

Computational Natural Language Processing

Multimodal Models

Hamidreza Mahyar
mahyarh@mcmaster.ca

What is multimodality?

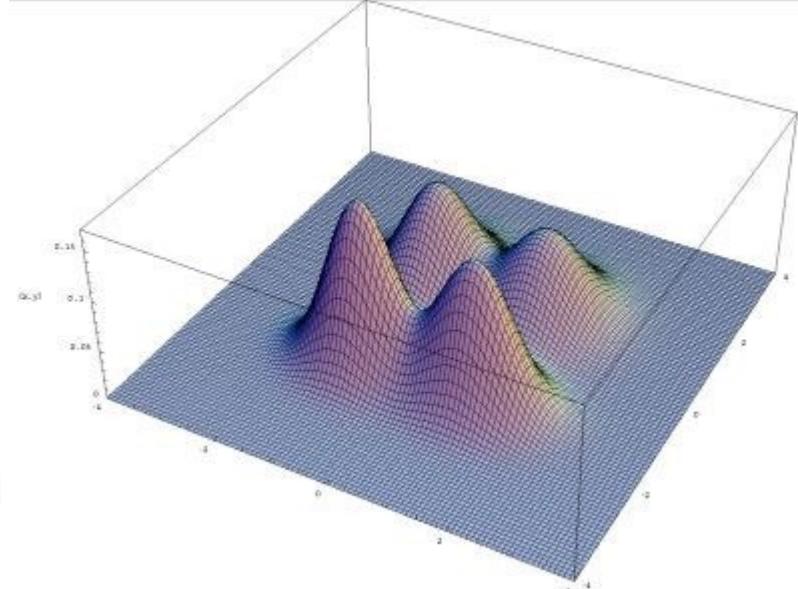
multimodal adjective

mul·ti·mod·al (məl-tē-'mō-dəl) -tī-

: having or involving several modes, modalities, or maxima

| *multimodal* distributions

| *multimodal* therapy



In our case, focusing on NLP: text + one or more other *modality* (images, speech, audio, olfaction, others). We'll mostly focus on images as the other modality.

Why does multimodality matter?

A range of very good reasons:

- Faithfulness: Human experience is multimodal
- Practical: The internet & many applications are multimodal
- Data efficiency and availability:
 - Efficiency: Multimodal data is rich and “high bandwidth” (compared to language; quoting LeCun, “an imperfect, incomplete, and low-bandwidth serialization protocol for the internal data structures we call thoughts”), so better for learning?
 - Scaling: More data is better, and we’re running out of high quality text data.

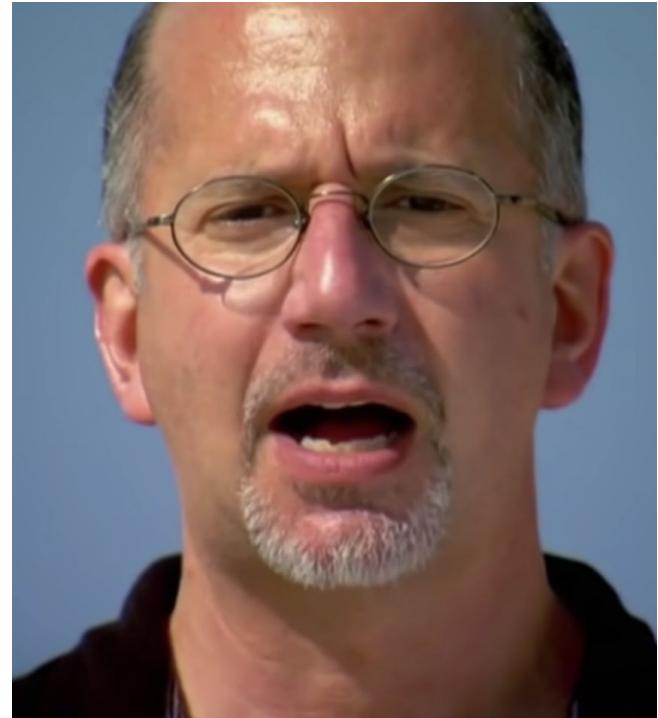


Multimodality is one of the main frontiers of the new foundation model revolution.

Multimodal brains

McGurk effect (McGurk & MacDonald, 1976)

<https://www.youtube.com/watch?v=2k8fHR9jKVM>



Multimodal applications

Let's say we're dealing with two modalities – text, and images:

- Retrieval (image <> text)
- Captioning (image -> text)
- Generation (text -> image)
- Visual question answering (image+text -> text)
- Multimodal classification (image+text -> label)
- Better understanding/generation (image+text -> label/text)

Multimodal is hot right now

.. and/but has been “the next big thing” for almost a decade!

Language Is Not All You Need: Aligning Perception with Language Models

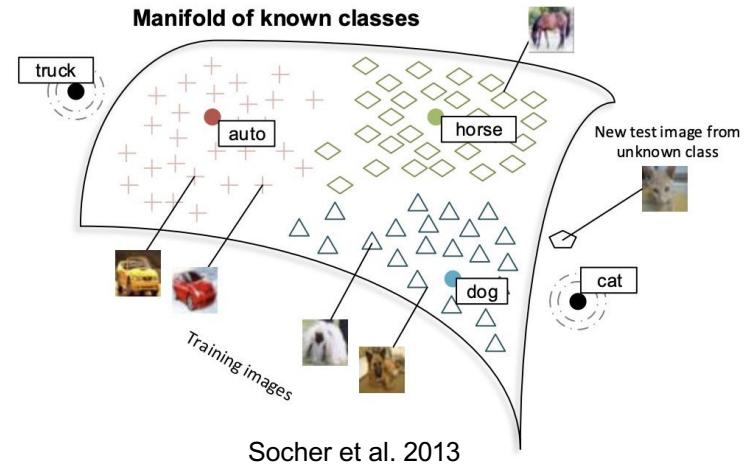
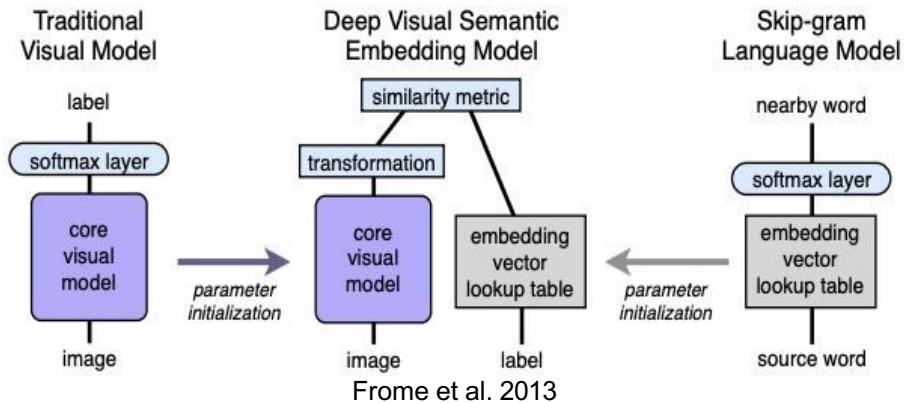
Shaohan Huang*, Li Dong*, Wenhui Wang*, Yaru Hao*, Saksham Singhal*, Shuming Ma*
Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi
Johan Bjorck, Vishrav Chaudhary, Subhajit Som, Xia Song, Furu Wei†
Microsoft

Outline

1. Early models
2. Features and fusion
3. Contrastive models
4. Multimodal foundation models
5. Evaluation
6. Beyond images: Other modalities
7. Where to next?

Cross-modal “Visual-Semantic Embeddings”

WSABI (Weston et al 2010), DeVise (Frome et al 2013),
Cross-Modal Transfer (Socher et al 2013)



$$\text{loss}(\text{image}, \text{label}) = \sum_{j \neq \text{label}} \max[0, \text{margin} - \vec{t}_{\text{label}}^T M \vec{v}(\text{image}) + \vec{t}_j^T M \vec{v}(\text{image})]$$

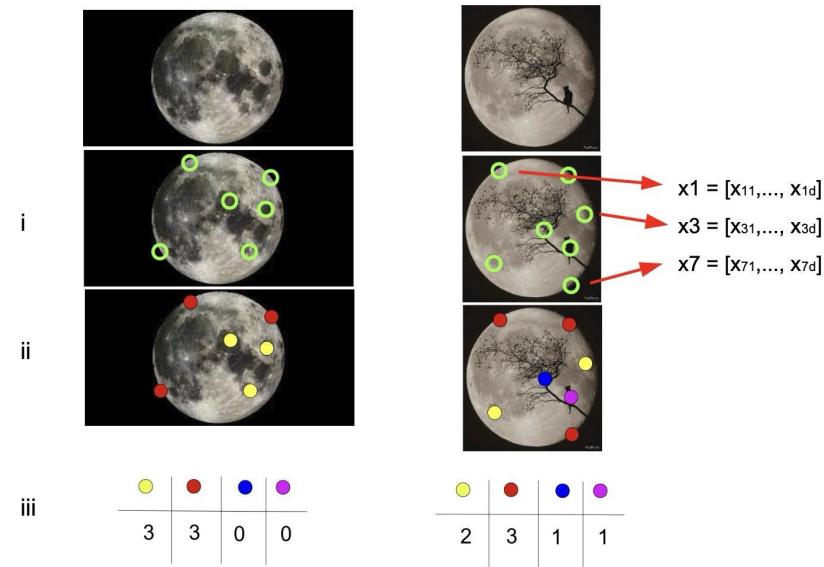
Socher et al. 2013

Multimodal distributional semantics (Bruni et al., 2014)

Algorithm:

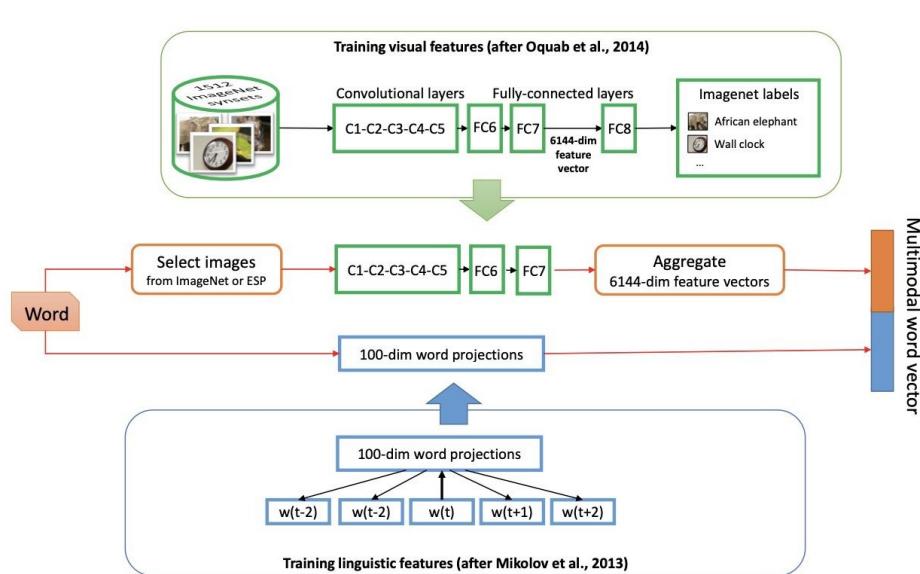
- Obtain visual “word vector” via BOVW:
 - Identify keypoints and get their descriptors
 - Cluster these and map to counts
- Concatenate with textual word vector
- Apply SVD to “fuse” information

This approach was shown to lead to better word representations on human similarity judgment datasets.

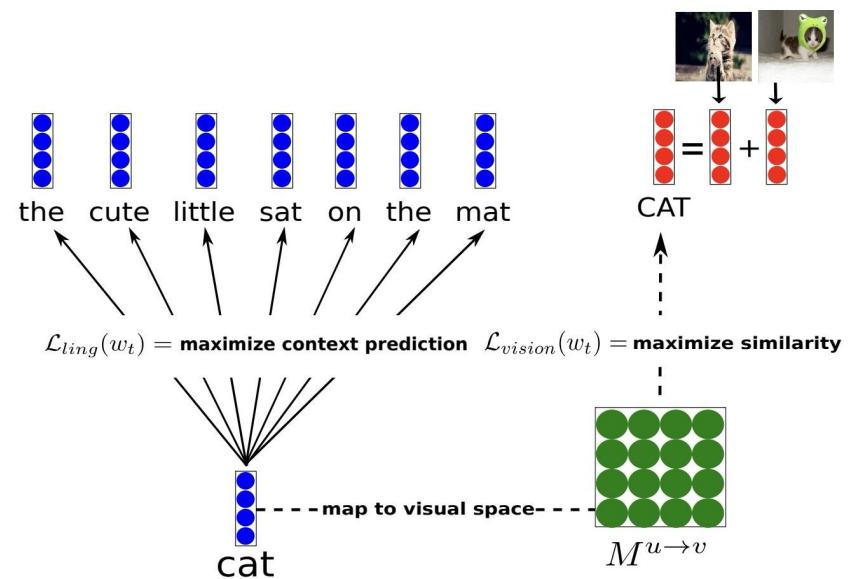


Bag of visual words (BOVW) illustration. Bruni et al., 2014.

Neural version (KB, 2014; Lazaridou et al., 2015)



Kiela & Bottou, 2014



Lazaridou et al., 2015

Beyond words: Sentence level alignment

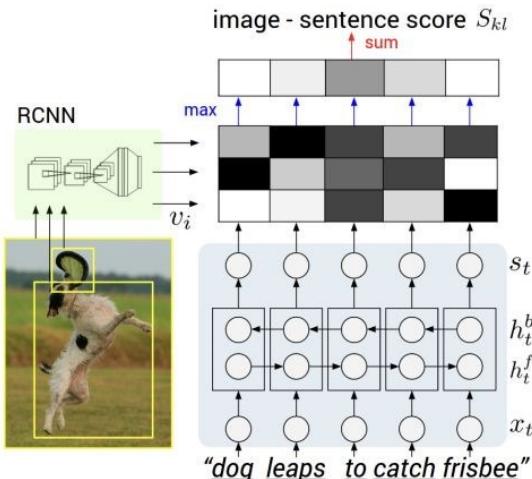
Grounded Compositional Semantics (Socher et al., 2013)

Visual-Semantic Embeddings (Kiros et al., 2014; Faghri et al., 2015)

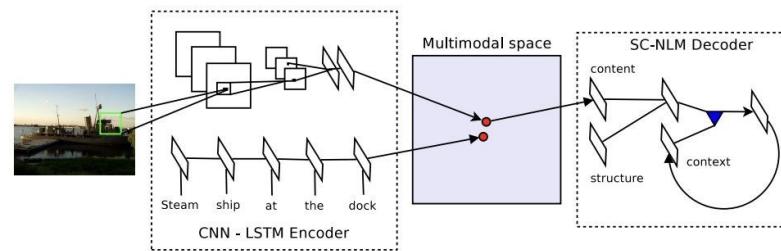
Visual-Semantic Alignments (Karpathy & Li, 2015)

Grounded Sentence Representations (Kiela et al., 2016)

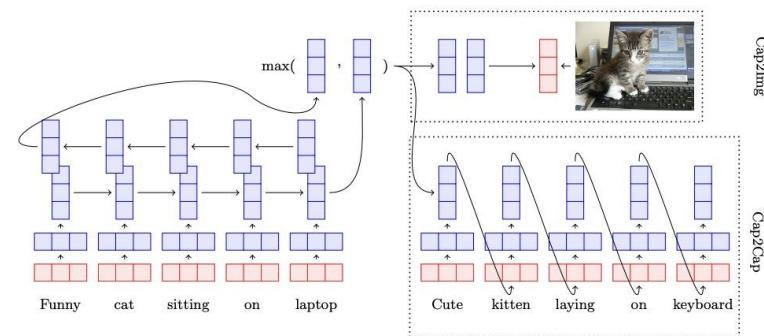
Hinge/margin-like loss as in WSABI/DeViSE.



Karpathy & Li, 2015



Kiros et al., 2014



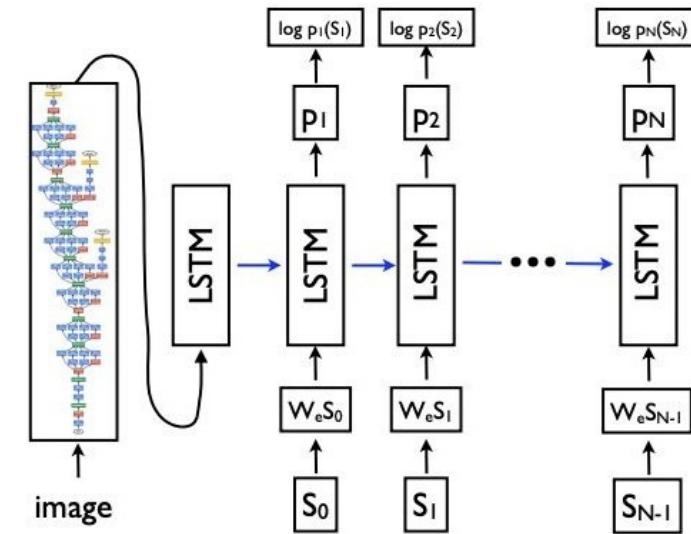
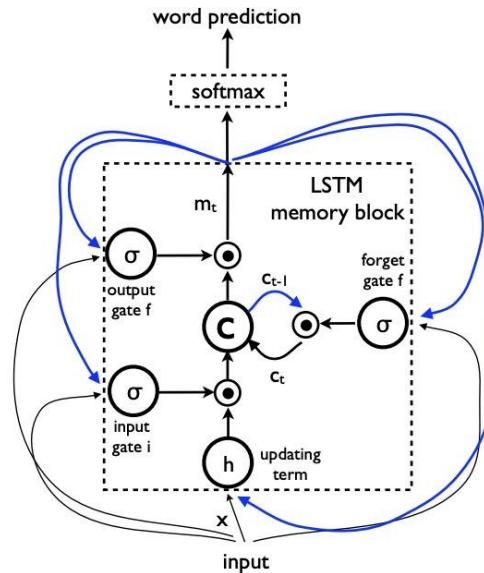
Cap2Img

Cap2Cap

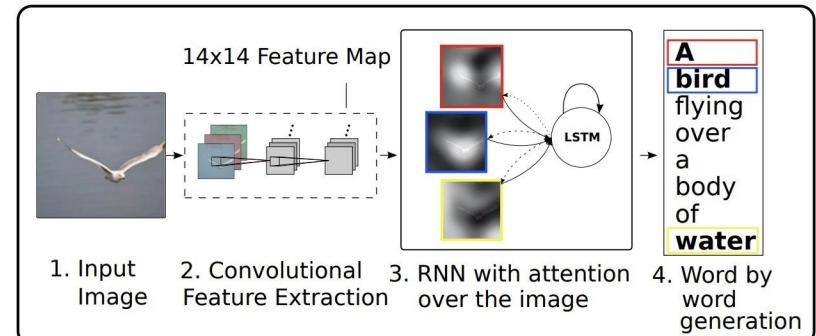
Image to text: Captioning

Show and tell (Vinyals et al., 2015)

Show, attend and tell (Xu et al., 2016)



Vinyals et al., 2015



Xu et al., 2016

Attention as visual-semantic alignment



A woman is throwing a frisbee in a park.



A dog is standing on a hardwood floor.



A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.



A group of people sitting on a boat in the water.



A giraffe standing in a forest with trees in the background.

Text to image: Conditional image synthesis

Generative adversarial nets (Goodfellow et al. 2014)



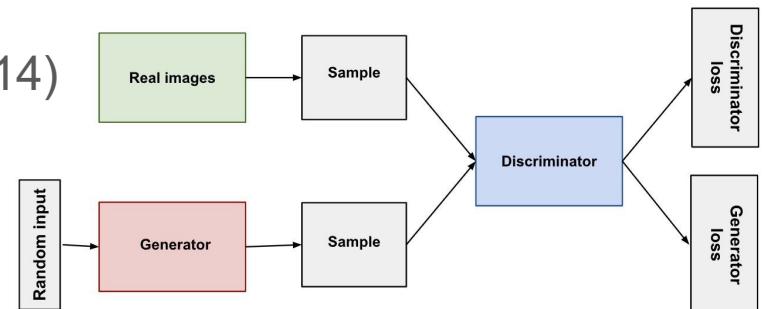
As training progresses, the generator gets closer to producing output that can fool the discriminator:



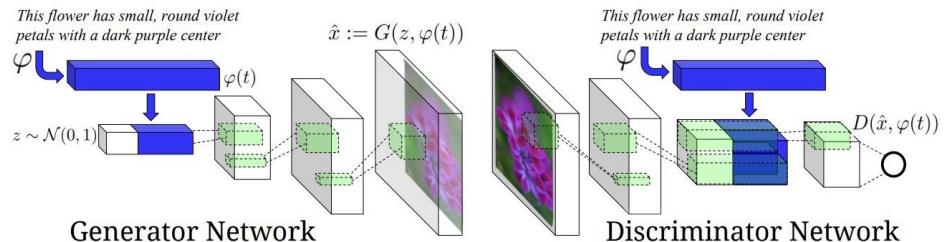
Finally, if generator training goes well, the discriminator gets worse at telling the difference between real and fake. It starts to classify fake data as real, and its accuracy decreases.



Source: [Google](#)



Source: [Google](#)



Reed et al., 2016

Outline

1. Early models
2. **Features and fusion**
3. Contrastive models
4. Multimodal foundation models
5. Evaluation
6. Beyond images: Other modalities
7. Where to next?

Problems with multimodality

If it's so important, why isn't every system multimodal from first principles?

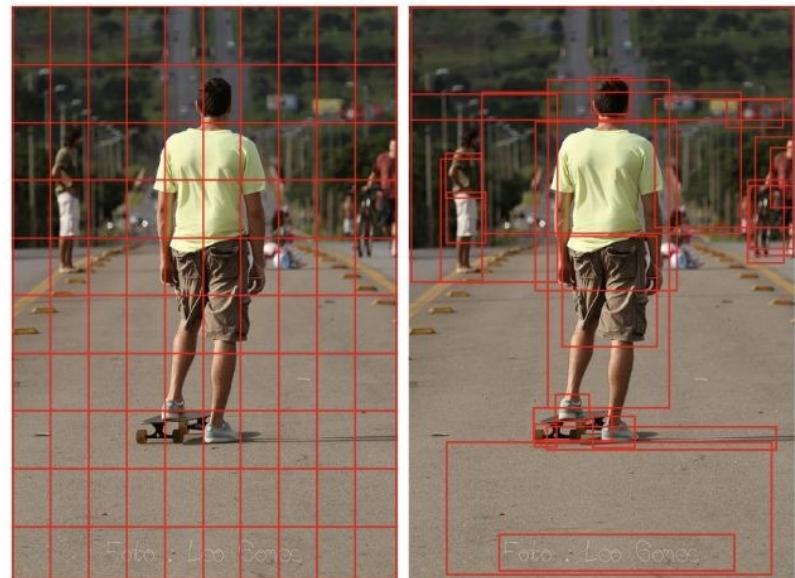
- One modality can dominate other modalities.
- Additional modalities can add noise.
- Full coverage over modalities is not guaranteed.
- We are (were) not ready.
- It's complicated.

Features

Featurizing text: Batch_size x Sequence_length x Hidden_size.

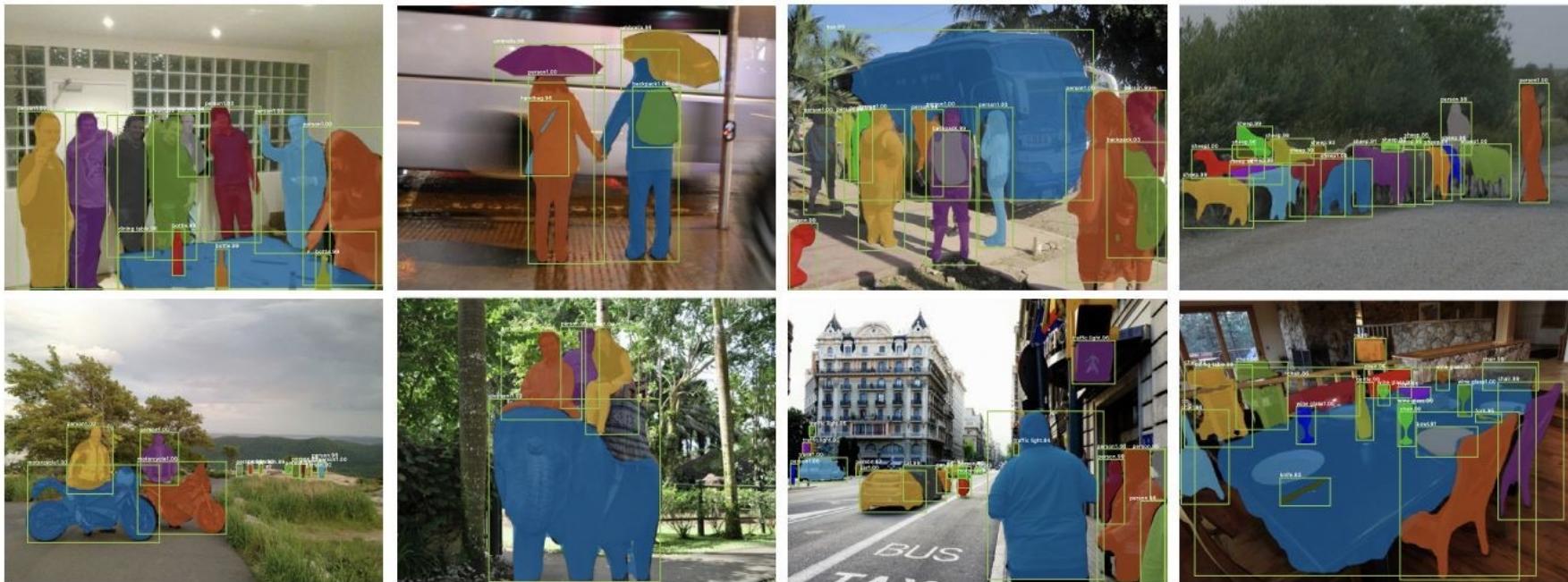
Featurizing images:

- Sparse “region” features:
 - Object detectors
- Dense features:
 - ConvNet layer(s) or feature maps
 - Vision Transformer layers

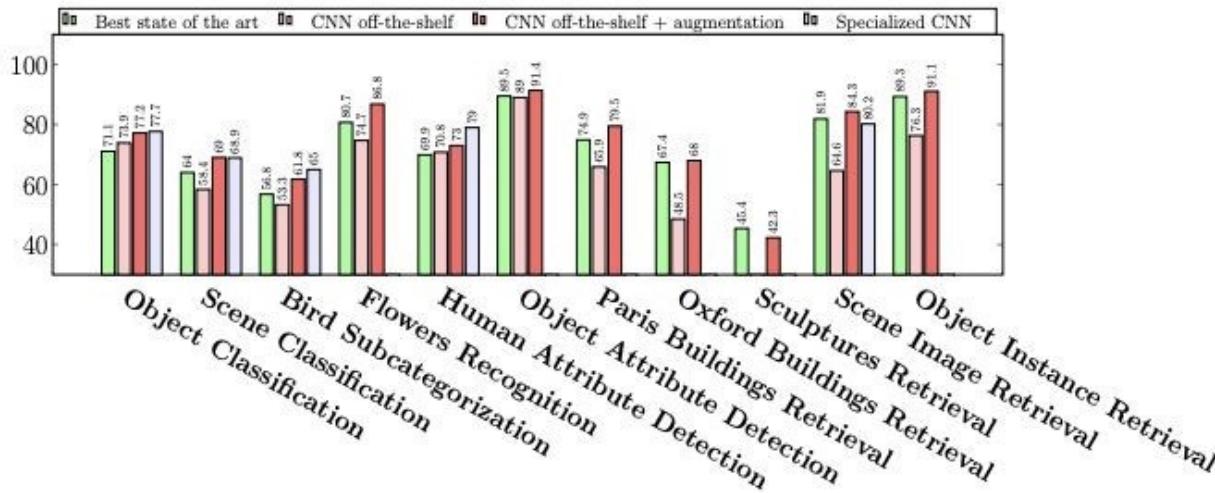
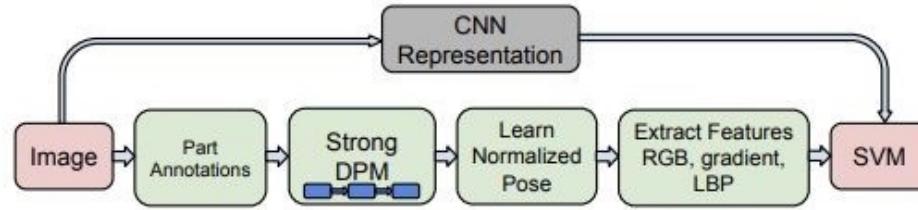


Region features

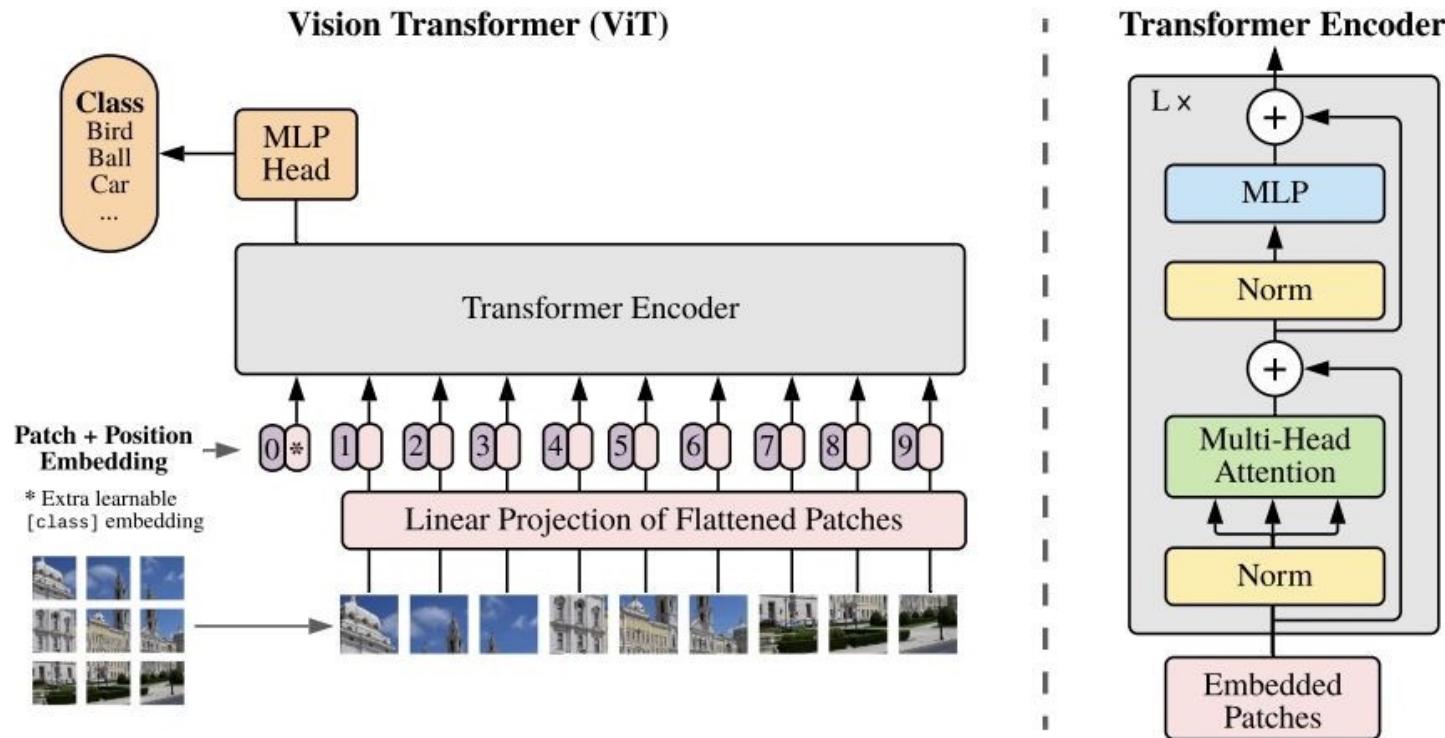
- R-CNN (Girshick et al., 2014); Fast R-CNN (Girshick, 2015); Faster R-CNN (Ren et al., 2015); YOLO (you only look once) vX..



“Off the shelf” ConvNet features (Razavian et al., 2014)



Vision Transformers (Dosovitskiy et al., 2020)



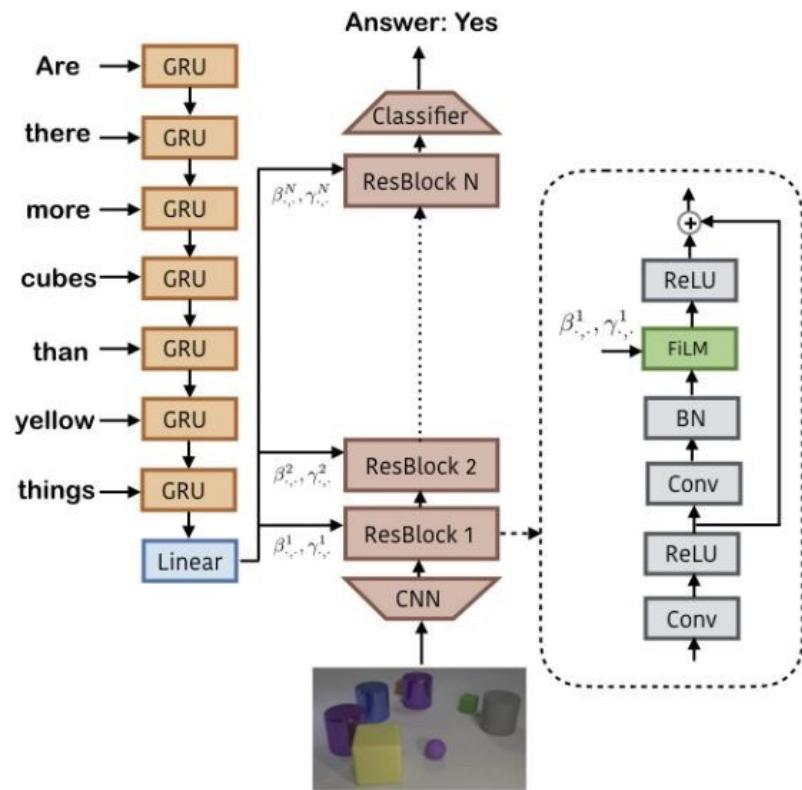
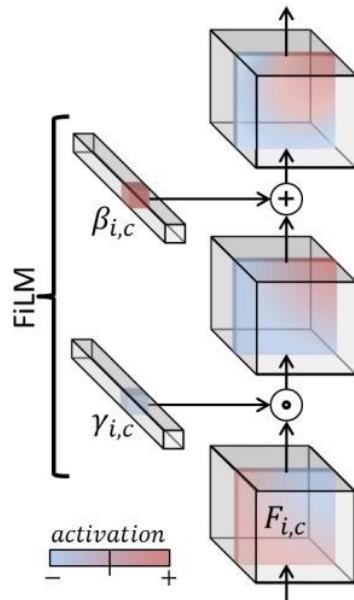
FiLM (Perez et al., 2017)

Modulate one modality, layerwise, by the other.

$$y_{i,c} = f_c(x_i)$$

$$\beta_{i,c} = h_c(x_i)$$

$$\text{FiLM}(F_{i,c} \mid \beta_{i,c}, y_{i,c}) = y_{i,c} F_{i,c} + \beta_{i,c}$$



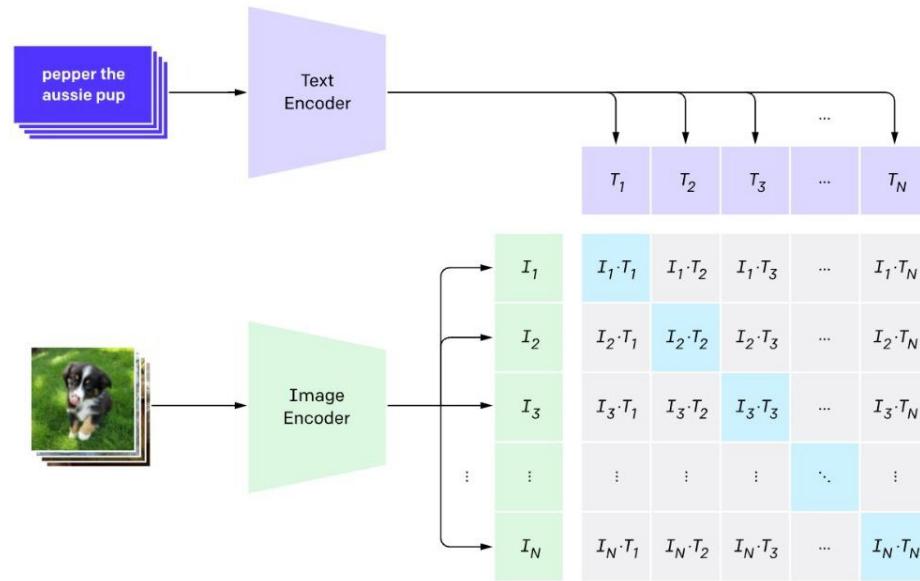
Outline

1. Early models
2. Features and fusion
- 3. Contrastive models**
4. Multimodal foundation models
5. Evaluation
6. Beyond images: Other modalities
7. Where to next?

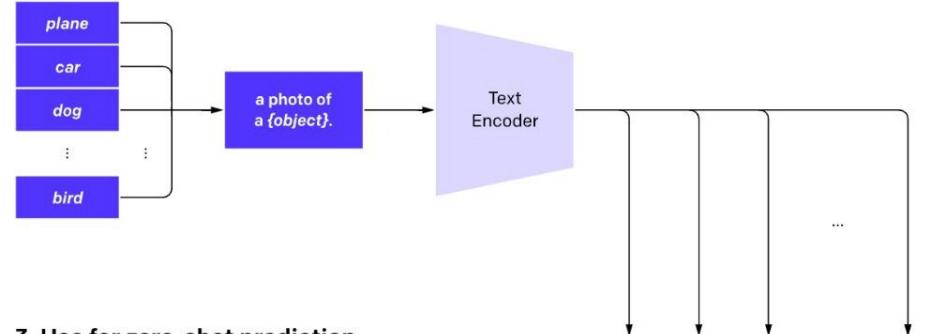
CLIP (Radford et al. 2021)

Exact same contrastive loss as earlier, but.. Transformers and *web data*!

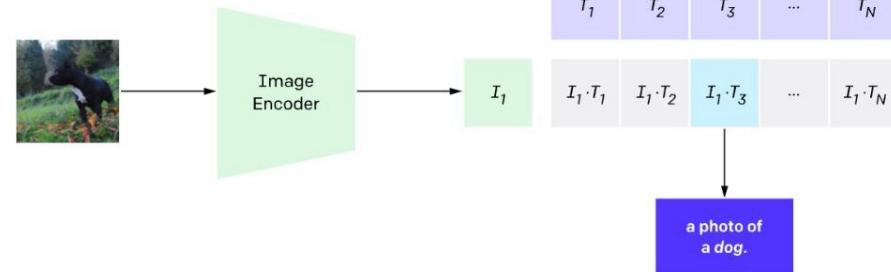
1. Contrastive pre-training



2. Create dataset classifier from label text



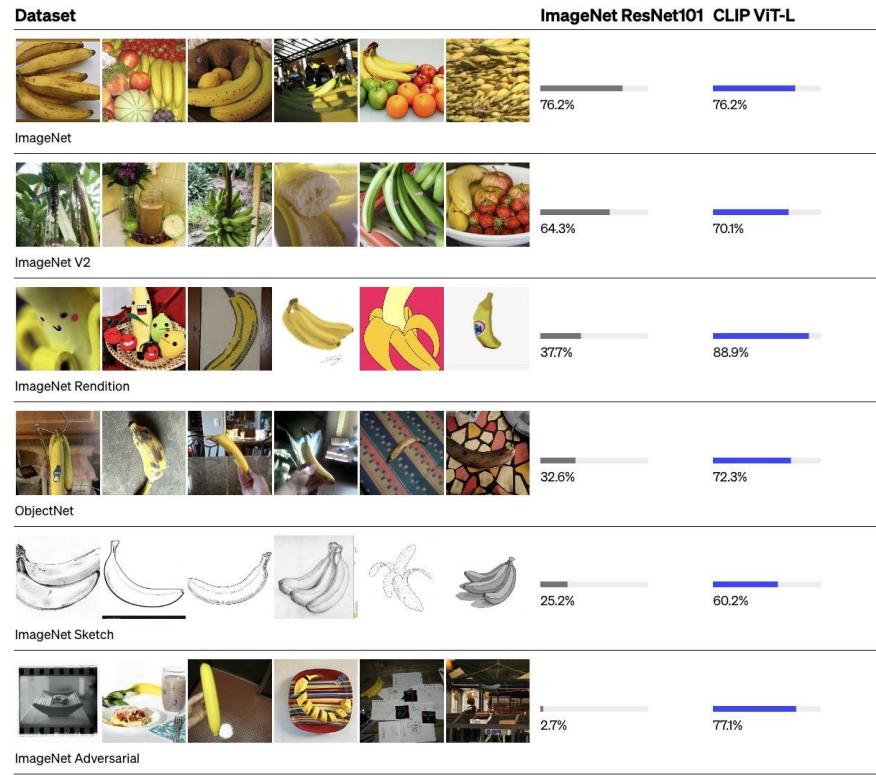
3. Use for zero-shot prediction



CLIP Robustness

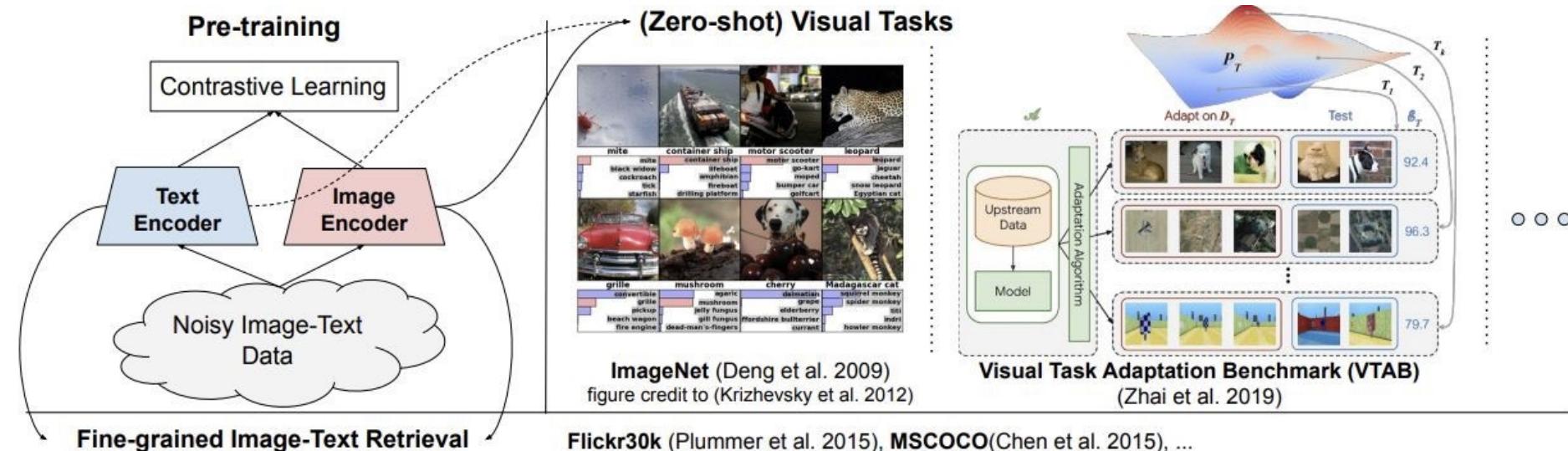
IMHO one of the best papers ever written in our field: extremely thorough, worth a close read.

Generalizes MUCH better →



ALIGN (Jia et al., 2021)

Same idea, but EVEN MORE data (JFT at 1.8B image-text pairs vs CLIP's 300m).



Aligned datasets

HUGE open source datasets of image-text pairs now exist.

Used to train eg StableDiffusion (Rombach et al., 2022).

LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI- MODAL DATASETS

by: Romain Beaumont, 31 Mar, 2022

<https://laion.ai/blog/laion-5b/>

Outline

1. Early models
2. Features and fusion
3. Contrastive models
4. **Multimodal foundation models**
5. Evaluation
6. Beyond images: Other modalities
7. Where to next?

BERT Refresher

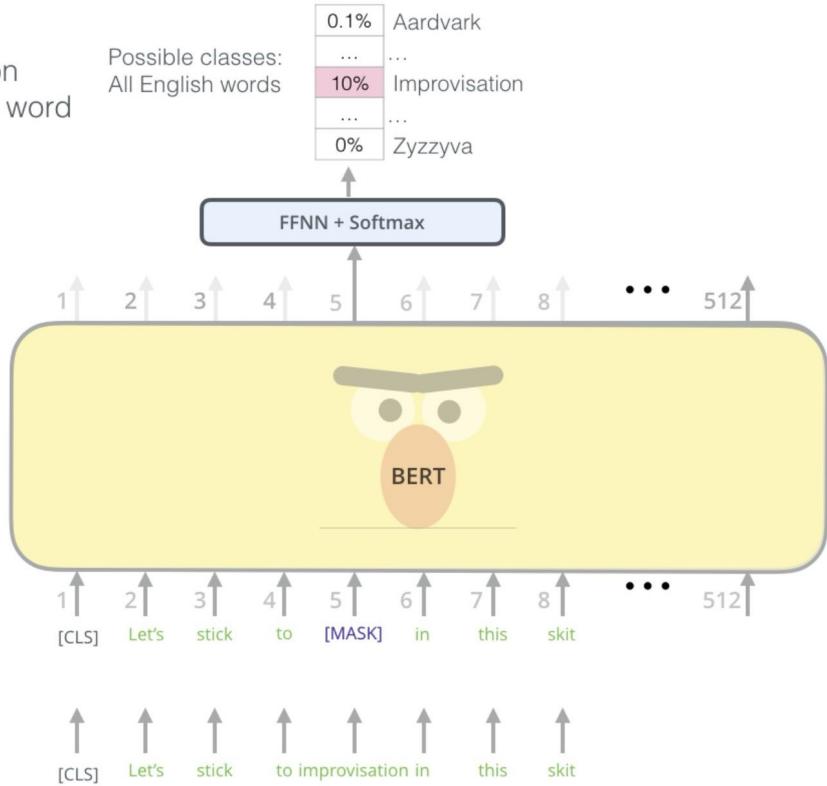
Use the output of the masked word's position to predict the masked word

0.1%	Aardvark
...	...
10%	Improvisation
...	...
0%	Zyzyva

How do we make this multimodal? →

Randomly mask
15% of tokens

Input

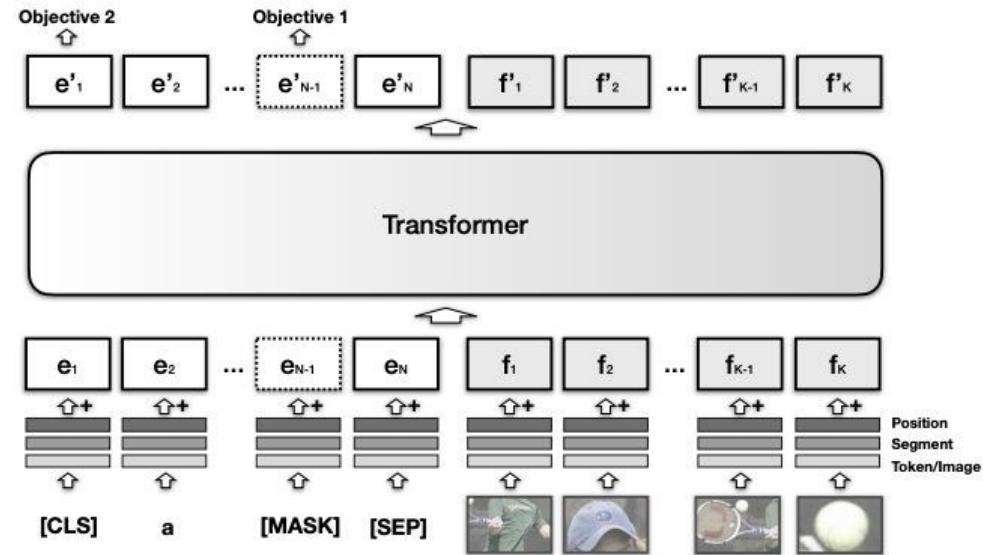


Alammar (2018), [Illustrated Bert](#)

Visual BERTs: VisualBERT

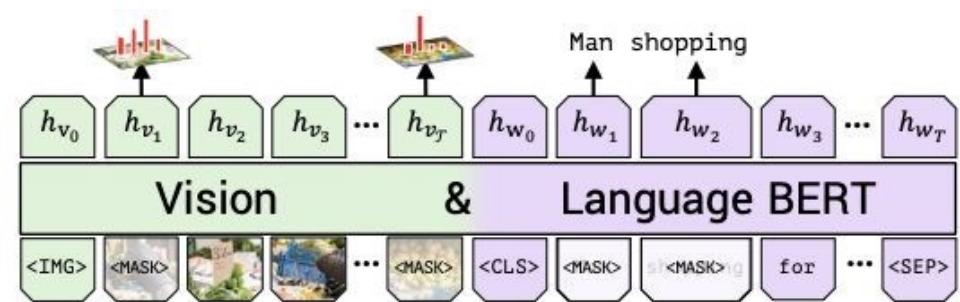
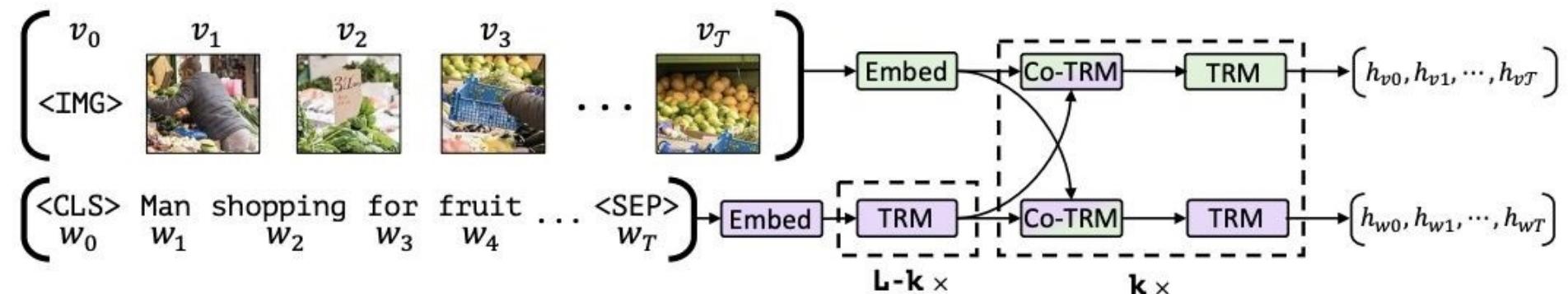


A person hits a ball with a tennis racket

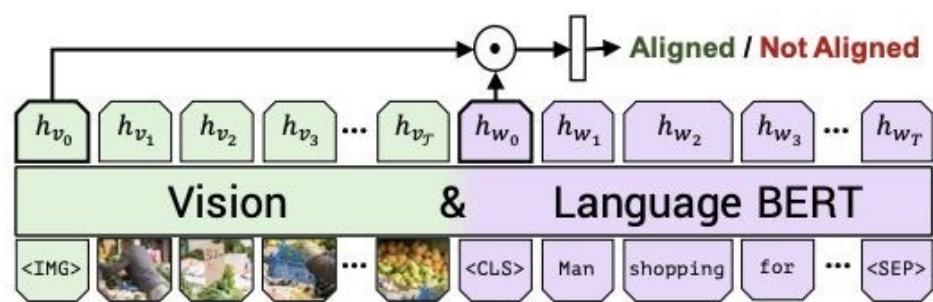


VisualBERT Li et al. 2019

Visual BERTs: ViLBERT



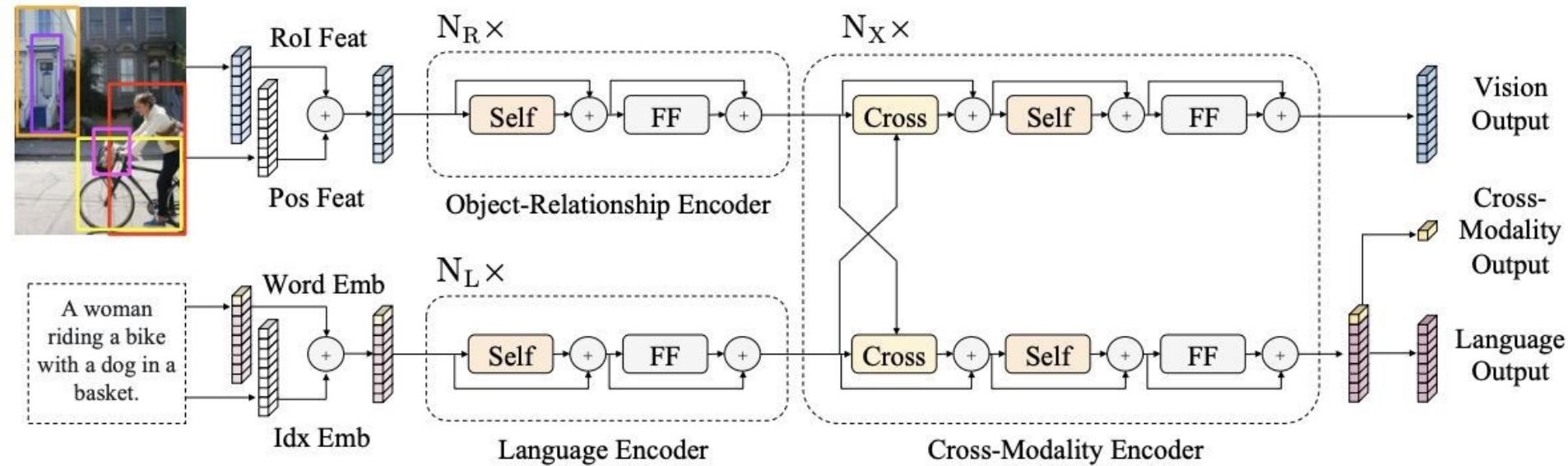
(a) Masked multi-modal learning



(b) Multi-modal alignment prediction

Visual BERTs: LXMERT

Learning Cross-Modality Encoder Representations from Transformers



Visual BERTs: Supervised Multimodal Bitransformers

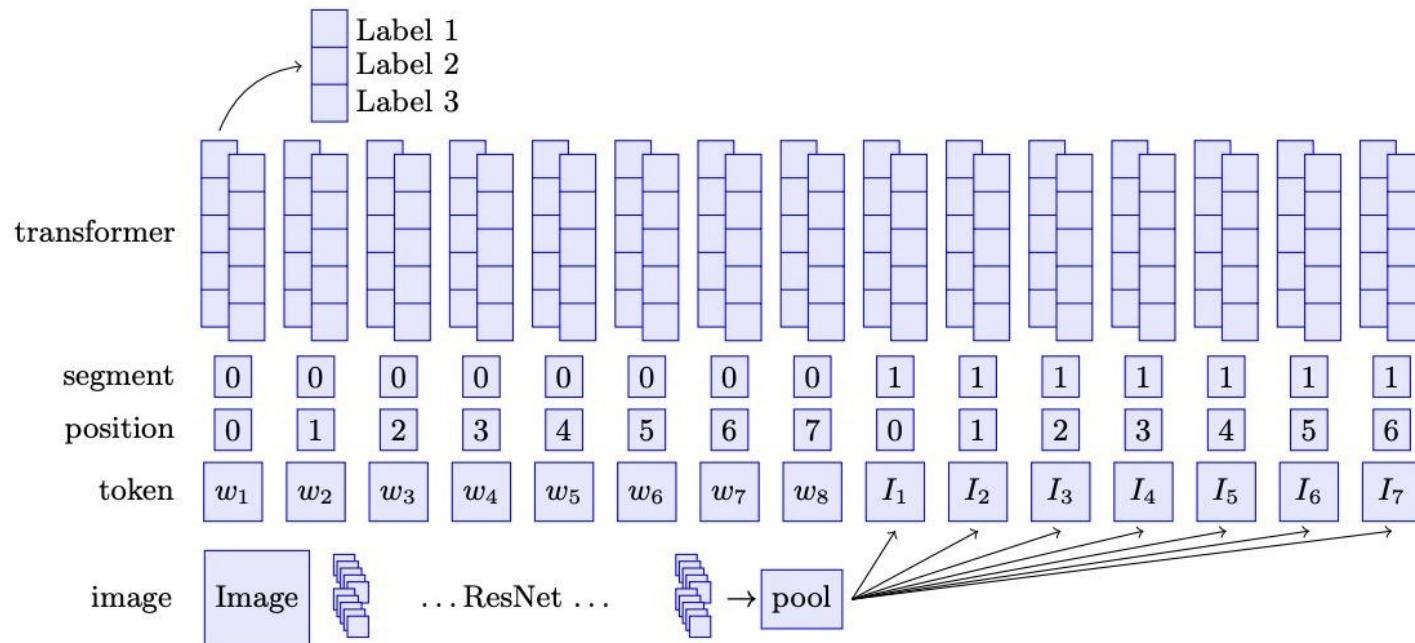
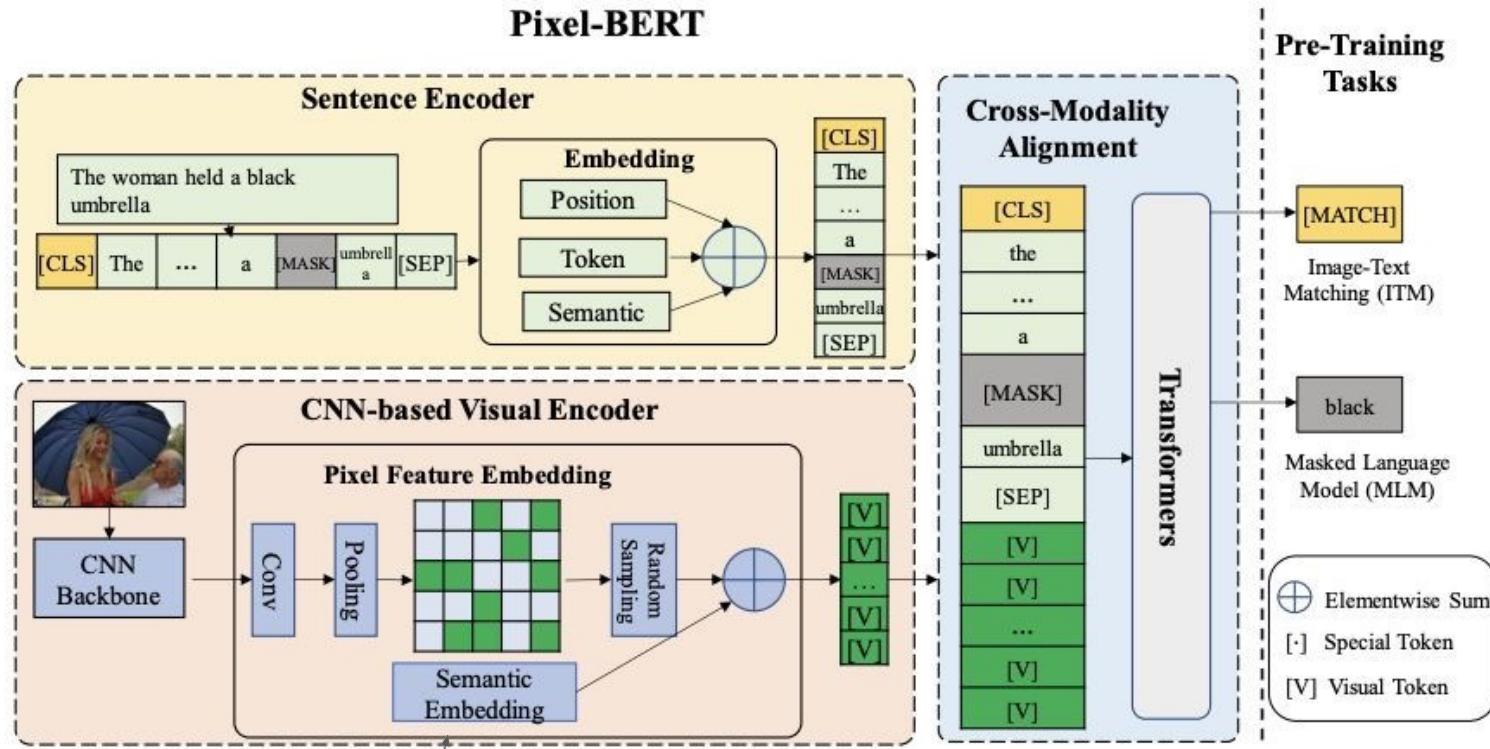


Figure 1: Illustration of the multimodal bitransformer architecture.

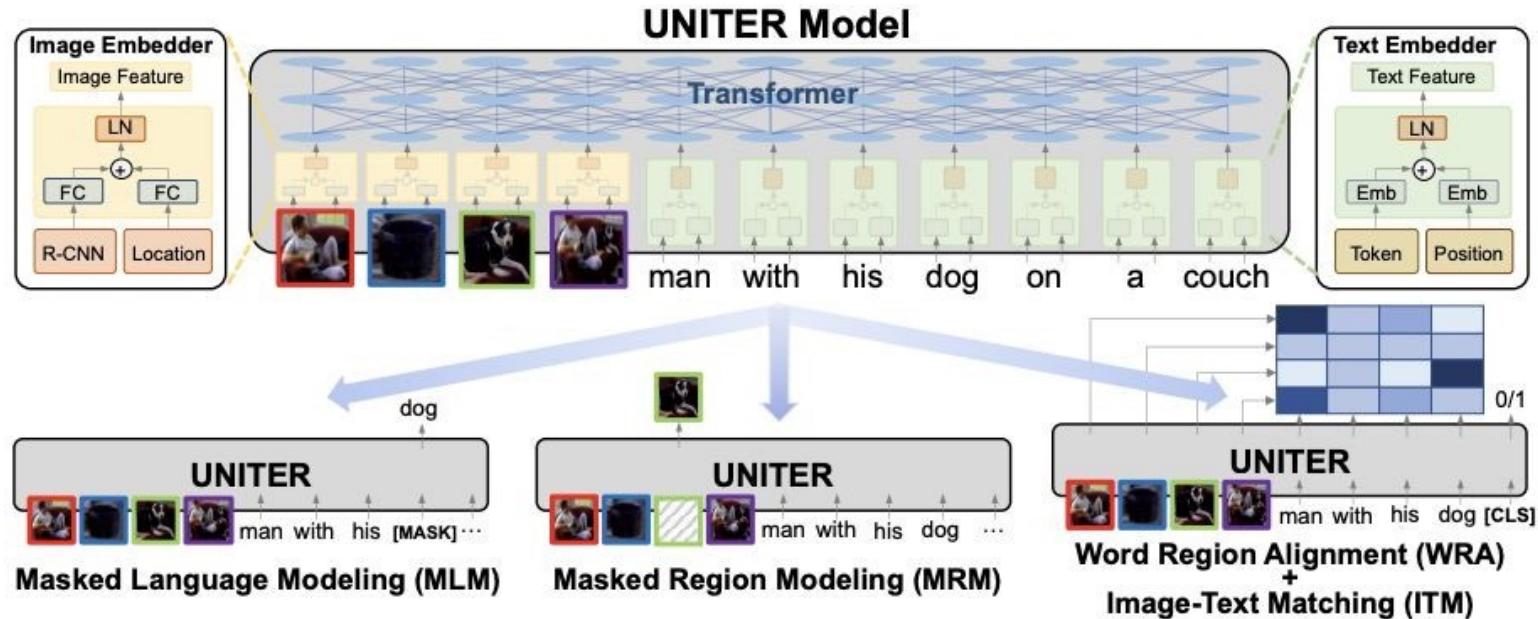
Visual BERTs: PixelBert



Misnomer: they mean segment embedding

PixelBert Huang et al. 2020

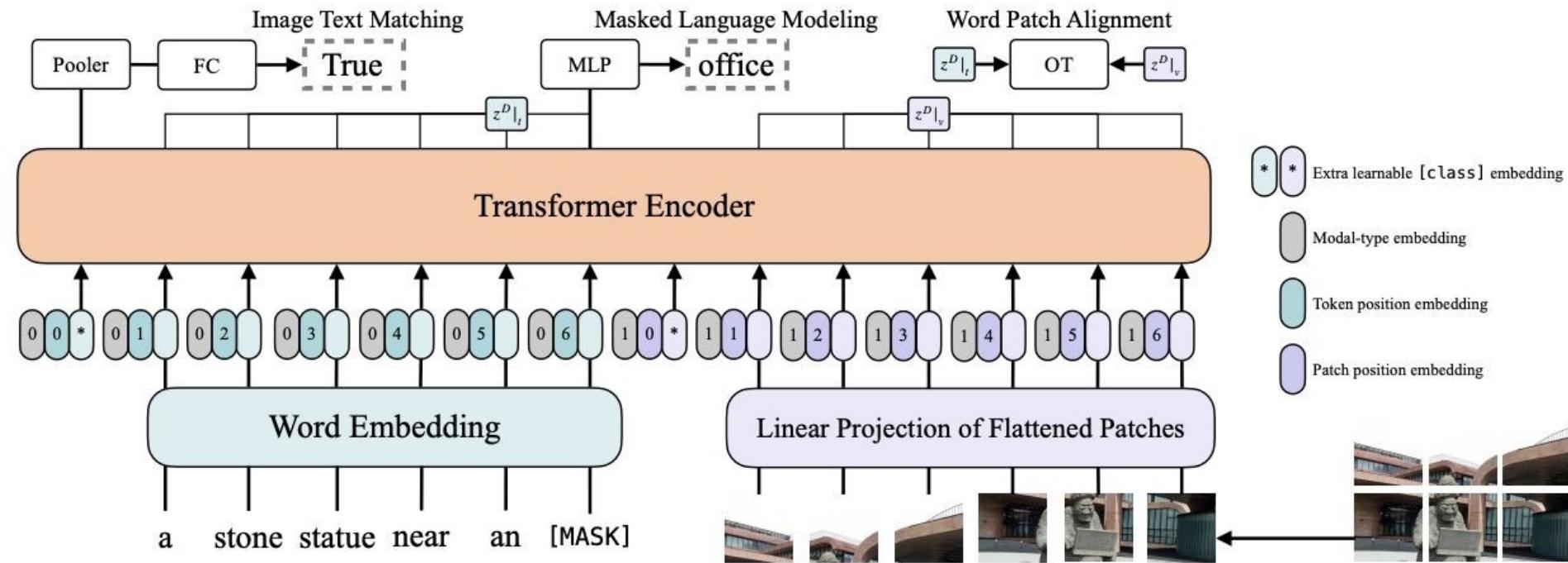
UNITER



Chen, Yi, Lu, et al. 2020

ViLT (Kim et al. 2021)

Feeding data directly to the transformer.



So many models

VL-PTM	Text encoder	Vision encoder	Fusion scheme	Pre-training tasks	Multimodal datasets for pre-training
Fusion Encoder					
VisualBERT [2019]	BERT	Faster R-CNN	Single stream	MLM+ITM	COCO
Uniter [2020]	BERT	Faster R-CNN	Single stream	MLM+ITM+WRA+MRFR+MRC	CC+COCO+VG+SBU
OSCAR [2020c]	BERT	Faster R-CNN	Single stream	MLM+ITM	CC+COCO+SBU+Flickr30k+VQA
InterBert [2020]	BERT	Faster R-CNN	Single stream	MLM+MRC+ITM	CC+COCO+SBU
ViLBERT [2019]	BERT	Faster R-CNN	Dual stream	MLM+MRC+ITM	CC
LXMERT [2019]	BERT	Faster R-CNN	Dual stream	MLM+ITM+MRC+MRFR+VQA	COCO+VG+VQA
VL-BERT [2019]	BERT	Faster R-CNN+ ResNet	Single stream	MLM+MRC	CC
Pixel-BERT [2020]	BERT	ResNet	Single stream	MLM+ITM	COCO+VG
Unified VLP [2020]	UniLM	Faster R-CNN	Single stream	MLM+seq2seq LM	CC
UNIMO [2020b]	BERT, RoBERTa	Faster R-CNN	Single stream	MLM+seq2seq LM+MRC+MRFR+CMCL	COCO+CC+VG+SBU
SOHO [2021]	BERT	ResNet + Visual Dictionary	Single stream	MLM+MVM+ITM	COCO+VG
VL-T5 [2021]	T5, BART	Faster R-CNN	Single stream	MLM+VQA+ITM+VG+GC	COCO+VG
XGPT [2021]	transformer	Faster R-CNN	Single stream	IC+MLM+DAE+MRFR	CC
Visual Parsing [2021]	BERT	Faster R-CNN + Swin transformer	Dual stream	MLM+ITM+MFR	COCO+VG
ALBEF [2021a]	BERT	ViT	Dual stream	MLM+ITM+CMCL	CC+COCO+VG+SBU
SimVLM [2021b]	ViT	ViT	Single stream	PrefixLM	C4+ALIGN
WenLan [2021]	RoBERTa	Faster R-CNN + EffcientNet	Dual stream	CMCL	RUC-CAS-WenLan
ViLT [2021]	ViT	Linear Projection	Single stream	MLM+ITM	CC+COCO+VG+SBU
Dual Encoder					
CLIP [2021]	GPT2	ViT, ResNet	CMCL		self-collected
ALIGN [2021]	BERT	EffcientNet	CMCL		self-collected
DeCLIP [2021b]	GPT2, BERT	ViT, ResNet, RegNetY-64GF	CMCL+MLM+CL		CC+self-collected
Fusion Encoder+ Dual Encoder					
VLMo [2021a]	BERT	ViT	Single stream	MLM+ITM+CMCL	CC+COCO+VG+SBU
FLAVA [2021]	ViT	ViT	Single stream	MMM+ITM+CMCL	CC+COCO+VG+SBU+RedCaps

Recommended paper

Bugliarello et al. (2021). *Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs*. Finding: in the same conditions, models perform *very* similarly.

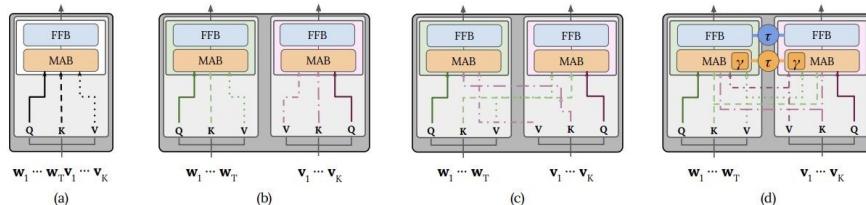


Figure 3: Visualisation of the (a) single-stream, (b) dual-stream intra-modal and (c) dual-stream inter-modal Transformer layers. (d) shows our gated bimodal layer. The inter-modal layer attends across modalities, while the intra-model layer attends within each modality. Ours can attend to either or both.

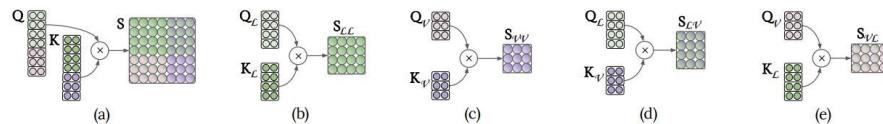
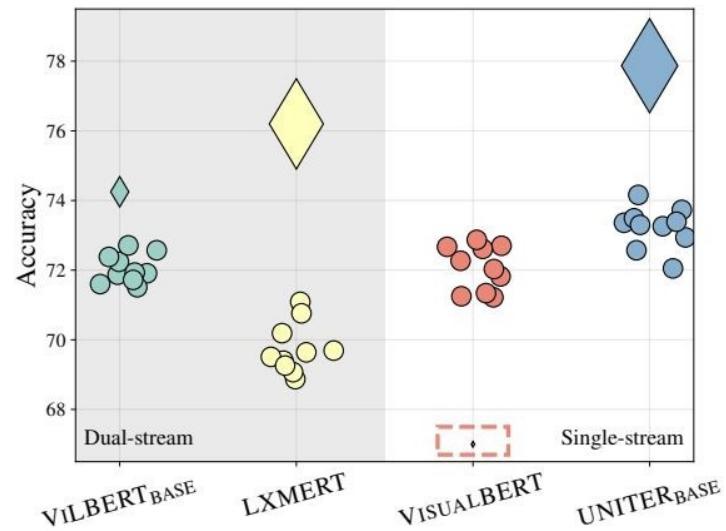


Figure 4: Visualisation of the score matrix for (a) single-stream, (b) text-text, (c) vision–vision, (d) text–vision, and (e) vision–text interactions. Shades of green denote the text modality, while purple ones denote the vision modality. Dual-stream scores are sub-matrices of the single-stream scores matrix.



FLAVA (Singh et al., 2021)

Holistic approach to multimodality.

One foundation model spanning V&L, CV and NLP.

Jointly pretrained on:

- unimodal text data (CCNews + BookCorpus)
- unimodal image data (ImageNet)
- public paired image-text data (70M)

All data/models are publicly released.



The PMD dataset

- 70M image-text pairs from public sources

COCO



A close up view of a pizza sitting on a table with a soda in the back.

Visual Genome



a lenovo laptop rebooting

SBU captions



Front view of basket 13, from the sidewalk in front of the basket.

Localized narratives



The woman is touching a utensil in front of her on the grill stand.

WIT



Typocerus balteatus,
Subfamily: Flower Longhorns

RedCaps



Deigdoh falls in india

CC12M



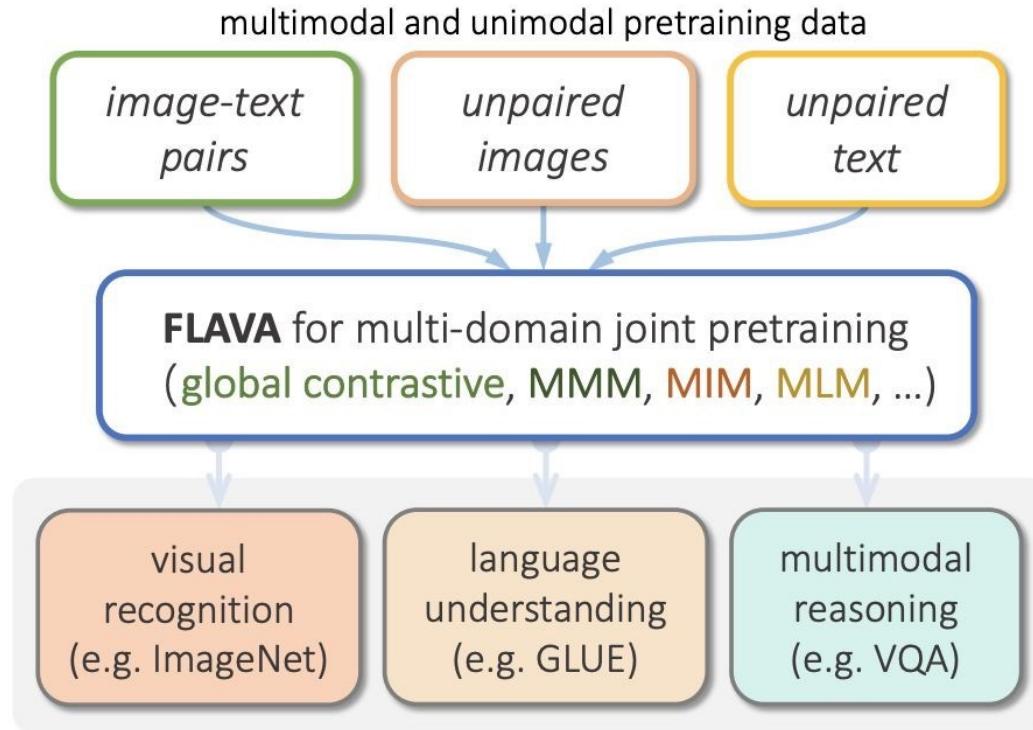
Jumping girl in a green summer dress stock illustration

YFCC filtered

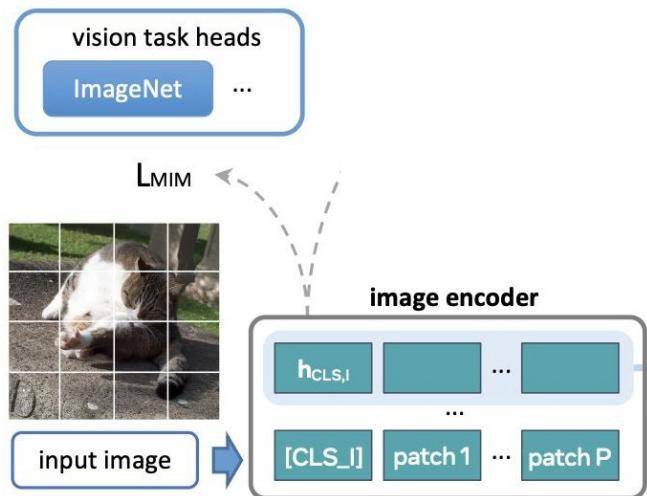


In the kitchen at the Muse Nissim de Camondo

Problem to solve



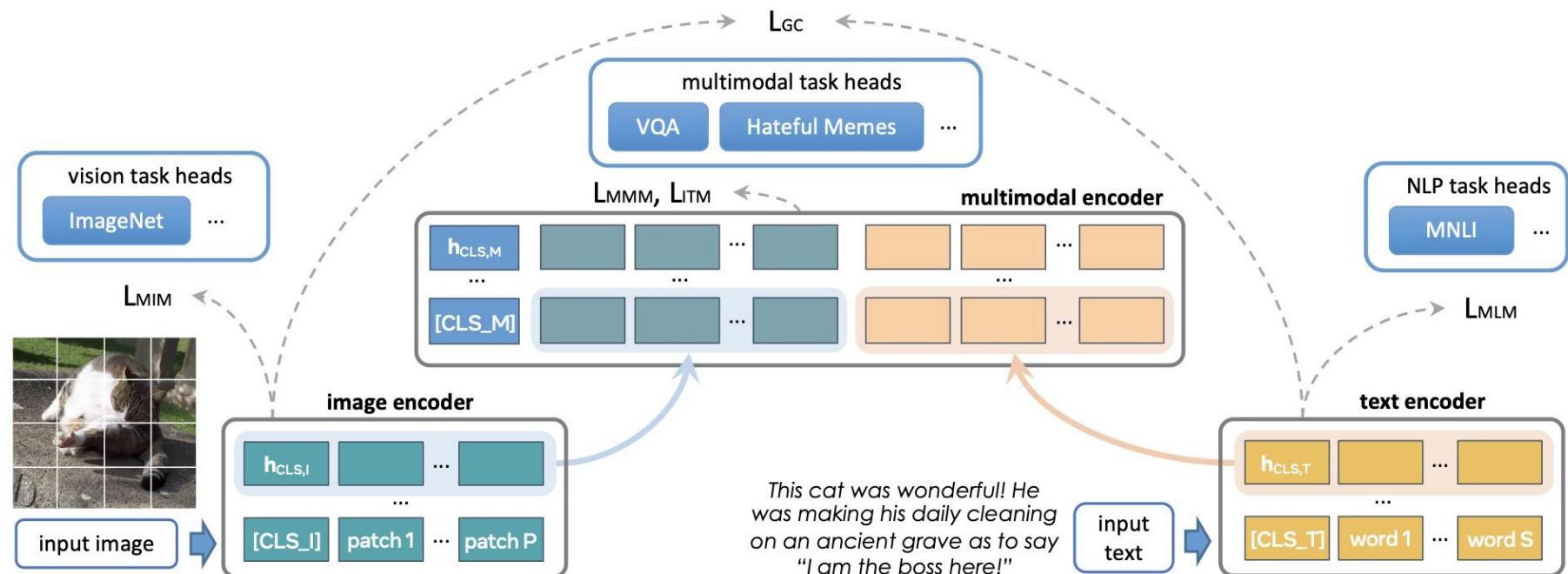
How does FLAVA work?



How does FLAVA work?

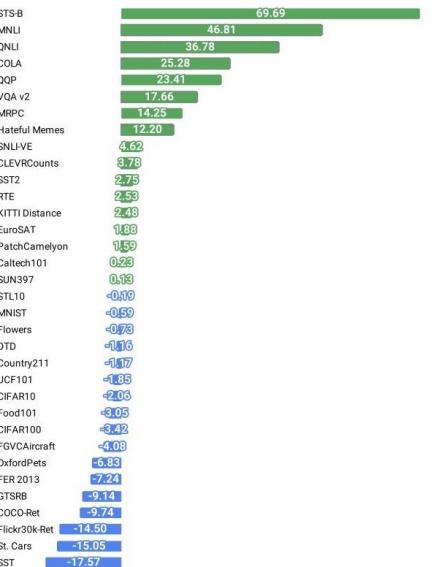


How does FLAVA work?



How well does it work?

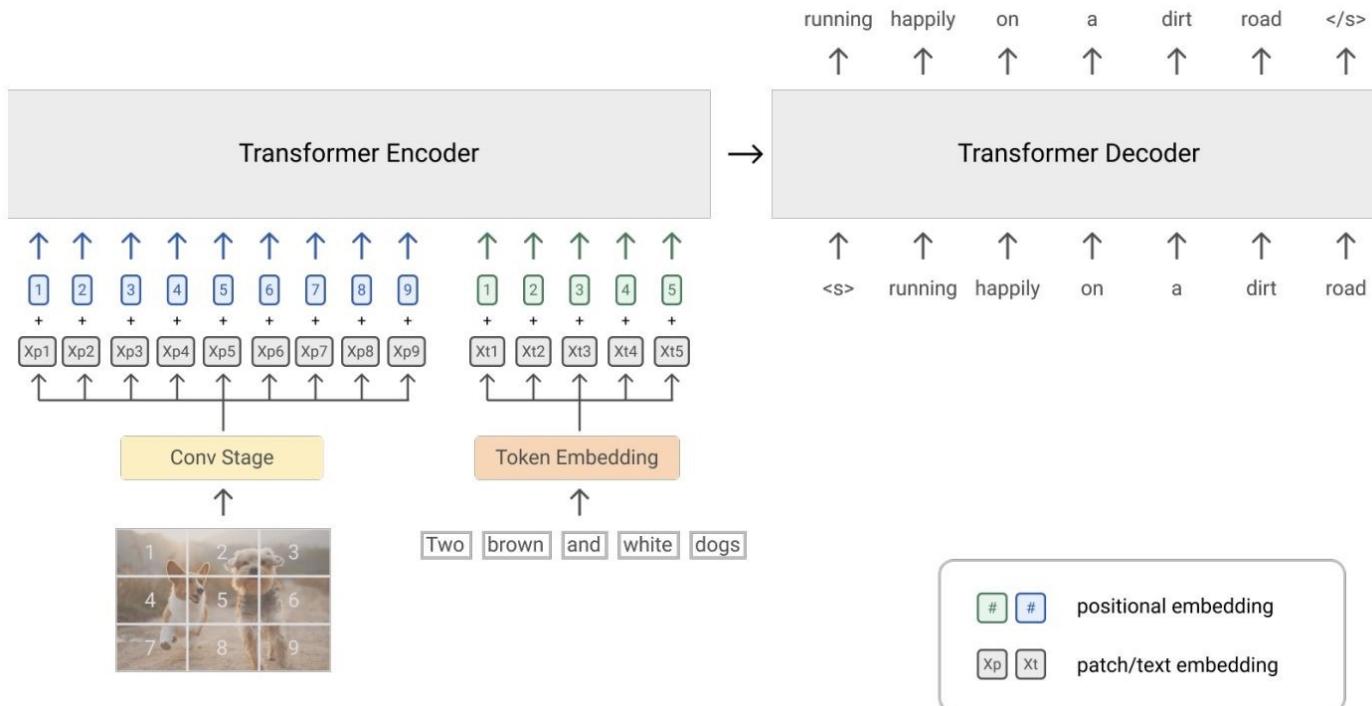
- On average, over 35 tasks, FLAVA obtains impressive performance



		MIM 1	MLM 2	FLAVA_C 3	FLAVA_MM 4	FLAVA w/o init 5	FLAVA 6	CLIP 7	CLIP 8
Datasets	Eval method	PMD	PMD	PMD	PMD	(PMD+IN-1k+CCNews+BC)	PMD	400M [83]	
MNLI [111]	fine-tuning	—	73.23	70.99	76.82	78.06	80.33	32.85	33.52
CoLA [110]	fine-tuning	—	39.55	17.58	38.97	44.22	50.65	11.02	25.37
MRPC [29]	fine-tuning	—	73.24	76.31	79.14	78.91	84.16	68.74	69.91
QQP [49]	fine-tuning	—	86.68	85.94	88.49	98.61	88.74	59.17	65.33
SST-2 [97]	fine-tuning	—	87.96	86.47	89.33	90.14	90.94	83.49	88.19
QNLI [88]	fine-tuning	—	82.32	71.85	84.77	86.40	87.31	49.46	50.54
RTE [7, 25, 36, 40]	fine-tuning	—	50.54	51.99	51.99	54.87	57.76	53.07	55.23
STS-B [1]	fine-tuning	—	78.89	57.28	84.29	83.21	85.67	13.70	15.98
NLP Avg.		—	71.55	64.80	74.22	75.55	78.19	46.44	50.50
ImageNet [90]	linear eval	41.79	—	74.09	74.34	73.49	75.54	72.95	80.20
Food101 [1]	linear eval	53.30	—	87.77	87.53	87.39	88.51	85.49	91.56
CIFAR10 [58]	linear eval	76.20	—	93.44	92.37	92.63	92.87	91.25	94.93
CIFAR100 [58]	linear eval	55.57	—	78.37	78.01	76.49	77.68	74.40	81.10
Cars [56]	linear eval	14.71	—	72.12	72.07	66.81	70.87	62.84	85.92
Aircraft [74]	linear eval	13.83	—	49.74	48.90	44.73	47.31	40.02	51.40
DTD [20]	linear eval	55.53	—	76.86	76.91	75.80	77.29	73.40	78.46
Pets [79]	linear eval	34.48	—	84.98	84.93	82.77	84.82	79.61	91.66
Caltech101 [32]	linear eval	67.36	—	94.91	95.32	94.95	95.74	93.76	95.51
Flowers102 [76]	linear eval	67.23	—	96.36	96.39	95.58	96.37	94.94	97.12
MNIST [60]	linear eval	96.40	—	98.39	98.58	98.70	98.42	97.38	99.01
STL10 [24]	linear eval	80.12	—	98.06	98.31	98.32	98.89	97.29	99.09
EuroSAT [41]	linear eval	95.48	—	97.00	96.98	97.04	97.26	95.70	95.38
GTSRB [100]	linear eval	63.14	—	78.92	77.93	77.71	79.46	76.34	88.61
KITTI [35]	linear eval	86.03	—	87.83	88.84	88.70	89.04	84.89	86.56
PCAM [106]	linear eval	85.10	—	85.02	85.51	85.72	85.31	83.99	83.72
UCF101 [98]	linear eval	46.34	—	82.69	82.90	81.42	83.32	77.85	85.17
CLEVR [52]	linear eval	61.51	—	79.35	81.66	80.62	79.66	73.64	75.89
FER 2013 [38]	linear eval	50.98	—	59.96	60.87	58.99	61.12	57.04	68.36
SUN397 [113]	linear eval	52.45	—	81.27	81.41	81.05	82.17	79.96	82.05
SST [83]	linear eval	57.77	—	56.67	59.25	56.40	57.11	56.84	74.68
Country211 [83]	linear eval	8.87	—	27.27	26.75	27.01	28.92	25.12	30.10
Vision Avg.		57.46	—	79.14	79.35	78.29	79.44	76.12	82.57
VQAv2 [39]	fine-tuning	—	—	67.13	71.69	71.29	72.49	59.81	54.83
SNLI-VE [114]	fine-tuning	—	—	73.27	78.36	78.14	78.89	73.53	74.27
Hateful Memes [33]	fine-tuning	—	—	55.58	70.72	77.45	76.09	56.59	63.93
Flickr30K [81]	TR R@1	zero-shot	—	68.30	69.30	64.50	67.70	60.90	82.20
Flickr30K [81]	TR R@5	zero-shot	—	93.50	92.90	90.30	94.00	88.90	96.60
Flickr30K [81]	IR R@1	zero-shot	—	60.56	63.16	60.04	65.22	56.48	62.08
Flickr30K [81]	IR R@5	zero-shot	—	86.68	87.70	86.46	89.38	83.60	85.68
COCO [66]	TR R@1	zero-shot	—	43.08	43.48	39.88	42.74	37.12	52.48
COCO [66]	TR R@5	zero-shot	—	75.82	76.76	72.84	76.76	69.48	76.68
COCO [66]	IR R@1	zero-shot	—	37.59	38.46	34.95	38.38	33.29	33.07
COCO [66]	IR R@5	zero-shot	—	67.28	67.68	64.63	67.47	62.47	58.37
Multimodal Avg.		—	—	66.25	69.11	67.32	69.92	62.02	67.29
Macro Avg.		19.15	23.85	70.06	74.23	73.72	75.85	61.52	66.78

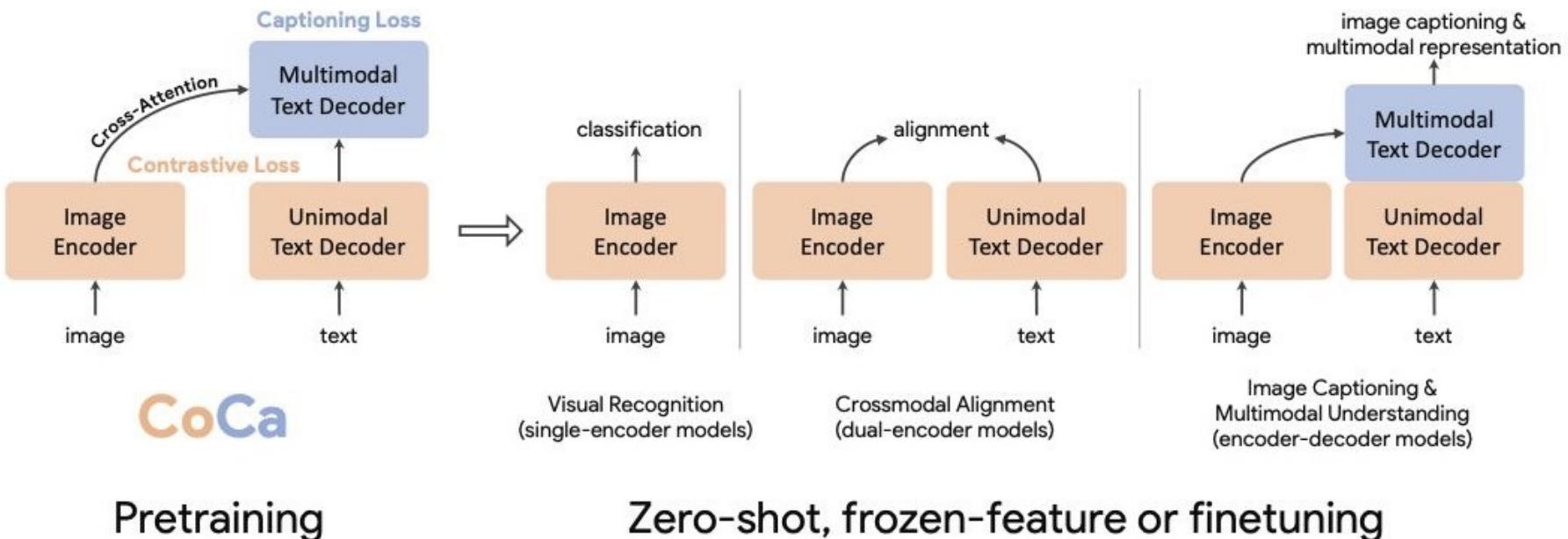
SimVLM (Wang et al., 2022)

Slowly moving from contrastive/discriminative to generative.



CoCa Contrastive Captioner (Yu et al., 2022)

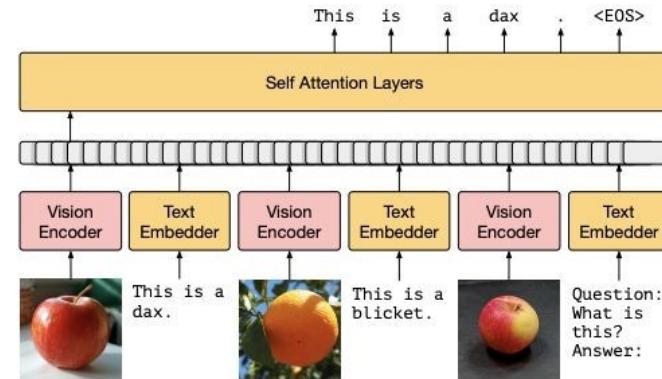
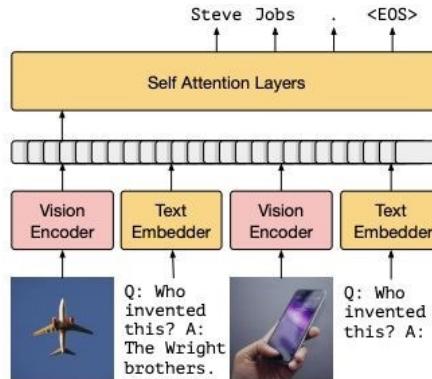
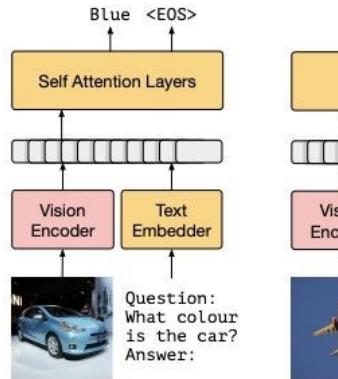
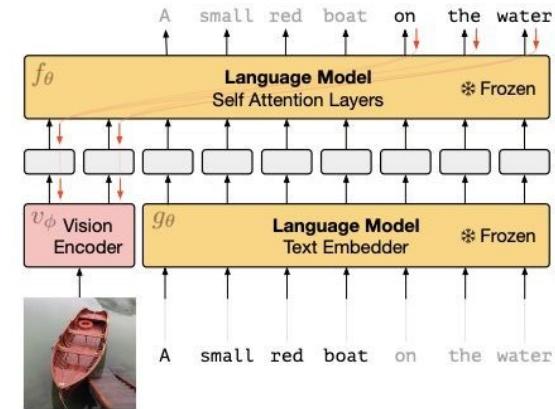
Best of both (contrastive and generative) worlds.



Frozen (Tsimpoukelli, Menick, Cabi, et al., 2021)

Kind of like MMBT but with a better LLM (T5) and a better vision encoder (NF-ResNet).

Multi-Modal Few-Shot Learners!



(a) 0-shot VQA

(b) 1-shot outside-knowledge VQA

(c) Few-shot image classification

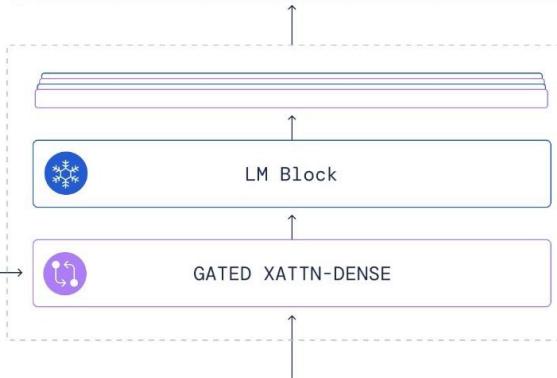
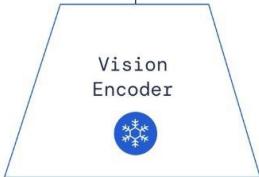
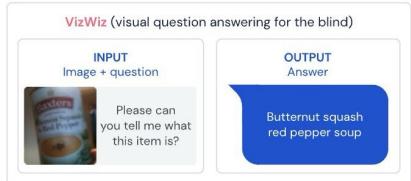
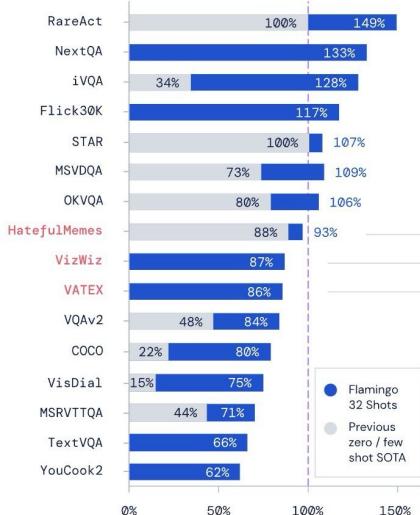


... a flamingo. They are found in the Caribbean.

Flamingo (Alayrac et al., 2022)

80b param model based on Chinchilla.
Multi-image.

Performance relative to SOTA



This is ...

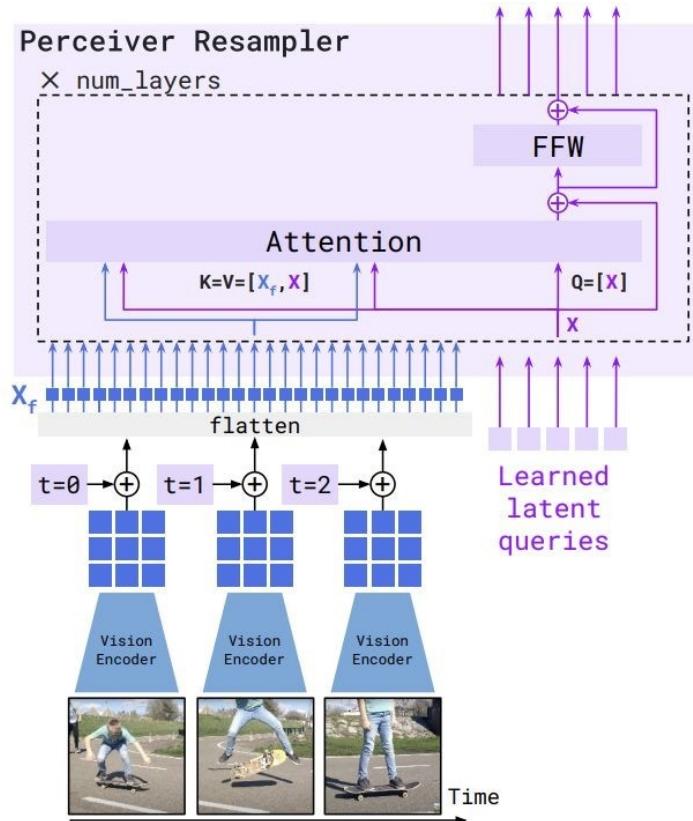
This is a chinchilla.
They are mainly found in Chile.

This is a shiba.
They are very popular in Japan.

This is ...

- Pretrained and frozen
- Trained from scratch during Flamingo training

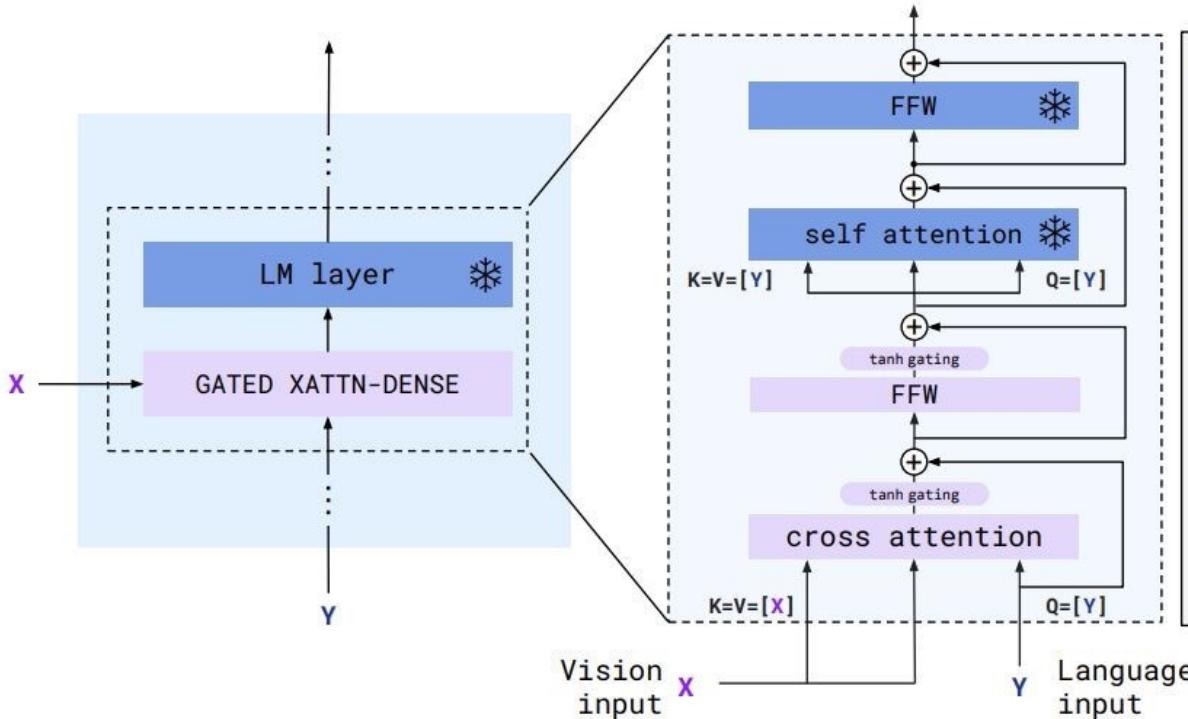
Perceiver Resampler



```
def perceiver_resampler(  
    x_f, # The [T, S, d] visual features (T=time, S=space)  
    time_embeddings, # The [T, 1, d] time pos embeddings.  
    x, # R learned latents of shape [R, d]  
    num_layers, # Number of layers  
):  
    """The Perceiver Resampler model."""  
  
    # Add the time position embeddings and flatten.  
    x_f = x_f + time_embeddings  
    x_f = flatten(x_f) # [T, S, d] -> [T * S, d]  
    # Apply the Perceiver Resampler layers.  
    for i in range(num_layers):  
        # Attention.  
        x = x + attention_i(q=x, kv=concat([x_f, x]))  
        # Feed forward.  
        x = x + ffw_i(x)  
    return x
```

Gated XATTN

Inject visual info directly into a frozen LM via cross-attention (remember FiLM?).



```
def gated_xattn_dense(
    y, # input language features
    x, # input visual features
    alpha_xattn, # xattn gating parameter - init at 0.
    alpha_dense, # ffw gating parameter - init at 0.
):
    """Applies a GATED XATTN-DENSE layer."""

    # 1. Gated Cross Attention
    y = y + tanh(alpha_xattn) * attention(q=y, kv=x)

    # 2. Gated Feed Forward (dense) Layer
    y = y + tanh(alpha_dense) * ffw(y)

    # Regular self-attention + FFW on language
    y = y + frozen_attention(q=y, kv=y)
    y = y + frozen_ffw(y)
    return y # output visually informed language features
```

Why is this funny?

Original image from Karpathy as a
“visual Turing test” →

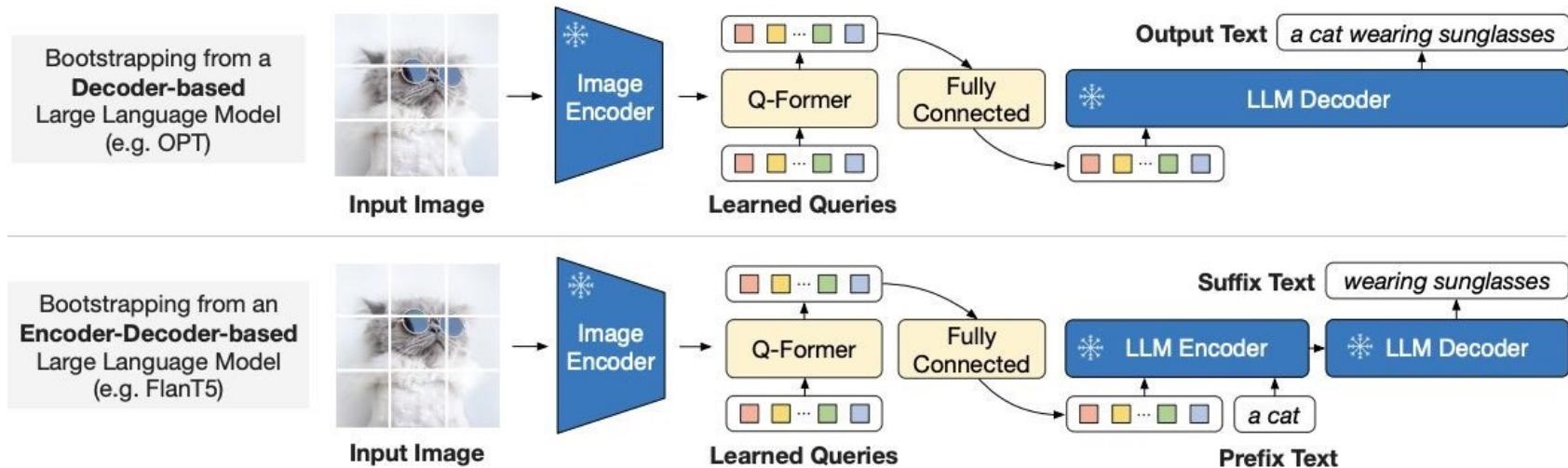


- This is a picture of Barack Obama. He is a former president of the United States.
- What is he doing?
- He is looking at the scale.
- Where is Obama's foot positioned?
- Obama's foot is positioned on the right side of the scale.
- Where was this picture taken?
- It was taken in a school.
- What happens as a result?
- The scale shows a higher weight.
- Is the person on the scale aware of it?
- I think he is not aware of it.
- Do you think that's why people are laughing?
- I think so.
- I think you are right.

BLIP/BLIP2 (Li et al., 2023)

Freeze it all (CLIP-ViT / OPT decoder / FlanT5 encoder-decoder)

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





Explain the advantages of this product.

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.



8

8



Tell me something about the history of this place.

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.



8

8



Write down the facts that you know about this flower.

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.



8

8



Is this photo unusual?



Yes, it's a house that looks like it's upside down.

8

8

How could someone get out of the house?



It has a slide on the side of the house.

8

8



What are shown in the photo?



A man and a chicken.

8

8

What does the man feel and why?



He is scared of the chicken because it is flying at him.

8

8



What are the ingredients I need to make this?



Pepperoni, mozzarella cheese, pizza sauce, olive oil, salt, pepper, basil.

8

8

What is the first step?
Place the pizza dough on a baking sheet, brush with olive oil, sprinkle with salt, pepper, and basil.



Which city is this?



Singapore.

8

8

Why do you think so?



The city has a statue of a merlion.



What happened at the end of this movie?



The titanic sank.

8

8

Did Leonardo DiCaprio's character survive?



No, he drowned.



What is in the photo?



A pizza that looks like a cat.

8

8

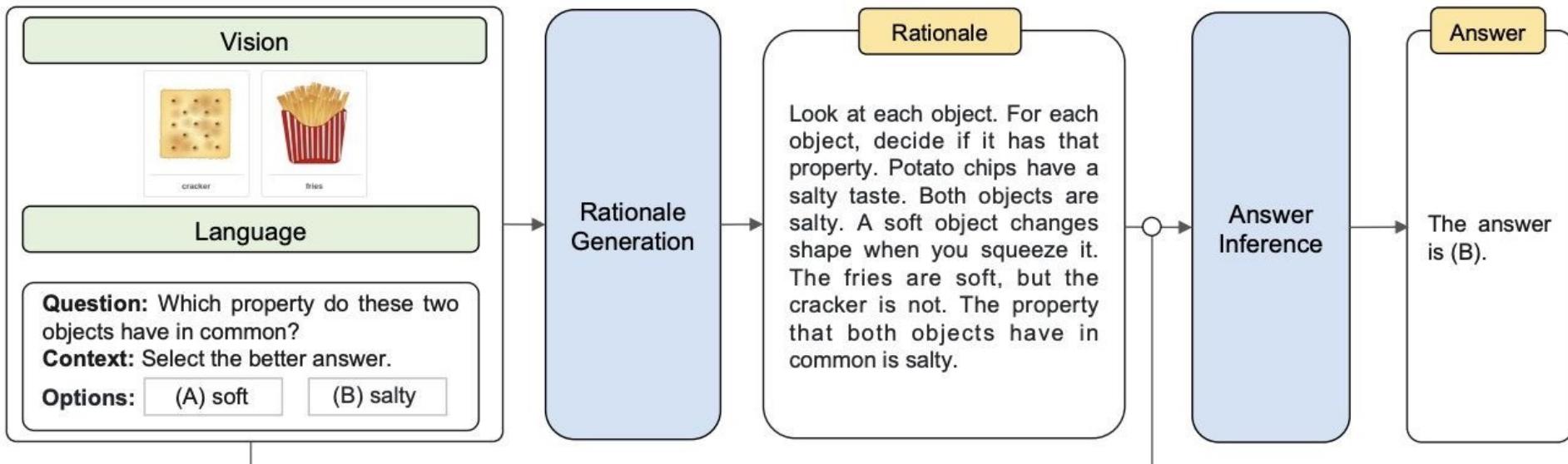
What is the nose made of?



A slice of pepperoni.

Multimodal “Chain of Thought” (Zhang et al., 2023)

Providing a rationale helps give the right answer.



KOSMOS-1 (Huang et al., 2023)

LLMs => MLLMs == FMs

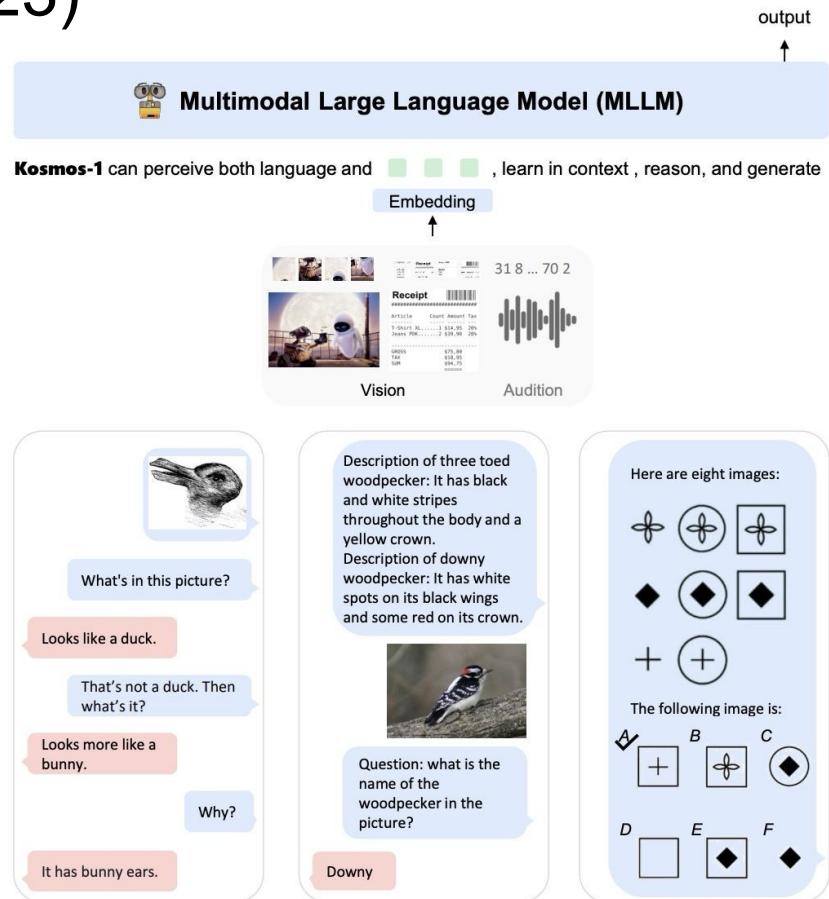
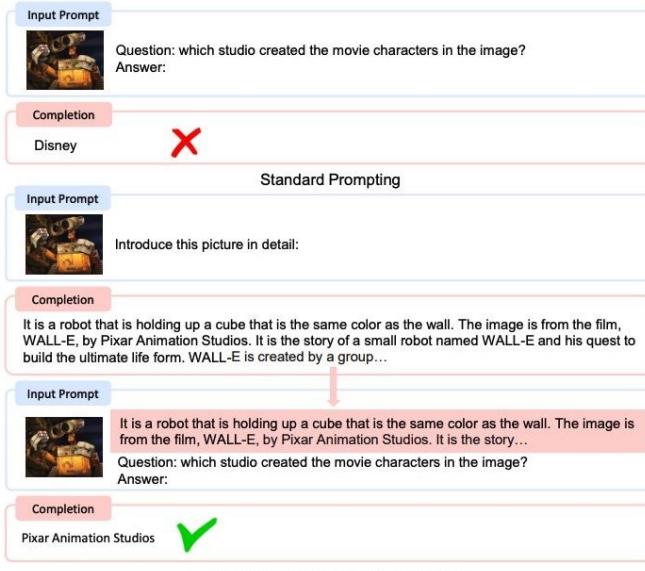


Figure 5: Multimodal Chain-of-Thought prompting enables KOSMOS-1 to generate a rationale first, then to tackle complex question-answering and reasoning tasks.

Outline

1. Early models
2. Features and fusion
3. Contrastive models
4. Multimodal foundation models
- 5. Evaluation**
6. Beyond images: Other modalities
7. Where to next?

What is COCO?

COCO - Common Objects in Context

Super impactful datasets (Lin et al. 2014; Chen et al. 2015)

Main multimodal tasks:

- Image captioning
- Image-caption retrieval

Similar datasets:

- Flickr30k, ConceptualCaptions, VisualGenome, SBU, RedCaps, LAION



The man at bat readies to swing at the pitch while the umpire looks on.



COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

- ✓ Object segmentation
- ✓ Recognition in context
- ✓ Superpixel stuff segmentation
- ✓ 330K images (>200K labeled)
- ✓ 1.5 million object instances
- ✓ 80 object categories
- ✓ 91 stuff categories
- ✓ 5 captions per image
- ✓ 250,000 people with keypoints

cocodataset.org

VQA - Visual Question Answering (Antol et al., 2015)

- The dominant task in vision and language.
 - VQA/VQAv2 citations: 4305+1684
 - COCO Captions: 1647
 - Flickr30k: 1228
- At first the “V” in VQA was found to not matter all that much, so a follow-up VQAv2 dataset was created (Goyal et al., 2017).
- There's also GQA (Hudson & Manning, 2019).



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

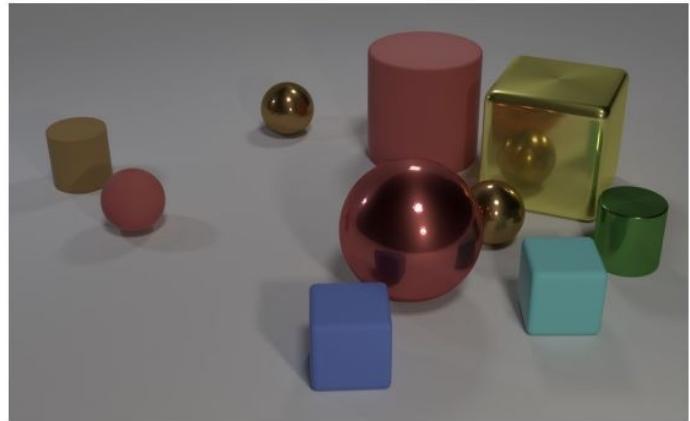


Does it appear to be rainy?
Does this person have 20/20 vision?

CLEVR (Johnson et al., 2016)

Compositional language and elementary visual reasoning diagnostics in a controlled setting.

Hand crafted for measuring compositionality.



- Q:** Are there an equal number of large things and metal spheres?
- Q:** What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
- Q:** How many objects are either small cylinders or metal things?

Figure 1. A sample image and questions from CLEVR. Questions test aspects of visual reasoning such as attribute identification, counting, comparison, multiple attention, and logical operations.

Hateful Memes (Kiela et al., 2020)

Motivated by the shortcomings of other V&L datasets: we need something that is harder, more realistic, and requires true multimodal reasoning and understanding.



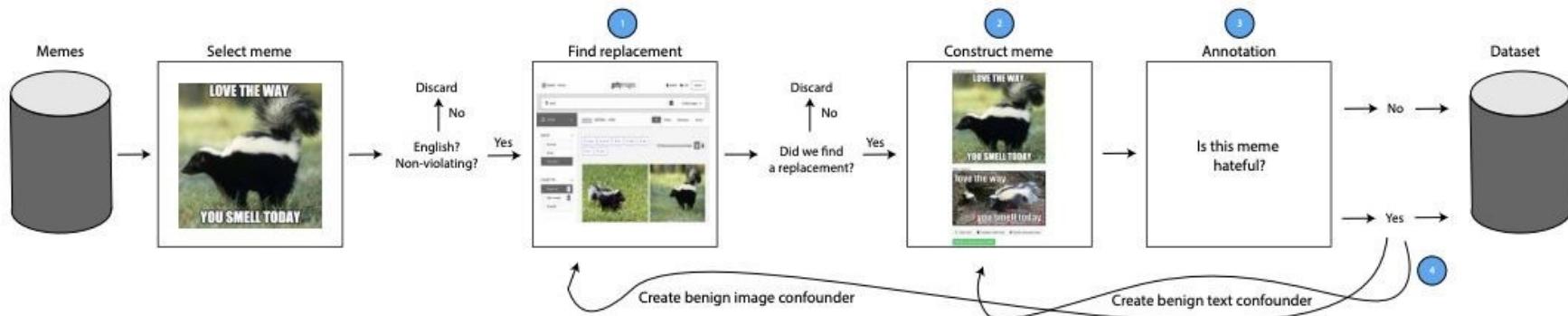
“Mean meme” examples for illustrative purposes – not actually in the dataset

Hateful Memes

Highly trained annotators, so: decent quality but small and expensive

Key concept: benign confounders

A “challenge set” for the community to do zero-shot/finetuning from pretrained



Hateful Memes

Findings in the paper:

- Big gap with human performance.
- Region features
(as opposed to grid) seem to help.
- Earlier fusion is better than middle,
is better than late.
- Multimodal pretraining doesn't
really work.

Type	Model	Validation		Test	
		Acc.	AUROC	Acc.	AUROC
	Human	-	-	84.70	-
Unimodal	Image-Grid	50.67	52.33	52.73±0.72	53.71±2.04
	Image-Region	52.53	57.24	52.36±0.23	57.74±0.73
	Text BERT	58.27	65.05	62.80±1.42	69.00±0.11
Multimodal (Unimodal Pretraining)	Late Fusion	59.39	65.07	63.20±1.09	69.30±0.33
	Concat BERT	59.32	65.88	61.53±0.96	67.77±0.87
	MMBT-Grid	59.59	66.73	62.83±2.04	69.49±0.59
	MMBT-Region	64.75	72.62	67.66±1.39	73.82±0.20
	ViLBERT	63.16	72.17	65.27±2.40	73.32±1.09
	Visual BERT	65.01	74.14	66.67±1.68	74.42±1.34
Multimodal (Multimodal Pretraining)	ViLBERT CC	66.10	73.02	65.90±1.20	74.52±0.06
	Visual BERT COCO	65.93	74.14	69.47±2.06	75.44±1.86

Hateful Memes Competition

After the paper came a \$100k competition on an unseen test set:

Type	Model	Unseen Dev		Unseen Test	
		Acc.	AUROC	Acc.	AUROC
Unimodal	Image-Region	61.48	53.54	60.28±0.18	54.64±0.80
	Text BERT	60.37	60.88	63.60±0.54	62.65±0.40
Multimodal (Unimodal Pretraining)	Late Fusion	61.11	61.00	64.06±0.02	64.44±1.60
	Concat BERT	64.81	65.42	65.90±0.82	66.28±0.66
	MMBT-Grid	67.78	65.47	66.85±1.61	67.24±2.53
	MMBT-Region	70.04	71.54	70.10±1.39	72.21±0.20
	ViLBERT	69.26	72.73	70.86±0.70	73.39±1.32
	Visual BERT	69.67	71.10	71.30±0.68	73.23±1.04
	ViLBERT CC	70.37	70.78	70.03±1.07	72.78±0.50
Multimodal (Multimodal Pretraining)	Visual BERT COCO	70.77	73.70	69.95±1.06	74.59±1.56

#	Team	AUROC	Acc.
1	Ron Zhu	0.844977	0.7320
2	Niklas Muennighoff	0.831037	0.6950
3	Team HateDetectron	0.810845	0.7650
4	Team Kingsterdam	0.805254	0.7385
5	Vlad Sandulescu	0.794321	0.7430

Winner characteristics: frameworks matter, SOTA pretrained models, ensembles, entities, faces and external knowledge.
STILL FAR FROM SOLVED.

Winoground

How good is CLIP really?

Some relevant ideas/findings from NLP:

- Winograd schemas
“The [trophy] doesn't fit in the [suitcase] because *it* is too [large/small]”
- Word order may not matter all that much



(a) some plants surrounding a lightbulb



(b) a lightbulb surrounding some plants

**Masked Language Modeling and the Distributional Hypothesis:
Order Word Matters Pre-training for Little**

Koustuv Sinha^{†‡} Robin Jia[†] Dieuwke Hupkes[†] Joelle Pineau^{†‡}

Adina Williams[†] Douwe Kiela[†]

Winoground

- Examples written by linguist experts
- Using Getty Images API
- Simple way to measure by comparing scores
- In some cases, very difficult and requiring world knowledge



(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



(a) the kid [with the magnifying glass] looks at them []

(c) the person with the ponytail [packs] stuff and other [buys] it

(e) there are [three] people and [two] windows



(b) the kid [] looks at them [with the magnifying glass]

(d) the person with the ponytail [buys] stuff and other [packs] it

(f) there are [two] people and [three] windows

Pragmatics

Series

Symbolic

Winoground Findings

SOTA models often perform *below chance*.

Model	Text	Image	Group
MTurk Human	89.50	88.50	85.50
Random Chance	25.00	25.00	16.67
VinVL	37.75	17.75	14.50
UNITER _{large}	38.00	14.00	10.50
UNITER _{base}	32.25	13.25	10.00
ViLLA _{large}	37.00	13.25	11.00
ViLLA _{base}	30.00	12.00	8.00
VisualBERT _{base}	15.50	2.50	1.50
ViLT (ViT-B/32)	34.75	14.00	9.25
LXMERT	19.25	7.00	4.00
ViLBERT _{base}	23.75	7.25	4.75
UniT _{ITM finetuned}	19.50	6.25	4.00
CLIP (ViT-B/32)	30.75	10.50	8.00
VSE++ _{COCO} (ResNet)	22.75	8.00	4.00
VSE++ _{COCO} (VGG)	18.75	5.50	3.50
VSE++ _{Flickr30k} (ResNet)	20.00	5.00	2.75
VSE++ _{Flickr30k} (VGG)	19.75	6.25	4.50
VSRN _{COCO}	17.50	7.00	3.75
VSRN _{Flickr30k}	20.00	5.00	3.50

STILL FAR FROM SOLVED.



Evan Morikawa
@EOM

...

DALL-E2 on Winoground I

DALL-E

Edit the detailed description

there is a mug in some grass, digital art

Surprise me Upload

Report issue ↗

This section shows a 2x5 grid of generated images. The first row contains five images of mugs (red, orange, blue, white, grey) placed in dense green grass. The second row contains five images of mugs (blue, white, brown, beige, yellow) also placed in grass, with some variations in lighting and shadowing.

More Winoground prompts. 1st run, no cherry-picking.
To all I added "digital art". That helps w/ the
composition (and aesthetic imo), particularly for less
common things

DALL-E

Edit the detailed description

there is some grass in a mug, digital art

Surprise me Upload

Report issue ↗

This section shows a 2x5 grid of generated images. The first row contains five images of mugs (white, grey, white, white, pink) with tufts of green grass growing out of their tops. The second row contains five images of mugs (white, grey, white, dark blue, teal) with similar grass growth, demonstrating a more specific and creative interpretation of the prompt.



Evan Morikawa
@EOM

...

DALL-E2 on Winoground II

DALL-E

Edit the detailed description

there are fewer forks than spoons, digital art

Surprise me Upload

Report issue

The interface shows a header with the DALL-E logo, a user profile, and three dots. Below is a text input field with the prompt. To the right are 'Surprise me' and 'Upload' buttons, and at the bottom right is a 'Report issue' link.

More Winoground prompts. 1st run, no cherry-picking.
To all I added "digital art". That helps w/ the
composition (and aesthetic imo), particularly for less
common things

DALL-E

Edit the detailed description

there are fewer spoons than forks, digital art

Surprise me Upload

Report issue

The interface is identical to the first one, with a header, user profile, and three dots. It includes a text input field with the prompt, 'Surprise me' and 'Upload' buttons, and a 'Report issue' link at the bottom right.

STILL NOT SOLVED

Outline

1. Early models
2. Features and fusion
3. Contrastive models
4. Multimodal foundation models
5. Evaluation
6. **Beyond images: Other modalities**
7. Where to next?

Speech / audio

Can EASILY do another full lecture just on this topic.

Recent cool example: Whisper, trained on 680,000 hours of multilingual multitask data.

We can also just treat audio as vision ;)

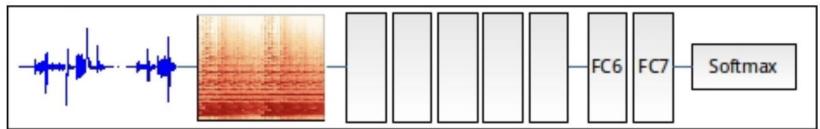
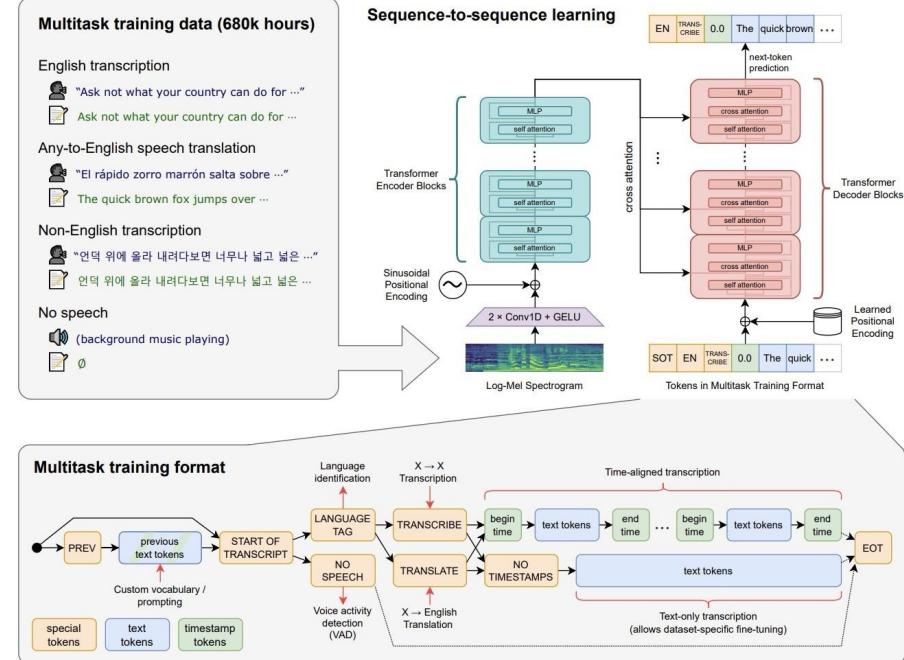
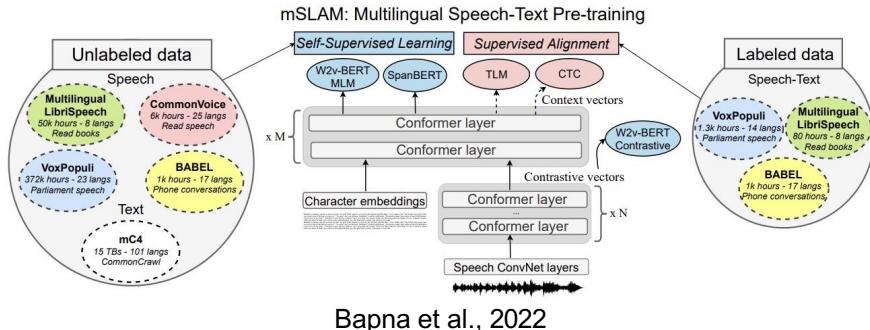


Figure 3: Illustration of the Neural Auditory Embedding method, using a convolutional neural network. The auditory signal is converted to a spectrogram which is fed to the neural network for classification. The pre-softmax layer, FC7, is transferred and taken as the neural audio embedding (NAE) for the given sound file.

Kiela et al., 2017



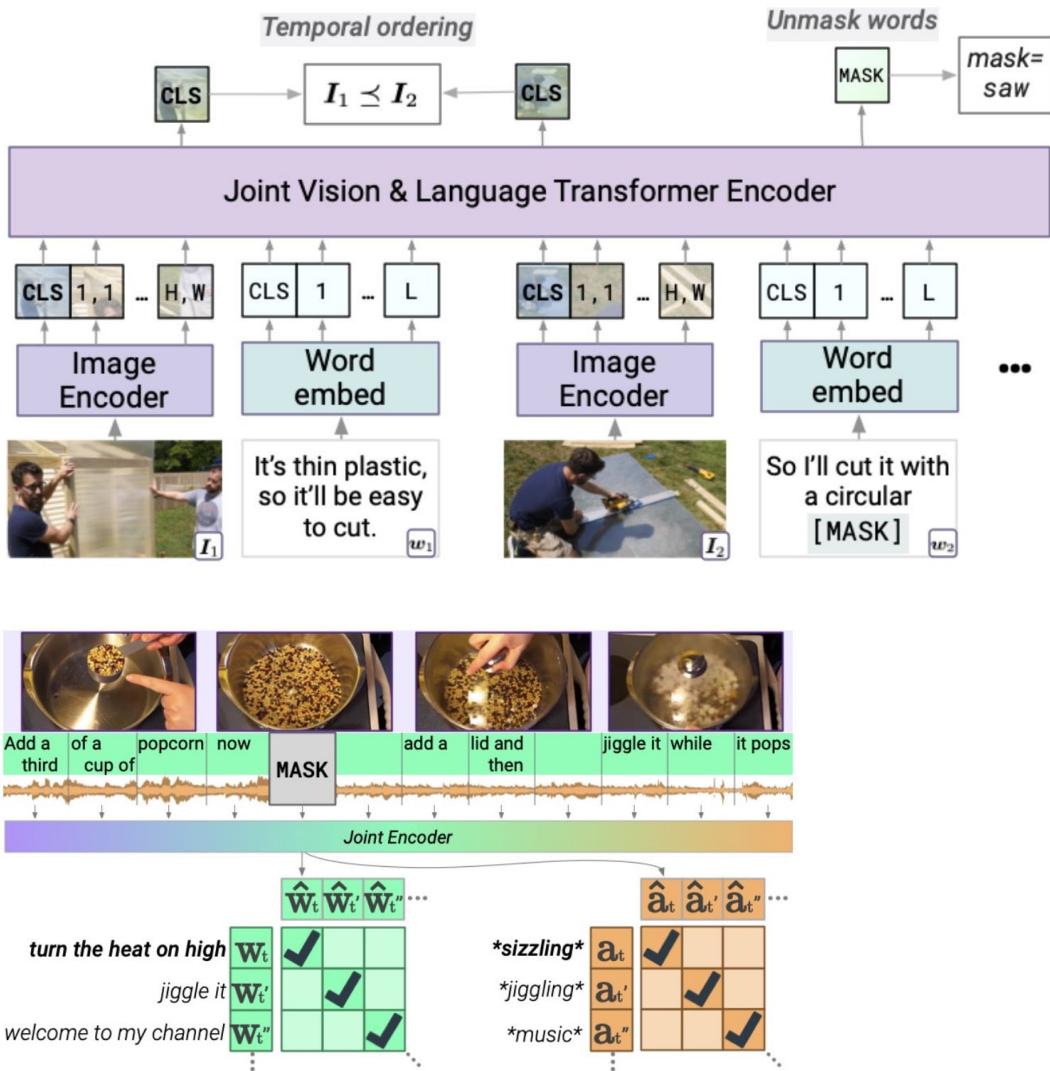
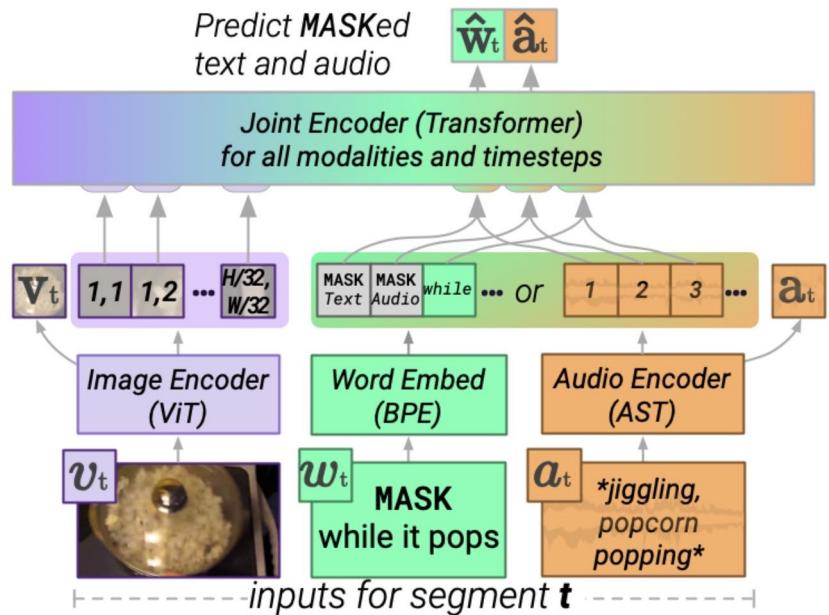
Radford et al., 2022



Bapna et al., 2022

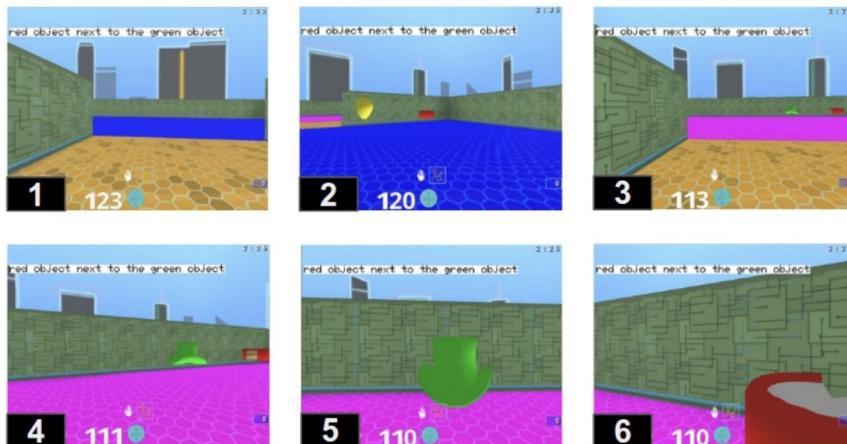
Video and text and audio

MERLOT (Zellers et al., 2021)
 MERLOT Reserve (idem, 2022)

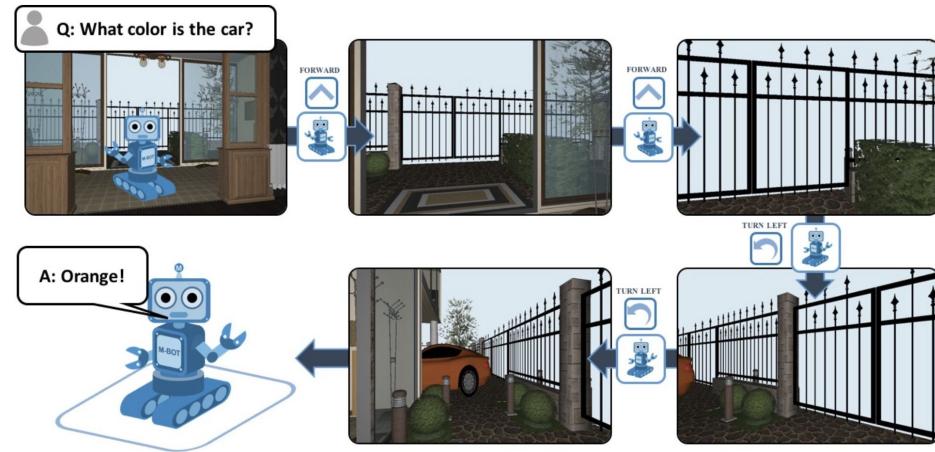


Grounded language learning in simulated environments

Agent view



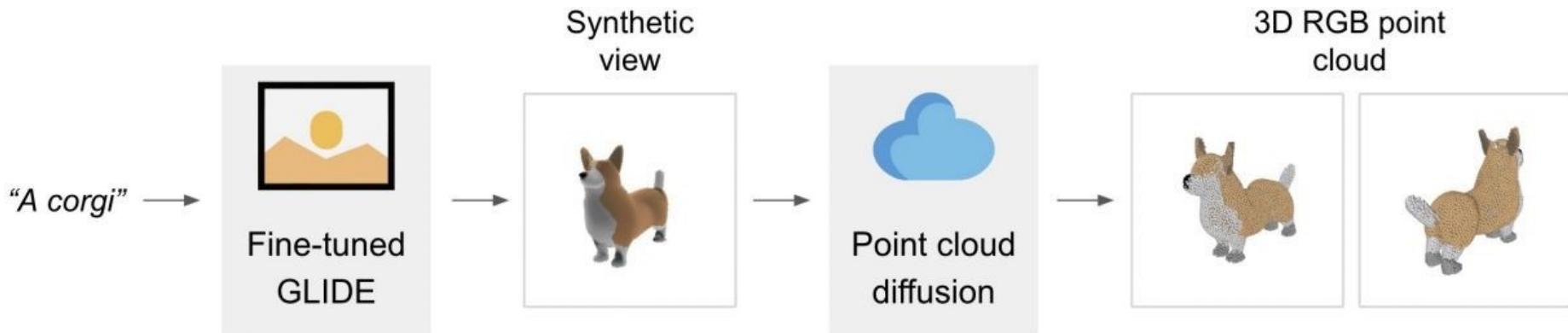
Hermann, Hill, et al. 2017



Das et al., 2018

Text to 3D

POINT-E (Nichol, Jun, et al., 2022)



Outline

1. Early models
2. Multimodal fusion
3. Contrastive models
4. Multimodal foundation models
5. Other modalities
6. Evaluation
7. **Where to next?**

One foundation model to rule them all

There will modality-agnostic foundation models that can read and generate many modalities.

These models can be bigger and be trained on vastly more data. Parameters will be shared in interesting ways.

Automatic alignment from unpaired unimodal data will become a big topic.

Multimodal scaling laws

We are just beginning to understand multimodal scaling laws, lots of interesting work to do here in understanding trade-offs.

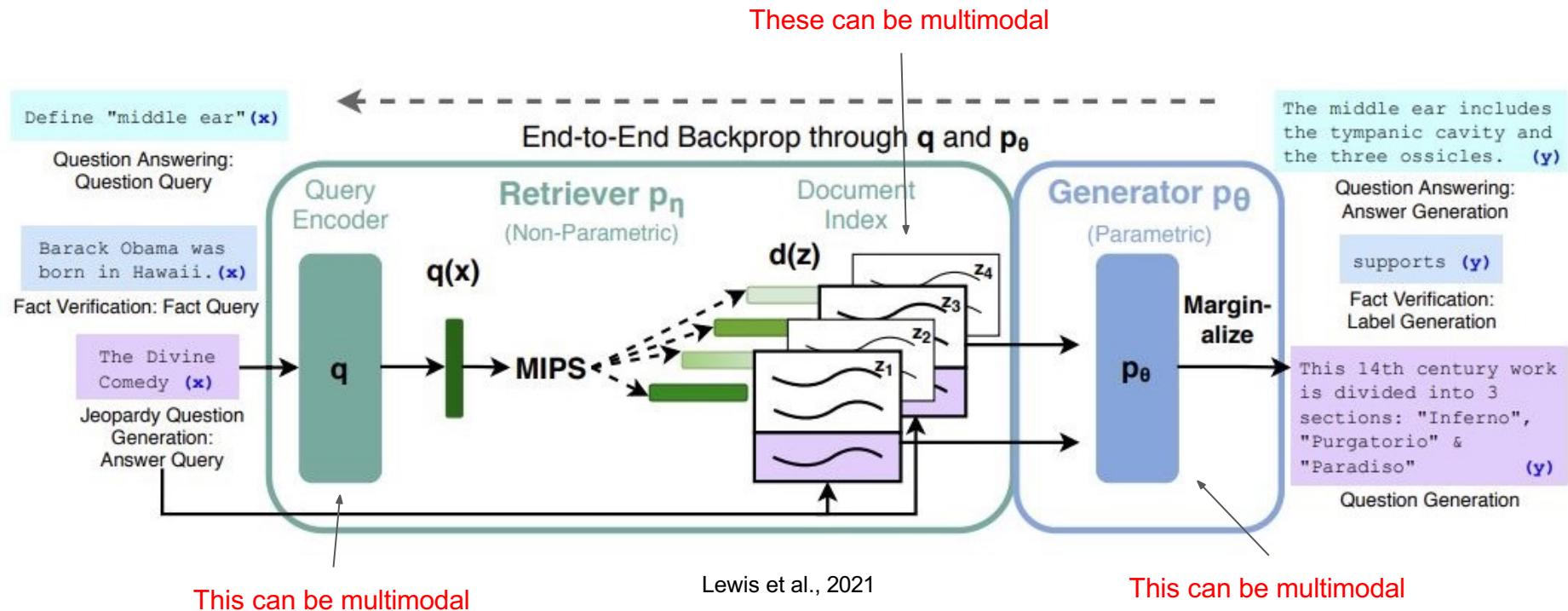
SCALING LAWS FOR GENERATIVE MIXED-MODAL LANGUAGE MODELS

Armen Aghajanyan^{*†}, Lili Yu^{*†}, Alexis Conneau[†], Wei-Ning Hsu[†]

Karen Hambardzumyan[◊], Susan Zhang[†], Stephen Roller[†], Naman Goyal[†]

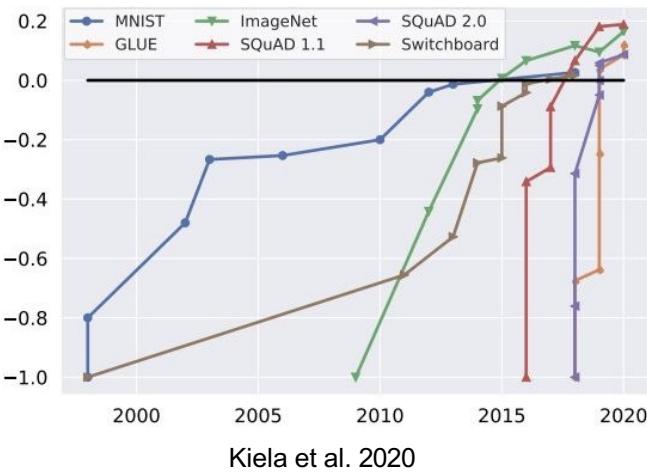
Omer Levy[†] & Luke Zettlemoyer^{†,♡}

Retrieval augmented generative multimodal models

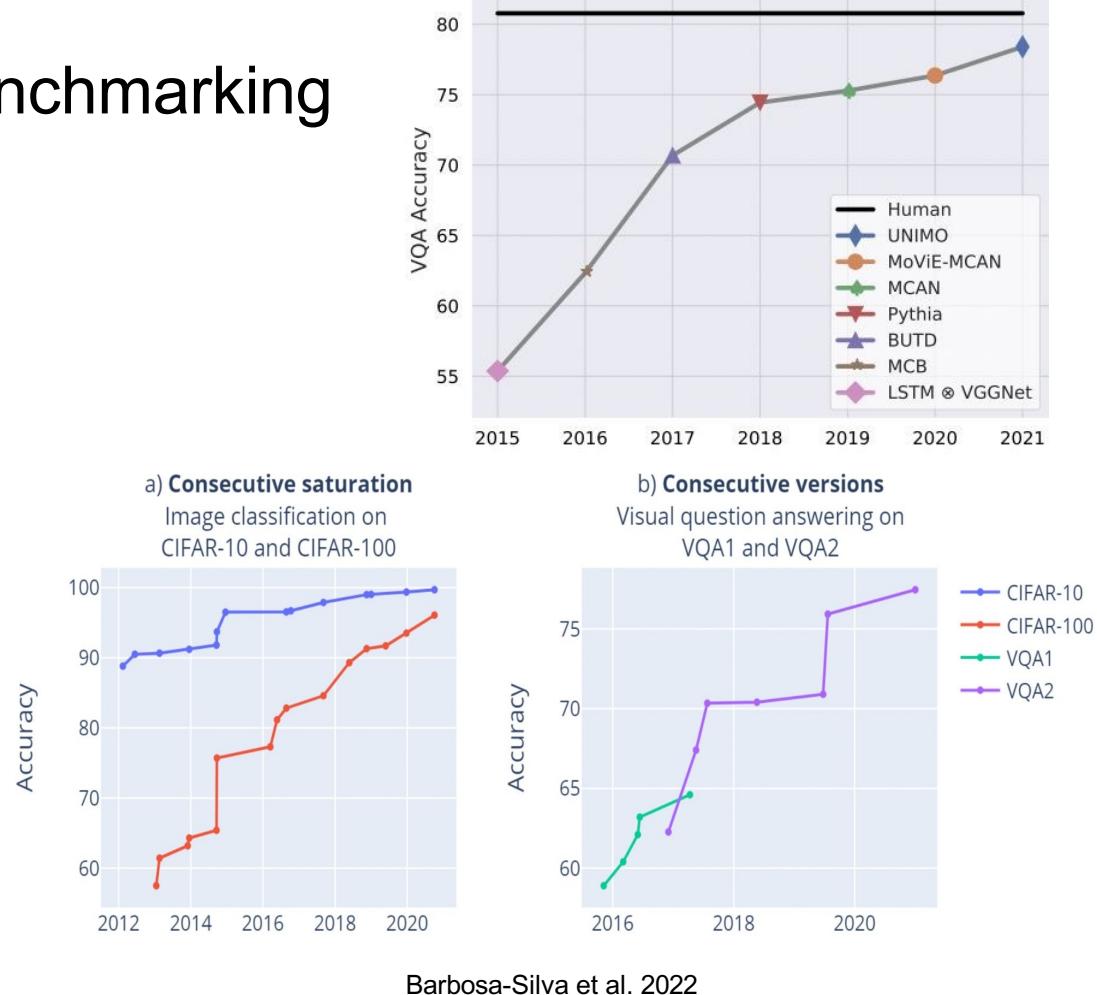


Better evaluation and benchmarking

We need better measurement.



Kiela et al. 2020



Thank you!