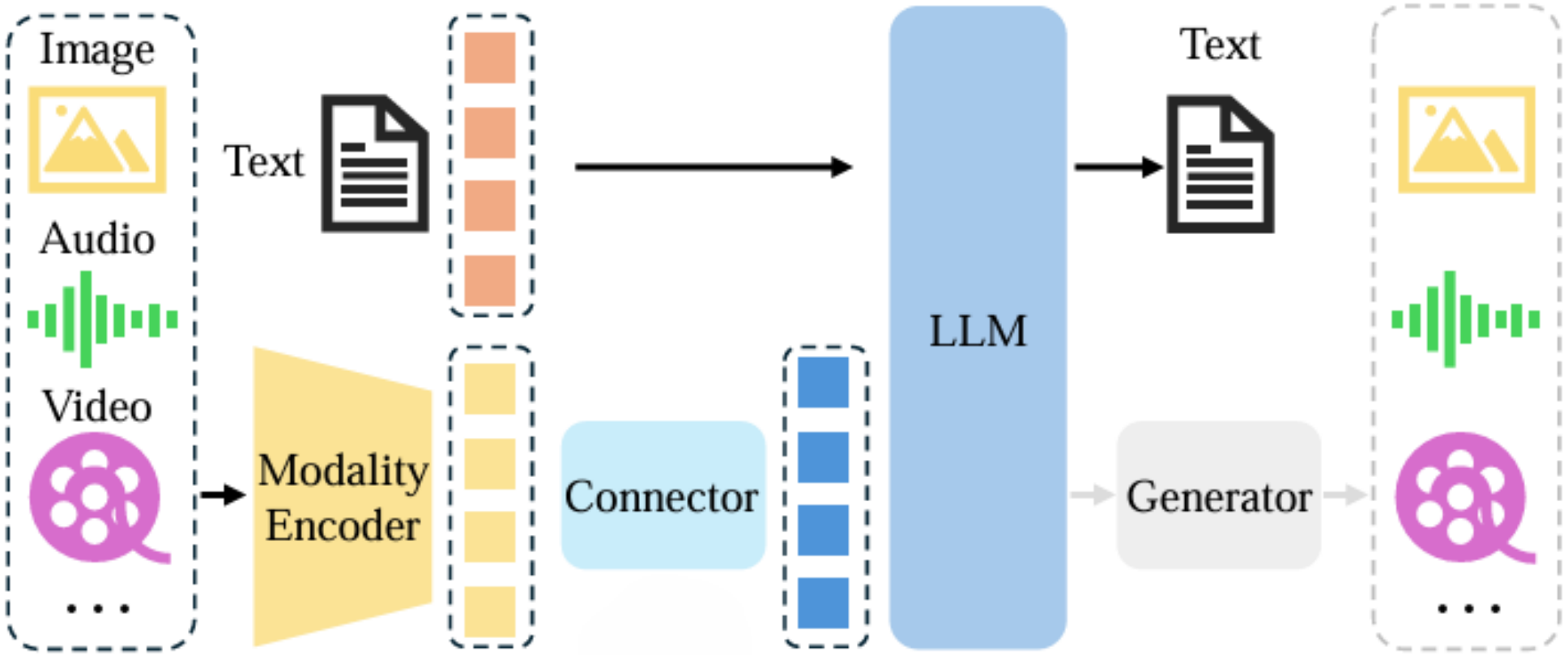
M-LLMs

On the way to AGI



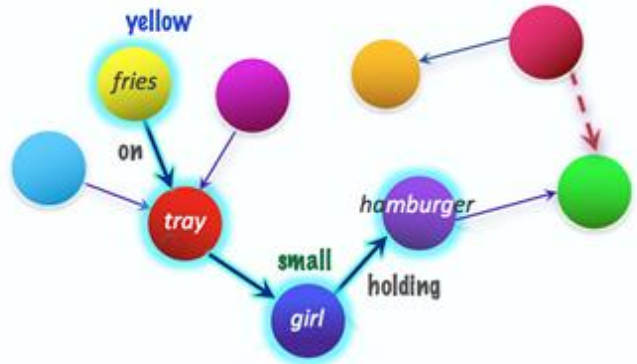
Source of image: A Survey on Multimodal Large Language Models (2023) https://arxiv.org/abs/2306.13549

Static Benchmarks: GQA

Pattern: What|Which <type> [do you think] <is> <dobject>, <attr> or <decoy>?
Program: **Select:** <dobject> → **Choose** <type>: <attr>|<decoy>
Reference: The food on the red object left of the small girl that is holding a hamburger
Decoy: brown

What color is the food on the red object left of the small girl that is holding a hamburger, yellow or brown?

Select: hamburger → **Relate:** girl,holding → **Filter** size: small → **Relate:** object, left → **Filter** color: red → **Relate:** food,on → **Choose** color: yellow | brown



Graph Normalization

- Ontology construction
- Edge Pruning
- Object Augmentation
- Global Properties

Question Generation

- Patterns Collection
- Compositional References
- Decoys Selection
- Probabilistic Generation

Sampling and Balancing

- Distribution Balancing
- Type-Based Sampling
- Deduplication

Entailments Relations

- Functional Programs
- Entailment Relations
- Recursive Reachability

New Metrics

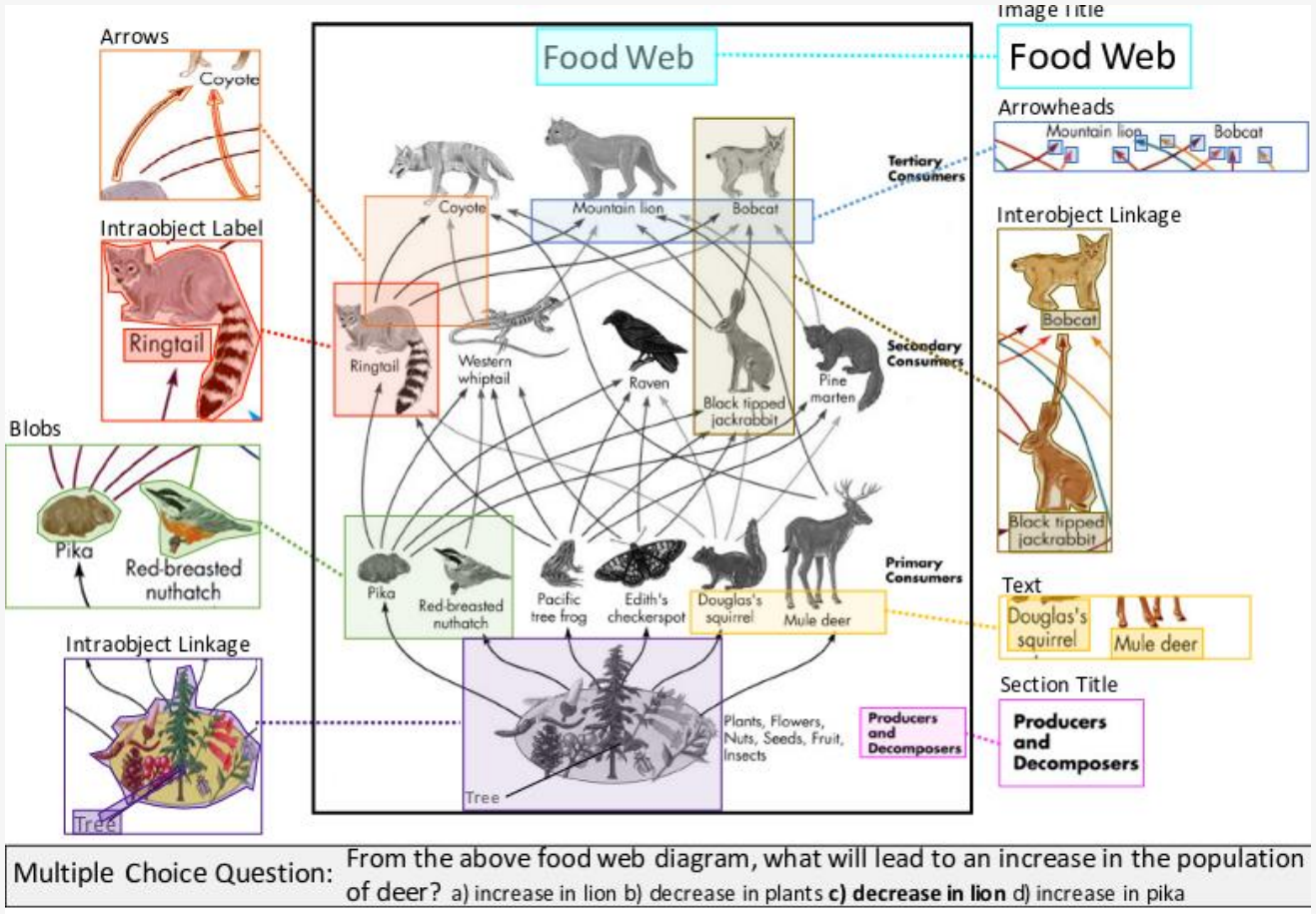
- Consistency
- Validity & Plausibility
- Distribution
- Grounding

GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering (2019) https://arxiv.org/abs/1902.09506

A1. Is the **tray** on top of the **table** black or light brown? light brown
A2. Are the **napkin** and the **cup** the same color? yes
A3. Is the small **table** both oval and wooden? yes
A4. Is there any **fruit** to the left of the **tray** the **cup** is on top of? yes
A5. Are there any **cups** to the left of the **tray** on top of the **table**? no
B1. What is the brown **animal** sitting inside of? **box**
B2. What is the large **container** made of? cardboard
B3. What **animal** is in the **box**? **bear**
B4. Is there a **bag** to the right of the green **door**? no
B5. Is there a **box** inside the plastic **bag**? no

Static Benchmarks: GQA

- Questions are generated using a graph of images
- 22.6M questions for 113K images
- Evaluation metrics: accuracy and 5 more



A Diagram Is Worth A Dozen Images (2016) https://arxiv.org/abs/1603.07396

AI2D

- 15K multiple choice questions for 5K school level diagrams
- Parsing graph is available
- Evaluation metric: accuracy

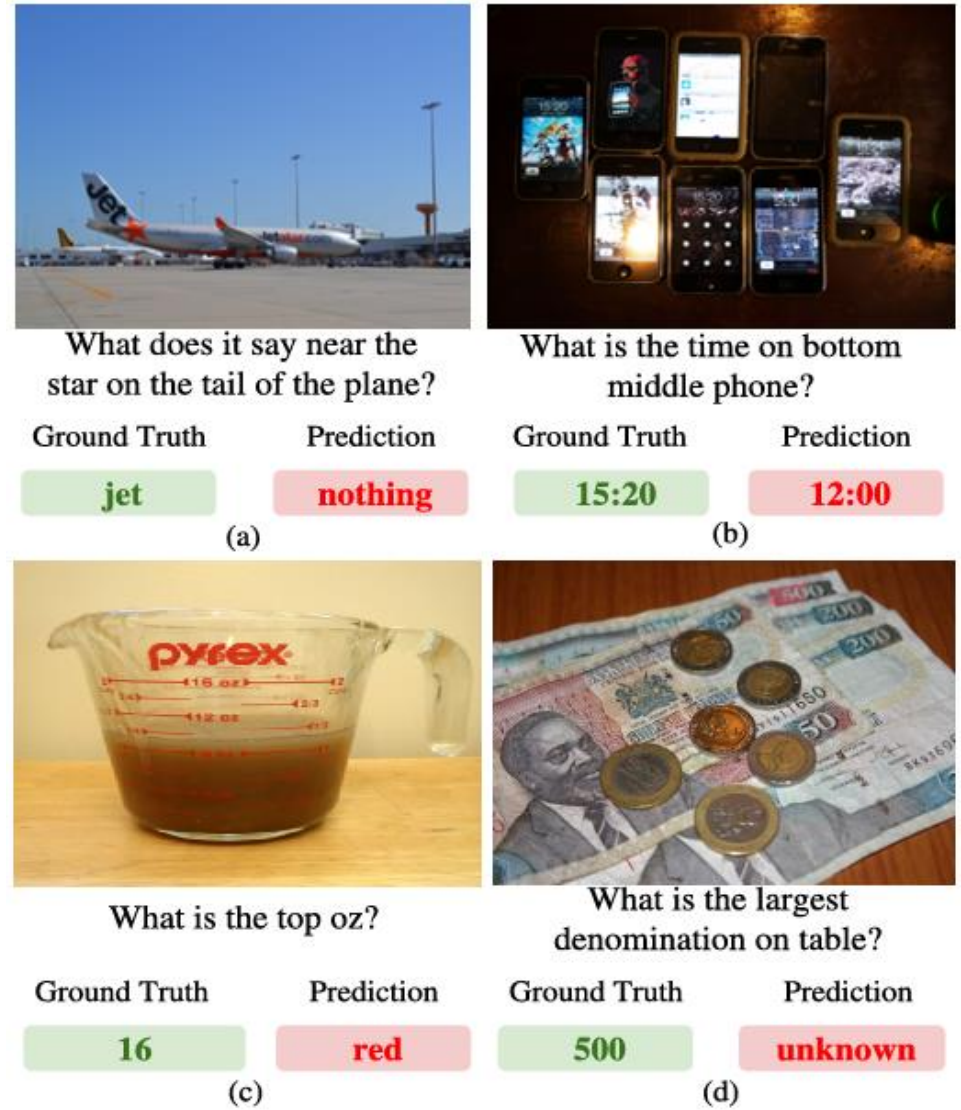
MMMU

- 11.5K questions from 6 university disciplines
- Answers were extracted using regexps
- Evaluation metric: accuracy

Art & Design	Business	Science
Question: Among the following harmonic intervals, which one is constructed incorrectly? **Options:** (A) Major third <*image 1*> (B) Diminished fifth <*image 2*> **(C) Minor seventh <*image 3*>** (D) Diminished sixth <*image 4*>	**Question:** ...The graph shown is compiled from data collected by Gallup <*image 1*>. Find the probability that the selected Emotional Health Index Score is between 80.5 and 82? **Options:** (A) 0 (B) 0.2142 **(C) 0.3571** (D) 0.5	**Question:** <*image 1*> The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R. **Options:** **(A)** $\int_0^{1.5}[f(x)-g(x)]dx$ (B) $\int_0^{1.5}[g(x)-f(x)]dx$ (C) $\int_0^{2}[f(x)-g(x)]dx$ (D) $\int_0^{2}[g(x)-x(x)]dx$
Subject: Music; **Subfield:** Music; **Image Type:** Sheet Music; **Difficulty:** Medium	**Subject:** Marketing; **Subfield:** Market Research; **Image Type:** Plots and Charts; **Difficulty:** Medium	**Subject:** Math; **Subfield:** Calculus; **Image Type:** Mathematical Notations; **Difficulty:** Easy
Health & Medicine	**Humanities & Social Science**	**Tech & Engineering**
Question: You are shown subtraction <*image 1*>, T2 weighted <*image 2*> and T1 weighted axial <*image 3*> from a screening breast MRI. What is the etiology of the finding in the left breast? **Options:** (A) Susceptibility artifact (B) Hematoma **(C) Fat necrosis** (D) Silicone granuloma	**Question:** In the political cartoon, the United States is seen as fulfilling which of the following roles? <*image 1*> **Option:** (A) Oppressor (B) Imperialist **(C) Savior** (D) Isolationist	**Question:** Find the VCE for the circuit shown in <*image 1*>. Neglect VBE **Answer:** 3.75 **Explanation:** ...IE = [(VEE) / (RE)] = [(5 V) / (4 k-ohm)] = 1.25 mA; VCE = VCC - IERL = 10 V - (1.25 mA) 5 k-ohm; VCE = 10 V - 6.25 V = 3.75 V
Subject: Clinical Medicine; **Subfield:** Clinical Radiology; **Image Type:** Body Scans: MRI, CT.; **Difficulty:** Hard	**Subject:** History; **Subfield:** Modern History; **Image Type:** Comics and Cartoons; **Difficulty:** Easy	**Subject:** Electronics; **Subfield:** Analog electronics; **Image Type:** Diagrams; **Difficulty:** Hard

MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI (2023)

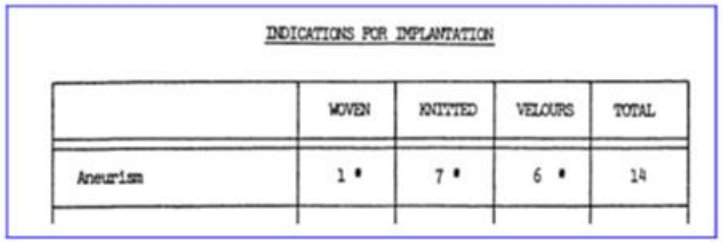
https://arxiv.org/abs/2311.16502



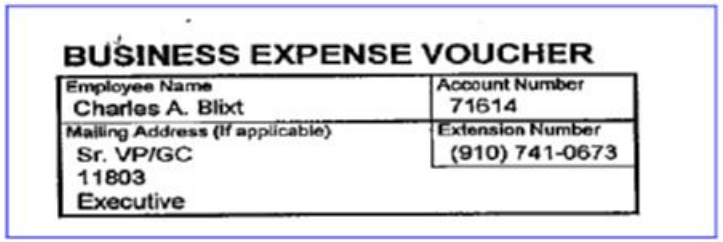
Towards VQA Models That Can Read (2019)
https://arxiv.org/abs/1904.08920

TextVQA

- 45K questions for 26K images
- 10 answers per question
- Evaluation metric: CQA accuracy (100% in case 3 people provided the same answer)



Q: What is the underlined heading just above the table?
GT: Indications for implantation
M4C best: indications for implantation
BERT best: total aneurism
Human: indications for implantation



Q: What is the Extension Number as per the voucher?
GT: (910) 741-0673
M4C best: 963.12
BERT best: (910) 741-0673
Human: (910) 741-0673

Q: How many boxed illustrations are there ?
GT: 9
M4C best: 4
BERT best: 4
Human: 9

DocVQA

- 50K questions for 12K images
- Documents from 1960-2000
- Evaluation metrics: Average Normalized Levenstein Similarity, Accuracy

DocVQA: A Dataset for VQA on Document Images (2021) https://arxiv.org/abs/2007.00398

Image

Model A

Given my horse's location on this map, what is the quickest route to reach it?

Based on the map, your current location is marked by the pin/stamp icon at Lookout Landing. The quickest route to reach your horse, which is marked with a horse icon at Rito Village, would be to head northwest along the roads/paths shown on the map. It looks like you'll need to traverse the area labeled "Tabantha Frontier" to get from your current spot to Rito Village where your horse is located. The exact path is a bit unclear from this zoomed out view, but generally heading northwest from Lookout Landing should get you to your horse's position at Rito Village by the most direct route possible given the geography shown.

Model B

Given my horse's location on this map, what is the quickest route to reach it?

I'm sorry, but I don't have access to your specific map or location information. Could you please provide more context or details so I can assist you better?

Rules

- Upload any image, and send any question to two anonymous models and vote for the better one!
- Or you could also click "Sample Input" to get a random example from public benchmarks such as VisIT-Bench.
- Single image multi-round chat is allowed, you can continue to send question until you identify a winner.

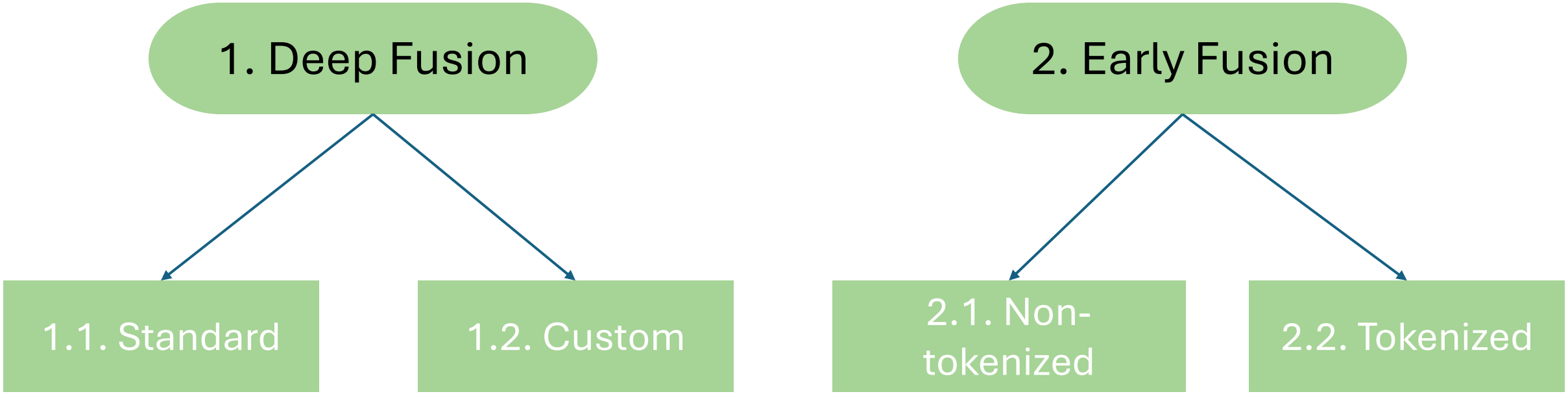
WildVision: Evaluating Vision-Language Models in the Wild with Human Preferences (2024)
https://arxiv.org/abs/2406.11069

Reason for your vote

Please input your reason for response preference here before clicking the model choice button.

A is better | B is better | Tie | Both are bad

High-Level Classification

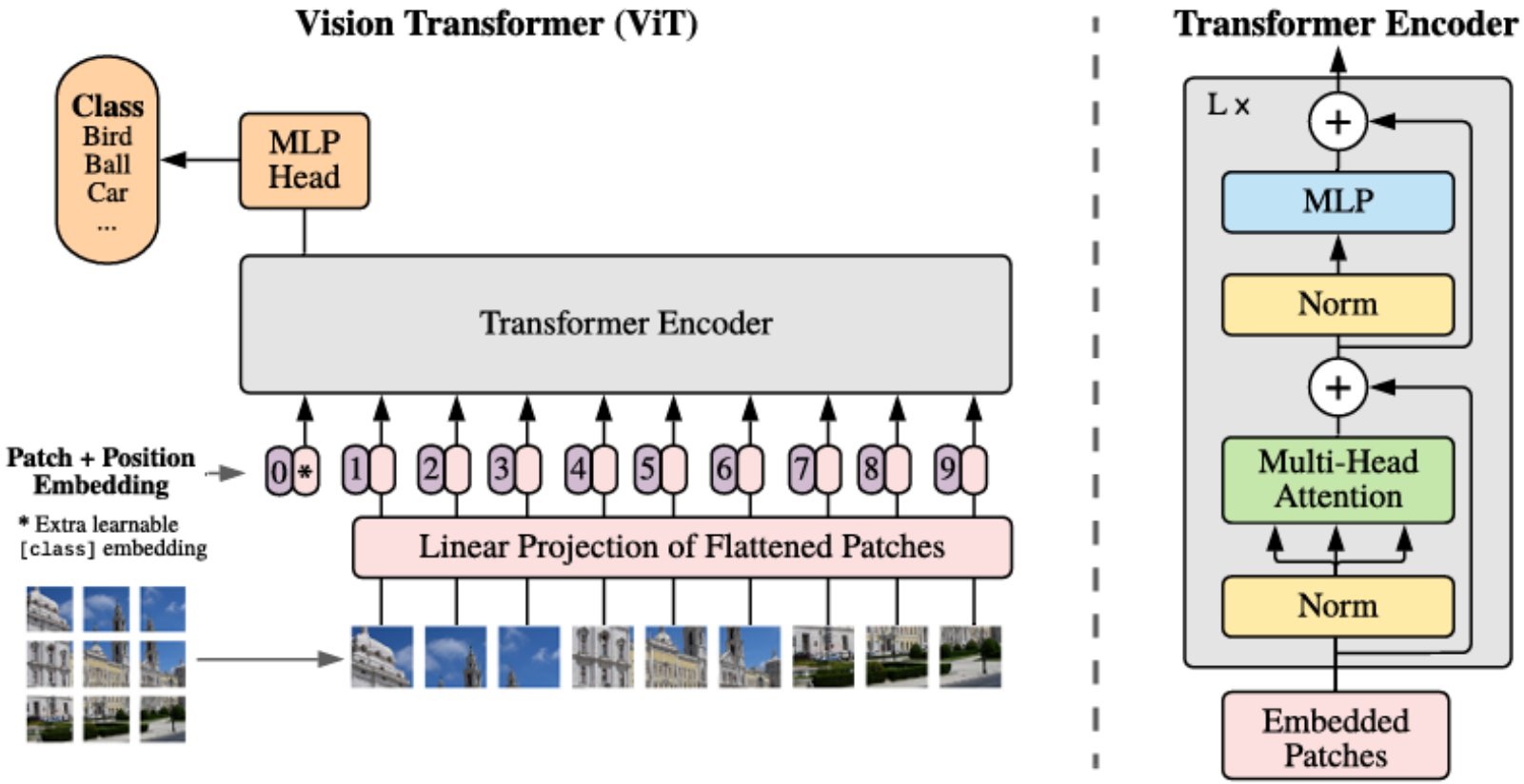


Cross-Attention
- Flamingo

Layers
- LLaVA
- miniGPT-4
- Gemini

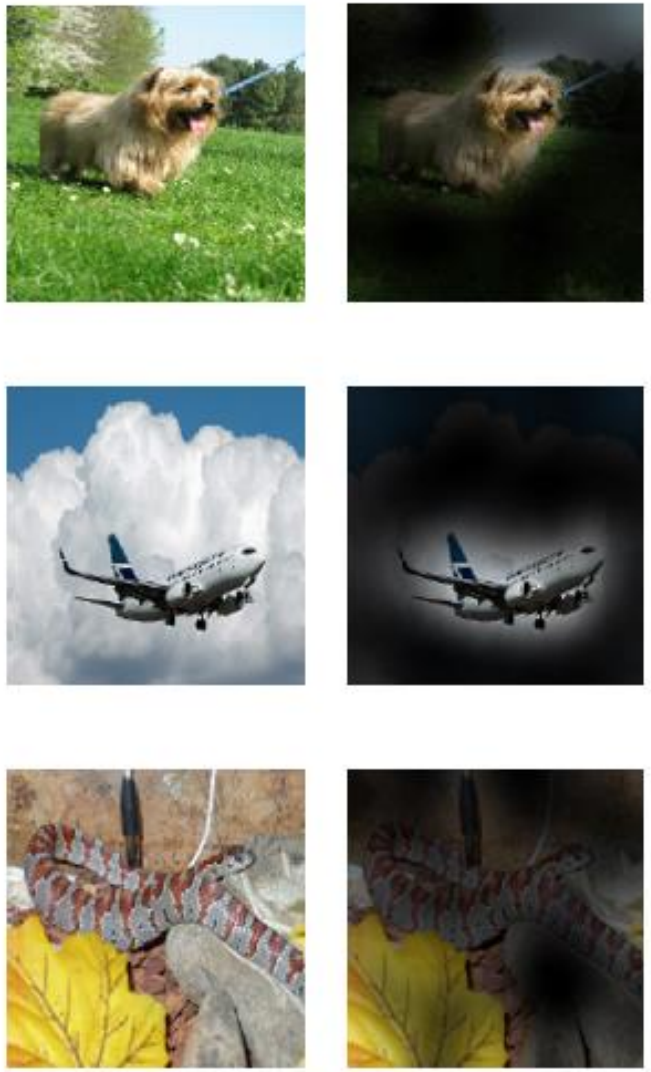
- BLIP

- Kosmos-1
- Fuyu



An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2021) https://arxiv.org/abs/2010.11929

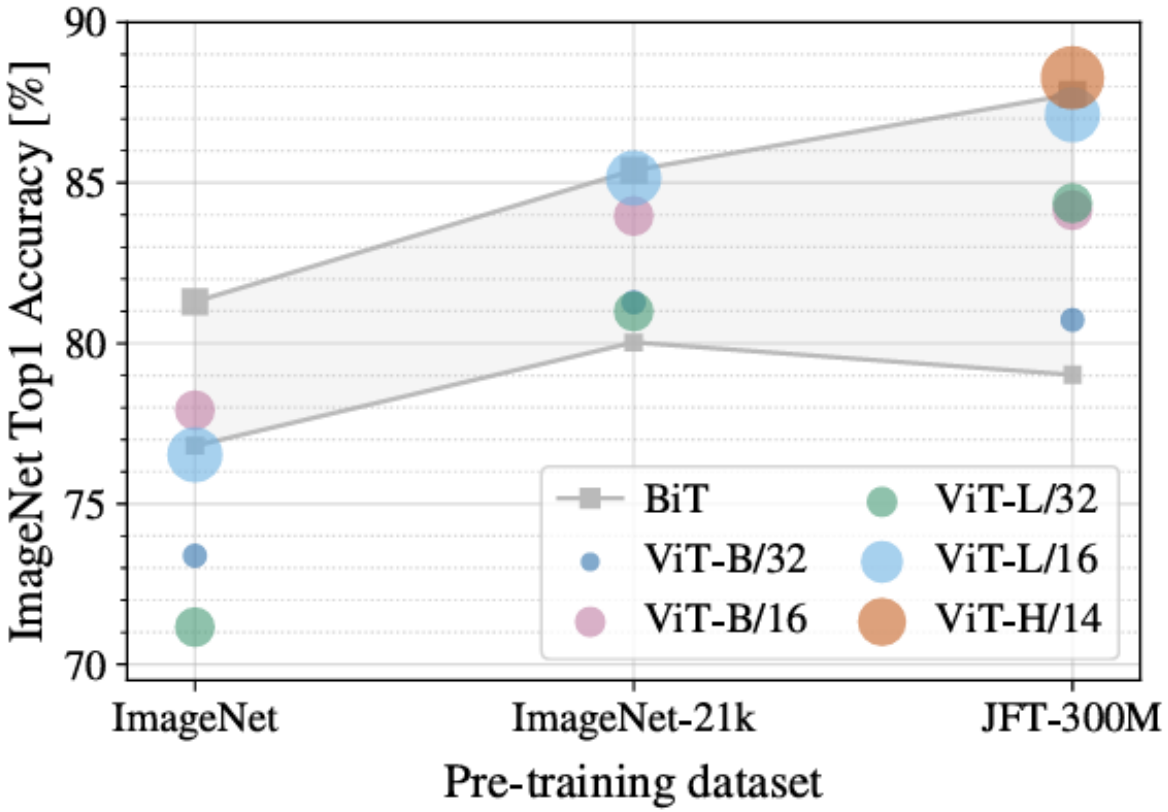
Input Attention

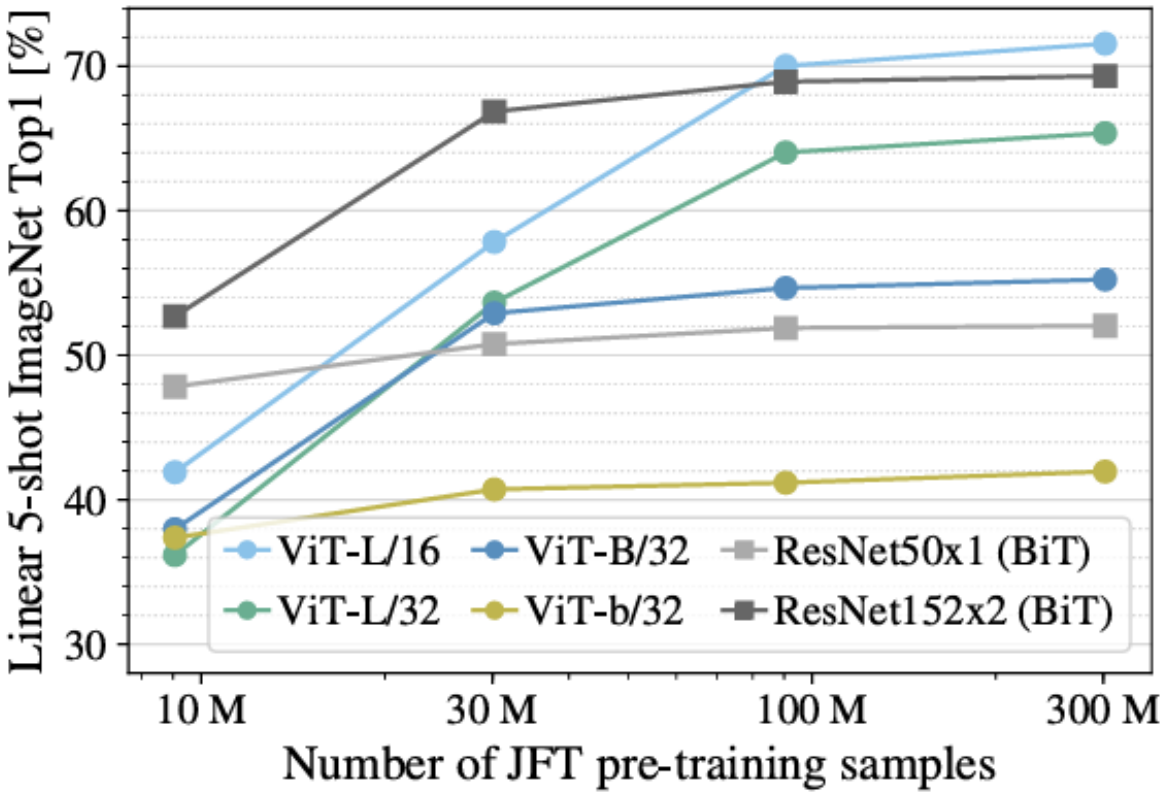


MSA visualization, 8×8 patches

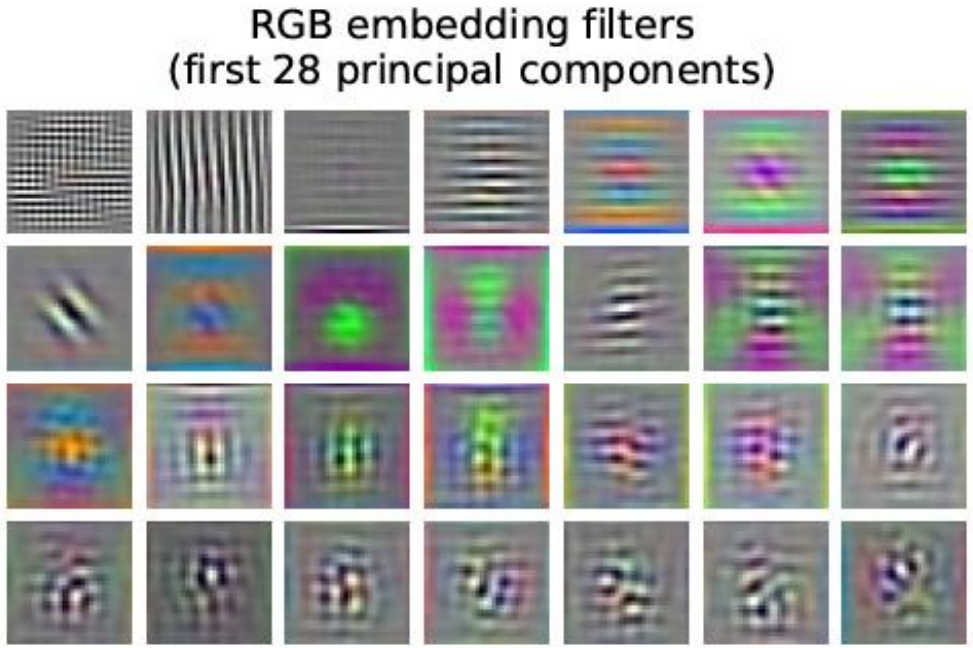
www.youtube.com/@cv-gml Lecture from 10/15/2024

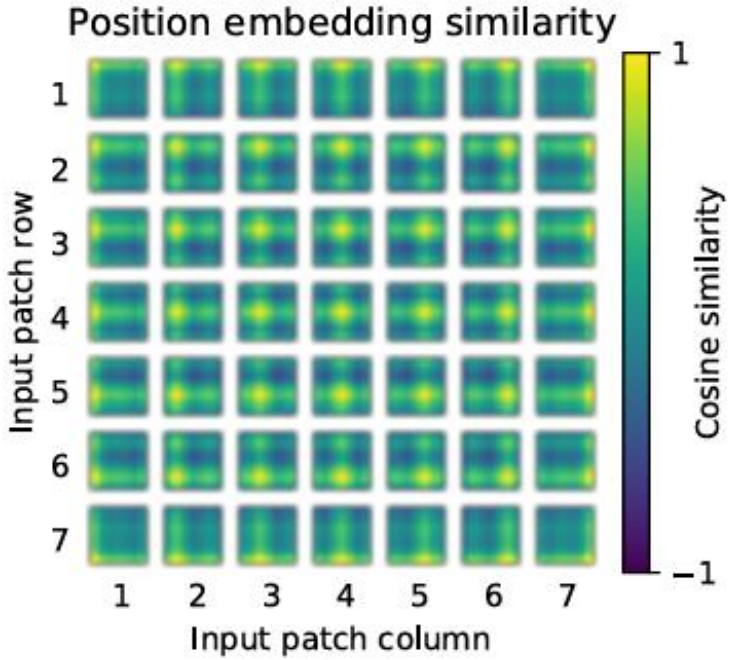
Vision Transformer

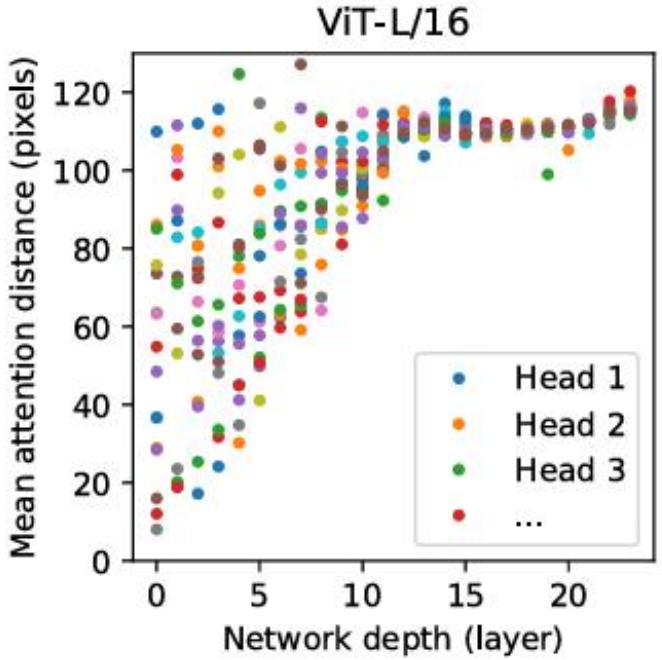




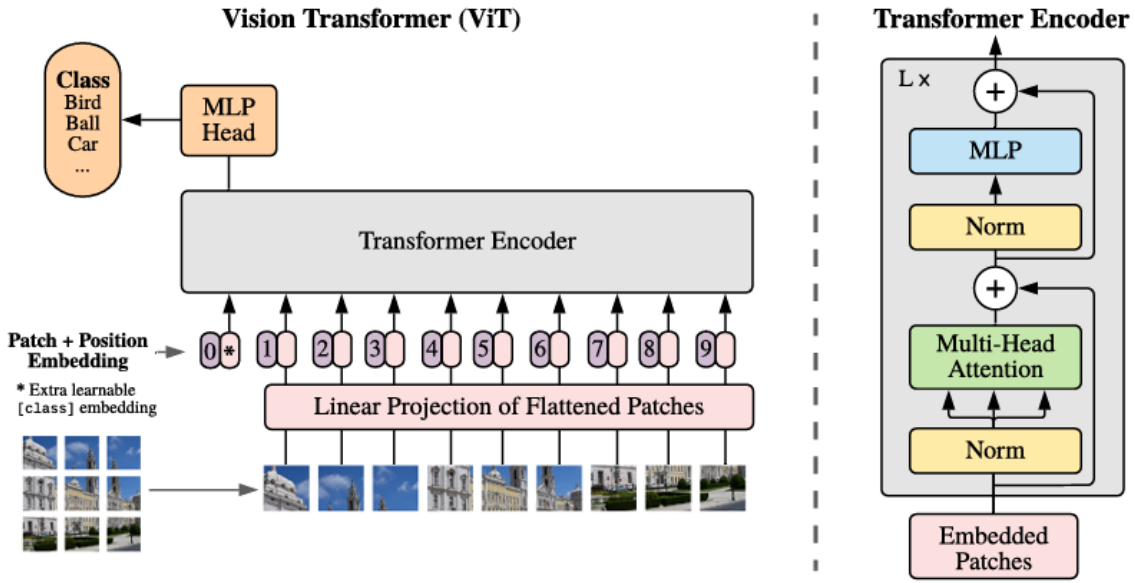
Vision Transformer

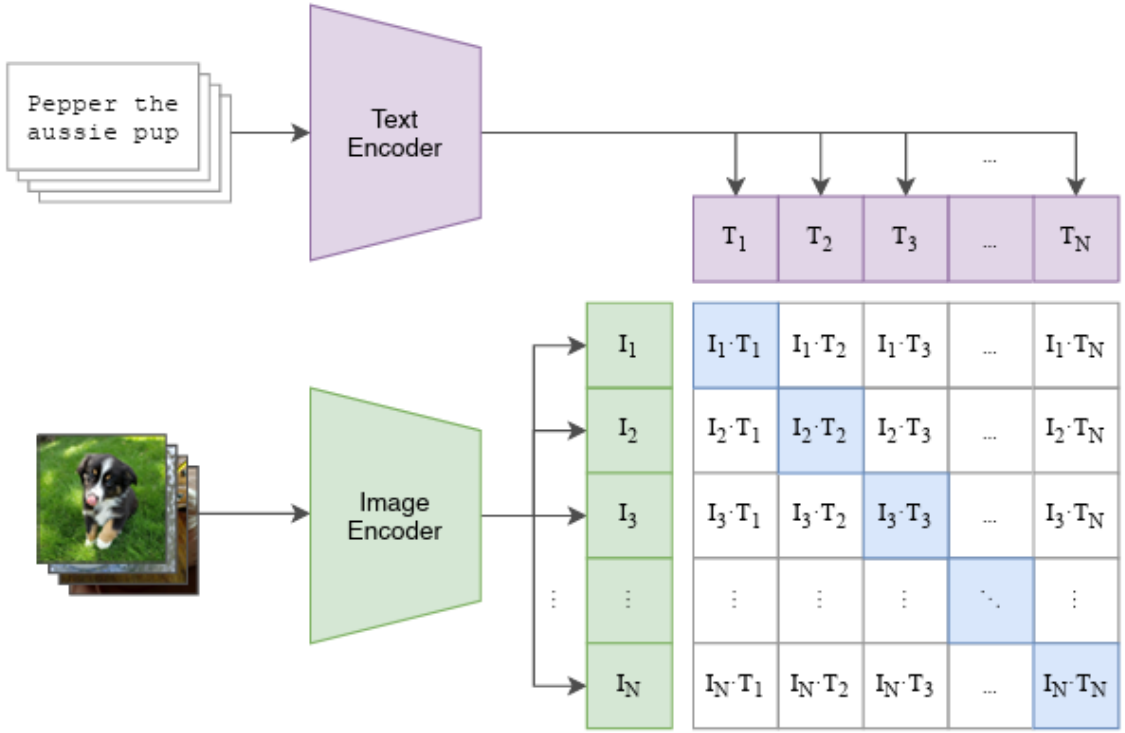






CLIP



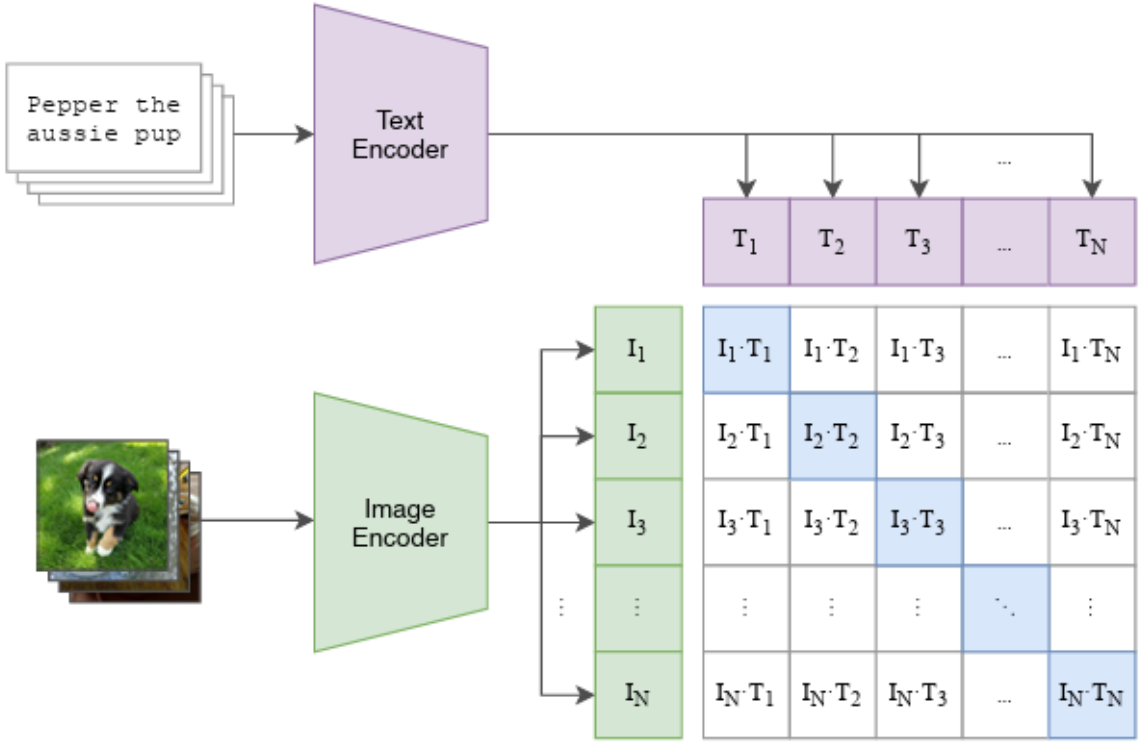


$$-\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_i\cdot\mathbf{y}_j}}}^{\text{image}\to\text{text softmax}}+\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_j\cdot\mathbf{y}_i}}}^{\text{text}\to\text{image softmax}}\right)$$

400M (image, text), 500 V100 GPU,

Learning Transferable Visual Models From Natural Language Supervision (2021) https://arxiv.org/abs/2103.00020

SigLIP



$$-\frac{1}{2|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_i\cdot\mathbf{y}_j}}}^{\text{image}\to\text{text softmax}}+\overbrace{\log\frac{e^{t\mathbf{x}_i\cdot\mathbf{y}_i}}{\sum_{j=1}^{|\mathcal{B}|}e^{t\mathbf{x}_j\cdot\mathbf{y}_i}}}^{\text{text}\to\text{image softmax}}\right)$$

$$-\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\frac{1}{1+e^{z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b)}}}_{\mathcal{L}_{ij}}$$

Sigmoid Loss for Language Image Pre-Training (2023) https://arxiv.org/abs/2303.15343