When Vision meets Language: A Glance of Multimodal Learning

19/10/2023

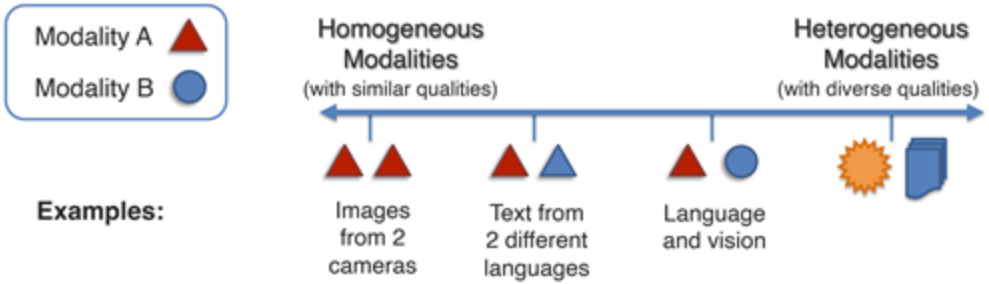
Agenda

- Introduction of Multimodal Learning
- Downstream tasks
- Vision-Language Pretraining
- Advanced Vision-Language Models
- Evaluation
- Future Challenges
- Takeaways

What is multimodal learning?

Multimodal is the scientific study of **heterogeneous** and **interconnected** data (Liang and Morency, 2023).

- Heterogeneous: Diverse qualities, structures and representations

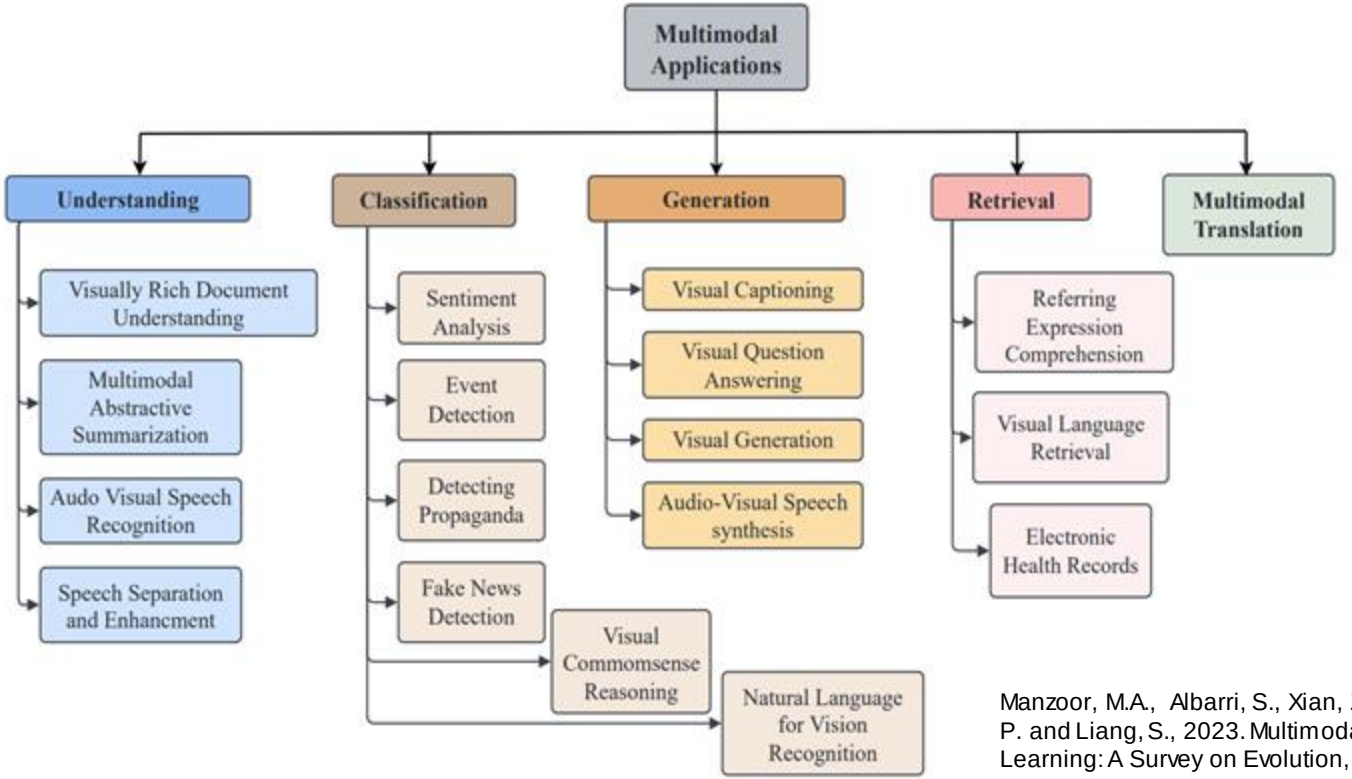


- Interconnected: Connected + Interacting
  - Connected: shared information that relates modalities
  - Interacting: process affecting each modality, creating new information

Why is multimodal learning hard?

- **Representation**: Learning representations that reflect cross-modal interactions across different modalities
- **Alignment**: Identifying and modelling cross-modal connections between all elements of multiple modalities, building from the data structure
- **Reasoning**: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure
- **Generation**: Learning a generative process to produce raw modalities that reflects cross-modal interactions, structure and coherence
- **Transference**: Transfer knowledge between modalities, usually to help the target modality which may be noisy or with limited resources
- **Quantification**: Empirical and theoretical study to better understand heterogeneity, cross-modal interactions and the multimodal learning process

Multimodal downstream tasks



Manzoor, M.A., Albarri, S., Xian, Z., Meng, Z., Nakov, P. and Liang, S., 2023. Multimodality Representation Learning: A Survey on Evolution, Pretraining and Its Applications. arXiv preprint arXiv:2302.00389.

Vision-Language Pretraining (VLP)

Large-scale transformer-based self-supervised pre-training

- Reuse the same pre-training weight as initialization point
- Separate output head and finetune model copies for different downstream tasks
- (relatively) Low data collection costs: automatically curated from Web
- State-of-the-art performance in many tasks across domains

Evolution of Generalized VLP Models

Specialized models are designed for different tasks

Pretraining on high-quality labeled multimodal datasets

VisualBERT
LXMert
ViLBERT
UNITER
OSCAR

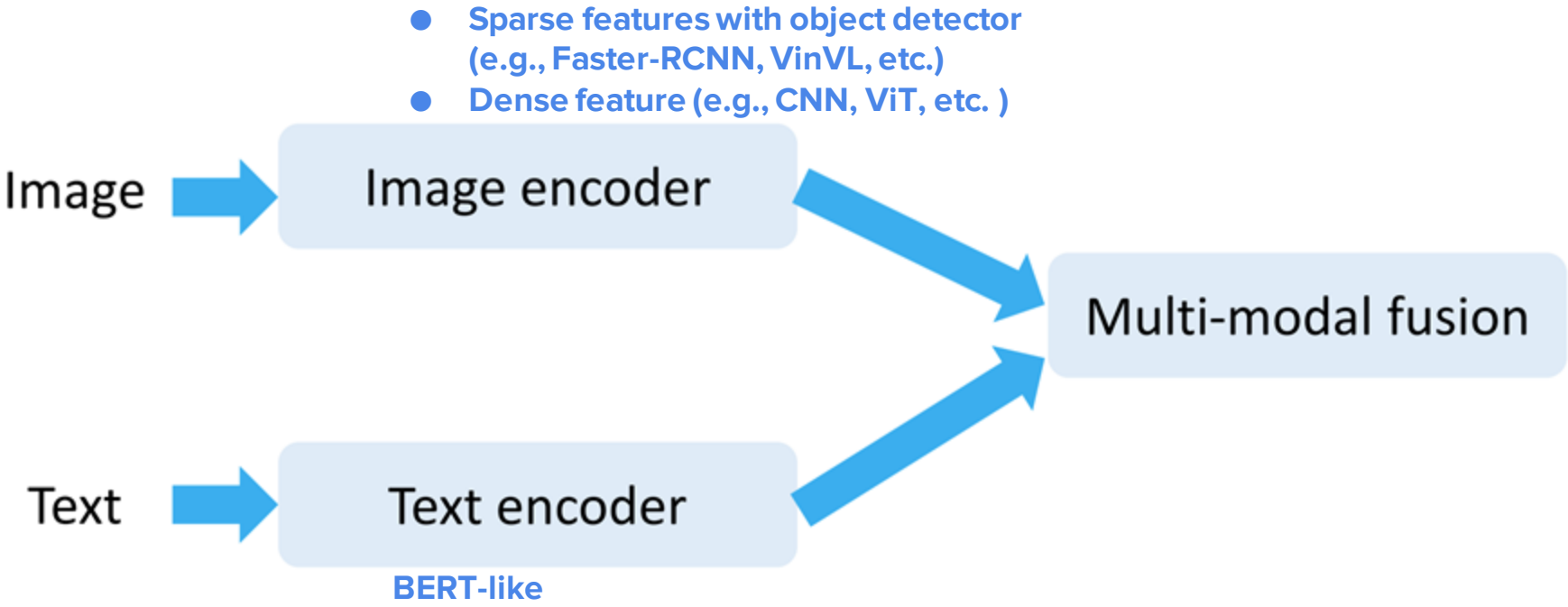
Pretraining on extremely large-scale weakly-labeled (noisy) datasets -> more general few-shot/zero-shot models

CLIP
ALIGN
Florence
FLAVA
Flamingo

Fused with LLMs and more modalities, it enables more efficient training and provides stronger cross-domain/zero-shot capabilities

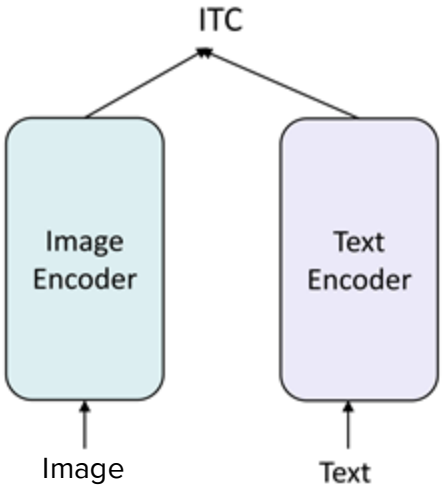
ImageBind
Multimodal-CoT
BLIP2
Prismer

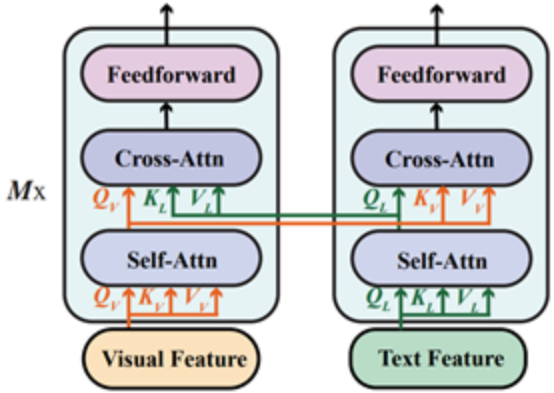
VLP Model Network Skeleton

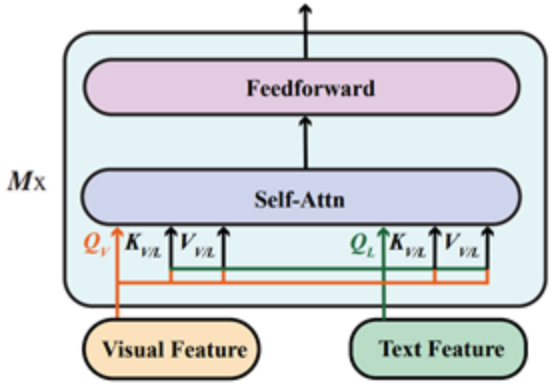


Multimodal Fusion

- Dual encoder: Cosine similarity
- Fusion Encoders: Transformer
  - co-attention
  - merged attention



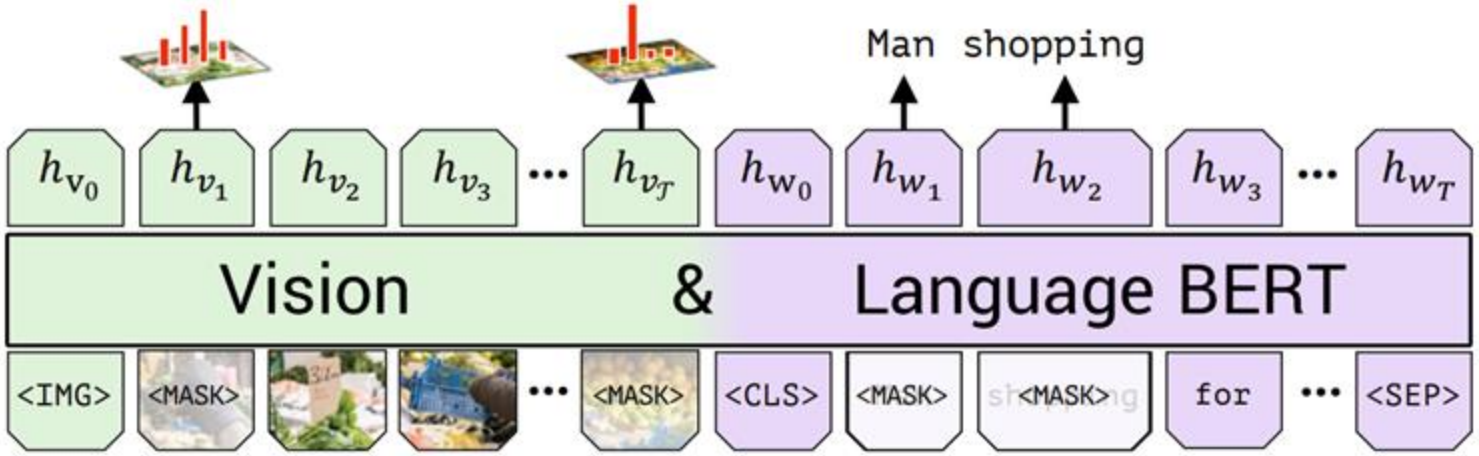




Pretraining Objectives

- Masked Language Modeling (MLM)
- Masked Image Modeling (MIM)
- Image-Text Matching (ITM)
- Image-Text Contrastive Learning (ITC)

Masked Language Modeling & Masked Region Modeling

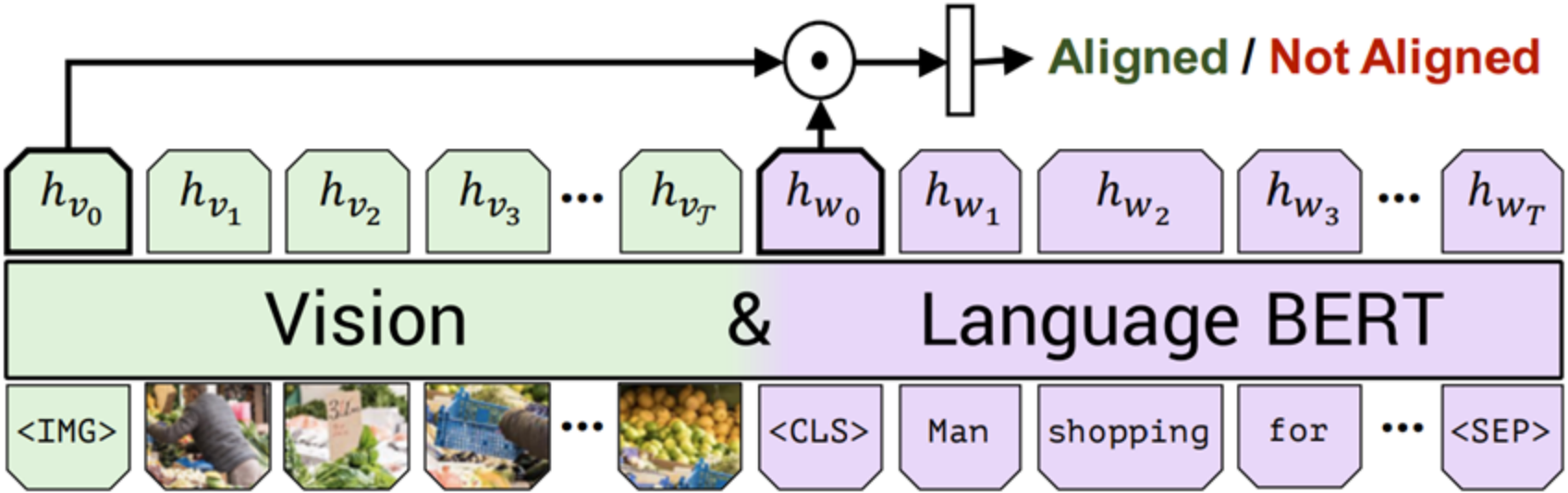


Masked Region Feature Regression
Masked Region Classification

$$\mathcal{L}_{\mathrm{MRM}}(\theta) = \mathbb{E}_{(\mathbf{w},\mathbf{v})\sim D} f_\theta(\mathbf{v_m}|\mathbf{v}_{\backslash\mathbf{m}}, \mathbf{w})$$

$$\mathcal{L}_{\mathrm{MLM}}(\theta) = -\mathbb{E}_{(\mathbf{w},\mathbf{v})\sim D} \log P_\theta(\mathbf{w_m}|\mathbf{w}_{\backslash\mathbf{m}}, \mathbf{v})$$

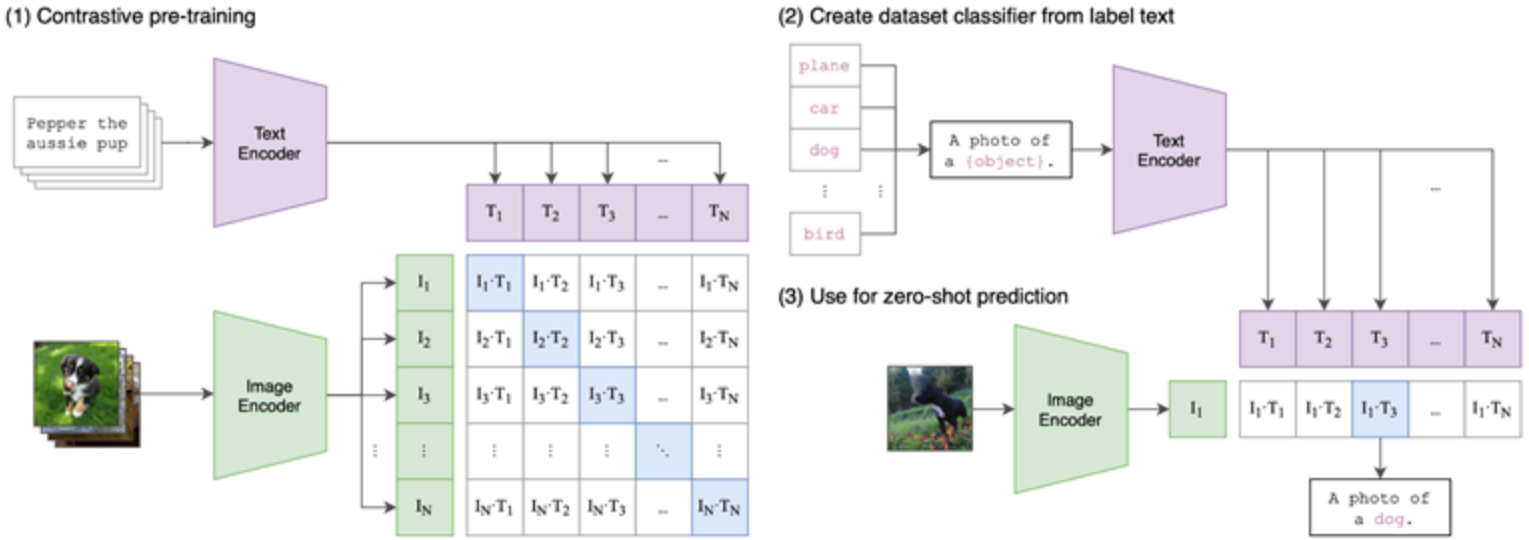
Image-Text Matching



$$\mathcal{L}_{\mathrm{ITM}}(\theta) = -\mathbb{E}_{(\mathbf{w},\mathbf{v})\sim D}[y \log s_\theta(\mathbf{w},\mathbf{v}) + (1-y)\log(1-s_\theta(\mathbf{w},\mathbf{v}))])$$

Image-Text Contrastive Learning

- Image-Text paired infoNCE



CLIP (Radford et al., 2021)

Advanced VLMs

- **Jointly Training Image and Text**

  VisualBERT (Li et al., 2019), SimVLM(Simple Visual Language Model; Wang et al., 2022), CM3 (Causally-Masked Multimodal Modeling; Aghajanyan, et al. 2022)

- **Learned Image Embedding as (Frozen) LM Prefix**

  ClipCap (Mokady et al., 2021), Frozen (Tsimpoukelli et al., 2021)

- **Text-Image Cross-Attention Fuse Mechanisms**

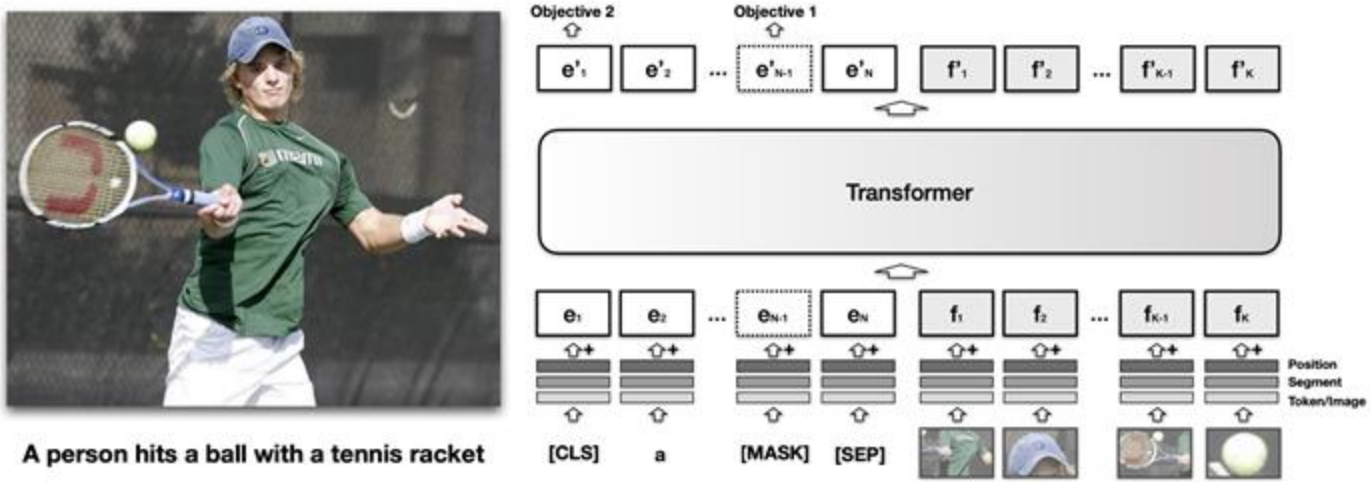
  VisualGPT (Chen et al., 2021), MERLOT (Zellers et al., 2021), Flamingo (Alayrac et al., 2022), CoCa (Contrastive Captioner; Yu & Wang et al., 2022)

- **No training**

  MAGiC (iMAge-Guided text generation with CLIP; Su et al., 2022), PICa (Prompts GPT-3 via the use of Image Captions; Yang et al., 2021), Socratic Models (Zeng et al., 2022)

Jointly Training with Image and Text

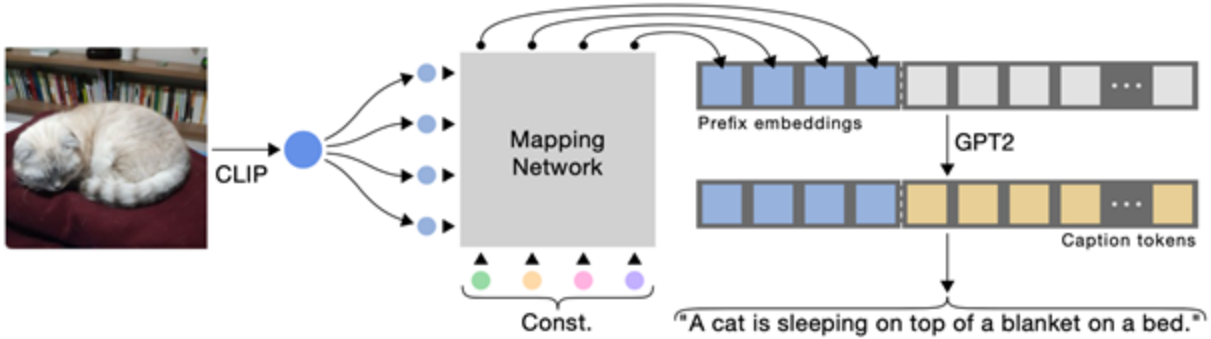
Images are divided into multiple smaller patches and each patch is treated as one object tag in the input sequence.



VisualBERT (Li et al., 2019)

Learned Image Embedding as (Frozen) LM Prefix

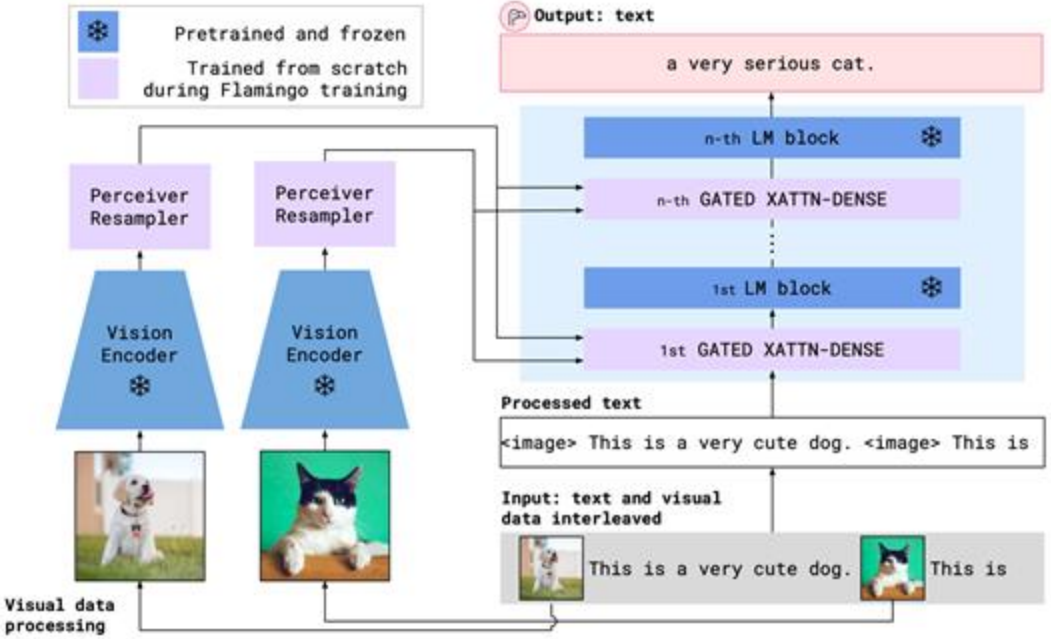
Inspired by prefix/ prompt tuning, only update the parameters of the vision module during training to produce vision prefix embeddings that can work with a pretrained, frozen language model.



ClipCap (Mokady et al., 2021)

Text-Image Cross-Attention Fuse Mechanisms

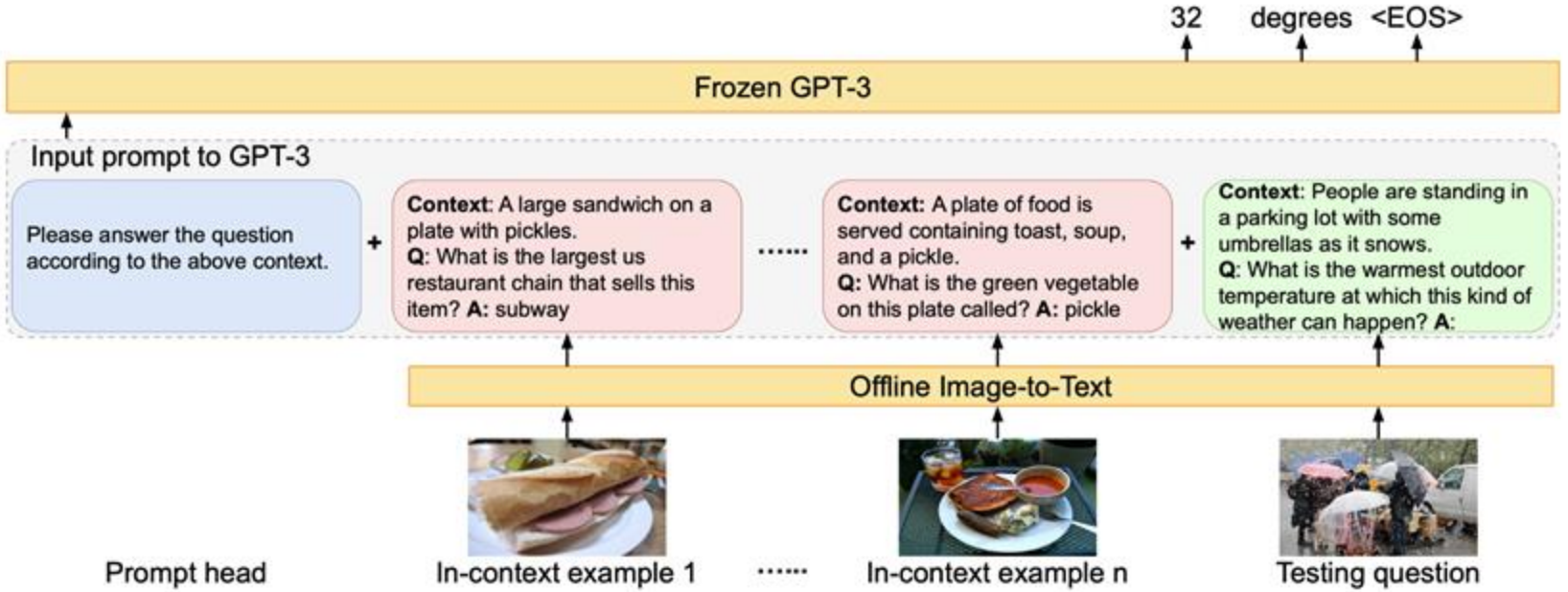
A specially designed cross-attention fuse mechanism to balance the mixture of text generation capacity and visual information.



Flamingo (Alayrac et al., 2022)

No Training

Language as Communication Interface



Socratic Models (Zeng et al., 2022)

Model Evaluation

- What are we doing now?
  - We evaluate on tasks/datasets like VQAv2, NLVR2, SNLI-VE, VCR, LSMDC, RefCOCO, etc.
  - Automatic evaluation, human evaluation
- Top on the leaderboard ≠ High-quality generations / predictions
- Some robustness analysis
  - Diagnostic tests: visio-linguistic compositionality, counting, rephrasing, image editing, reasoning
  - OOD generalization
  - Adversarial attacks: human-adversarial VQA
  - Probing

Gan, Z., Li, L., Li, C., Wang, L., Liu, Z. and Gao, J., 2022. Vision-language pre-training: Basics, recent advances, and future trends. Foundations and Trends® in Computer Graphics and Vision, 14(3–4), pp.163-352.

Future Challenges

- Unification: model architecture, tasks, etc.
- Efficient Computing: model compression, efficient pretraining, faster inference
- Training Strategy: multi-stage, multi-task
- Evaluation

Takeaways

- Multimodal is the scientific study of heterogeneous and interconnected data
- Inspired by the great success of Transformers in NLP, VLP models have attracted an increasing attention
- The VL model is consist of image encoder, text encoder and multimodal fusion module
- Typical pretraining tasks in VLP models are masked language /image modeling, image-text matching and image-text contrastive learning
- VL model evaluation is hard