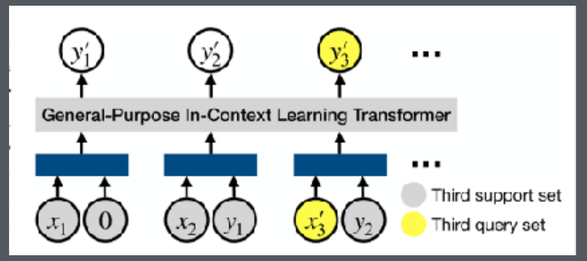
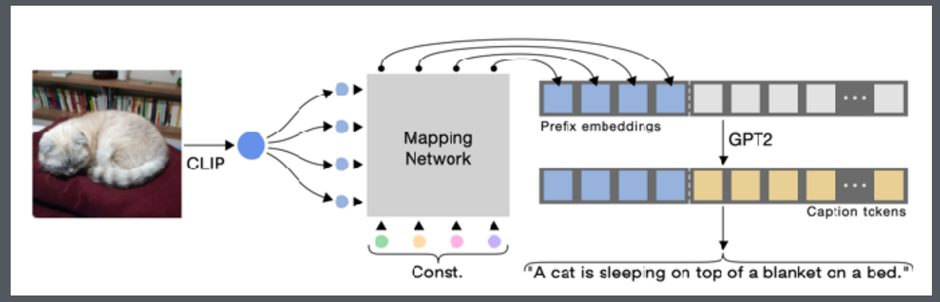
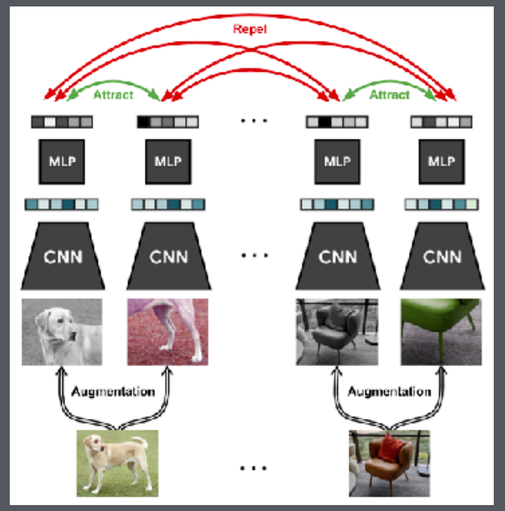
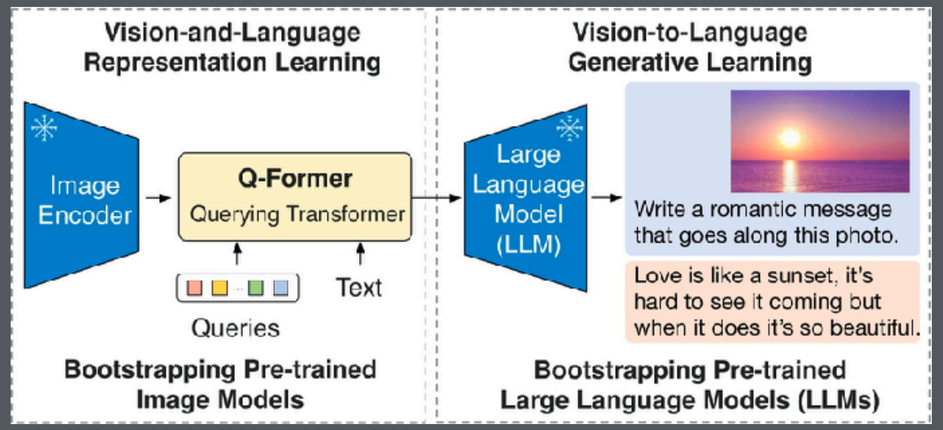
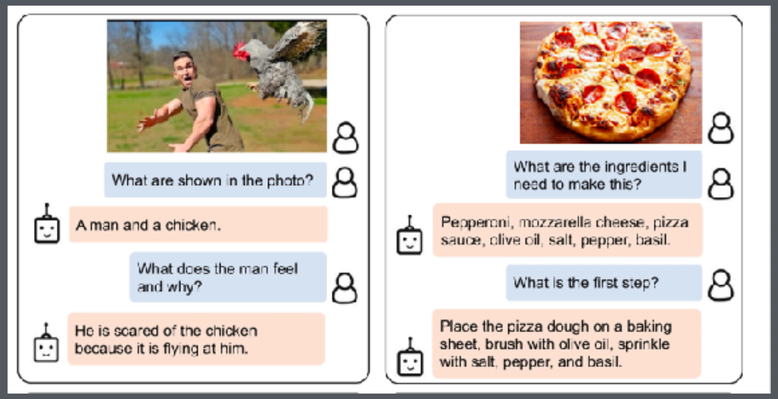
Self-supervised and vision-language learning











@ DEEP LEARNING 2

YUKI M. ASANO
VIS LAB & QUVA LAB

Orga:

I've re-recorded the first lecture (hopefully somethings are more clear now)

Everyone should know which team they are on, and who their TA is

Hope you enjoy the work! The undirected, creative part of this project is hopefully the most fun part of this course

Don't sit alone – some discussion parts in today's lecture where you work with your neighbor🤗

Caveat:

As with the last lecture, this lecture only presents a slice of the research that is ongoing as we speak.

I will not try to be comprehensive here, but instead give insights into selected topics and papers.

Recap

Single-modal self-supervised pretraining methods (MAE, SimCLR, GPT)

Multi-modal pretraining (CLIP, ALIGN, CoCa)

Beyond contrastive (BLIP, ClipCap)

GPT + X (Socratic, TeachText)

Longer context (FROMAGe, Frozen, Flamingo)

Tasks (VQA, VisDial)

GPT-3: "Language models are few-shot learners"

more on this later

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

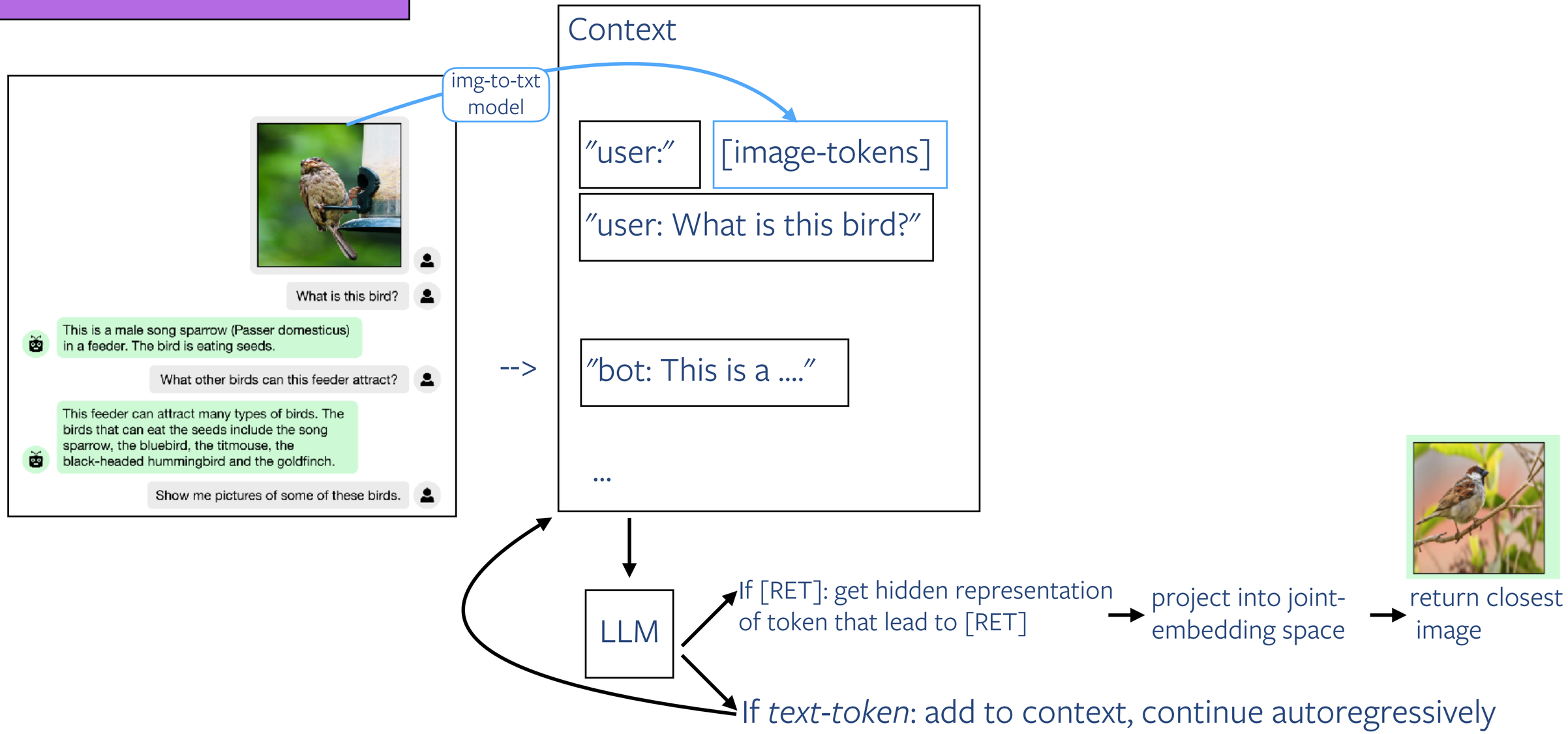
One emergent capability of large language models is *in-context learning*.

Here, the "task" is defined within the language model's context, and the model *picks up the task* and *solves it* for the given sample both during a single forward pass

Brown et al. Language models are few-shot learners. NeurIPS 2020.

Recap

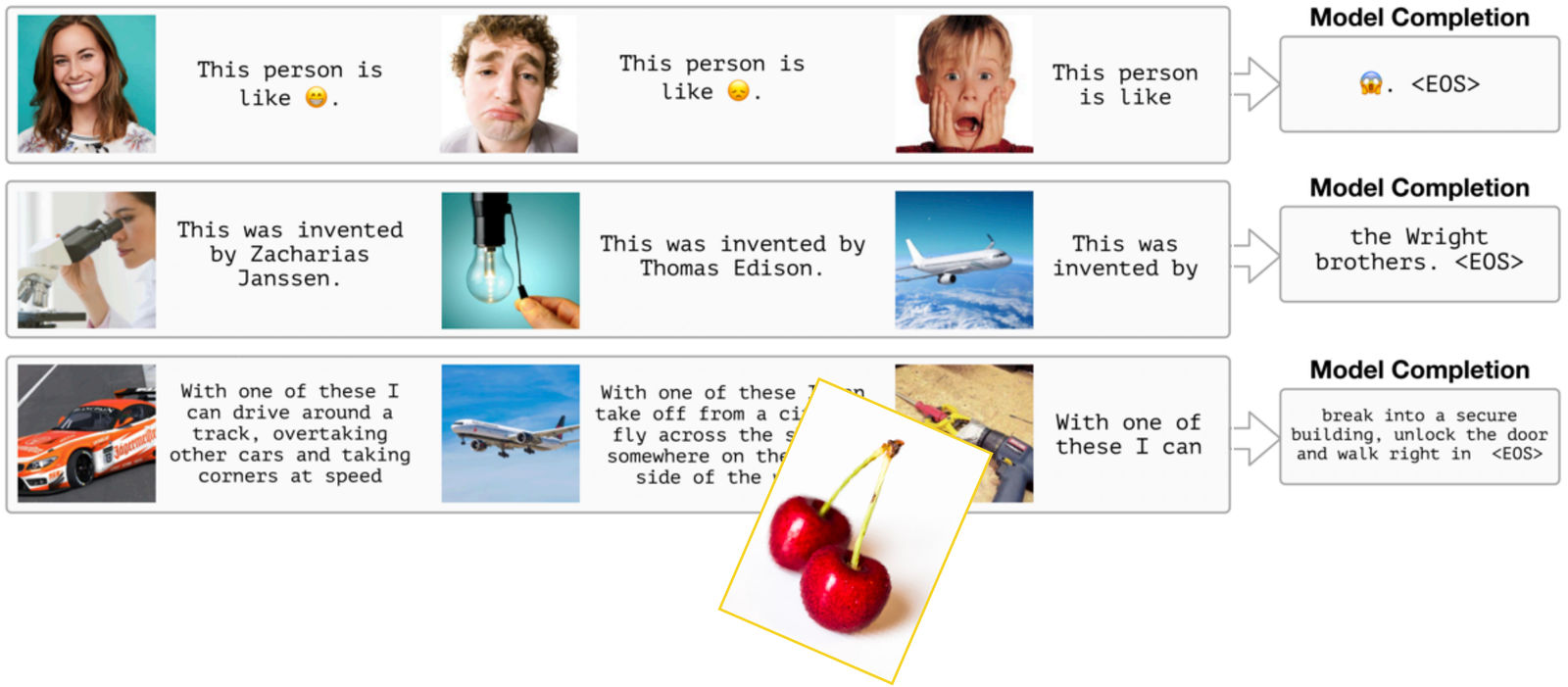
Visual Language Models: *everything is a language-token*



Koh et al. Grounding Language Models to Images for Multimodal Generation. 2023

Recap

Frozen: Multimodal Few-Shot Learning with Frozen Language Models



Method:

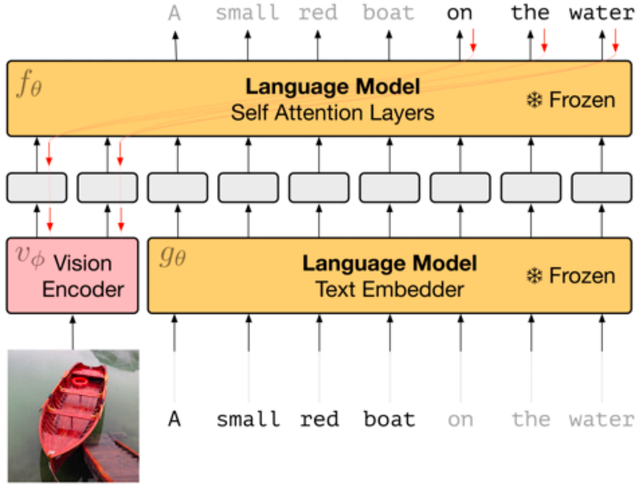
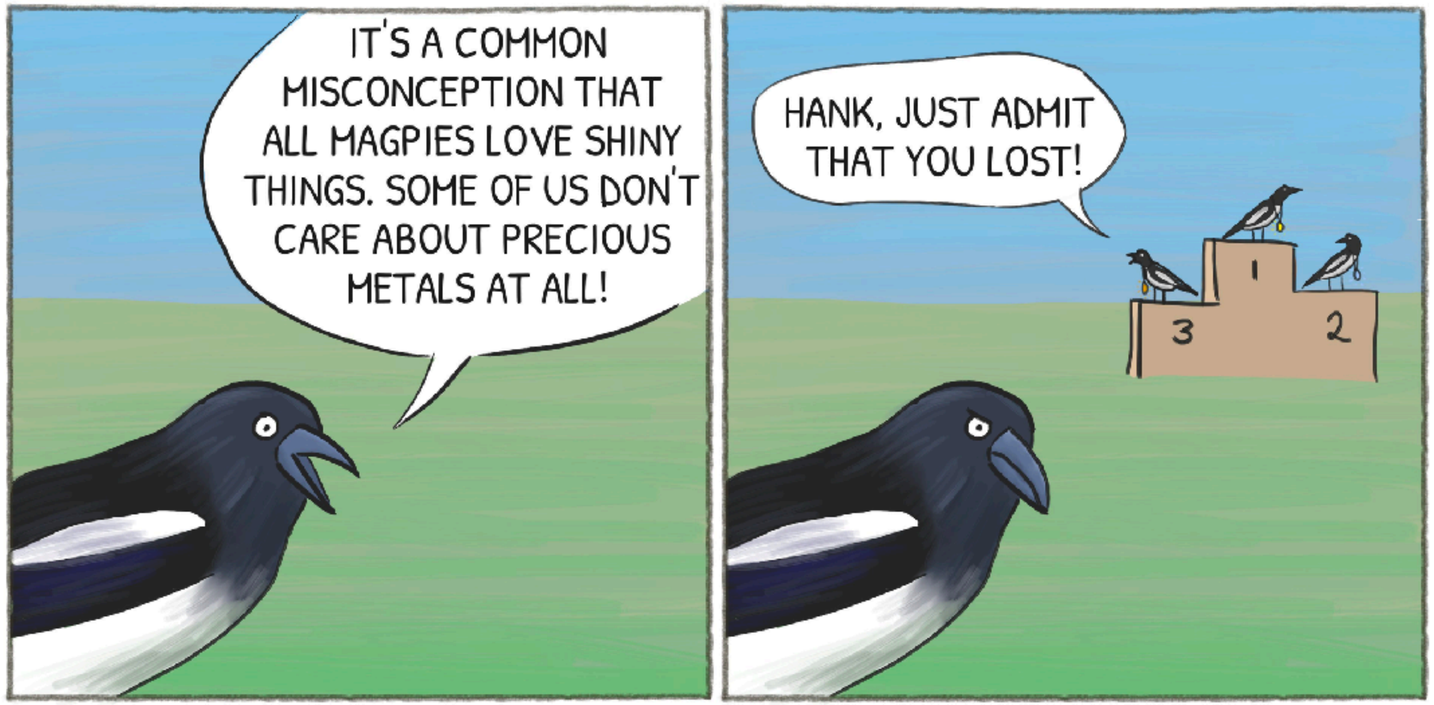


Figure 2: Gradients through a frozen language model's self attention layers are used to train the vision encoder.

Tsimpoukelli et al. Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021.

In-context learning

towards more useful systems



context matters

Vision-language in-context learning (ICL)

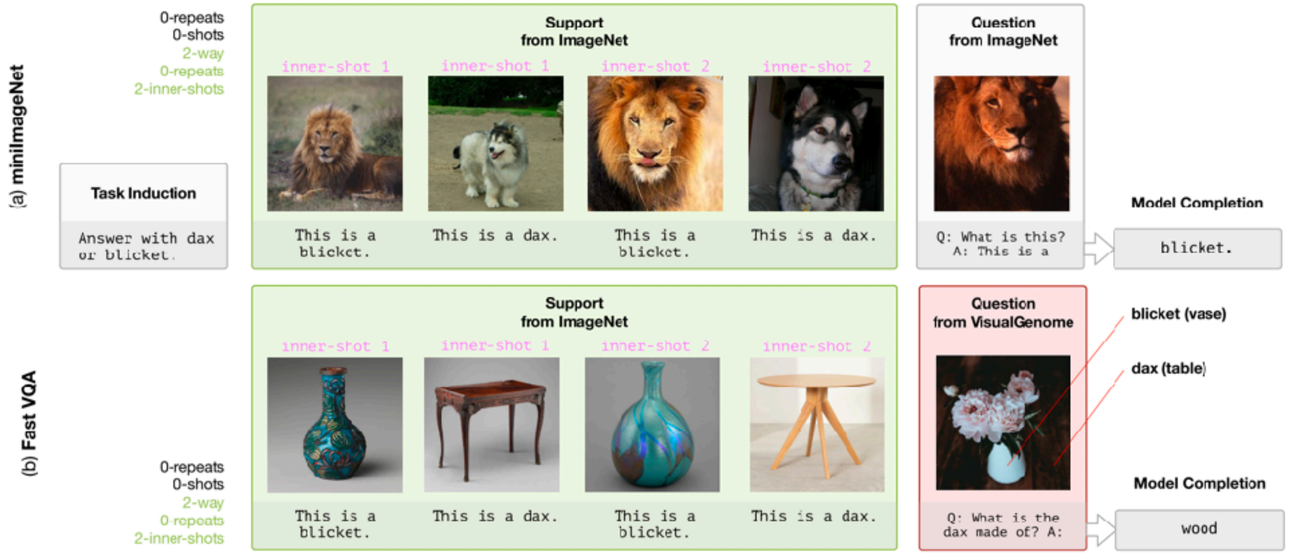


Figure 4: Examples of (a) the Open-Ended miniImageNet evaluation (b) the Fast VQA evaluation.

- Here, ICL is short for something like "open-ended vision-language few-shot evaluation"
- Open-ended: it needs to infer what it's supposed to do & what the answer style is.
- Vision-language: it needs to process both the image & the text info
- Few-shot: few-shot samples "support set" are provided as input, along with the test sample
- "fast-binding": text & image are associated within the single forward pass

Tsimpoukelli et al. Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021.

Vision-language in-context learning (ICL) in Frozen

Task Induction	✗	✓	✓	✓	✓	✓	✓
Inner Shots	1	1	3	5	1	1	1
Repeats	0	0	0	0	1	3	5
Frozen	29.0	53.4	57.9	58.9	51.1	57.7	58.5
Frozen (Real-Name)	1.7	33.7	66	66	63	65	63.7

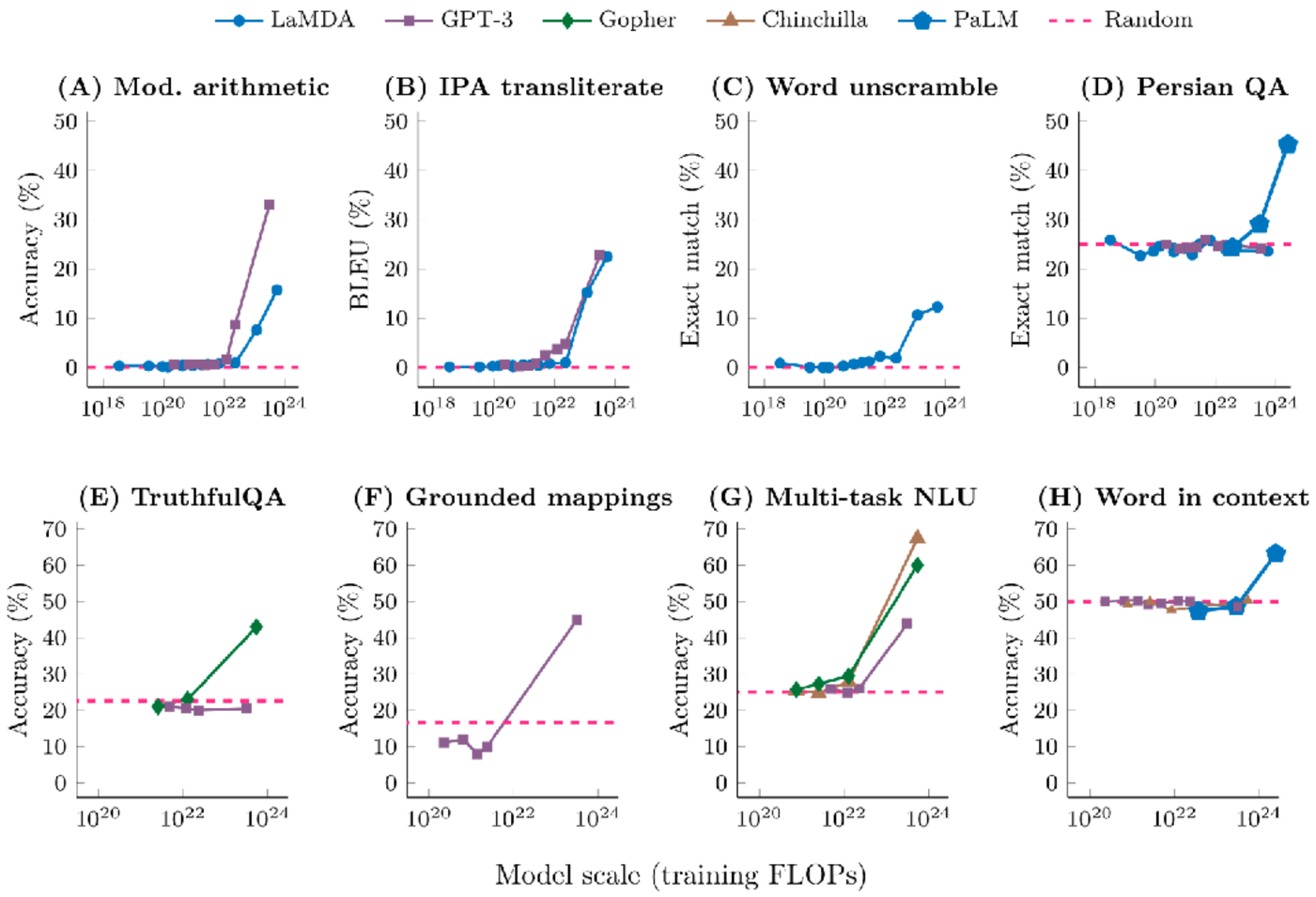
Task Induction	✗	✓	✓	✓	✓	✓	✓
Inner Shots	1	1	3	5	1	1	1
Repeats	0	0	0	0	1	3	5
Frozen	18.0	20.2	22.3	21.3	21.4	21.6	20.9
Frozen (Real-Name)	0.9	14.5	34.7	33.8	33.8	33.3	32.8

For ImageNet few-shot evaluation, using ″dax″ and ″blicket″ works better than using the real names of the animals/objects.

Not for their version of VQA: [but note the non-zero performance of blind models]

	Fast-VQA				Real-Fast-VQA			
Inner Shots	0	1	3	5	0	1	3	5
Frozen	1.6	2.8	7.0	7.9	3.7	7.8	10.1	10.5
Frozen train-blind	0.7	0.3	1.3	0.4	1.9	2.3	3.7	3.7

In-context learning emerges in LLMs with scale



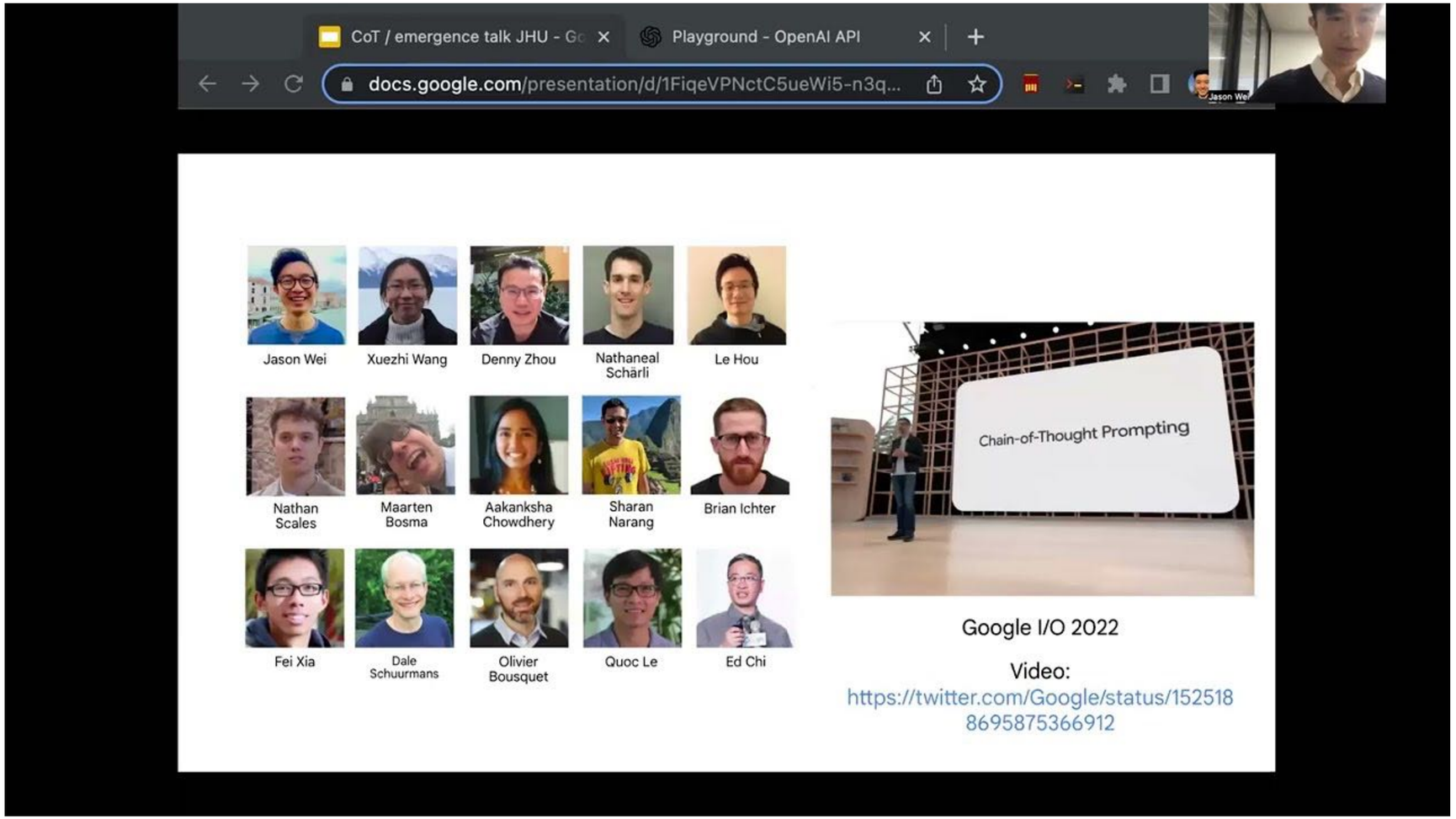
- ″An ability is emergent if it is not present in smaller models but is present in larger models.″
- ″Emergence is when quantitative changes in a system result in qualitative changes in behaviour″
- sort of a 0 to 1 change

- Why/how does it emerge?
  - Some problems require memorising
  - Also note: evaluation metrics only evaluate strict string matching (e.g. ″A panda″, vs. ″A panda bear″)
  - But generally: ???

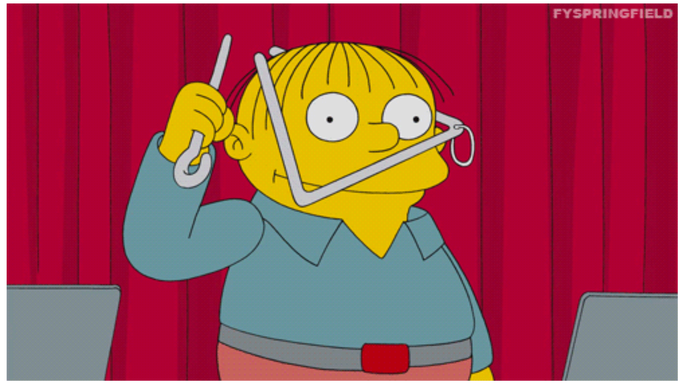
Wei et al. Emergent Abilities of Large Language Models. TMLR 2022

recommended read

Emergence and reasoning in large language models - Jason Wei



Careful about AI hype + news



- Easy to complain about this sort of stuff (see also DL1)
- But why does this keep happening?

My two cents:

- Deep learning in industry is increasingly a marketing battle
- Moreover, companies do not have much incentive to really document/ analyse their training data

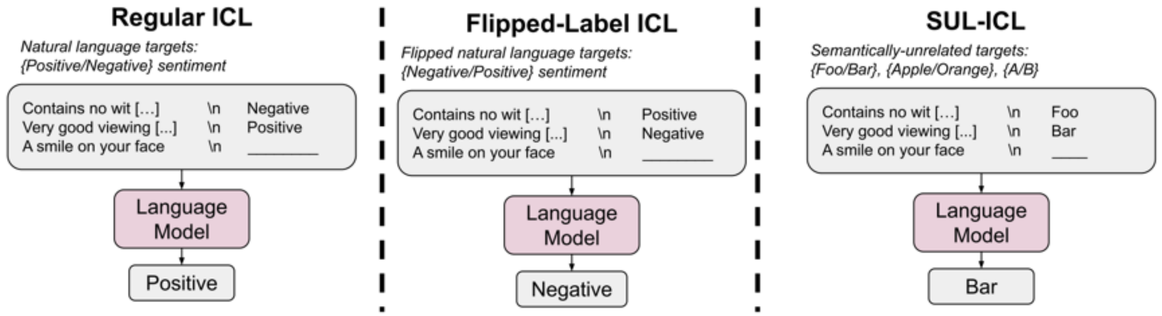
more on this: https://twitter.com/mmitchell_ai/status/1648029417497853953

With that caveat in mind, in-context learning in vision-language models is still impressive

- Especially because: models like Frozen, Flamingo, FROMAGe weren't explicitly trained for in-context learning.
- While Flamingo and CM3 were trained with websites,
  - in-context like samples are frequent
- So these VL models obtain a significant (and useful part) of their ability from the language models
-

--> studying language models (and related papers) useful!

”Larger language models do in-context learning differently”



- Test abstraction & overring abilities
- ability seems to increase with scale

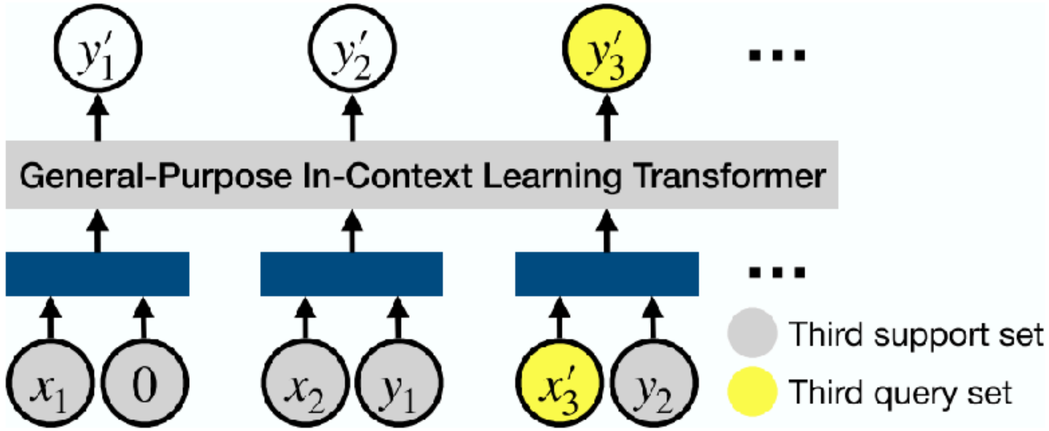
[google authors:]
For this reason, we consider all GPT-3 models to be “small” models because they all behave similarly to each other this way.🔥🔥🔥🔥

...but they also write:

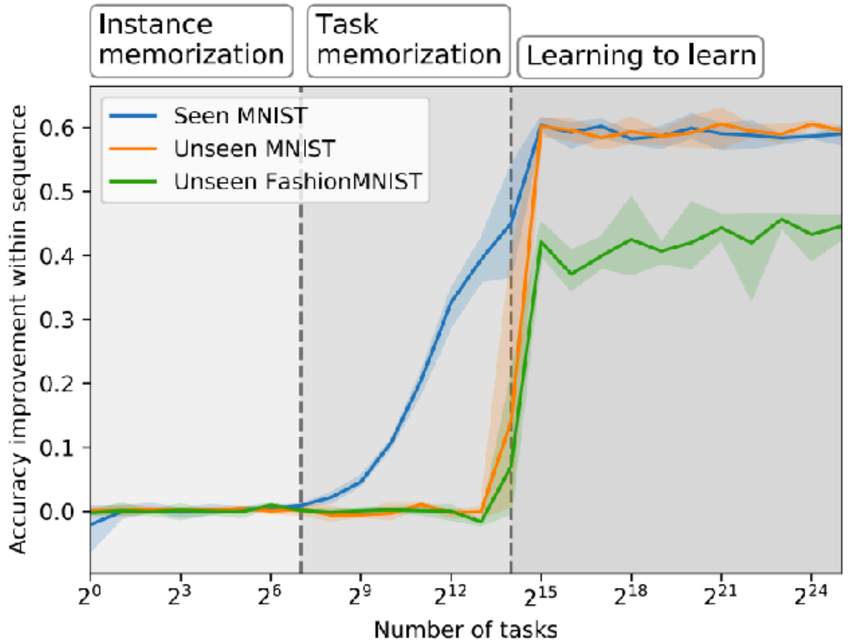
this ability to override prior knowledge with input–label mappings only appears in large models, we conclude that it is an emergent phenomena unlocked by model scaling (Wei et al., 2022b).

soo 🤷‍♂️

ICL as a learning-to-learn algorithm:

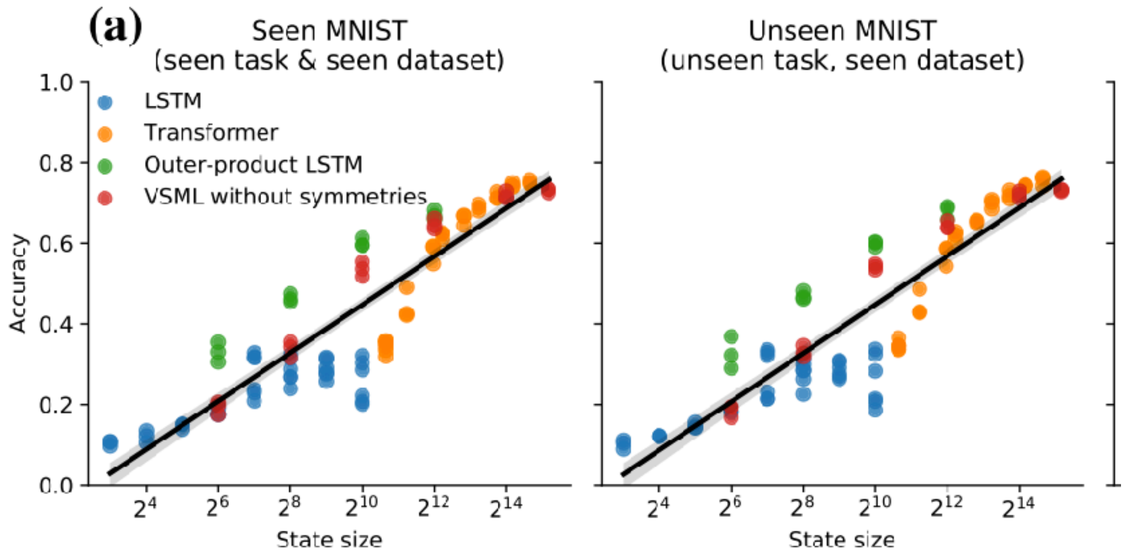


Train this simple arch on MNIST by constructing many (x1,y1; x2,y2, x3 pairs)

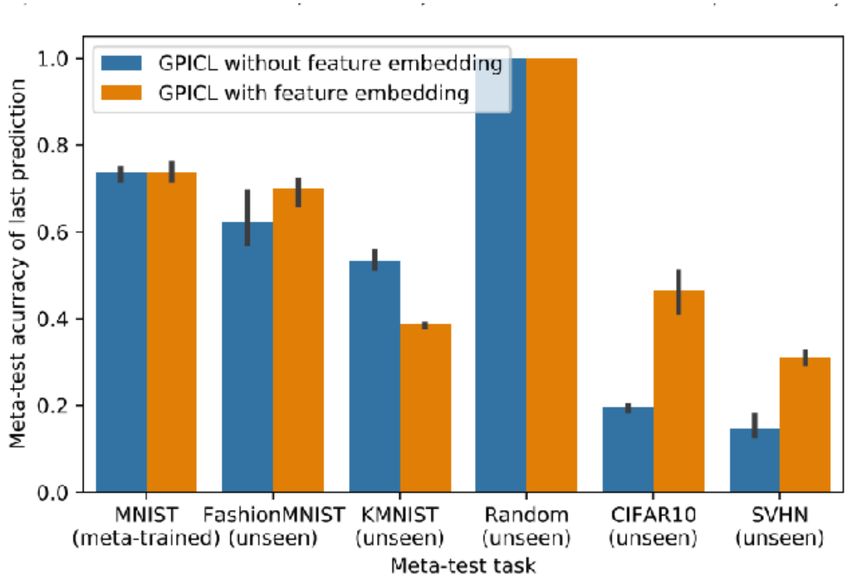


Performance on unseen data jumps when trained with sufficient diversity

What makes ICL a learning-to-learn algorithm:



Important is number of tasks & the transformer's state size (== memory), not parameter count

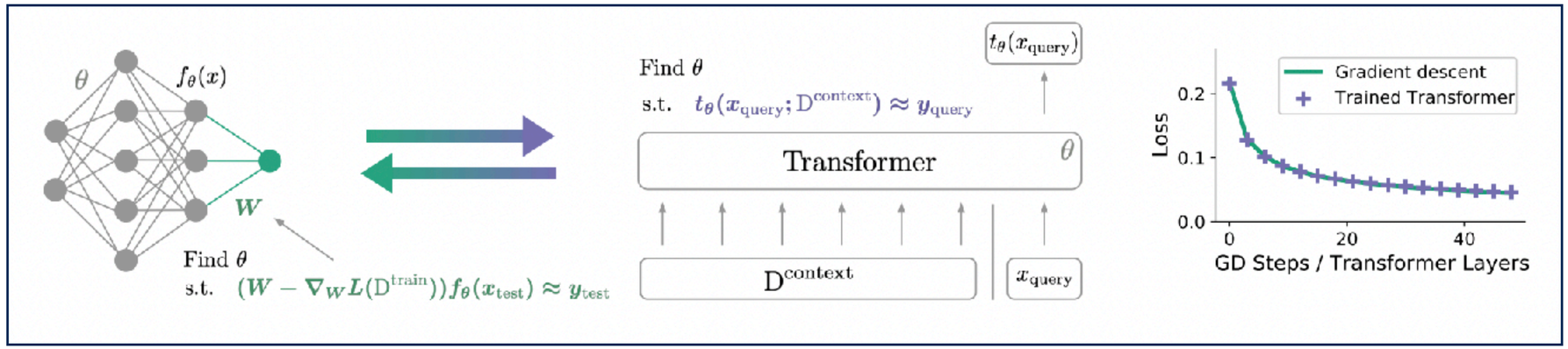


This learning-to-learn even generalises well to unseen datasets

Kirsch et al. General-Purpose In-Context Learning by Meta-Learning Transformers. 2022

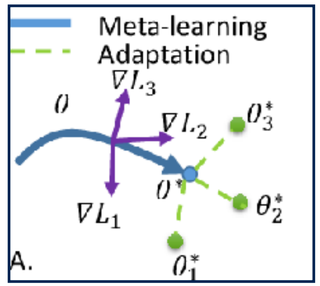
ICL implicitly implements well-known learning algorithms

[*investigations on small/toy transformers*]



Auto-regressive learning ~ Outer meta-learning loop

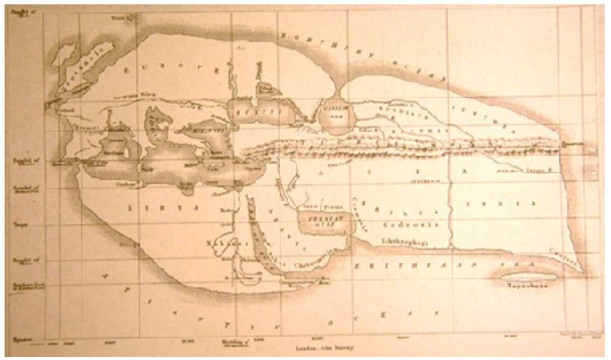
ICL ~ GD step from meta-learned model



Transformers learn to learn by gradient descent based on their context.

For linear regression case, depending on noise, ICL ~ OLS or Ridge regression

... still lots to explore!



Von Oswald et al. Transformers learn in-context by gradient descent. 2022
Akyürek et al. What learning algorithm is in-context learning? Investigations with linear models. ICLR 2023

Quiz: turn to your neighbour and briefly explain the core idea behind in-context learning

Food for thought:

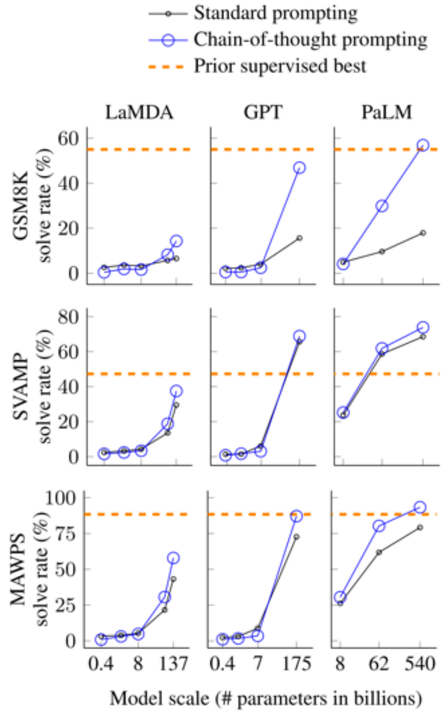
What are the core principles and ideas?
What is intuitive? What is (so far) unclear?

Another ability:
Chain-of-thought reasoning

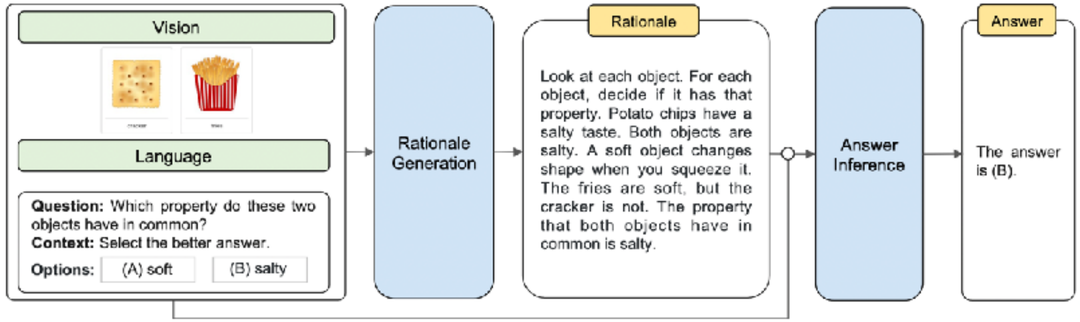
Multimodal Chain-of-Thought Reasoning in Language Models

Chain-of-thought prompting for NLP

- Either few-shot
- Or zero-shot by adding "let's think step by step"



	MultiArith	GSM8K
Zero-Shot	**17.7**	**10.4**
Few-Shot (2 samples)	33.7	15.6
Few-Shot (8 samples)	33.8	15.6
Zero-Shot-CoT	**78.7**	**40.7**



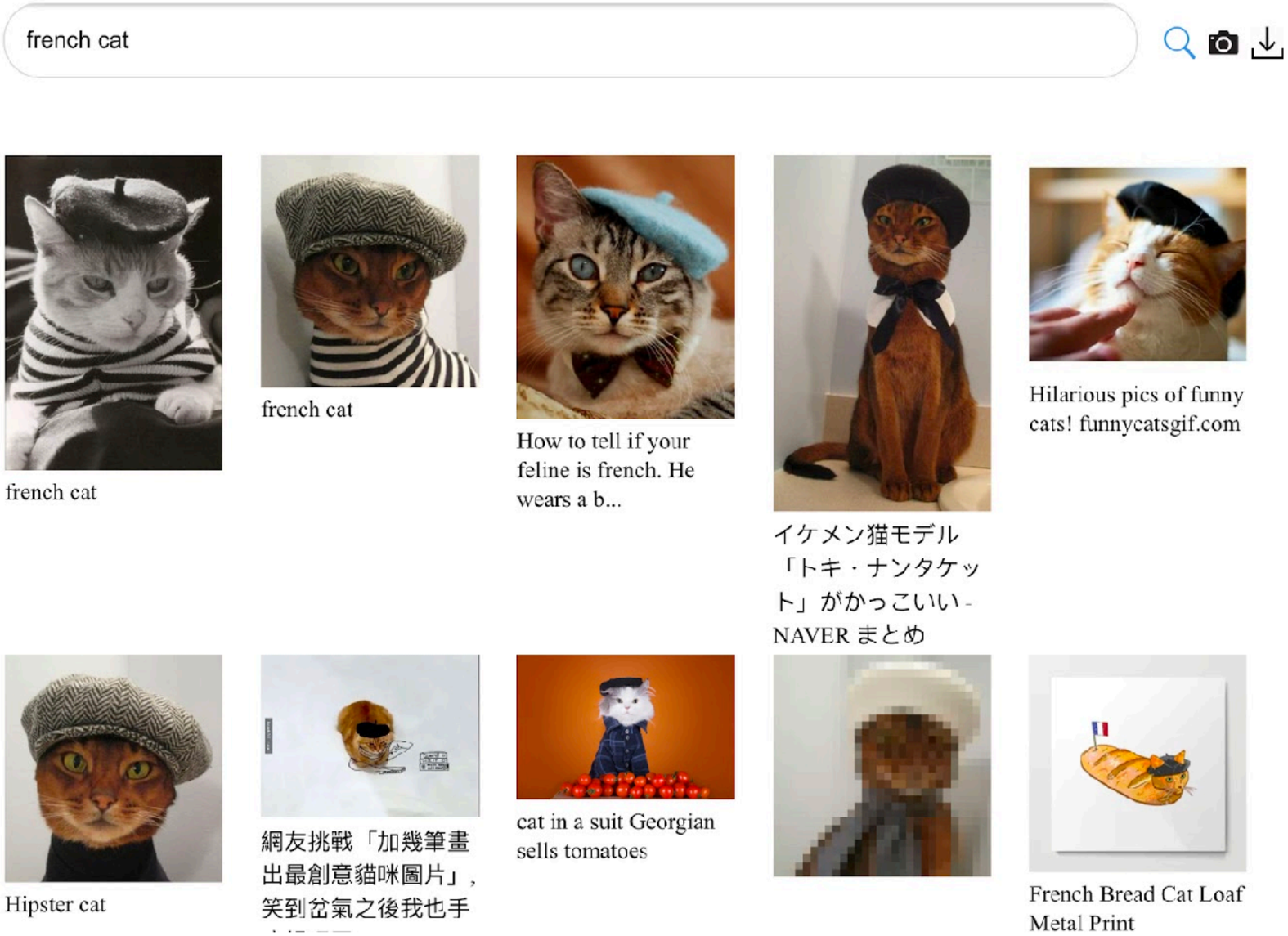
Chain-of-thought prompting for VL

- First generate rationale, then the answer (both supervised finetuning 😮‍💨)
- "small" LLMs (ie not zero- or few-shot)

Wei et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS 2022
Kojima et al. Large Language Models are Zero-Shot Reasoners. NeurIPS 2022
Wang et al. Self-consistency improves chain of thought reasoning in language models. ICLR 2023
Zhang et al. Multimodal Chain-of-Thought Reasoning in Language Models. 2023

Quiz: turn to your (other) neighbour and discuss *why* things like chain-of-thought helps

Vision-Language Datasets



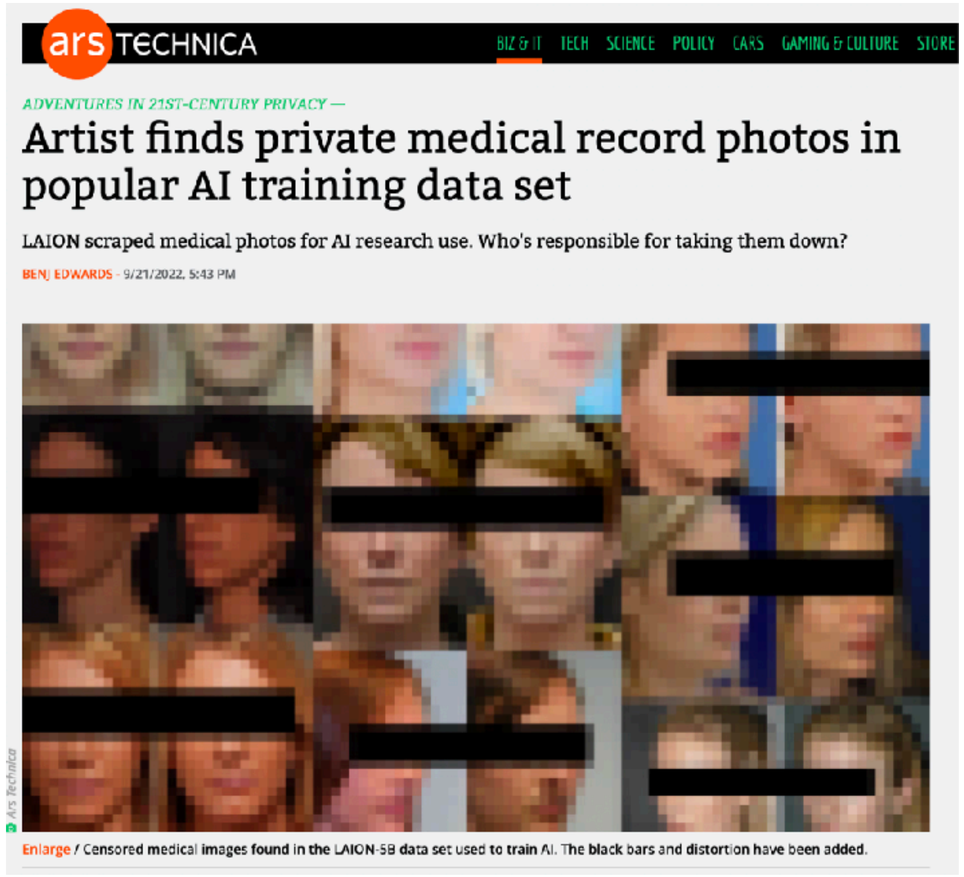
LAION: Large-scale Artificial Intelligence Open Network

Use "dump of internet": Common Crawl

CLIP-based filtering ~90% removed, yielding ~6 billion

Further filtering of NSFW, watermarked images

Training dataset for generative models like Stable Diffusion



Schumann et al. LAION-5B: An open large-scale dataset for training next generation image-text models. NeurIPS-Data 2022
https://arstechnica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data-set/

Demo

https://rom1504.github.io/clip-retrieval

Explore some search terms. What sort of content do you find?

After the break: discuss with your neighbor the pros and cons of the dataset.

For this, assign the role of advocate vs opposer beforehand and write down your reasons. (6min)

Conceptual Captions (CC3M, CC12M)

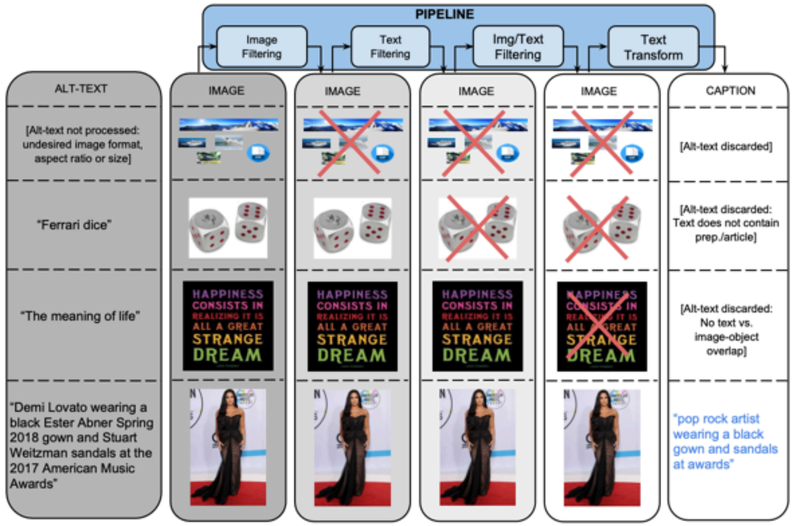


Figure 2: Conceptual Captions pipeline steps with examples and final output.

Clean based on: alt-text:

* high unique word ratio covering various POS tags
* remove ones with high rate of token repetition
* Capitalisation is good indicator
* Filter based on NSFW
* ... -> 3% remains
* further filtering with supervised image classifier

Finally: replace with hypernyms (e.g. "actor"), remove locations etc.

Sharma et al. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. ACL 2016
Changpinyo et al. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. CVPR 2021

Multimodal C4: An open, billion-scale corpus of images interleaved with text.

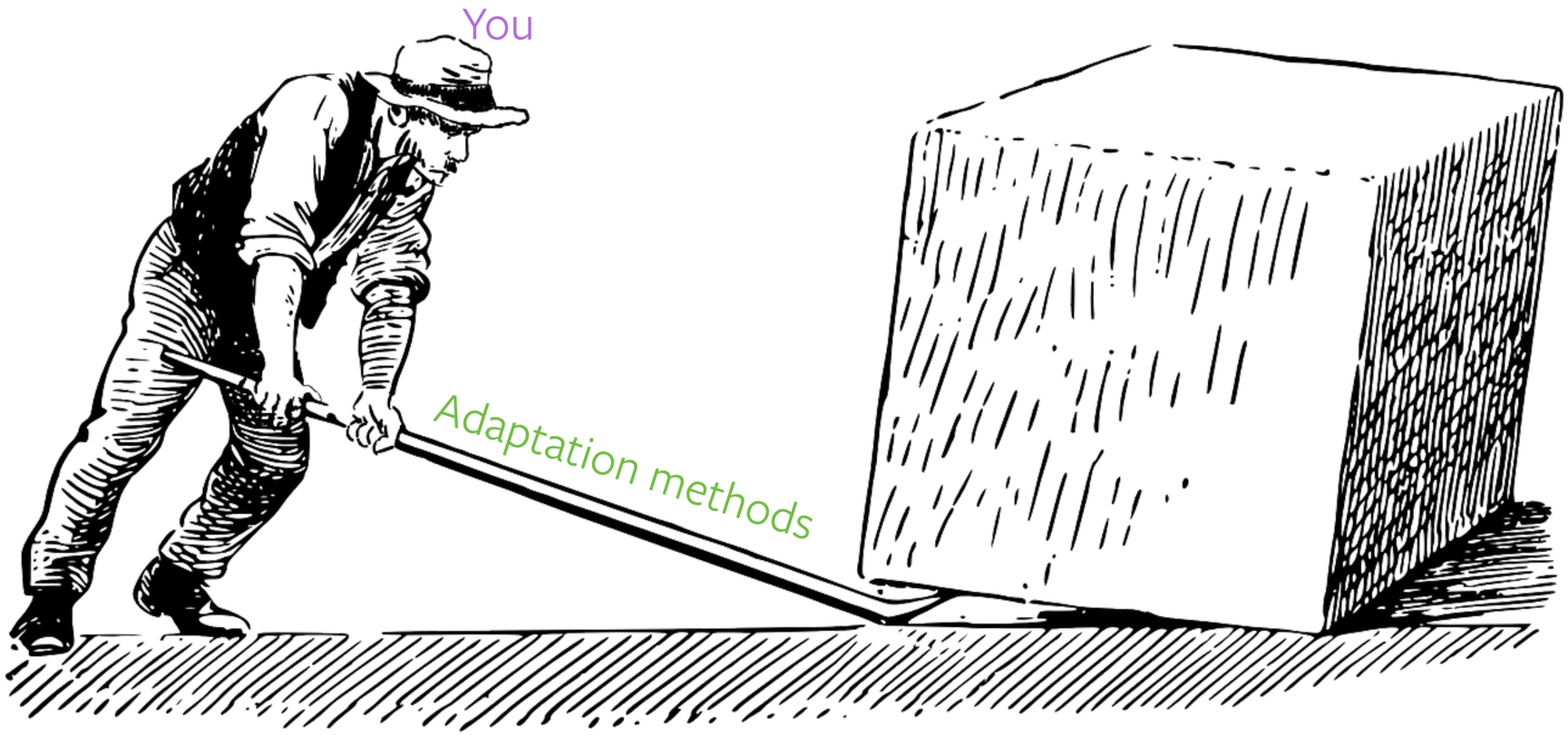
	# images	# docs	# tokens	Public?
M3W (Flamingo) [2]	185M	43M	-	✗
Interleaved training data for CM3 [1]	25M	61M	223B	✗
Interleaved training data for KOSMOS-1 [13]	⩽ 355M	71M	-	✗
Multimodal C4 (`mmc4`)	585M	103M	43B	✓
Multimodal C4 fewer-faces (`mmc4-ff`)	385M	79M	34B	✓
`mmc4` core (`mmc4-core`)	30.5M	7.4M	2.5B	✓
`mmc4` core fewer-faces (`mmc4-core-ff`)	22.9M	5.6M	1.8B	✓

- Large dataset
- Several manual and CLIP based filters

Sentence	Image	CLIP Similarity
`Our new service for teams to manage their fleets for racing.`		
`Getting boats has never been this easy.`		
`Get a step ahead with the planning for your team and get all the boats you need for next season races.`		23.51
`Our new service for teams to manage their fleets for racing.`		22.40
`As easy as adding boats to a list, this service aims to be the simplest way to rent boats, no extra knowledge needed and with full support from our staff.`		
`Get all the features of a Nelo boat, from having great equipment to our service team for a fraction of the price of a new boat.`		28.76
`All our rental boats for racing are carefully maintained and revised between each race so each boat is as good as new.`		

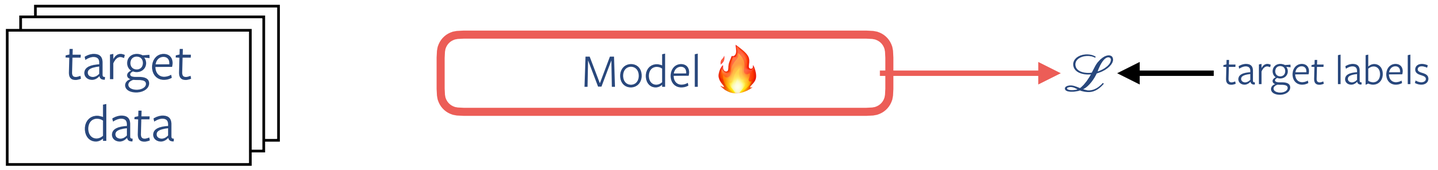
Table 5: An example document from `mmc4` with interleaved sentences and images, together with the CLIP ViT/-14 image-text similarities. This document contains two logo-related images (the 2nd & 3rd images with "NELO") that are relevant to the content of this document, and are therefore excluded from the category of advertisement.

Adaptation of large models

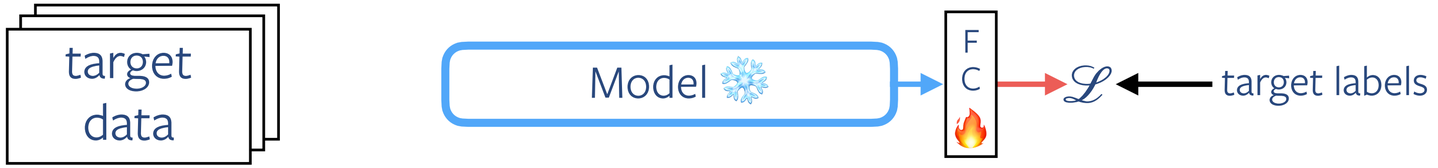


Main ways of adapting models (1/2)

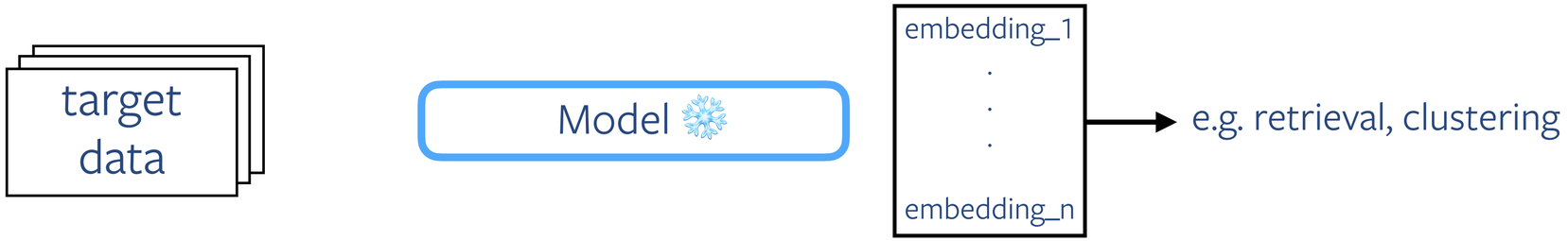
Full-finetuning



Limited-finetuning (e.g. linear probing)

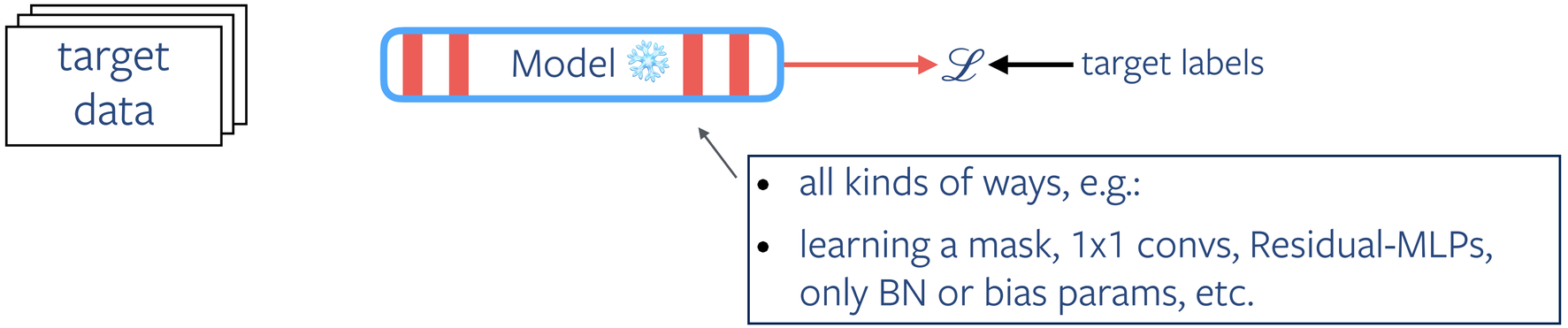


No-finetuning (e.g. used for retrieving similar instances)



Main ways of adapting models (2/2)

Adapters



- all kinds of ways, e.g.:
- learning a mask, 1x1 convs, Residual-MLPs, only BN or bias params, etc.

Prompt/prefix learning

- similar to prompt manual engineering [like "step-by-step" or "trending on artstation"]

Rebuffi et al. Learning multiple visual domains with residual adapters. NeurIPS 2017, Houlsby et al. Parameter-Efficient Transfer Learning for NLP. ICML 2019
Elsayed et al. Adversarial reprogramming of neural networks. ICLR 2019
Li et al. Prefix-tuning: Optimizing continuous prompts for generation. ACL 2021

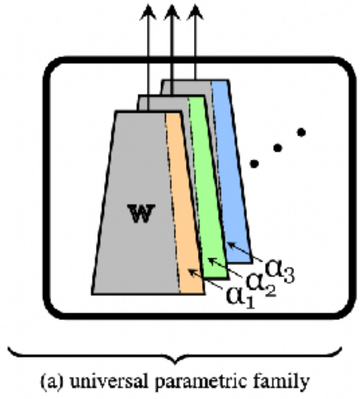
Some examples

Adapters: any modification "in the middle" of NNs

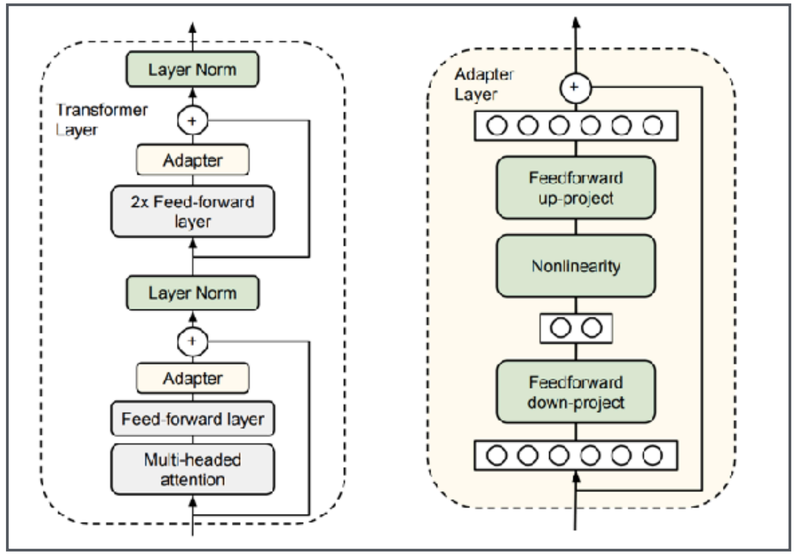
Simplest form: residual adapters

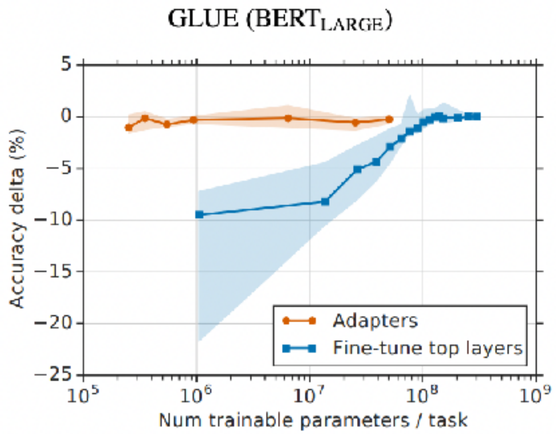
$$g(x;\alpha) = x + \alpha * x.$$

limit α to e.g. 1x1 conv



(a) universal parametric family



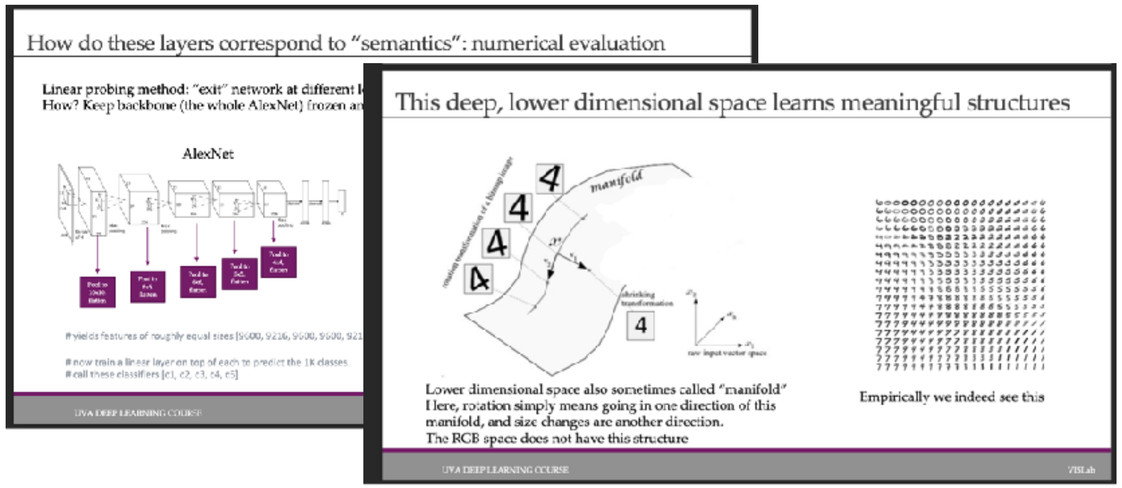


- (-) makes computation graph more complex; adds inference time
- (+) doesn't require much memory to store
- (+) very expressive/performant and fast to learn

Rebuffi et al. Learning multiple visual domains with residual adapters. NeurIPS 2017, Houlsby et al. Parameter-Efficient Transfer Learning for NLP. ICML 2019

LoRA: adapting matrix multiplies in efficiently / "a generalisation of full-finetuning"

$$\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$$



Remember from DL1:

- real data ~ lies on lower dimensional manifold,
- DNNs map from RGB space gradually to more semantic space.

Normal fully connected layer:

$$h = W_0 x$$

LoRA adapted:

$$h = W_0 x + \Delta W x = W_0 x + BAx$$

BA is low-rank matrix.

"Low-rank"
--> think of it as outer-product of few vectors

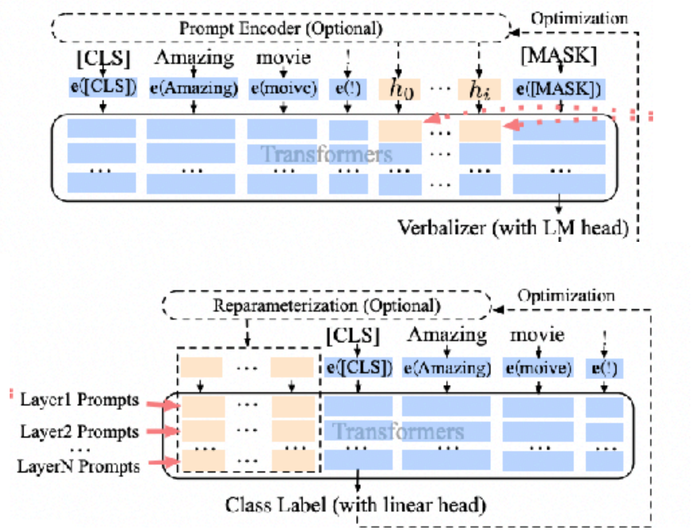
- (-) not as expressive as adapters
- (+) linear op, so after training can be fused with original weights --> same speed

Hu et al. LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022
Aghajanyan et al. Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. ACL 2021

In LoRA: Why are two matrices A & B learned, even though they are only used as A*B?

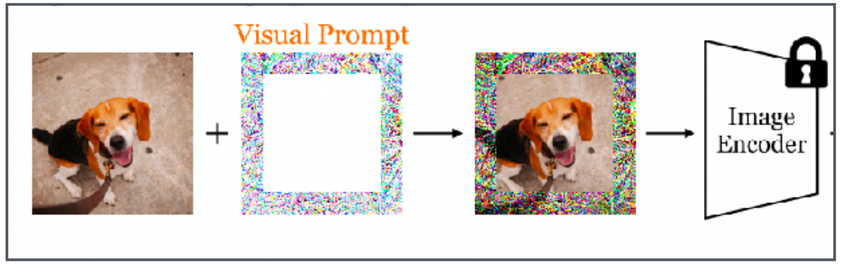
1) Learning two easy linear transformations is easier than one complicated one
2) Enforcing low-rank ness via singular value decomposition is expensive
3) By allowing A and B to be square-sized we obtain more optimised matrix multiplies
4) Actually, rather than two linear layers, a single bigger one would also do it.

Prompt learning: per task

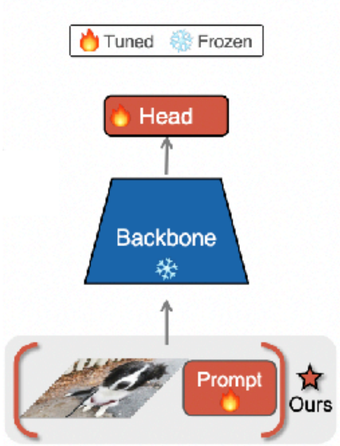


Li et al. Prefix-tuning: Optimizing continuous prompts for generation. ACL 2021

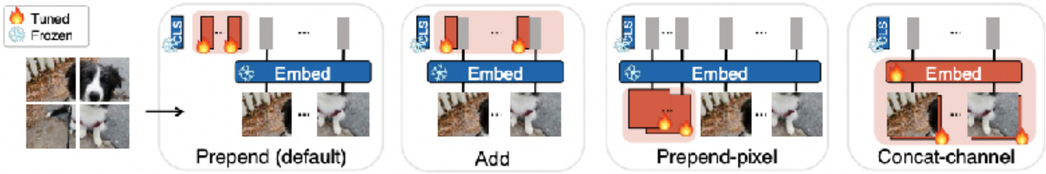
- prefixed are just learnable vectors
- 🧐: are reparameterised as an MLP that gets a fixed input (″more stable″)
- Extend this: ″deep prompt tuning″



- Works also for CNNs
- Strictly input-only

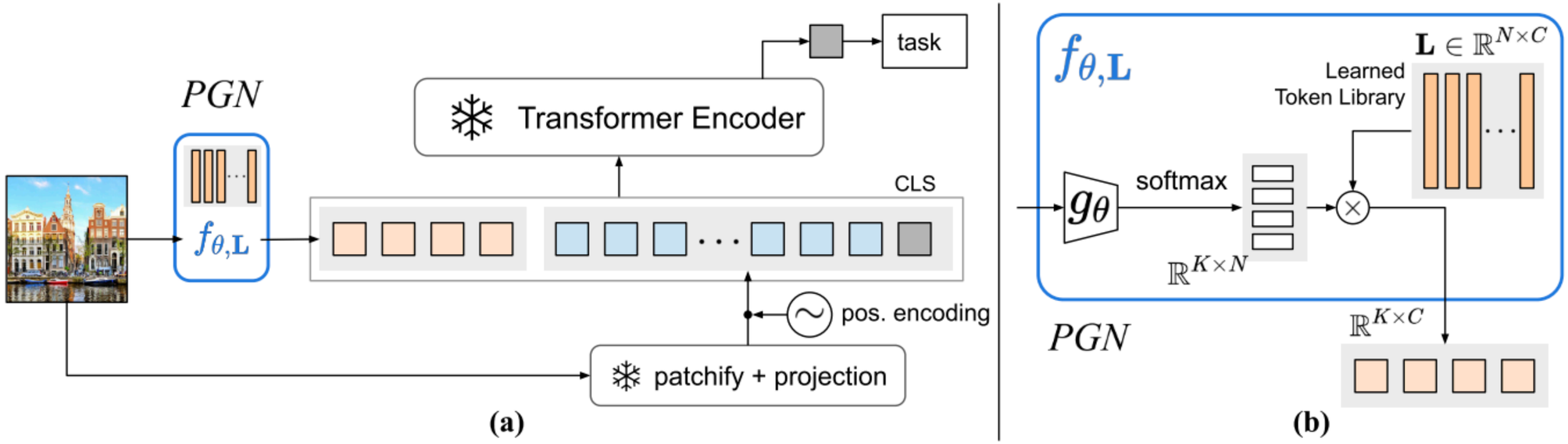


- Actually also trains linear layer on top
- 👇 explore various ways of prompting inputs for visual inputs



Li et al. Prefix-tuning: Optimizing continuous prompts for generation. ACL 2021
Liu et al. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. ACL 2022
Bhang et al. Exploring Visual Prompts for Adapting Large-Scale Models. 2022
Zhou et al. Learning to Prompt for Vision-Language Models.IJCV 2021l, Zhou et al. Conditional Prompt Learning for Vision-Language Models. CVPR 2022
Jia et al. Visual Prompt Tuning.. ECCV 2022

Prompt learning: per datum: *Prompt Generation Networks*



- Learn a input-to-prompt mini-network
- Generate prompts from a set of learnable prompts
- Prompts (learned in space after first conv1), can be made to be input-only (convs are linear operation!)

	PGN backbone (alone)	CLIP	CLIP with PGN
CIFAR-100	63.7	63.1	79.3

PGN learns what's missing in CLIP

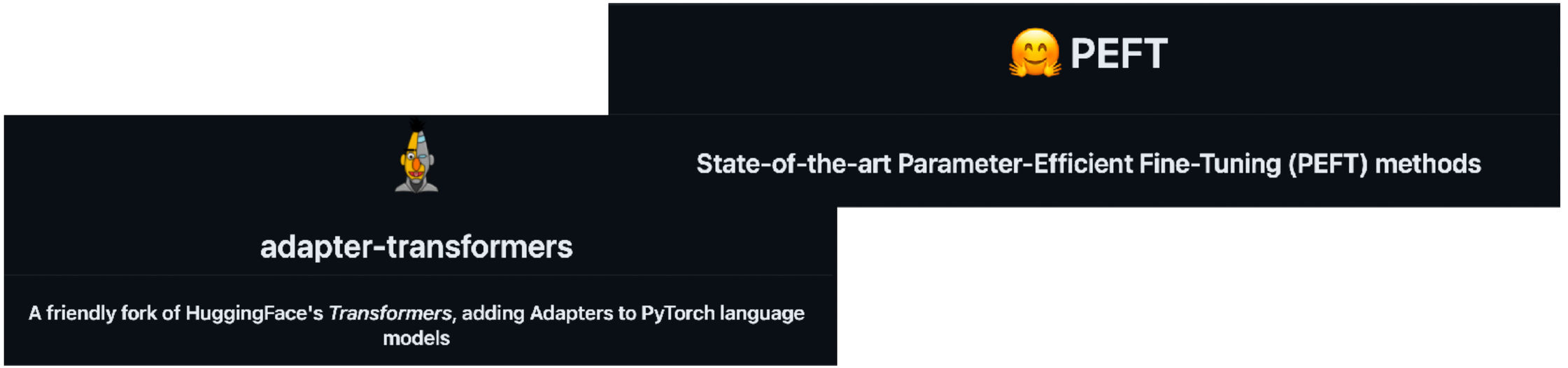
Method	ImageNet	A	R	V2	Sketch
PGN	66.0	22.8	62.5	56.7	36.5
LP	67.0	10.6	38.1	1.0	36.1

More robust compared to linear probing (LP)

See also:

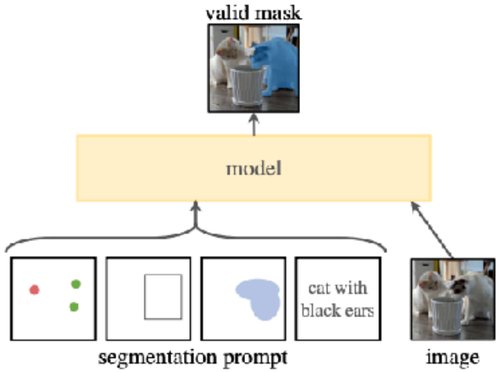
https://github.com/adapter-hub/adapter-transformers

https://github.com/huggingface/peft

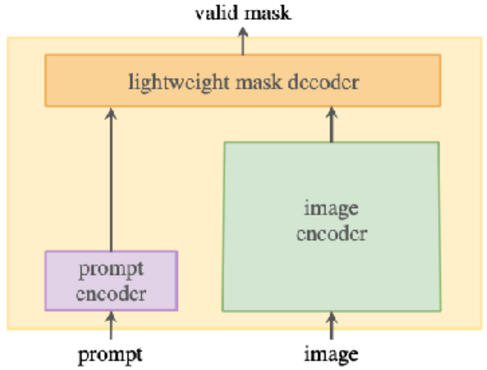


Other interesting developments

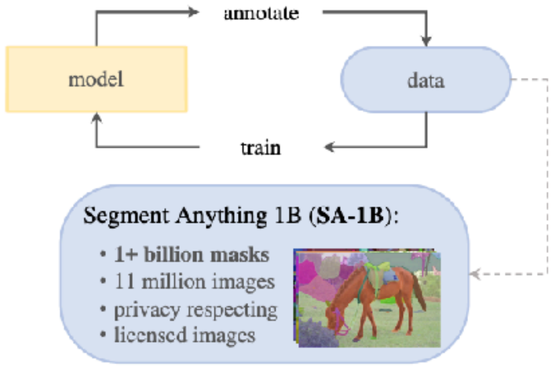
Segment Anything Model (SAM)



(a) **Task**: promptable segmentation

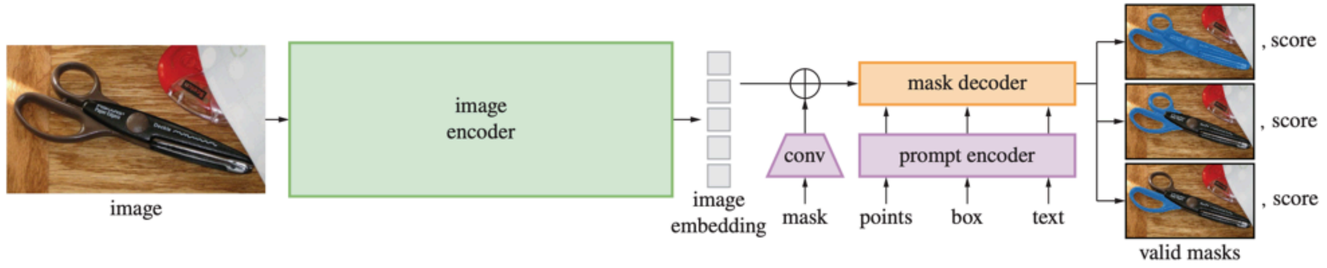


(b) **Model**: Segment Anything Model (**SAM**)



(c) **Data**: data engine (top) & dataset (bottom)

Annotation collection stages
1) 4.3M masks from 120k images
2) 11M masks in 300k images
3) 1B masks in 11M images



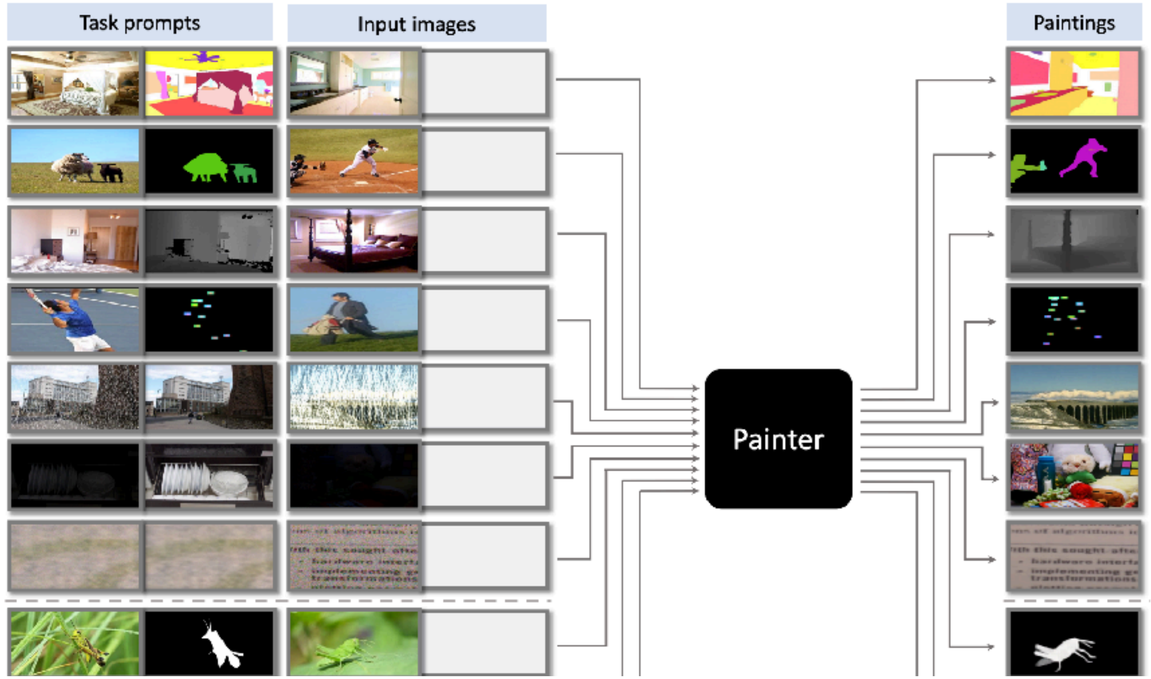
Results

	mIoU at 1 point	mIoU at 3 points
perceived gender presentation		
feminine	54.4 ±1.7	90.4 ±0.6
masculine	55.7 ±1.7	90.1 ±0.6
perceived age group		
older	62.9 ±6.7	92.6 ±1.3
middle	54.5 ±1.3	90.2 ±0.5
young	54.2 ±2.2	91.2 ±0.7

	mIoU at 1 point	mIoU at 3 points
perceived skin tone		
1	52.9 ±2.2	91.0 ±0.9
2	51.5 ±1.4	91.1 ±0.5
3	52.2 ±1.9	91.4 ±0.7
4	51.5 ±2.7	91.7 ±1.0
5	52.4 ±4.2	92.5 ±1.4
6	56.7 ±6.3	91.2 ±2.4

Table 2: SAM's performance segmenting people across perceived gender presentation, age group, and skin tone. 95% confidence intervals are shown. Within each grouping, all confidence intervals overlap except older *vs*. middle.

Painter/SegGPT: if we visualise outputs via colors masks, can we treat them instead as input-images?



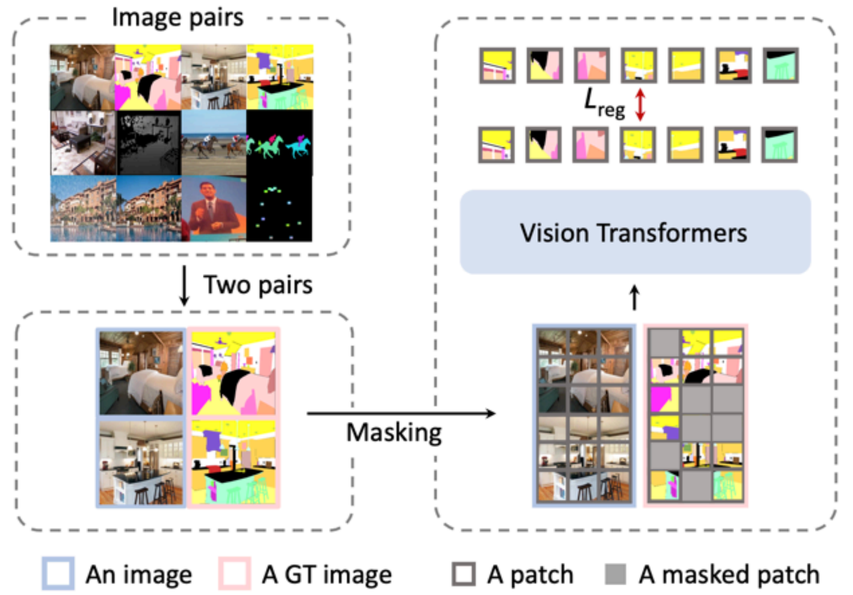
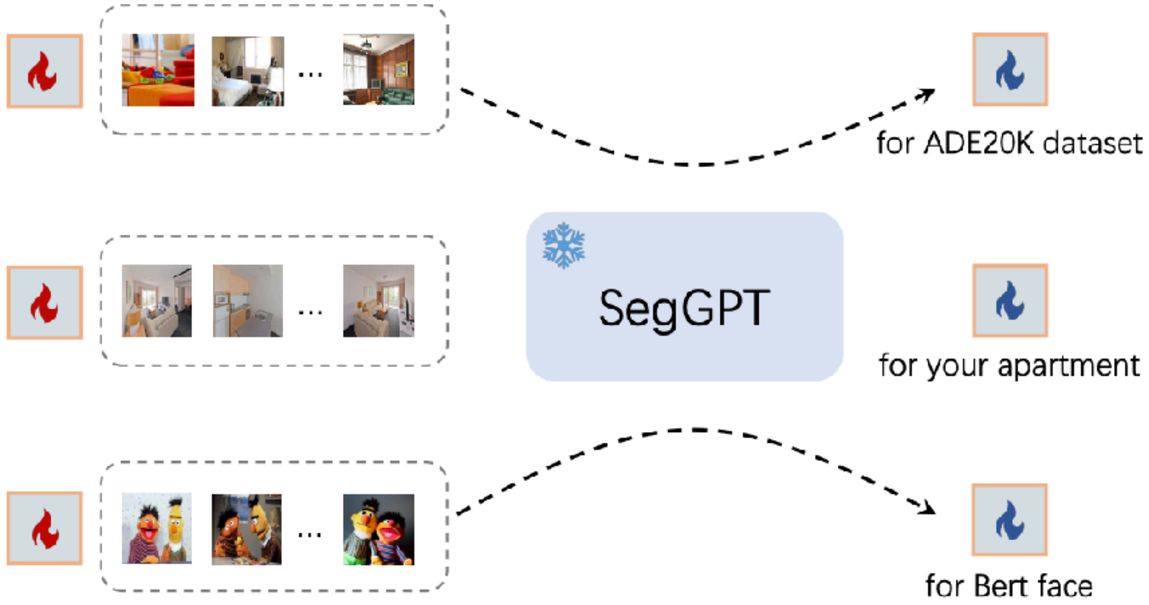
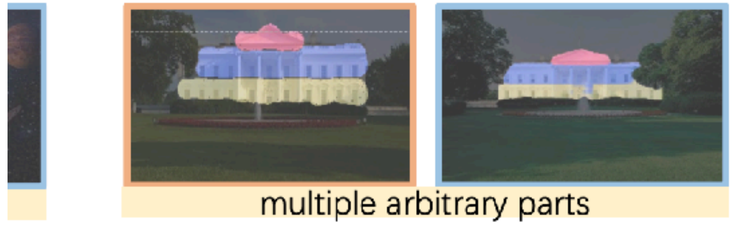


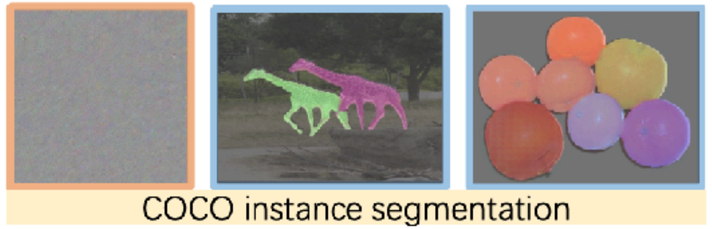
Figure 2. The training pipeline of the masked image modeling (MIM) framework.

Wang et al. SegGPT: Segmenting Everything In Context. Apr 2023
Wang et al. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. CVPR 2023

Finetune via prompt learning







Wang et al. SegGPT: Segmenting Everything In Context. Apr 2023
Wang et al. Images Speak in Images: A Generalist Painter for In-Context Visual Learning. CVPR 2023