SmoothQuant & AWQ for Vision Language Models

Rachelle Hu, Tahmid Jamal, Zoe Wong

Vision Language Models (VLMs)

Tasks:

- Image Captioning
- Visual Question Answering
- Text-to-Image Search

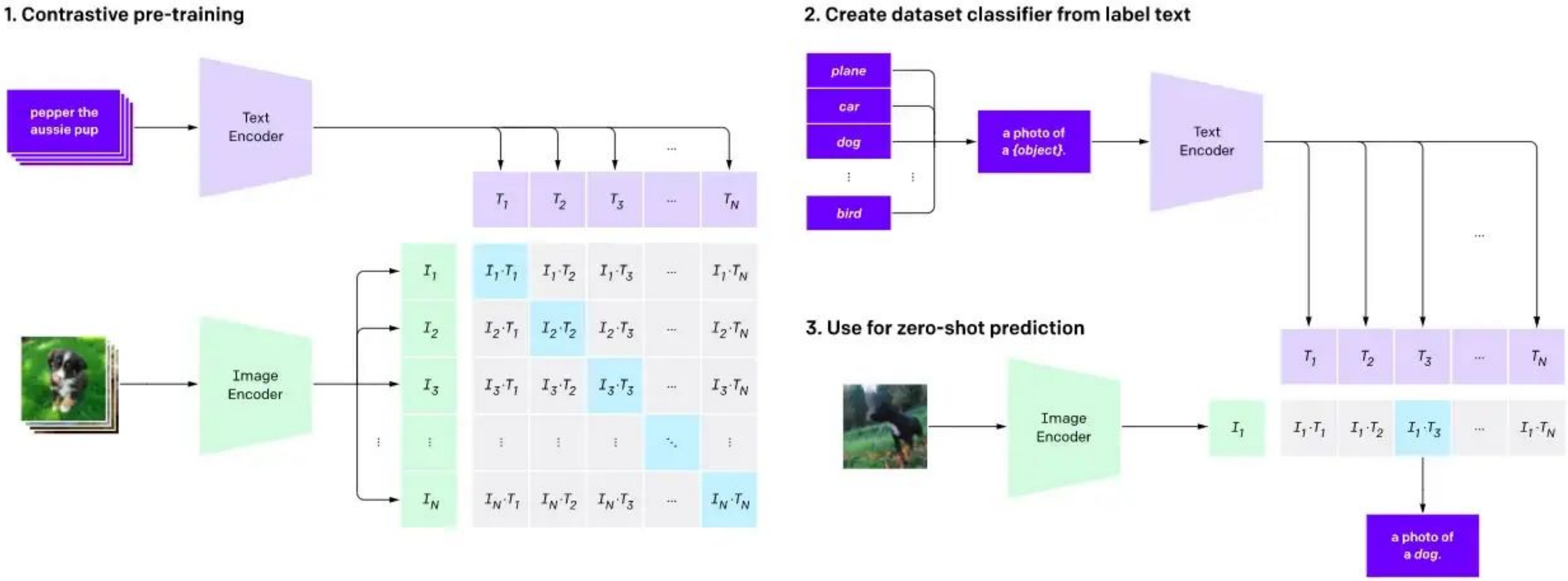
Prompt: “a girl smiling and holding a cat”

Vision Language Models (VLMs)

Pre-training Objectives:

- Contrastive Learning
- Multi-modal Fusing with Cross Attention
- PrefixLM
- Masked-Language Modeling / Image-Text Matching

CLIP



1. Contrastive pre-training

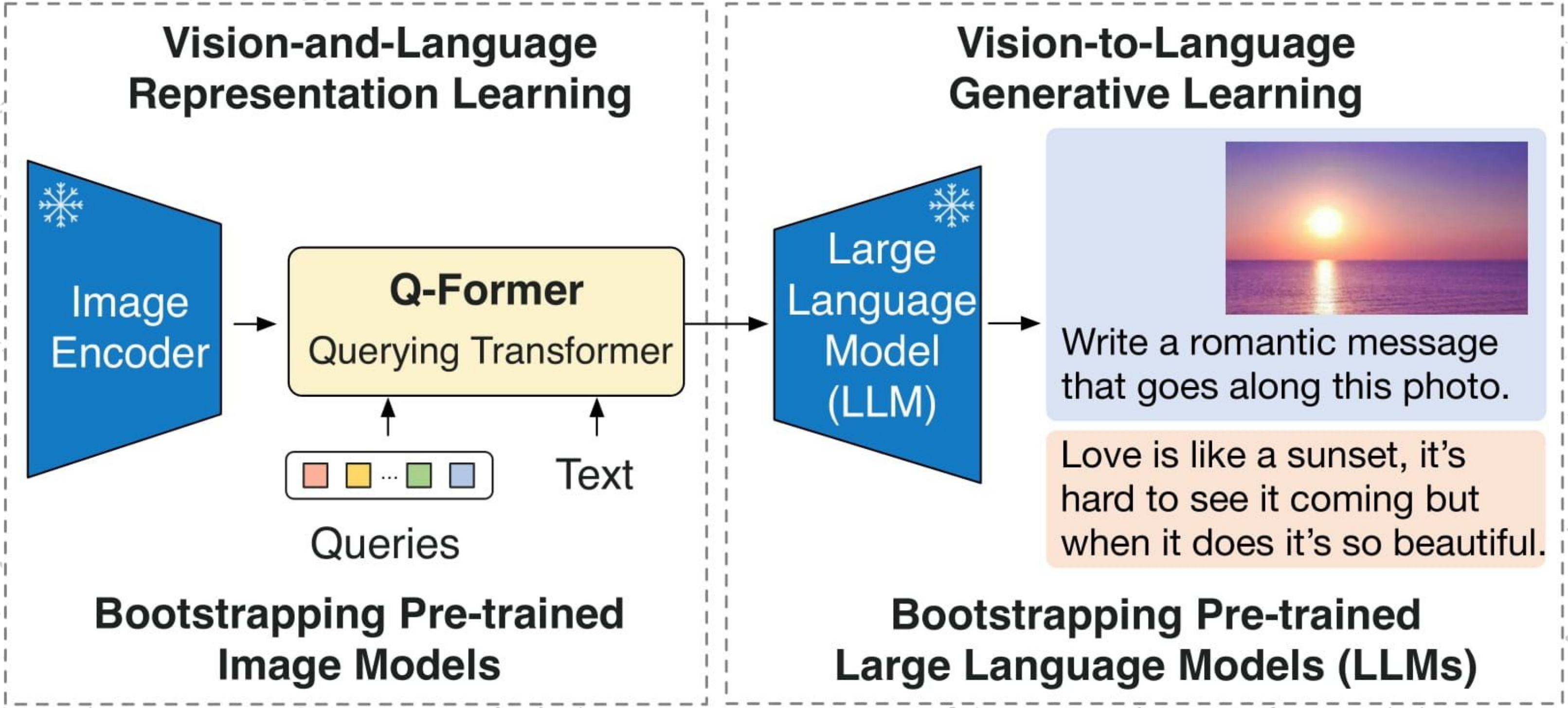
	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

2. Create dataset classifier from label text

3. Use for zero-shot prediction

Image Encoder
I_1
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
a photo of a dog.

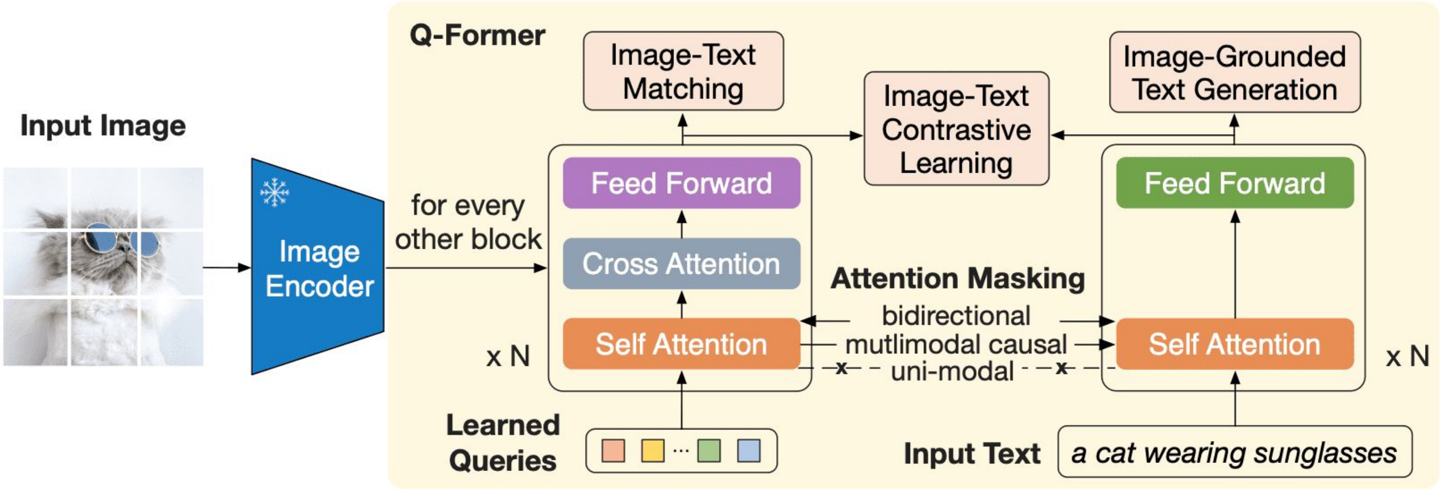
BLIP-2 Model



Write a romantic message that goes along this photo.

Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.

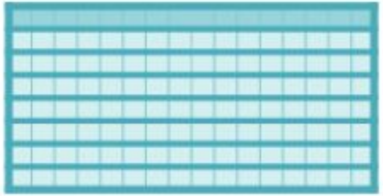
BLIP-2 Model: Q-Former

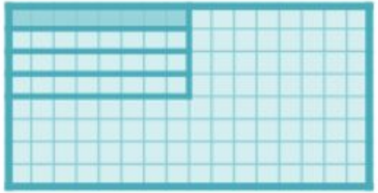


Quantization Methods On LLM of VLM

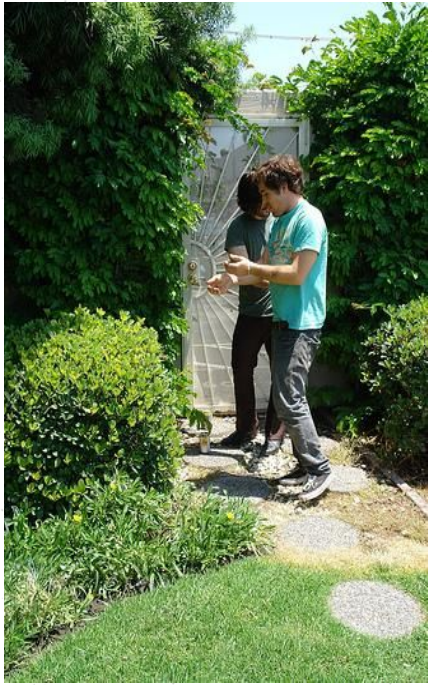
- Activation Aware Weight Only Quantization
  - Non-quantized activations, A16W4
  - Per-channel activation awareness (A16W4)

- SmoothQuant
  - Per-tensor (A8W8)
  - Per-channel (W8) & per-token (A8)
  - Per-group (A4W4 and A8W8)





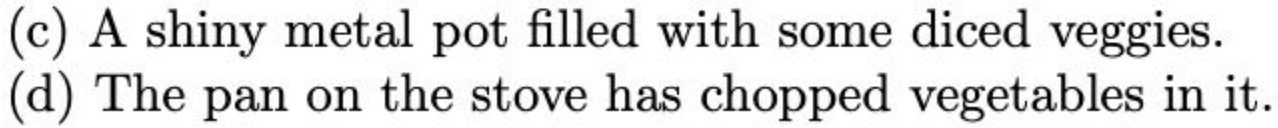
Dataset: Flickr30k



“Two young guys with shaggy hair look at their hands while hanging out in the yard.”	0
“Two young, white males are outside near many bushes.”	1
“Two men in green shirts are standing in a yard.”	2
“A man in blue shirt standing in a garden.”	3
“Two friends enjoy time spent together.”	4

Evaluation Metrics

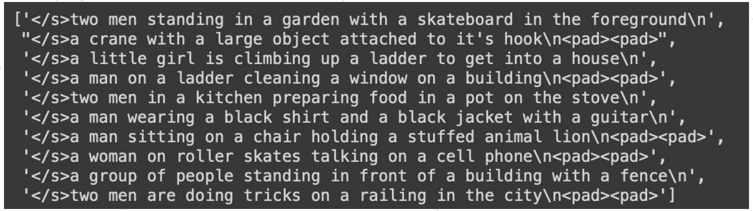
- *CIDEr*: Consensus-based Image Description Evaluation
- *SPICE*: Semantic Propositional Image Caption Evaluation



(c) A shiny metal pot filled with some diced veggies.
(d) The pan on the stove has chopped vegetables in it.

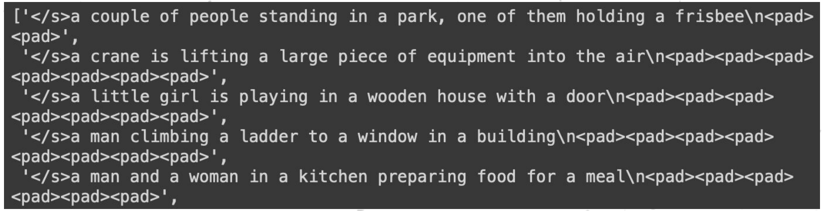
Preliminary Findings

- Grid search for factor vs fixed scale factor to protect salient weights



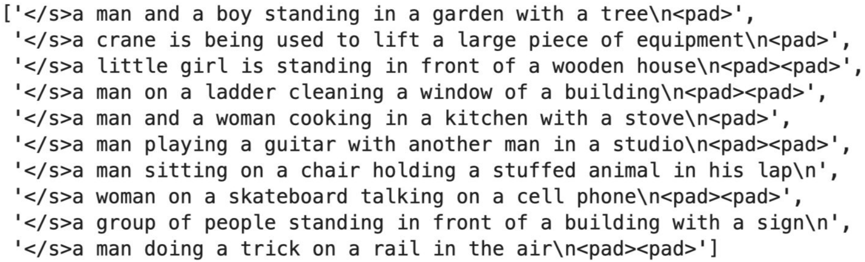
```
['</s>two men standing in a garden with a skateboard in the foreground\n',
 "</s>a crane with a large object attached to it's hook\n<pad><pad>",
 '</s>a little girl is climbing up a ladder to get into a house\n',
 '</s>a man on a ladder cleaning a window on a building\n<pad><pad>',
 '</s>two men in a kitchen preparing food in a pot on the stove\n',
 '</s>a man wearing a black shirt and a black jacket with a guitar\n',
 '</s>a man sitting on a chair holding a stuffed animal lion\n<pad><pad>',
 '</s>a woman on roller skates talking on a cell phone\n<pad><pad>',
 '</s>a group of people standing in front of a building with a fence\n',
 '</s>two men are doing tricks on a railing in the city\n<pad><pad>']
```

Original Model



```
['</s>a couple of people standing in a park, one of them holding a frisbee\n<pad>
<pad>',
 '</s>a crane is lifting a large piece of equipment into the air\n<pad><pad><pad>
<pad><pad><pad><pad>',
 '</s>a little girl is playing in a wooden house with a door\n<pad><pad><pad>
<pad><pad><pad><pad>',
 '</s>a man climbing a ladder to a window in a building\n<pad><pad><pad><pad>
<pad><pad><pad><pad>',
 '</s>a man and a woman in a kitchen preparing food for a meal\n<pad><pad><pad>
<pad><pad><pad>',
```

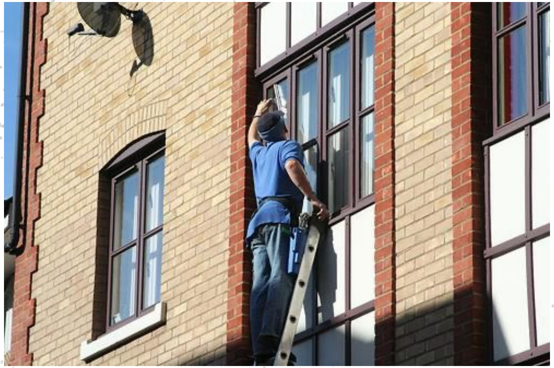
s = 2



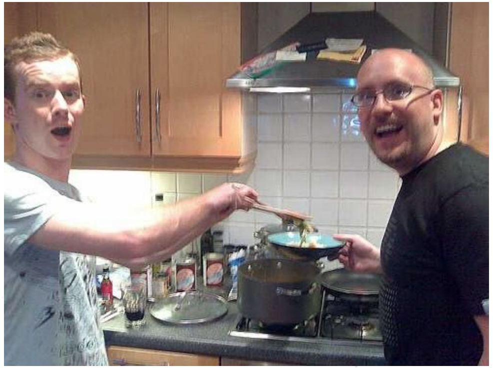
```
['</s>a man and a boy standing in a garden with a tree\n<pad>',
 '</s>a crane is being used to lift a large piece of equipment\n<pad>',
 '</s>a little girl is standing in front of a wooden house\n<pad><pad>',
 '</s>a man on a ladder cleaning a window of a building\n<pad><pad>',
 '</s>a man and a woman cooking in a kitchen with a stove\n<pad>',
 '</s>a man playing a guitar with another man in a studio\n<pad><pad>',
 '</s>a man sitting on a chair holding a stuffed animal in his lap\n',
 '</s>a woman on a skateboard talking on a cell phone\n<pad><pad>',
 '</s>a group of people standing in front of a building with a sign\n',
 '</s>a man doing a trick on a rail in the air\n<pad><pad>']
```

Grid Search for *s*

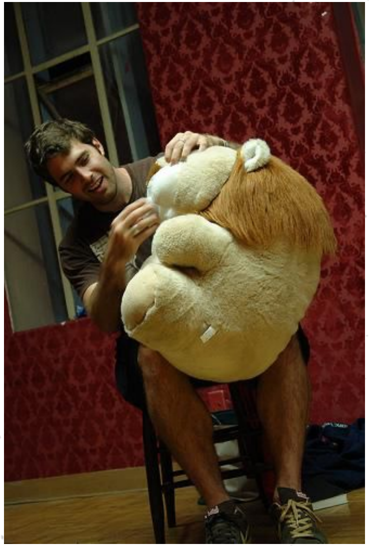
Original BLIP2 Descriptions



"a man on a ladder cleaning a window on a building"

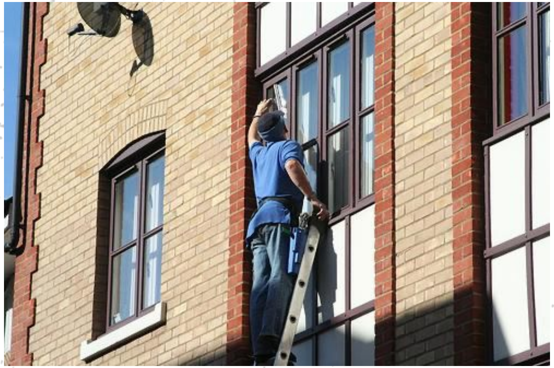


“two men in a kitchen preparing food in a pot on the stove”

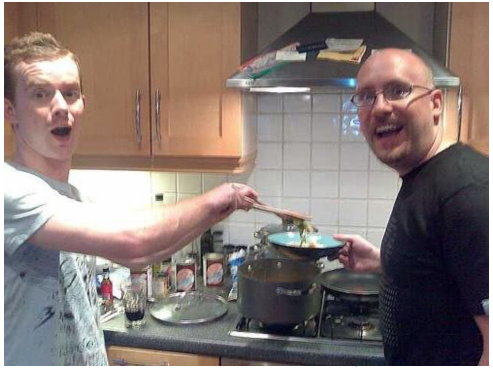


“a man sitting on a chair holding a stuffed animal lion”

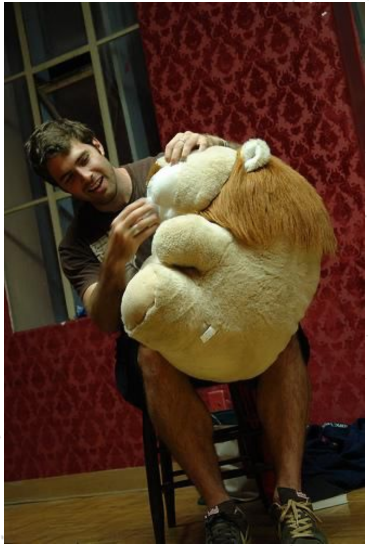
AWQ Descriptions, A16W4



“a man on a ladder cleaning a window of a building ”



“a man and a woman cooking in a kitchen with a stove”



“a man sitting on a chair holding a stuffed animal in his lap”