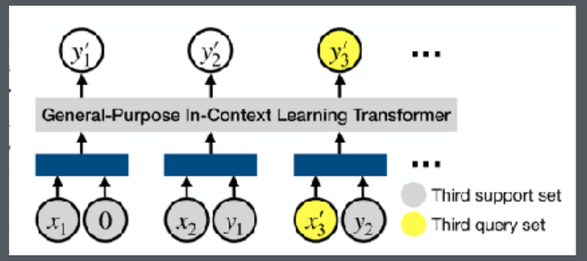
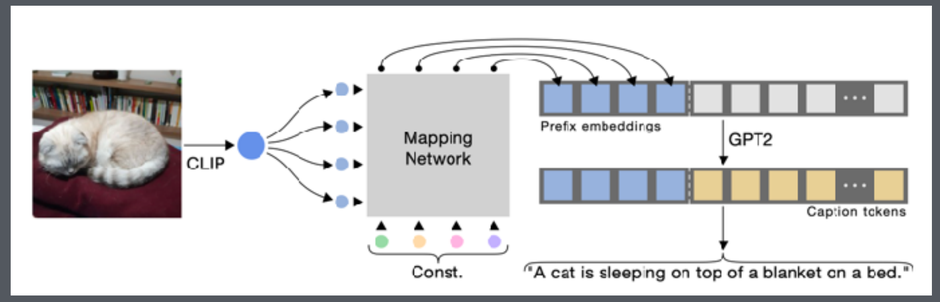
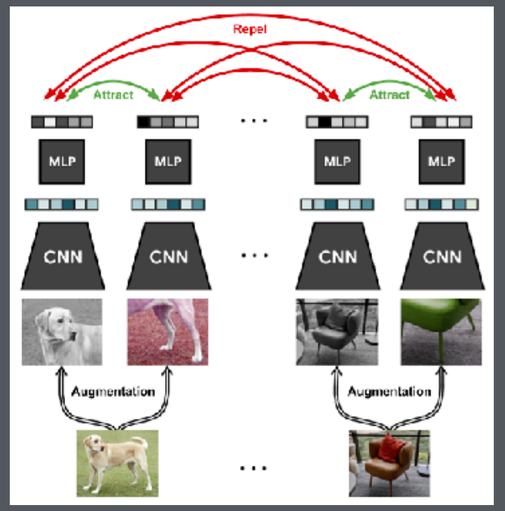
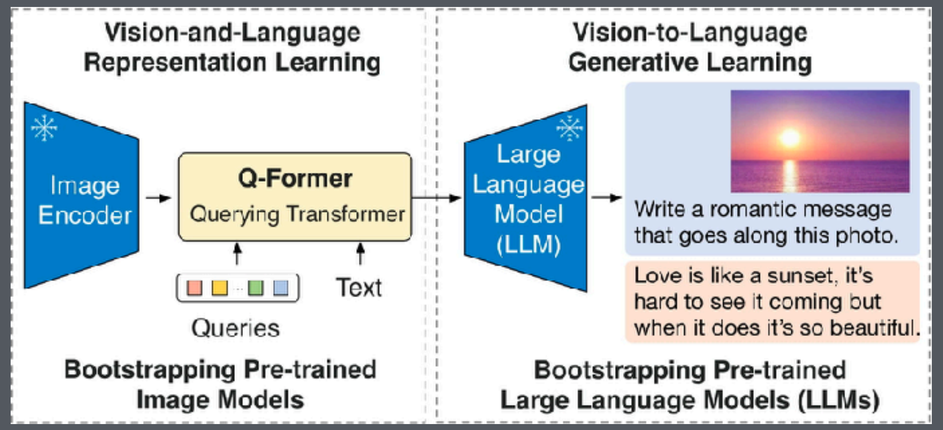
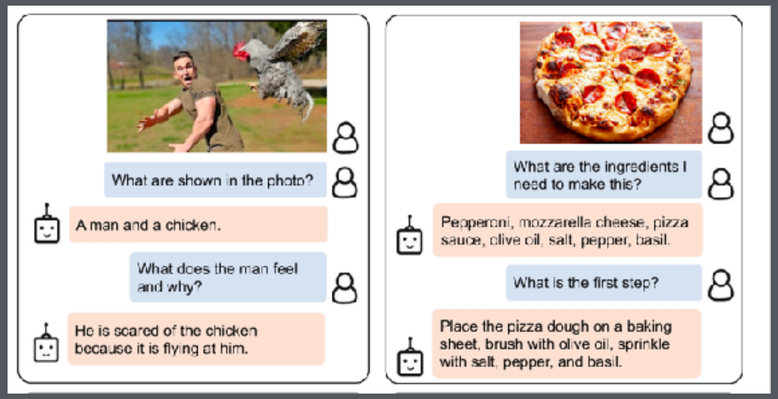
Self-supervised and vision-language learning











@ DEEP LEARNING 2 – LECTURE (1/2)

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

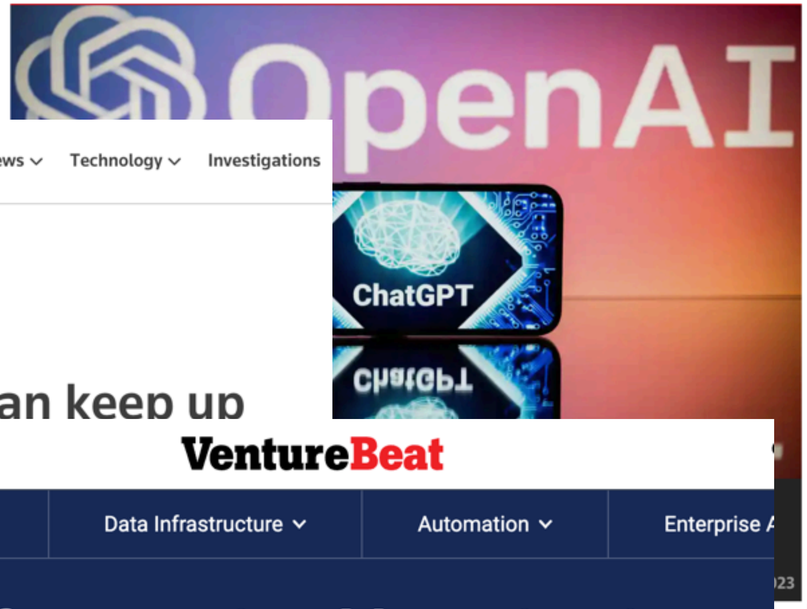
Support us

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
- In-context learning
- Chain-of-thought

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.

Organisational note:

next lecture: 25th April: 9-11am

Tutorial: 25th April 11am-1pm

Meetings with Ivona/Mohammad/Pengwan (the TAs):
on 18th April via zoom, more meetings on-demand

Self-supervised Pretraining

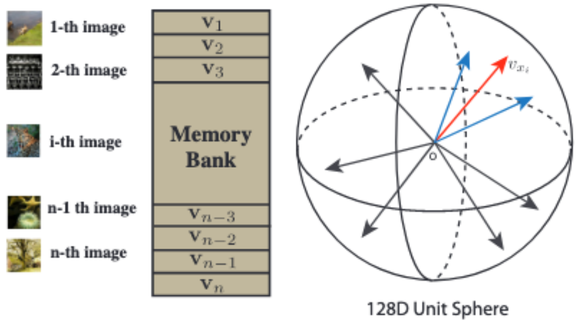
Q: How does one learn without labels?

A: Need to generate a loss that provides gradients.

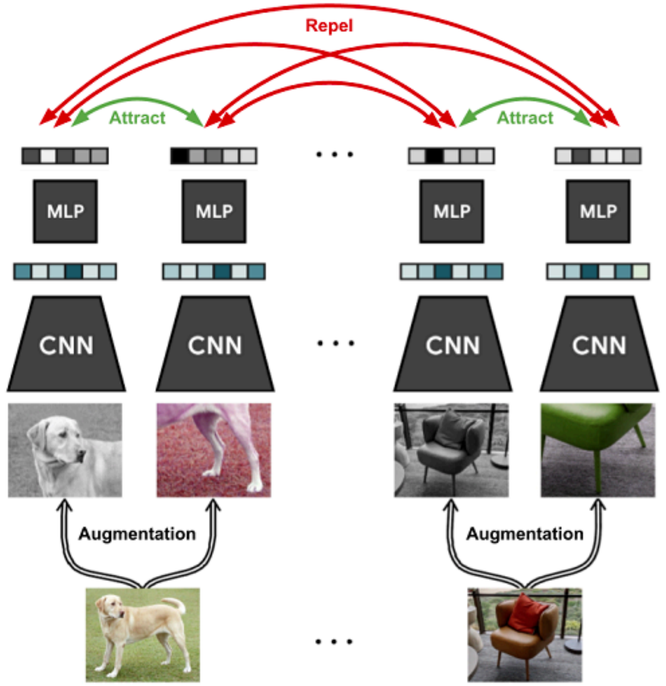
We will focus on signals from

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

Modern Noise-contrastive self-supervised learning



NPID

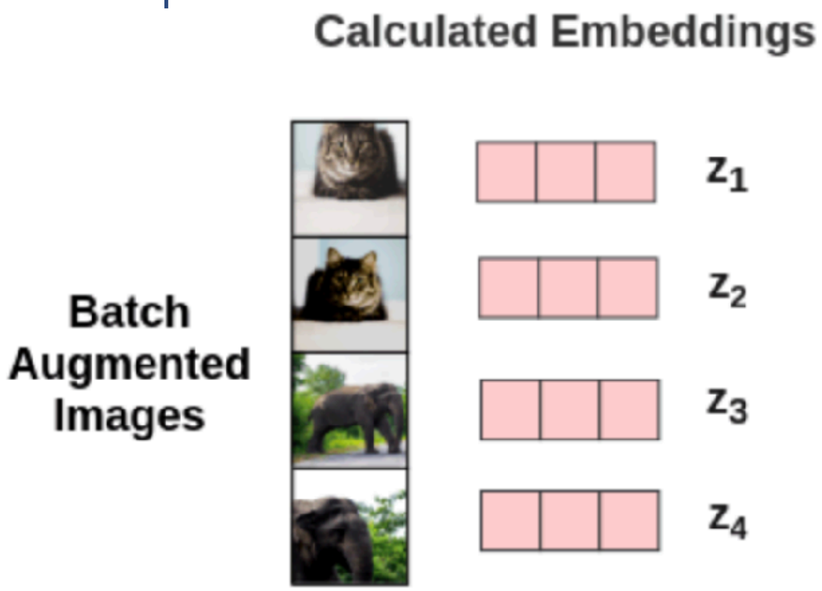


SimCLR

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

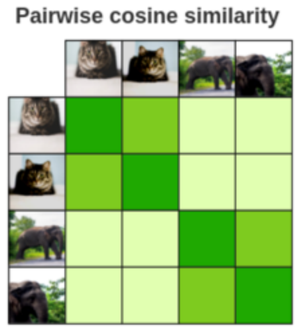
Similarity Calculation of Augmented Images

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

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

- 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

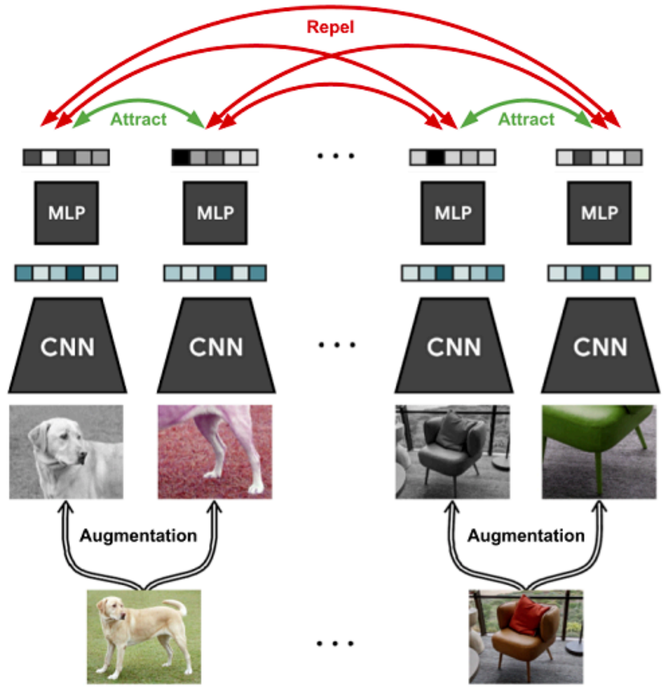


Loss: *relatively* increase similarity for pairs, decrease rest

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

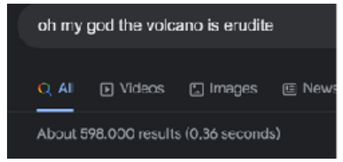
https://thegradient.pub/understanding-evaluation-metrics-for-language-models/

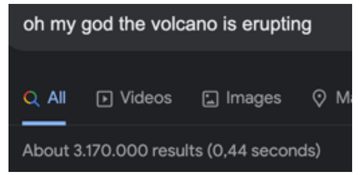
Language Modelling via next-word prediction

Why "erudite" is not a good guess

Factor the probability of a datapoint (w_1,..., w_n):

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





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)

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

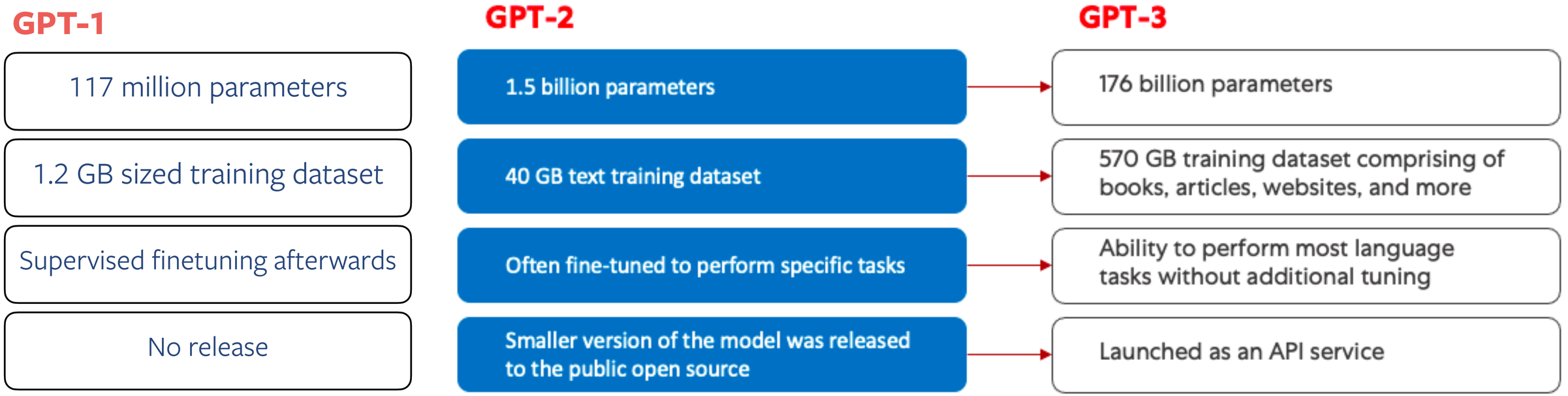
$$h_0 = UW_e + W_p$$
$$h_l = \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n]$$
$$P(u) = \texttt{softmax}(h_n W_e^T)$$

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



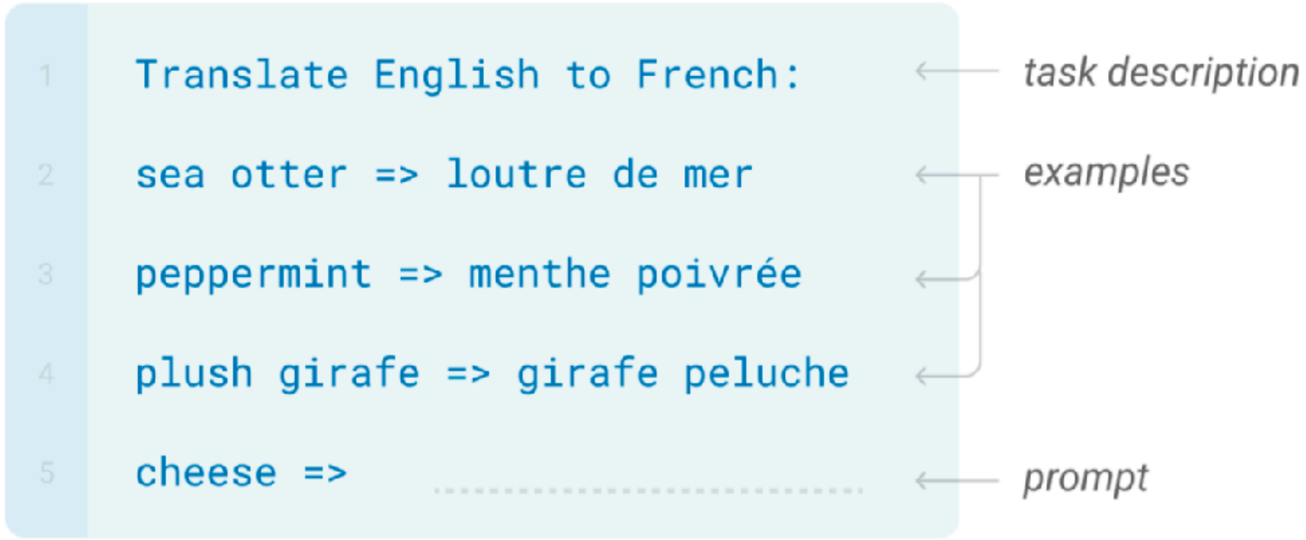
https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286

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.

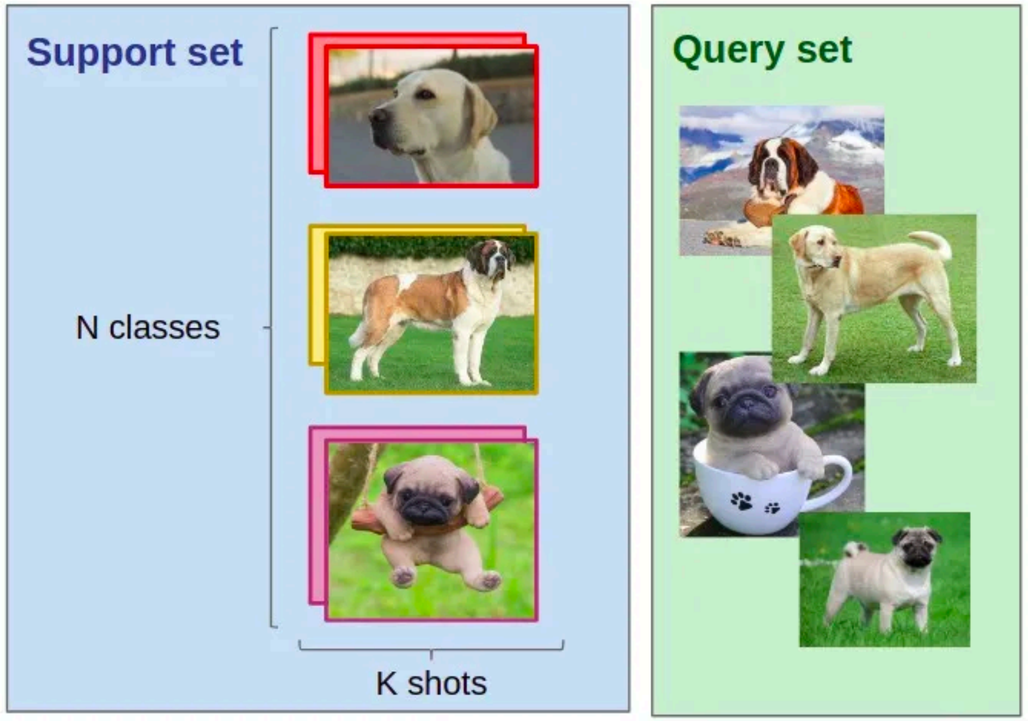


One emergent capability of large language models is *in-context learning*.

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.

[Compare to "usual" few-shot learning in e.g. vision]



When classifying between N classes: "N-way"
With K examples: "K-shot"

e.g. 5-way 5-shot.

The shots and classes that are tested need to be deterministic in the test set

Previous methods often rely on pretraining for this objective, extract embeddings & find neared neighbors or finetune models.

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

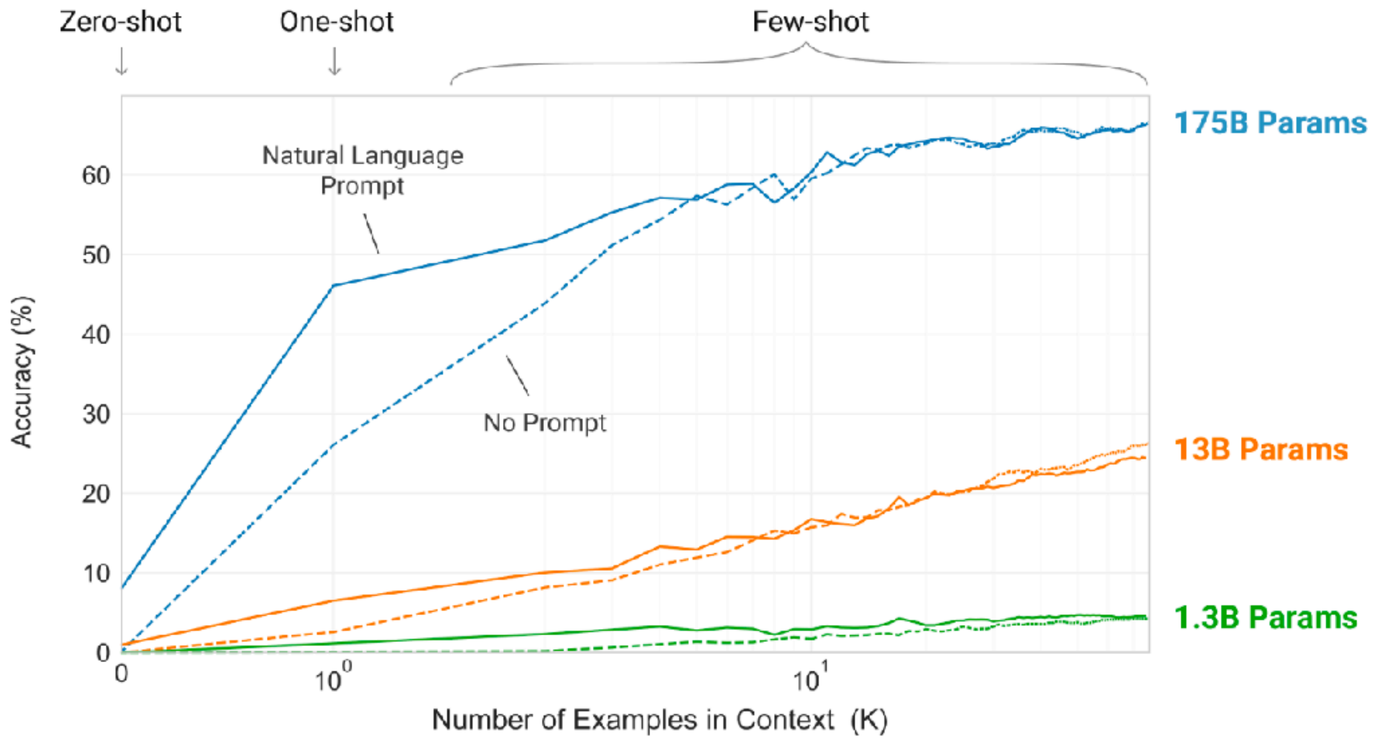
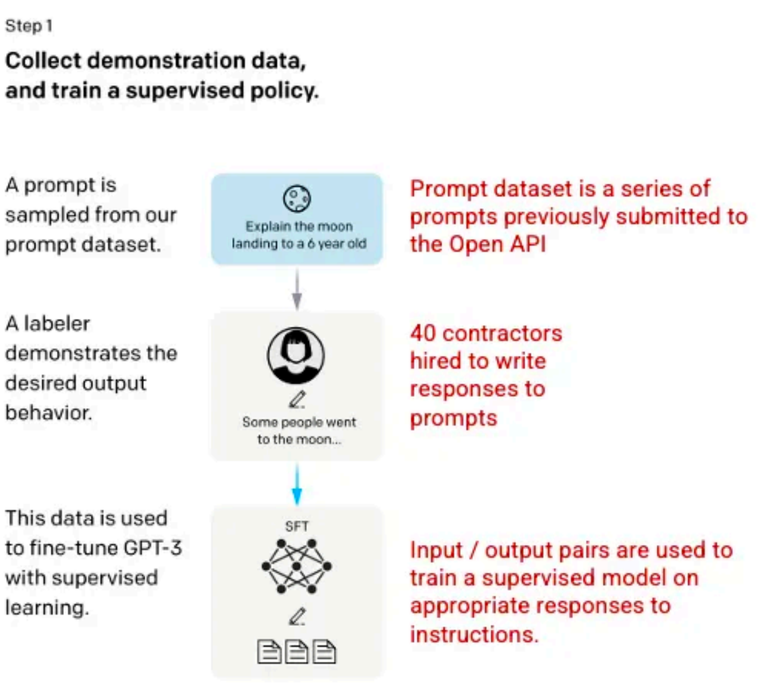
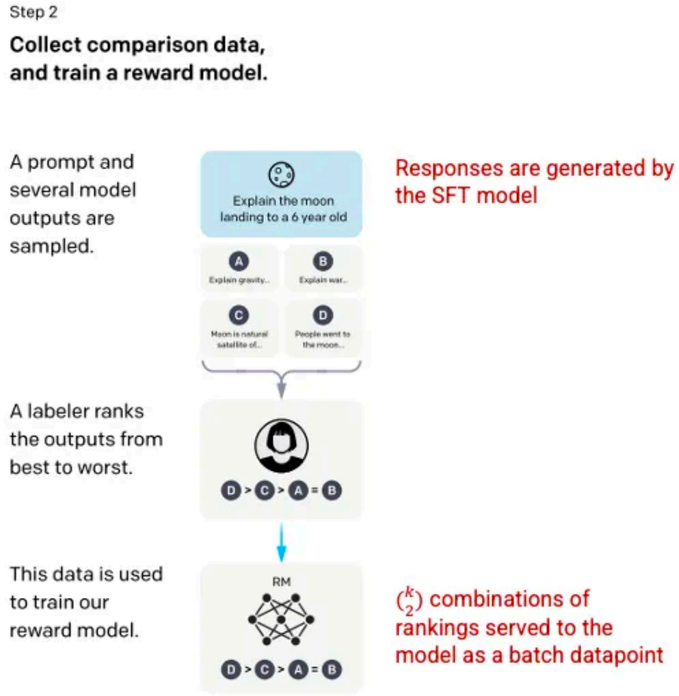


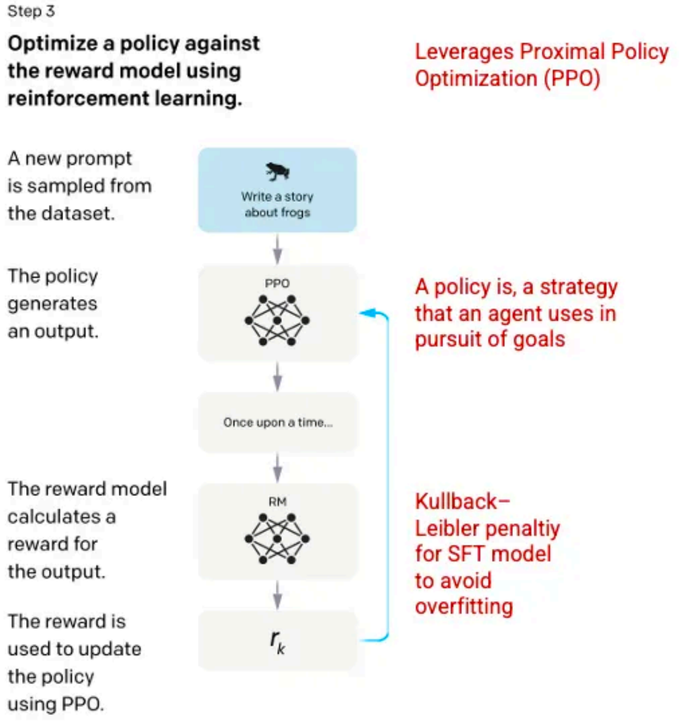
Figure 1.2: Larger models make increasingly efficient use of in-context information. We show in-context learning performance on a simple task requiring the model to remove random symbols from a word, both with and without a natural language task description (see Sec. 3.9.2). The steeper "in-context learning 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

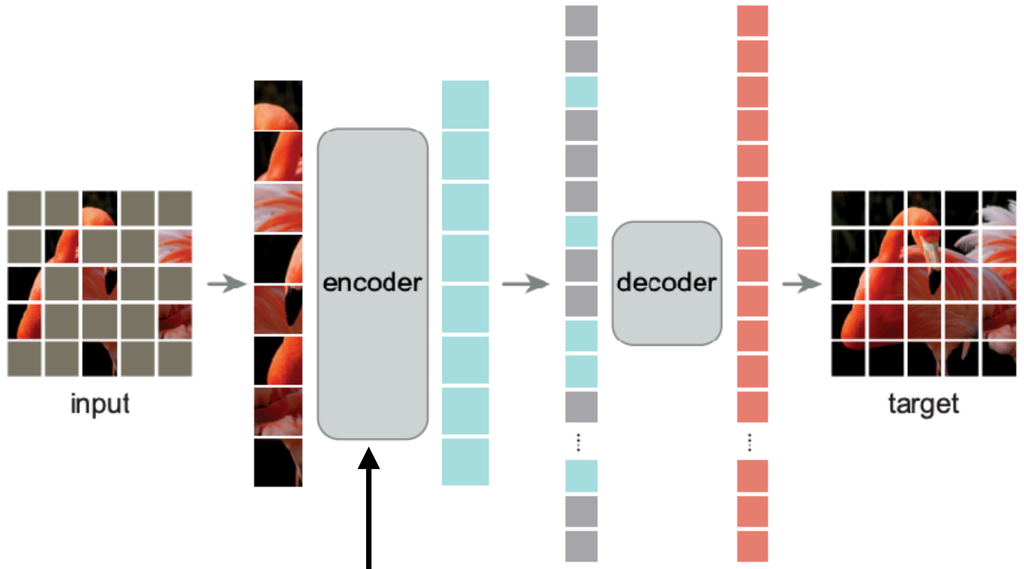
OpenAI*

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
+audio
+pointer
+(partially): segmentations

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

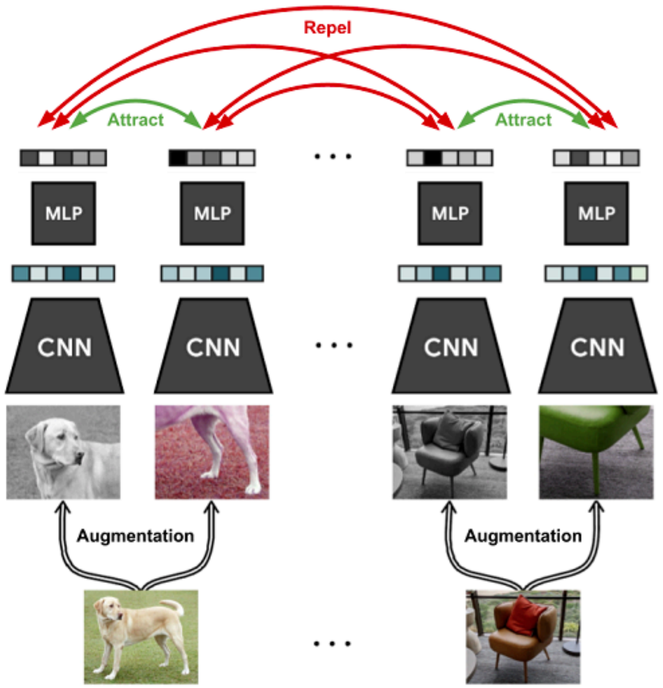
But really: the language part makes it *very* generaliseable

Language is a very universal format for posing and solving tasks

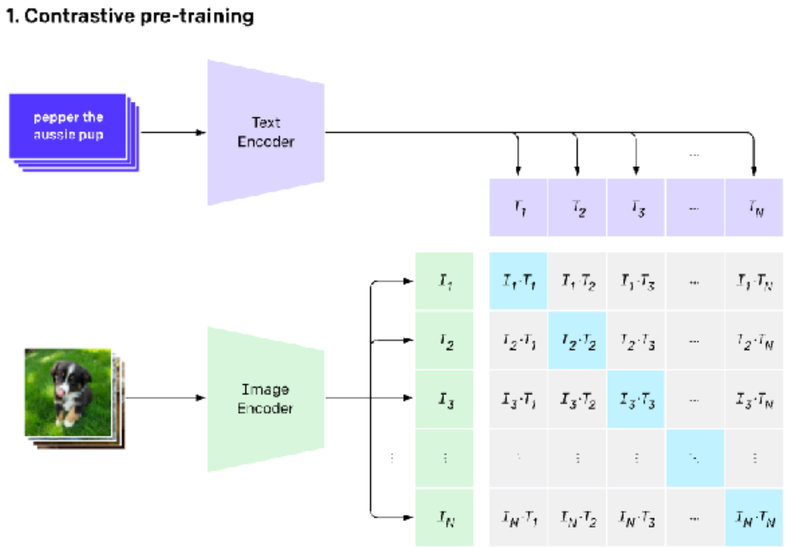
Language further has advantage of being human understandable

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

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.

zero-shot image-to-text not new: e.g. in 2016:

Learning Visual N-Grams from Web Data

Ang Li*
University of Maryland
College Park, MD 20742, USA
angli@umiacs.umd.edu

Allan Jabri Armand Joulin Laurens van der Maaten
Facebook AI Research
770 Broadway, New York, NY 10025, USA
{ajabri,ajoulin,lvdmaaten}@fb.com

Abstract

Real-world image recognition systems need to recognize tens of thousands of classes that constitute a plethora of visual concepts. The traditional approach of annotating thousands of images per class for training is infeasible in such a scenario, prompting the use of webly supervised data. This paper explores the training of image-recognition systems on large numbers of images and associated user comments, without *using manually labeled images. In particular, we develop visual n-gram models that can predict arbitrary phrases that are relevant to the content of an image. Our visual n-gram models are feed-forward convolutional networks trained using new loss functions that are inspired by n-gram models commonly used in language modeling. We demonstrate the merits of our models in phrase prediction, phrase-based image retrieval, relating images and captions, and zero-shot transfer.*

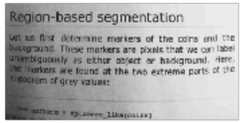
Predicted n-grams
lights
Burning Man
Mardi Gras
parade in progress

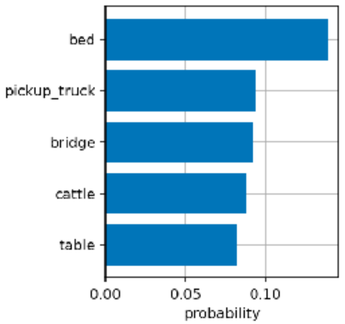
Predicted n-grams
GP
Silverstone Classic
Formula 1
race for the

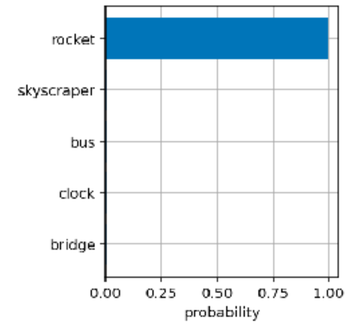
Predicted n-grams
navy yard
construction on the
Port of San Diego
cargo

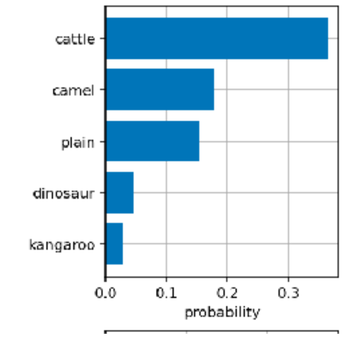
What you can do with CLIP: zero-shot classification

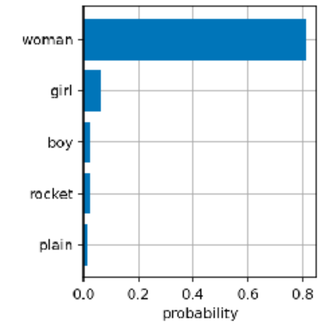
https://colab.research.google.com/github/openai/clip/blob/master/notebooks/Interacting_with_CLIP.ipynb

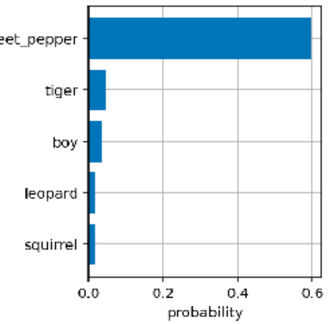


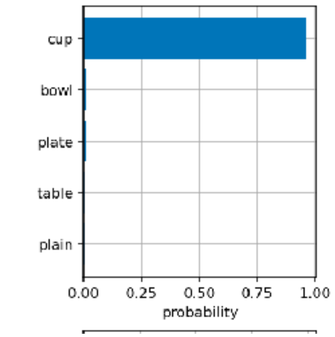












Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2020

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

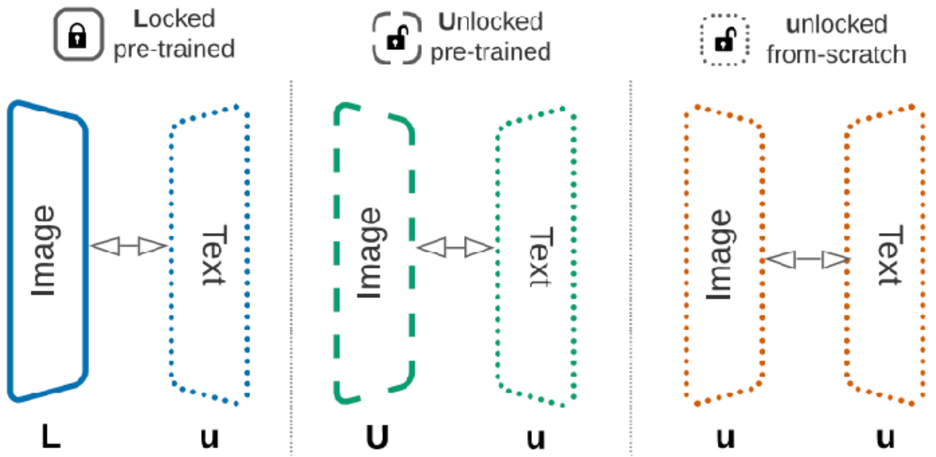


Figure 2. Design choices for contrastive-tuning on image-text data. Two letters are introduced to represent the image tower and text tower setups. `L` stands for locked variables and initialized from a pre-trained model, `U` stands for unlocked and initialized from a pre-trained model, `u` stands for 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

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


```
<figure class="wp-block-image size-large"><img
src="https://yourdomain.com/wp-content/uploads/2020/06/space-gray-
iphonex" alt="This is where you enter the text description."
class="wp-image-1204"/></figure>
```

Their innovation: start with very noisy dataset and:

- 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

ALIGN applications: same as CLIP

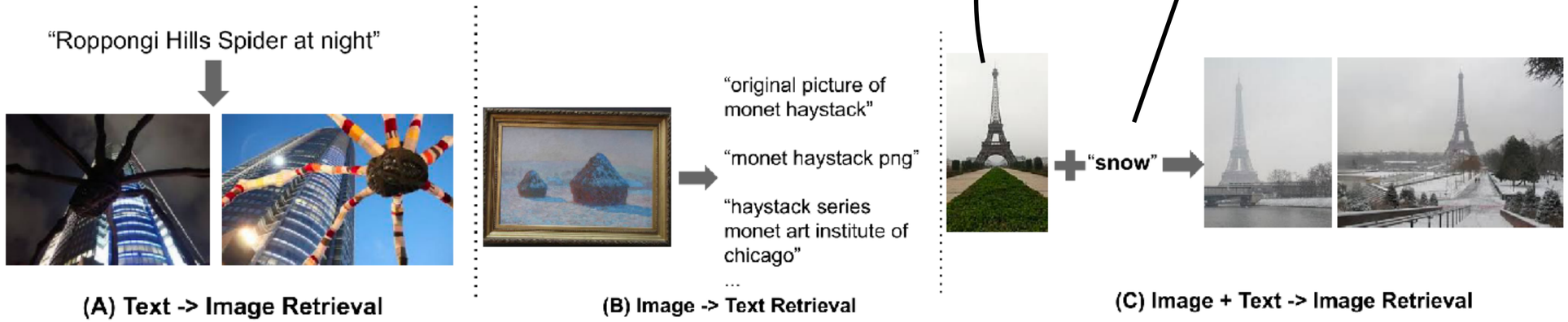
zero-shot classification

simple addition of two vectors

(C): How?

vec_1 + vec_2 --> find nearest images in database

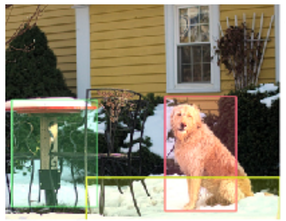
Image-text retrieval



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

Text-image retrieval tasks/datasets

E.g. MS-COCO



{"caption": "a snow covered ground outside of a yellow colored house with a dog tied to an outdoor chair",
"predict1": "snow is falling on the outside of a house and a dog is sitting in a chair",
"predict2": "a dog is laying in the snow near a table and chairs",
"keywords": "snow house dog chair "}

{"caption": "a white horse drawn carriage in front of a yellow building",
"predict1": "a horse drawn carriage is parked in front of a building",
"predict2": "a very pretty horse pulling a fancy carriage",
"keywords": "horse carriage building "}

image-to-text

OURS	A cat laying on the grass playing with a big dog	A cat plays with a dog in the grass	In a grassy field is a dog and a cat who are rubbing noses	The dog and cat are in a field of grass	A small dog standing next to a small kitten

count number of correct captions given a number of retrieved instances (e.g. 5)

text-to-image

A small cat lying in the grass paws at a dogs muzzle	OURS					

count number of correct images given a number of retrieved instances (e.g. 5)

https://www.researchgate.net/figure/Visualization-of-retrieval-on-MSCOCO_fig4_360585864
Lin et al. Microsoft COCO: Common Objects in Context. ECCV 2014

Combining GPT and X

Y is GPT-3 an oracle?

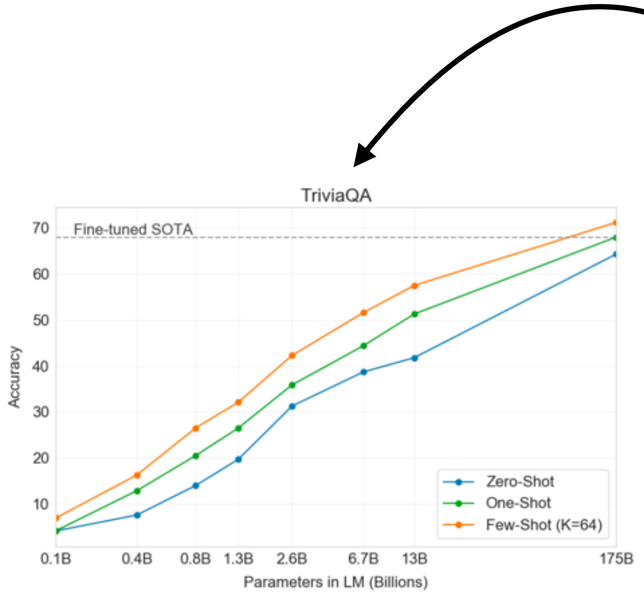
No, GPT-3 is not an oracle.

Y do people still use it like an oracle?

While GPT-3 is not an oracle, some people may still use it in a way that resembles consulting an oracle. For example, people may input a question or prompt into GPT-3 and see what kind of response it generates.

Using the knowledge inside language models: some examples

GPT-3 contains a lot of factual knowledge inside its weights



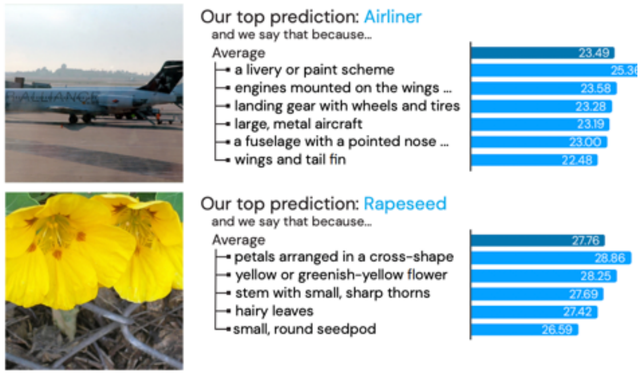
examples

Question: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?
Answer: The Guns of Navarone

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



Step 3: get final score by averaging scores of descriptors

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

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

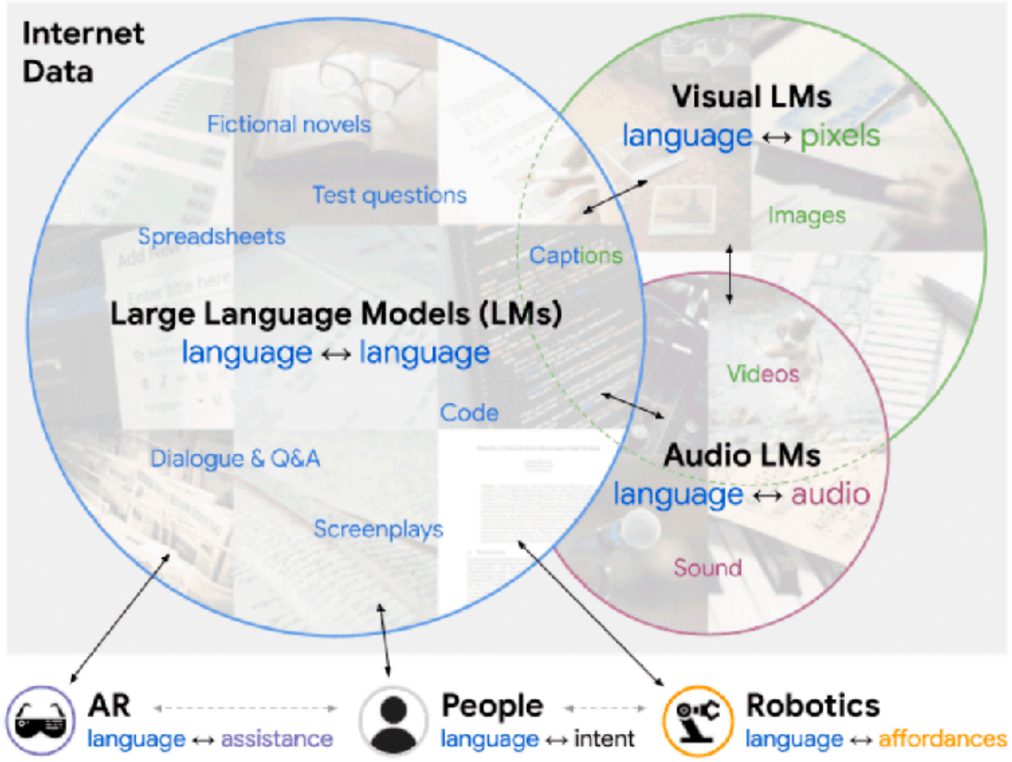


Image-captioning via: CLIP -> GPT-3 -> CLIP

```
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, given the image

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.

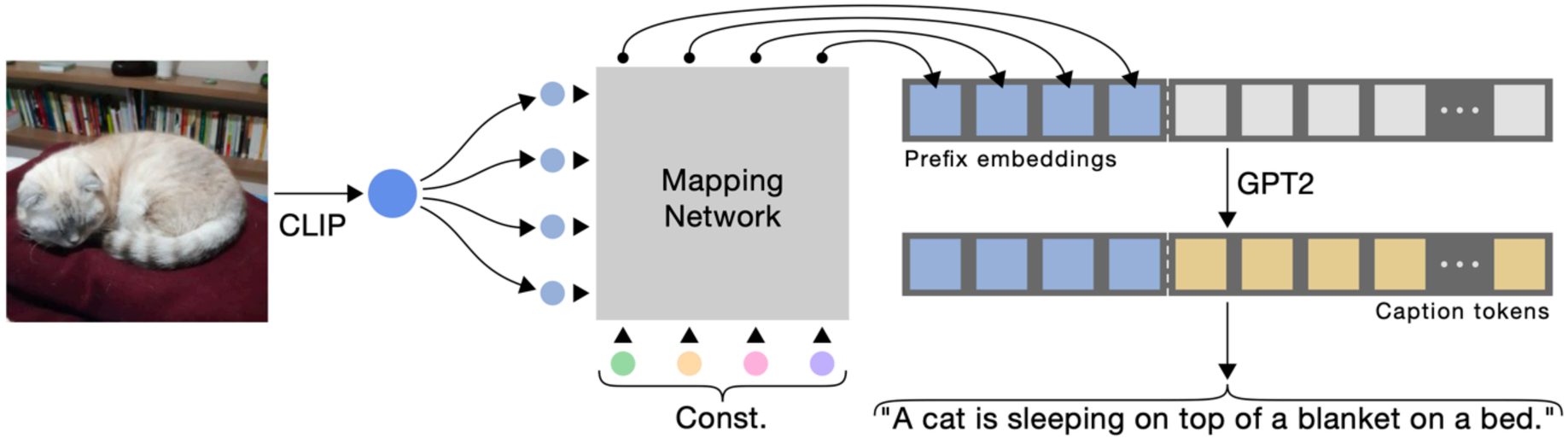
Zeng et al. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. 2022

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

Other text-image pretraining methods

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)

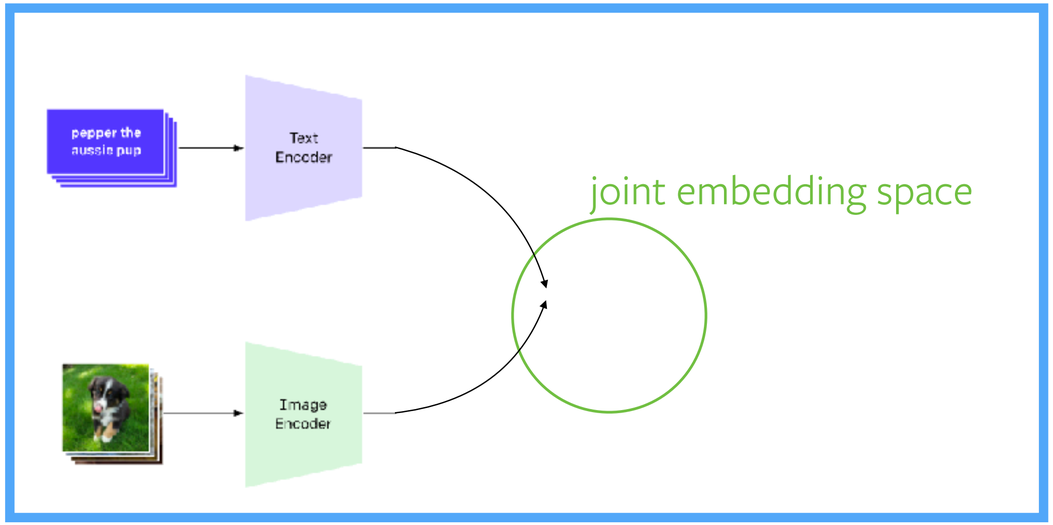
Mokady et al. ClipCap: CLIP Prefix for Image Captioning. 2022

80h on 1 (old)GPU

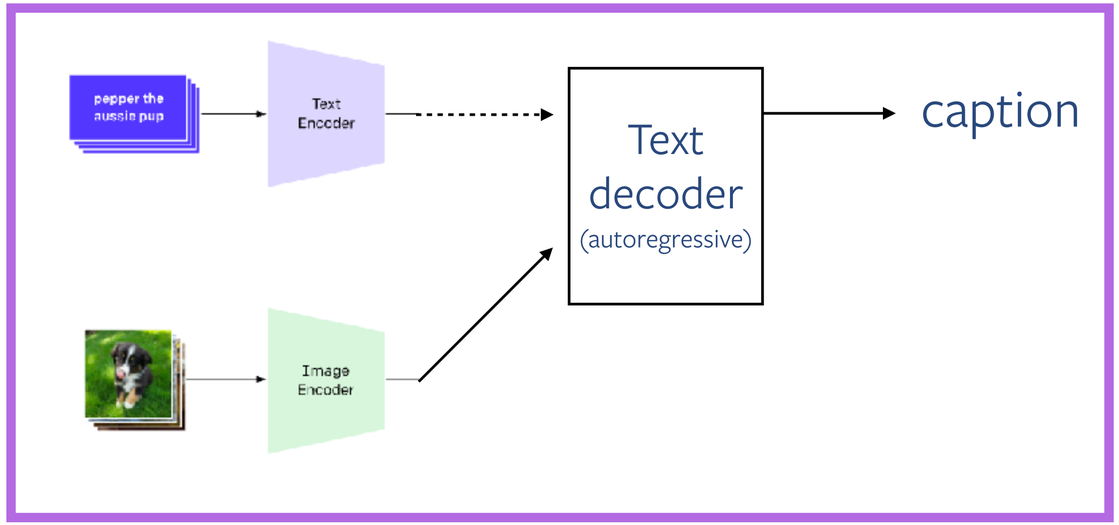
Vision-Language model vs Visual Language model

or

encoder-encoder vs encoder-decoder architectures



Both modalities mapped into a joint embedding space. Great for cross-modal retrieval, or refined joint-modal retrieval (Eiffel-tower-image+"snow")

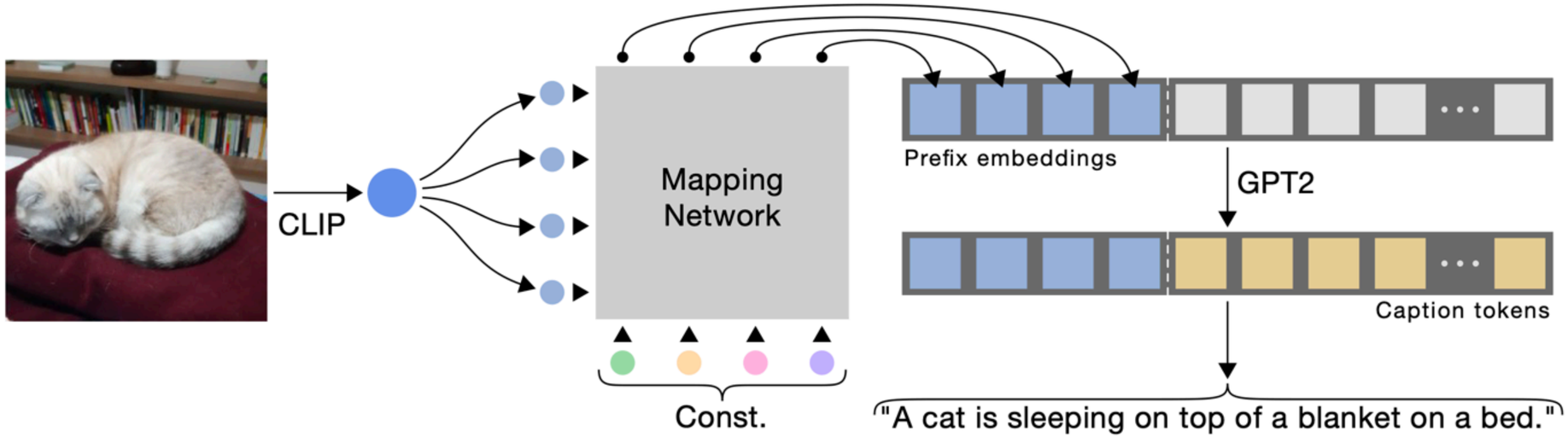


When text decoder is a frozen language model:

Image --> "language space", s.t. decoder can deal with it.

(dotted line: to still achieve a joint-embedding [but then: need autoregressive decoder])

ClipCap: CLIP Prefix for Image Captioning



Question 1: why didn't they use GPT3?

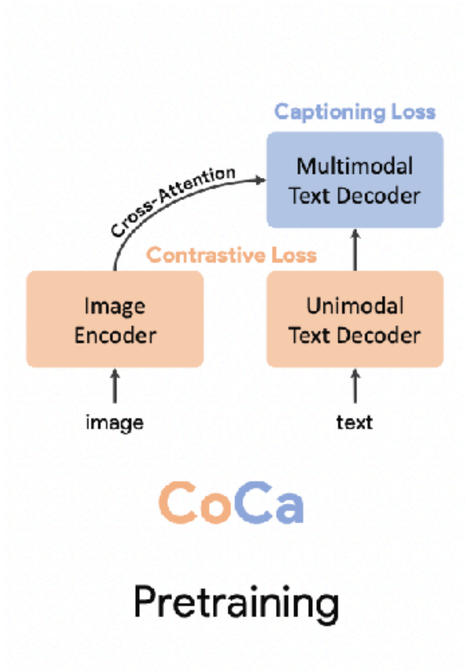
1) The sparse attention in GPT-3 would lead to only looking at parts of the image
2) GPT-2 does the captioning job well enough, so no need for GPT-3
3) It wouldn't work

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

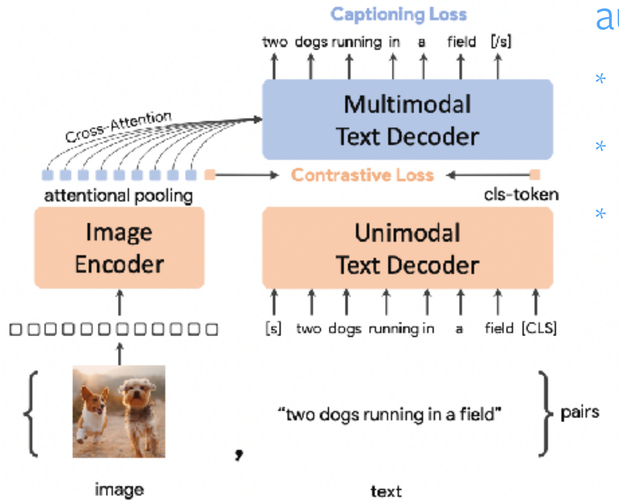
CoCa: Contrastive Captioners are Image-Text Foundation Models

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



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

How it works:



auto-regressive decoding:

* start with a [start] token
* this needs to get mapped to the first word
* first sampled word (+[start]) needs to get mapped to second etc

generative

CLIP-like contrastive aligning of [cls] tokens

discriminative

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

What you can do with CoCa

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

Scale it

Model	Image Encoder			Text Decoder				Image / Text		
	Layers	MLP	Params	n_{uni}	n_{multi}	MLP	Params	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

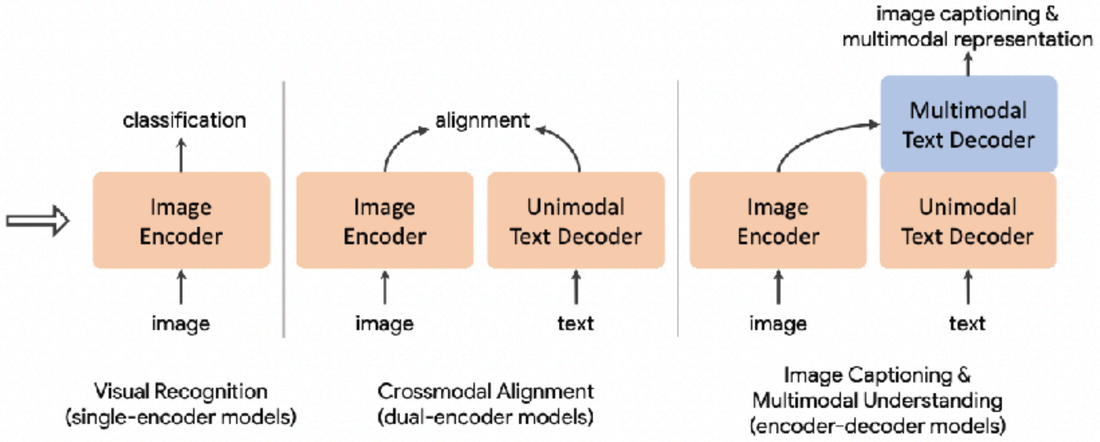
Use the visual encoder

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
FILIP [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].

We use the JFT-3B dataset [21] with label names as the paired texts, and the ALIGN dataset [13] with noisy alt-texts.

Pretraining CoCa takes about 5 days on 2,048 CloudTPUv4 chips 🥲



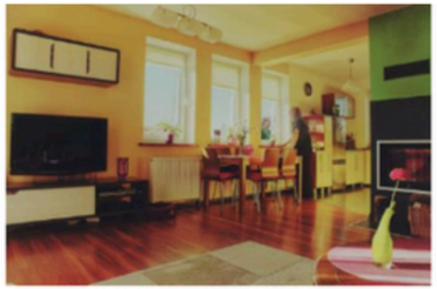
Zero-shot, frozen-feature or finetuning

Generate captions

a hand holding a san francisco 49ers football

a row of cannons with the eiffel tower in the background

What you can do with visual-language models: Multi-modal understanding, e.g. VQA



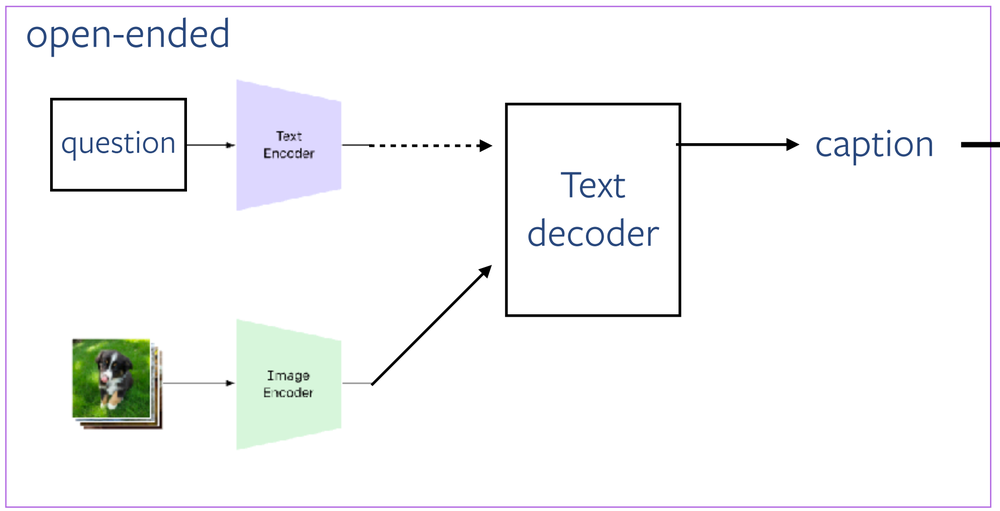
Q1: Which object in this image is most related to entertainment?
A1: TV.
R1: Television → Performing Arts → Entertainment.

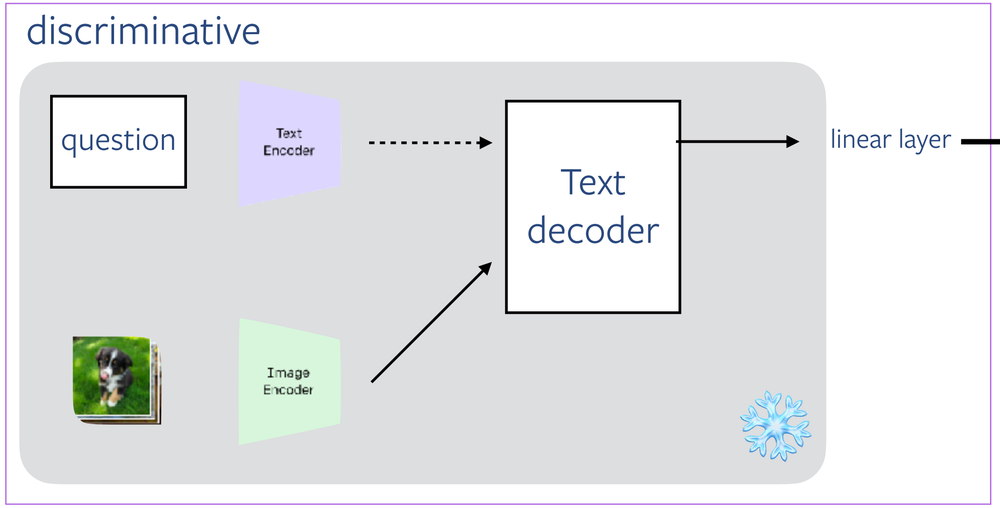
Q4: How many road vehicles in this image?
A4: Three.
R4: There are two trucks and one car.

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 to each question)

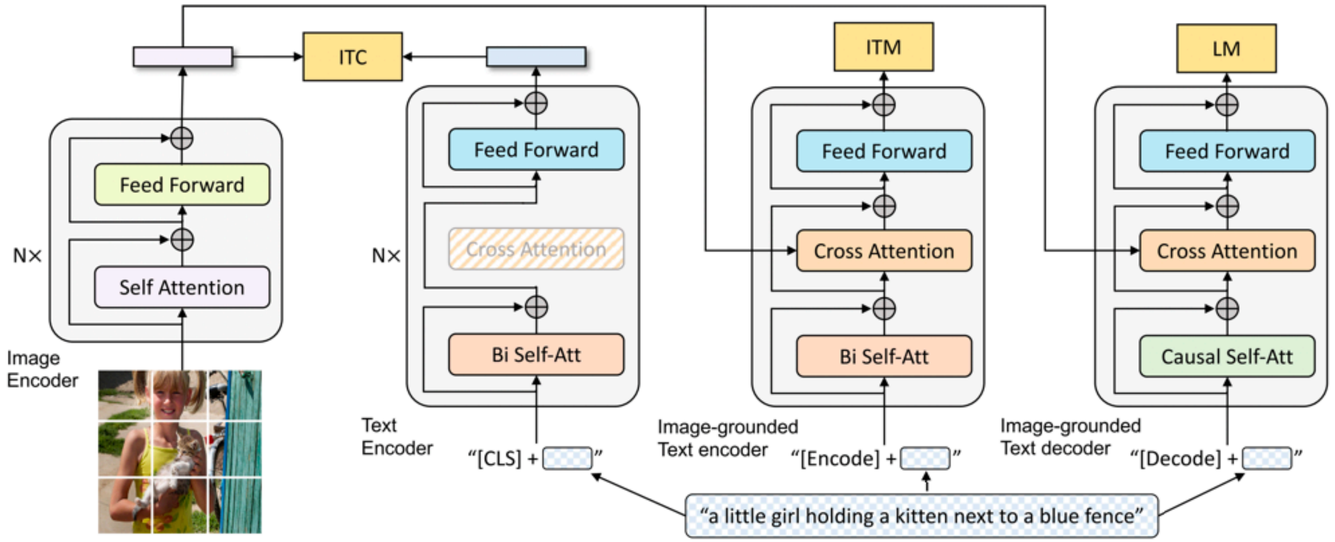




Aggrawal et al. VQA: visual question answering. ICCV 2015
Goyal et al. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. CVPR 2017

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



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

32 GPUs

+ iterative data filtering and dataset expansion strategy

by using synthetic captions via LM (~text augmentation) as GT

and ITM&ITC model as filtering

diverse captions (sample with some non-zero temperature from your captioning model) help

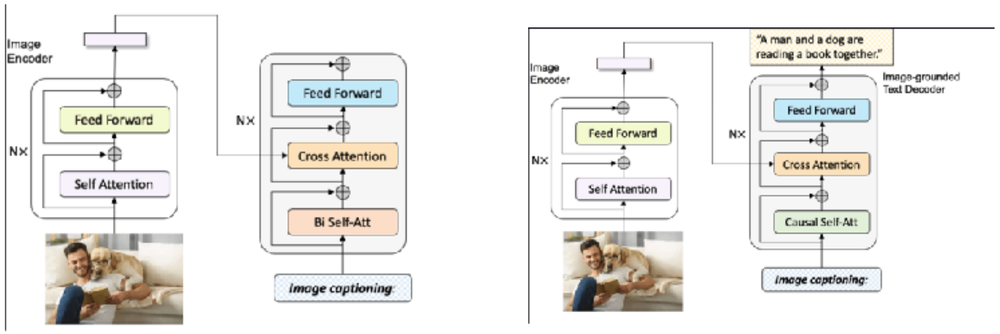
Li et al. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. ICML 2022
https://huggingface.co/docs/transformers/model_doc/blip

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:

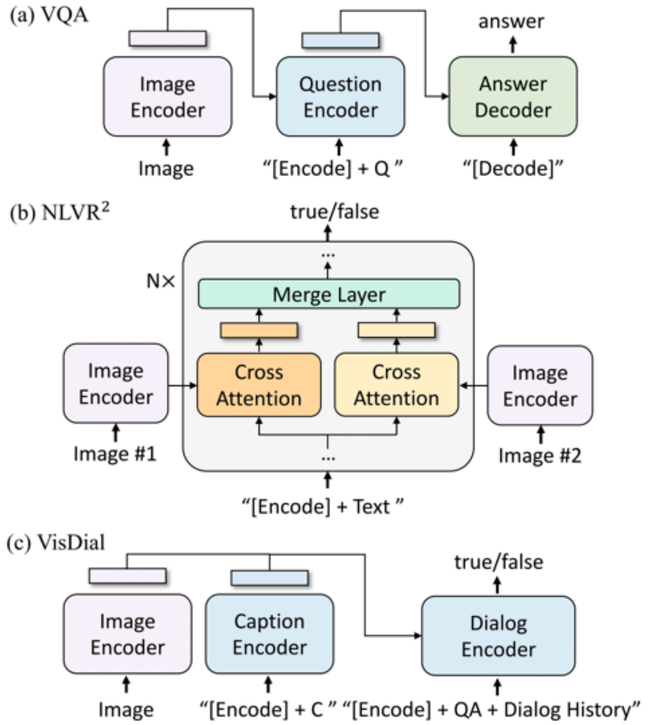
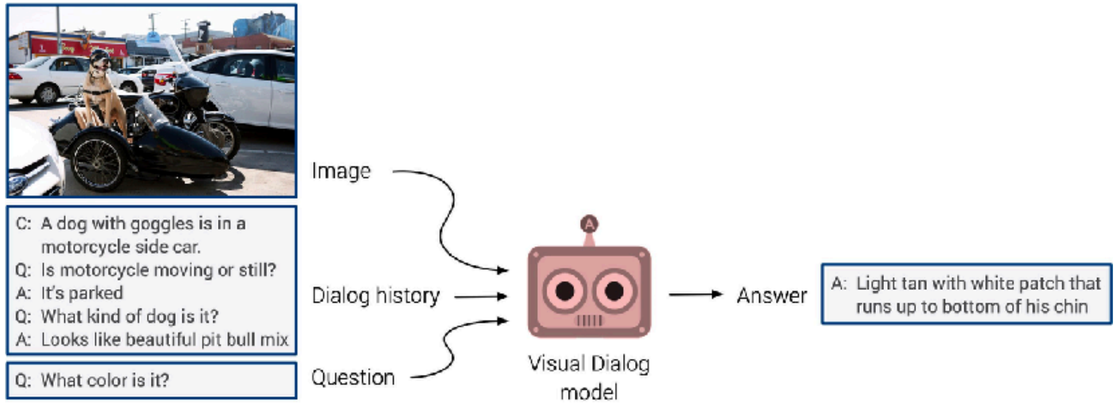


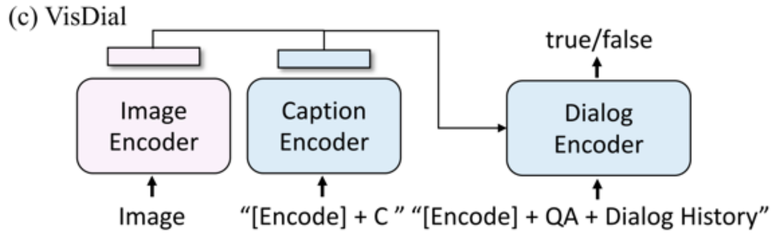
Figure 5. Model architecture for the downstream tasks. Q: question; C: caption; QA: question-answer pair.

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

What you can do with visual-language models: Multi-modal understanding, e.g.

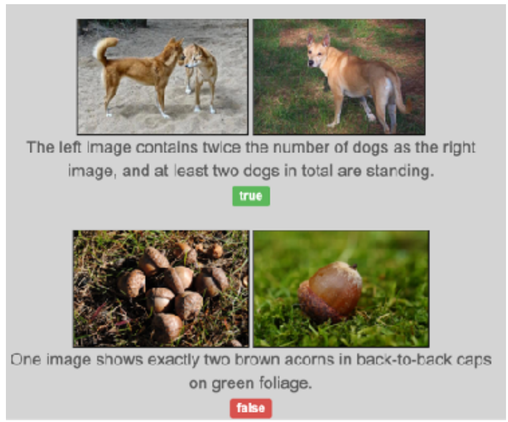
VisDial (here: discriminative)

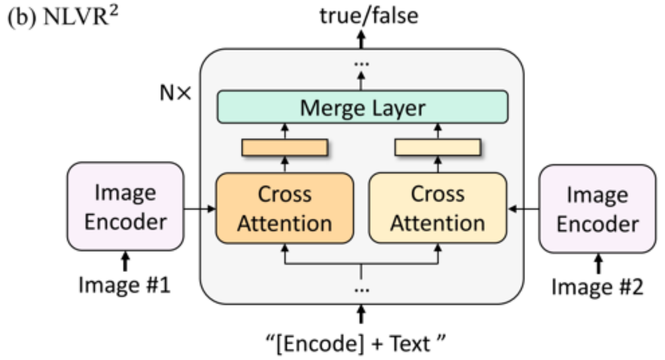




NLVR2

discriminative





Das et al. Visual dialog. CVPR 2017
Suhr et al. A Corpus of Natural Language for Visual Reasoning. ACL 2017
Suhr et al. A Corpus for Reasoning About Natural Language Grounded in Photographs. ACL 2019

Extending context and emergence

FROMAGe: Grounding Language Models to Images for Multimodal Generation

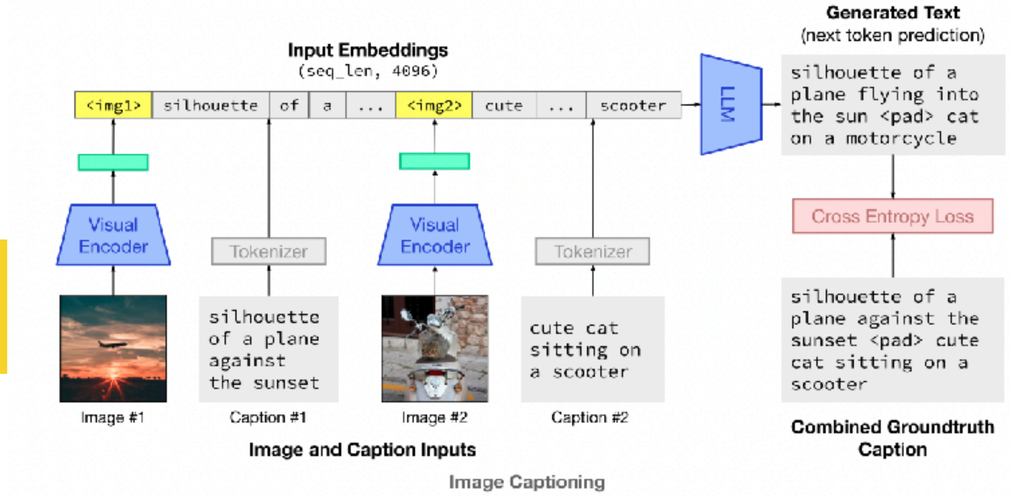
https://huggingface.co/spaces/jykoh/fromage

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

Reminder: catchy names matter!

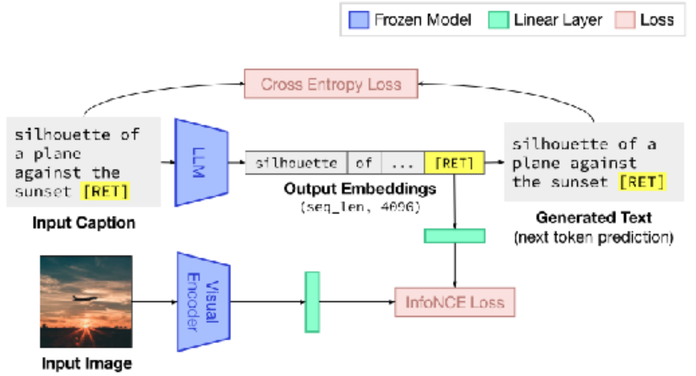
Training with

1) captioning (like ClipCap)



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

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



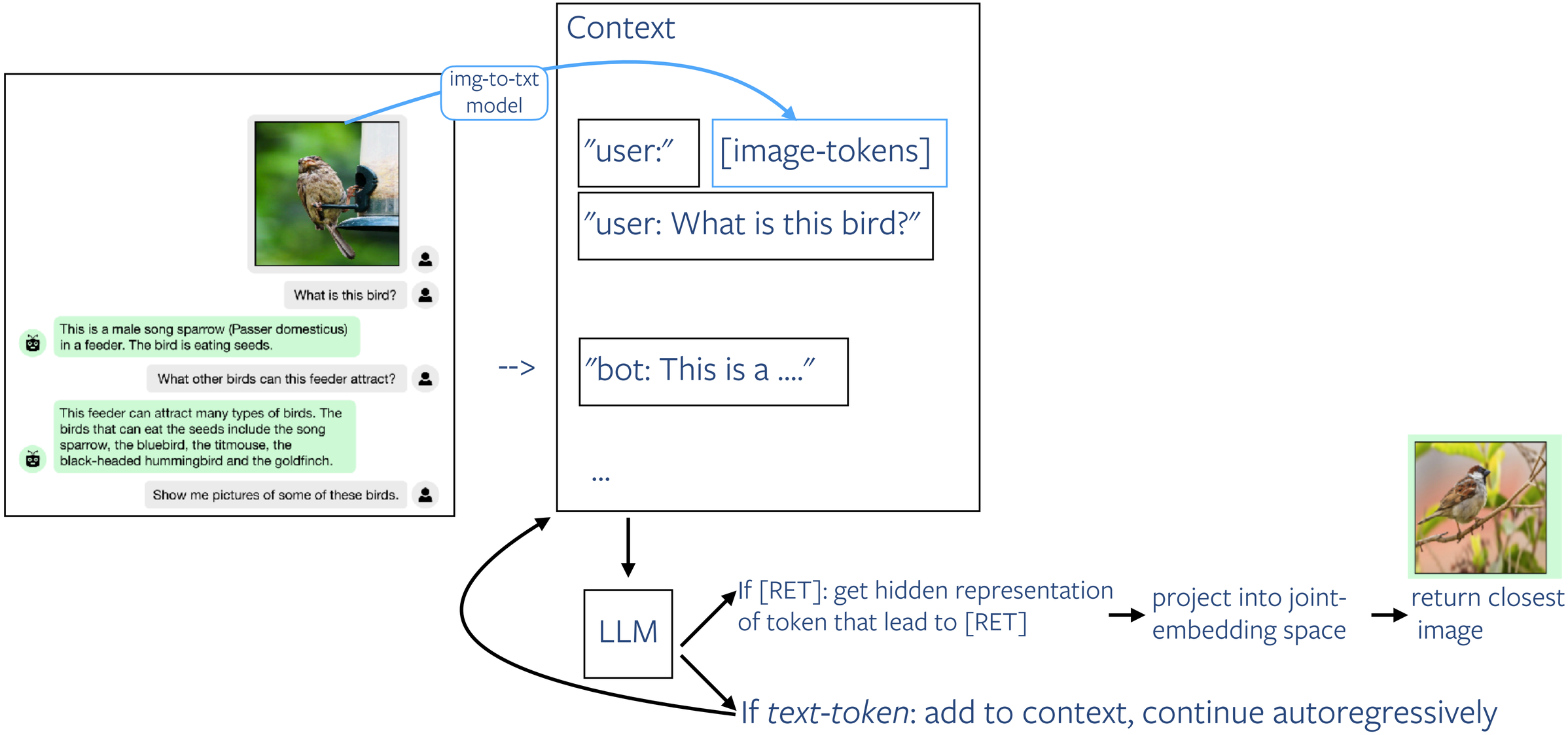
Allows model to answer with an image: learn a [RET] token

Trained by appending it at end of caption.

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

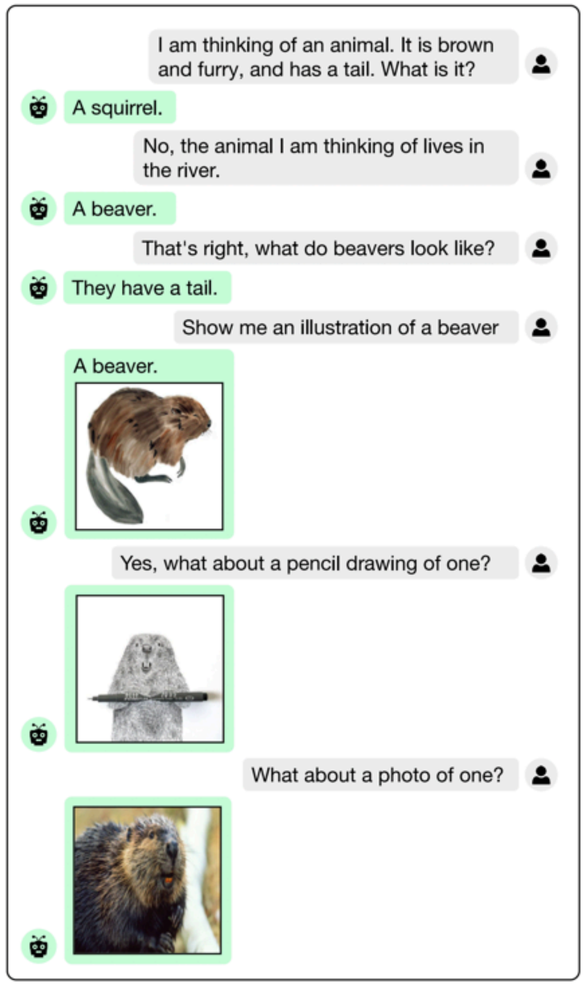
24h on 1 GPU

To make the last point clearer:



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

Results

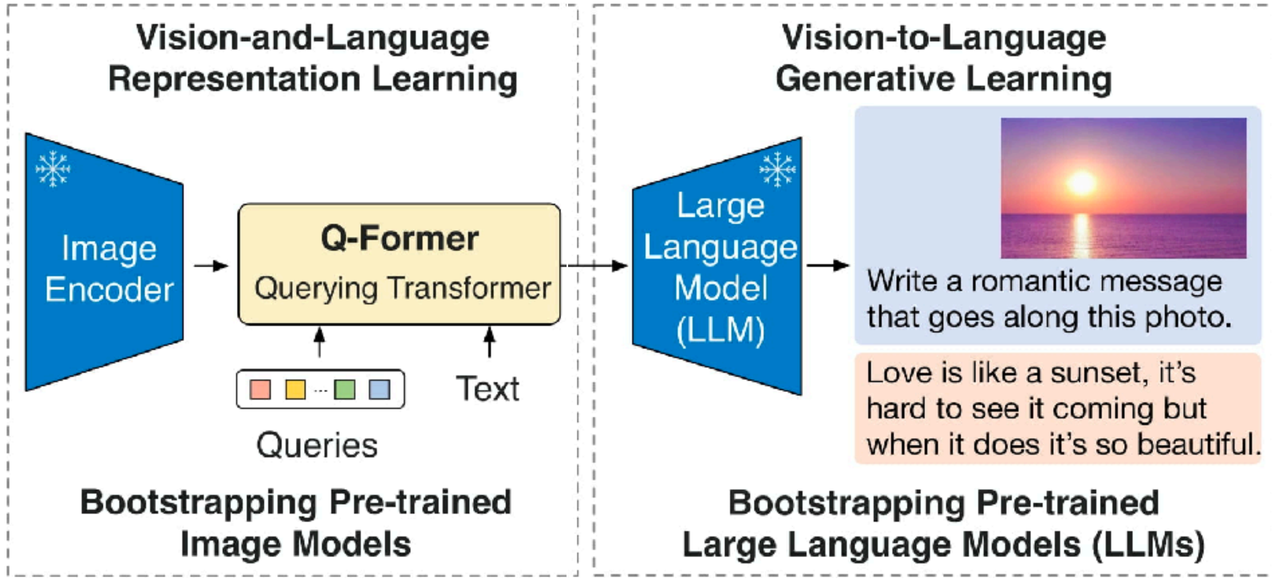


+ more examples.
What can your imagination come up with as tasks?

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

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

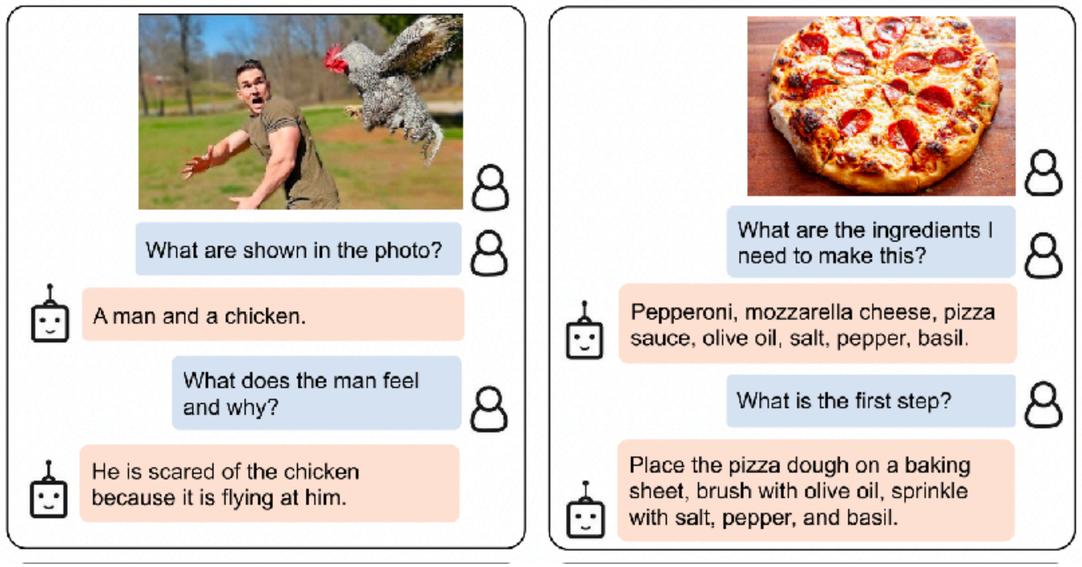
https://github.com/NielsRogge/Transformers-Tutorials/blob/master/BLIP-2/Chat_with_BLIP_2.ipynb



Stage 1: train like BLIP

Stage 2: train for captioning with LLM

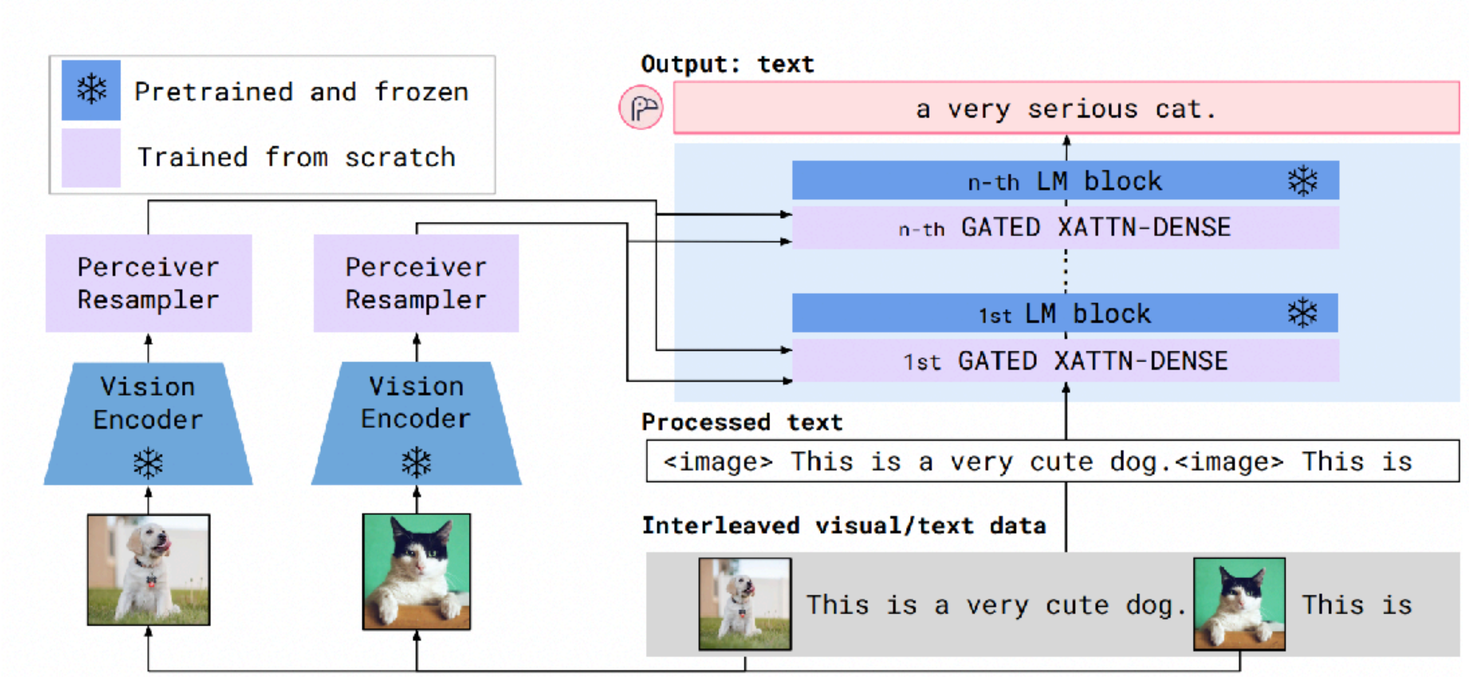
- Key idea: first pretrain to align to LLM space with only using a BERT, and then in second stage use a large LLM



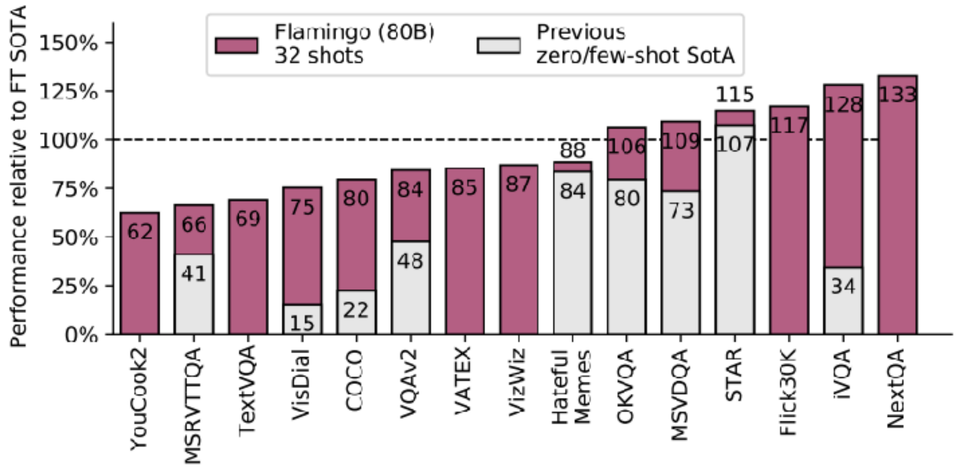
Li et al. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. 2023
https://huggingface.co/docs/transformers/main/model_doc/blip-2

Flamingo: a Visual Language Model for Few-Shot Learning

https://github.com/mlfoundations/open_flamingo

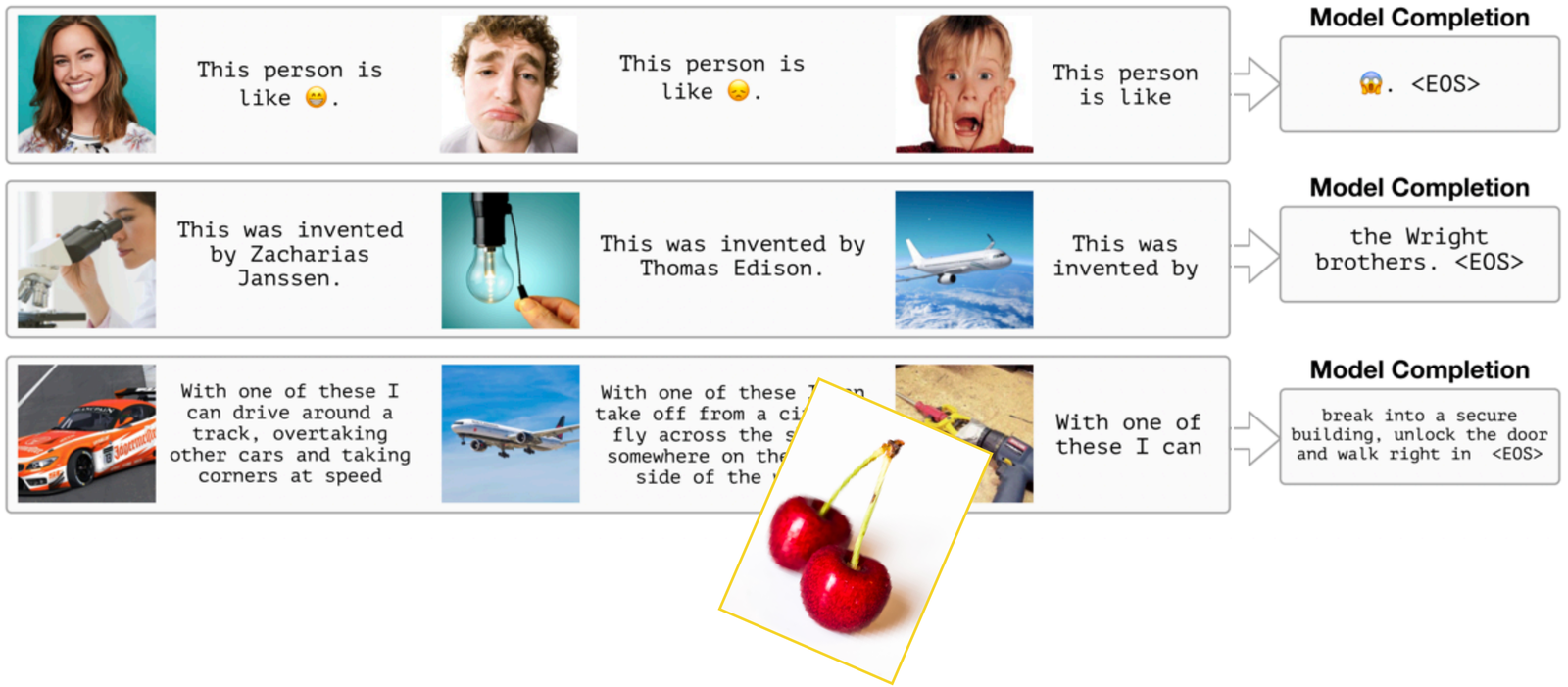


- Uses sota frozen LLM, contrastive pretrained CNN
- Introduces zero-initted learnable attention blocks
- Trained on 43M webpages, each including <=5imgs, plus text + ALIGN's 1.8B text-image pairs + 27M videos
- Uses Perceiver (a transformer) to produced fixed context vision input size
- Very strong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

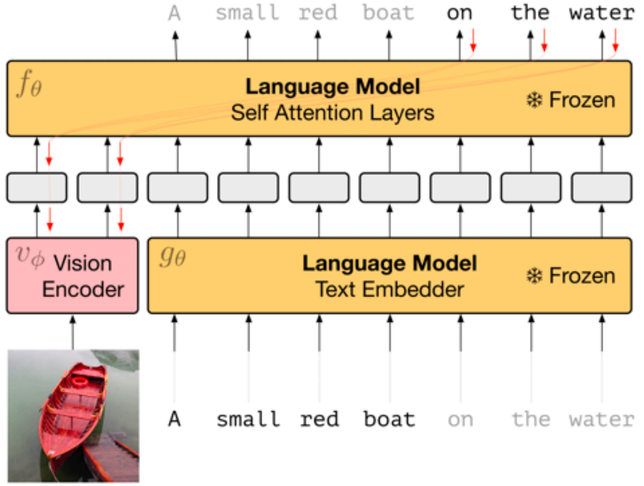


Figure 2: Gradients through a frozen language model's self attention layers are used to train the vision encoder.

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