Vision-Language Pretraining: Current Trends and the Future

Aida Nematzadeh

IPM Summer School
Sep 2022

Why Multimodal Pretraining?

Train once, use multiple times. Multimodal features are useful across a range of multimodal tasks and applications.

Q: What are the people waiting for?

A: bus

[Images are taken from https://vision-explorer.allenai.org/visual_question]

Why Multimodal Pretraining?

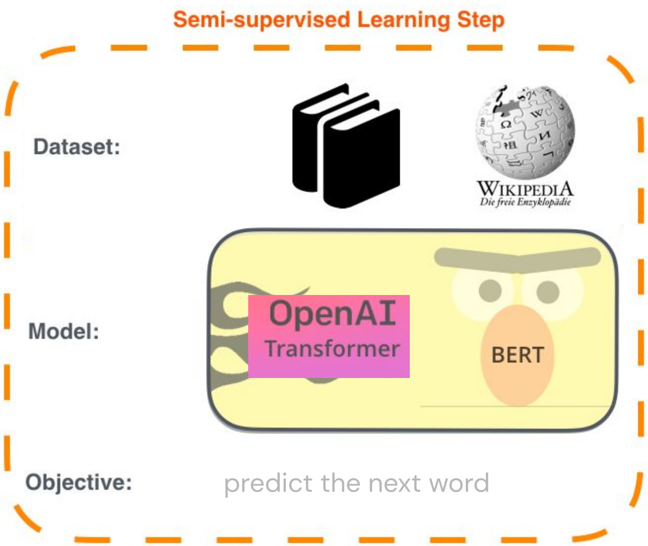
The ability to ground language to vision–multimodal pretraining–is a fundamental aspect of both language & vision.

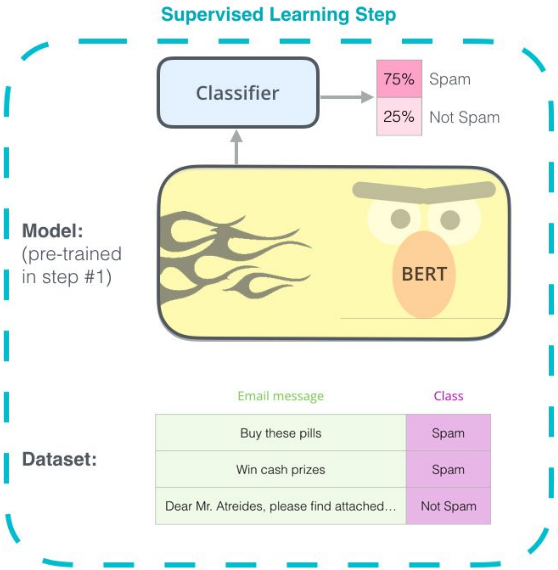
Q: What are the people waiting for?

A: bus

[Images are taken from https://vision-explorer.allenai.org/visual_question]

Success of Pretraining in NLP





Performance gain is due to architecture innovations & larger data. [Peters et al., 2018; Howard & Ruder, 2018; Devlin et al., 2018; Radford et al., 2018; Raffel et al., 2019; Rae et al. 2022]

[Images are from http://jalammar.github.io/illustrated-bert/]

Similar Models for Multimodal Pretraining?

Dataset:

"The scenic route through mountain ranges includes these unbelievably coloured mountains.

Model:

Objective:

predict the next word

Other objectives?

Dataset: image-text pairs where a given text describes its image.

Model: attention mechanisms over both image and text; preprocessing images to "visual tokens".

Objective: loss functions specific to the image modality and image-text pairs.

[Images are taken from http://jalammar.github.io/illustrated-bert/ , https://ai.googleblog.com/2018/09/conceptual-captions-new-dataset-and.html, https://visualqa.org/]

Multimodal Pretraining: How it Started

BERT
Devlin et al. NAACL 2019
(arxiv 2018)

VideoBert
Sun et al. ICCV 2019.

LXMert
Tan et al. EMNLP 2019.

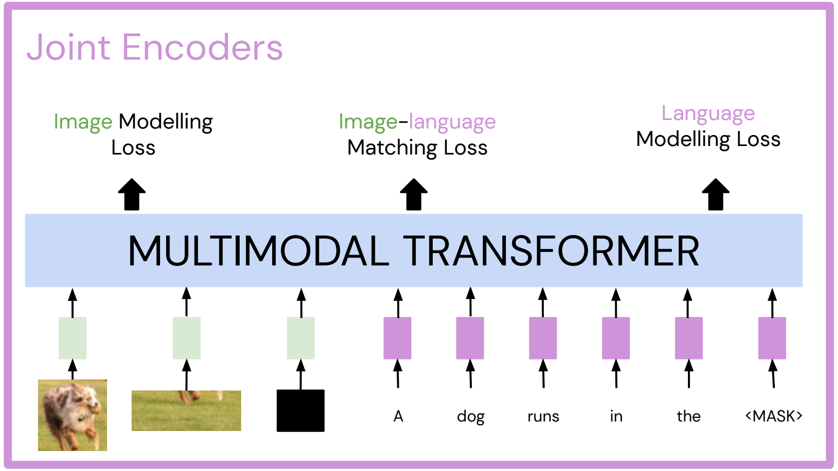
VilBERT
Lu et al. NeurIPS 2019.

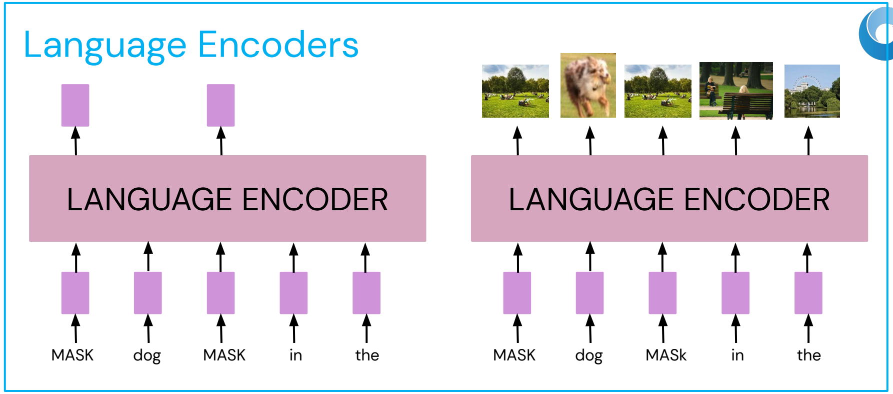
CBT
Sun et al. Arxiv 2019.

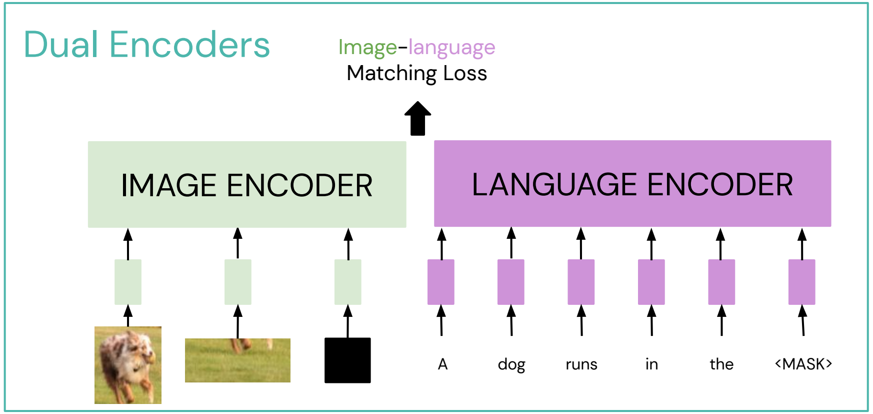
UNITER
Chen et al. ECCV 2020.

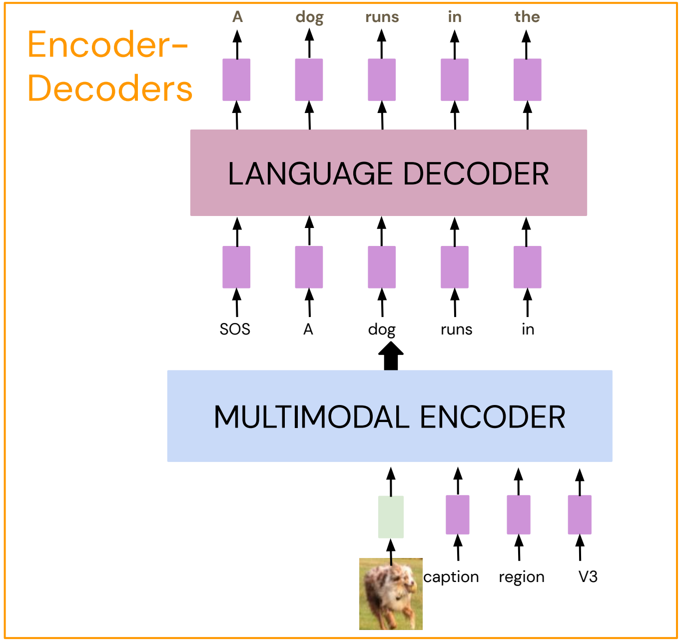
Models share the same "backbone" with slight differences in loss design, preprocessing, etc.

They achieve the state-of-the-art results in a range of tasks.

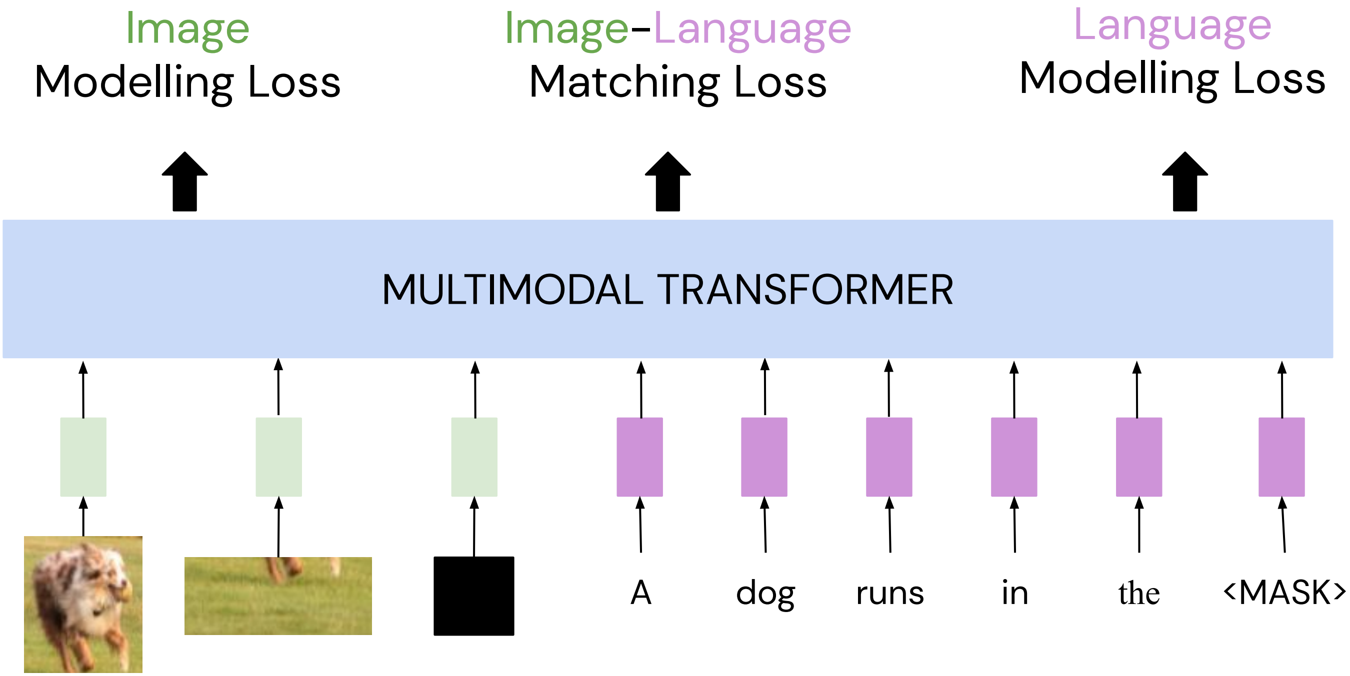

Joint Encoders
Image Modelling Loss
Image-language Matching Loss
Language Modelling Loss
MULTIMODAL TRANSFORMER
A
dog
runs
in
the
<MASK>


Language Encoders
LANGUAGE ENCODER
MASK
dog
MASK
in
the
LANGUAGE ENCODER
MASK
dog
MASk
in
the


Dual Encoders
Image-language Matching Loss
IMAGE ENCODER
LANGUAGE ENCODER
A
dog
runs
in
the
<MASK>

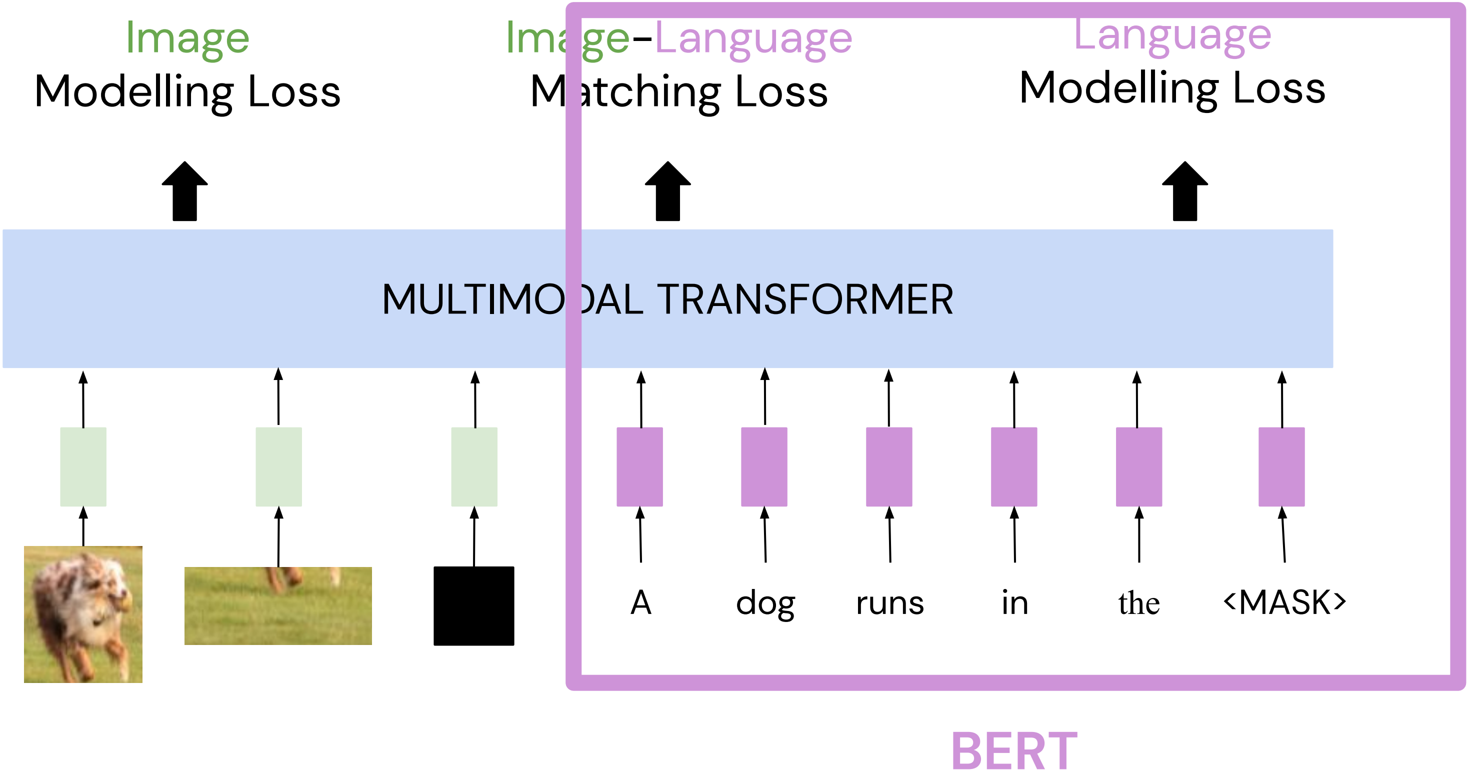

Encoder-Decoders
A
dog
runs
in
the
LANGUAGE DECODER
SOS
A
dog
runs
in
MULTIMODAL ENCODER
caption
region
V3

Multimodal Transformers (Joint Encoders)

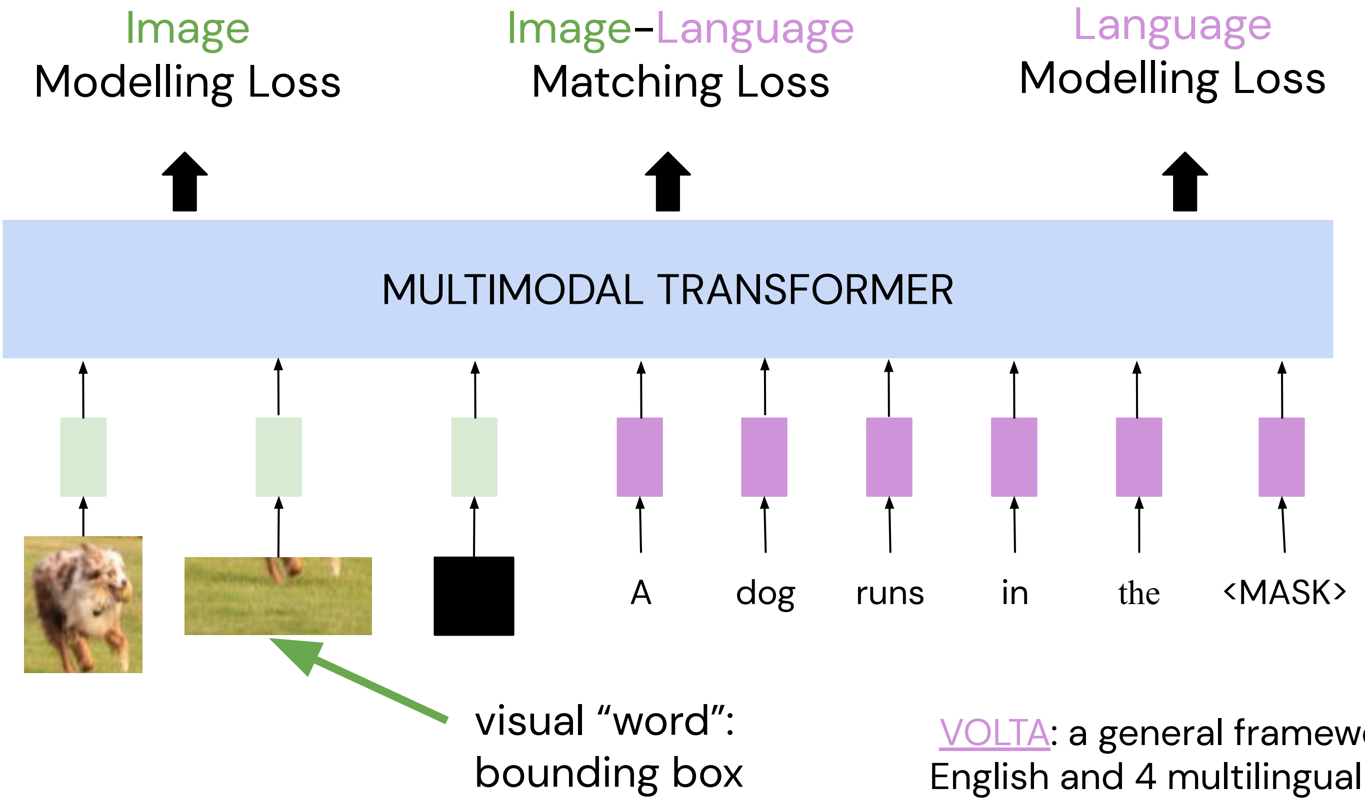


Thanks Lisa Anne Hendricks for sharing slides on MMT.

Multimodal Transformers (Joint Encoders)



Multimodal Transformers (Joint Encoders)



VOLTA: a general framework (with 5 English and 4 multilingual MMT).

What Contributes to these Models' Success?

Are results due to advances in the architecture or large pretraining datasets?

Are the "adopted" losses from language models good enough?

Is the cross-talk between modalities (via attention) important?

What makes a good pretraining dataset?

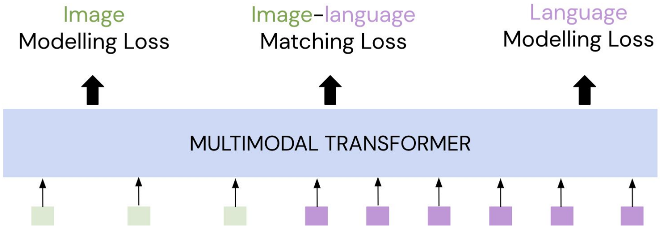
Evaluation: Zero-Shot Image Retrieval

Zero-shot image retrieval directly evaluates the goodness of **pretrained** representations.

Image Retrieval (IR)

"Grey haired man in black and yellow tie."

Typical Loss Functions



Language/image modeling: masked language/region modeling

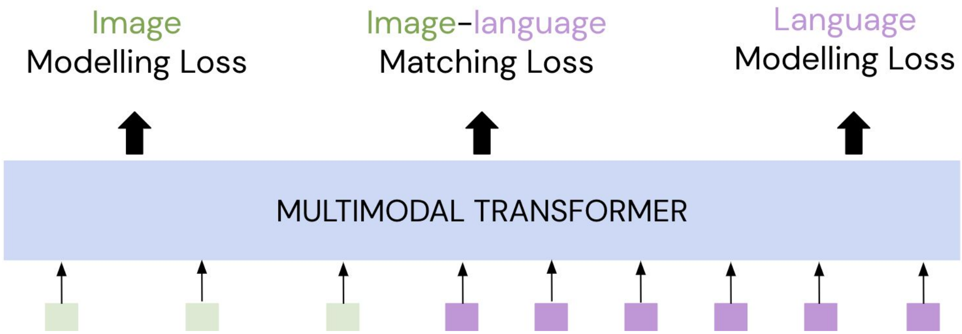
Image–language matching: binary classification or contrastive formulation

$$L_{i2t} = -\frac{1}{N}\sum_{i}^{N}\log\frac{\exp(x_i^\top y_i/\sigma)}{\sum_{j=1}^{N}\exp(x_i^\top y_j/\sigma)} \qquad L_{t2i} = -\frac{1}{N}\sum_{i}^{N}\log\frac{\exp(y_i^\top x_i/\sigma)}{\sum_{j=1}^{N}\exp(y_i^\top x_j/\sigma)}$$

[Taken from ALIGN]

Are All Losses Needed? [Hendricks et al. TACL 2021]

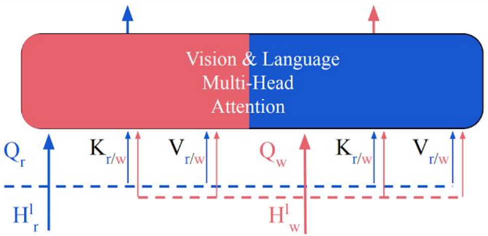
R@1	No Language Modeling Loss	No Image Modeling Loss	All Loses
Zeroshot Flickr	15.0	**41.1**	**40.7**



With the right hyper-parameters, **image modeling loss** is not needed.

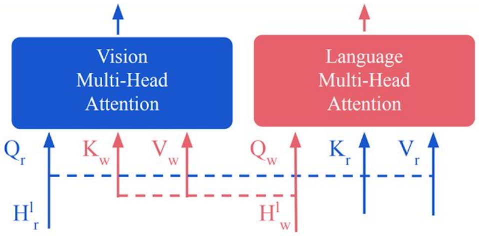
Vision-and-Language
or
Vision-for-Language?
[Frank et al, 2021]

Different Attention Mechanisms [Hendricks et al. TACL 2021]



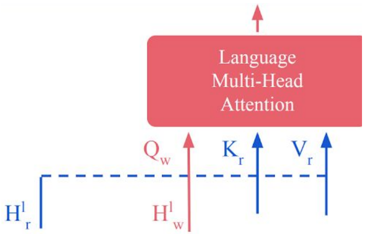
Merged attention

Each modality *attends* to **both** modalities.



Coattention

Each modality *attends* **only** to the other modality (two asymmetric attentions).

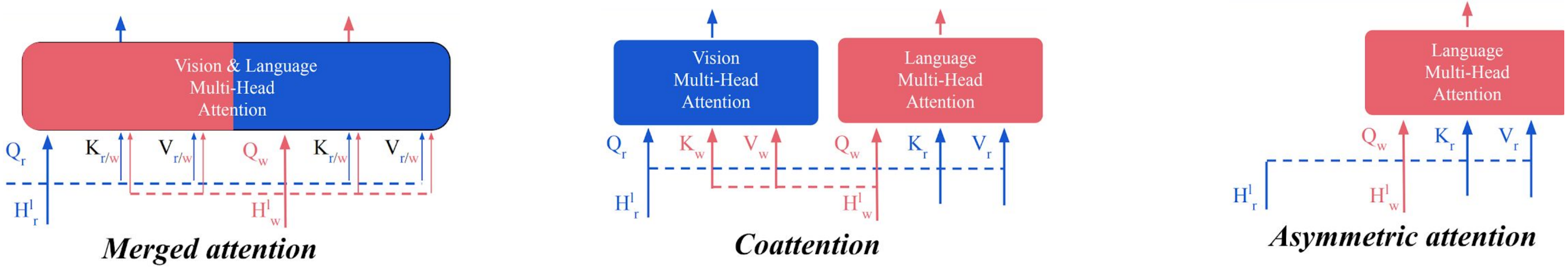


Asymmetric attention

Only one modality (*e.g.*, language) *attends* to the **other** modality (*e.g.*, image).

multimodal attention

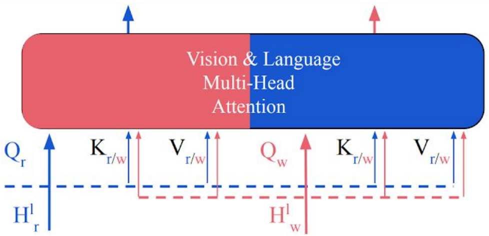
Different Attention Mechanisms [Hendricks et al. TACL 2021]



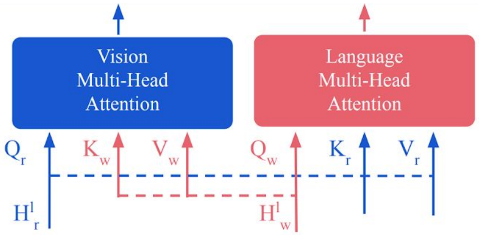
Similar performance

R@1	merged	coattention
Zeroshot Flickr	40.0	**41.9**

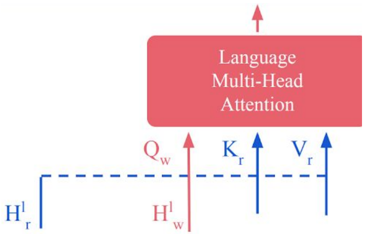
Different Attention Mechanisms [Hendricks et al. TACL 2021]



Merged attention



Coattention



Asymmetric attention

R@1	merged	coattention	asymmetric (language)	asymmetric (image)
Zeroshot Flickr	40.0	**41.9**	33.6	31.6

Multimodal Attention > Depth / Size [Hendricks et al. TACL 2021]

6 multimodal layers & 12 attention heads

R@1	coattention	asymmetric (language)	asymmetric (image)	coattention with 1 multimodal layer	coattention with 6 attention heads
Zeroshot Flickr	**41.9**	33.6	31.6	37.2	39.9

Depth and number of parameters alone are not enough.

What Contributes to these Models' Success?

Are results due to advances in the architecture or large pretraining datasets?

Are the "adopted" losses from language models good enough? No, we need better image modeling losses.

Is the cross-talk between modalities (via attention) important? Yes, multimodal attention is important.

What makes a good pretraining dataset?

Pretraining Datasets

MSCOCO

"The two people are walking down the beach."

MSCOCO/OI Narratives

"In this image we can see a bridge and sea. In the background, we can see trees and the sky. We can see so many people on the bridge. At the bottom of the image, we can see two people. We can see stairs in the right bottom of the image ..."

Visual Genome

small round yellow frisbee, man has cast on his arm, concrete trail path in the park, man wearing black sunglasses

manually annotated

Conceptual Captions

"The ***scenic route*** through mountain ranges includes these unbelievably coloured mountains.

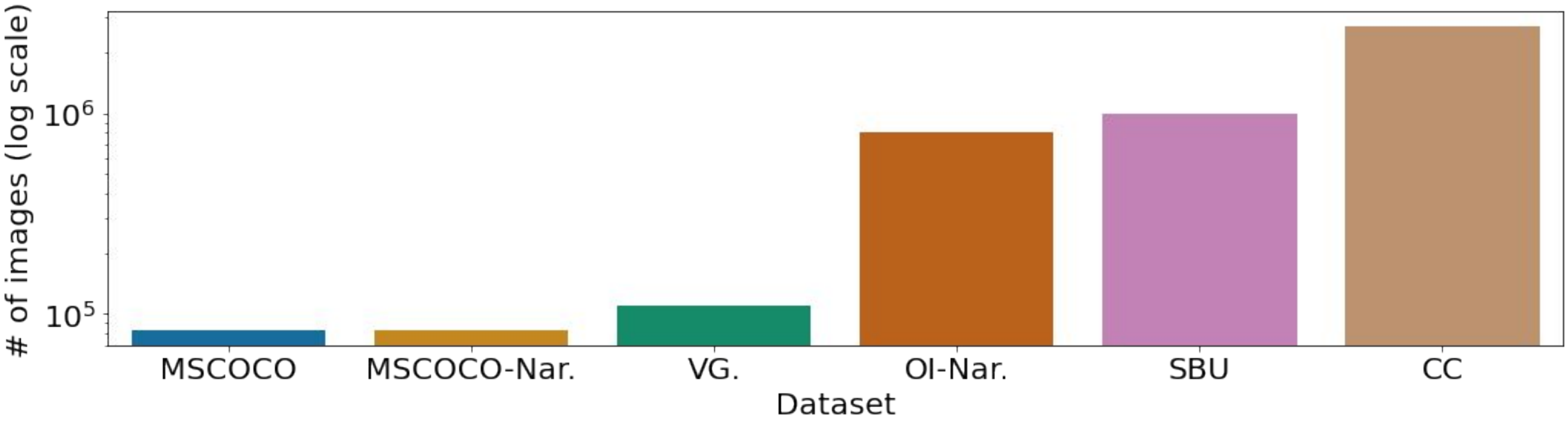
SBU Captions

"***King Arthur's*** beheading rock - right on the sidewalk in the middle of ***town***".

from "the wild"

Noisier image-text correspondence but larger

Dataset Considerations: Size



Dataset Considerations: Language

COCO style caption: “Single black dog sitting on the grass”

Narratives style caption: “The dog is black and brown. The collar is red. … The dog is on the grass. …”

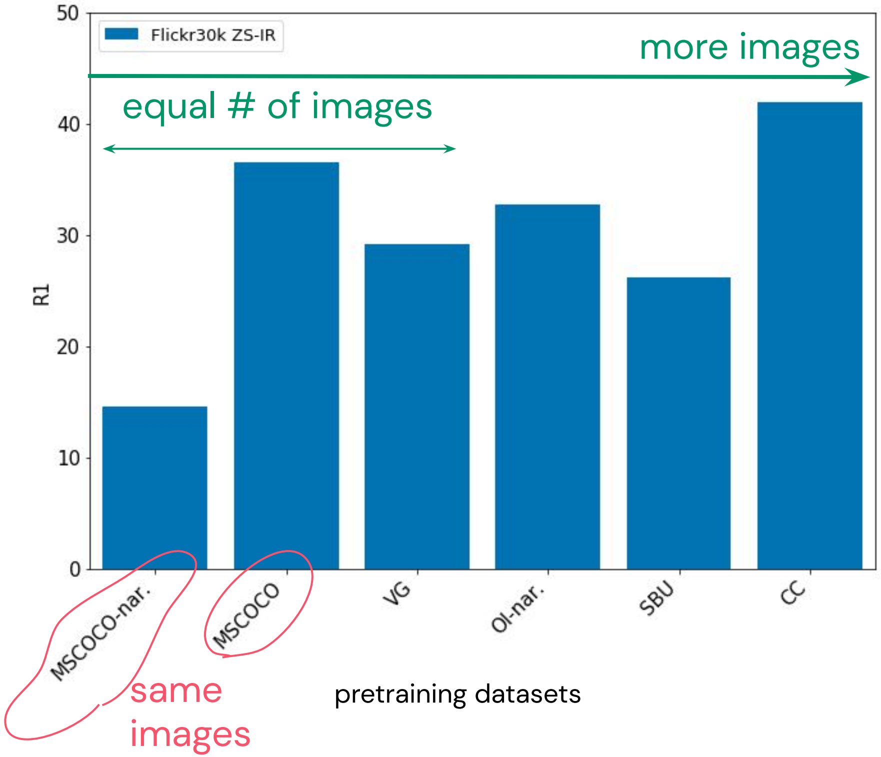
Genome style caption: “Black dog”

Dataset Considerations: Noise

"Single ***black dog sitting*** on the ***grass***"

"A ***person*** takes a ***dog*** on a ***walk*** near the ***river***."

Image Retrieval: Language Similarity Matters

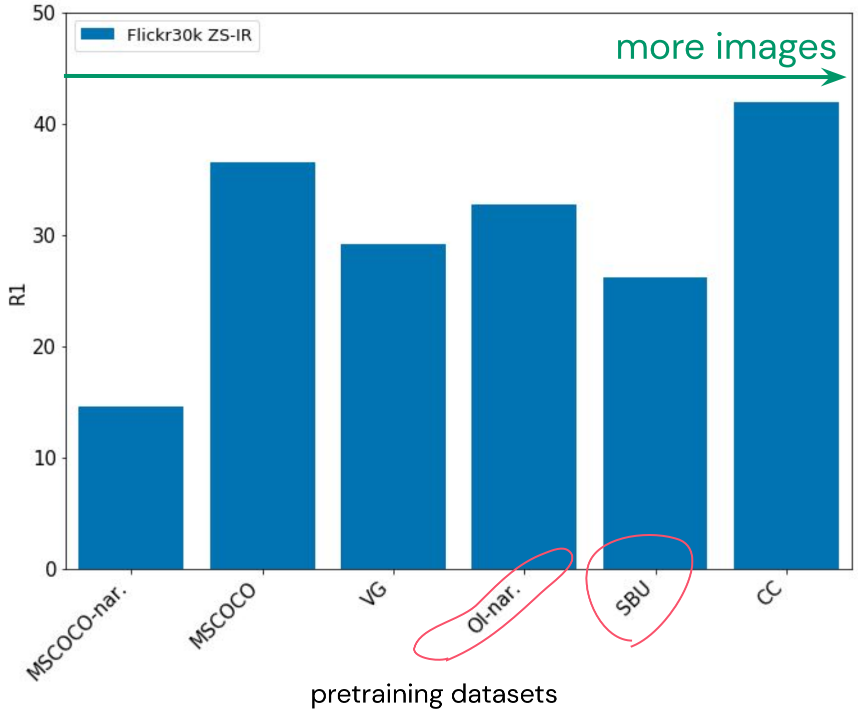


Performance is **not** directly correlated with the number of images.

Language similarity in pretraining & test (measured by perplexity) explains the difference in the results.

Hendricks et al. "Data, Architecture, or Losses: What Contributes Most to Multimodal Transformer Success?" TACL 2021.

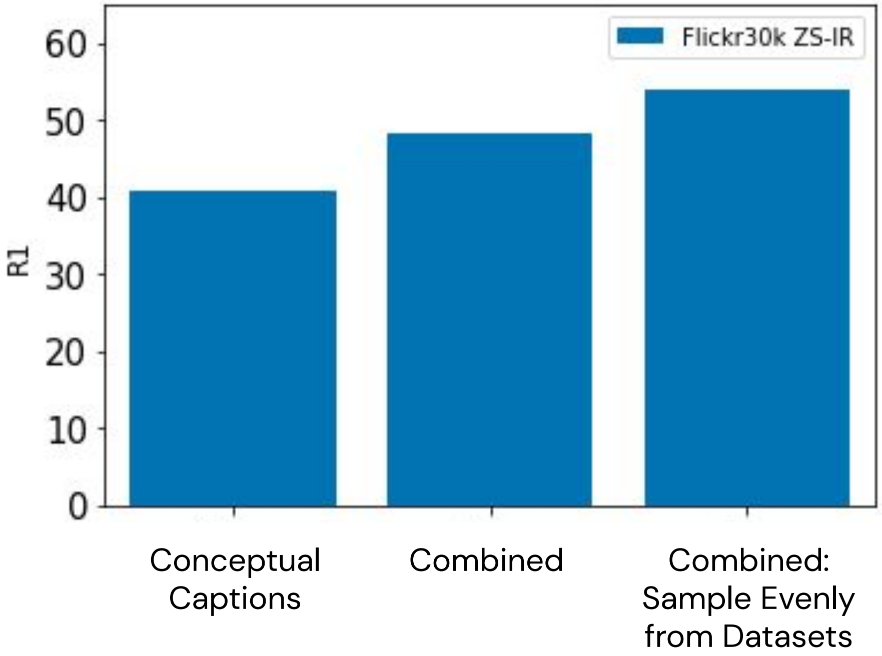
Image Retrieval: Noise Matters



SBU is larger than Open Images and has lower perplexity, but is still worse. However, SBU has more noise, meaning the language does not always describe the image content.

Hendricks et al. "Data, Architecture, or Losses: What Contributes Most to Multimodal Transformer Success?" TACL 2021.

Image Retrieval: Dataset Sampling Matters

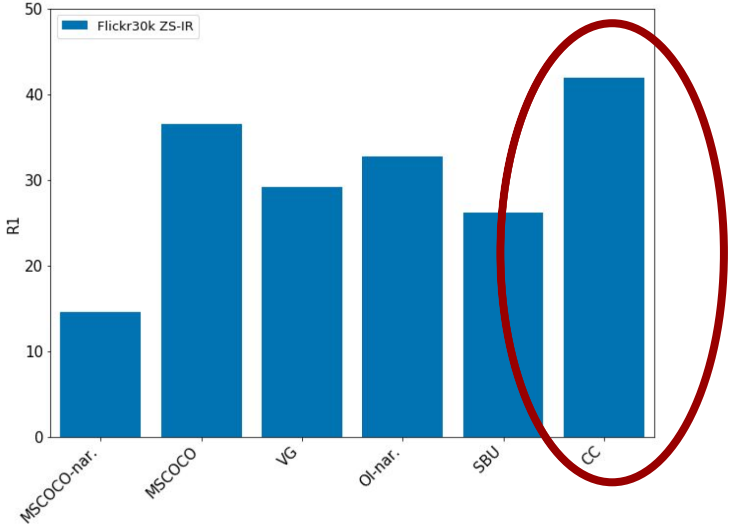


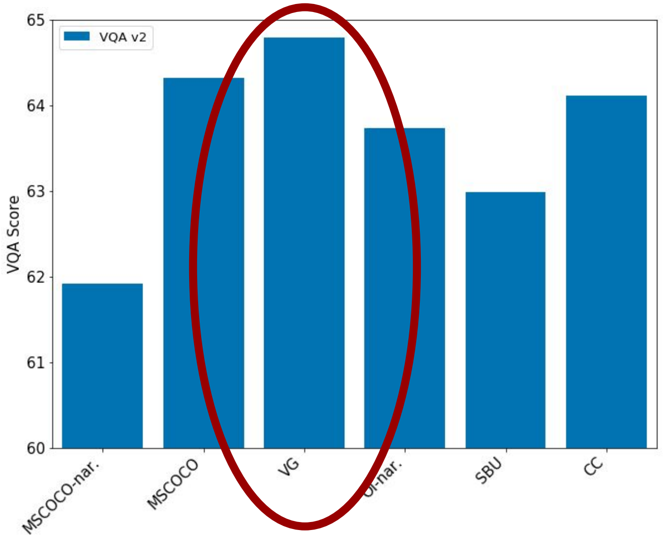
Combining datasets does lead to better results, but how we sample from combined datasets matter.

MSCOCO is a good dataset for pretraining; sampling method which weights MSCOCO images higher does better.

Hendricks et al. "Data, Architecture, or Losses: What Contributes Most to Multimodal Transformer Success?" TACL 2021.

"Best" Dataset is Task Dependent





Best datasets are different for IR (Conceptual Captions is best) and VQA (VG is best)

What Contributes to these Models' Success?

Are results due to advances in the architecture or large pretraining datasets?

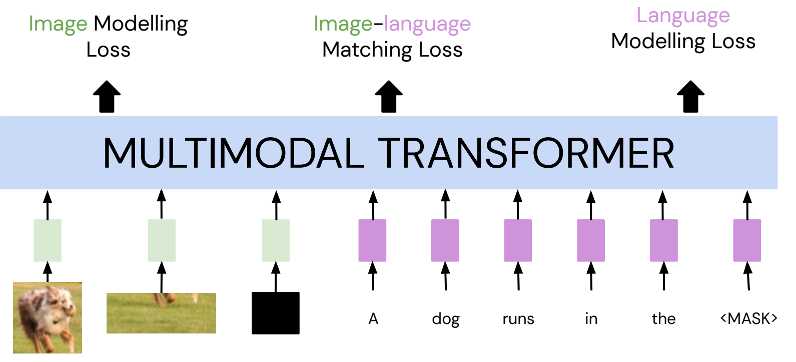
Are the "adopted" losses from language models good enough? No, we need better image modeling losses.

Is the cross-talk between modalities (via attention) important? Yes, multimodal attention is important.

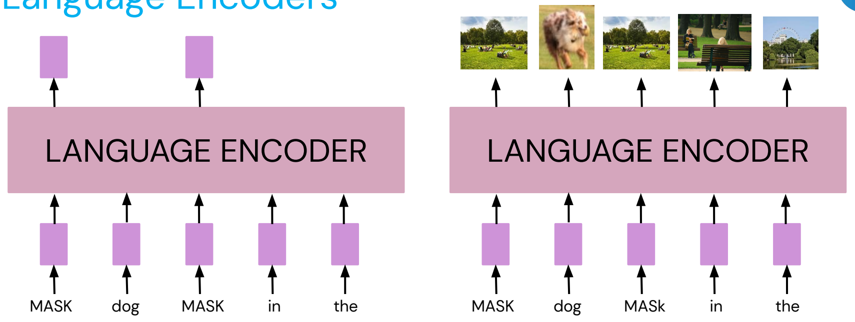
What makes a good pretraining dataset? The level of noise and language matter.

We have released our pretrained models! 🎉🎉

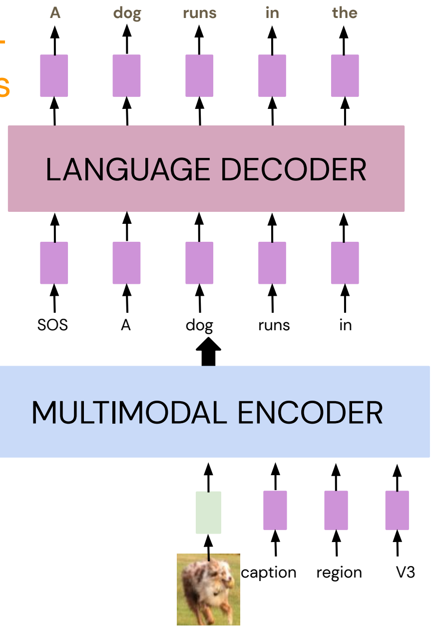
Joint Encoders



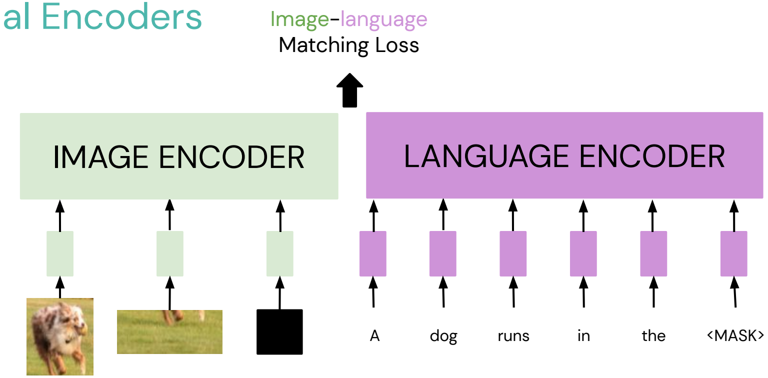
Language Encoders



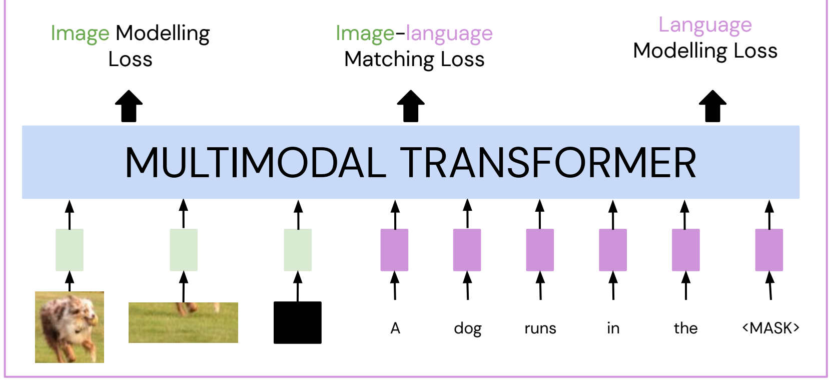
Encoder-Decoders



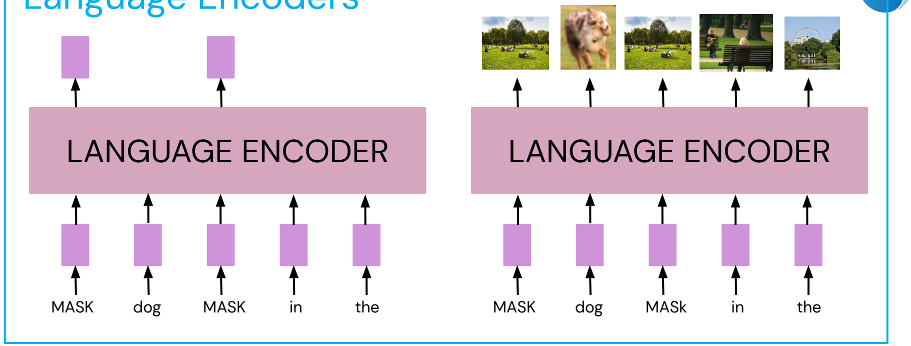
Dual Encoders



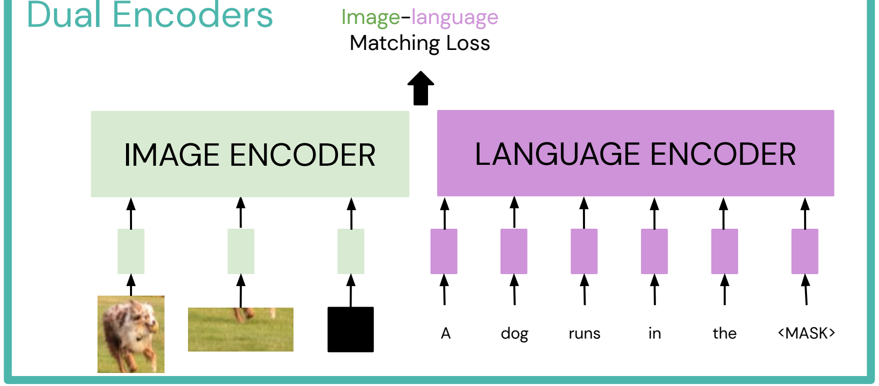
Joint Encoders



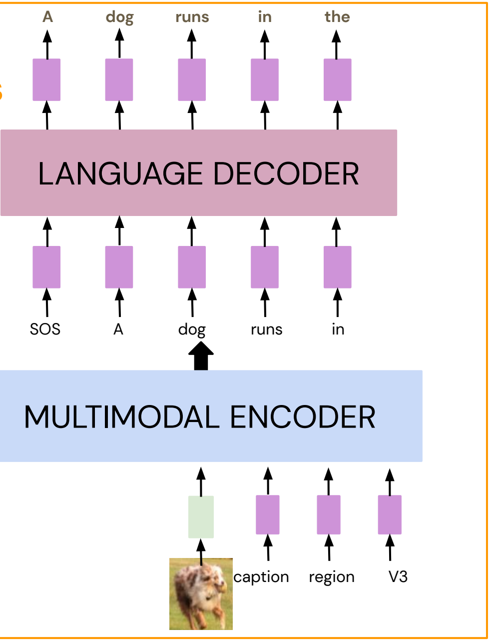
Language Encoders



Dual Encoders



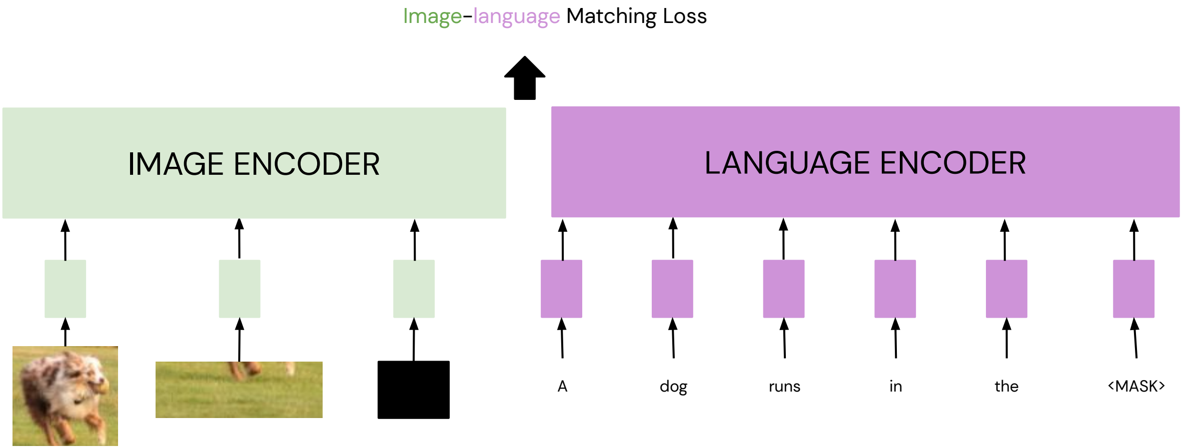
Encoder-Decoders



Dual Encoders

Two separate encoders for image and language modalities; no cross-talk between the two. [Weston et al., 2011; Frome et al., 2013; Kiros et al., 2014]

Very successful for retrieval tasks [Chowdhury et al., 2018; Miech, Alayrac, et al.2020]



Recent Large-Scale Dual Encoders [Radford et al, 2021; Jia et al, 2021]

CLIP [Radford et al, 2021] and ALIGN [Jia et al, 2021]: Larger models & datasets

How to collect large-scale datasets?

Pretraining Datasets: Refresher

MSCOCO

"The two people are walking down the beach."

MSCOCO/OI Narratives

"In this image we can see a bridge and sea. In the background, we can see trees and the sky. We can see so many people on the bridge. At the bottom of the image, we can see two people. We can see stairs in the right bottom of the image ..."

Visual Genome

small round yellow frisbee, man has cast on his arm, concrete trail path in the park, man wearing black sunglasses

manually annotated

Conceptual Captions

"The ***scenic route*** through mountain ranges includes these unbelievably coloured mountains.

SBU Captions

"***King Arthur's*** beheading rock - right on the sidewalk in the middle of ***town***".

from "the wild"

Noisier image-text correspondence but larger

Recent Large-Scale Dual Encoders [Radford et al, 2021; Jia et al, 2021]

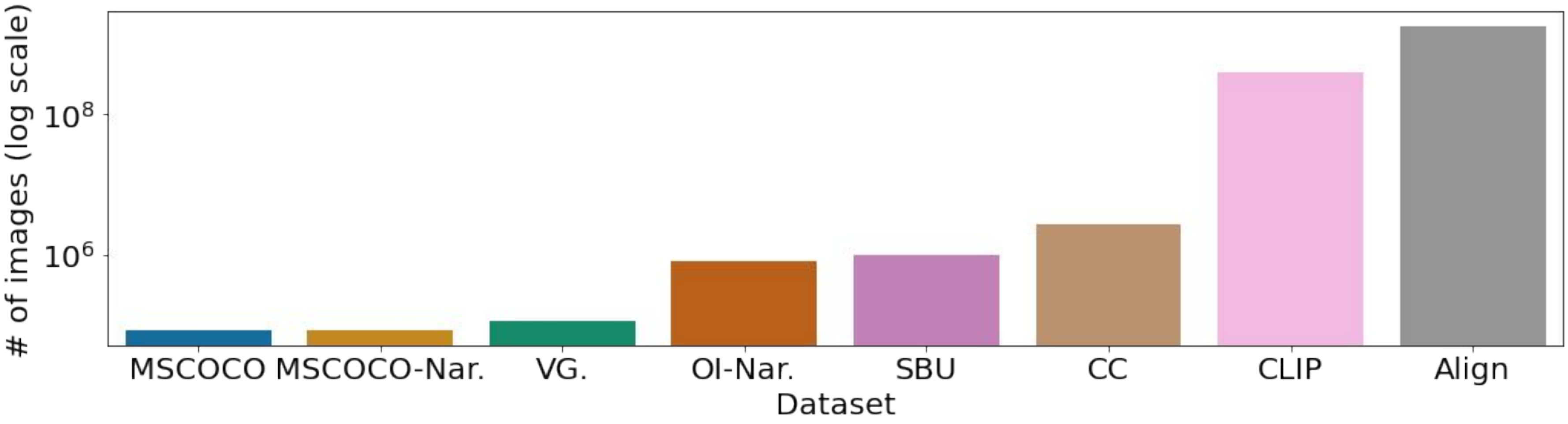
CLIP [Radford et al, 2021] and ALIGN [Jia et al, 2021]: Larger models & datasets

How to collect large-scale datasets?

- ALIGN removes any filtering to increase the size (1.8B) → noisier.
  - The same pipeline as Conceptual Captions (CC).
- CLIP uses heuristics to clean the data (400M).

Tradeoff between data size & noise: CC (3M) > ALIGN (3M/6M) on MSCOCO retrieval. [Jia et al, 2021]

Dataset Considerations: Size



Recent Large-Scale Dual Encoders [Radford et al, 2021; Jia et al, 2021]

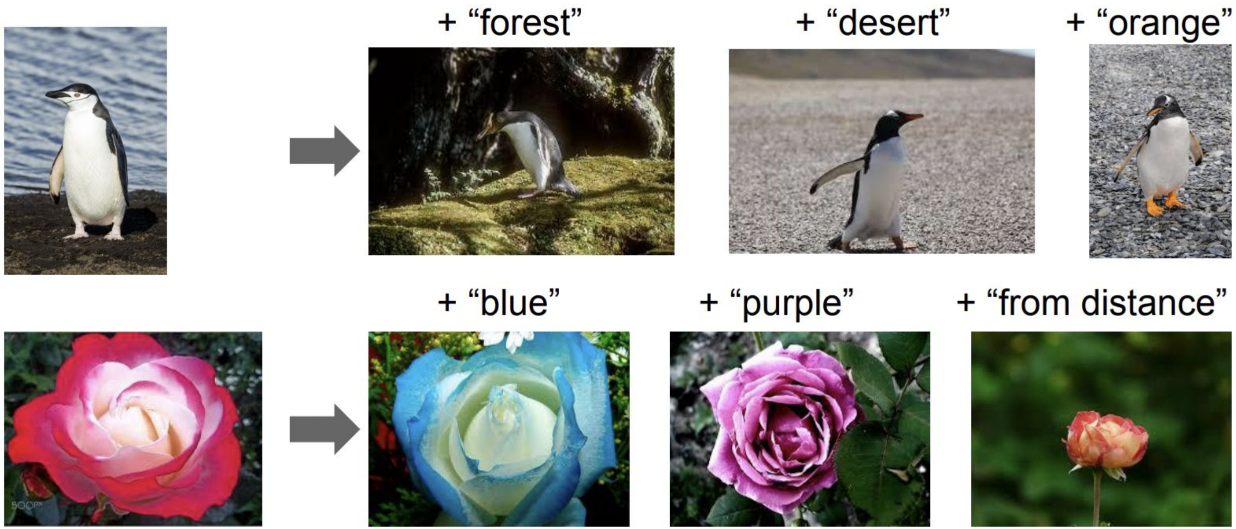
CLIP [Radford et al, 2021] and ALIGN [Jia et al, 2021]: Larger models & datasets

Use similar contrastive losses; ALIGN uses label smoothing that can be helpful with dataset noise.

Perform zero-shot image classification as a image-text retrieval task.

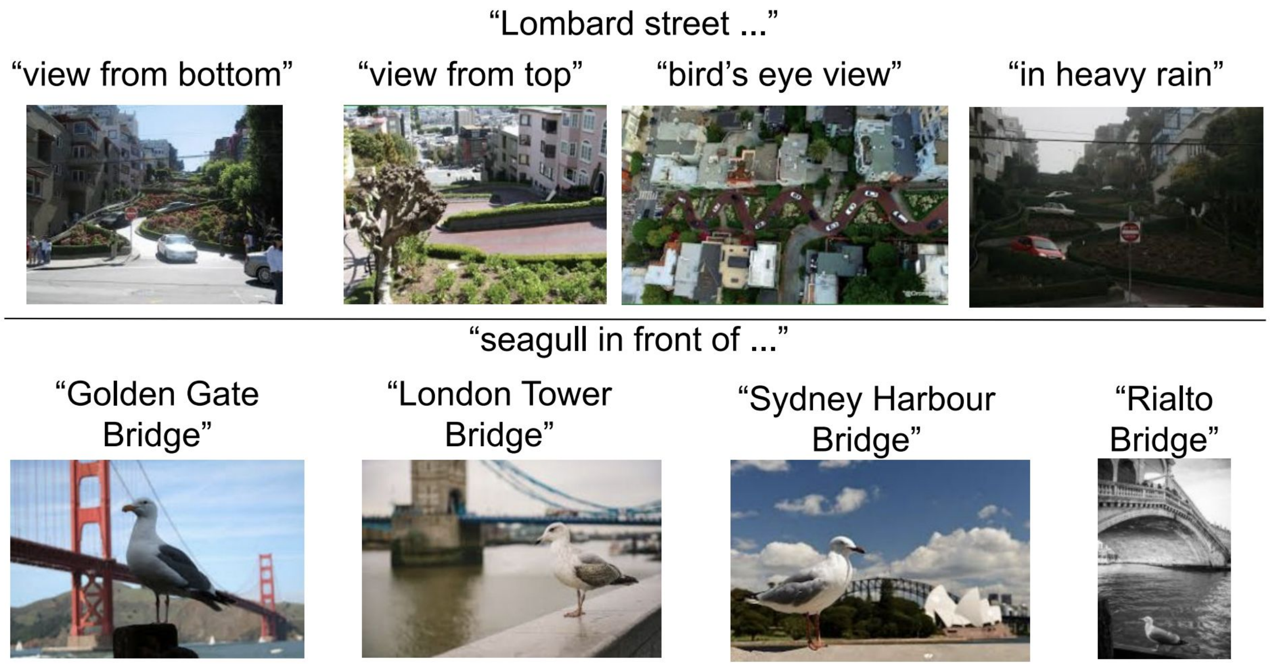
Qualitative Examples from ALIGN

Image retrieval with image +/- text queries



Qualitative Examples from ALIGN

Image retrieval with fine-grained queries.

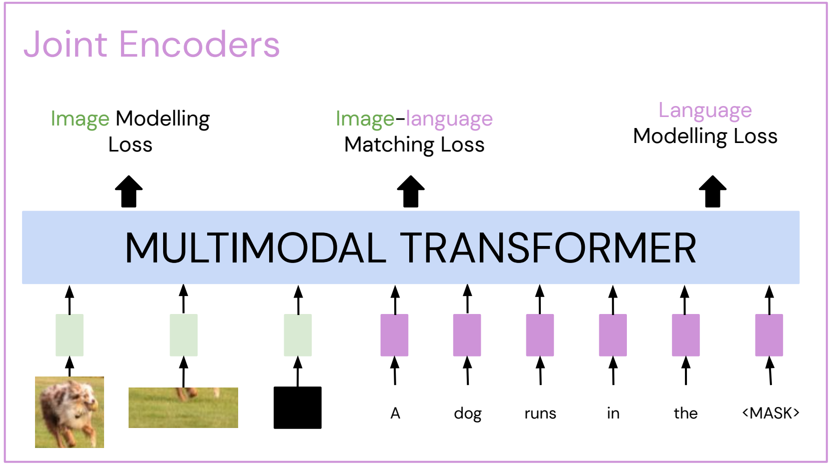


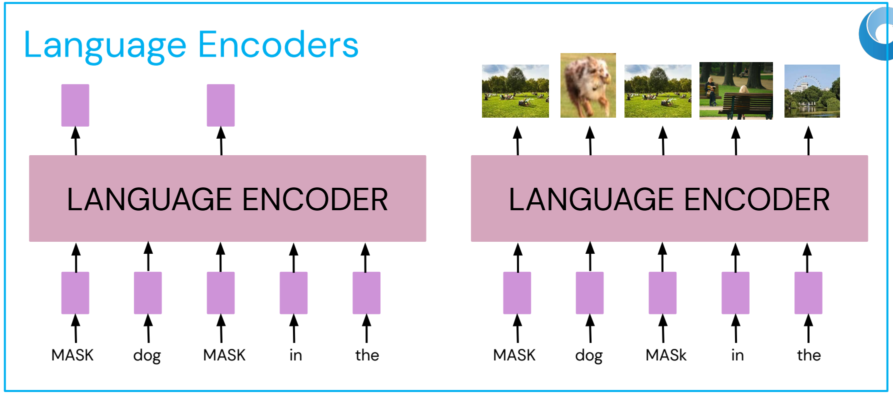
Dual Encoders Vs. Multimodal Transformers

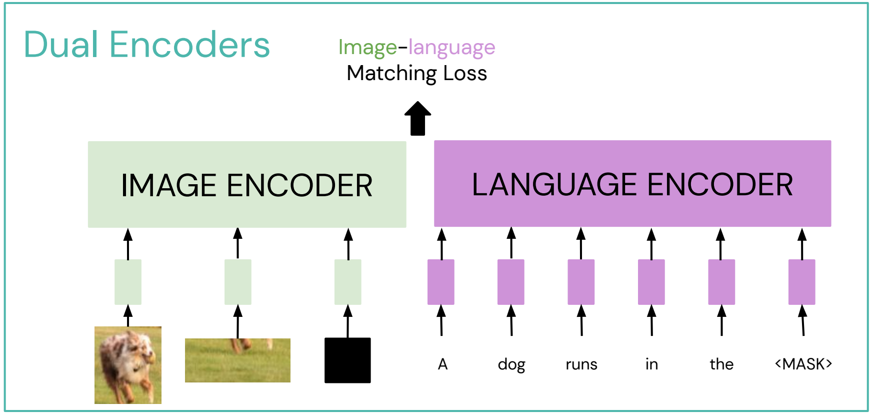
Dual encoders are easier to scale since they can reuse image/language features across pairs.

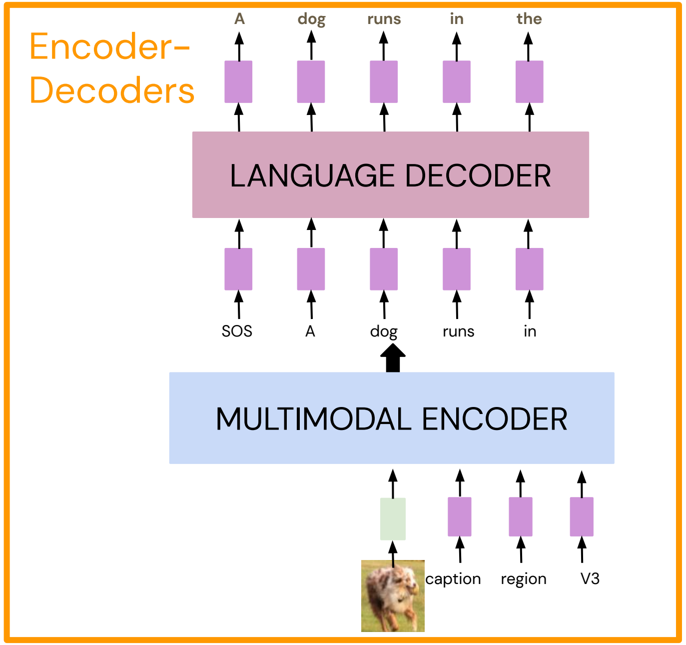
But they are not sample efficient.

BOW-DE: [Miech & Alayrac et al. CVPR 2021]
MMT: [Hendricks et al. TACL 2021]
UNITER: [Chen et al. ECCV 2020]
CLIP: [Radford et al. Arxiv 2021]
ALIGN: [Jia et al. Arxiv 2021]


Joint Encoders
Image Modelling Loss
Image-language Matching Loss
Language Modelling Loss
MULTIMODAL TRANSFORMER
A
dog
runs
in
the
<MASK>


Language Encoders
LANGUAGE ENCODER
MASK
dog
MASK
in
the
LANGUAGE ENCODER
MASK
dog
MASk
in
the


Dual Encoders
Image-language Matching Loss
IMAGE ENCODER
LANGUAGE ENCODER
A
dog
runs
in
the
<MASK>


Encoder-Decoders
A
dog
runs
in
the
LANGUAGE DECODER
SOS
A
dog
runs
in
MULTIMODAL ENCODER
caption
region
V3

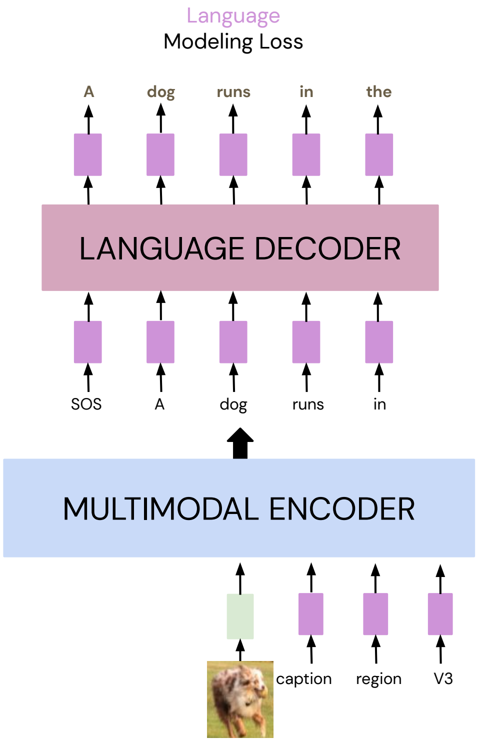
Encoder-Decoders

An image captioning setup:

- Replace the image encoder with a multimodal one
- Virtex, VL-BART(T5), SimVLM [Desai & Johnson, 2020; Cho *et al*, 2021; Wang *et al*, 2022]

Uses language as supervision for vision or multimodal pretraining.

Requires less images than the image-classification setting.

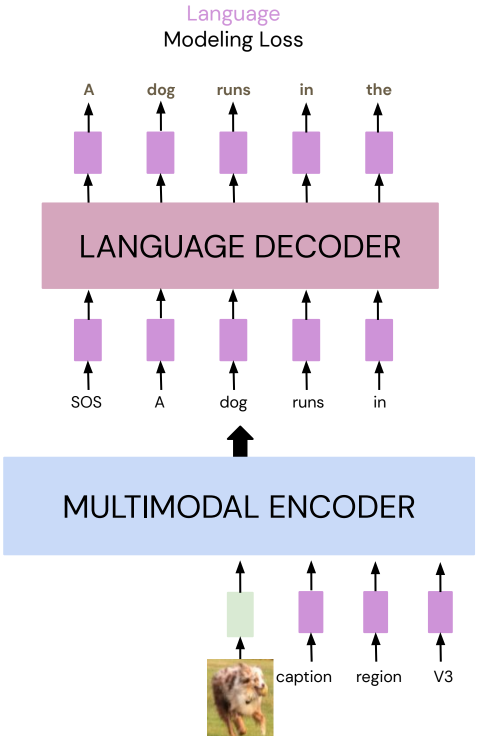


Encoder-Decoders

An image captioning setup:

- Replace the image encoder with a multimodal one
- Virtex, VL-BART(T5), SimVLM [Desai & Johnson, 2020; Cho *et al*, 2021; Wang *et al*, 2022]

Uses language as supervision for vision or multimodal pretraining.



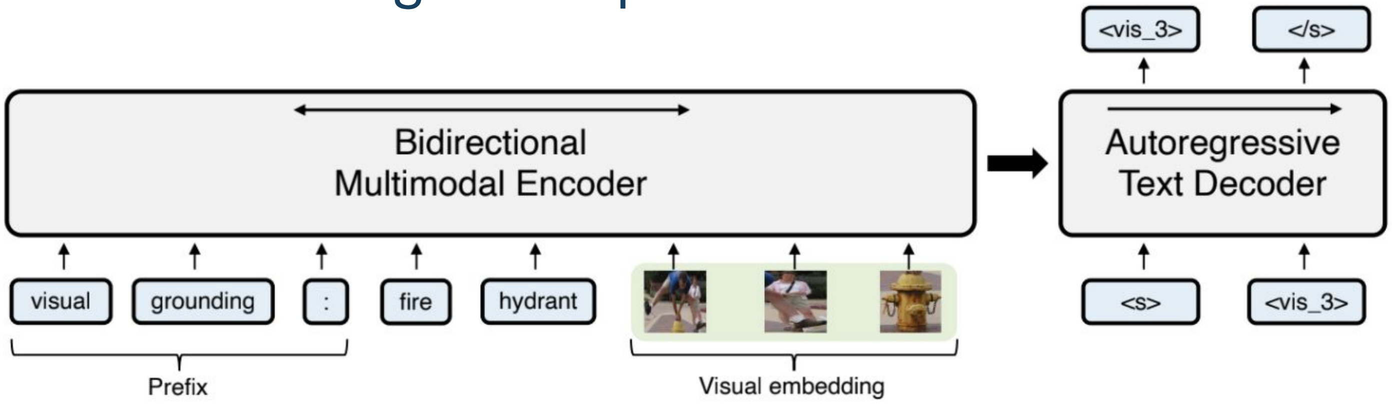
VL-BART (VL-T5) [Cho et al, 2021]

Unifies tasks as text generation (w/ task-specific prefixes).

- Parameters for each task are separately optimized.

Builds on a pretrained language model (BART or T5).

Trains on ~9.18M image–text pairs.

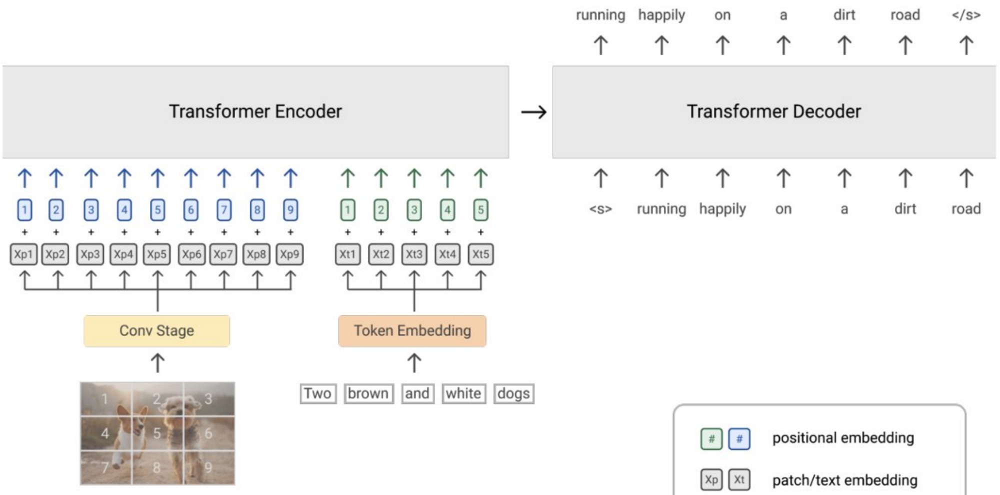


SimVLM [Wang et al, 2022]

Unifies tasks as text generation.

Removes object detection supervision.

Trains on large-scale noisy image-text data (ALIGN).



Comparing the Three Approaches

Given the same amount of pretraining data, and evaluated on zero-shot retrieval:

captioning model

Encoder Decoders ≅ Multimodal Transformers ≫ Dual Encoders

[rh/P1776]

[rh/P1713]

[rh/P1776]

Combining Frozen (Pretrained) Models

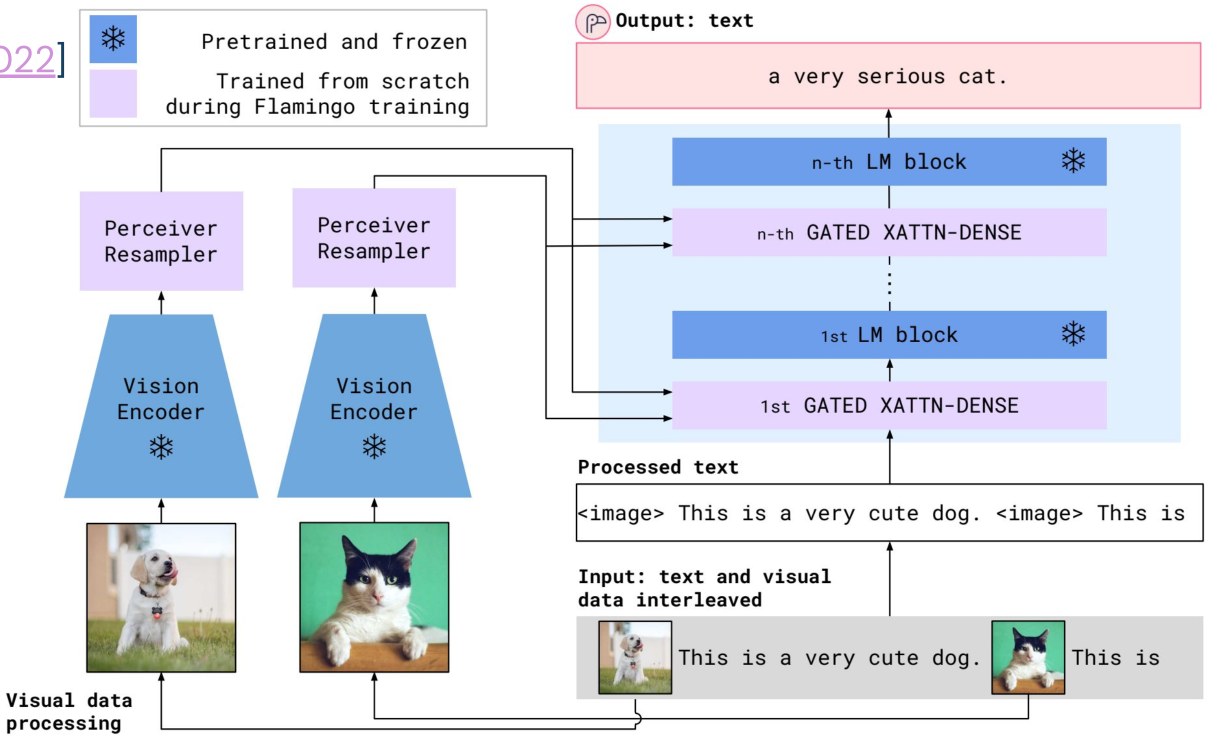
Given the cost of pretraining large models, can we reuse and combine existing vision and/or language models?

- Frozen [Tsimpoukelli et al, 2021]
- MAGMA [Eichenberg et al, 2021]
- Flamingo [Alayrac et al, 2022]

But, need to map different feature spaces (*e.g.*, image features to the language ones) → Train adapter layers

Combining Frozen (Pretrained) Models

Flamingo [Alayrac et al, 2022]



Moving Towards Generative Models

Pros.

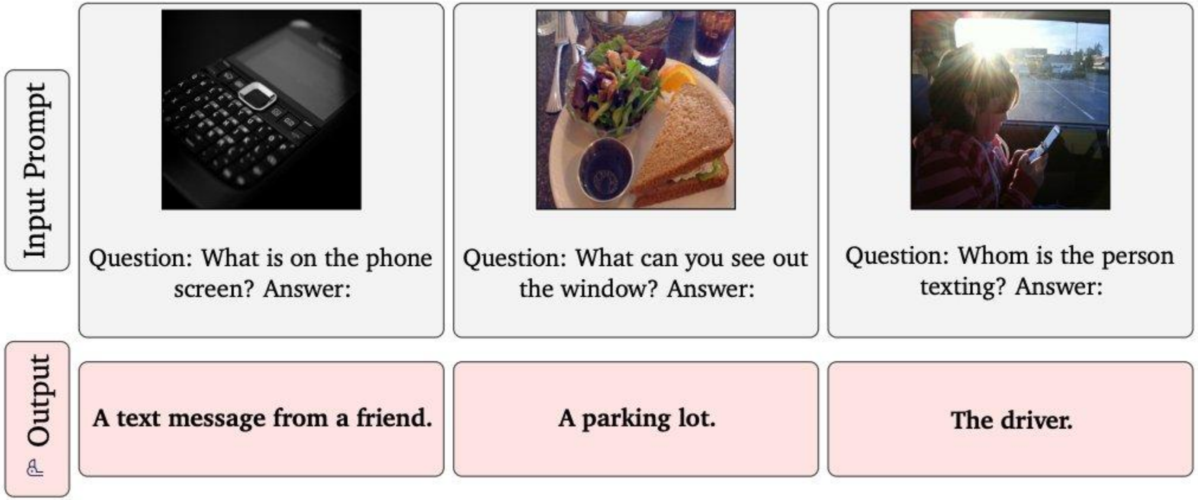
- Unified modeling of vision–language tasks
- Better out-of-distribution generalization in VQA [Agrawal et al, 2022]

Cons.

- How to accurately evaluate the generated text? [Agrawal et al, 2022]
- Inheriting existing biases of pretrained language (and vision) models.

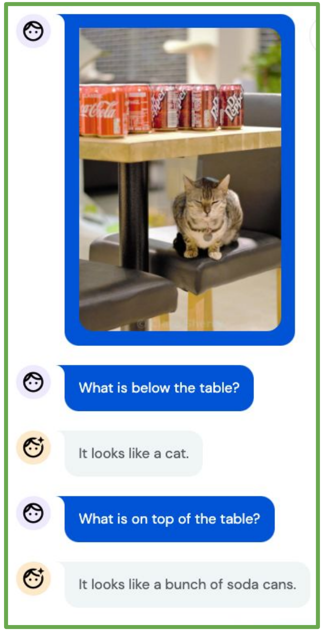
Qualitative Examples from Flamingo

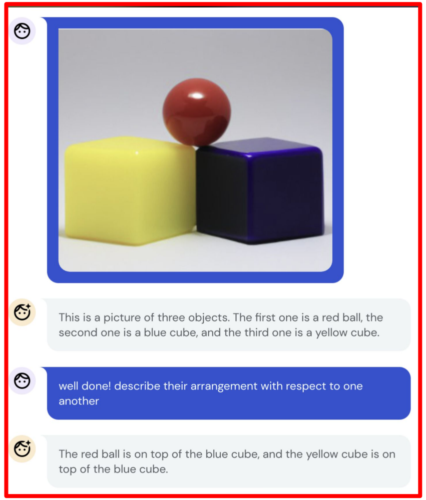
Language model bias can result in hallucinations.

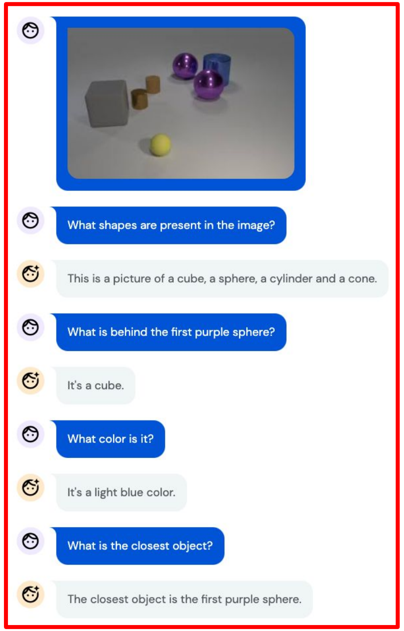


Qualitative Examples from Flamingo

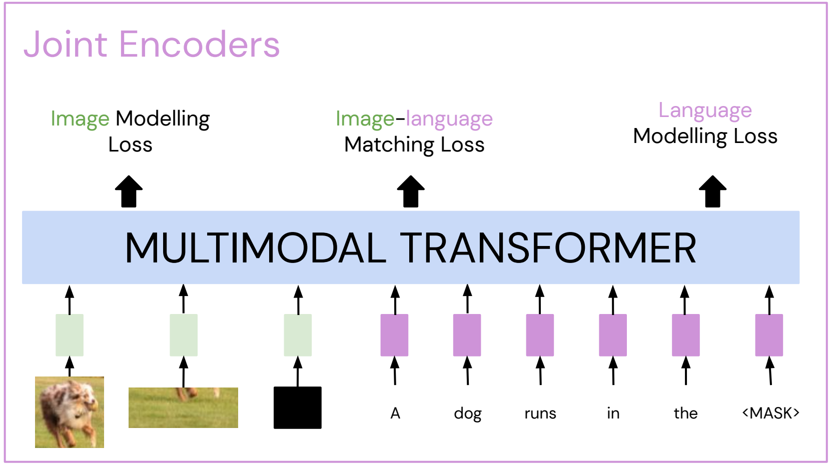
Mistakes in spatial understanding but correct language use.

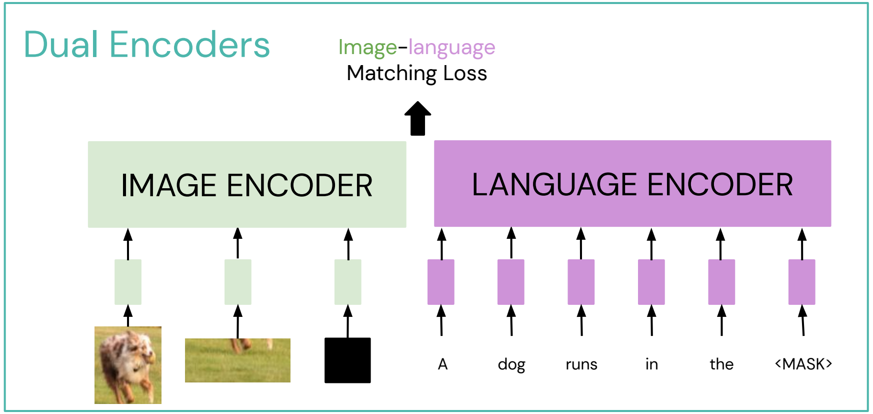


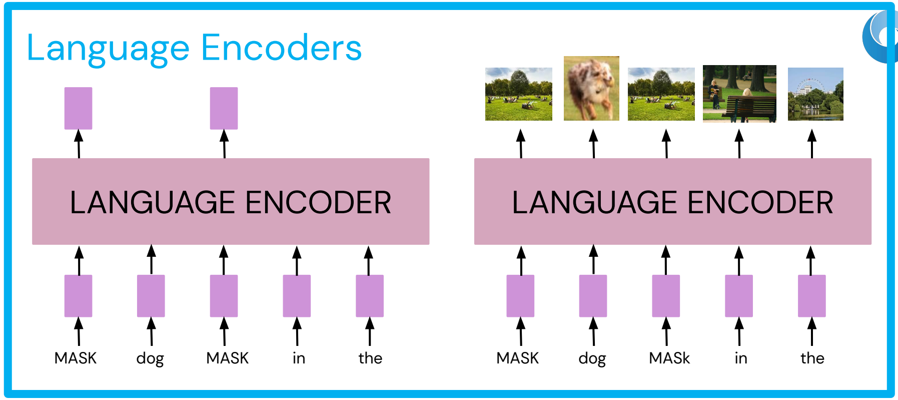


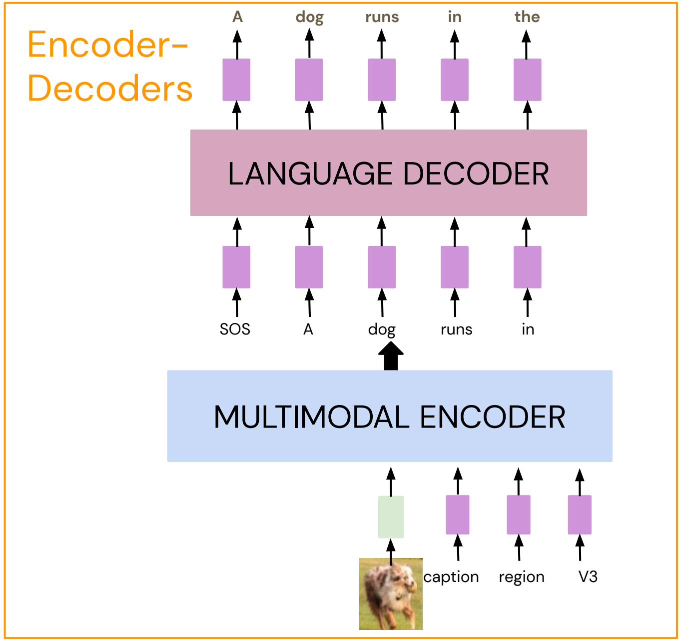


[Examples from JB Alayrac]


Joint Encoders
Image Modelling Loss
Image-language Matching Loss
Language Modelling Loss
MULTIMODAL TRANSFORMER
A
dog
runs
in
the
<MASK>


Dual Encoders
Image-language Matching Loss
IMAGE ENCODER
LANGUAGE ENCODER
A
dog
runs
in
the
<MASK>

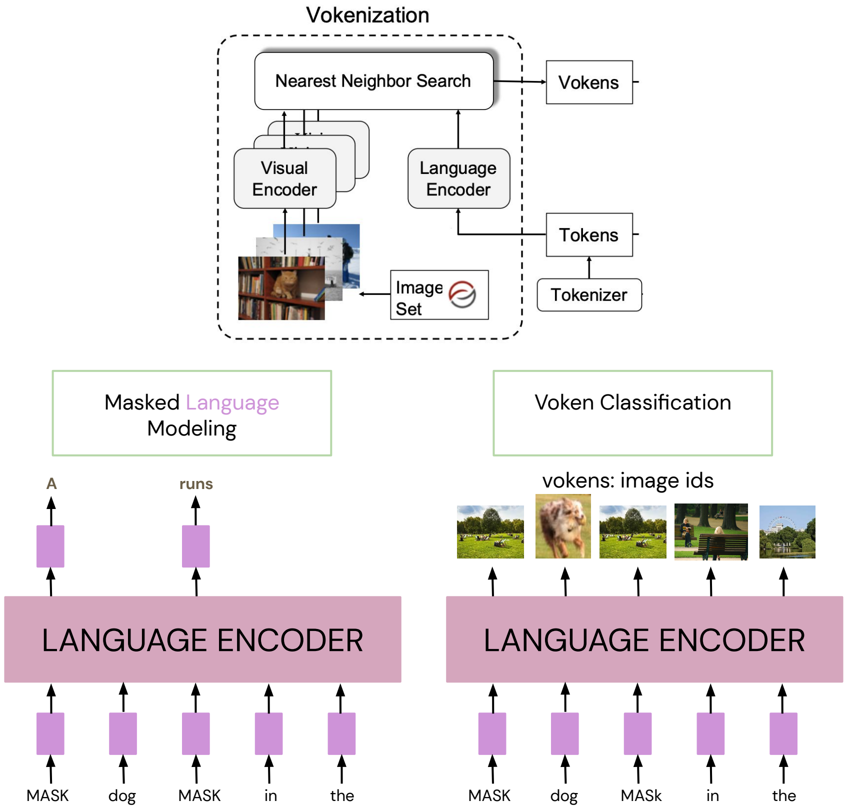

Language Encoders
LANGUAGE ENCODER
MASK
dog
MASK
in
the
LANGUAGE ENCODER
MASK
dog
MASk
in
the


Encoder-Decoders
A
dog
runs
in
the
LANGUAGE DECODER
SOS
A
dog
runs
in
MULTIMODAL ENCODER
caption
region
V3

Language Encoders

A language modeling setup:

- Vokenization: map each language token to a visual token (voken) [Tan & Bansal, 2020]



Uses vision as supervision for language pretraining.

Summary of Different Approaches

How to evaluate pretrained models?

- Use task-specific heads for each downstream task (e.g., ViLBERT, LXMERT, UNITER, OSCAR, VinVL).
- Treat all downstream tasks as language generation with no task-specific head (e.g., VL-T5, VL-BART, SimVLM).

Summary of Different Approaches

How are the features used (other than vision-language tasks)?

- In vision tasks (e.g., VirTex, CLIP, ALIGN)
- In language tasks, including multilingual data (e.g., Vokenization, M3P, VL-T5, SimVLM)

Towards a Better Evaluation of Pretrained Models

Performance after fine-tuning depends on the the size of fine-tuning data and other experimental set-up [Yogatama et al., 2019].

Recent work has shifted focus to few- and zero-shot evaluation.

Other approaches

- Evaluate for out-of-distribution generalization (transfer)
- Probe for certain capabilities (e.g., verb understanding)

See my talk on evaluation if you are interested!

Answering Questions from Blind People

Q: What are the people waiting for?
A: bus

Q: What is this?
A: 10 euros.

VizWiz is a benchmark curated from visually-impaired users.

Evaluate in a Transfer Setting [Agrawal et al, 2022]

Fine-tune a multimodal transformer on one dataset (VQAv2), test on another one (VizWiz): we observe ~26 drop in accuracy.

Q: What are the people waiting for?
A: bus

Q: What is this?
A: 10 euros.

Winoground: Visio-Linguistic Compositionality

[Thrush et al, 2022]

(a) there is [a mug] in [some grass]

(c) a person [sits] and a dog [stands]

(e) it's a [truck] [fire]

(b) there is [some grass] in [a mug]

(d) a person [stands] and a dog [sits]

(f) it's a [fire] [truck]

Object *Relation* *Both*

What Image Retrieval Tests

Order images with respect to their match to a sentence.

A person is **riding** a horse.

Subject Verb Object

Does not require fine-grained multimodal understanding.

What SVO-Probes Tests [Hendricks *et al.*, Findings of ACL 2021]

A person is **riding** a horse

❌

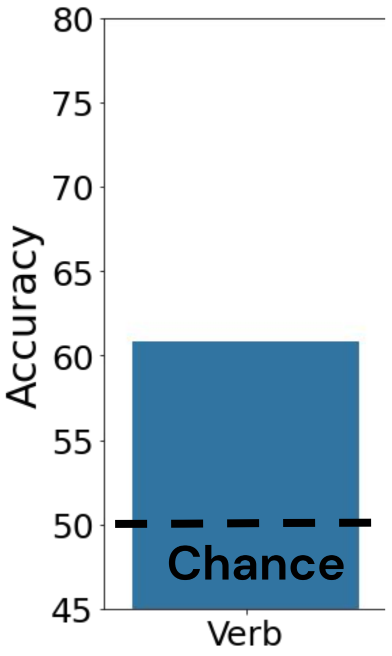
✔

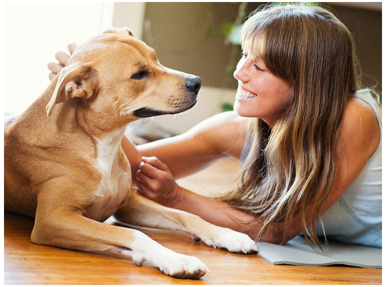
Correctly classify both the **positive** & **negative** examples.

We have released our dataset! 🎉🎉

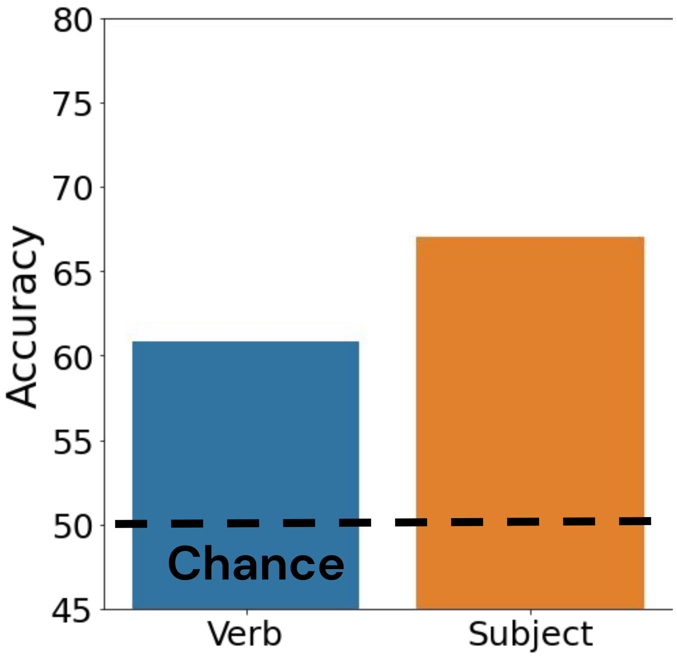
Do MMTs Have Fine-grained Verb Understanding?

A woman **lying** with a dog



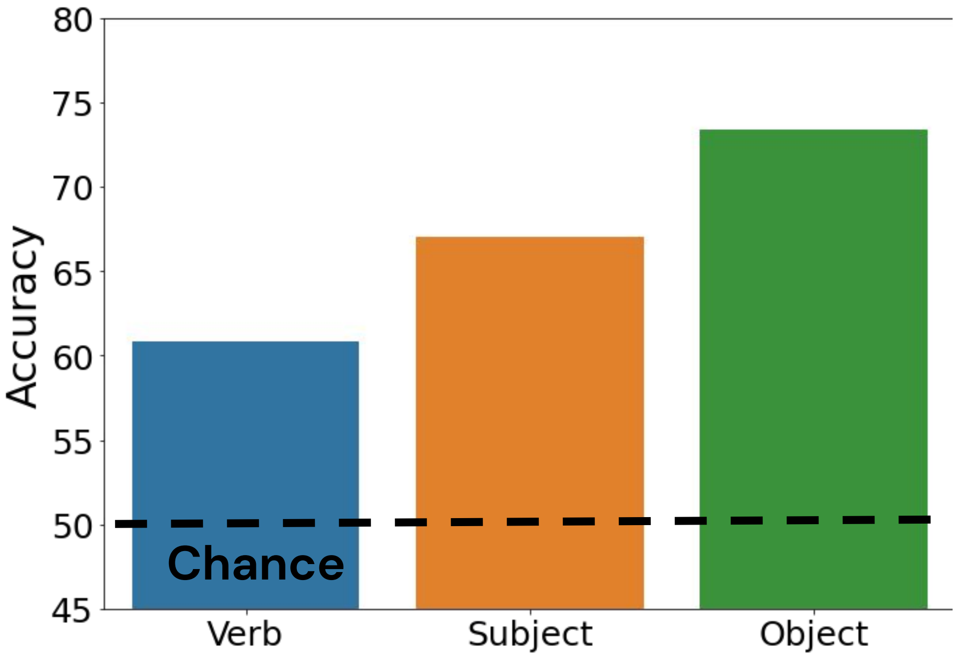


Do MMTs Have Fine-grained Verb Understanding?

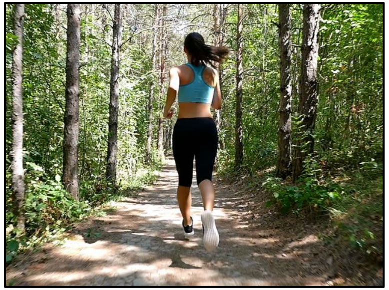


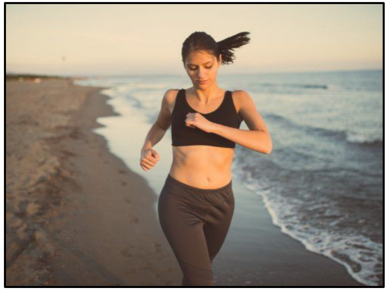
A **animal** lays in the grass

Do MMTs Have Fine-grained Verb Understanding?

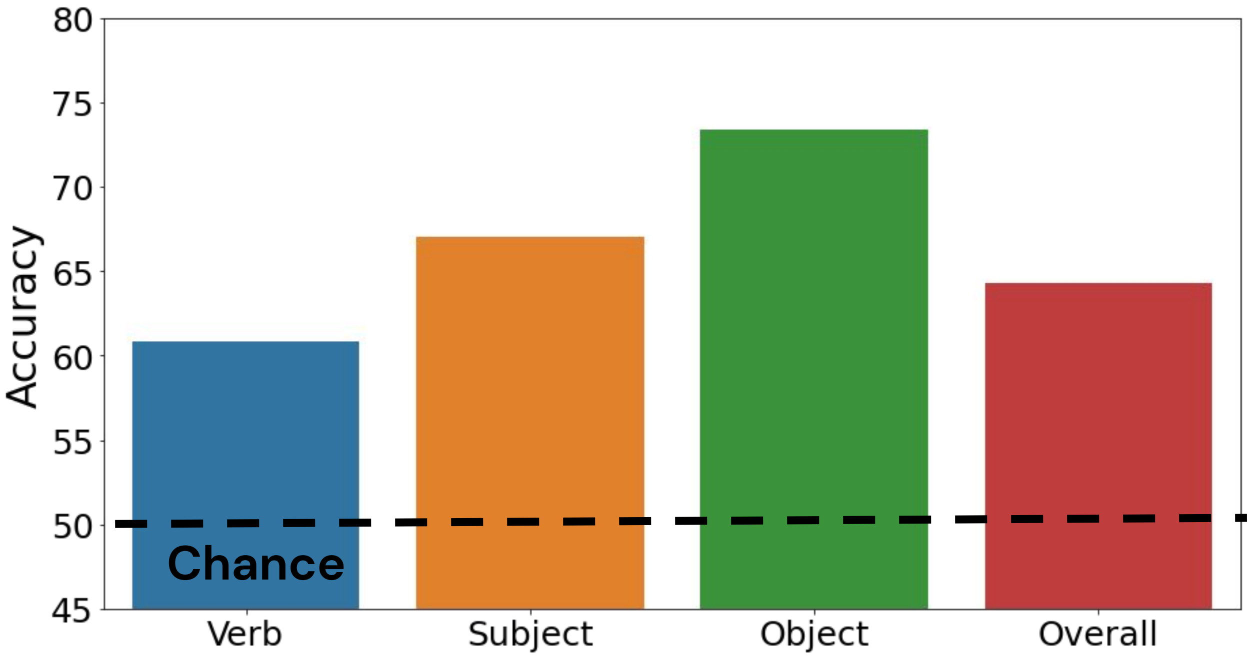


A woman jogs on the **beach**





Do MMTs Have Fine-grained Verb Understanding?



Overall MMT performance 64.3 -- lots of room for improvement!

To build stronger models, we need to better evaluate them first.

Thanks for listening!

Questions?