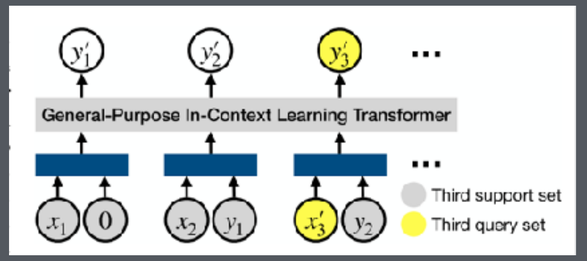
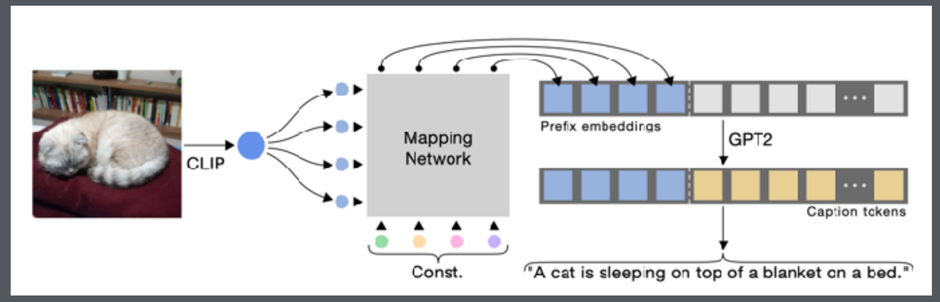
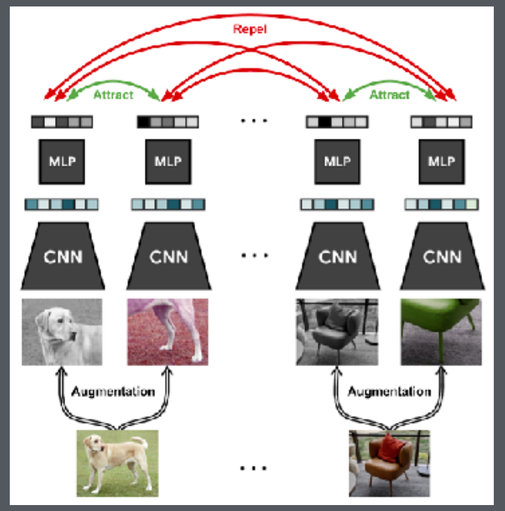
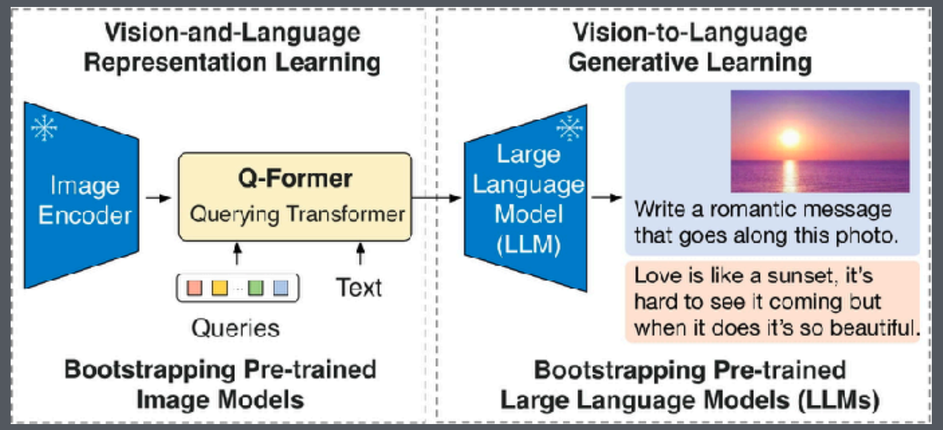
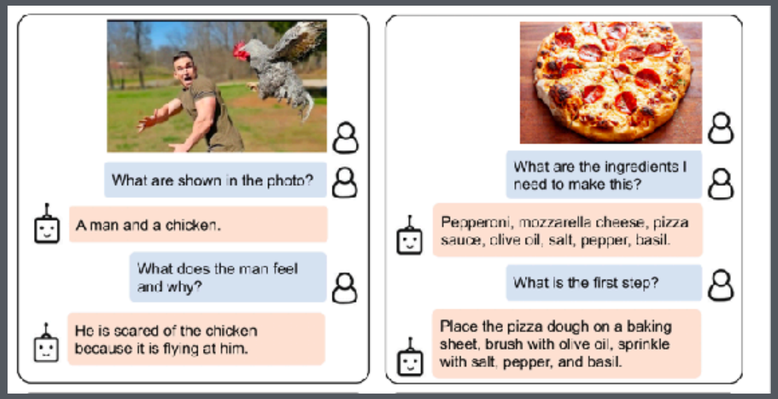
Self-supervised and vision-language learning











@ DEEP LEARNING 2

YUKI M. ASANO
VIS LAB & QUVA LAB

Self-supervised and vision-language learning is everywhere now

FINANCIAL TIMES

US COMPANIES TECH MARKETS CLIMATE OPINION WORK & CAREERS LIFE & ARTS HTSI

Artificial intelligence + Add to myFT

GPT-4 from OpenAI shows advances – ar
moneymaking potential

Microsoft-backed group shifts towards showing less openness amid race to com
AI systems

The Guardian

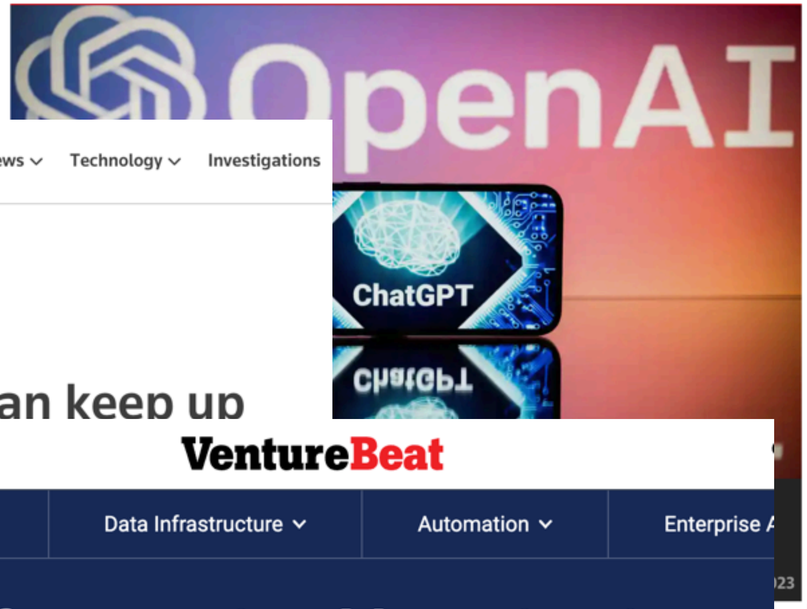
Support us

News Opinion Sport Culture Lifestyle

World UK Coronavirus Climate crisis Environment Science Global development More

ChatGPT

April 2023



deVolkskrant

Log in

Topverhalen vandaag Opinie Cultuur & Media Podcasts Beter Leven

ZES VRAGEN

Nieuwe 'turbo-versie' van ChatGPT is een stuk veelzijdiger en kan ook omgaan met plaatjes

Source: Imgur (Reddit)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

REUTERS®

World Business Markets Legal Breakingviews Technology Investigations

Disrupted

3 minute read · March 15, 2023 7:17 PM GMT+1 · Last Updated a month ago

Bar exam score shows AI can keep up
with 'human lav

By Karen Sloan

VentureBeat

Security Data Infrastructure Automation Enterprise A

Why self-supervised learning is a medical AI game-changer

What will we cover?

Topics/Models

- GPT-2,3,4, ChatGPT
- MAE
- CLIP, LiT
- ALIGN, ClipCap
- BLIP, BLIP
- Socratic Models
- Flamingo
- FROMAGe
- Kosmos-1
- LAION, Conceptual Captions
- LoRA
- PGN

Intended Learning Outcomes (ILOs)

- The student can present and explain the crucial works in the recent large-scale vision-language learning domain
- The student can explain the difference and commonalities between previous and recent vision-language approaches and self-supervised learning
- The student can describe in-context learning
- The student can interpret, critically analyse and judge scientific publications that combine pretrained language models with visual learning
- The student can apply and develop large language models via various API calls into their own code

Philosophy of these two lectures

There's a lot going on.

This is not an exhaustive enumeration, but instead meant to showcase a number of important works that represent the different research directions.

I've achieved my goal if after these two lectures you think:
"combining self-supervised/vision-language learning is exciting and the lectures gave me ideas for my own future creative research ideas"

Note: if you wish to unleash your creative ideas,
I will be offering some MSc projects in this direction this year.

Self-supervised Pretraining

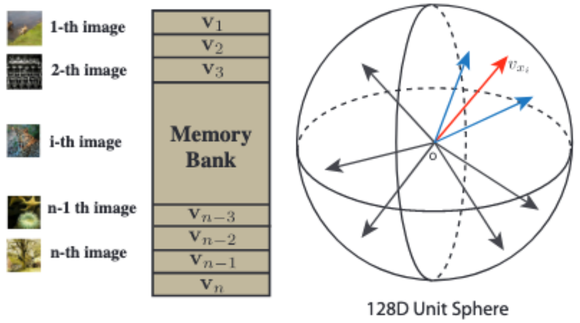
Q: How does one learn without labels?

A: Need to generate a loss that provides gradients.

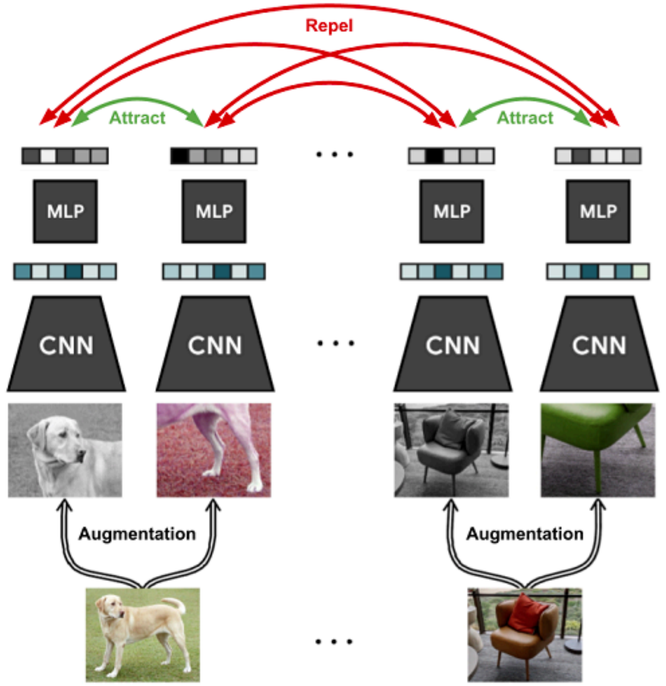
We will focus on signals from

- Image uniqueness + augmentation invariance (”contrastive learning”)
- Reconstruction (Masked Image Modelling)

Modern Noise-contrastive self-supervised learning



NPID



SimCLR

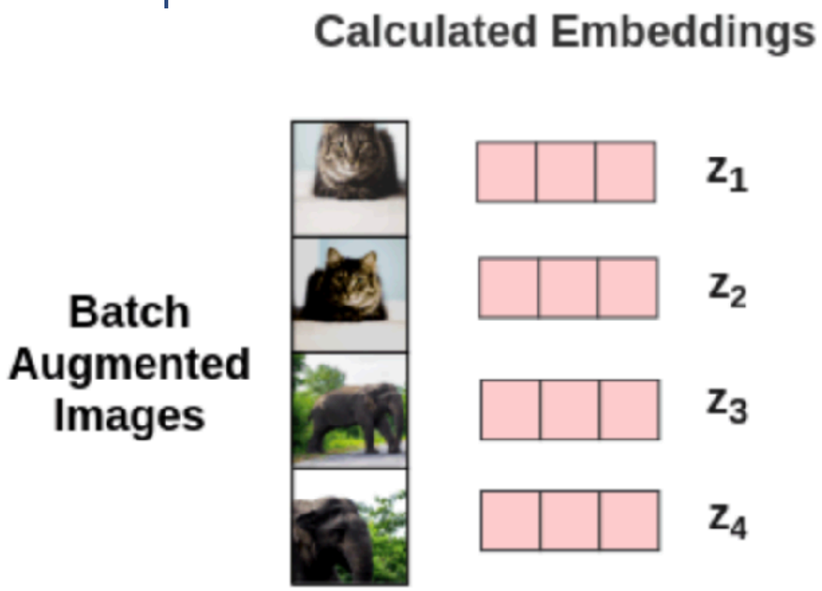
Enforces image-uniqueness and

enforces augmentation-invariance

Wu et al. *Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination.* CVPR 2018
Chen et al. *A Simple Framework for Contrastive Learning of Visual Representations.* ICML 2020

How SimCLR works in detail

Step 1



Step 2

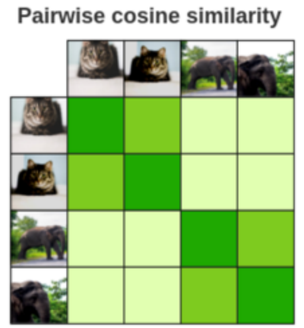
Similarity Calculation of Augmented Images

$$\text{similarity}(x_i, x_j) = \text{cosine similarity}(z_i, z_j)$$

$$s_{i,j} = \frac{z_i^T z_j}{(\tau ||z_i|| ||z_j||)}$$

- τ is the adjustable temperature parameter. It can scale the inputs and widen the range `[-1, 1]` of cosine similarity
- ||z_i|| is the norm of the vector.

Step 3

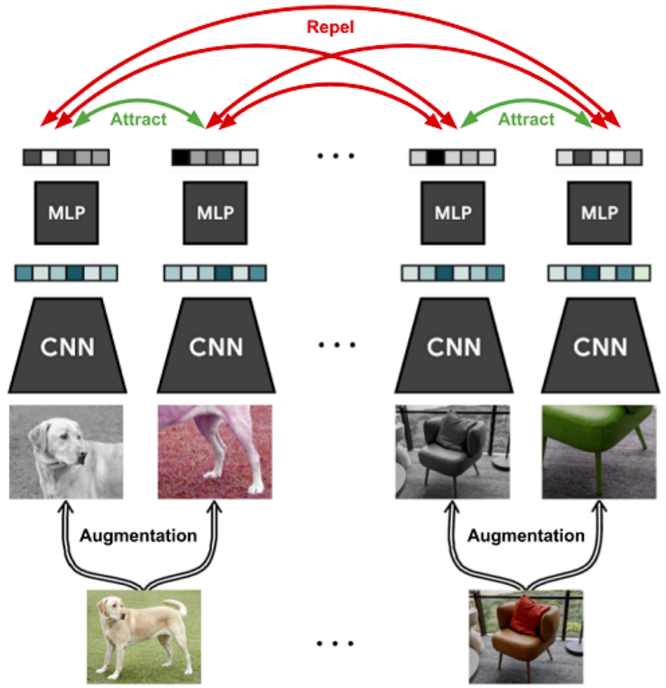


Loss: *relatively* increase similarity for pairs, decrease rest

What happens if you only try to increase the diagonal?

https://amitness.com/2020/03/illustrated-simclr/

Putting it into a loss function



SimCLR

The contrastive loss for positive pairs i,j:

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} {}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)},$$

with z_i, z_j embeddings for images i and j,

τ a temperature, sim() *is the dot-product*

"non-parametric" softmax

Enforces image-uniqueness and

enforces augmentation-invariance

Wu et al. *Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination.* CVPR 2018
Chen et al. *A Simple Framework for Contrastive Learning of Visual Representations.* ICML 2020

Language Modelling

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

https://thegradient.pub/understanding-evaluation-metrics-for-language-models/

Language Modelling via next-word prediction

Why "erudite" is not a good guess

Factor the probability of a datapoint:

$$P_{(w_1,w_2,\ldots,w_n)} = p(w_1)p(w_2|w_1)p(w_3|w_1,w_2)...p(w_n|w_1,w_2,..,w_{n-1})$$
$$= \prod_{i=1}^{n} p(w_i|w_1,...,w_{i-1})$$

S = Where are we going

Previous words (Context)

Word being predicted

P(S) = P(Where) x P(are | Where) x P(we | Where are) x P(going | Where are we)

https://thegradient.pub/understanding-evaluation-metrics-for-language-models/

Radford et al. Improving Language Understanding by Generative Pre-Training . 2018

Generative Pretrained Transformer (GPT)
simply does language modelling with a Transformer (decoder)

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

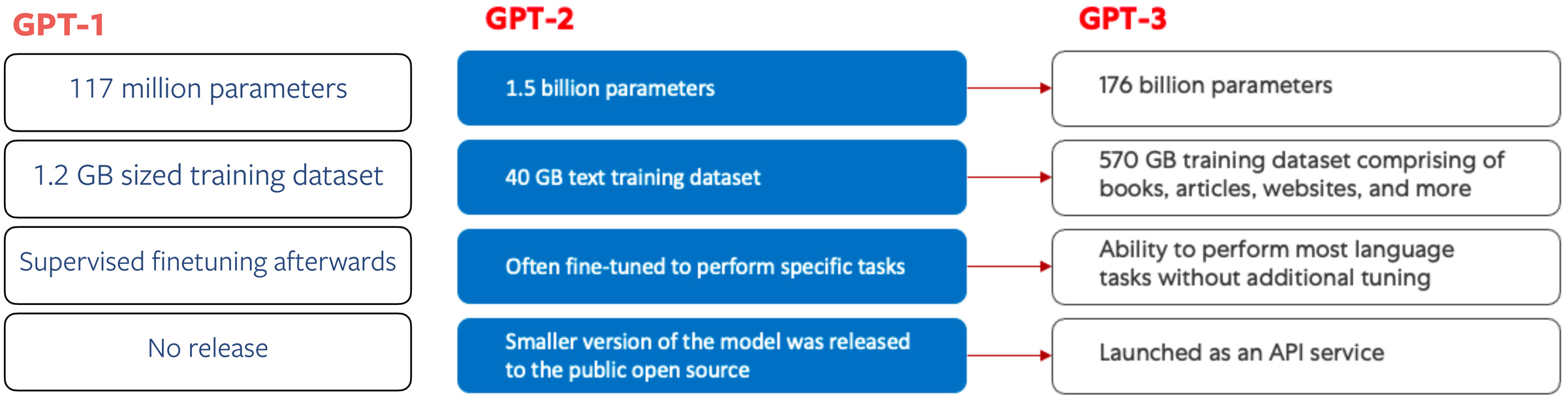
$$h_0 = UW_e + W_p$$
$$h_l = \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n]$$
$$P(u) = \texttt{softmax}(h_n W_e^T)$$

U = (u_k, . . . , u_1) is the context vector of tokens,
n is the number of layers,
W_e is the token embedding matrix,
W_p is the position embedding matrix

in practice: "causal" (left-to-right) context via masking

Radford et al. Improving Language Understanding by Generative Pre-Training . 2018

GPT-1,2,3: same loss. different training data and model sizes



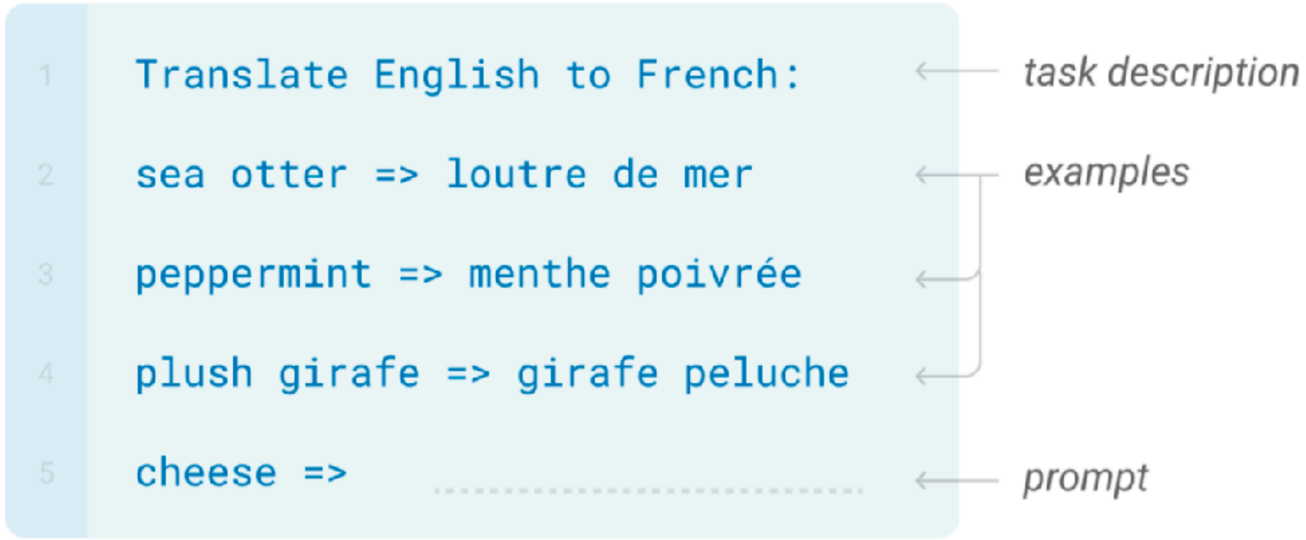
https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286

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.



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.

In-context Learning: benefitting from more examples in the input

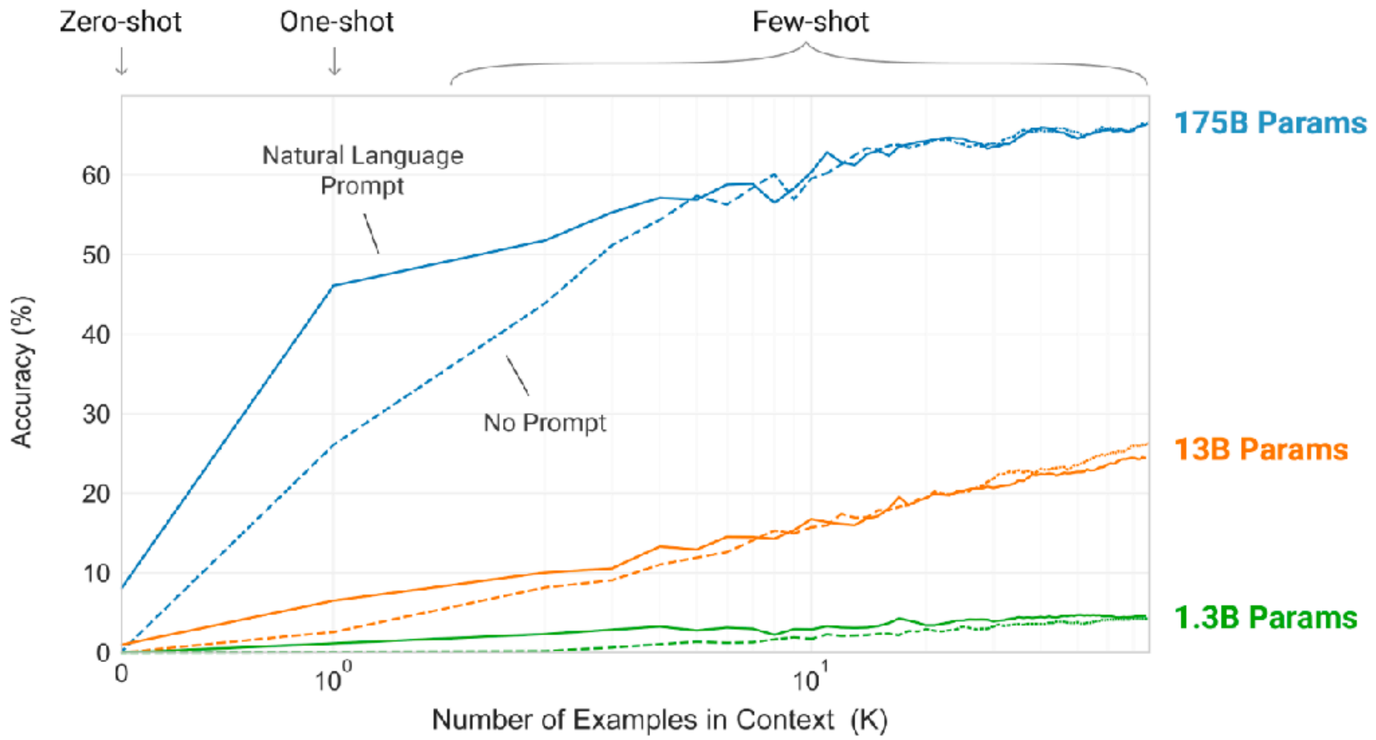
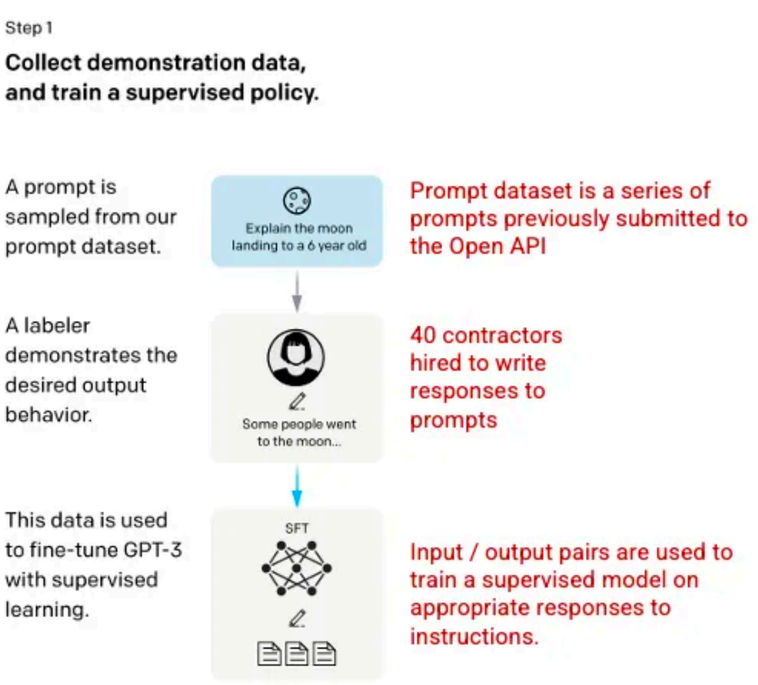
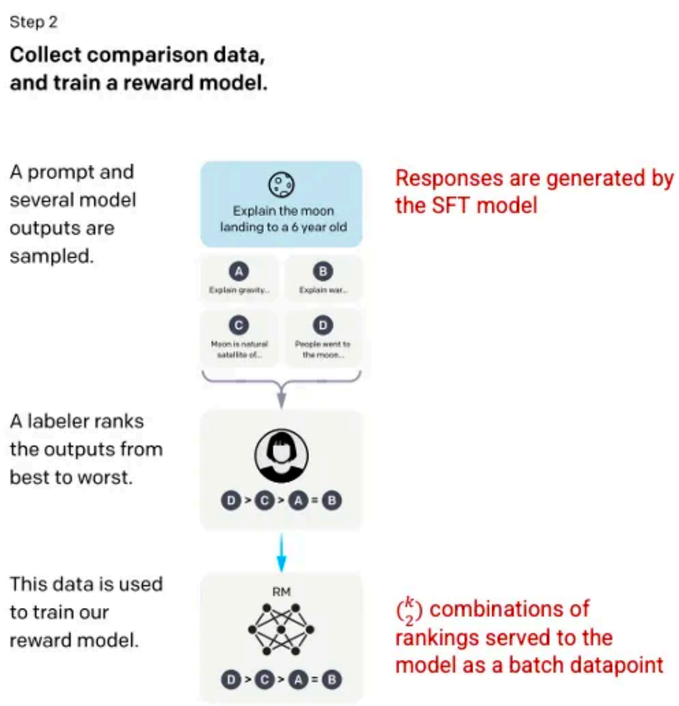


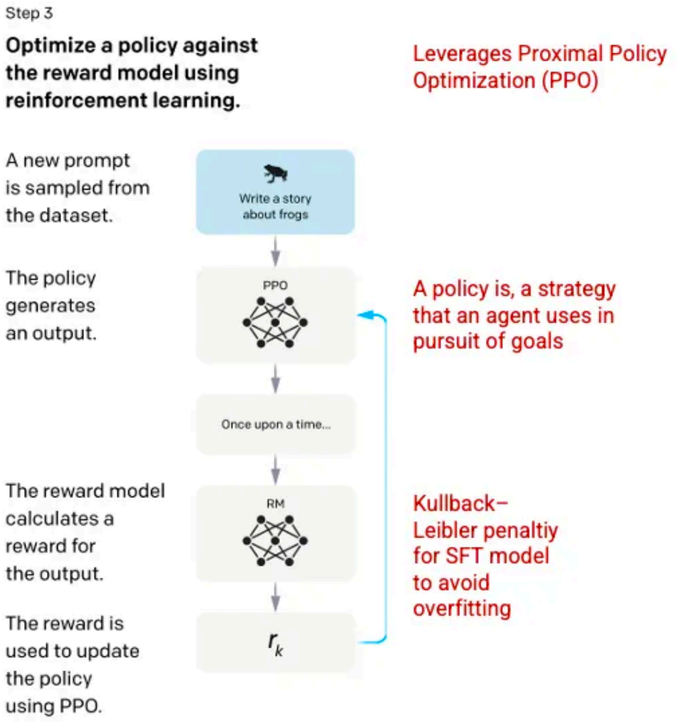
Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper "in-context learning curves" for large models demonstrate improved ability to learn a task from contextual information. We see qualitatively similar behavior across a wide range of tasks.

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

[btw: ChatGPT uses GPT-3 to then supervisedly finetune on human preferences & then learn these to enable reinforcement learning]







Ouyang et al. Training language models to follow instructions with human feedback. NeurIPS 2022.
https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286

GPT-4

better.
bigger.

GPT-4 Technical Report

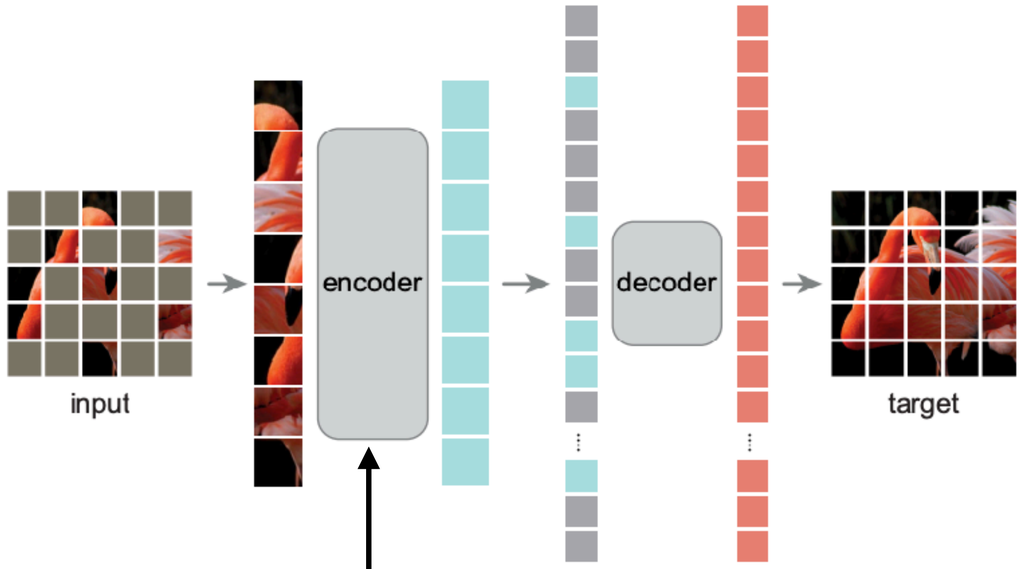
OpenAI*

Abstract

we used python

There’s more language models! Llama / (Flan)-T5 etc.
they are even open-source available and quite big! up to 65B params

Masked Image Modelling ~ same, but instead of words use image patches



Vision Transformer

Something to think:

- Words present the "atoms" of masked language modelling and have individual meaning
- Yet image-patches do not carry an individual meaning
- The output of a language model is vast (from writing a joke to your assignments)
- The output of an image model is "just" pixels
- --> what is missing for vision models?

He et al. *Masked Autoencoders Are Scalable Vision Learners*. CVPR'21
Xie et al. *SimMIM: A Simple Framework for Masked Image Modeling. ArXiv*
Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. ICLR'21
https://www.sbert.net/examples/unsupervised_learning/MLM

Multi-modal Learning

+ captions/ thoughts?

What modalities does Deep Learning (mostly) deal with?

- Generally: anything on the internet
- Images
- Text
- Speech audio
- LiDAR points
- 3D models
- ....

Multiple modalities

- Videos (RGB frames + audio + audio transcriptions if there's speech)
- Image-text (e.g. images with captions, images with alt text)
- ...

Chen et al. *A Simple Framework for Contrastive Learning of Visual Representations.* ICML 2020

What makes multi-modal learning interesting? e.g. vision-language

Text is like an "augmentation" / broader description

The man at bat readies to swing at the pitch while the umpire looks on.

The meaning depends on both modalities (rarer)

Text can also be very detailed

In the front portion of the picture we can see a dried grass area with dried twigs. There is a woman standing wearing a light blue jeans and ash colour long sleeve length shirt. This woman is holding a black jacket in her hand. On the other hand she is holding a balloon which is peach in colour. on the top of the picture we can see a clear blue sky with clouds. The hair colour of the woman is brownish.

850k images with such descriptions

Pont-Tuset et al. Connecting Vision and Language with Localized Narratives. ECCV 2020

But really: the language part makes it *very* generaliseable

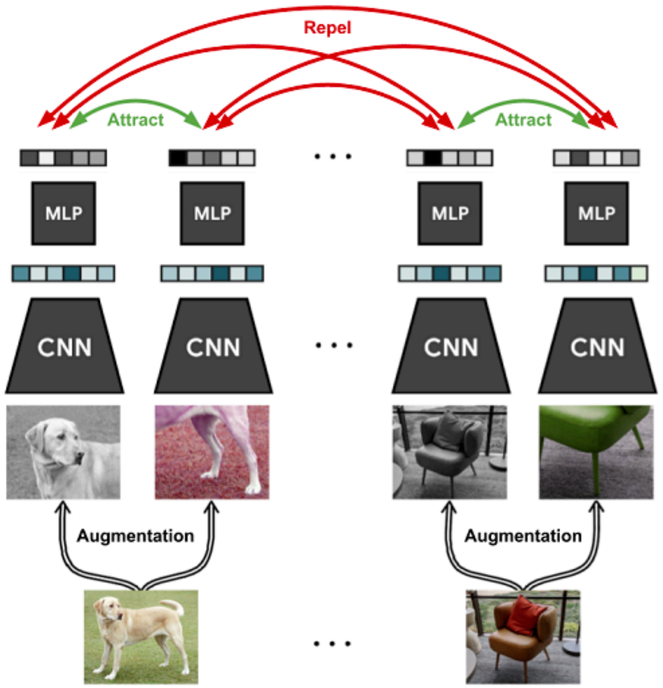
Language is a very universal format for posing and solving tasks

Language further has advantage of being human understandable

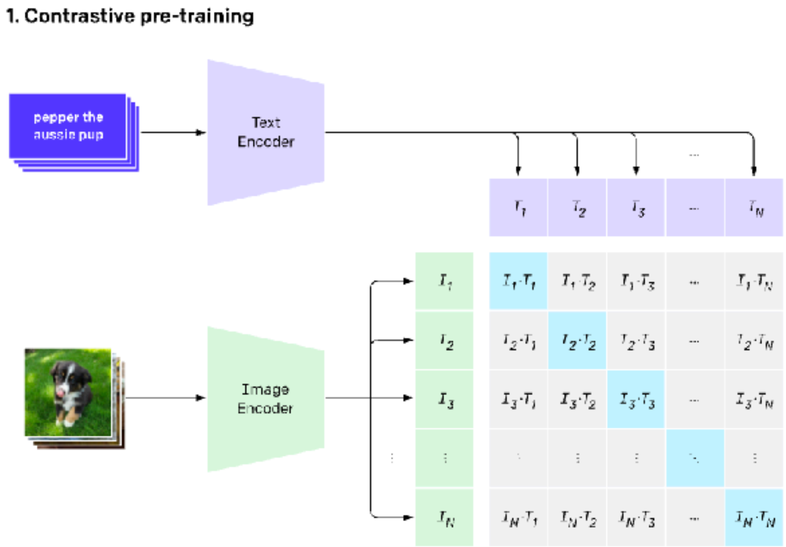
Language models are few-shot (in-context) learners

Chen et al. *A Simple Framework for Contrastive Learning of Visual Representations.* ICML 2020

CLIP from DL 1 Lecture 9 simply applies SimCLR across modalities



SimCLR

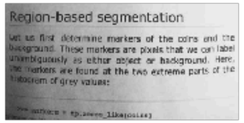


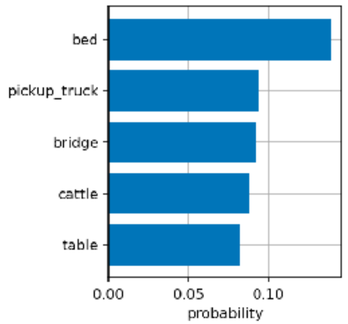
CLIP: instead of augmentation, uses an image caption (the magic is in the training data)

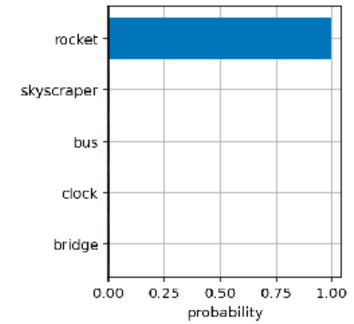
Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2020
but also: Desai & Johnson VirTex: Learning Visual Representations from Textual Annotations. CVPR 2021 etc.

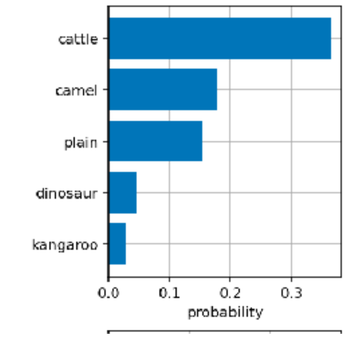
What you can do with CLIP: zero-shot classification

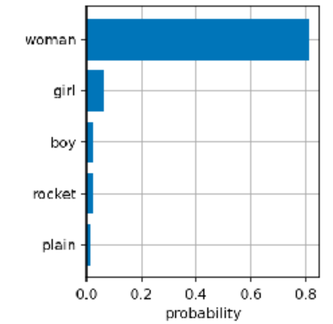
https://colab.research.google.com/github/openai/clip/blob/master/notebooks/Interacting_with_CLIP.ipynb

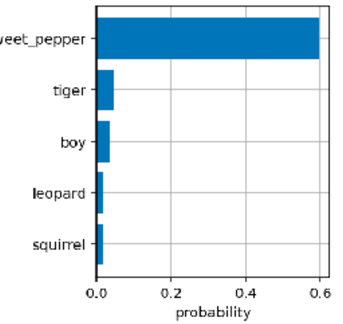


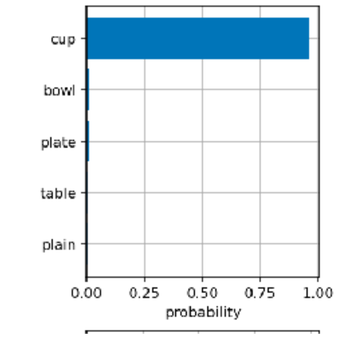












Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2020

When comparing pretrained image and language models, which one needs to adapt (more?)

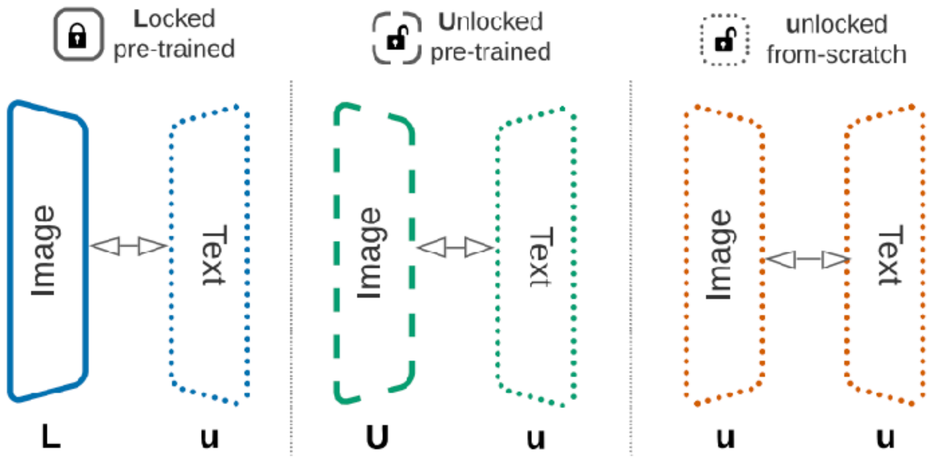


Figure 2. Design choices for contrastive-tuning on image-text data. Two letters are introduced to represent the image tower and text tower setups. `L` stands for locked variables and initialized from a pre-trained model, `U` stands for unlocked and initialized from a pre-trained model, `u` stands for unlocked and randomly initialized. `Lu` is named as "Locked-image Tuning" (LiT).

Method	ImgNet	ImgNet-v2	Cifar100	Pets
`Lu`	70.1	61.7	70.9	88.1
`Uu`	57.2	50.2	62.1	74.8
`uu`	50.6	43.3	47.9	70.3

Locking the image model is better.

Table 3: Zero-shot transfer results on ImageNet (variants).

Model	IN	IN-v2	IN-R	IN-A	ObjNet	ReaL
CLIP	76.2	70.1	88.9	77.2	72.3	-
ALIGN	76.4	70.1	92.2	75.8	72.2	-
BASIC	85.7	80.6	95.7	85.6	78.9	-
CoCa	86.3	80.7	96.5	90.2	82.7	-
LiT-g/14	85.2	79.8	94.9	81.8	82.5	88.6
LiT-e/14	85.4	80.6	96.1	88.0	84.9	88.4
LiT-22B	85.9	80.9	96.0	90.1	87.6	88.6

With only requiring one forward pass for getting image embeddings, can combine this with using a 22B parameter ViT

Zhai et al. LiT: Zero-Shot Transfer with Locked-image text Tuning. CVPR 2022
Dheghani et al. Scaling Vision Transformers to 22 Billion Parameters. 2023

Combining GPT and X

CLIP is a vision-language model GPT-3 is a language model. So they can interact via language.

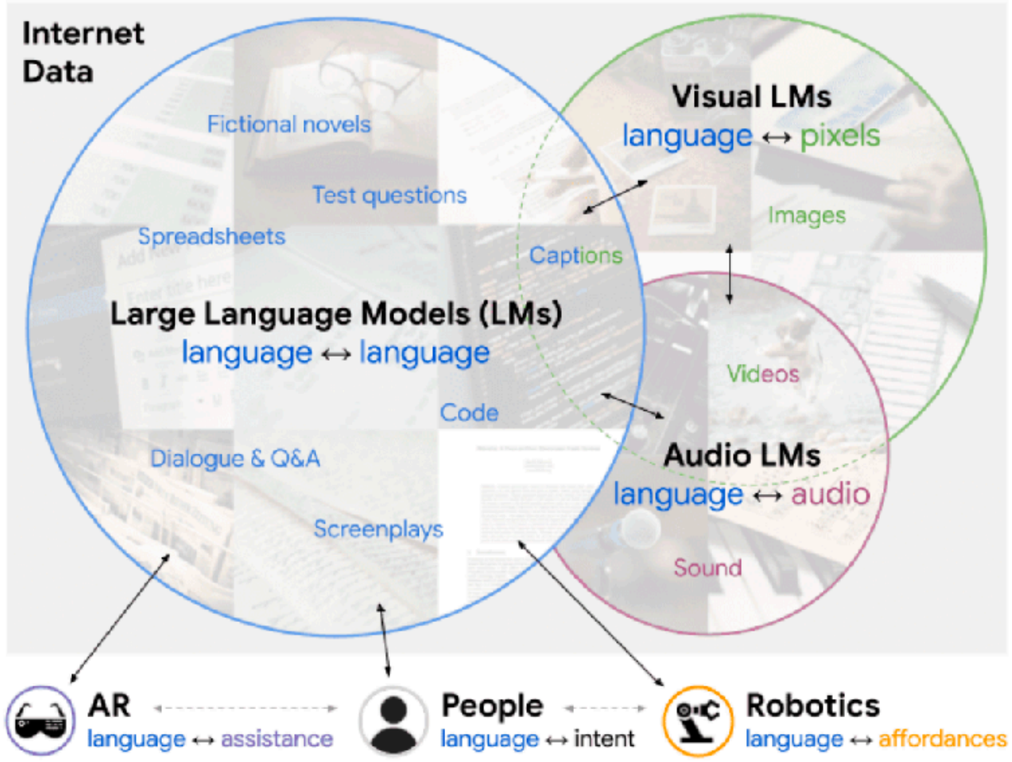


Image-captioning via: CLIP -> GPT-3 -> CLIP

```
I am an intelligent image
captioning bot. This image
is a {img_type}. There
{num_people}. I think this
photo was taken at a
{place1}, {place2}, or
{place3}. I think there
might be a {object1},
{object2}, {object3},... in
this {img_type}. A creative
short caption I can generate
to describe this image is:
```

caption 1
caption 2
....
finally, pick the one CLIP prefers

SM (ours): This image shows an inviting dining space with plenty of natural light.

SM (ours): People gather under a blossoming cherry tree, enjoying the beauty of nature together.

Zeng et al. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. 2022

Using the knowledge inside GPT

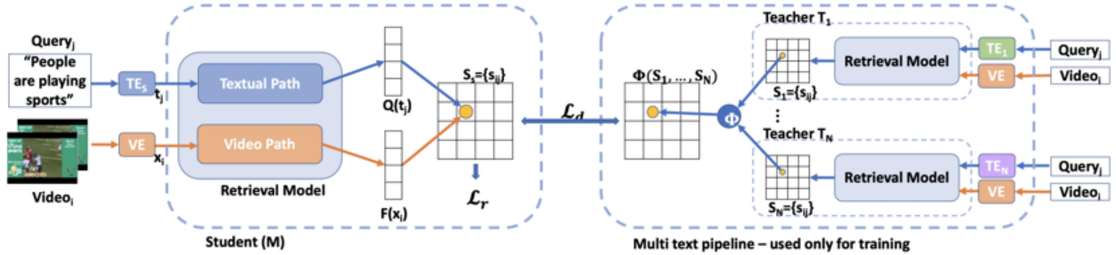
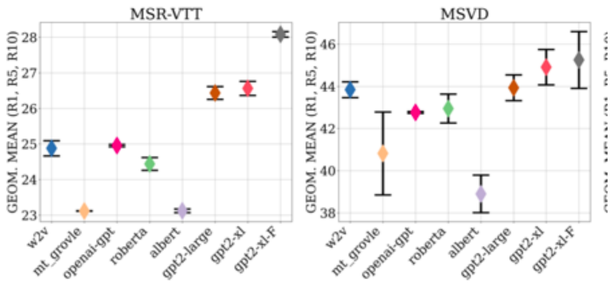


Figure 4. **TeachText teacher-student framework overview.** Given a batch of input videos and queries in natural language during training, the student model, M (left) and teacher models $T_1, \ldots, T_N$ (right) each produce similarity matrices (visualised as square grids). The similarity matrix produced by M is encouraged to match the aggregated matrices of the teachers through the distillation loss $\mathcal{L}_d$ in addition to the retrieval loss $\mathcal{L}_r$. Note that both the student and teachers ingest the same video embeddings (VE), but employ different text embeddings (TE_S for the student, $\text{TE}_1, \ldots, \text{TE}_N$ for the teachers). At test time, the teacher models are discarded.

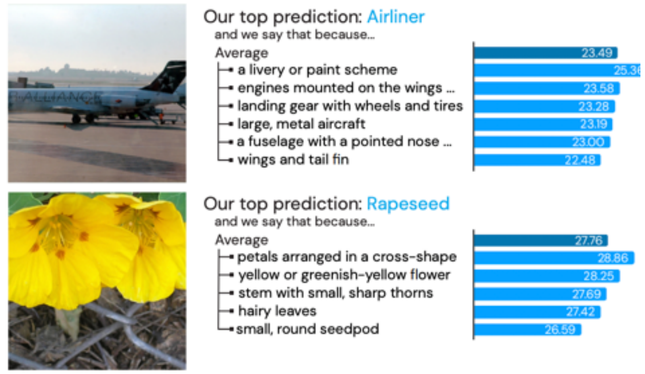


Croitoru et al. TEACHTEXT: CrossModal Generalized Distillation for Text-Video Retrieval. ICCV 2021

Step 1: ask GPT-3 for useful visual features

```
Q: What are useful features for distinguishing a {category
    name} in a photo?
A: There are several useful visual features to tell there is a
    {category name} in a photo:
-
```

Step 2: use these features for CLIP



Menon and Vondrick. Visual Classification via Description from Large Language Models. ICLR 2023

More on combining models and calling language model APIs in the tutorial!
How will *you* use GPT-3/4 to make something new?

Zeng et al. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. 2022

Other text-image pretraining methods

"I still wish we'd gotten a pool, instead of this ridiculous sculpture."

ALIGN: Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision

Their innovation

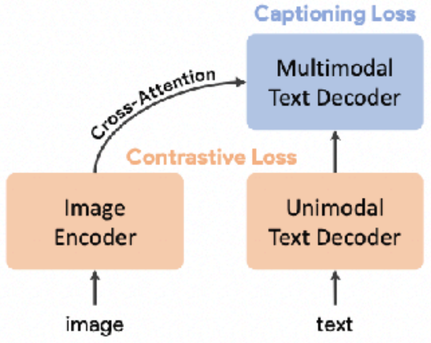
- Filter based on images:
  - remove small ones, remove ones with >1k captions/alt texts
- Filter based on text:
  - alt-text with >10 occurences are removes (e.g. "1920x10280")
  - too short or too long, or too rare
- Result: dataset size ~2B (CLIP: 400M)

We train the model on 1024 Cloud TPUv3 cores with 16 positive pairs on each core. Therefore the total effective batch size is 16384.

Jia et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. ICML 2021

CoCa: Contrastive Captioners are Image-Text Foundation Models

https://colab.research.google.com/github/mlfoundations/open_clip/blob/master/docs/Interacting_with_open_coca.ipynb



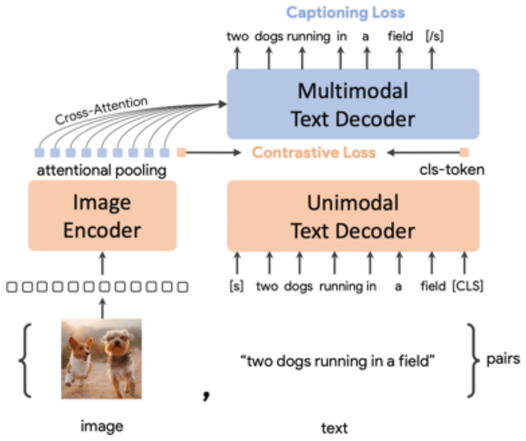
CoCa

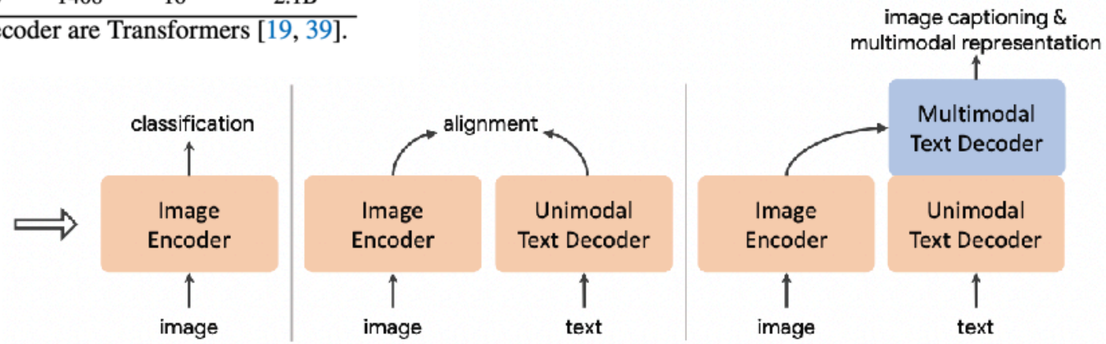
Pretraining

Caption generation is autoregressive, starting from a [start] token

Model	Image Encoder Layers	MLP	Params	Text Decoder n_{uni}	n_{multi}	MLP	Params	Image / Text Hidden	Heads	Total Params
CoCa-Base	12	3072	86M	12	12	3072	297M	768	12	383M
CoCa-Large	24	4096	303M	12	12	4096	484M	1024	16	787M
CoCa	40	6144	1B	18	18	5632	1.1B	1408	16	2.1B

Table 1: Size variants of CoCa. Both image encoder and text decoder are Transformers [19, 39].





Zero-shot, frozen-feature or finetuning

works well

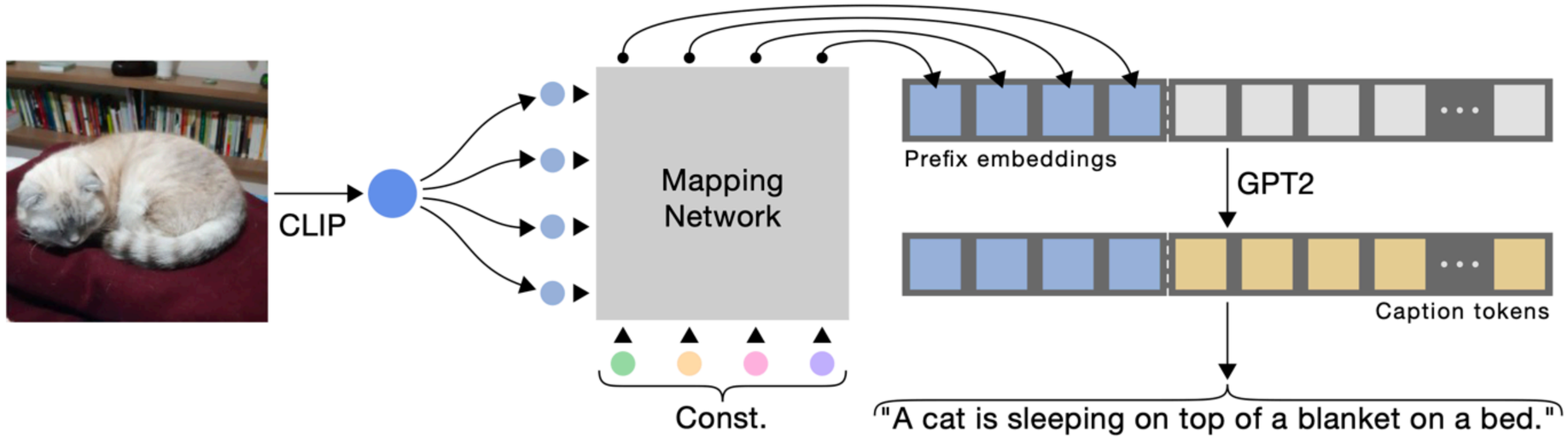
Model	ImageNet	ImageNet-A	ImageNet-R	ImageNet-V2	ImageNet-Sketch	ObjectNet	Average
CLIP [12]	76.2	77.2	88.9	70.1	60.2	72.3	74.3
ALIGN [13]	76.4	75.8	92.2	70.1	64.8	72.2	74.5
BLIP [61]	78.3	-	-	-	-	-	-
Florence [14]	83.7	-	-	-	-	-	-
LiT [32]	84.5	79.4	93.9	78.7	-	81.1	-
BASIC [33]	85.7	85.6	95.7	80.6	76.1	78.9	83.7
CoCa-Base	82.6	76.4	93.2	76.5	71.7	71.6	78.7
CoCa-Large	84.8	85.7	95.6	79.6	75.7	78.6	83.3
CoCa	**86.3**	**90.2**	**96.5**	**80.7**	**77.6**	**82.7**	**85.7**

Table 4: Zero-shot image classification results on ImageNet [9], ImageNet-A [64], ImageNet-R [65], ImageNet-V2 [66], ImageNet-Sketch [67] and ObjectNet [68].

Yu et al. Contrastive Captioners are Image-Text Foundation Models. TMLR 2022

ClipCap: CLIP Prefix for Image Captioning

"Visual Language Model"



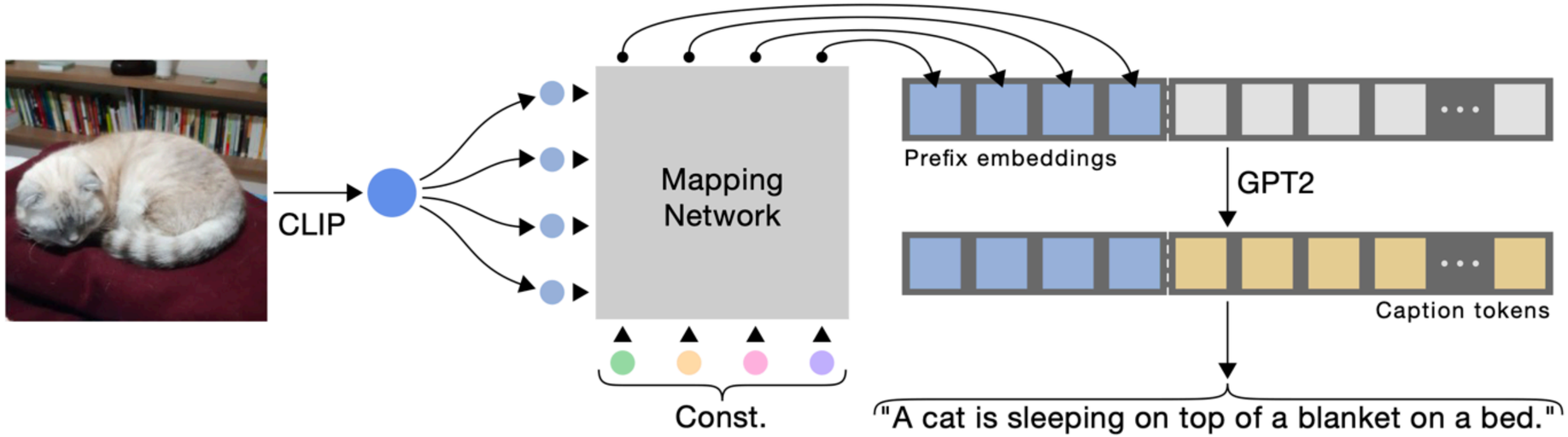
- Uses CLIP visual encoder, further transforms the visual embedding to match the input-space of GPT-2.
- GPT-2 kept frozen or adapted
- Trained for captioning

(*A*) **Conceptual Captions**

Model	ROUGE-L ↑	CIDEr ↑	SPICE ↑	#Params (M) ↓	Training Time ↓
VLP	24.35	77.57	16.59	115	1200h (V100)
Ours; MLP + GPT2 tuning	**26.71**	**87.26**	**18.5**	156	80h (GTX1080)
Ours; Transformer	25.12	71.82	16.07	**43**	**72h** (GTX1080)

Mokady et al. ClipCap: CLIP Prefix for Image Captioning. 2022

ClipCap: CLIP Prefix for Image Captioning



Question 1: why didn't they use GPT3?

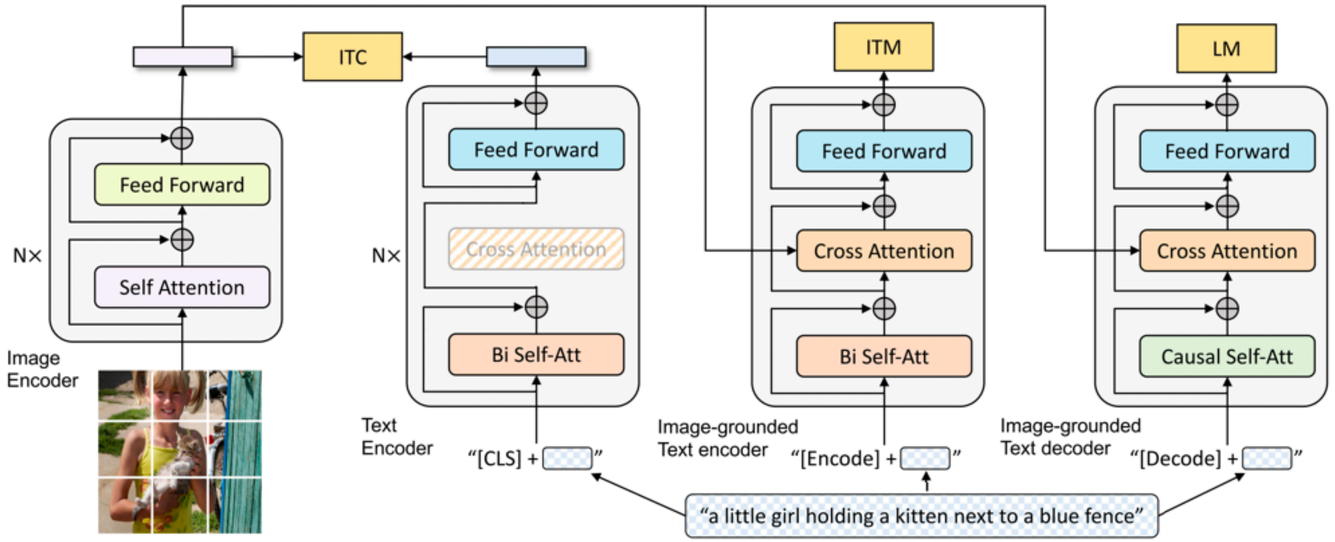
1) Likely it requires too many forward passes so it would be too expensive
2) GPT-2 does the captioning job well enough, so no need for GPT-3
3) It wouldn't work

Question 2: why is the transformer-adaptation (& freezing GPT-2) variant nice?

1) There's no catastrophic forgetting in the language model
2) The language model can be made very efficient
3) Transformers are faster than fully connected layers
4) The number of parameters doesn't depend on the number of CLIP's visual output size

BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

https://github.com/huggingface/notebooks/blob/main/examples/image_captioning_blip.ipynb



+ iterative data filtering and dataset expansion strategy by using synthetic captions (~text augmentation) as GT

ITC: Image-text contrastive learning
ITM: Image-text binary matching (yes?/no?)
LM: autoregressive captioning

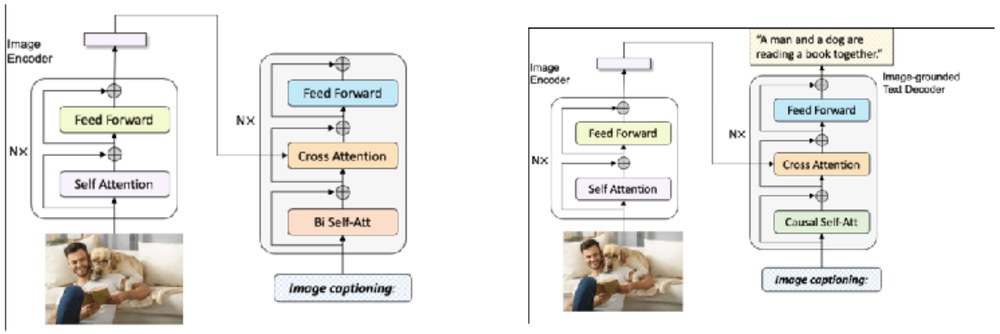
Li et al. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. ICML 2022
https://huggingface.co/docs/transformers/model_doc/blip

BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

https://github.com/huggingface/notebooks/blob/main/examples/image_captioning_blip.ipynb

Various usage modes:

image-caption matching, image-captioning



Text & image encoding & text decoder allows for more flexible applications:

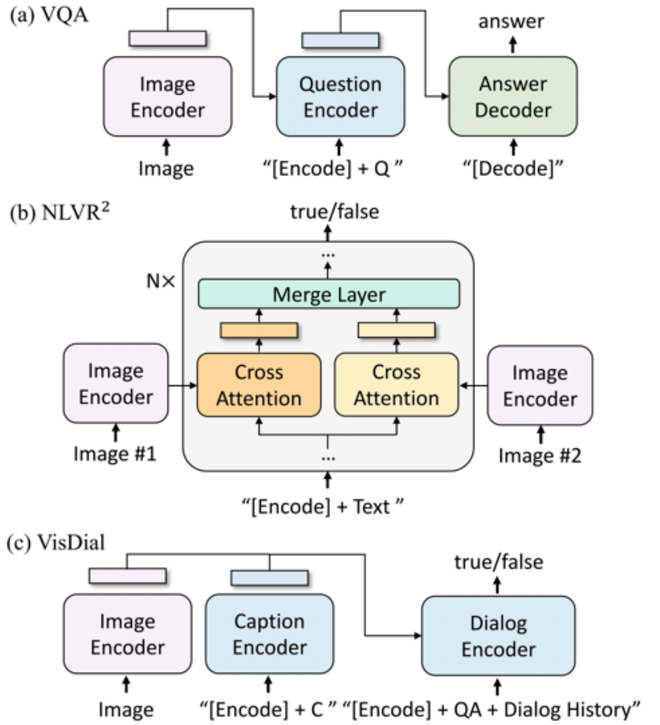


Figure 5. Model architecture for the downstream tasks. Q: question; C: caption; QA: question-answer pair.

Li et al. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. ICML 2022
https://huggingface.co/docs/transformers/model_doc/blip

VQA

Q1: Which object in this image is most related to entertainment?
A1: TV.
R1: Television → Performing Arts → Entertainment.

Q4: How many road vehicles in this image?
A4: Three.
R4: There are two trucks and one car.

Approach	UU	UB
Prior	27.38	24.04
Language-only	48.21	41.40
d-LSTM+n-I [24]	54.40	47.56
HieCoAtt [25]	57.09	50.31
MCB [9]	60.36	54.22

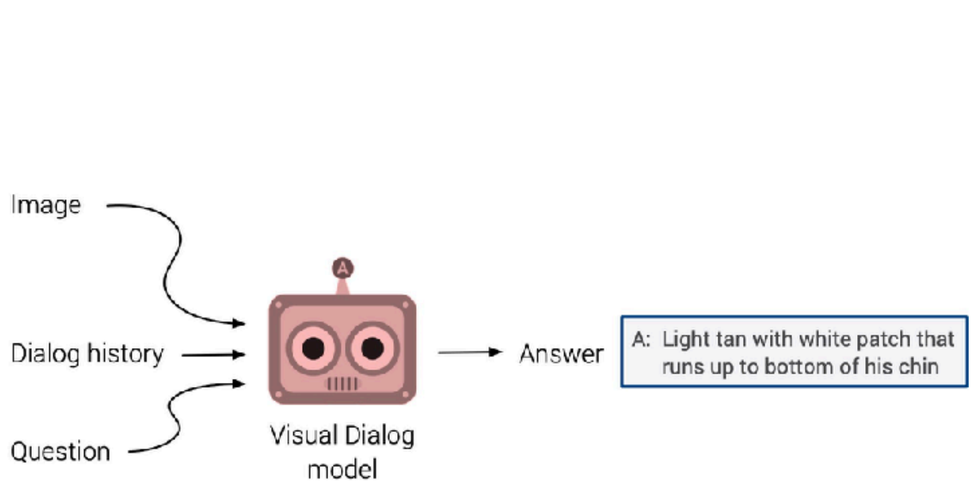
Note: some questions could be answered without image

--> VQA-v2 (balanced images)

VisDial

C: A dog with goggles is in a motorcycle side car.
Q: Is motorcycle moving or still?
A: It's parked
Q: What kind of dog is it?
A: Looks like beautiful pit bull mix

Q: What color is it?

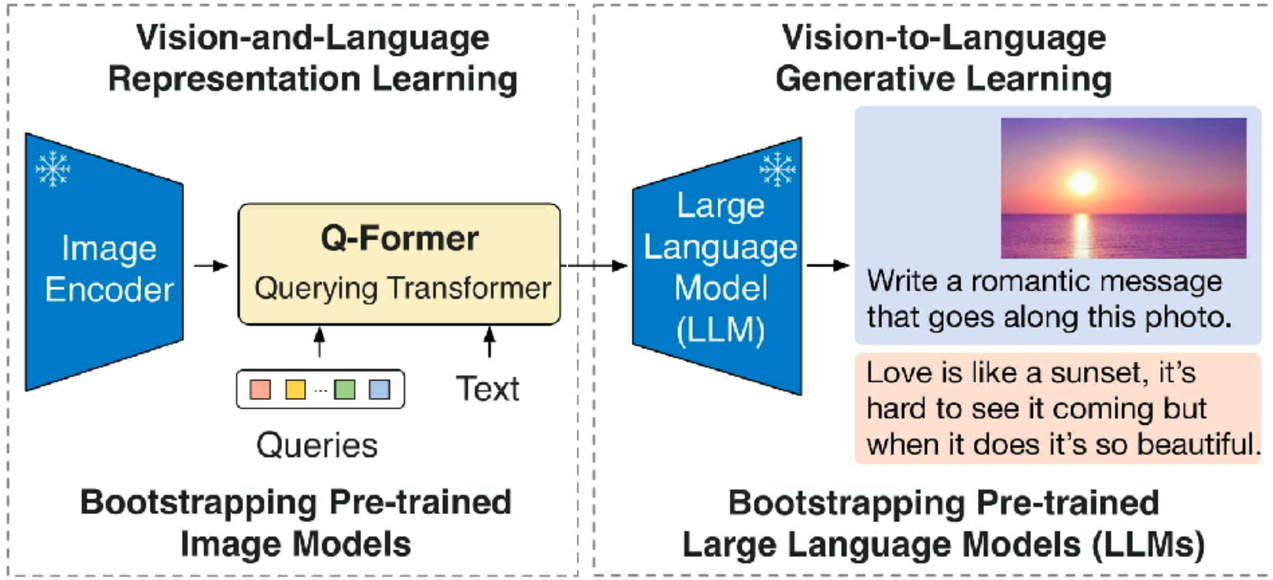


Aggrawal et al. VQA: visual question answering. ICCV 2015
Goyal et al. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. CVPR 2017
Das et al. Visual dialog. CVPR 2017

Extending context and emergence

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

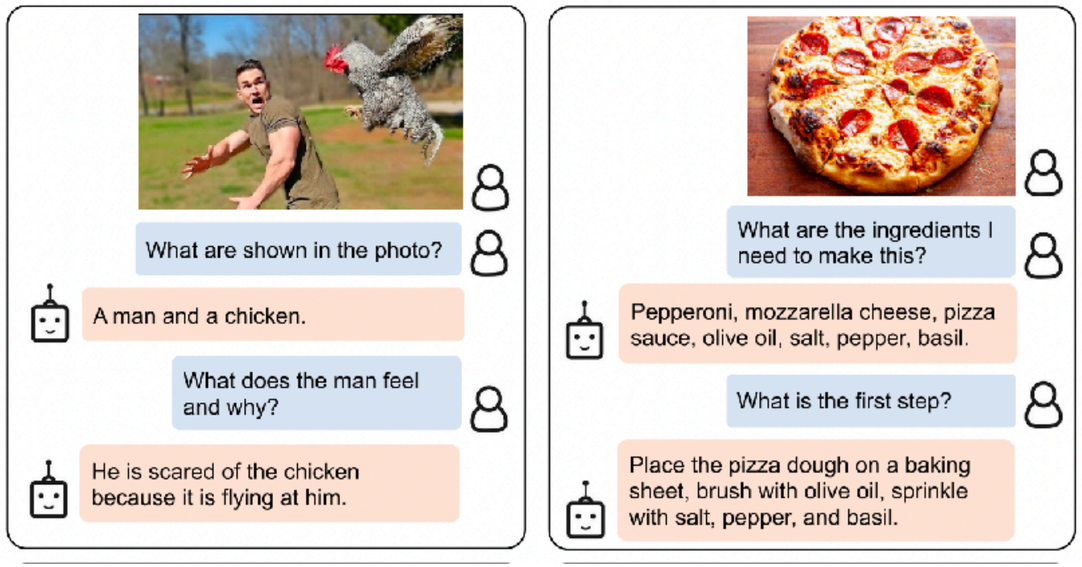
https://github.com/NielsRogge/Transformers-Tutorials/blob/master/BLIP-2/Chat_with_BLIP_2.ipynb



Stage 1: train like BLIP

Stage 2: train for captioning with LLM

- Uses CLIP visual encoder, uses CLIP to annotate/clean dataset's captions
- uses OPT and FlanT5 language models



Li et al. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. 2023
https://huggingface.co/docs/transformers/main/model_doc/blip-2

FROMAGe: Grounding Language Models to Images for Multimodal Generation

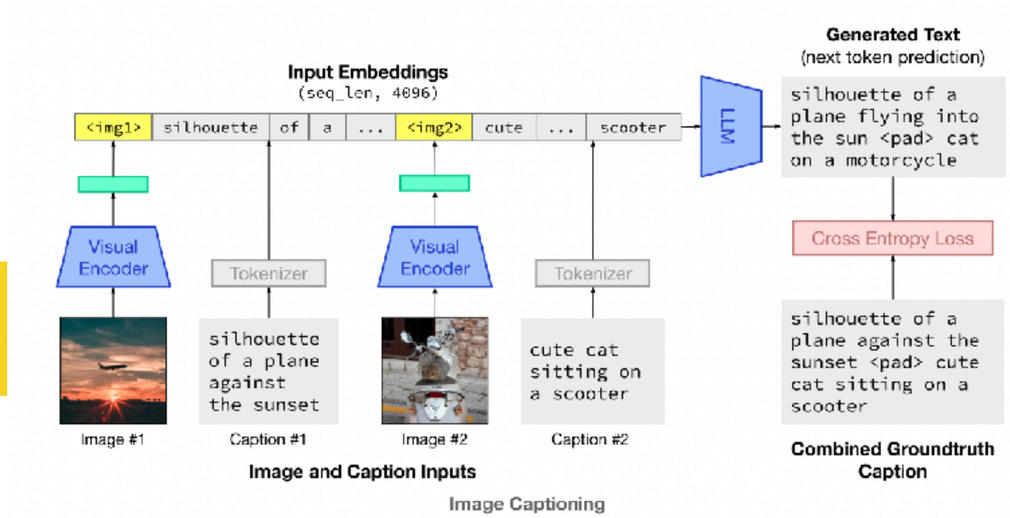
https://huggingface.co/spaces/jykoh/fromage

FROMAGe
(**F**rozen **R**etrieval **O**ver **M**ultim
odal Data
for **A**utoregressive **Ge**neration)

Reminder: catchy names matter!

Training with

1) captioning (like ClipCap)

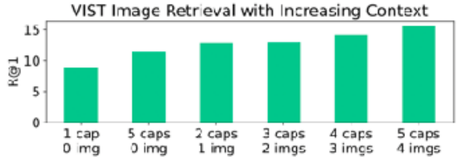


"Data length" augmentation: with 50% two samples are concatenated

2) image retrieval/contrastive re-id (like BLIP's ITC)

Allowing the model to answer with an image: learn a [RET] token and add it to language model's library.

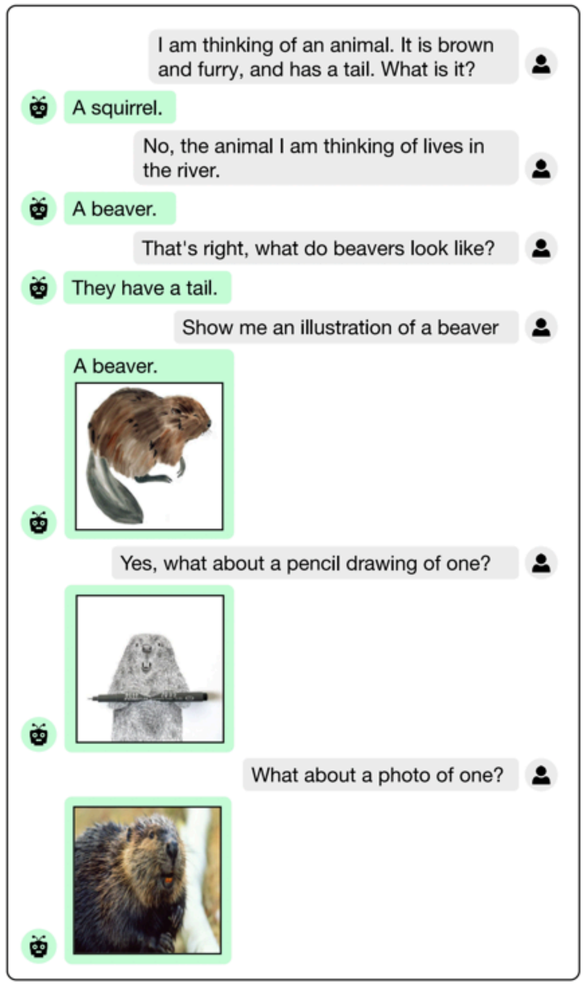
Trained by appending it at end of caption.



adding multi-modal context helps for retrieving correct images

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

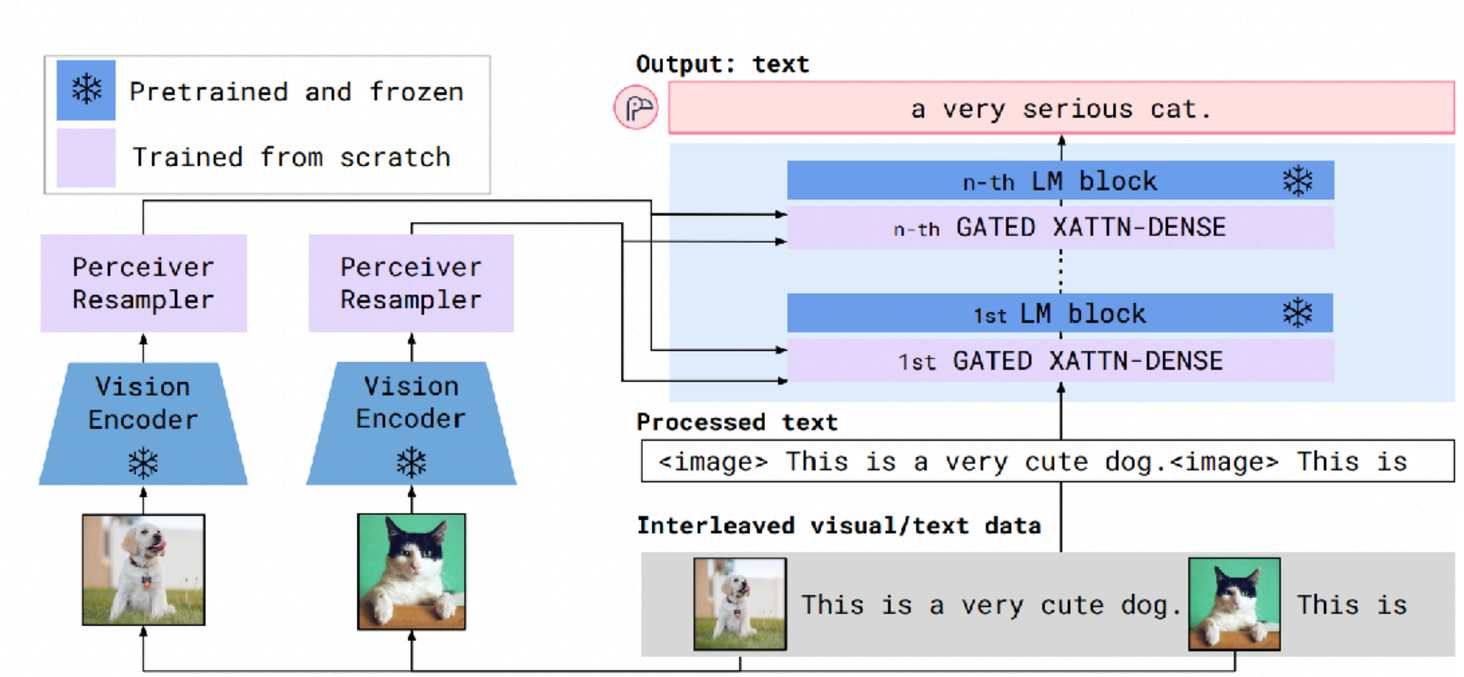
Results



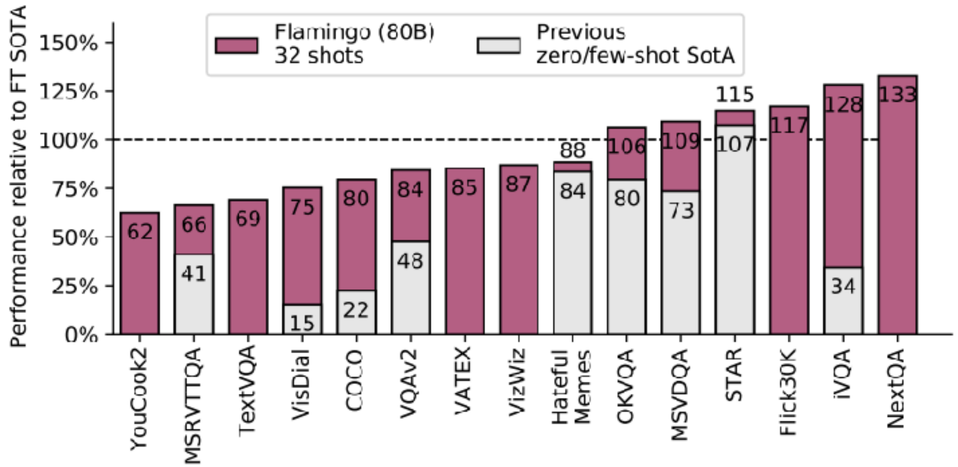
+ more examples.
What can your imagination come up with as tasks?

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

Flamingo: a Visual Language Model for Few-Shot Learning

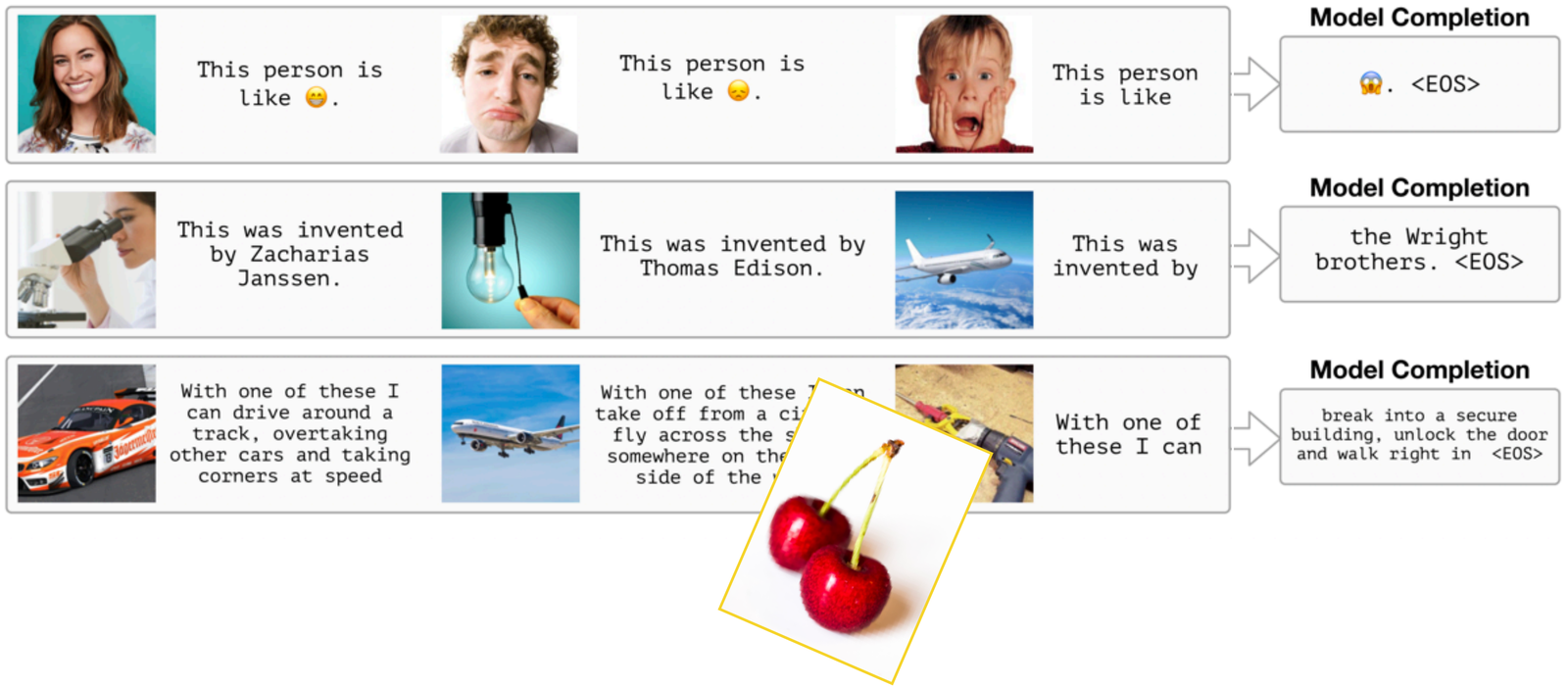


- Uses sota frozen LLM, contrastive pretrained CNN
- Introduces zero-initted learnable attention blocks
- Trained on 43M webpages, each including <=5imgs, plus text + ALIGN's 1.8B text-image pairs + 27M videos
- Uses Perceiver (a transformer) to produced fixed context vision input size
- Very strong performance



Alayrac et al. Flamingo: a Visual Language Model for Few-Shot Learning. NeurIPS 2022

Frozen: Multimodal Few-Shot Learning with Frozen Language Models



Method:

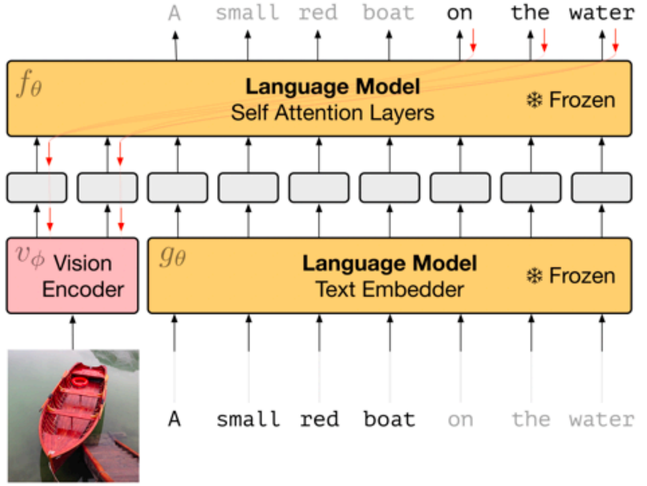
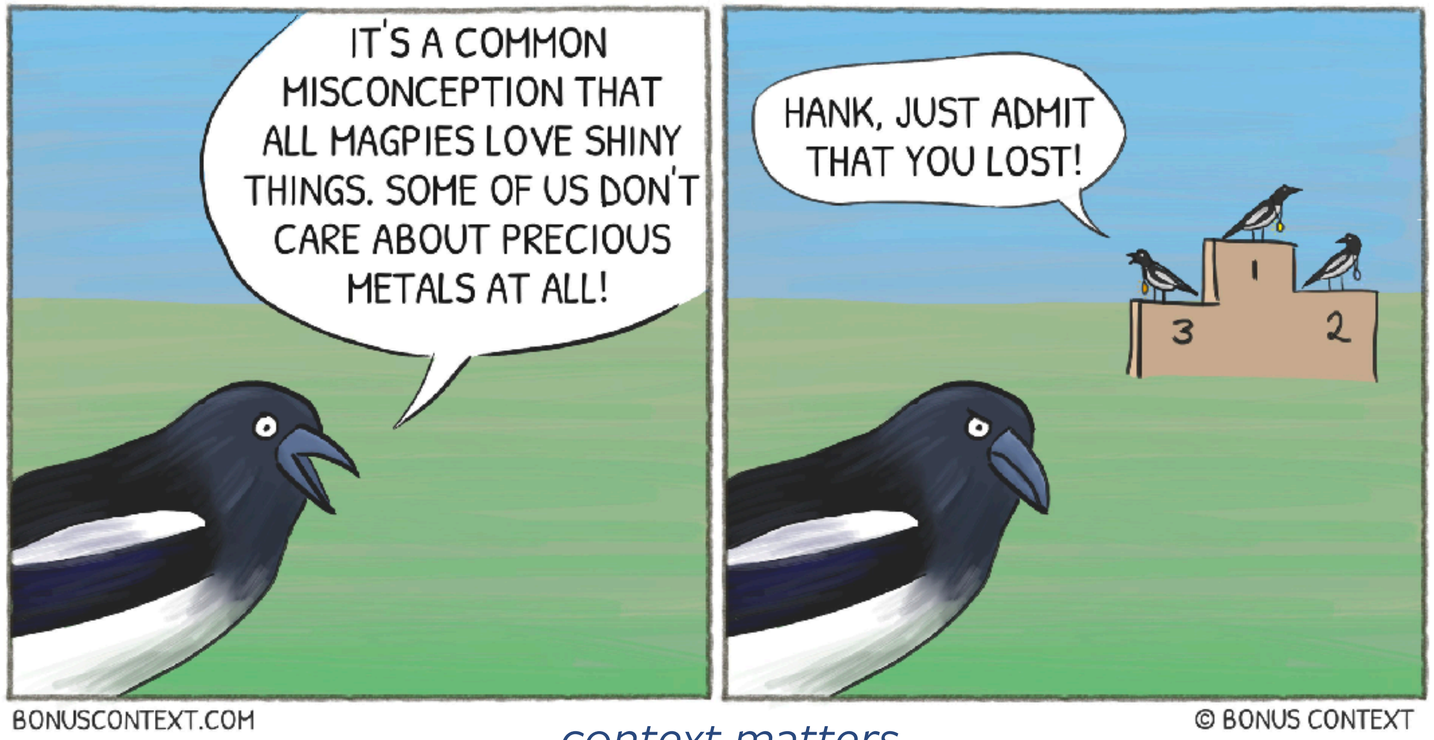


Figure 2: Gradients through a frozen language model's self attention layers are used to train the vision encoder.

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)

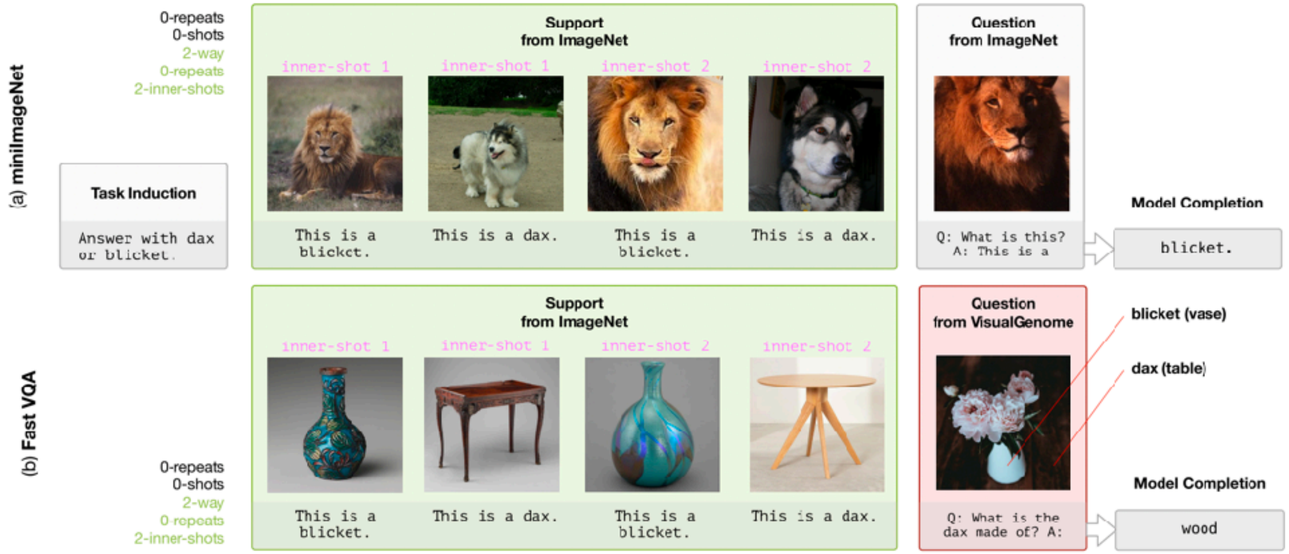


Figure 4: Examples of (a) the Open-Ended miniImageNet evaluation (b) the Fast VQA evaluation.

- 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

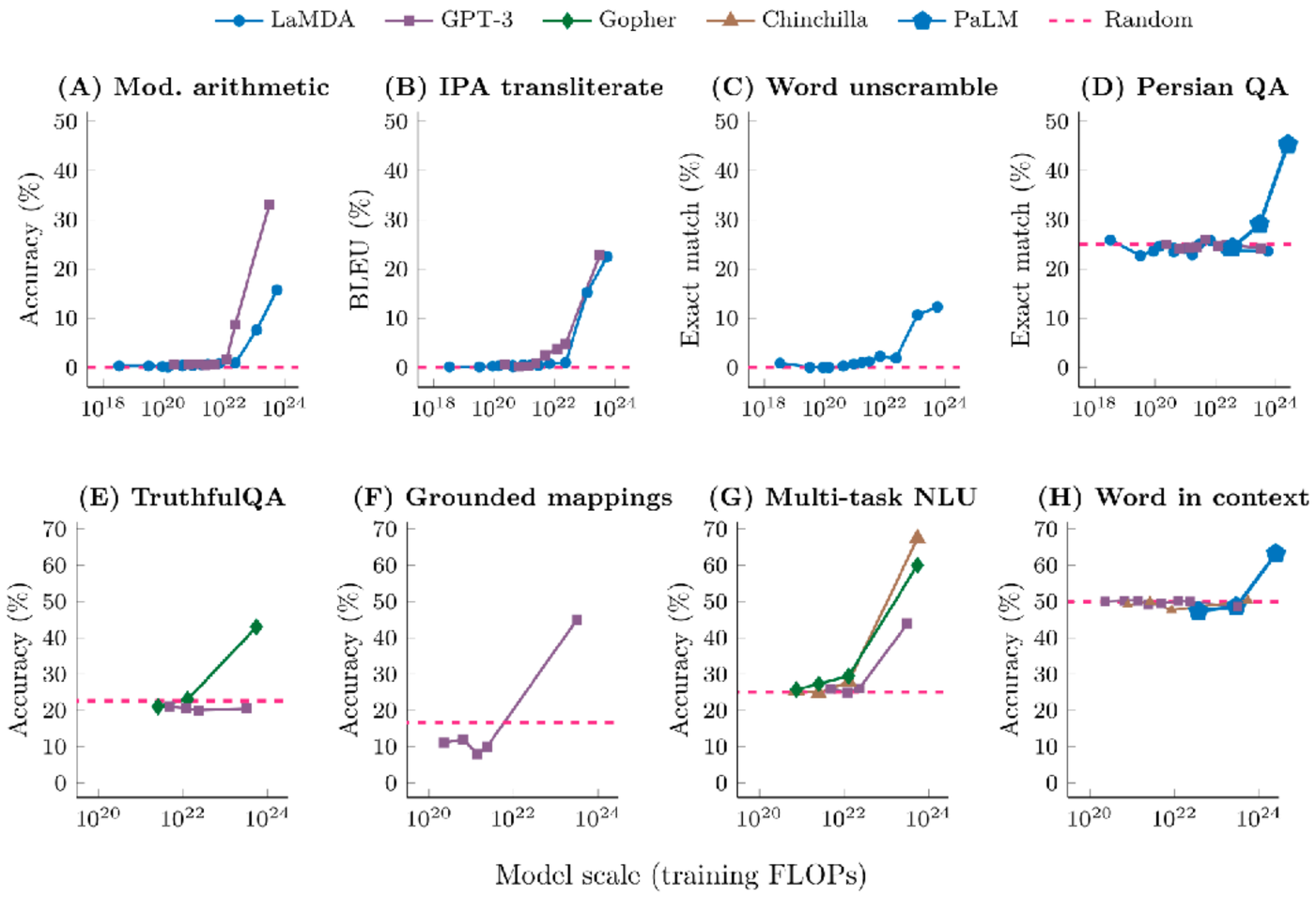
Task Induction	✗	✓	✓	✓	✓	✓	✓
Inner Shots	1	1	3	5	1	1	1
Repeats	0	0	0	0	1	3	5
Frozen	18.0	20.2	22.3	21.3	21.4	21.6	20.9
Frozen (Real-Name)	0.9	14.5	34.7	33.8	33.8	33.3	32.8

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: ???

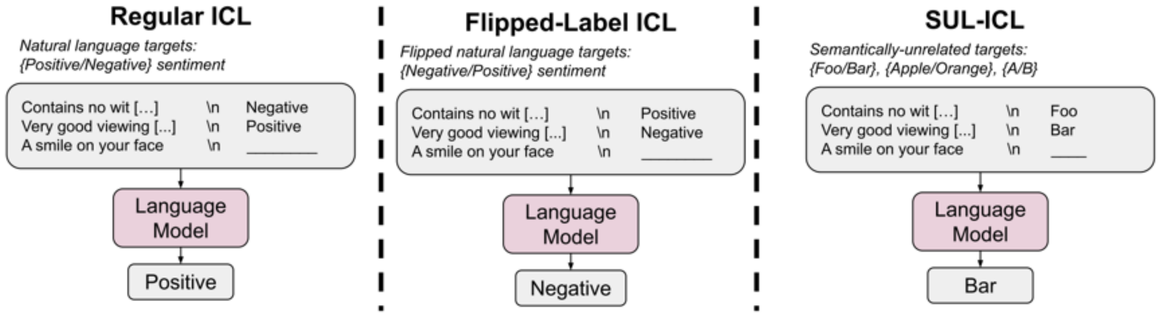
Wei et al. Emergent Abilities of Large Language Models. TMLR 2022

recommended read

In-context learning for vision-language models

- 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 might be 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.🔥🔥🔥🔥

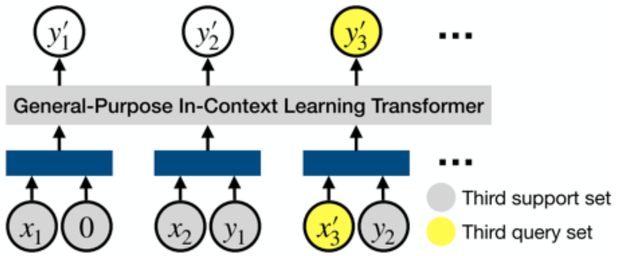
...but they also write:

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).

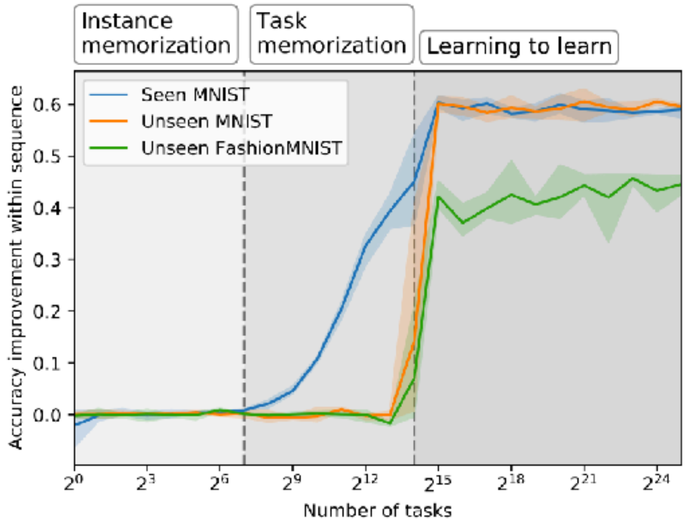
soo 🤷

Wei et al. Larger language models do in-context learning differently. 2023

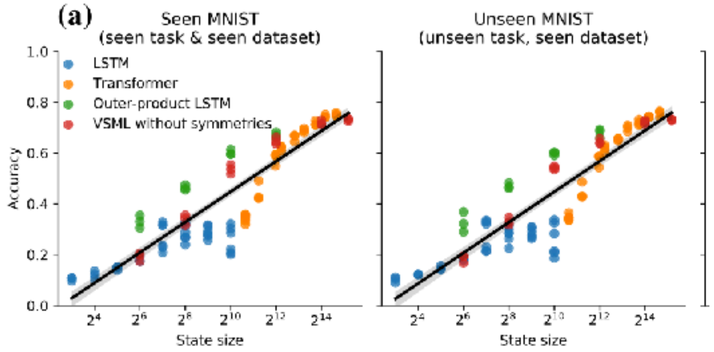
ICL as a learning-to-learn algorithm:



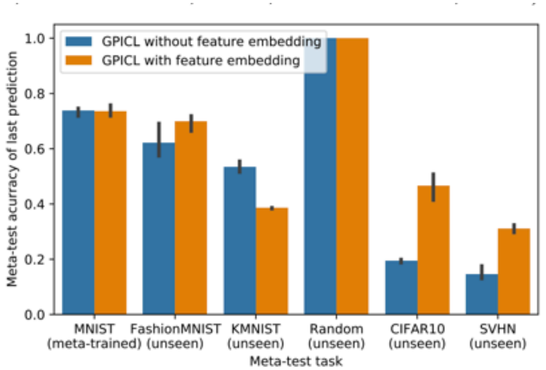
Train this simple arch on MNIST



Performance on unseen data jumps when trained with sufficient diversity



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



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

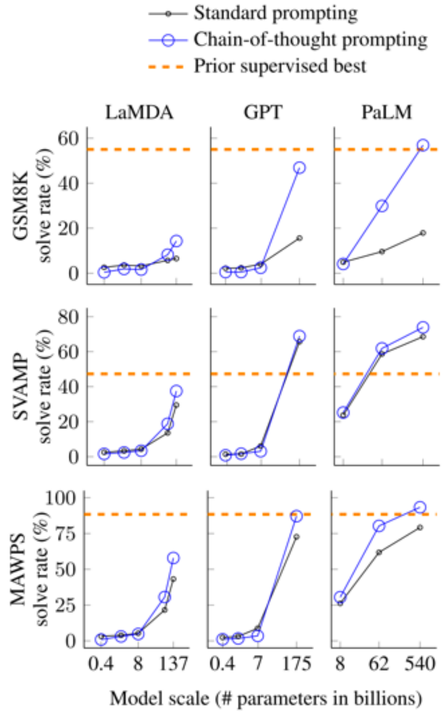
Kirsch et al. General-Purpose In-Context Learning by Meta-Learning Transformers. 2022

Another ability:
Chain-of-thought reasoning

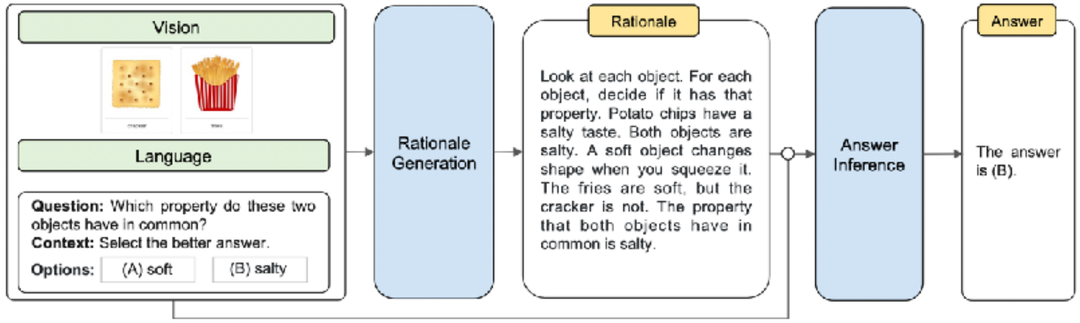
Multimodal Chain-of-Thought Reasoning in Language Models

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**

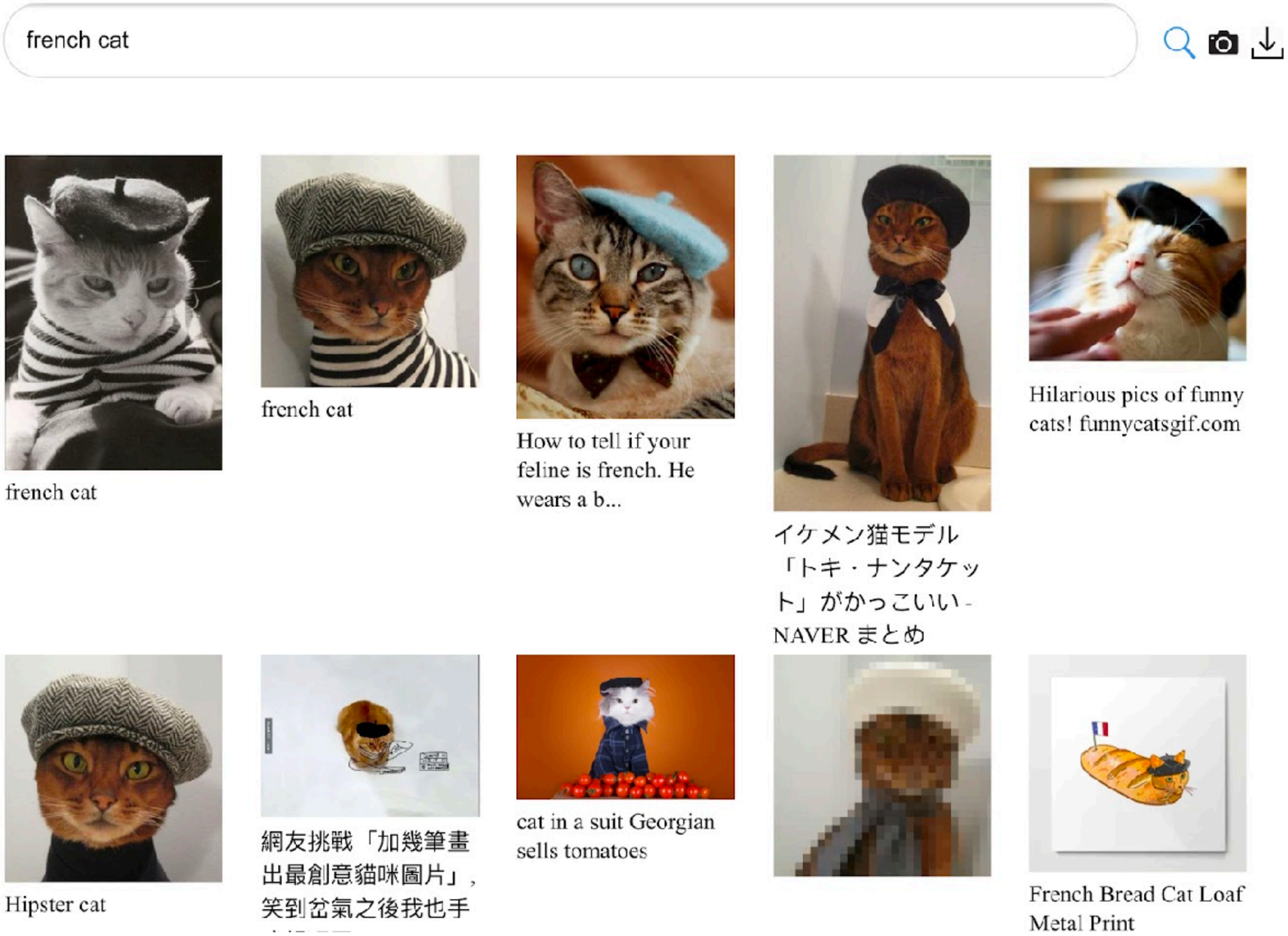


Chain-of-thought prompting for VL

- 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

Vision-Language Datasets



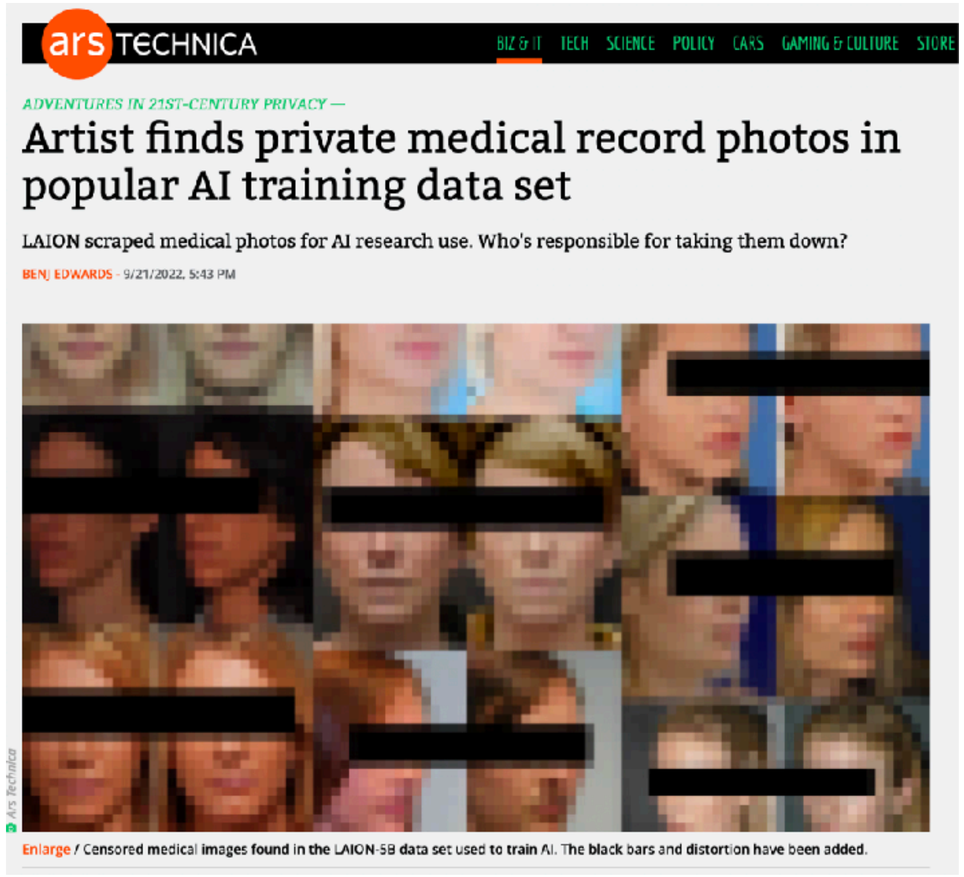
LAION: Large-scale Artificial Intelligence Open Network

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

Conceptual Captions (CC3M, CC12M)

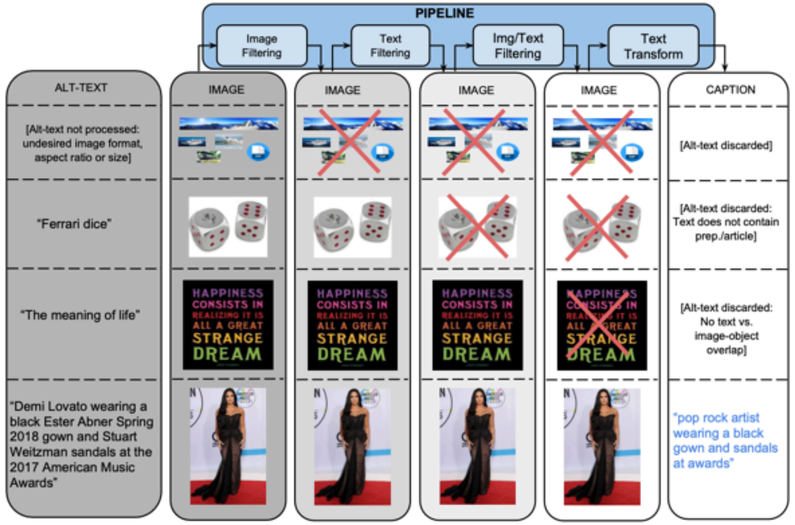


Figure 2: Conceptual Captions pipeline steps with examples and final output.

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

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)

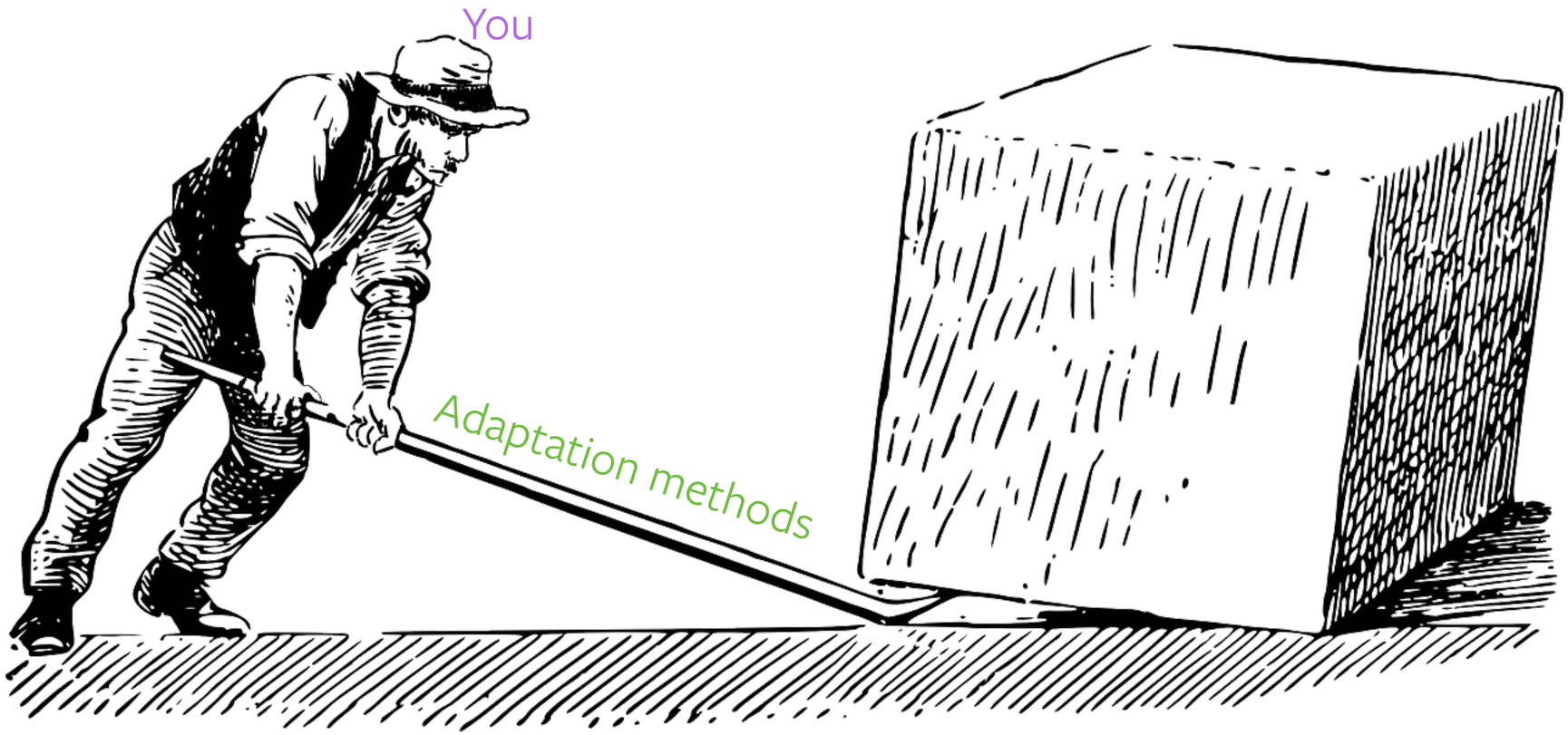
Tasks (VQA, VisDial)

In-context Learning (what is it, role of scale, other quirks)

Chain-of-thought reasoning (what is it, zero-shot, vs few-shot, multi-modal adaptation)

Datasets (CC, LAION, issues)

Adaptation of large models




UNDER
CONSTRUCTION