Deep Learning for Computer Vision

Beyond CLIP: BLIP, BLIP-2 and CoCA

Vineeth N Balasubramanian

Department of Computer Science and Engineering
Indian Institute of Technology, Hyderabad

Where does CLIP fail?

- Model perspective
  - Encoder models are not straightforward to transfer to text generation tasks

Where does CLIP fail?

- Model perspective
  - Encoder models are not straightforward to transfer to text generation tasks
- Data perspective
  - Noisy web scraped text is sub-optimal for vision-language learning

T_w: "a week spent at our rented beach house in Sandbridge"

T_s: "an outdoor walkway on a grass covered hill"

T_w: "that's what a sign says over the door"

T_s: "the car is driving past a small old building"

T_w: "hand held through the glass in my front bedroom window"

T_s: "a moon against the night sky with a black background"

T_w: "stunning sky over walney island, lake district, july 2009"

T_s: "an outdoor walkway on a grass covered hill"

T_w: "living in my little white house"

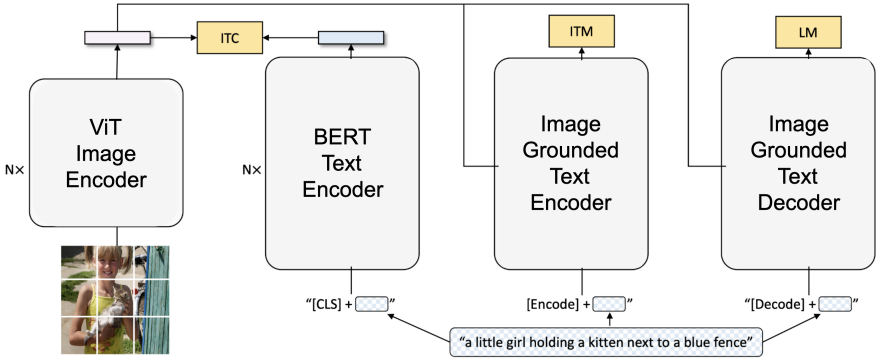
T_s: "a tiny white flower with a bee in it"

T_w: "the pink rock from below"

T_s: "some colorful trees that are on a hill in the mountains"

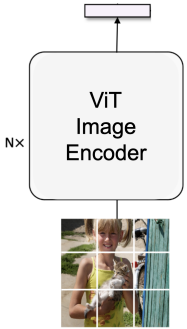
T_w: Text scraped from web. T_s: Synthetic text

BLIP[1]: Solving the Model Problem



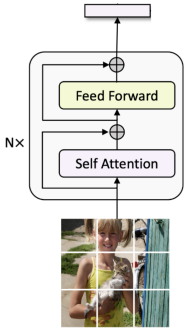
[1]Li et al, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation", ICML 2022

BLIP[1]: Solving the Model Problem



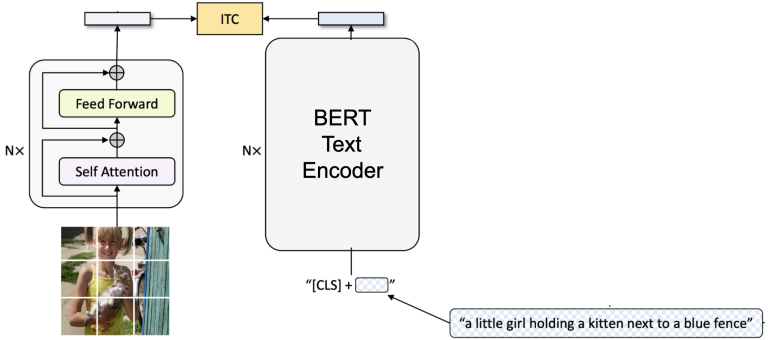
[1]Li et al, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation", ICML 2022

BLIP[1]: Solving the Model Problem



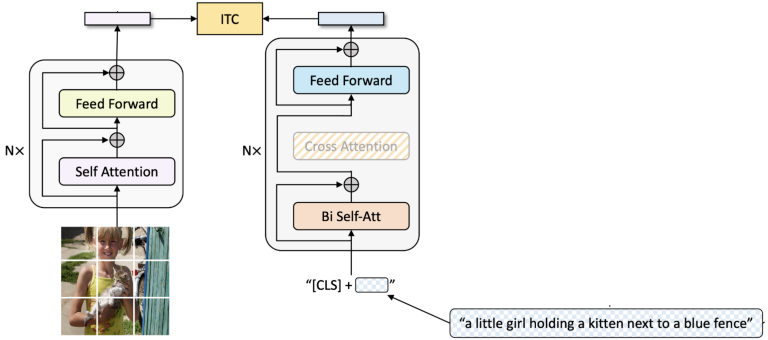
[1]Li et al, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation", ICML 2022

BLIP[1]: Solving the Model Problem



[1]Li et al, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation", ICML 2022

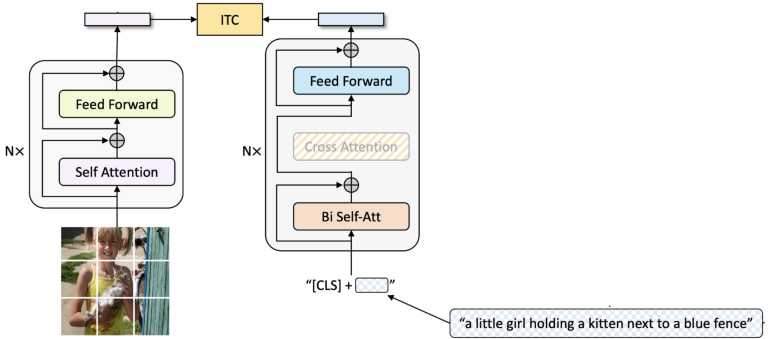
BLIP[1]: Solving the Model Problem



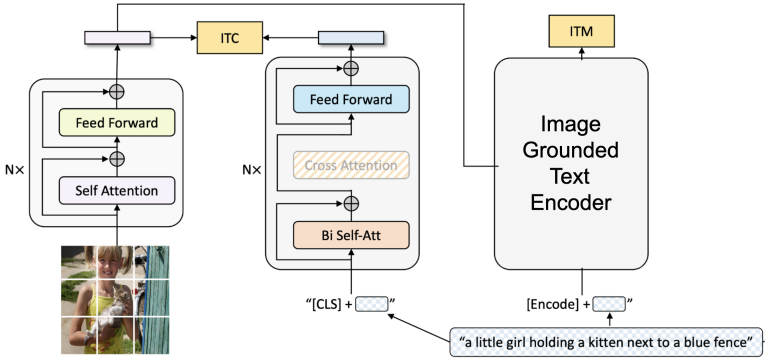
[1]Li et al, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation", ICML 2022

BLIP: Solving the Model Problem

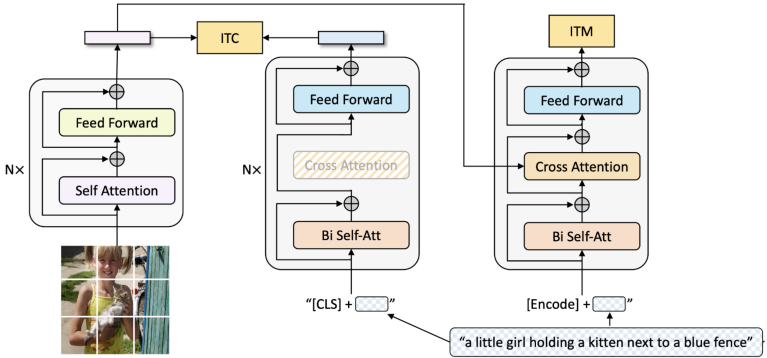
Image-Text Contrastive loss (L_{ITC}) aims to align the feature space of the visual transformer and the text transformer by encouraging positive image-text pairs to have similar representations in contrast to the negative pairs



BLIP: Solving the Model Problem

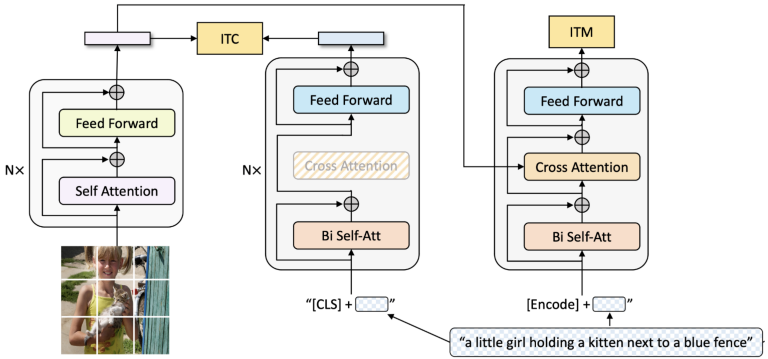


BLIP: Solving the Model Problem

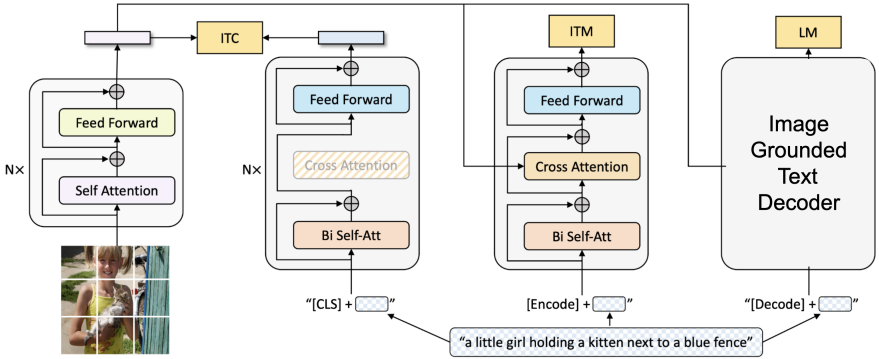


BLIP: Solving the Model Problem

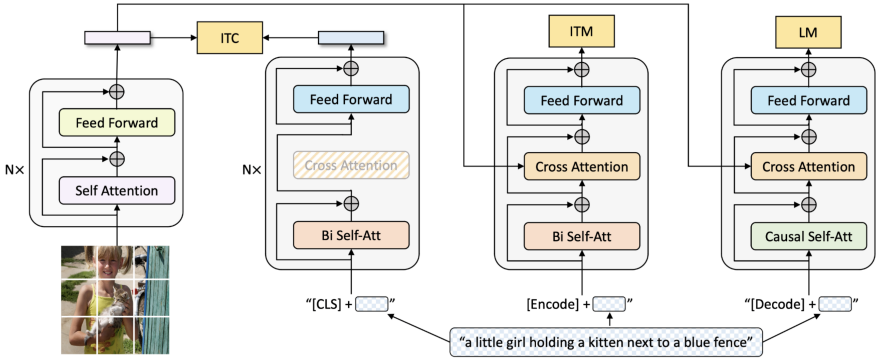
Image-Text Matching loss (L_{ITM}) aims to learn image-text multimodal representation that captures the fine-grained alignment between vision and language



BLIP: Solving the Model Problem

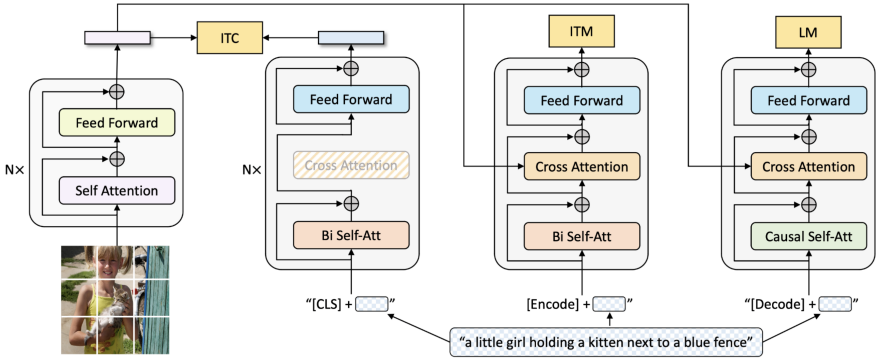


BLIP: Solving the Model Problem

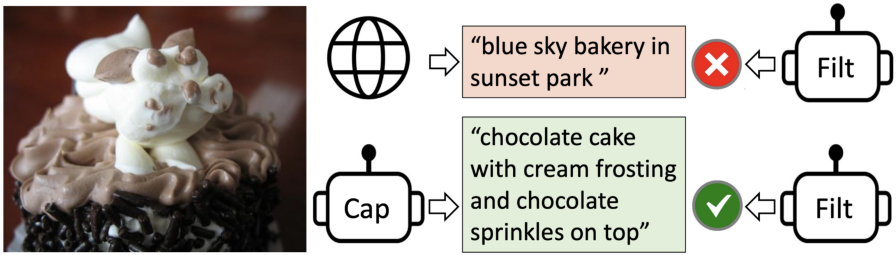


BLIP: Solving the Model Problem

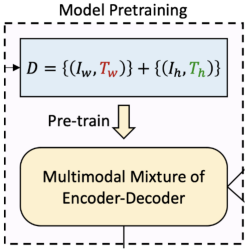
Language Modeling loss (L_{LM}) aims to generate textual descriptions given an image. It optimizes a cross entropy loss which trains the model to maximize the likelihood of the text in an autoregressive manner



BLIP: Solving the Noisy Text Problem - CapFilt



BLIP: Combining Both Components



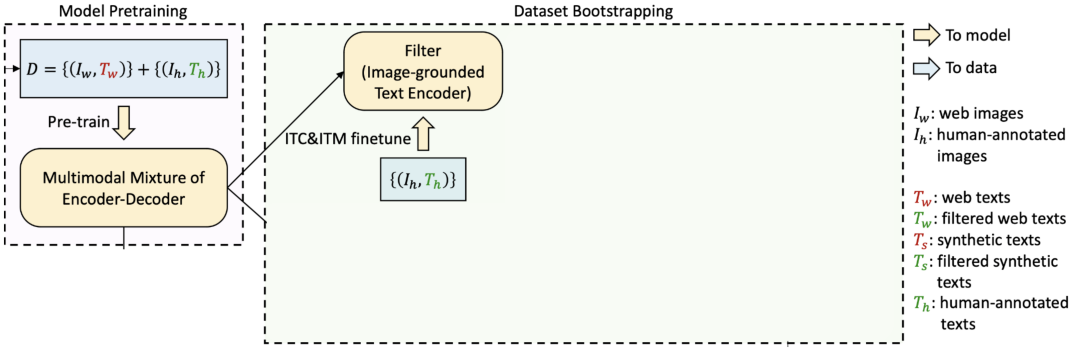
To model

To data

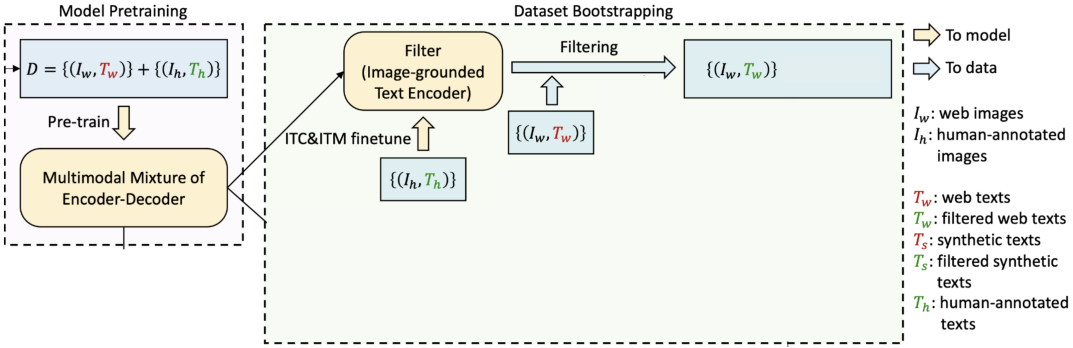
I_w: web images
I_h: human-annotated images

T_w: web texts
T_w: filtered web texts
T_s: synthetic texts
T_s: filtered synthetic texts
T_h: human-annotated texts

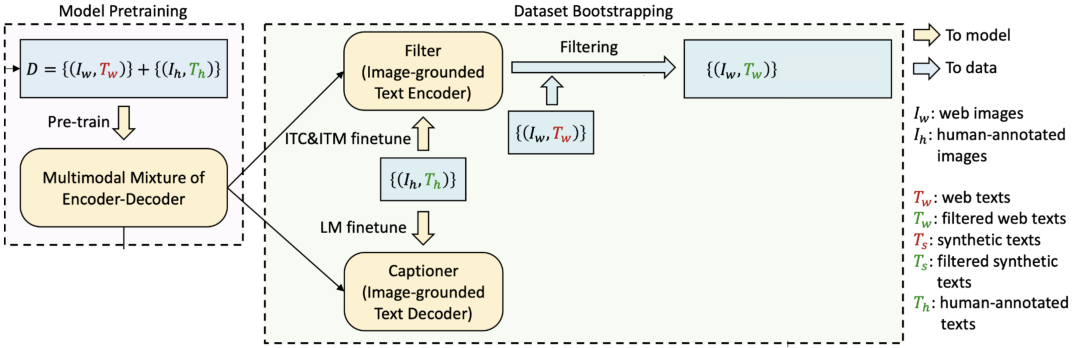
BLIP: Combining Both Components



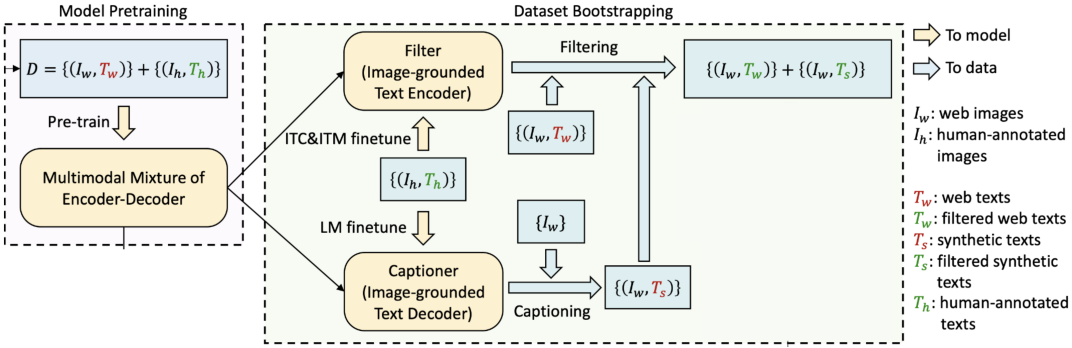
BLIP: Combining Both Components



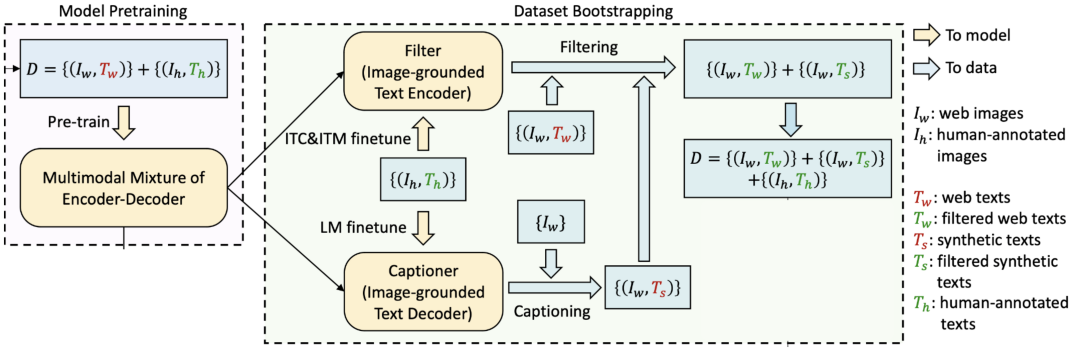
BLIP: Combining Both Components



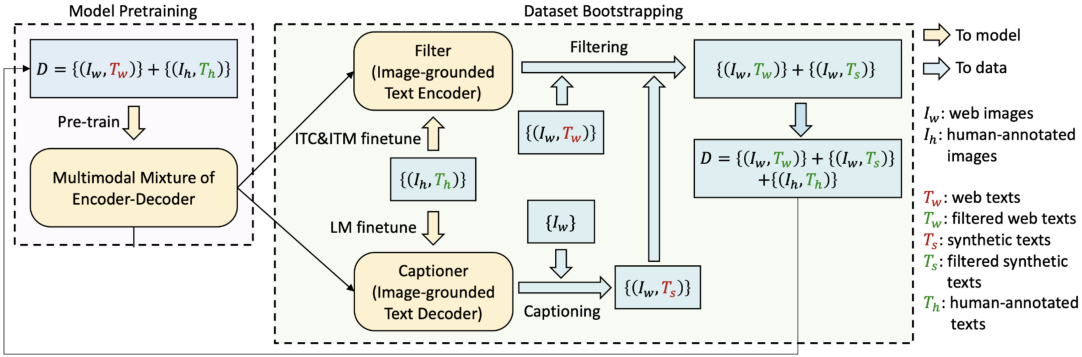
BLIP: Combining Both Components



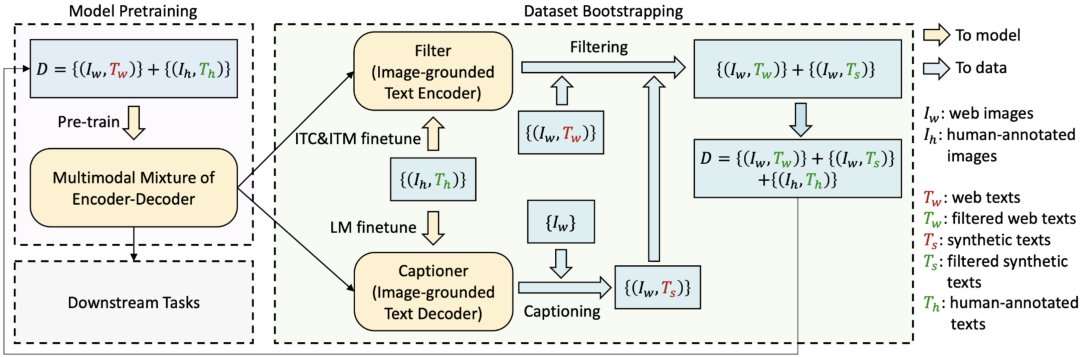
BLIP: Combining Both Components



BLIP: Combining Both Components



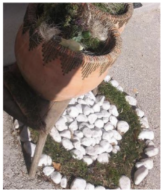
BLIP: Combining Both Components



BLIP: Qualitative Examples

T_w: "from bridge near my house"

T_s: "a flock of birds flying over a lake at sunset"



T_w: "in front of a house door in Reichenfels, Austria"

T_s: "a potted plant sitting on top of a pile of rocks"

T_w: "the current castle was built in 1180, replacing a 9th century wooden castle"

T_s: "a large building with a lot of windows on it"

BLIP-2[2]: Contributions

- Employing frozen models
  - Cost of vision-and-language pre-training has become increasingly prohibitive due to end-to-end training of large-scale models
  - BLIP-2, as a first attempt, employs frozen backbones for both vision and language tasks

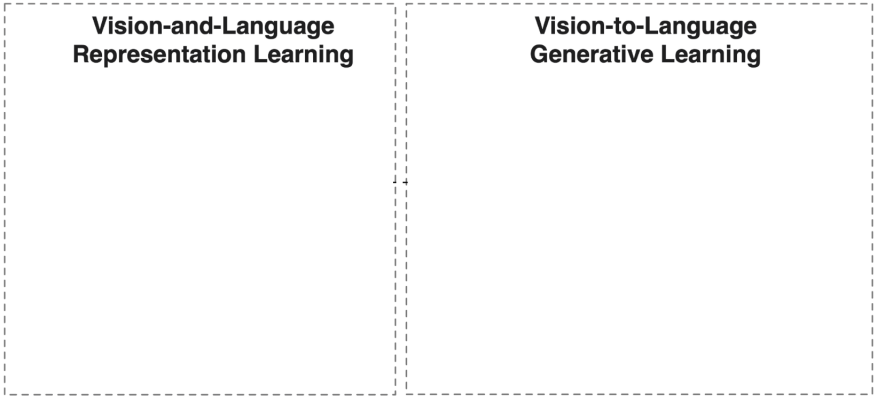
[2] Li et al, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models", ICML 2023

BLIP-2[2]: Contributions

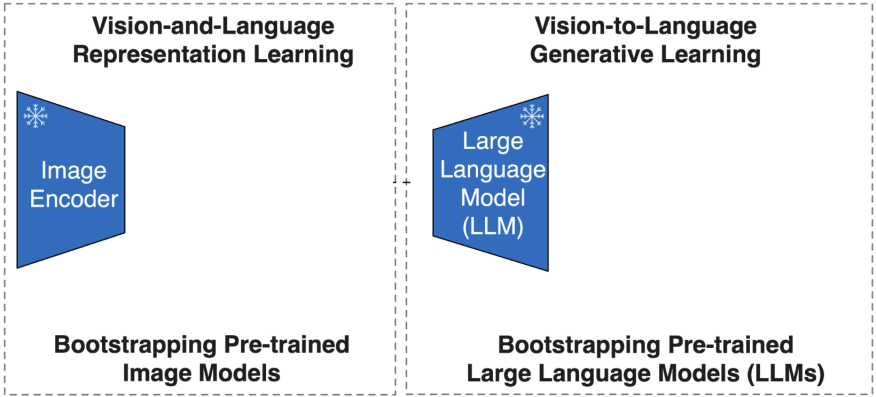
- Employing frozen models
  - Cost of vision-and-language pre-training has become increasingly prohibitive due to end-to-end training of large-scale models
  - BLIP-2, as a first attempt, employs frozen backbones for both vision and language tasks
- Addressing modality gap
  - Utilizing frozen pre-trained models may introduce a modality gap
  - BLIP-2 tackles this challenge through a dedicated pipeline

[2] Li et al, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models", ICML 2023

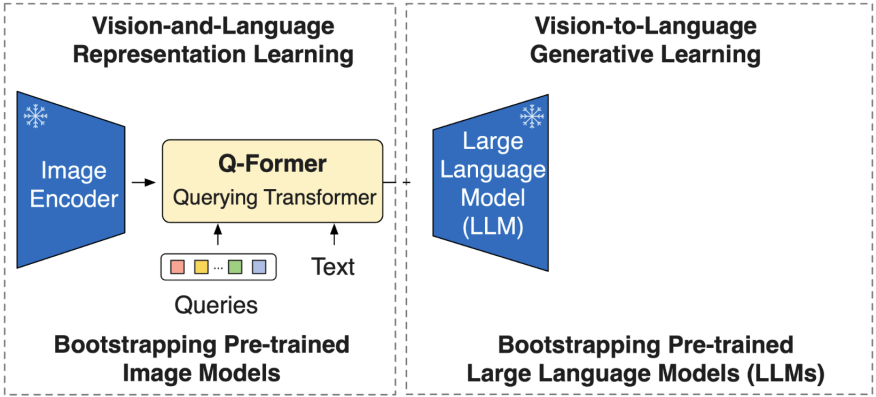
BLIP-2: Overview



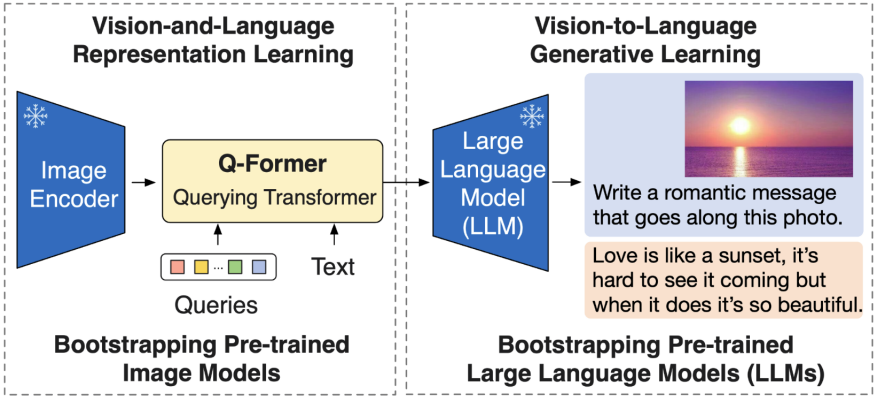
BLIP-2: Overview



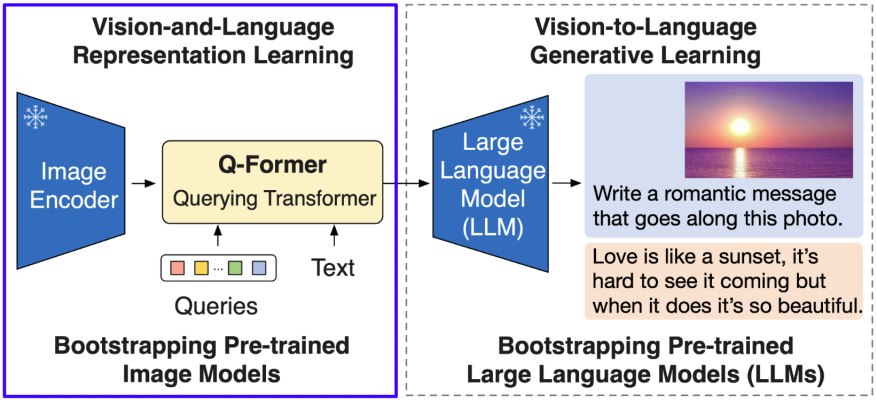
BLIP-2: Overview



BLIP-2: Overview

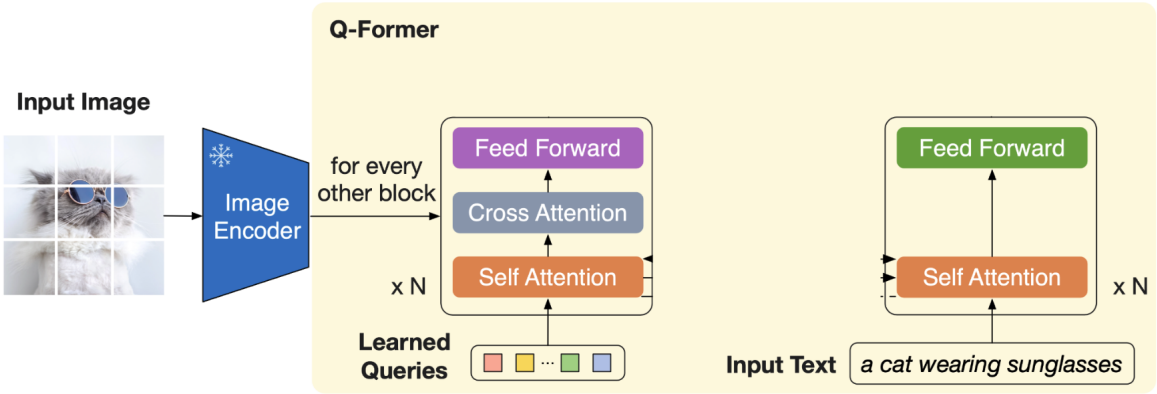


BLIP-2: Overview

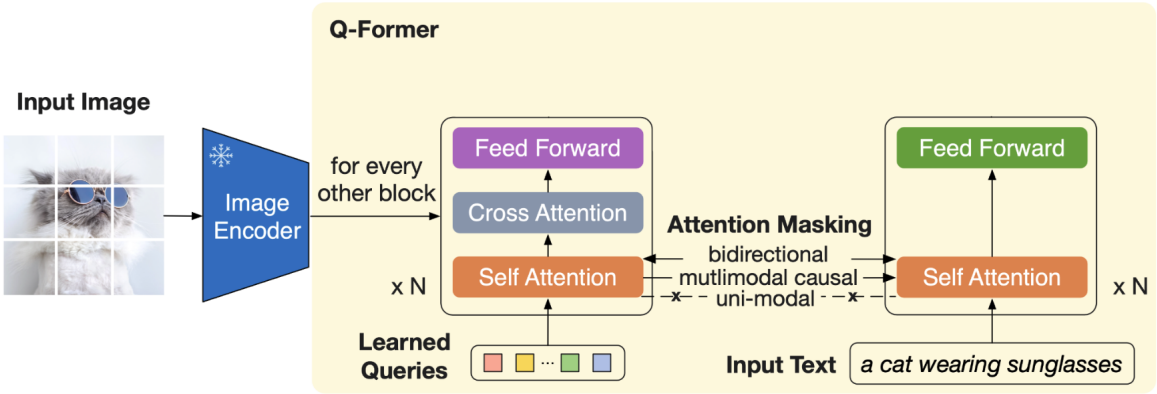


BLIP-2: Representation Learning

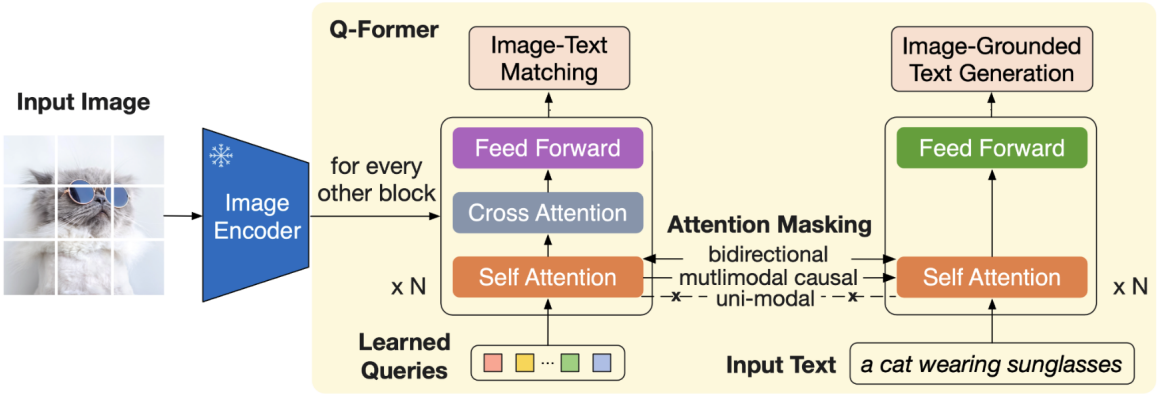
BLIP-2: Representation Learning



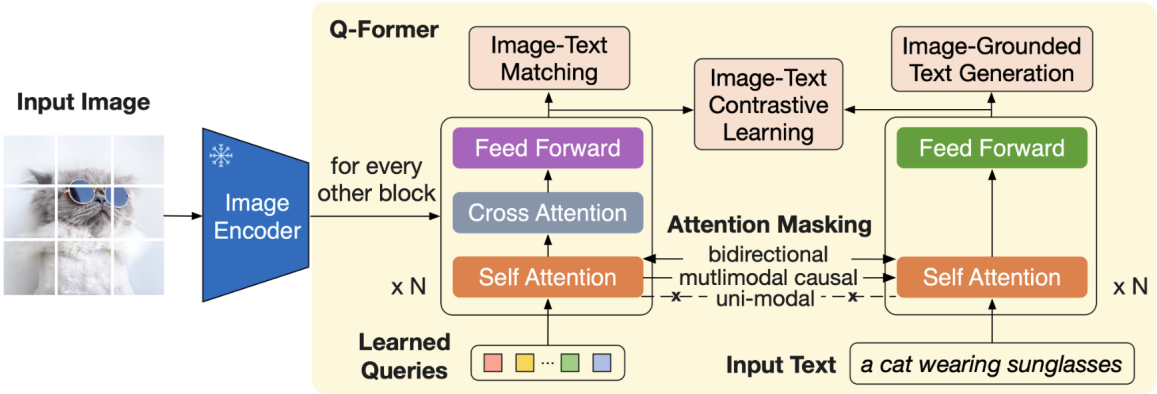
BLIP-2: Representation Learning



BLIP-2: Representation Learning



BLIP-2: Representation Learning



BLIP-2: Representation Learning - Masking Strategies

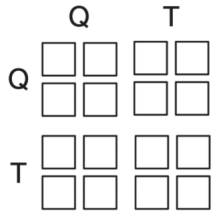
Q: query token positions; **T**: text token positions.
■ masked □ unmasked

Bi-directional Self-Attention Mask	Multi-modal Causal Self-Attention Mask	Uni-modal Self-Attention Mask

BLIP-2: Representation Learning - Masking Strategies

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

■ masked □ unmasked



Bi-directional Self-Attention Mask

Image-Text Matching

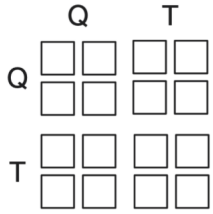
Multi-modal Causal Self-Attention Mask

Uni-modal Self-Attention Mask

BLIP-2: Representation Learning - Masking Strategies

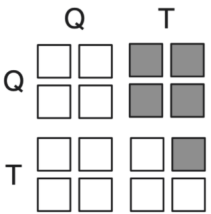
Q: query token positions; **T**: text token positions.

■ masked □ unmasked



Bi-directional Self-Attention Mask

Image-Text Matching



Multi-modal Causal Self-Attention Mask

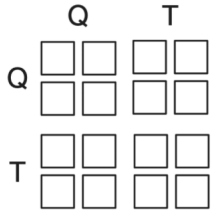
Image-Grounded Text Generation

Uni-modal Self-Attention Mask

BLIP-2: Representation Learning - Masking Strategies

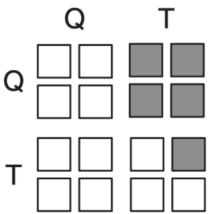
Q: query token positions; **T**: text token positions.

■ masked □ unmasked



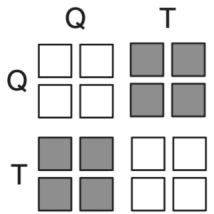
Bi-directional Self-Attention Mask

Image-Text Matching



Multi-modal Causal Self-Attention Mask

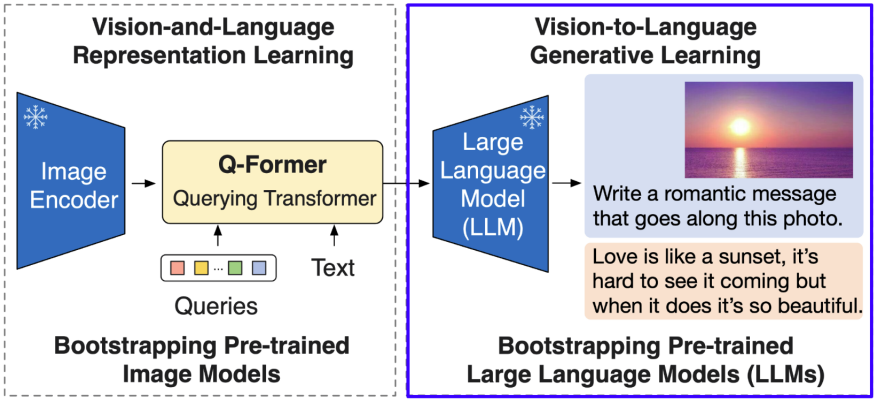
Image-Grounded Text Generation



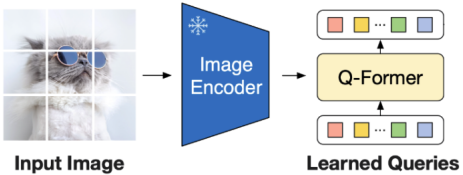
Uni-modal Self-Attention Mask

Image-Text Contrastive Learning

BLIP-2: Generative Learning

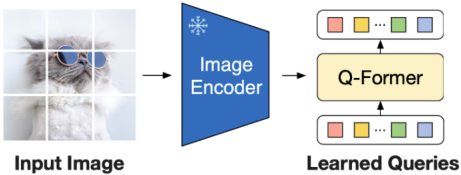


BLIP-2: Generative Learning

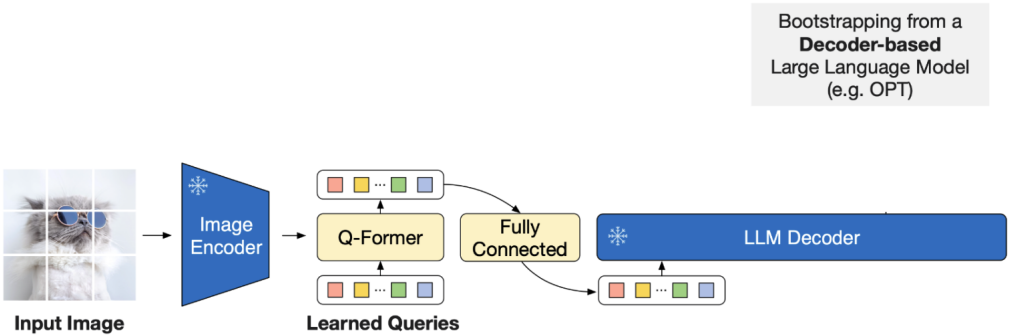


BLIP-2: Generative Learning

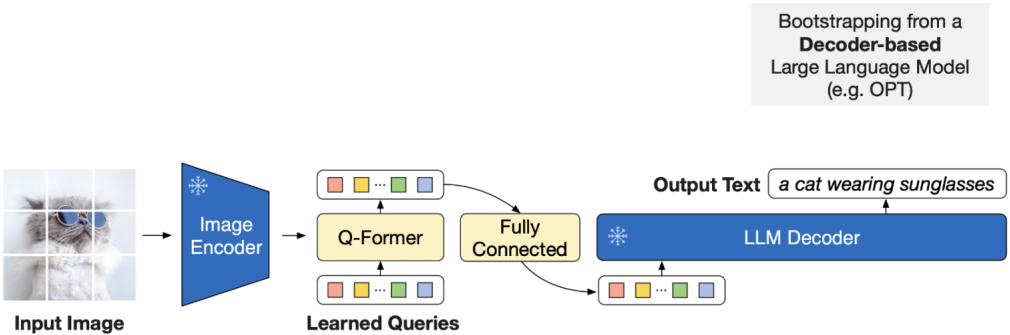
Bootstrapping from a **Decoder-based** Large Language Model (e.g. OPT)



BLIP-2: Generative Learning

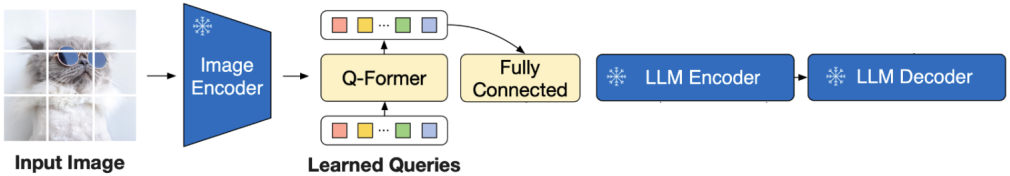


BLIP-2: Generative Learning

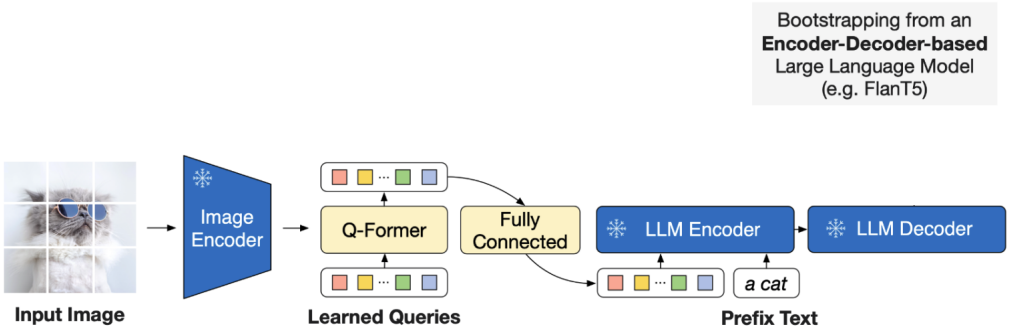


BLIP-2: Generative Learning

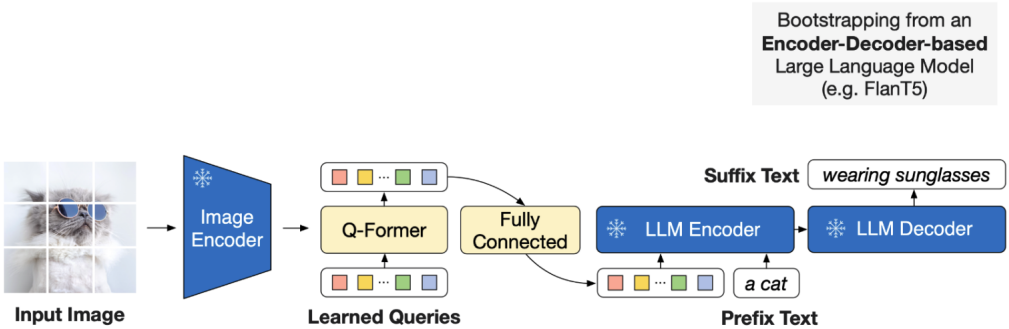
Bootstrapping from an **Encoder-Decoder-based** Large Language Model (e.g. FlanT5)



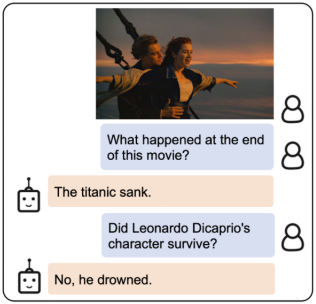
BLIP-2: Generative Learning



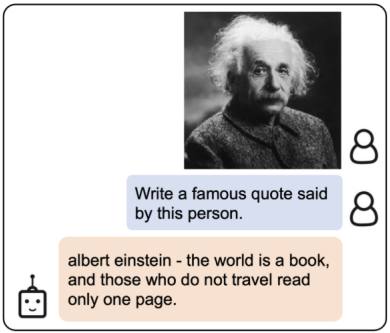
BLIP-2: Generative Learning



BLIP-2: Qualitative Examples



BLIP-2: Qualitative Examples



Inaccurate knowledge
(quote is from a different person)

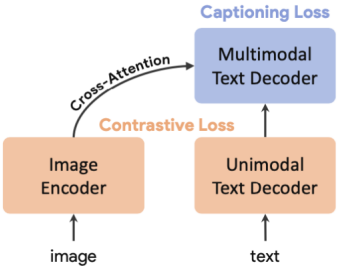
Incorrect reasoning path
(should have considered weather)

CoCa[3]: Contributions

- Unifying Paradigms
  - Introduces a novel approach that amalgamates single-encoder, dual-encoder, and encoder-decoder paradigms
  - It involves training a single image-text foundation model that encompasses the functionalities of all three approaches

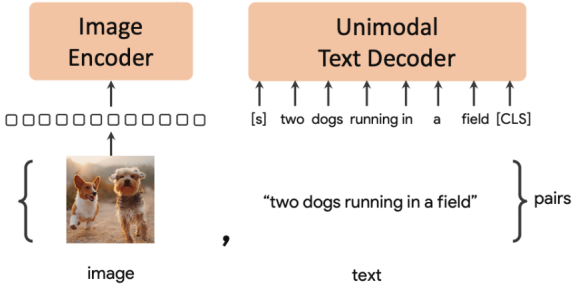
[3] Yu et al, "CoCa: Contrastive Captioners are Image-Text Foundation Models", TMLR 2022

CoCa: Overview

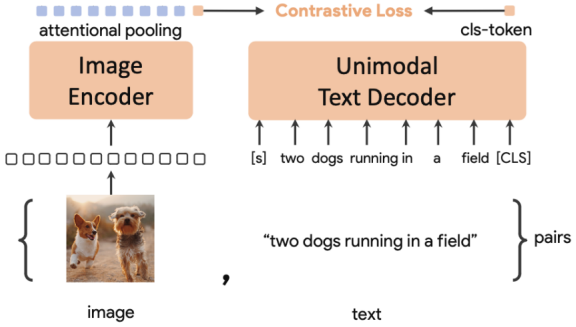


CoCa: Pre-training

CoCa: Pre-training

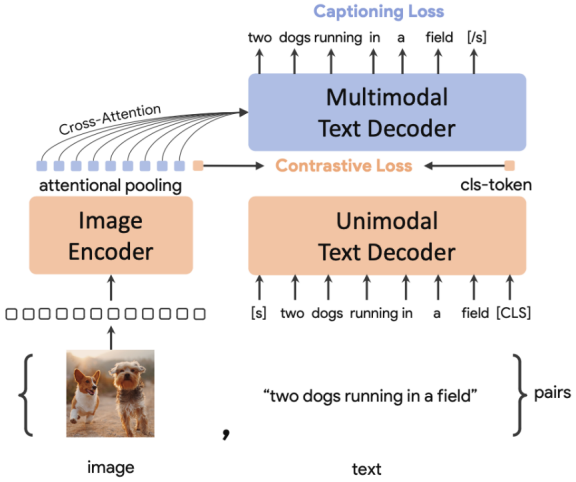


CoCa: Pre-training



$$\mathcal{L}_{\text{Con}} = -\frac{1}{N}\Big(\underbrace{\sum_{i}^{N}\log\frac{\exp(x_i^\top y_i/\sigma)}{\sum_{j=1}^{N}\exp(x_i^\top y_j/\sigma)}}_{\text{image-to-text}} + \underbrace{\sum_{i}^{N}\log\frac{\exp(y_i^\top x_i/\sigma)}{\sum_{j=1}^{N}\exp(y_i^\top x_j/\sigma)}}_{\text{text-to-image}}\Big)$$

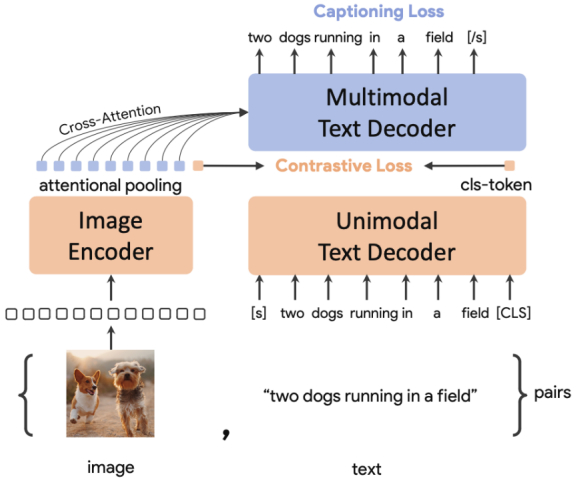
CoCa: Pre-training



$$\mathcal{L}_{\text{Con}} = -\frac{1}{N}\Bigg(\underbrace{\sum_{i}^{N}\log\frac{\exp(x_i^\top y_i/\sigma)}{\sum_{j=1}^{N}\exp(x_i^\top y_j/\sigma)}}_{\text{image-to-text}} + \underbrace{\sum_{i}^{N}\log\frac{\exp(y_i^\top x_i/\sigma)}{\sum_{j=1}^{N}\exp(y_i^\top x_j/\sigma)}}_{\text{text-to-image}}\Bigg)$$

$$\mathcal{L}_{\text{Cap}} = -\sum_{t=1}^{T}\log P_\theta(y_t|y_{<t}, x)$$

CoCa: Pre-training

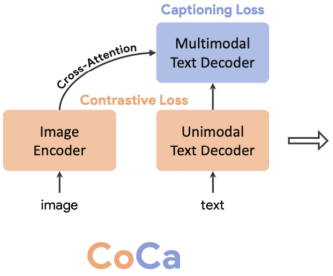


$$\mathcal{L}_{\text{Con}} = -\frac{1}{N}\Bigg(\underbrace{\sum_{i}^{N}\log\frac{\exp(x_i^\top y_i/\sigma)}{\sum_{j=1}^{N}\exp(x_i^\top y_j/\sigma)}}_{\text{image-to-text}} + \underbrace{\sum_{i}^{N}\log\frac{\exp(y_i^\top x_i/\sigma)}{\sum_{j=1}^{N}\exp(y_i^\top x_j/\sigma)}}_{\text{text-to-image}}\Bigg)$$

$$\mathcal{L}_{\text{Cap}} = -\sum_{t=1}^{T}\log P_\theta(y_t|y_{<t}, x)$$

$$\mathcal{L}_{\text{CoCa}} = \lambda_{\text{Con}}\cdot\mathcal{L}_{\text{Con}} + \lambda_{\text{Cap}}\cdot\mathcal{L}_{\text{Cap}}$$

CoCa: Finetuning

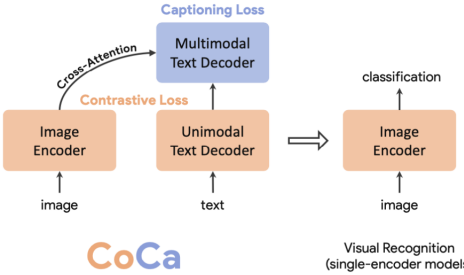


Pretraining

Zero-shot, frozen-feature or finetuning

CoCa: Finetuning

$$\mathcal{L}_{\text{Cls}} = -p(y) \log q_\theta(x)$$

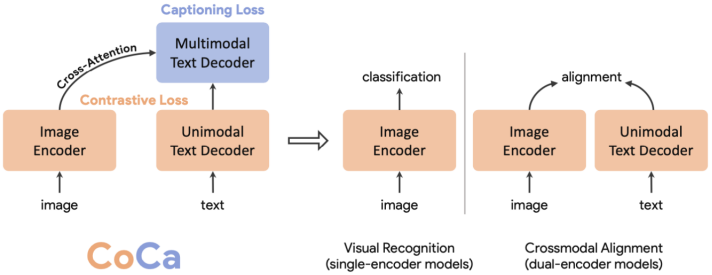


Pretraining

Zero-shot, frozen-feature or finetuning

CoCa: Finetuning

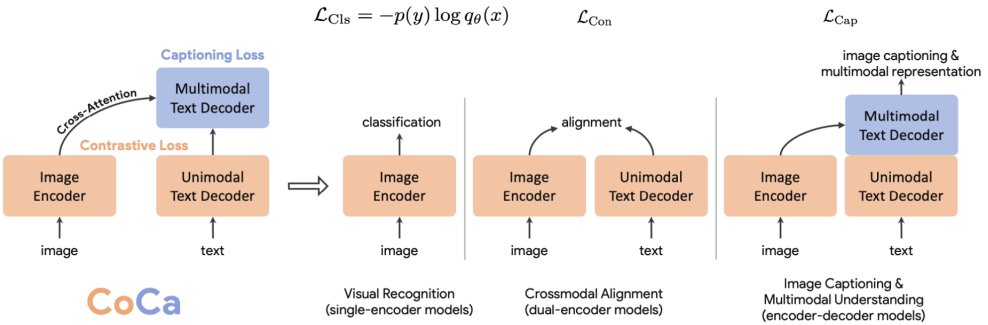
$$\mathcal{L}_{\text{Cls}} = -p(y) \log q_\theta(x) \qquad \mathcal{L}_{\text{Con}}$$



Pretraining

Zero-shot, frozen-feature or finetuning

CoCa: Finetuning



CoCa: Qualitative Examples

Curated samples of text captions generated by CoCa with NoCaps images as input.

Homework

Readings

- Lilian Weng, Generalized Visual Language Models
- Hugging Face, A Dive into Vision-Language Models

References