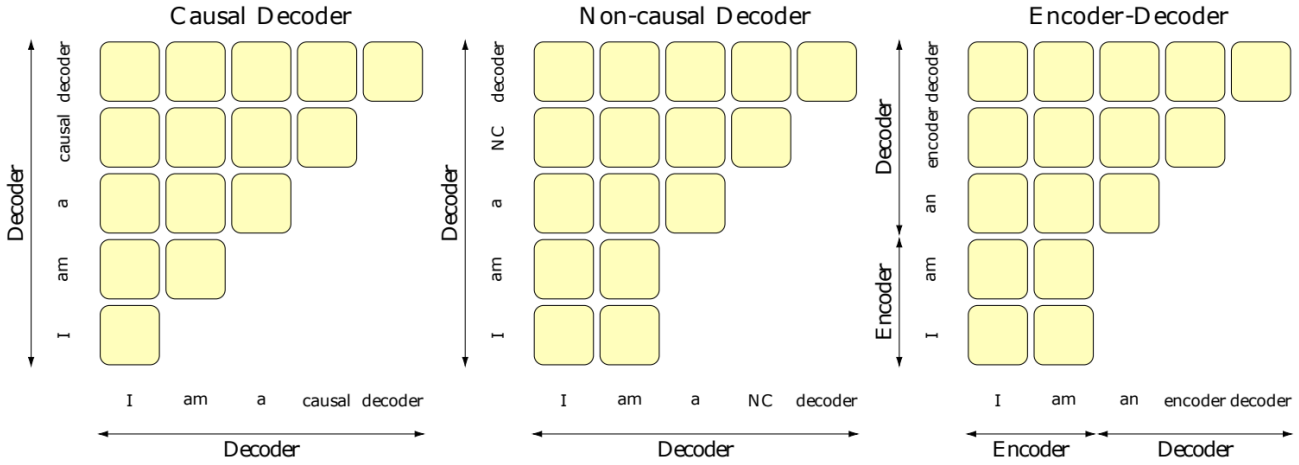
PEGASUS / T5



Architecture	Objective	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	Denoising	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Enc-dec, shared	Denoising	P	M	82.81	18.78	**80.63**	**70.73**	26.72	39.03	**27.46**
Enc-dec, 6 layers	Denoising	P	$M/2$	80.88	18.97	77.59	68.42	26.38	38.40	26.95
Language model	Denoising	P	M	74.70	17.93	61.14	55.02	25.09	35.28	25.86
Prefix LM	Denoising	P	M	81.82	18.61	78.94	68.11	26.43	37.98	27.39
Encoder-decoder	LM	$2P$	M	79.56	18.59	76.02	64.29	26.27	39.17	26.86
Enc-dec, shared	LM	P	M	79.60	18.13	76.35	63.50	26.62	39.17	27.05
Enc-dec, 6 layers	LM	P	$M/2$	78.67	18.26	75.32	64.06	26.13	38.42	26.89
Language model	LM	P	M	73.78	17.54	53.81	56.51	25.23	34.31	25.38
Prefix LM	LM	P	M	79.68	17.84	76.87	64.86	26.28	37.51	26.76

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer." *arXiv preprint arXiv:1910.10683* (2019).

Decoder-only Transformer is dominant



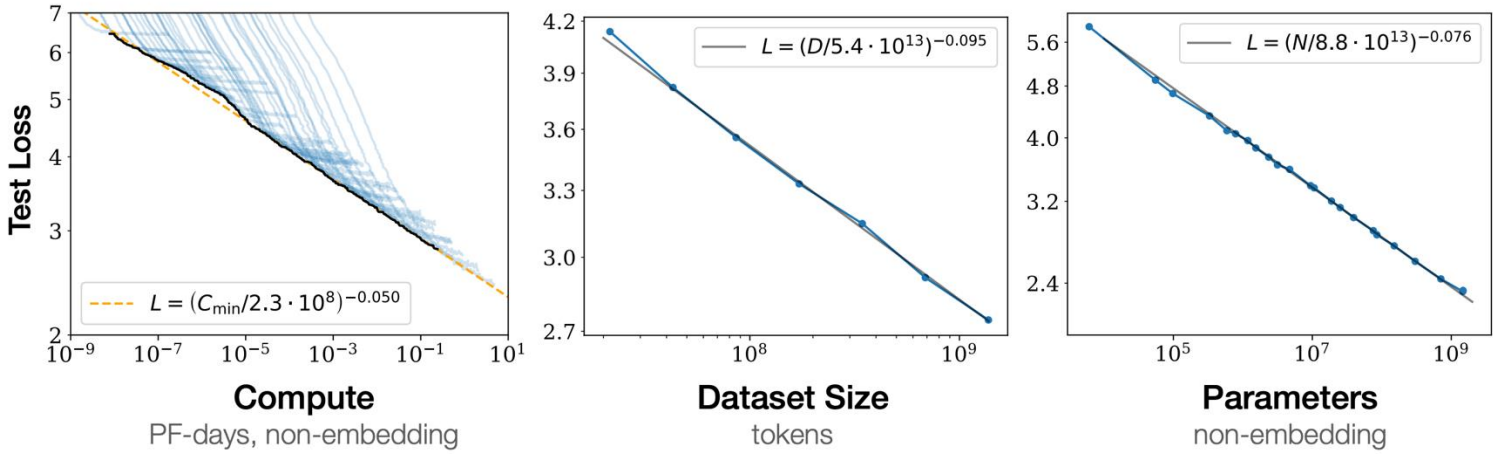
	EAI-EVAL	T0-EVAL
Causal decoder	**44.2**	**42.4**
Non-causal decoder	43.5	41.8
Encoder-decoder	39.9	41.7
Random baseline	32.9	41.7

Finding 1. **Causal decoder-only** models pretrained with a **full language modeling** objective achieve best zero-shot generalization when evaluated immediately after **self-supervised pretraining**, in line with current common practices for large language models.

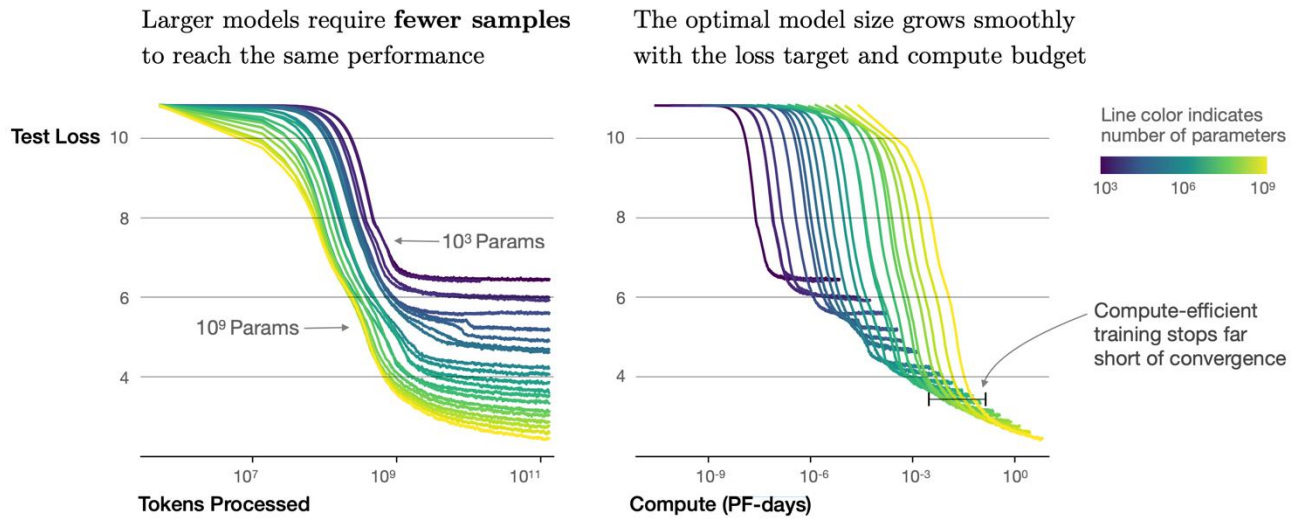
Finding 2. **Encoder-decoder** models pretrained with **masked language modeling** achieve the best zero-shot performance after **multitask finetuning**. More broadly, approaches that perform well in the single-task finetuning setting perform well on multi-task finetuning.

Wang, Thomas, et al. "What language model architecture and pretraining objective works best for zero-shot generalization?." *International Conference on Machine Learning*. PMLR, 2022.

The scaling laws is not just for architectures



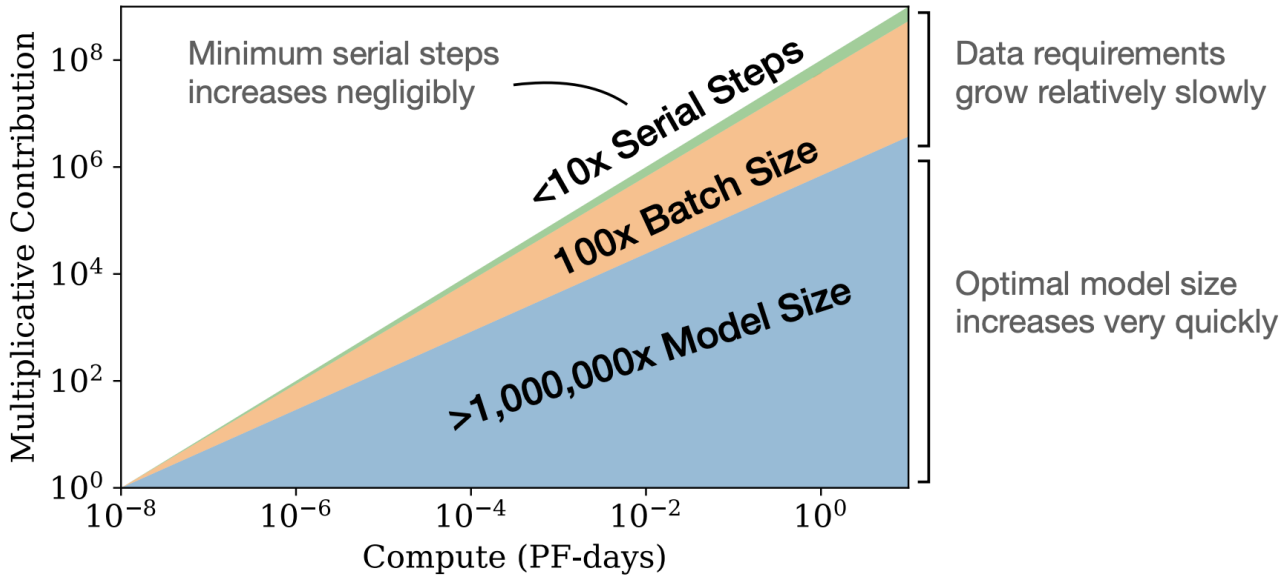
The scaling laws for compute, data, and parameters



The bigger the model, The easier it is to learn

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

The scaling laws is not just for architectures

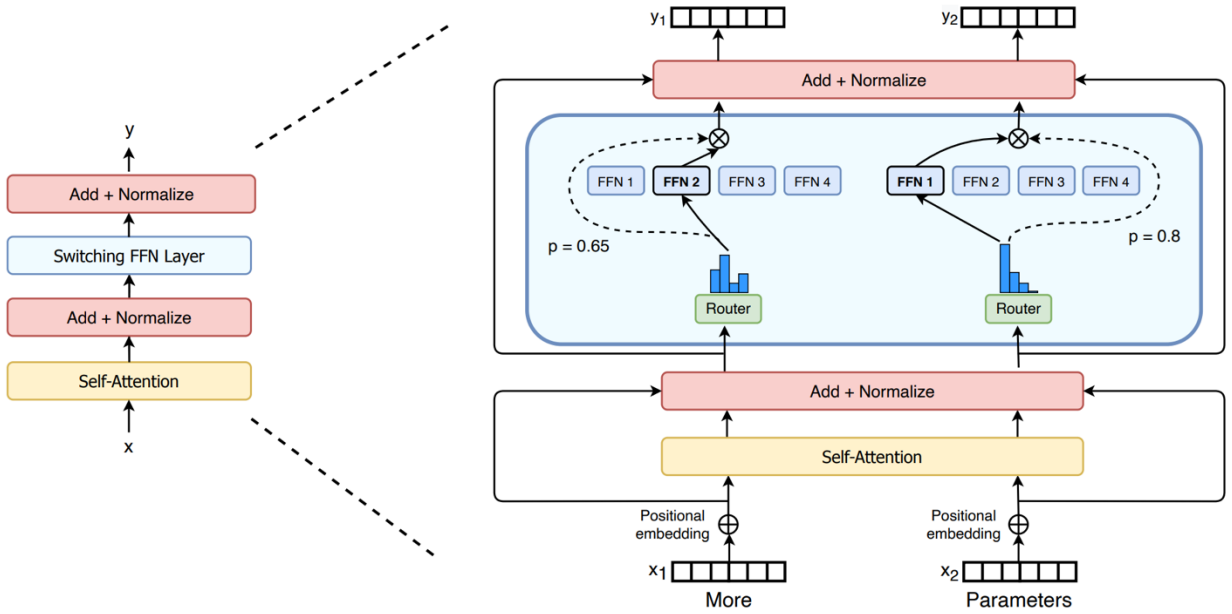


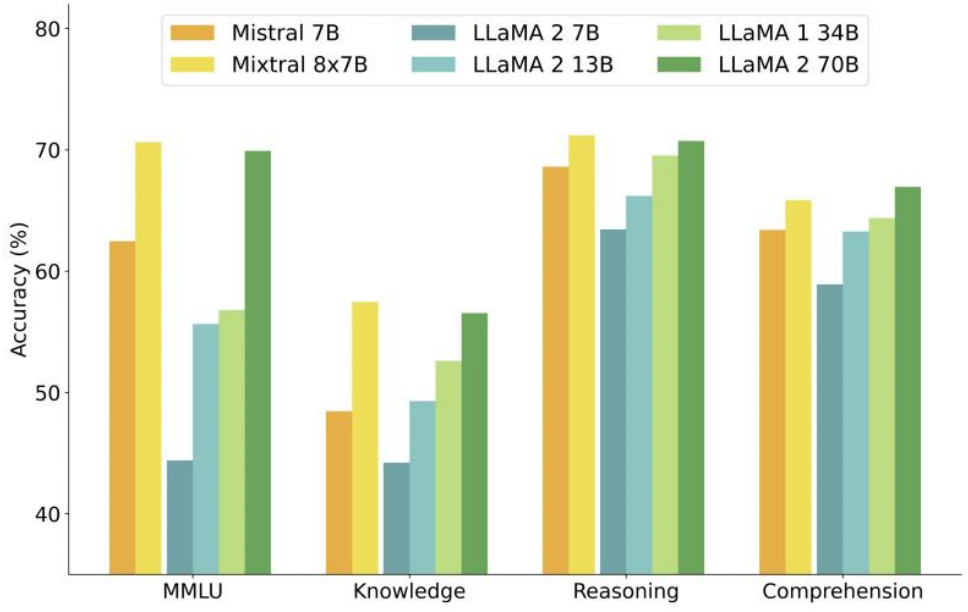
Increase model size as much as possible

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

Sardana, Nikhil, and Jonathan Frankle. "Beyond chinchilla-optimal: Accounting for inference in language model scaling laws." *arXiv preprint arXiv:2401.00448* (2023).

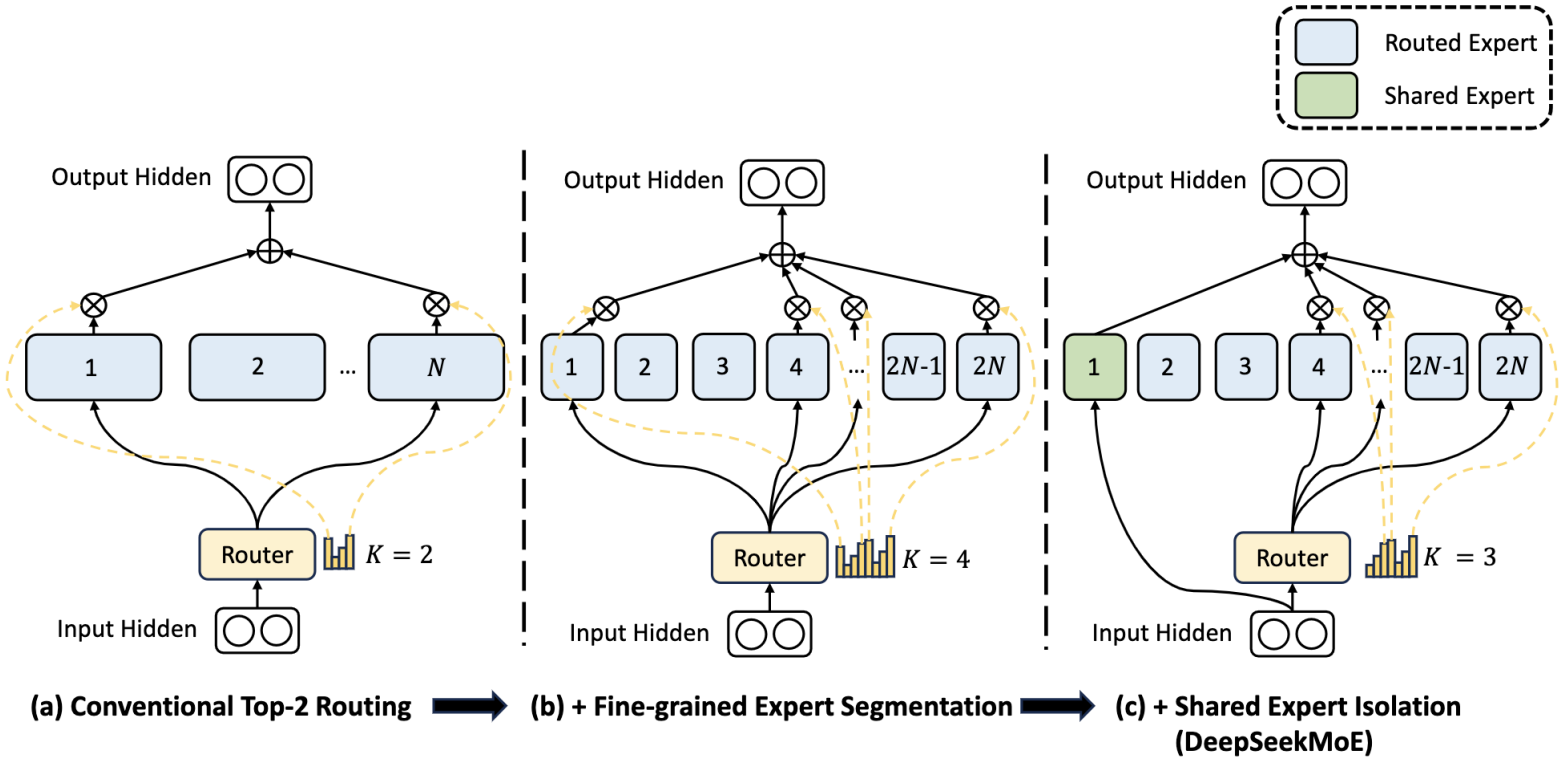
Mixture of Experts (MoEs): Scaling up the # of parameters





Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. The Journal of Machine Learning Research, 23(1), 5232-5270.

Deepseek MOE: a better load-balancing method



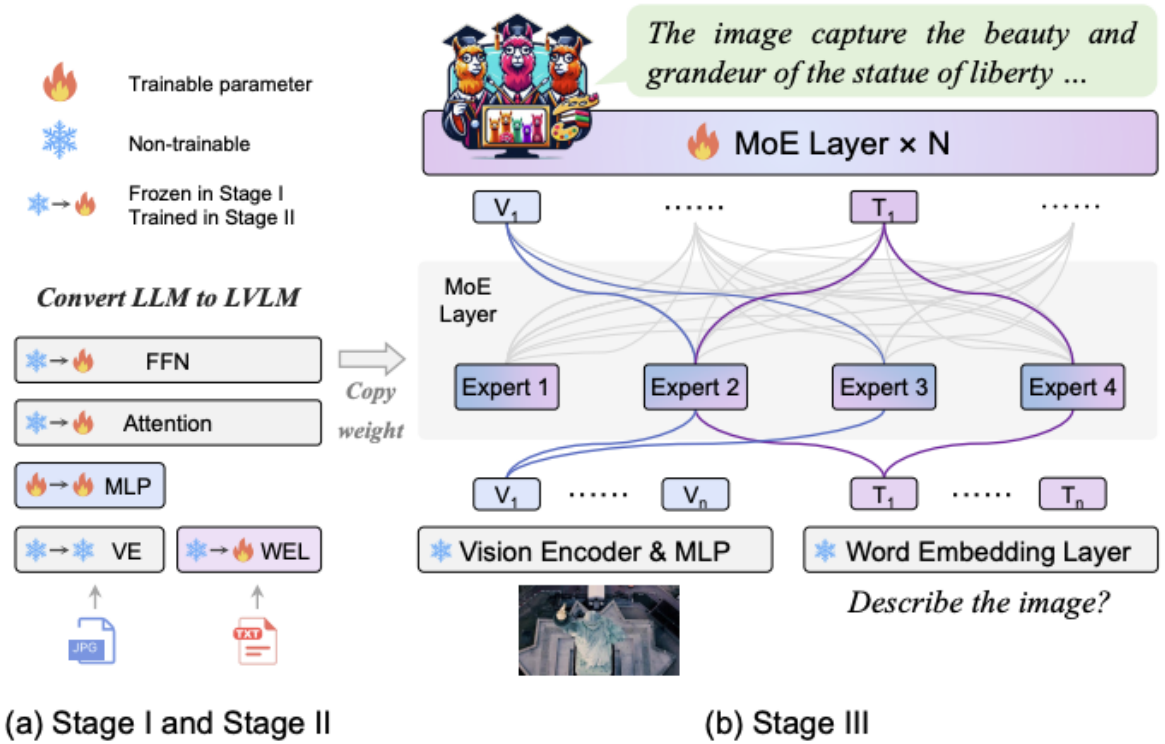
Dai, Damai, et al. "Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models." *arXiv preprint arXiv:2401.06066* (2024).

Deepseek MOE: a better load-balancing method

Metric	# Shot	GShard×1.5	Dense×16	DeepSeekMoE
Relative Expert Size	N/A	1.5	1	0.25
# Experts	N/A	0 + 16	16 + 0	1 + 63
# Activated Experts	N/A	0 + 2	16 + 0	1 + 7
# Total Expert Params	N/A	2.83B	1.89B	1.89B
# Activated Expert Params	N/A	0.35B	1.89B	0.24B
FLOPs per 2K Tokens	N/A	5.8T	24.6T	4.3T
# Training Tokens	N/A	100B	100B	100B
Pile (Loss)	N/A	1.808	1.806	1.808
HellaSwag (Acc.)	0-shot	54.4	55.1	54.8
PIQA (Acc.)	0-shot	71.1	71.9	72.3
ARC-easy (Acc.)	0-shot	47.3	51.9	49.4
ARC-challenge (Acc.)	0-shot	34.1	33.8	34.3
RACE-middle (Acc.)	5-shot	46.4	46.3	44.0
RACE-high (Acc.)	5-shot	32.4	33.0	31.7
HumanEval (Pass@1)	0-shot	3.0	4.3	4.9
MBPP (Pass@1)	3-shot	2.6	2.2	2.2
TriviaQA (EM)	5-shot	15.7	16.5	16.6
NaturalQuestions (EM)	5-shot	4.7	6.3	5.7

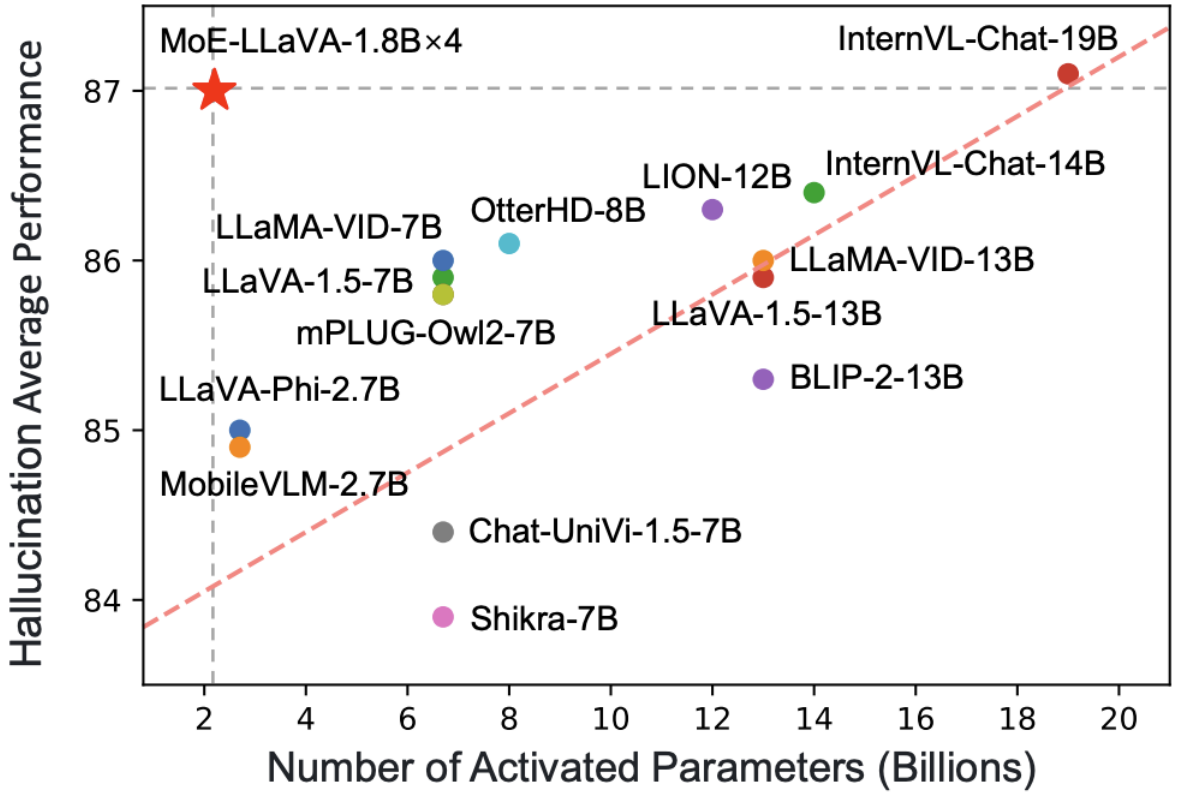
Dai, Damai, et al. "Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models." *arXiv preprint arXiv:2401.06066* (2024).

Mixture of Experts (MoEs): Scaling up the # of parameters



(a) Stage I and Stage II

(b) Stage III

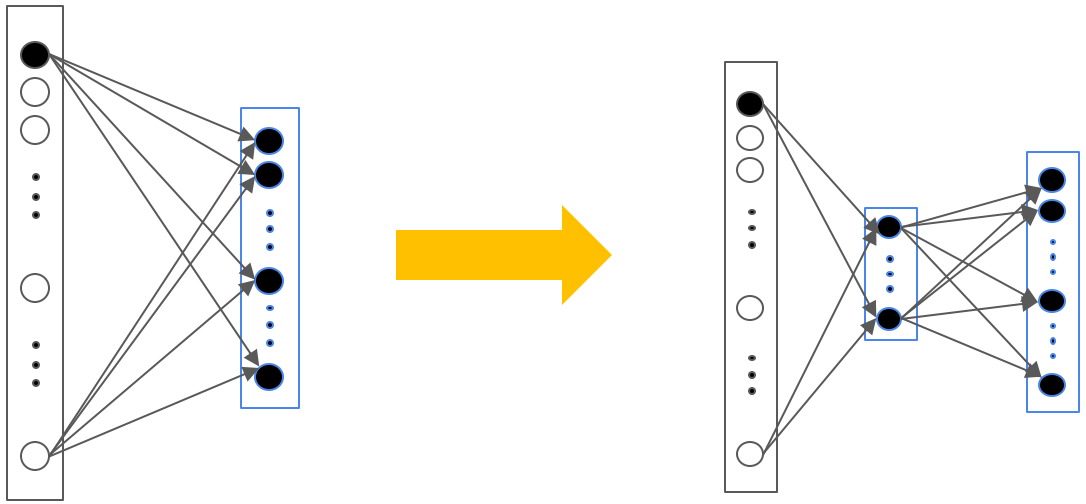


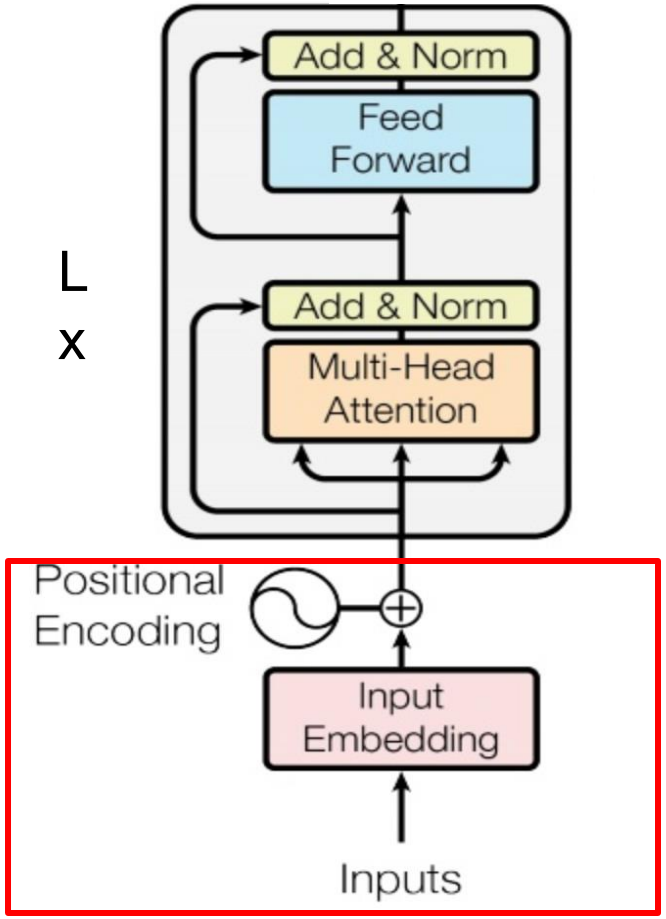
MoE-LLaVA: Mixture of Experts for Large Vision-Language Models

Low-rank word embedding: smarter parameter usage

- Token embeddings are context independent while hidden layer embeddings are context dependent.
- Token embeddings are sparsely updated.

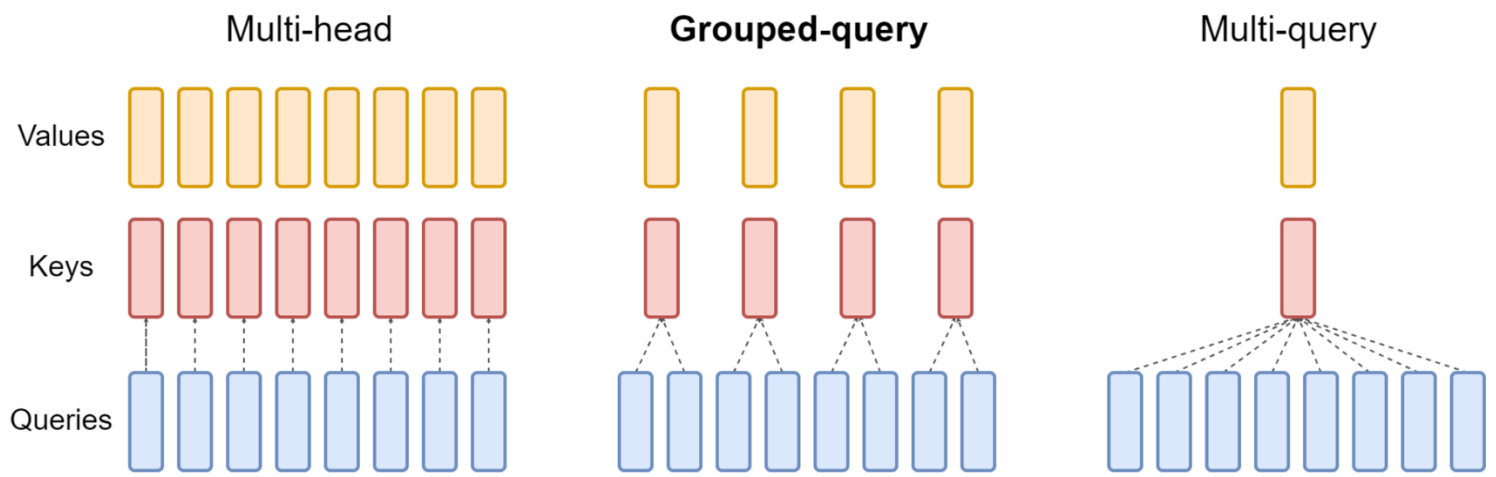
O (V x H) → O (V x E + E x H) where E << H





Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

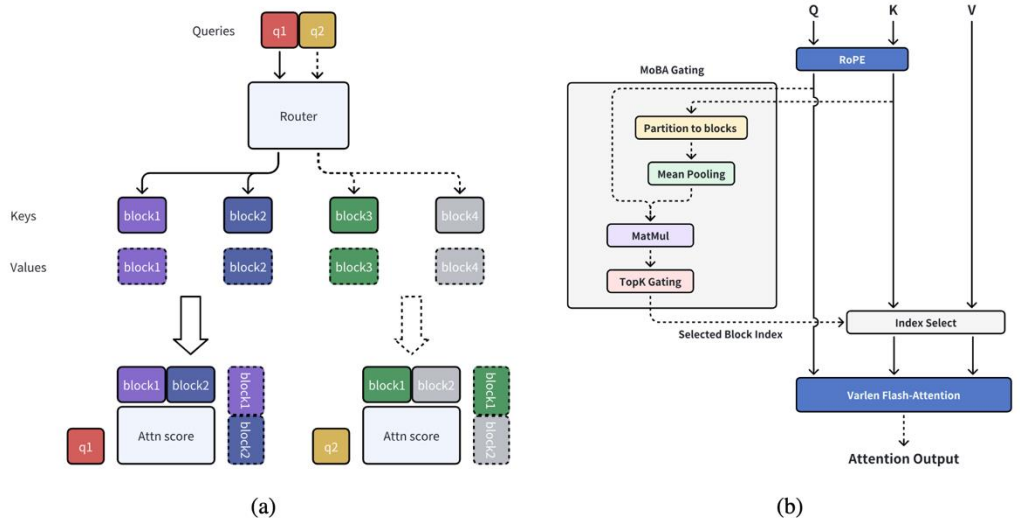
Multi-query attention and multi-group attention: smarter parameter usage



Model	T_{infer}	Average	CNN	arXiv	PubMed	MediaSum	MultiNews	WMT	TriviaQA
	s		R_1	R_1	R_1	R_1	R_1	BLEU	F1
MHA-Large	0.37	46.0	42.9	44.6	46.2	35.5	46.6	27.7	78.2
MHA-XXL	1.51	47.2	43.8	45.6	47.5	36.4	46.9	28.4	81.9
MQA-XXL	0.24	46.6	43.0	45.0	46.9	36.1	46.5	28.5	81.3
GQA-8-XXL	0.28	47.1	43.5	45.4	47.7	36.3	47.2	28.4	81.6

Ainslie, Joshua, et al. "GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints." arXiv preprint arXiv:2305.13245 (2023).

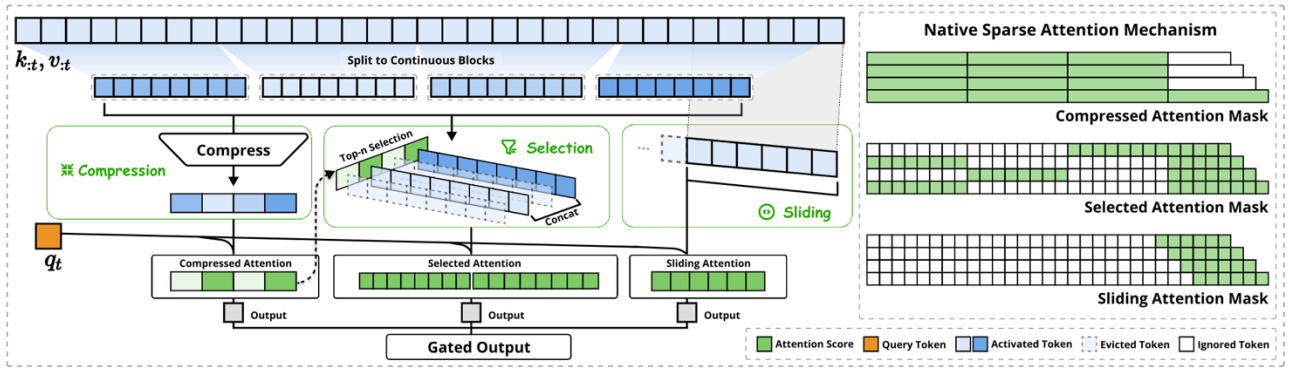
MoBA: Mixture of Block Attention for Long-Context LLMs

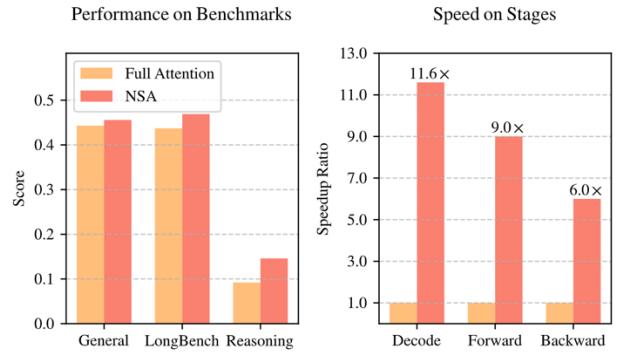


L(C)	MoBA	Full
LM loss (seqlen=8K)	$2.625 \times C^{-0.063}$	$2.622 \times C^{-0.063}$
Trailing LM loss (seqlen=32K, last 2K)	$1.546 \times C^{-0.108}$	$1.464 \times C^{-0.097}$

Lu, Enzhe, et al. "MoBA: Mixture of Block Attention for Long-Context LLMs." arXiv preprint arXiv:2502.13189 (2025).

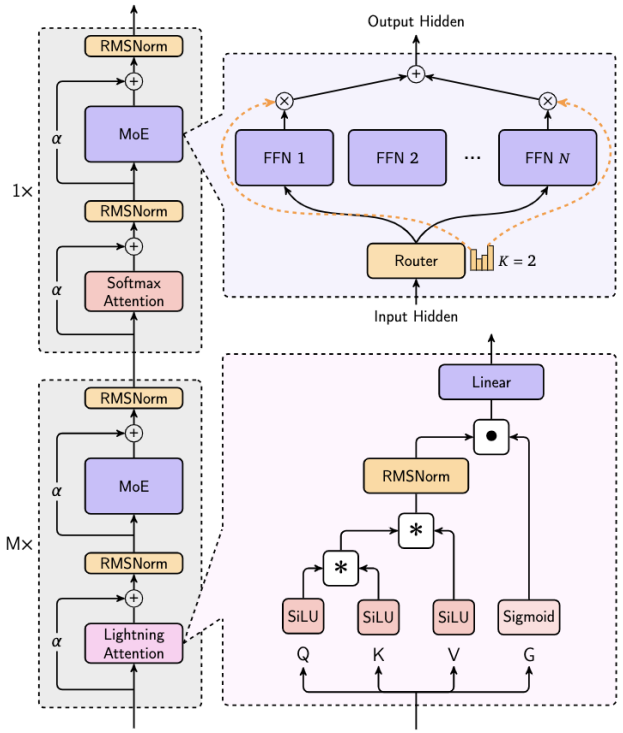
Native Sparse Attention: Hardware-Aligned and Natively Trainable Sparse Attention

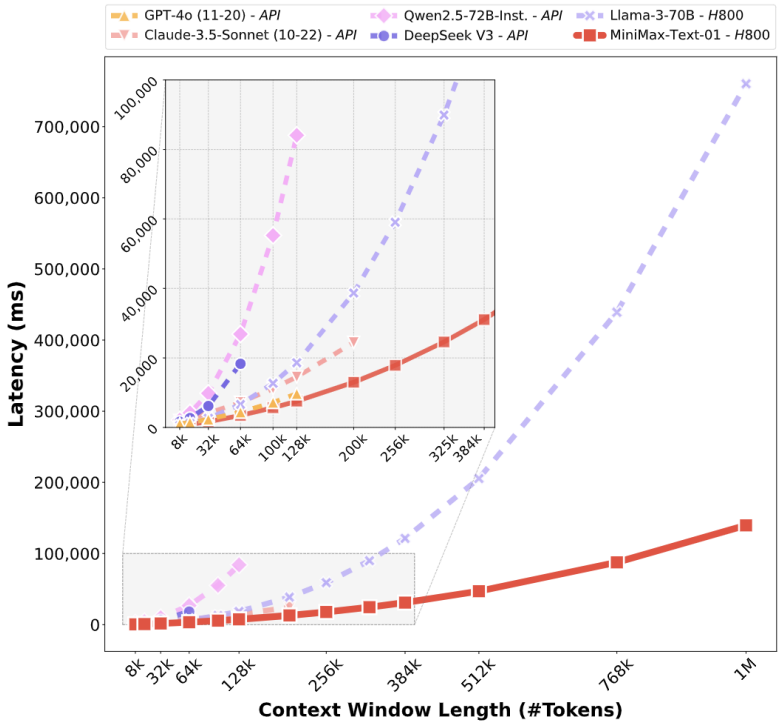




Yuan, Jingyang, et al. "Native sparse attention: Hardware-aligned and natively trainable sparse attention, 2025." URL https://arxiv.org/abs/2502.11089.

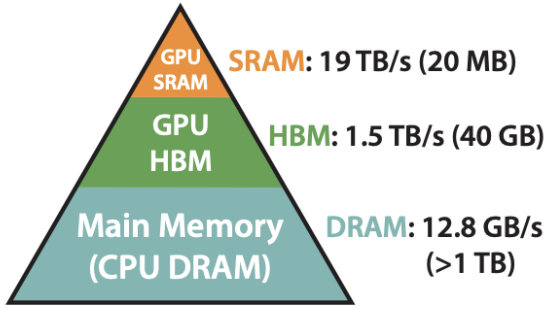
Hybrid attention





Li, Aonian, et al. "Minimax-01: Scaling foundation models with lightning attention." arXiv preprint arXiv:2501.08313 (2025).
Team, Gemma, et al. "Gemma 3 Technical Report." arXiv preprint arXiv:2503.19786 (2025).

Flash attention: faster computation



Memory Hierarchy with Bandwidth & Memory Size

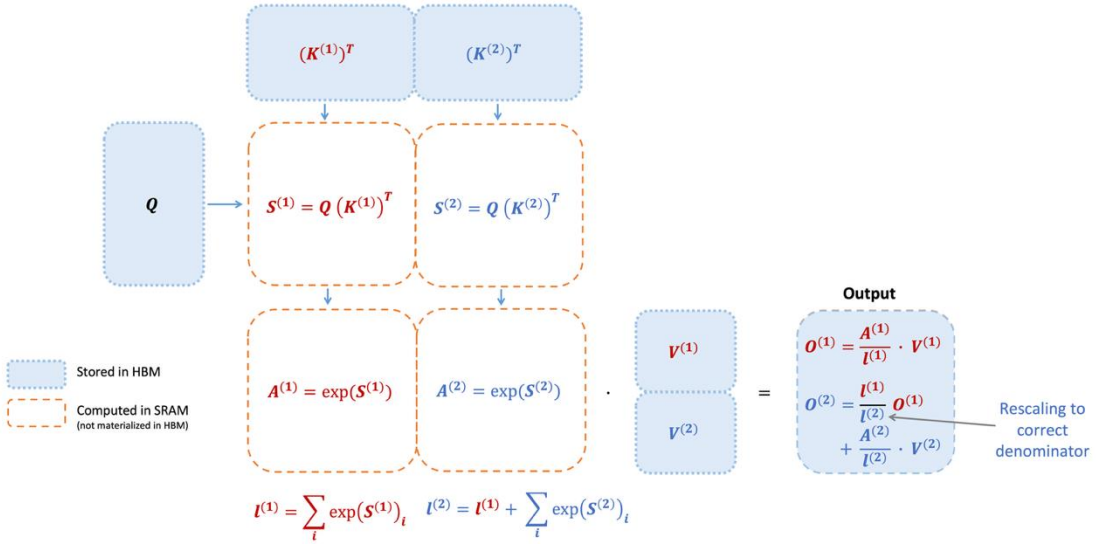
Algorithm 0 Standard Attention Implementation

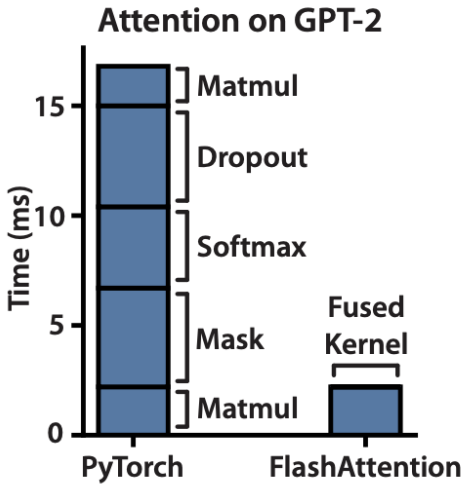
Require: Matrices $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{N \times d}$ in HBM.

1: Load $\mathbf{Q}, \mathbf{K}$ by blocks from HBM, compute $\mathbf{S} = \mathbf{Q}\mathbf{K}^\top$, write $\mathbf{S}$ to HBM.
2: Read $\mathbf{S}$ from HBM, compute $\mathbf{P} = \mathrm{softmax}(\mathbf{S})$, write $\mathbf{P}$ to HBM.
3: Load $\mathbf{P}$ and $\mathbf{V}$ by blocks from HBM, compute $\mathbf{O} = \mathbf{P}\mathbf{V}$, write $\mathbf{O}$ to HBM.
4: Return $\mathbf{O}$.

Dao, Tri, et al. "Flashattention: Fast and memory-efficient exact attention with io-awareness." *Advances in Neural Information Processing Systems* 35 (2022): 16344-16359.

Flash attention: faster computation





Dao, Tri, et al. "Flashattention: Fast and memory-efficient exact attention with io-awareness." *Advances in Neural Information Processing Systems* 35 (2022): 16344-16359.

Flash attention: faster computation

Model implementations	OpenWebText (ppl)	Training time (speedup)
GPT-2 small - Huggingface [87]	18.2	9.5 days (1.0×)
GPT-2 small - Megatron-LM [77]	18.2	4.7 days (2.0×)
GPT-2 small - FlashAttention	18.2	**2.7 days (3.5×)**
GPT-2 medium - Huggingface [87]	14.2	21.0 days (1.0×)
GPT-2 medium - Megatron-LM [77]	14.3	11.5 days (1.8×)
GPT-2 medium - FlashAttention	14.3	**6.9 days (3.0×)**

Dao, Tri, et al. "Flashattention: Fast and memory-efficient exact attention with io-awareness." *Advances in Neural Information Processing Systems* 35 (2022): 16344-16359.

Mixed precision training: faster computation

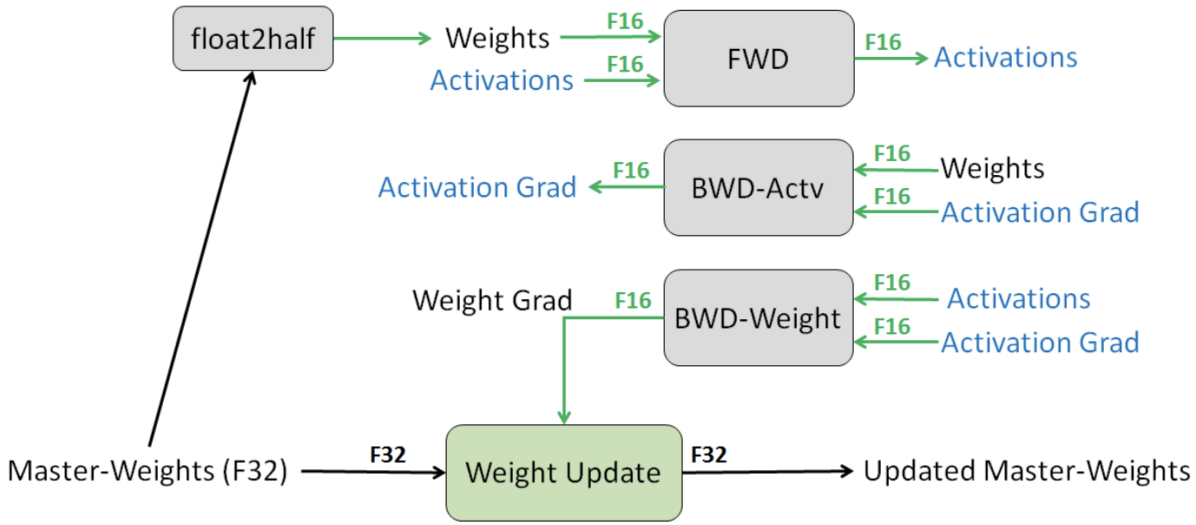


Figure 1: Mixed precision training iteration for a layer.

Model	Baseline	Mixed Precision
AlexNet	56.77%	56.93%
VGG-D	65.40%	65.43%
GoogLeNet (Inception v1)	68.33%	68.43%
Inception v2	70.03%	70.02%
Inception v3	73.85%	74.13%
Resnet50	75.92%	76.04%

Micikevicius, Paulius, et al. "Mixed precision training." *arXiv preprint arXiv:1710.03740* (2017).

Foundation for foundation models

Principle 3 (the scaling law): AI methods that leverage computation are ultimately the most effective way of improvements (from "The bitter lesson" by Rich Sutton)

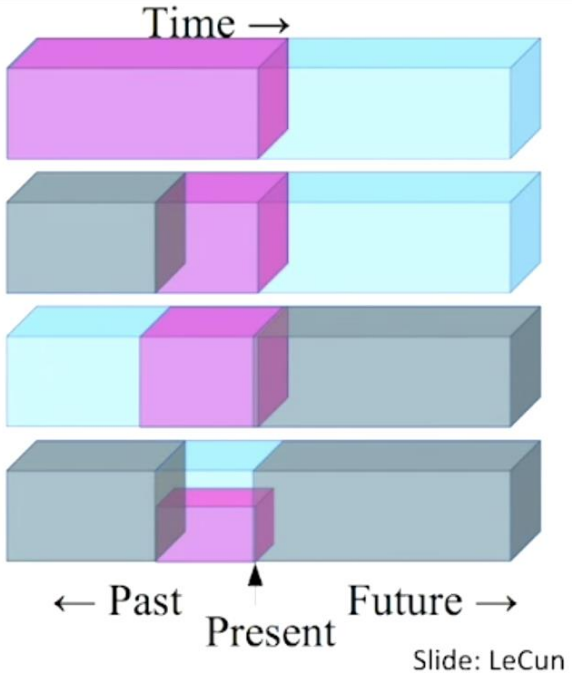
What is the most effective network architecture to leverage computation?
(Comparison between CNNs and Transformer)

Principle 4 (the data law): Data is the ultimate way of regularization

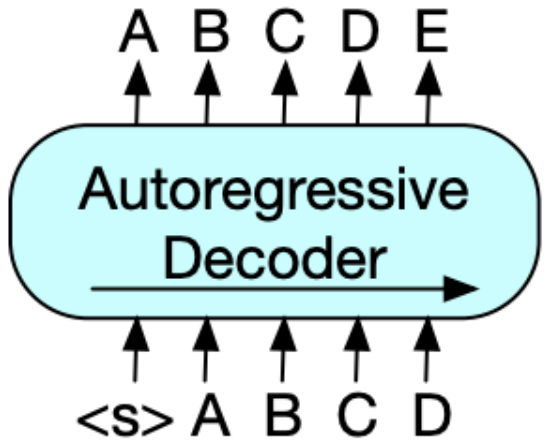
What is the most effective way of (pre-)training the network?

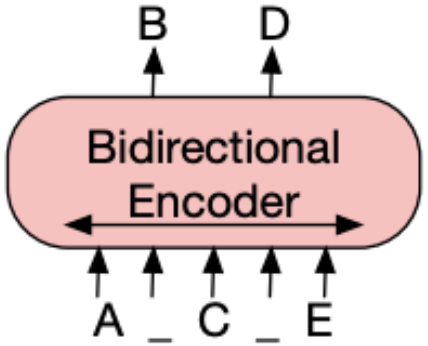
Self-supervised learning

- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the future from the recent past.
- Predict the past from the present.
- Predict the top from the bottom.
- Predict the occluded from the visible
- Pretend there is a part of the input you don't know and predict that.

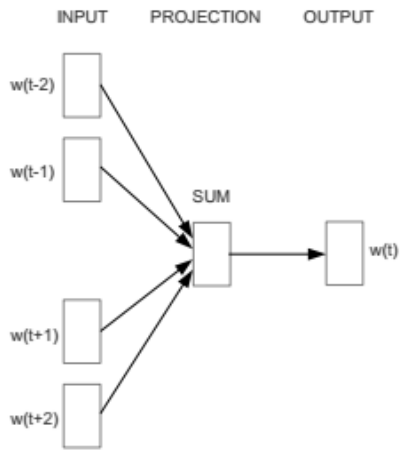


Slide: LeCun

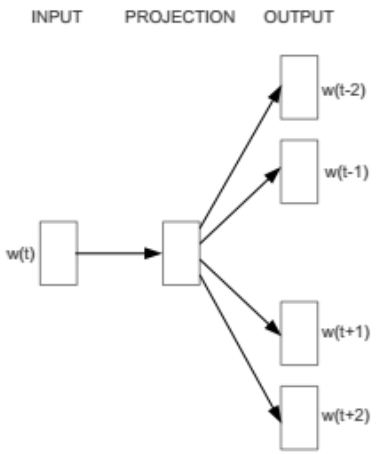




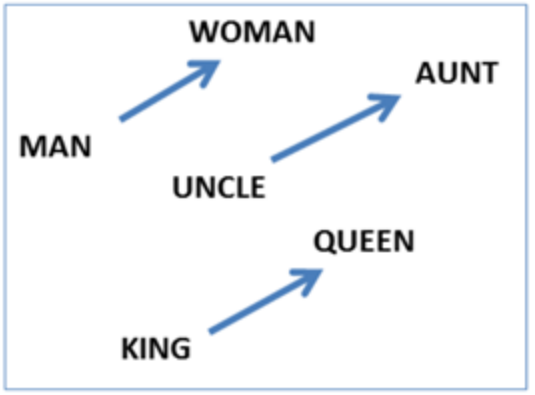
Word2Vec: word level SSL



CBOW

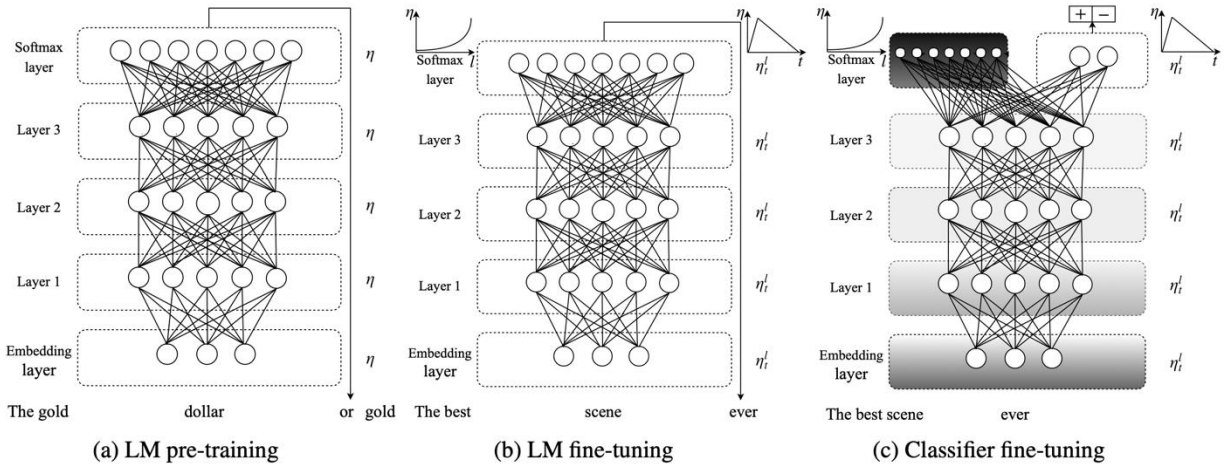


Skip-gram



Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *NIPS* 2013.

Universal Language model fine-tuning (ULMFiT): major learning paradigm shift



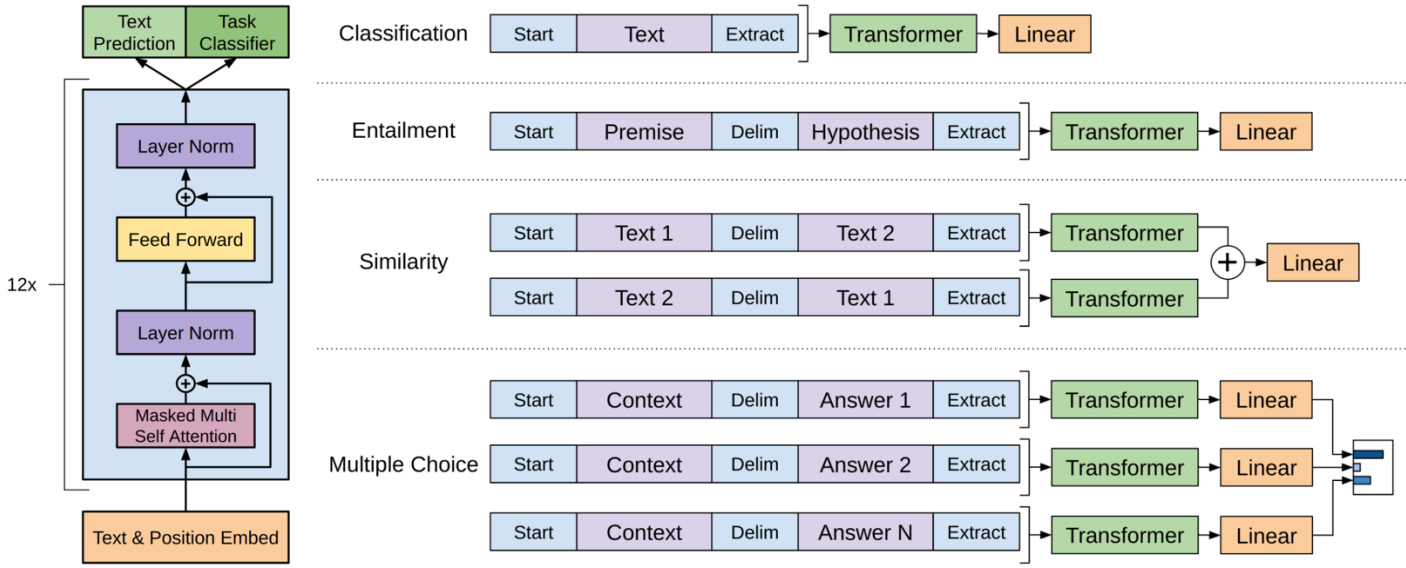
(a) LM pre-training (b) LM fine-tuning (c) Classifier fine-tuning

Pretraining	IMDb	TREC-6	AG
Without pretraining	5.63	10.67	5.52
With pretraining	**5.00**	**5.69**	**5.38**

First direct finetuning concept : end2end

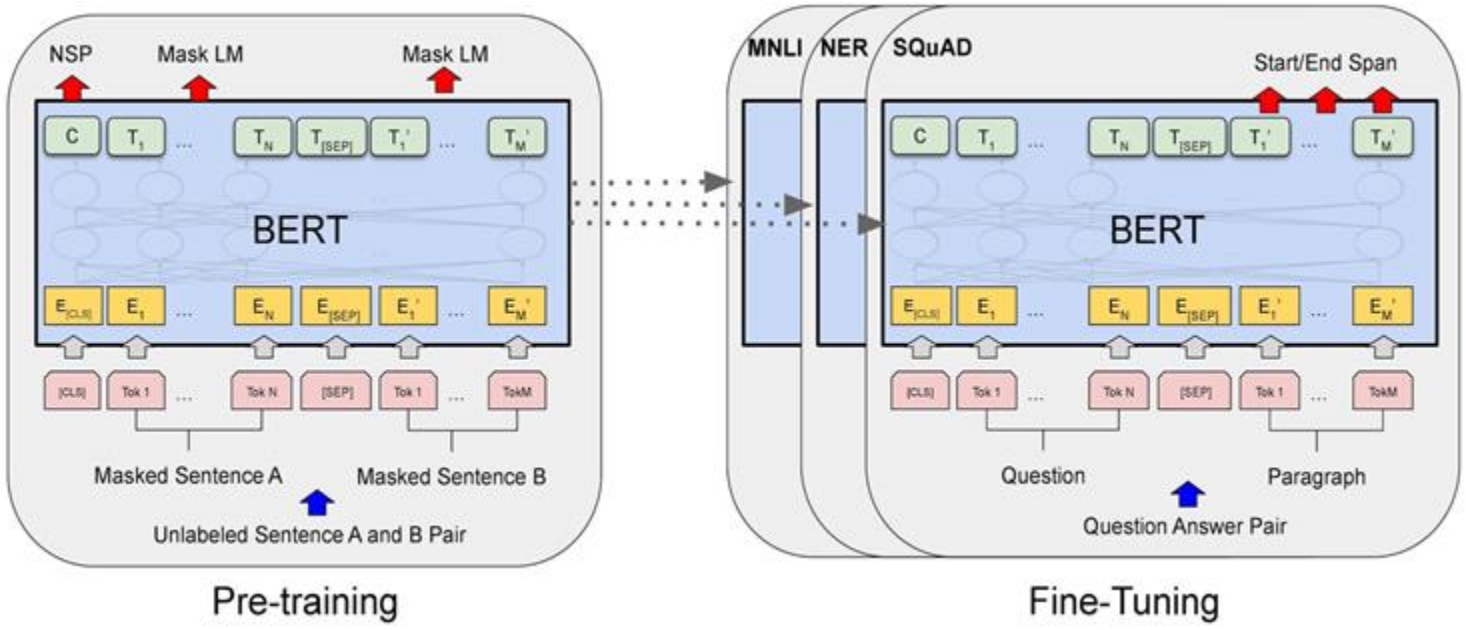
Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.

Generative Pre-training: general pre-training



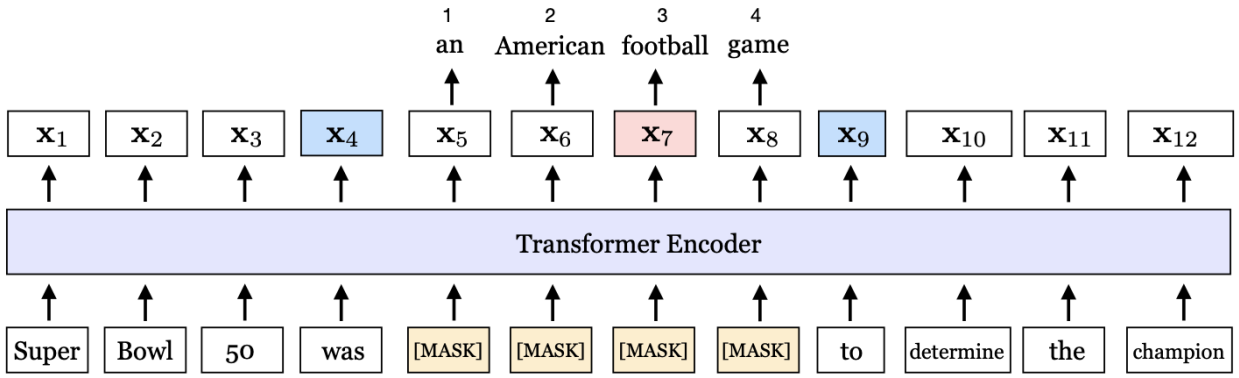
Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. (GPT-1)

BERT: pretraining through masked language model



Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

Span masking is better than random masking



	NewsQA	TriviaQA	SearchQA	HotpotQA	Natural Questions	Avg.
Google BERT	68.8	77.5	81.7	78.3	79.9	77.3
Our BERT	71.0	79.0	81.8	80.5	80.5	78.6
Our BERT-1seq	71.9	80.4	84.0	80.3	81.8	79.7
SpanBERT	**73.6**	**83.6**	**84.8**	**83.0**	**82.5**	**81.5**

Joshi, Mandar, et al. "Spanbert: Improving pre-training by representing and predicting spans." *Transactions of the Association for Computational Linguistics* 8 (2020): 64-77.

Smart span masking

Sentence	Harry	Potter	is	a	series	of	fantasy	novels	written	by	British	author	J.	K.	Rowling
Basic-level Masking	[mask]	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	J.	[mask]	Rowling
Entity-level Masking	Harry	Potter	is	a	series	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]
Phrase-level Masking	Harry	Potter	is	[mask]	[mask]	[mask]	fantasy	novels	[mask]	by	British	author	[mask]	[mask]	[mask]

mask strategy	dev Accuracy	test Accuracy
word-level(chinese character)	77.7%	76.8%
word-level&phrase-level	78.3%	77.3%
word-level&phrase-leve&entity-level	78.7%	77.6%

Sun, Yu, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. "Ernie: Enhanced representation through knowledge integration." *arXiv preprint arXiv:1904.09223* (2019).

Next sentence prediction vs Sentence Order prediction

Sentences are next to each other

1st google is an american multinational technology company that specializes in internet-related services.

2nd it is considered one of the big four technology companies, alongside amazon, apple and facebook.

Sentences are from different documents

1st google is an american multinational technology company that specializes in internet-related services.

2nd california is a state in the pacific region of the united states.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Next sentence prediction vs Sentence Order prediction

Sentences are next to each other

1st google is an american multinational technology company that specializes in internet-related services.

2nd it is considered one of the big four technology companies, alongside amazon, apple and facebook.

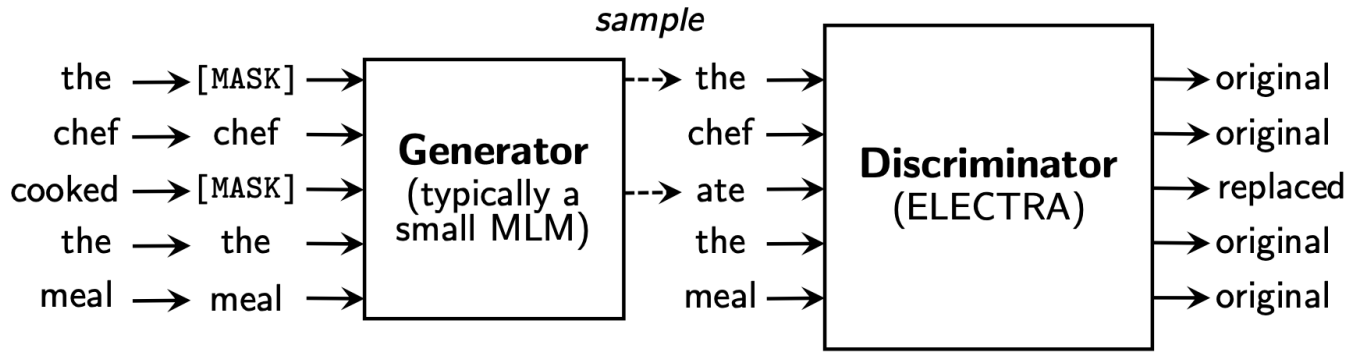
Simply reverse the order

1st it is considered one of the big four technology companies, alongside amazon, apple and facebook.

2nd google is an american multinational technology company that specializes in internet-related services.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

ELECTRA improves training efficiency

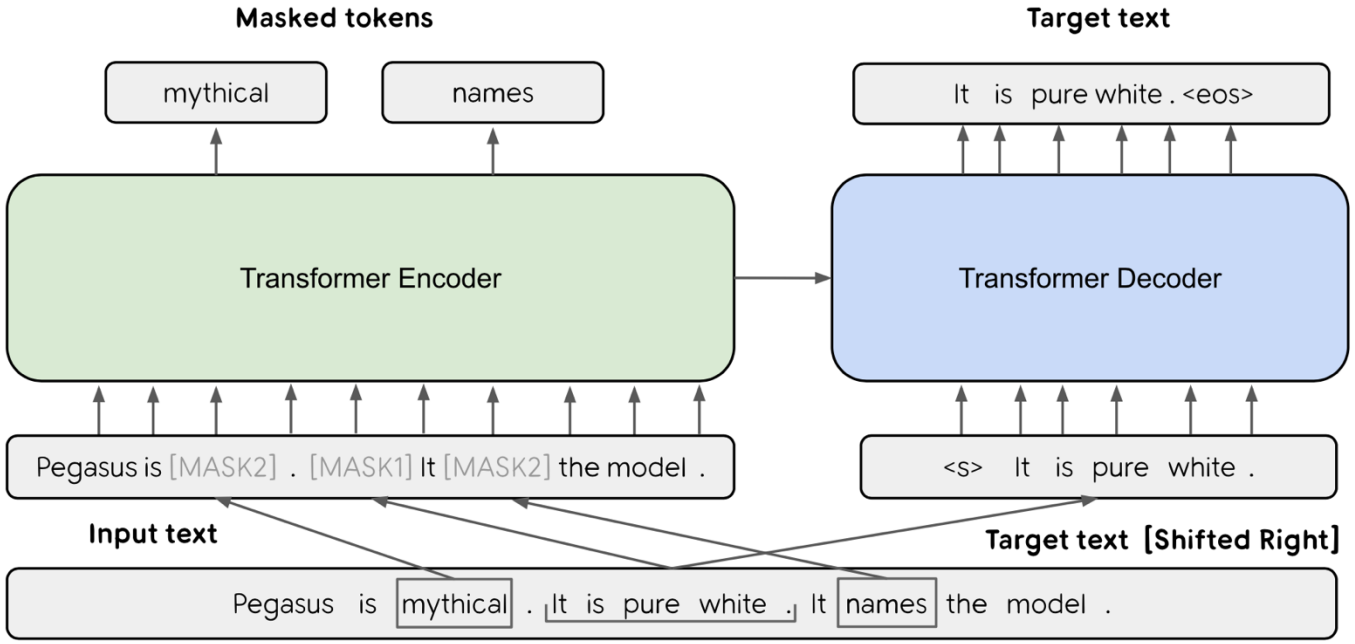


Model	ELECTRA	All-Tokens MLM	Replace MLM	ELECTRA 15%	BERT
GLUE score	85.0	84.3	82.4	82.4	82.2

Inconsistency?

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

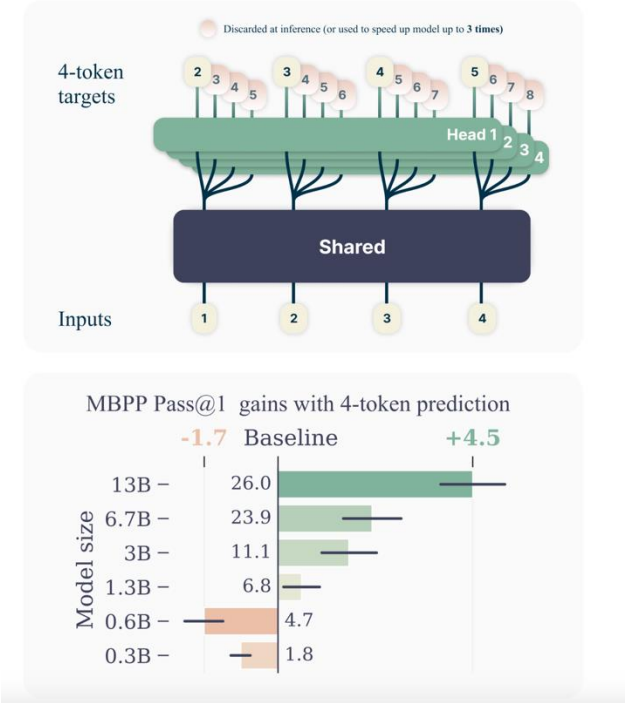
PEGASUS improves both language understanding and generation



Zhang, Jingqing, et al. "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization." *arXiv preprint arXiv:1912.08777* (2019).

He, Pengcheng, et al. "Deberta: Decoding-enhanced bert with disentangled attention." *arXiv preprint arXiv:2006.03654* (2020).

Next few tokens prediction improves training efficiency



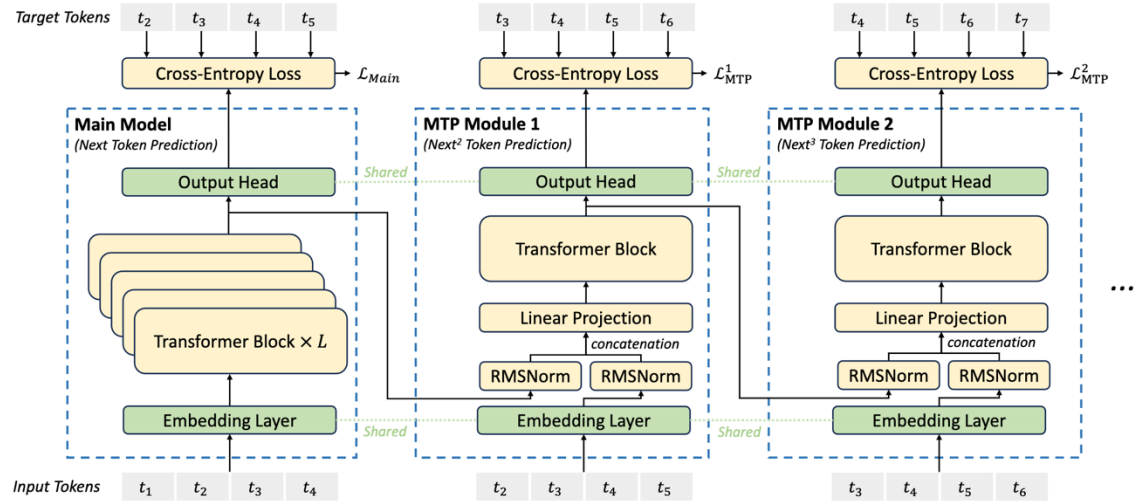


Figure 3 | Illustration of our Multi-Token Prediction (MTP) implementation. We keep the complete causal chain for the prediction of each token at each depth.

Gloeckle, Fabian, et al. "Better & faster large language models via multi-token prediction." *arXiv preprint arXiv:2404.19737* (2024).
Liu, Aixin, et al. "Deepseek-v3 technical report." *arXiv preprint arXiv:2412.19437* (2024).

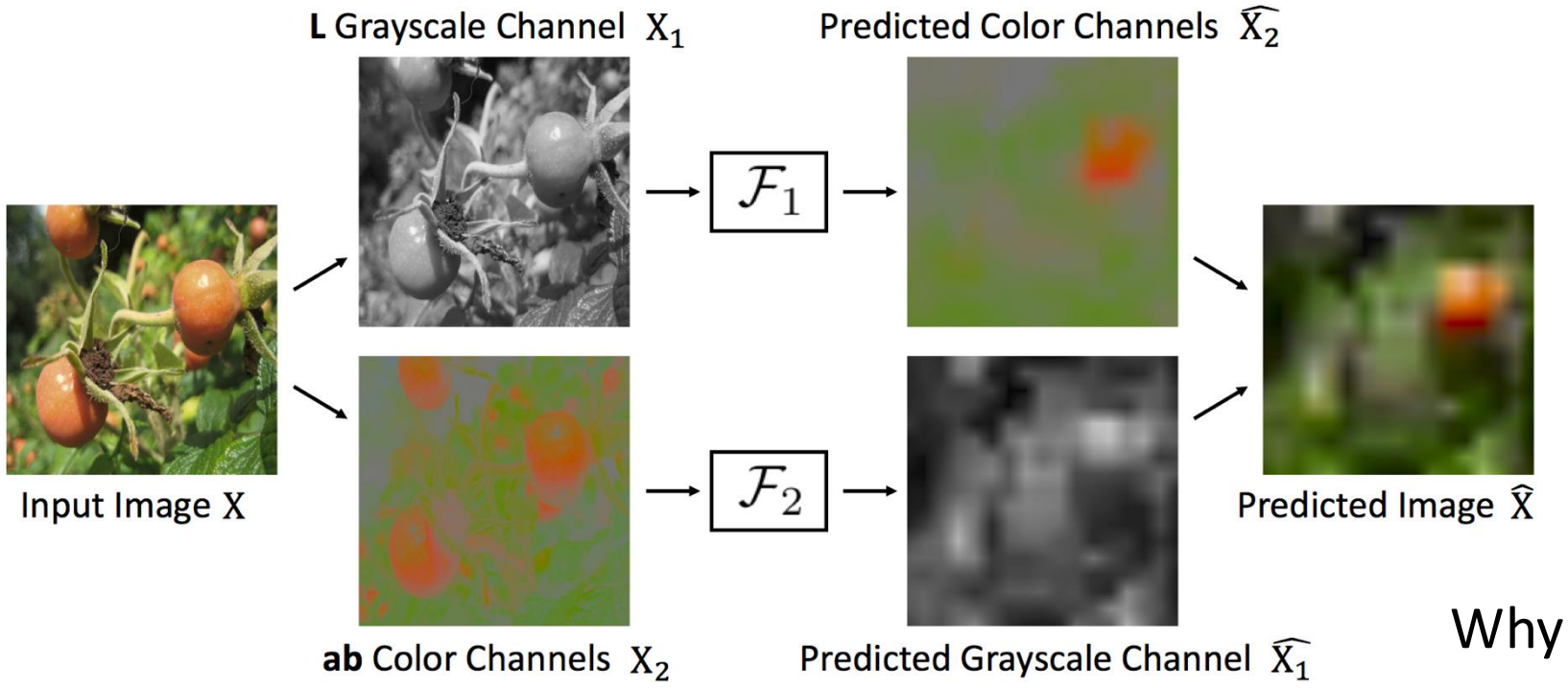
Next few tokens prediction improves training efficiency

Benchmark (Metric)	# Shots	Small MoE Baseline	Small MoE w/ MTP	Large MoE Baseline	Large MoE w/ MTP
# Activated Params (Inference)	-	2.4B	2.4B	20.9B	20.9B
# Total Params (Inference)	-	15.7B	15.7B	228.7B	228.7B
# Training Tokens	-	1.33T	1.33T	540B	540B
Pile-test (BPB)	-	**0.729**	**0.729**	0.658	**0.657**
BBH (EM)	3-shot	39.0	**41.4**	70.0	**70.7**
MMLU (EM)	5-shot	50.0	**53.3**	**67.5**	66.6
DROP (F1)	1-shot	39.2	**41.3**	68.5	**70.6**
TriviaQA (EM)	5-shot	56.9	**57.7**	**67.0**	**67.3**
NaturalQuestions (EM)	5-shot	**22.7**	22.3	27.2	**28.5**
HumanEval (Pass@1)	0-shot	20.7	**26.8**	44.5	**53.7**
MBPP (Pass@1)	3-shot	35.8	**36.8**	61.6	**62.2**
GSM8K (EM)	8-shot	25.4	**31.4**	72.3	**74.0**
MATH (EM)	4-shot	10.7	**12.6**	38.6	**39.8**

Table 4 | Ablation results for the MTP strategy. The MTP strategy consistently enhances the model performance on most of the evaluation benchmarks.

Gloeckle, Fabian, et al. "Better & faster large language models via multi-token prediction." *arXiv preprint arXiv:2404.19737* (2024).
Liu, Aixin, et al. "Deepseek-v3 technical report." *arXiv preprint arXiv:2412.19437* (2024).

Color channel as supervision



Why blurring prediction?

Zhang, R., Isola, P., & Efros, A. A. (2017). Split-brain autoencoders: Unsupervised learning by cross-channel prediction. CVPR 2017

SimCLR: a simple framework for contrastive learning

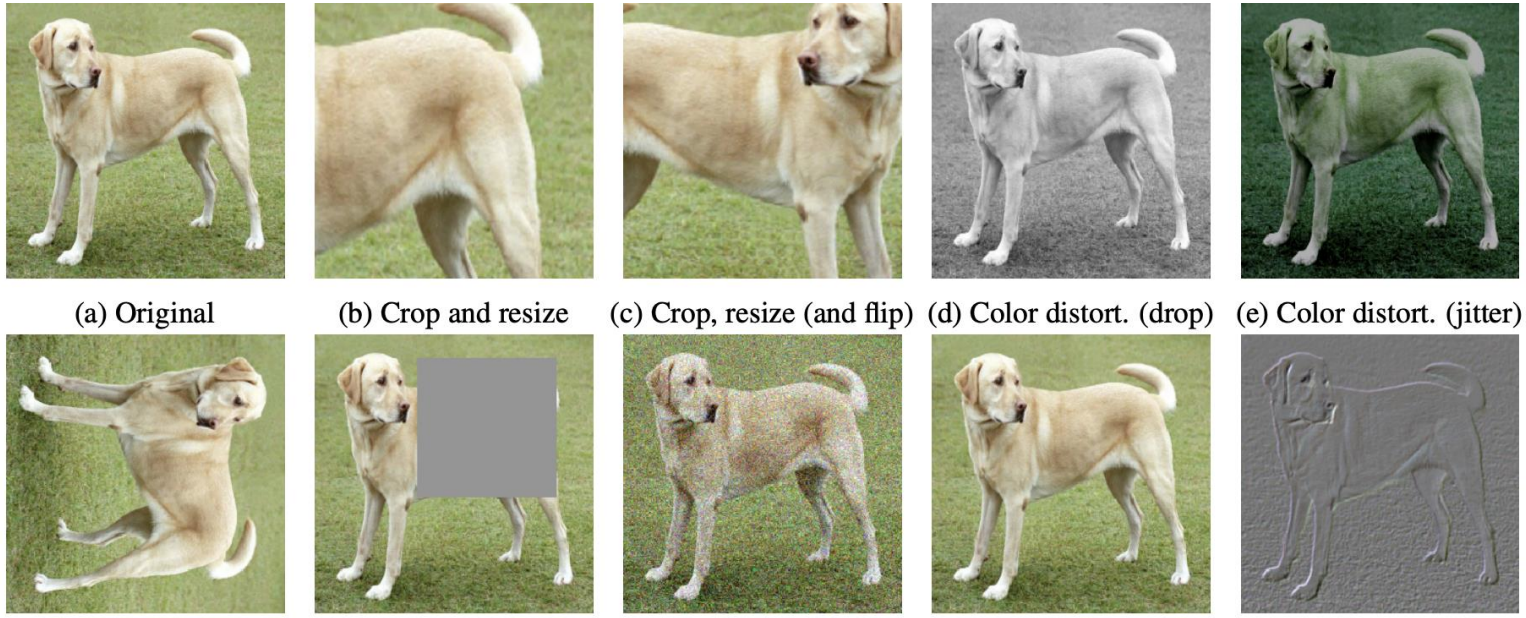
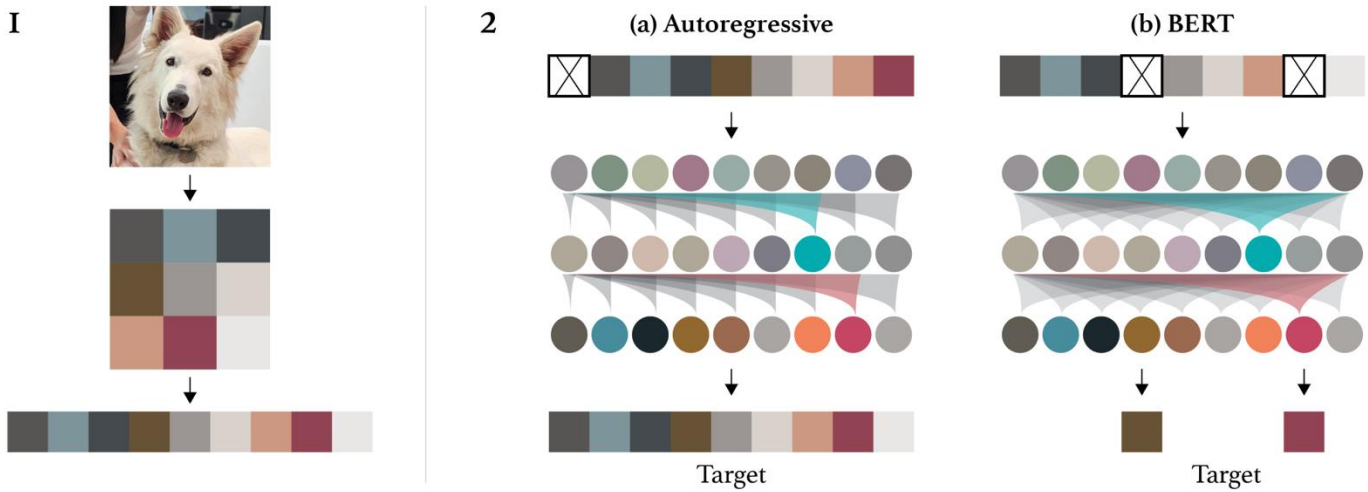


Figure 4. Illustrations of the studied data augmentation operators. Each augmentation can transform data stochastically with some internal parameters (e.g. rotation degree, noise level). Note that we *only* test these operators in ablation, the *augmentation policy used to train our models* only includes *random crop (with flip and resize)*, *color distortion*, and *Gaussian blur*. (Original image cc-by: Von.grzanka)

Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." ICML2020

Image GPT: image tokens as supervision



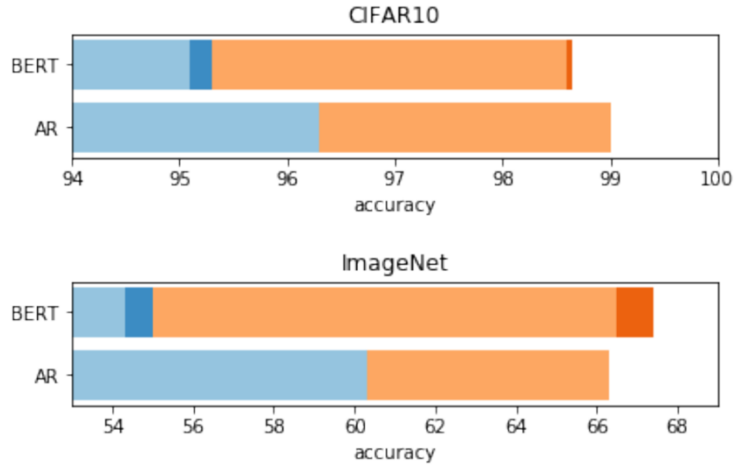
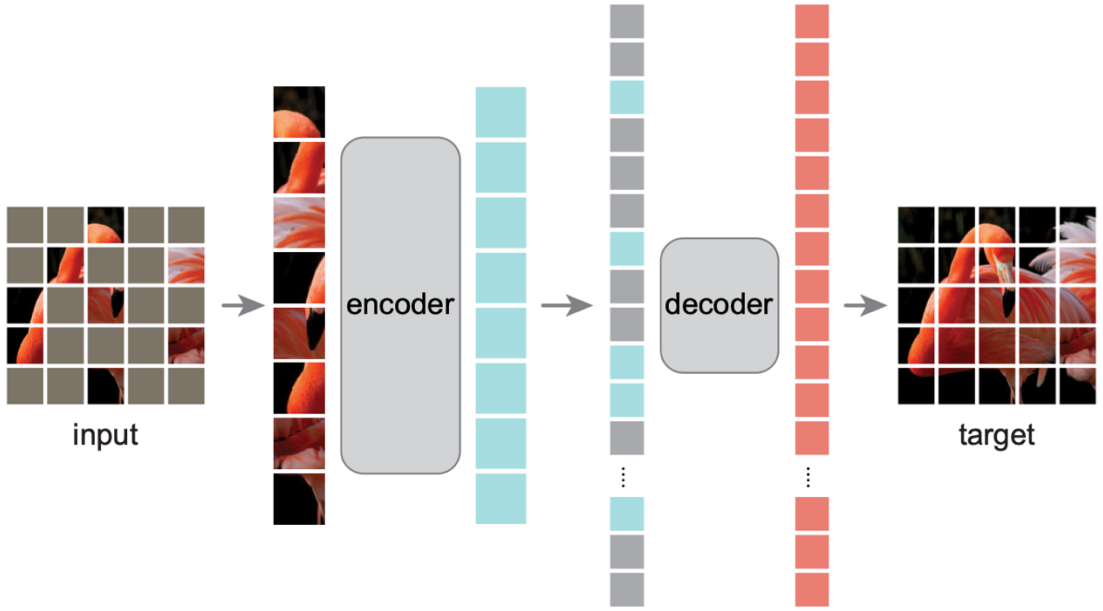
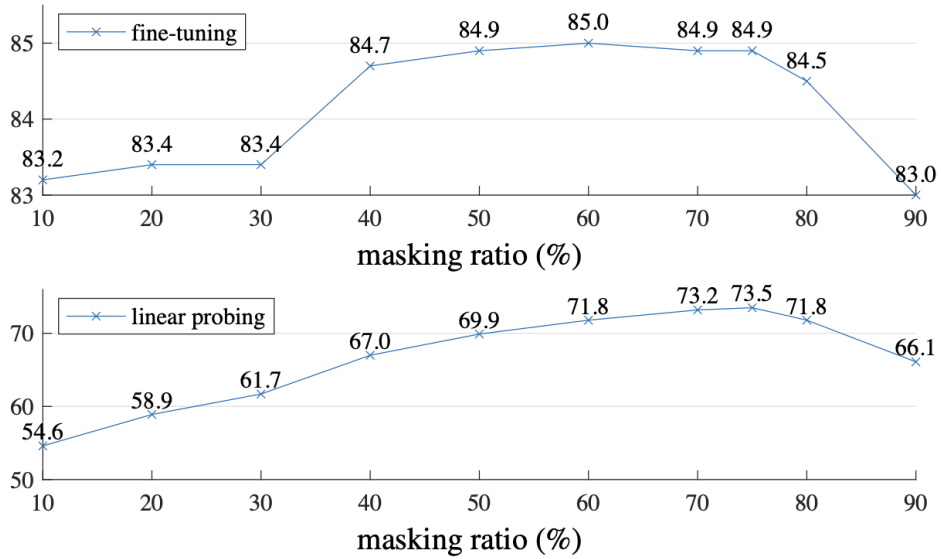


Figure 4. Comparison of auto-regressive pre-training with BERT pre-training using iGPT-L at an input resolution of $32^2 \times 3$. Blue bars display linear probe accuracy and orange bars display fine-tune accuracy. Bold colors show the performance boost from ensembling BERT masks. We see that auto-regressive models produce much better features than BERT models after pre-training, but BERT models catch up after fine-tuning.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. Generative pretraining from pixels. ICML 2020

Masked Autoencoder (MAE)

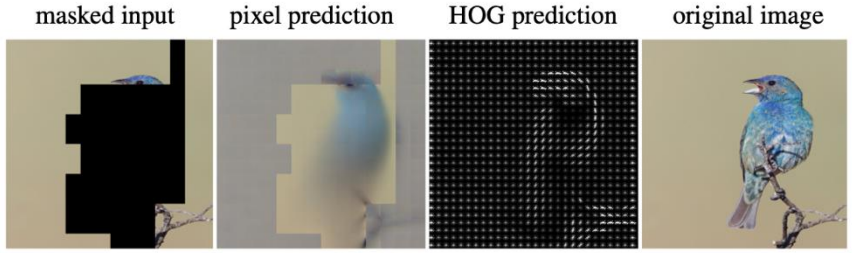




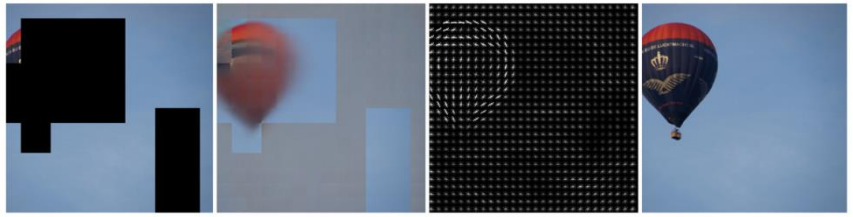
Large encoder, small decoder

He, Kaiming, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. "Masked autoencoders are scalable vision learners." arXiv preprint arXiv:2111.06377 (2021).

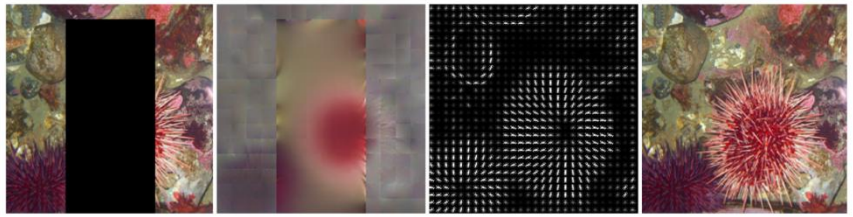
Masked feature prediction (MaskFeat)：feature as supervision



Both two predictions make good sense given a small visible region at the bird's head.



Pixel with **color ambiguity**: Though pixel prediction makes a sensible guess on the balloon, the loss penalty is large because of unmatched color (red *vs*. black).



Pixel with **texture ambiguity**: Pixel prediction is blurry in texture-rich area because of ambiguity, while HOG successfully characterizes major edge directions.

ratio	20%	40%	60%	80%
top-1	81.9 (-0.3)	**82.2**	**82.2**	82.0 (-0.2)

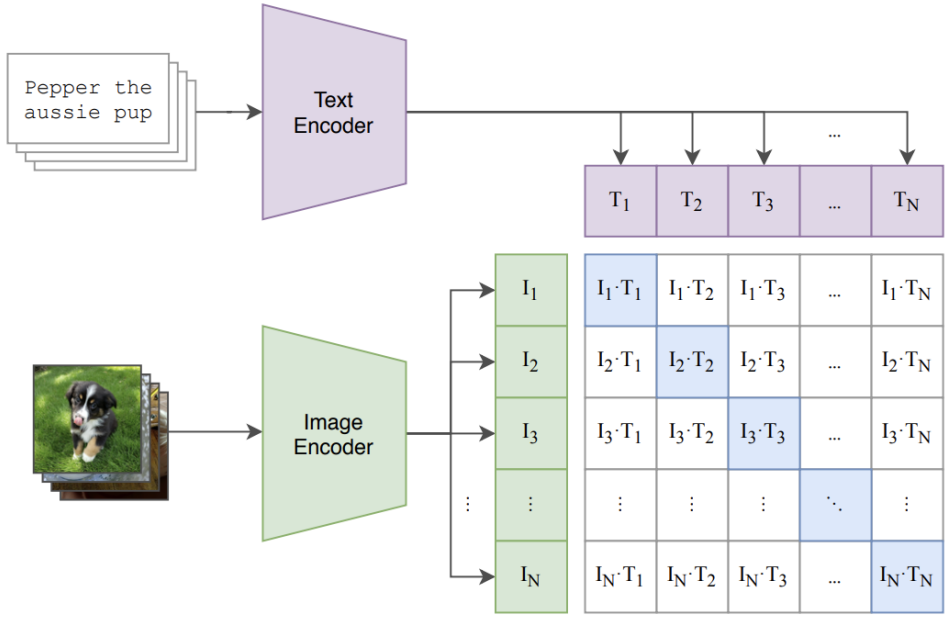
Table 9. **Masking ratio.** Varying the percentage of masked patches. MaskFeat is robust to masking ratio in video domain.

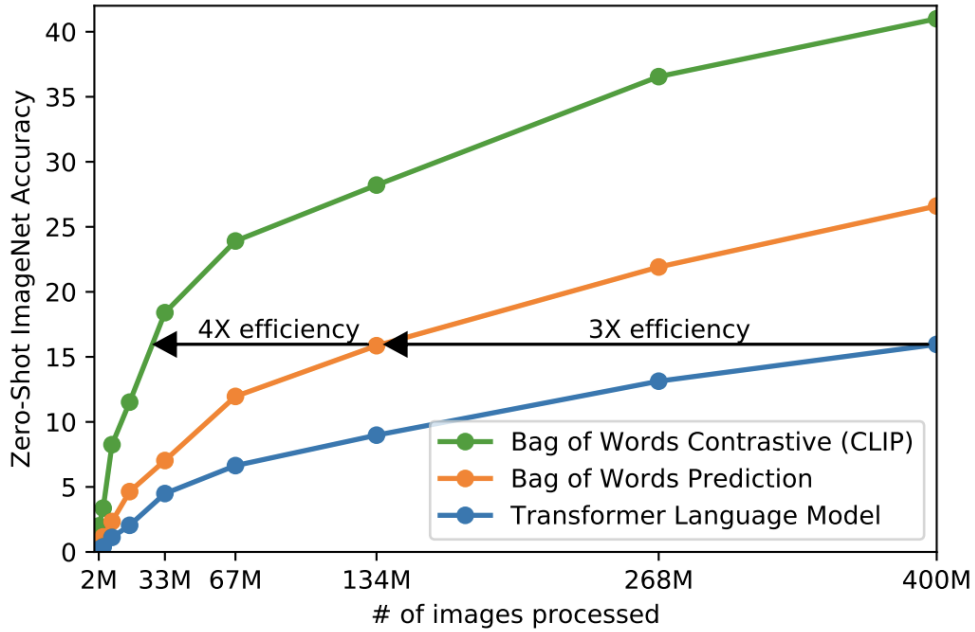
targets	pixel	HOG	pixel + HOG
top-1	82.5 (-1.1)	**83.6**	82.3 (-1.3)

Table 13. **Multi-tasking.** Simply combining two targets with two separate linear prediction heads results in a drop, suggesting conflict in the objectives. The default entry is marked as gray.

Wei, Chen, et al. "Masked feature prediction for self-supervised visual pre-training." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022

CLIP: leverage large scale image-text pairs

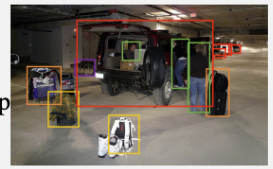




Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I.. "Learning transferable visual models from natural language supervision." ICML, 2021.

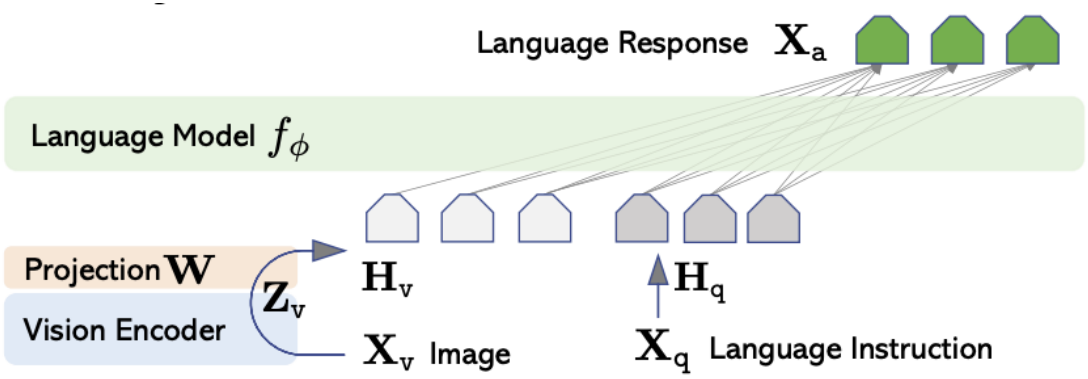
Multi-modality pretraining

Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.



Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>
Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>



Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

User: What is unusual about this image?
LLaVA: The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
[*Start a new conversation, and clear the history*]
User: What's happening in the scene?
LLaVA: The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.

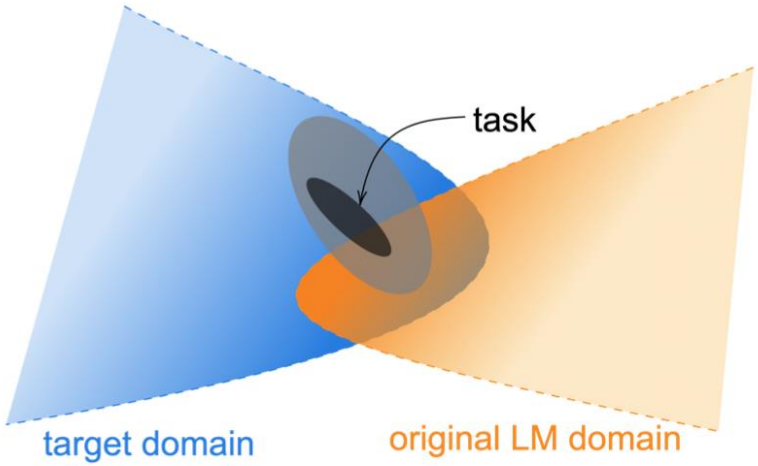
Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing systems* 36 (2024).

Don’t stop training

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

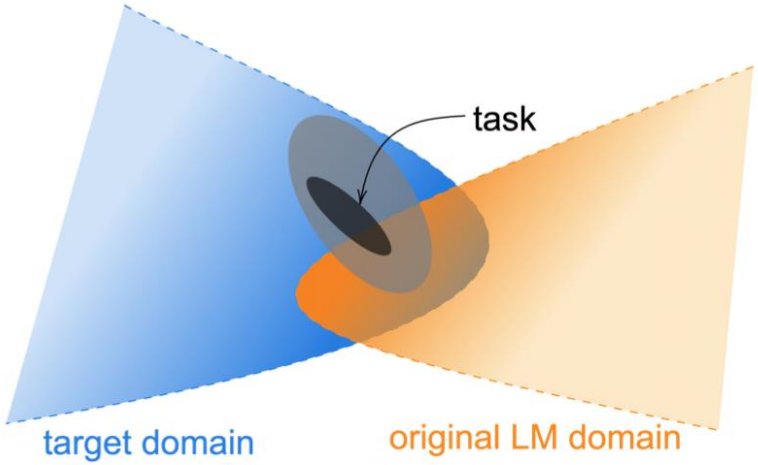
Don’t stop training



Domain	Task	RoBERTa	Additional Pretraining Phases		
			DAPT	TAPT	DAPT + TAPT
BioMed	ChemProt	$81.9_{1.0}$	$84.2_{0.2}$	$82.6_{0.4}$	$\mathbf{84.4}_{0.4}$
	†RCT	$87.2_{0.1}$	$87.6_{0.1}$	$87.7_{0.1}$	$\mathbf{87.8}_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	$75.4_{2.5}$	$67.4_{1.8}$	$\mathbf{75.6}_{3.8}$
	SciERC	$77.3_{1.9}$	$80.8_{1.5}$	$79.3_{1.5}$	$\mathbf{81.3}_{1.8}$
News	HyperPartisan	$86.6_{0.9}$	$88.2_{5.9}$	$\mathbf{90.4}_{5.2}$	$90.0_{6.6}$
	†AGNews	$93.9_{0.2}$	$93.9_{0.2}$	$94.5_{0.1}$	$\mathbf{94.6}_{0.1}$
Reviews	†Helpfulness	$65.1_{3.4}$	$66.5_{1.4}$	$68.5_{1.9}$	$\mathbf{68.7}_{1.8}$
	†IMDB	$95.0_{0.2}$	$95.4_{0.1}$	$95.5_{0.1}$	$\mathbf{95.6}_{0.1}$

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv preprint arXiv:2004.10964*.

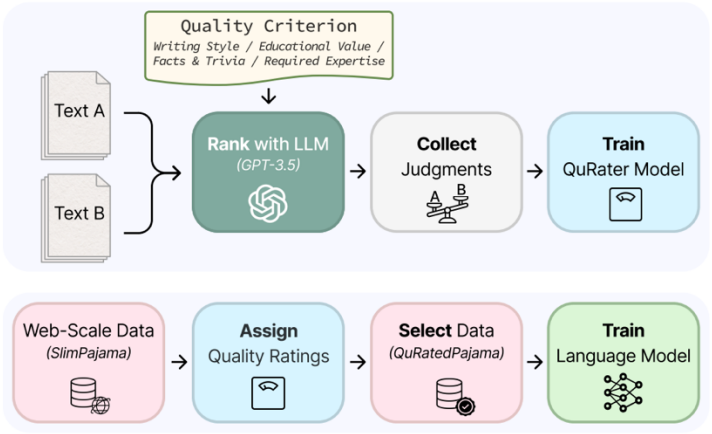
Don’t stop training



Domain	Task	RoBERTa	Additional Pretraining Phases		
			DAPT	TAPT	DAPT + TAPT
BioMed	ChemProt	$81.9_{1.0}$	$84.2_{0.2}$	$82.6_{0.4}$	$\mathbf{84.4}_{0.4}$
	†RCT	$87.2_{0.1}$	$87.6_{0.1}$	$87.7_{0.1}$	$\mathbf{87.8}_{0.1}$
CS	ACL-ARC	$63.0_{5.8}$	$75.4_{2.5}$	$67.4_{1.8}$	$\mathbf{75.6}_{3.8}$
	SciERC	$77.3_{1.9}$	$80.8_{1.5}$	$79.3_{1.5}$	$\mathbf{81.3}_{1.8}$
News	HyperPartisan	$86.6_{0.9}$	$88.2_{5.9}$	$\mathbf{90.4}_{5.2}$	$90.0_{6.6}$
	†AGNews	$93.9_{0.2}$	$93.9_{0.2}$	$94.5_{0.1}$	$\mathbf{94.6}_{0.1}$
Reviews	†Helpfulness	$65.1_{3.4}$	$66.5_{1.4}$	$68.5_{1.9}$	$\mathbf{68.7}_{1.8}$
	†IMDB	$95.0_{0.2}$	$95.4_{0.1}$	$95.5_{0.1}$	$\mathbf{95.6}_{0.1}$

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv preprint arXiv:2004.10964*.

Clean up your data



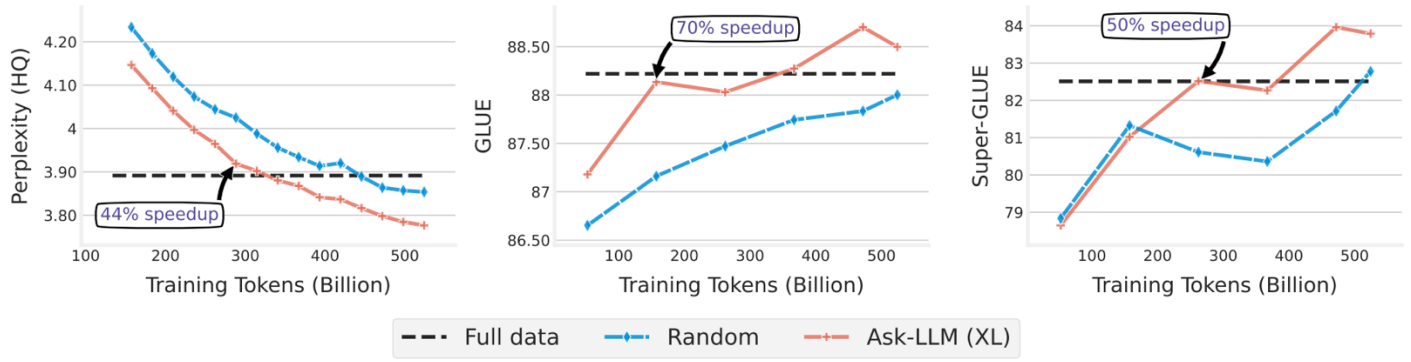
Ask-LLM prompt

###
This is a pretraining .... datapoint.
###

Does the previous paragraph demarcated within ### and ### contain informative signal for pre-training a large-language model? An informative datapoint should be well-formatted, contain some usable knowledge of the world, and strictly NOT have any harmful, racist, sexist, etc. content.

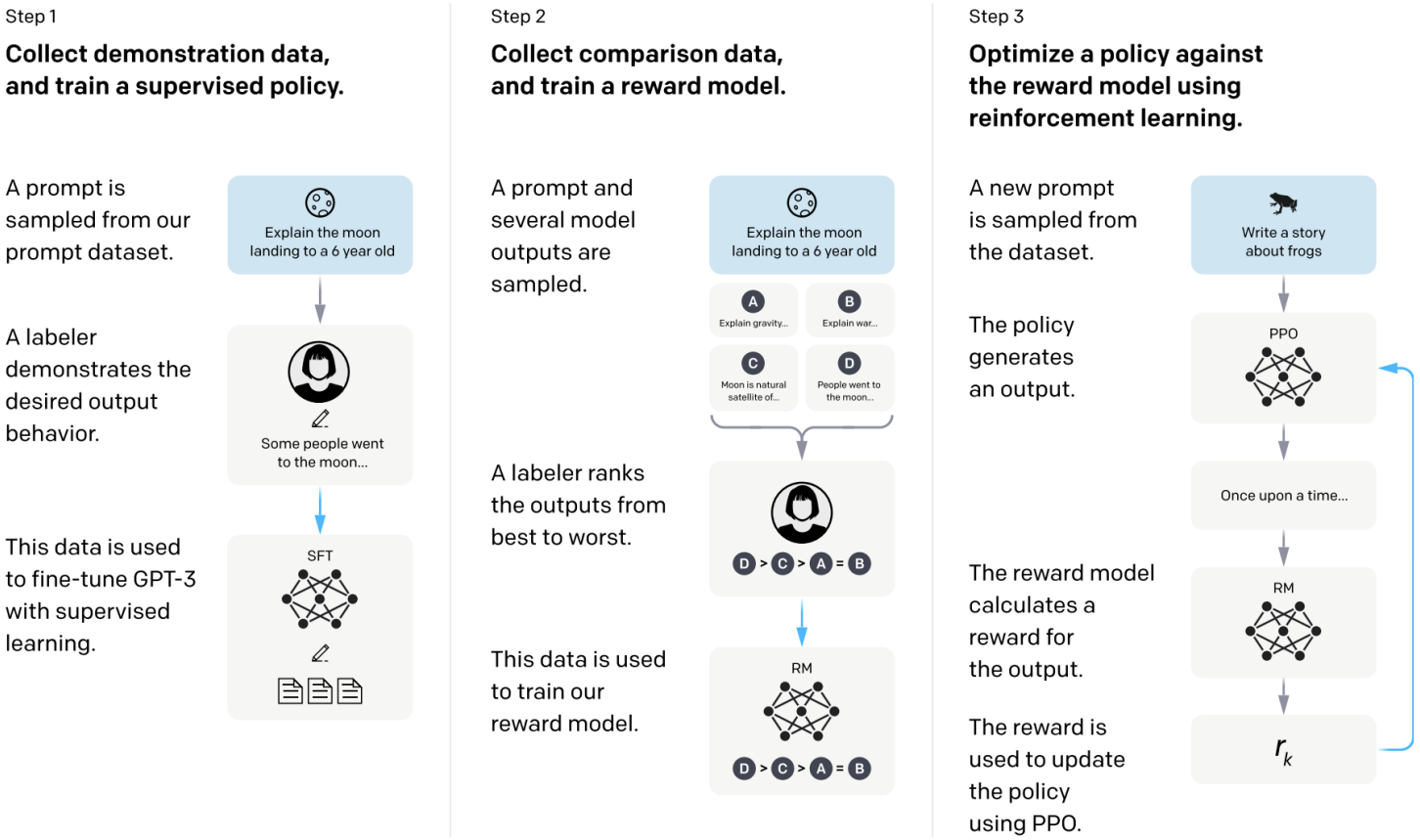
OPTIONS:
- yes
- no

Sampling score = P("yes" | prompt)



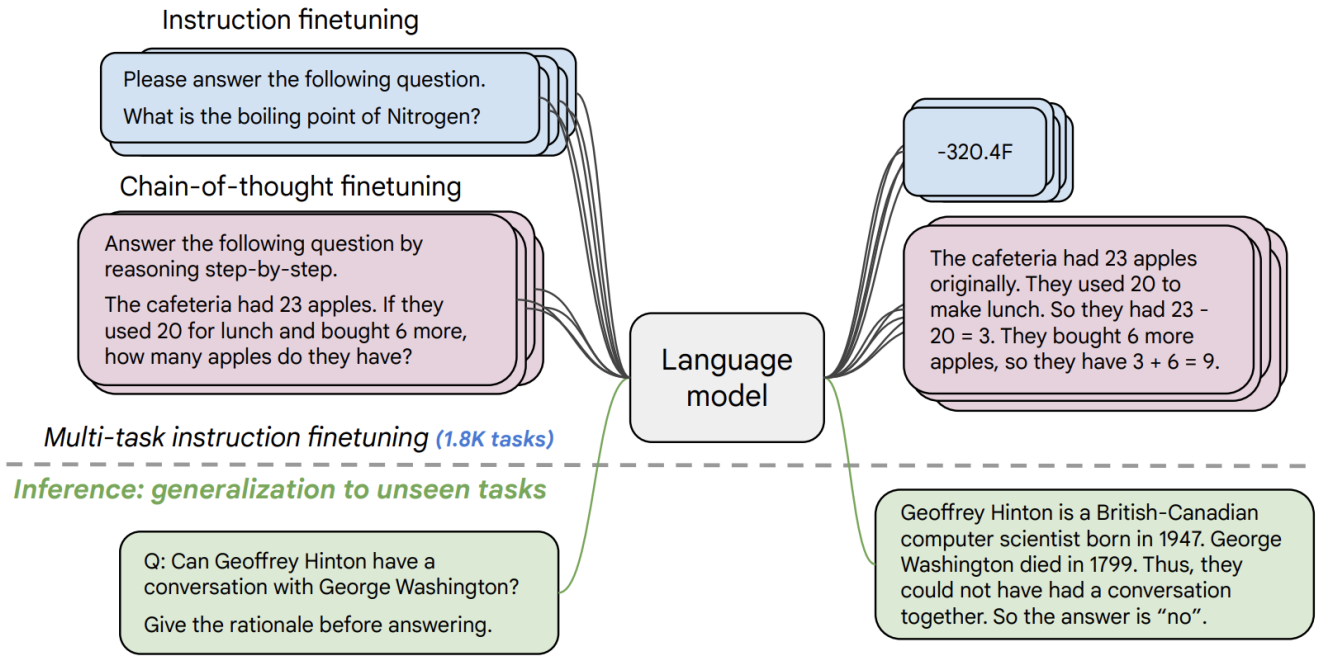
Wettig, Alexander, et al. "QuRating: Selecting High-Quality Data for Training Language Models." *arXiv preprint arXiv:2402.09739* (2024).
Sachdeva, Noveen, et al. "How to Train Data-Efficient LLMs." *arXiv preprint arXiv:2402.09668* (2024).

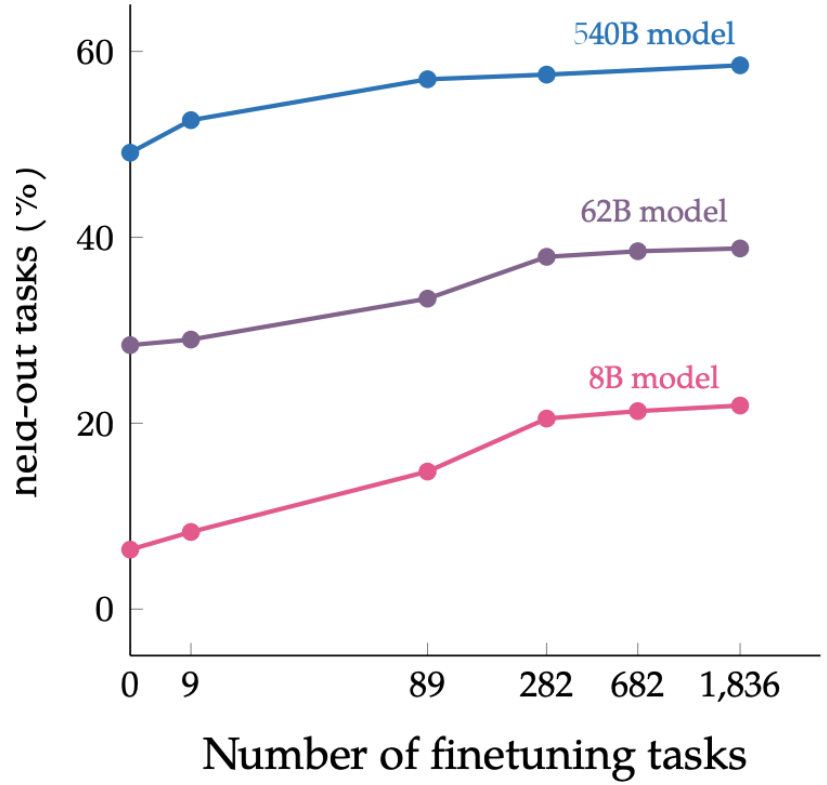
Alignment foundation models with human preference



Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *Advances in neural information processing systems* 35 (2022): 27730-27744.

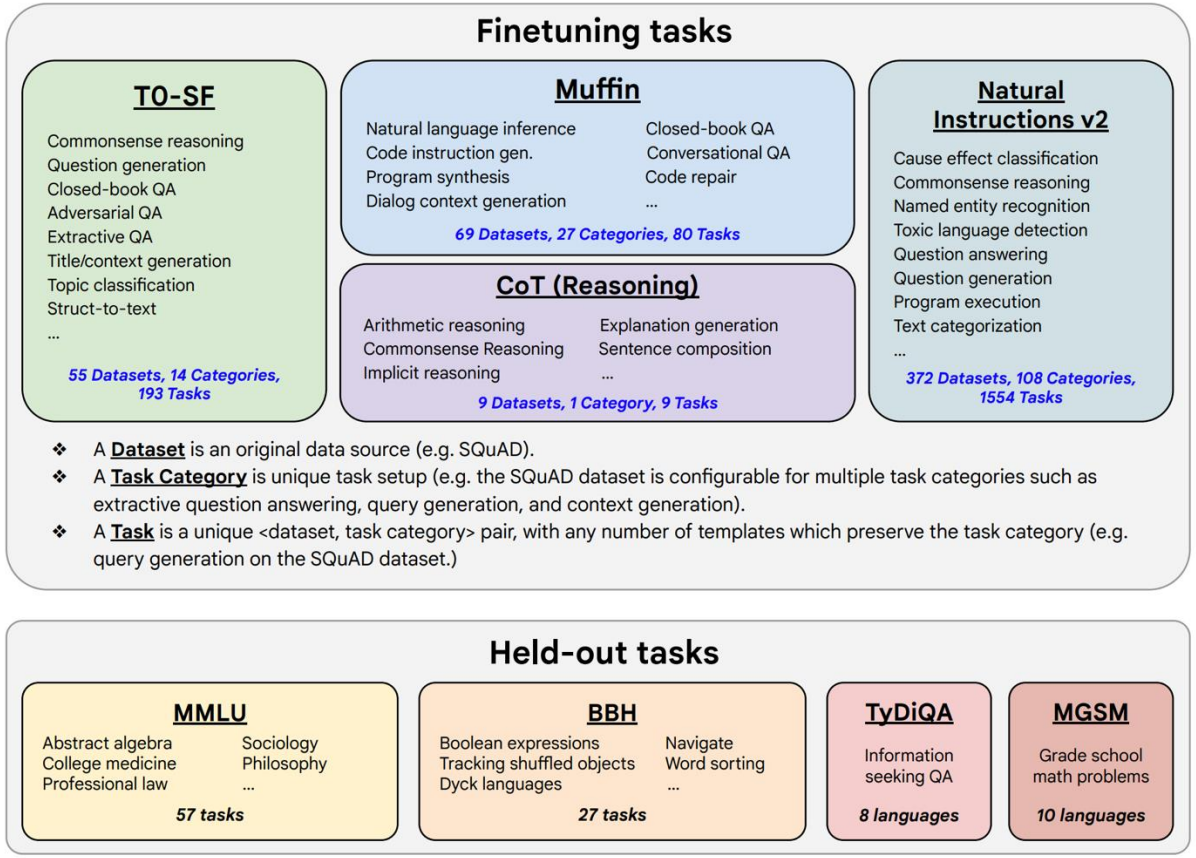
Alignment foundation models with human preference

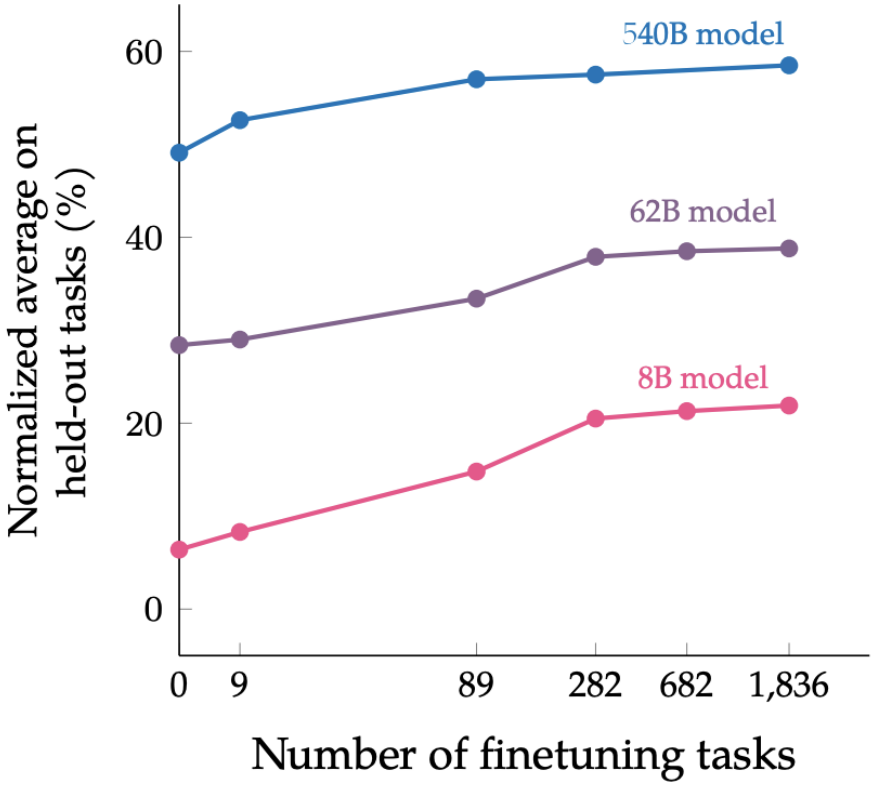




Chung, Hyung Won, et al. "Scaling instruction-finetuned language models." *arXiv preprint arXiv:2210.11416* (2022).

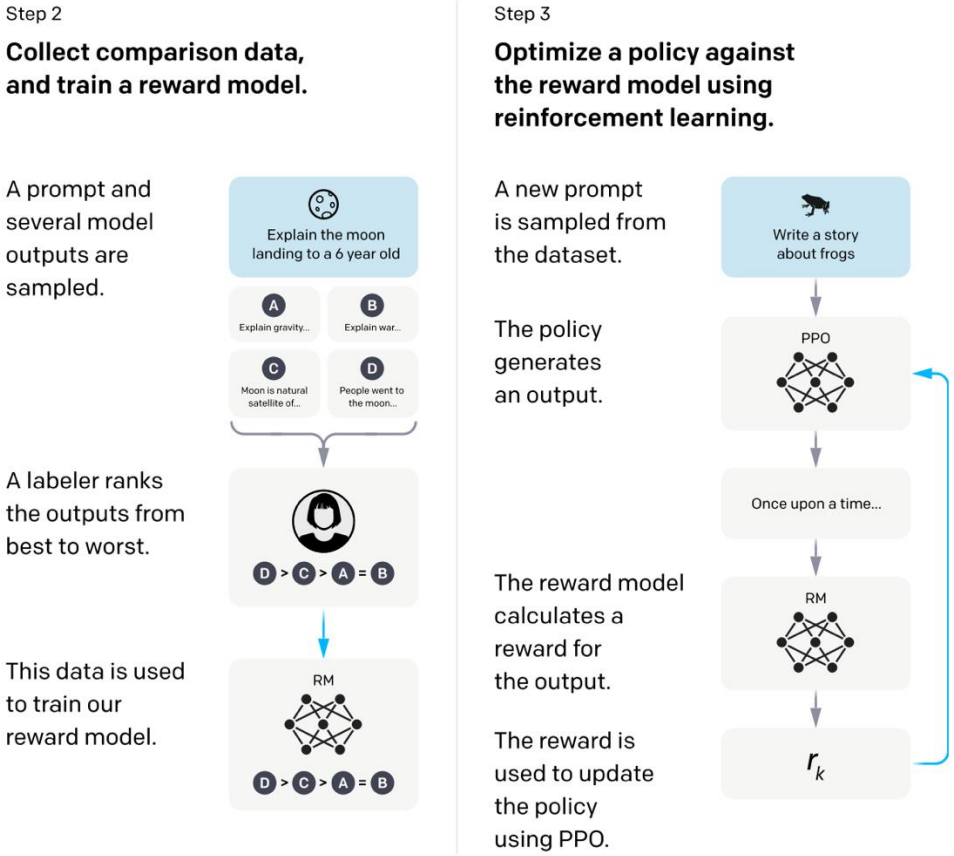
Alignment foundation models with human preference

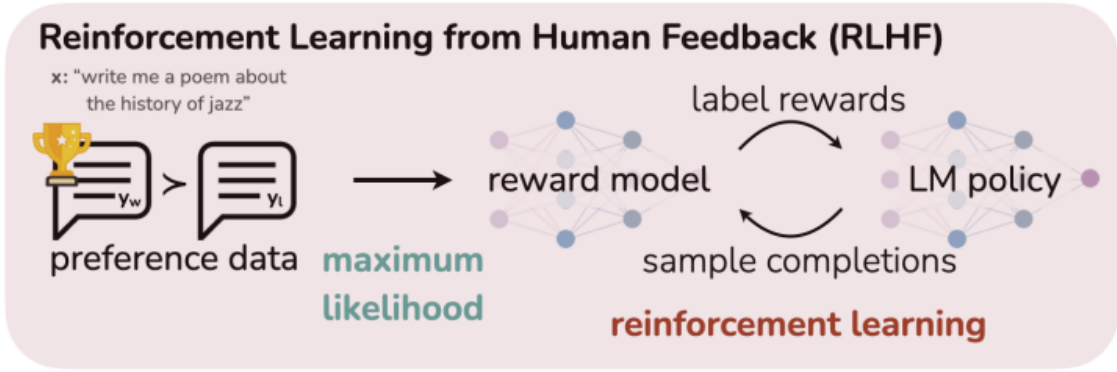




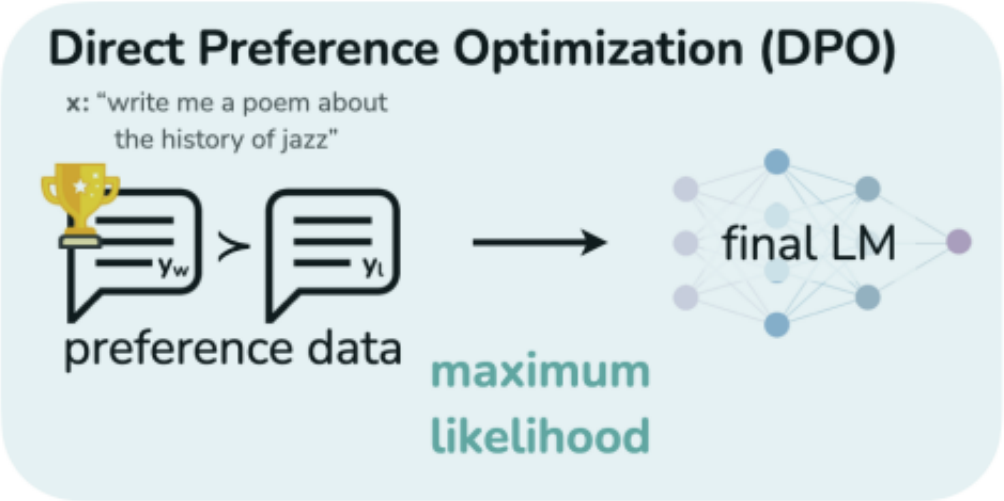
Chung, Hyung Won, et al. "Scaling instruction-finetuned language models." *arXiv preprint arXiv:2210.11416* (2022).

Direct preference optimization (DPO)





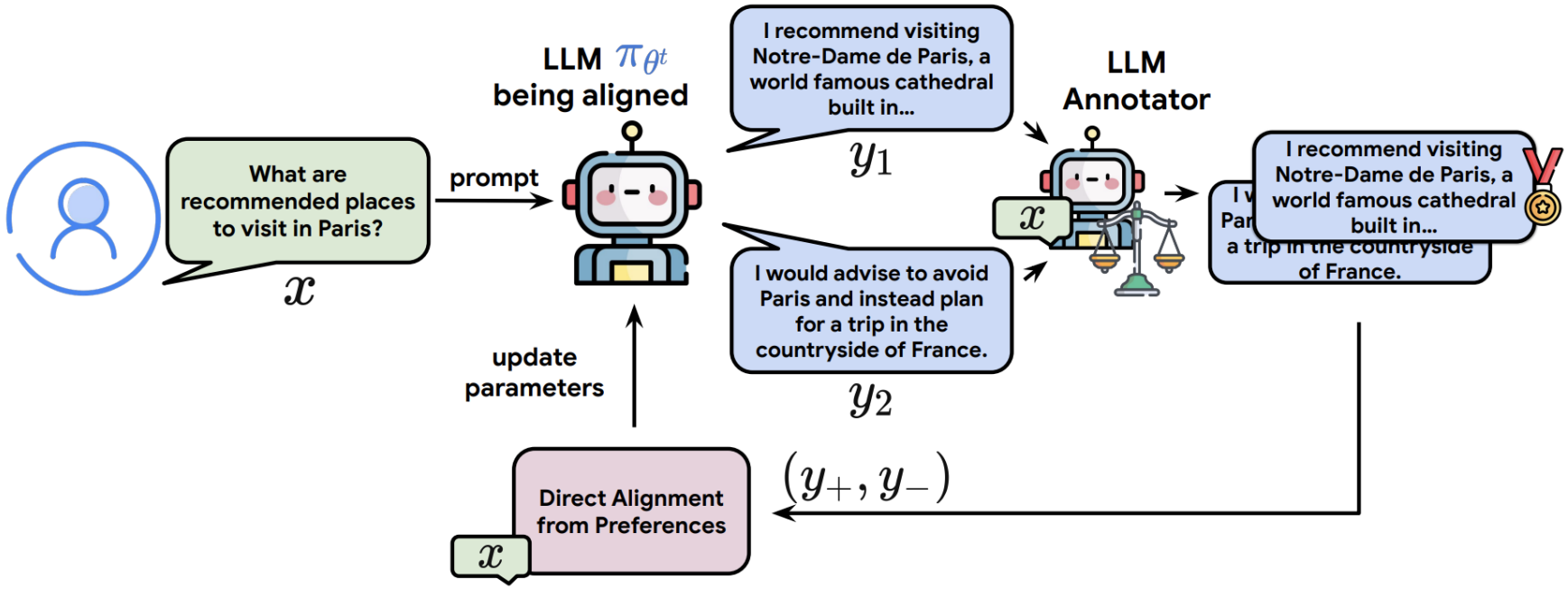
$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)}\left[r_\phi(x,y)\right] - \beta\mathbb{D}_{\mathrm{KL}}\left[\pi_\theta(y \mid x) \,||\, \pi_{\mathrm{ref}}(y \mid x)\right]$$



$$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta; \pi_{\mathrm{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\mathrm{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\mathrm{ref}}(y_l \mid x)}\right)\right]$$

Rafailov, Rafael, et al. "Direct preference optimization: Your language model is secretly a reward model." *Advances in Neural Information Processing Systems* 36 (2024).

Direct preference optimization from Online AI Feedback

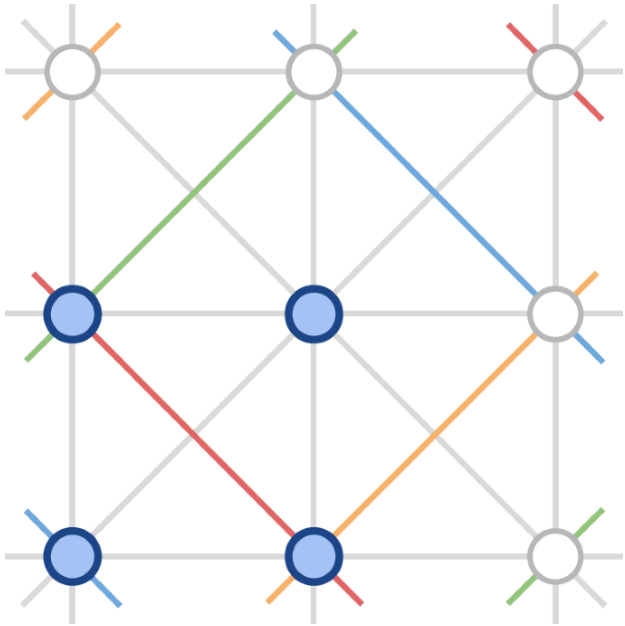


Guo, Shangmin, et al. "Direct language model alignment from online ai feedback." *arXiv preprint arXiv:2402.04792* (2024).

What foundation models can do

	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (incl. benchmark-specific tuning)
MMLU [49] Multiple-choice questions in 57 subjects (professional & academic)	**86.4%** 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM [50]	75.2% 5-shot Flan-PaLM [51]
HellaSwag [52] Commonsense reasoning around everyday events	**95.3%** 10-shot	85.5% 10-shot	84.2% LLaMA (validation set) [28]	85.6 ALUM [53]
AI2 Reasoning Challenge (ARC) [54] Grade-school multiple choice science questions. Challenge-set.	**96.3%** 25-shot	85.2% 25-shot	85.2% 8-shot PaLM [55]	86.5% ST-MOE [18]
WinoGrande [56] Commonsense reasoning around pronoun resolution	**87.5%** 5-shot	81.6% 5-shot	85.1% 5-shot PaLM [3]	85.1% 5-shot PaLM [3]
HumanEval [43] Python coding tasks	**67.0%** 0-shot	48.1% 0-shot	26.2% 0-shot PaLM [3]	65.8% CodeT + GPT-3.5 [57]
DROP [58] (F1 score) Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot	70.8 1-shot PaLM [3]	**88.4** QDGAT [59]
GSM-8K [60] Grade-school mathematics questions	**92.0%*** 5-shot chain-of-thought	57.1% 5-shot	58.8% 8-shot Minerva [61]	87.3% Chinchilla + SFT+ORM-RL, ORM reranking [62]

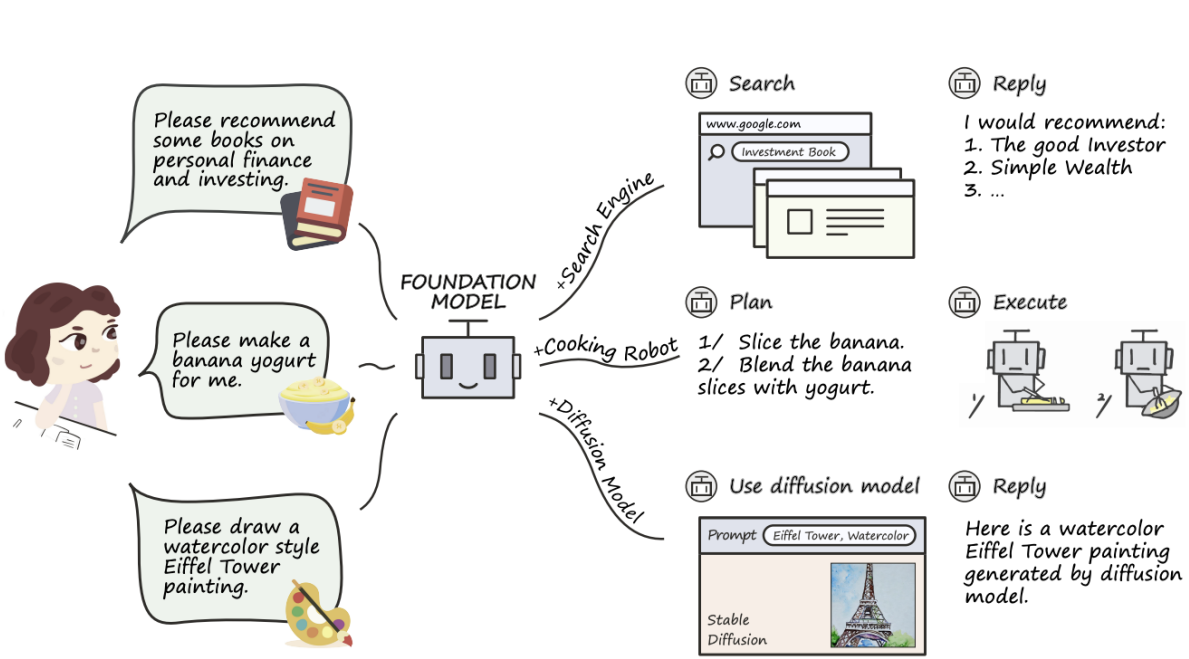
Achiam, Josh, et al. "Gpt-4 technical report." *arXiv preprint arXiv:2303.08774* (2023).

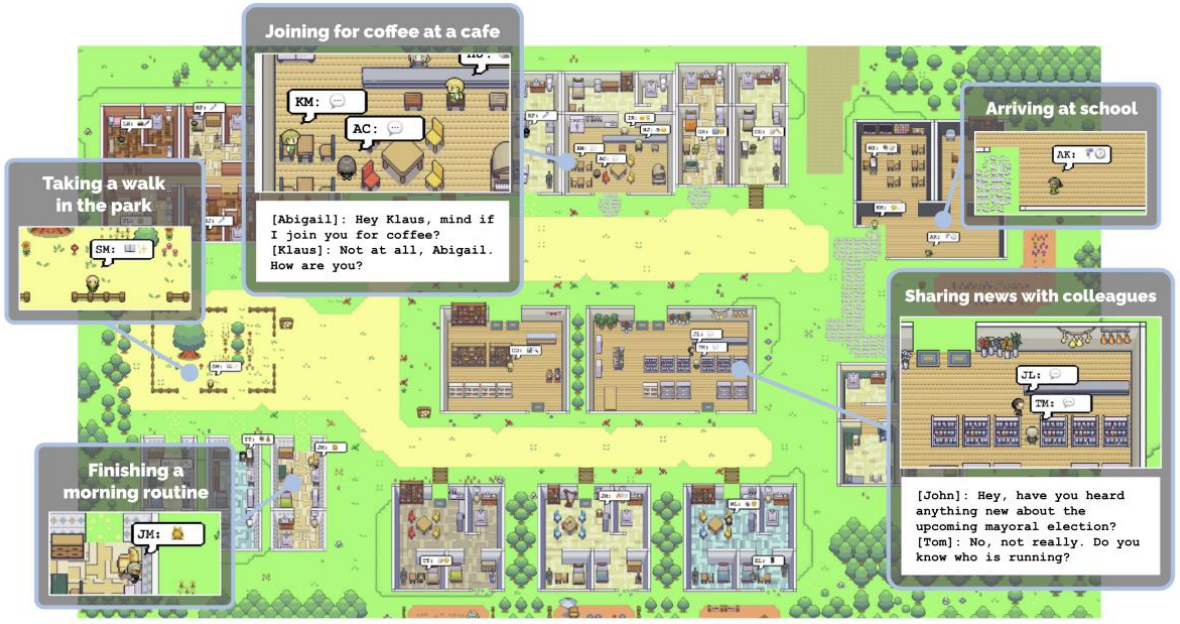


n	3	4	5	6	7	8
Best known	9	20	45	112	236	496
FunSearch	9	20	45	112	236	512

Romera-Paredes, Bernardino, et al. "Mathematical discoveries from program search with large language models." *Nature* 625.7995 (2024): 468-475.

Planning and execution





Qin, Yujia, et al. "Tool learning with foundation models." *arXiv preprint arXiv:2304.08354* (2023

Park, Joon Sung, et al. "Generative agents: Interactive simulacra of human behavior." *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 2023.

Physical world simulator

Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

https://openai.com/research/video-generation-models-as-world-simulators

Open questions for foundation models

1. **Long-context Understanding**:
   How can LLMs effectively utilize long-context information to improve their understanding and generation capabilities?
2. **Hallucination**:
   How can we mitigate hallucination, where LLMs generate text that is not supported by the input or is factually incorrect?
3. **Memory Augmented Models**:
   How can memory-augmented architectures like RAG be further developed to improve the ability of LLMs to store and retrieve information over long sequences?
4. **Consistency and Coherence**:
   How can LLMs be trained to maintain consistency and coherence over long sequences of text, especially in tasks requiring multi-turn dialogue or narrative generation?
5. **Evaluation Metrics**:
   What are the most appropriate metrics for assessing the performance of LLMs in tasks involving memory, hallucination, and long-context understanding?
6. **Scaling-up, Continual Learning, Bias and Fairness, Interpretability, Safety…**

Key takeaways

Principle 3 (the scaling law): AI methods that leverage computation are ultimately the most effective way of improvements (from "The bitter lesson" by Rich Sutton)

Principle 4 (the data law): Data is the ultimate way of regularization

1. Transformer and its improvements
2. Different kinds of SSL methods
3. Application of foundation models

Reading materials

- Foundation model papers
- Code: transformers; diffusers
- Leaderboard: LMSYS Chatbot Arena Leaderboard; Open LLM Leaderboard
- How to Train Really Large Models on Many GPUs
- Rotary Position Embedding