







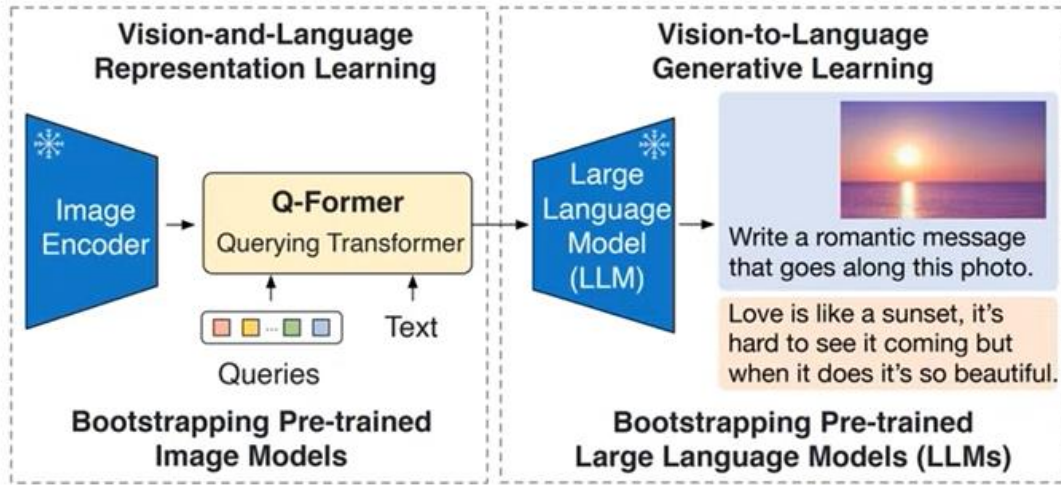
What can the BLIP-2 model do?

 <p>Explain the advantages of this product.</p> <p>The audi e-tron quattro concept is a plug-in hybrid electric sports car that has a range of 310 miles and can accelerate from 0 to 60 mph in just 4 seconds.</p>	 <p>Tell me something about the history of this place.</p> <p>The great wall of china was built in 221 bc by emperor qin shihuang to protect the emperor's capital from invaders from the north.</p>	 <p>Write down the facts that you know about this flower.</p> <p>It is a genus of flowering plants in the orchid family, orchidaceae. It is native to eastern asia and south east asia. It is a perennial flowering plant that grows up to a height of 1 m.</p>
 <p>Is this photo unusual?</p> <p>Yes, it's a house that looks like it's upside down.</p> <p>How could someone get out of the house?</p> <p>It has a slide on the side of the house.</p>	 <p>What are shown in the photo?</p> <p>A man and a chicken.</p> <p>What does the man feel and why?</p> <p>He is scared of the chicken because it is flying at him.</p>	 <p>What are the ingredients I need to make this?</p> <p>Pepperoni, mozzarella cheese, pizza sauce, olive oil, salt, pepper, basil.</p> <p>What is the first step?</p> <p>Place the pizza dough on a baking sheet, brush with olive oil, sprinkle with salt, pepper, and basil.</p>

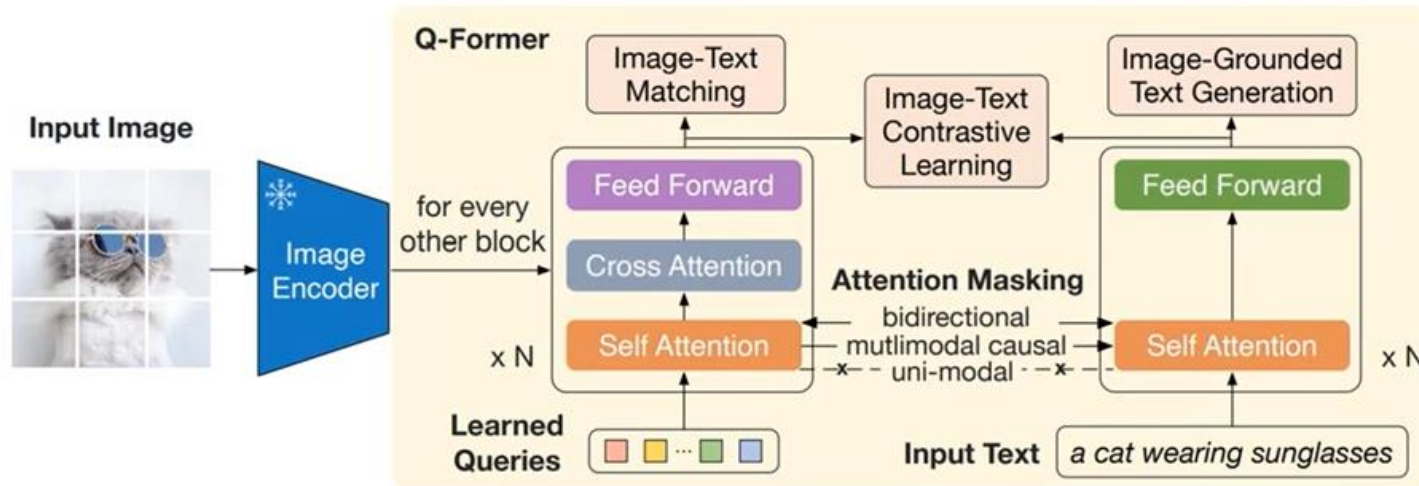
- Instructed zero-shot image-to-text generation
 - Visual conversation
 - Visual knowledge reasoning
 - Visual commonsense reasoning
 - Storytelling
 - Personalized image-to-text generation

Li, Junnan, Dongxu Li, Silvio Savarese, and Steven Hoi. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." arXiv:2301.12597 (2023).

How is the BLIP-2 model pretrained?

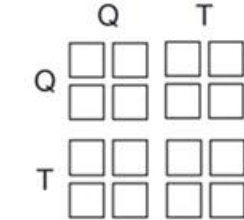


- Querying Transformer (Q-Former) pre-trained in two stages:
 - Vision-language representation learning stage with a frozen image encoder
 - Vision-to-language generative learning stage with a frozen LLM.
- Q-Former=image transformer+ text transformer
- Queries interact with
 - each other and optionally text through self-attention layers
 - frozen image features through cross-attention layers



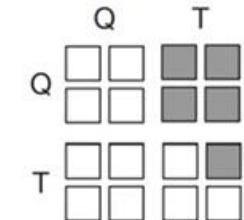
Q: query token positions; T: text token positions.

■ masked □ unmasked



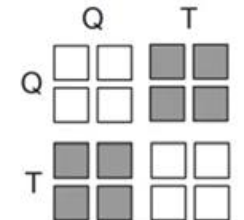
Bi-directional Self-Attention Mask

Image-Text Matching



Multi-modal Causal Self-Attention Mask

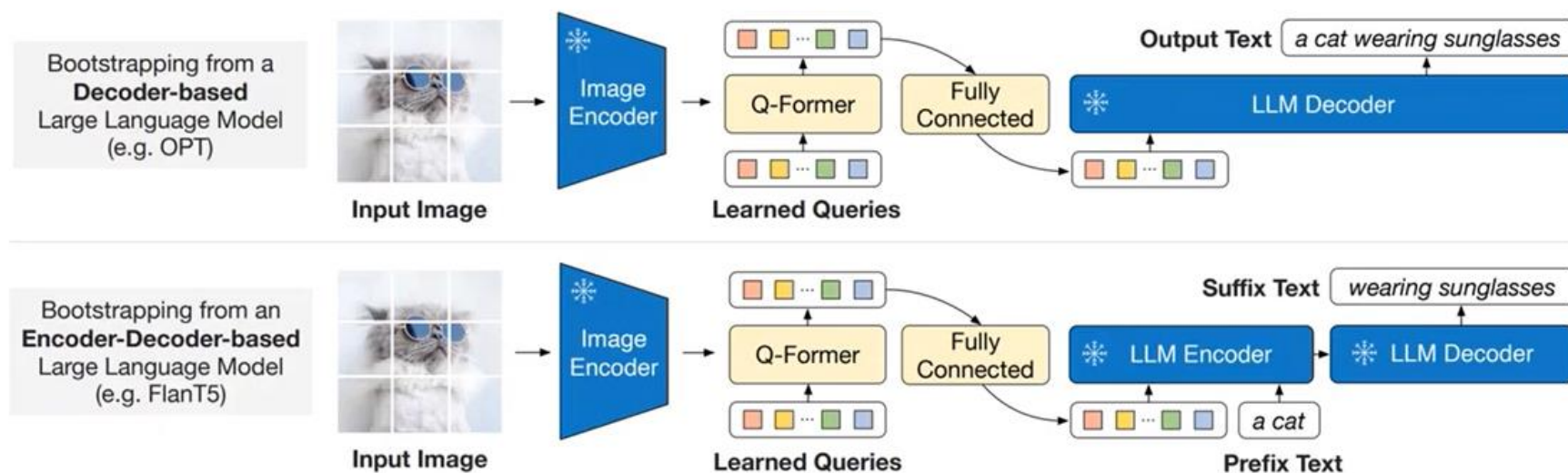
Image-Grounded Text Generation



Uni-modal Self-Attention Mask

Image-Text Contrastive Learning

How is the BLIP-2 model pretrained?



- Pre-training dataset
 - Same as BLIP
 - 129M images from COCO, Visual Genome, CC3M, CC12M, SBU
 - 115M images from LAION400M dataset
 - CapFilt method to create synthetic captions for the web images.
- Pre-trained image encoder
 - ViT-L/14 from CLIP
 - ViT-g/14 from EVA-CLIP
- Frozen language model
 - OPT for decoder-based LLMs
 - FlanT5 for encoder-decoder-based LLMs.

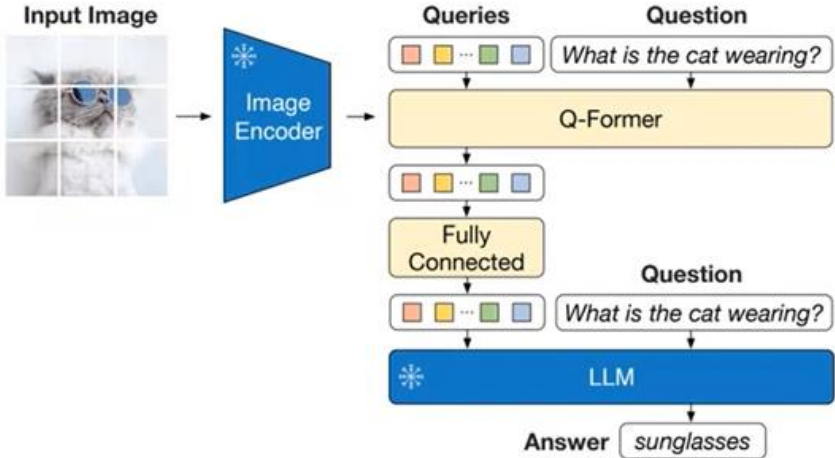
[Li, Junnan, Dongxu Li, Silvio Savarese, and Steven Hoi. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." arXiv:2301.12597 \(2023\).](https://arxiv.org/abs/2301.12597)

How does BLIP-2 model perform?

Models	#Trainable Params	Open-sourced?	Visual Question Answering	Image Captioning		Image-Text Retrieval	
			VQAv2 (test-dev) VQA acc.	NoCaps (val) CIDEr	SPICE	Flickr (test) TR@1	IR@1
BLIP (Li et al., 2022)	583M	✓	-	113.2	14.8	96.7	86.7
SimVLM (Wang et al., 2021b)	1.4B	✗	-	112.2	-	-	-
BEIT-3 (Wang et al., 2022b)	1.9B	✗	-	-	-	94.9	81.5
Flamingo (Alayrac et al., 2022)	10.2B	✗	56.3	-	-	-	-
BLIP-2	188M	✓	65.0	121.6	15.8	97.6	89.7

Zero-shot vision-language tasks

- OPT prompt “Question: {} Answer:” FlanT5 prompt “Question: {} Short answer:”
- BLIP-2 > Flamingo80B by 8.7% on VQAv2, despite having 54x fewer trainable parameters.

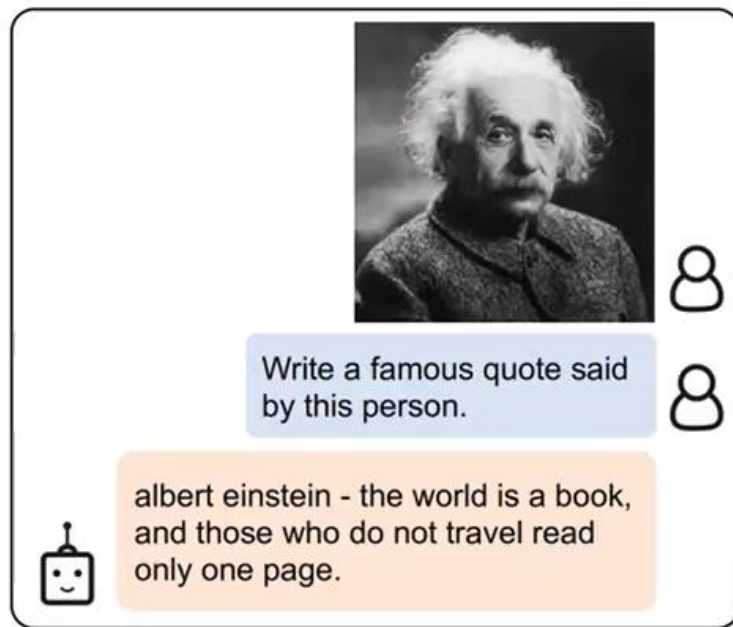


Models	#Trainable Params	NoCaps Zero-shot (validation set)								COCO Fine-tuned Karpathy test	
		in-domain C	in-domain S	near-domain C	near-domain S	out-domain C	out-domain S	overall C	overall S	B@4	C
OSCAR (Li et al., 2020)	345M	-	-	-	-	-	-	80.9	11.3	37.4	127.8
VinVL (Zhang et al., 2021)	345M	103.1	14.2	96.1	13.8	88.3	12.1	95.5	13.5	38.2	129.3
BLIP (Li et al., 2022)	446M	114.9	15.2	112.1	14.9	115.3	14.4	113.2	14.8	40.4	136.7
OFA (Wang et al., 2022a)	930M	-	-	-	-	-	-	-	-	43.9	145.3
Flamingo (Alayrac et al., 2022)	10.6B	-	-	-	-	-	-	-	-	-	138.1
SimVLM (Wang et al., 2021b)	~1.4B	113.7	-	110.9	-	115.2	-	112.2	-	40.6	143.3
BLIP-2 ViT-g OPT _{2.7B}	1.1B	<u>123.0</u>	<u>15.8</u>	117.8	<u>15.4</u>	123.4	15.1	119.7	<u>15.4</u>	<u>43.7</u>	145.8
BLIP-2 ViT-g OPT _{6.7B}	1.1B	123.7	<u>15.8</u>	<u>119.2</u>	15.3	<u>124.4</u>	14.8	<u>121.0</u>	15.3	43.5	145.2
BLIP-2 ViT-g FlanT5 _{XL}	1.1B	123.7	16.3	120.2	15.9	124.8	15.1	121.6	15.8	42.4	144.5

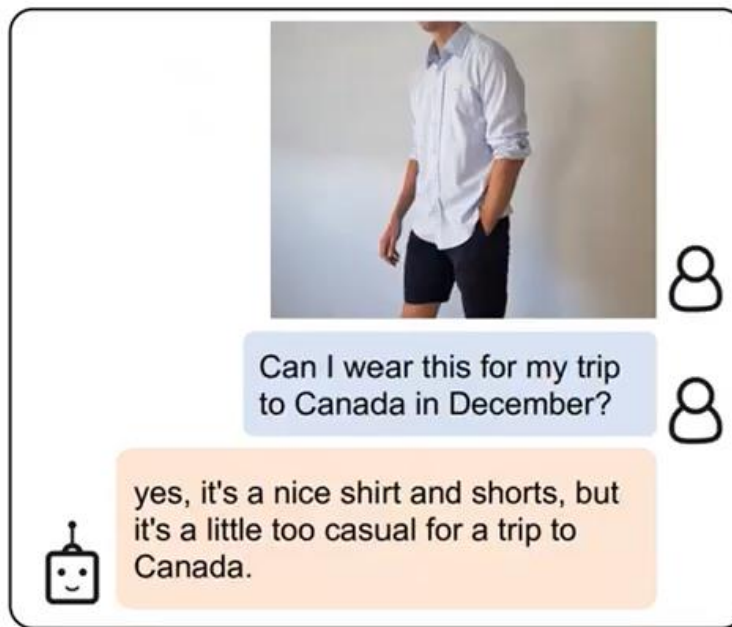
Models	#Trainable Params	VQAv2	
		test-dev	test-std
<i>Open-ended generation models</i>			
ALBEF (Li et al., 2021)	314M	75.84	76.04
BLIP (Li et al., 2022)	385M	78.25	78.32
OFA (Wang et al., 2022a)	930M	82.00	82.00
Flamingo80B (Alayrac et al., 2022)	10.6B	82.00	82.10
BLIP-2 ViT-g FlanT5 _{XL}	1.2B	81.55	81.66
BLIP-2 ViT-g OPT _{2.7B}	1.2B	81.59	81.74
BLIP-2 ViT-g OPT _{6.7B}	1.2B	82.19	82.30

Li, Junnan, Dongxu Li, Silvio Savarese, and Steven Hoi. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." arXiv:2301.12597 (2023).

How does BLIP-2 model perform?



Inaccurate knowledge
(quote is from a different person)



Incorrect reasoning path
(should have considered weather)



Information not up-to-date
(this is iphone 14)

- Inaccurate knowledge from the LLM
- Activating the incorrect reasoning path
- Not having up-to-date information about new image content.

[Li, Junnan, Dongxu Li, Silvio Savarese, and Steven Hoi. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." arXiv:2301.12597 \(2023\).](https://arxiv.org/abs/2301.12597)

Summary

- BLIP-2: a generic and compute-efficient method for vision-language pre-training that leverages frozen pretrained image encoders and LLMs.
- SOTA on various vision-language tasks while having a small amount of trainable parameters during pre-training.
- Emerging capabilities in zero-shot instructed image-to-text generation.
- <https://github.com/salesforce/LAVIS/tree/main/projects/blip2>
- Thanks for watching!
- LinkedIn: <http://aka.ms/manishgupta>
- HomePage: <https://sites.google.com/view/manishg/>