

Multimodal Machine Learning:

- Flamingo
- CLIP
- ViLBERT

Chris Winsor
UMass Lowell CS5300: Deep Learning for NLP
4/2023

What is Multimodal Machine Learning (MMML)?


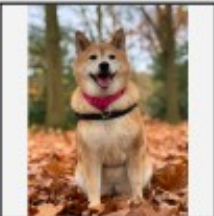







- MMML is machine learning that involves multiple sensory modalities. Most commonly (but not exclusively):
 - Natural language
 - Images or videos
- MMML can offer the following advantages:
 - Performance of each modality often outperforms single-modal baselines
 - Allows for a wide variety of downstream tasks and data formats

Flamingo

Flamingo

- DeepMind 2022 - a “Visual Language Model”
- How different:
 - Can accept arbitrarily interleaved text and images as input
 - Can perform different tasks and be “prompted” on task to be performed
 - Outputs free-form text
 - Requires only a handful of examples to fine tune

Interleaved Text/Image Prompting


Input Prompt						Completion
	This is a chinchilla. They are mainly found in Chile.		This is a shiba. They are very popular in Japan.		This is	a flamingo. They are found in the Caribbean and South America.
	What is the title of this painting? Answer: The Hallucinogenic Toreador.		Where is this painting displayed? Answer: Louvres Museum, Paris.		What is the name of the city where this was painted? Answer:	Arles.
	Output: "Underground"		Output: "Congress"		Output:	"Soulomes"

Multi-image visual dialog




What is the common thing about these three images?




 They are all flamingos.

What is the difference between these three images?




 The first one is a cartoon, the second one is a real flamingo, and the third one is a 3D model of a flamingo.



 This is an apple with a sticker on it.


What does the sticker say?



 The sticker says "iPod".


Where is the photo taken?



 It looks like it's taken in a backyard.


Do you think it is printed or handwritten?




 It looks like it's handwritten.

What color is the sticker?




 It's white.



 This is a cityscape. It looks like Chicago.

What makes you think this is Chicago?




 I think it's Chicago because of the Shedd Aquarium in the background.



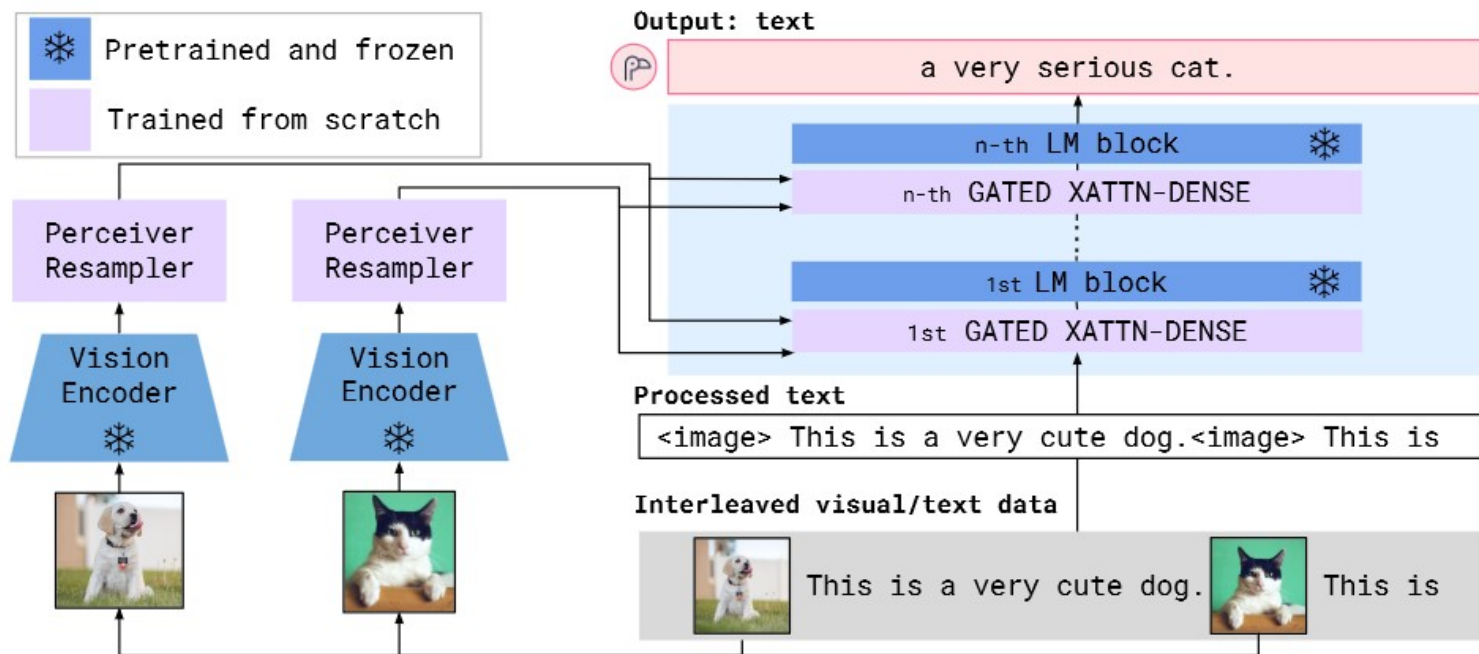
What about this one? Which city is this and what famous landmark helped you recognise the city?



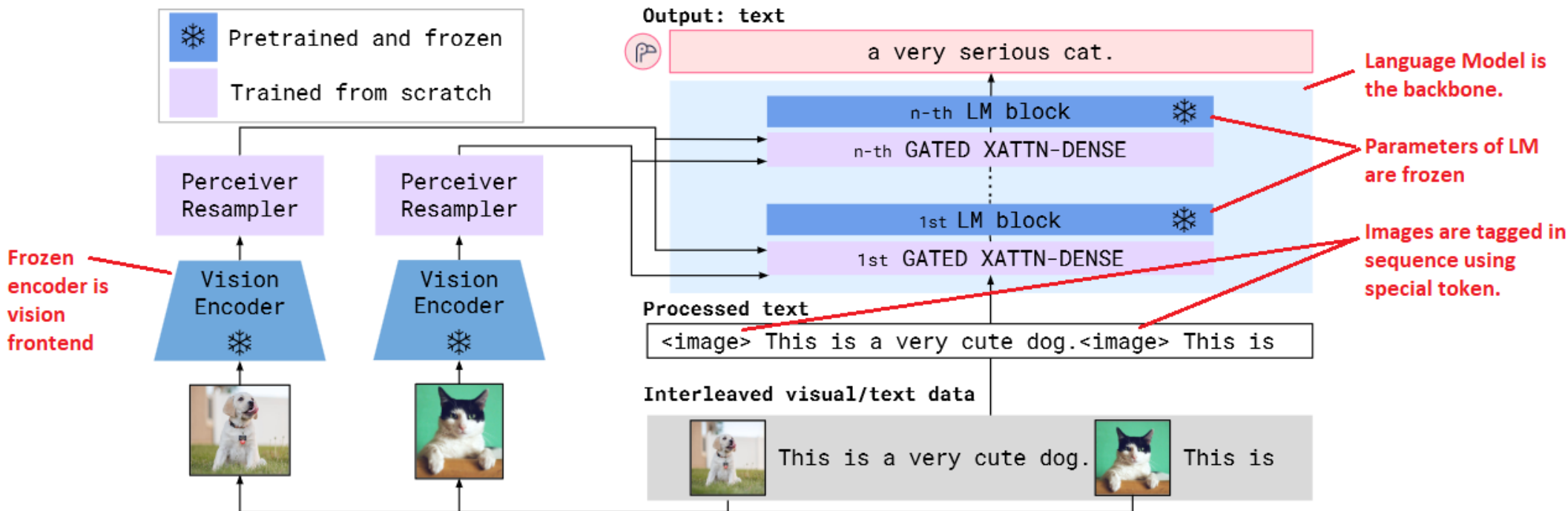
 This is Tokyo. I think it's Tokyo because of the Tokyo Tower.

Flamingo Architecture

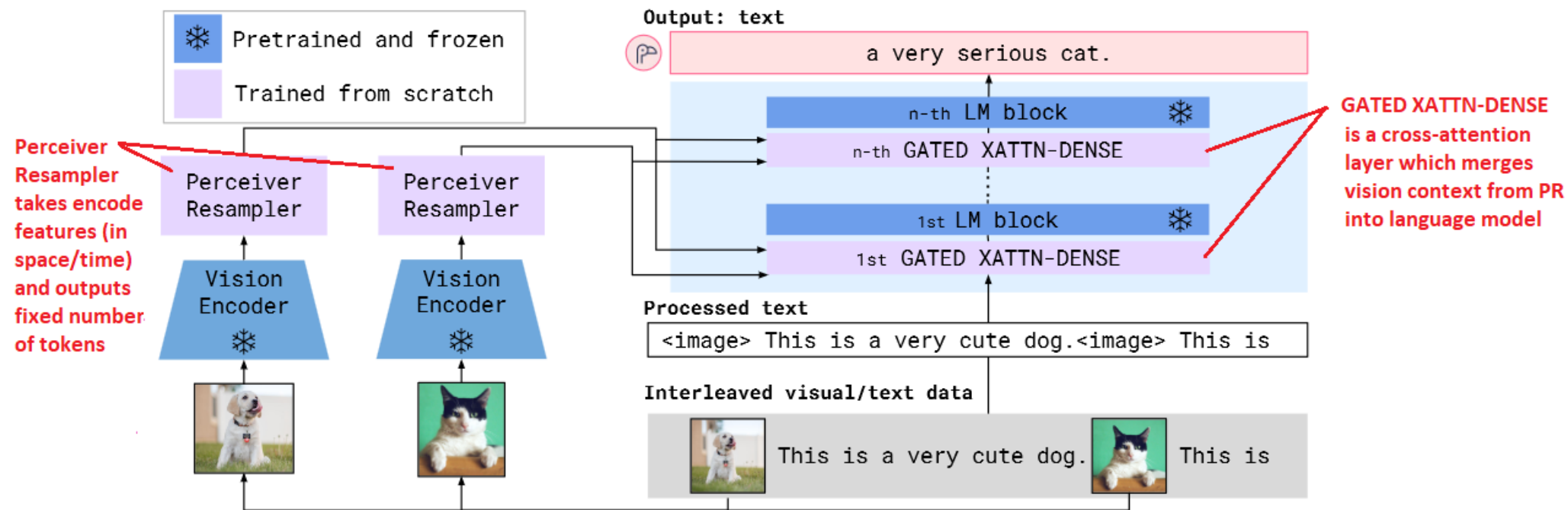
- Starts with pretrained language and vision models (frozen)
- Add trainable GATED XATTN-DENSE and Perceiver Resampler



Pretrained and Frozen Components



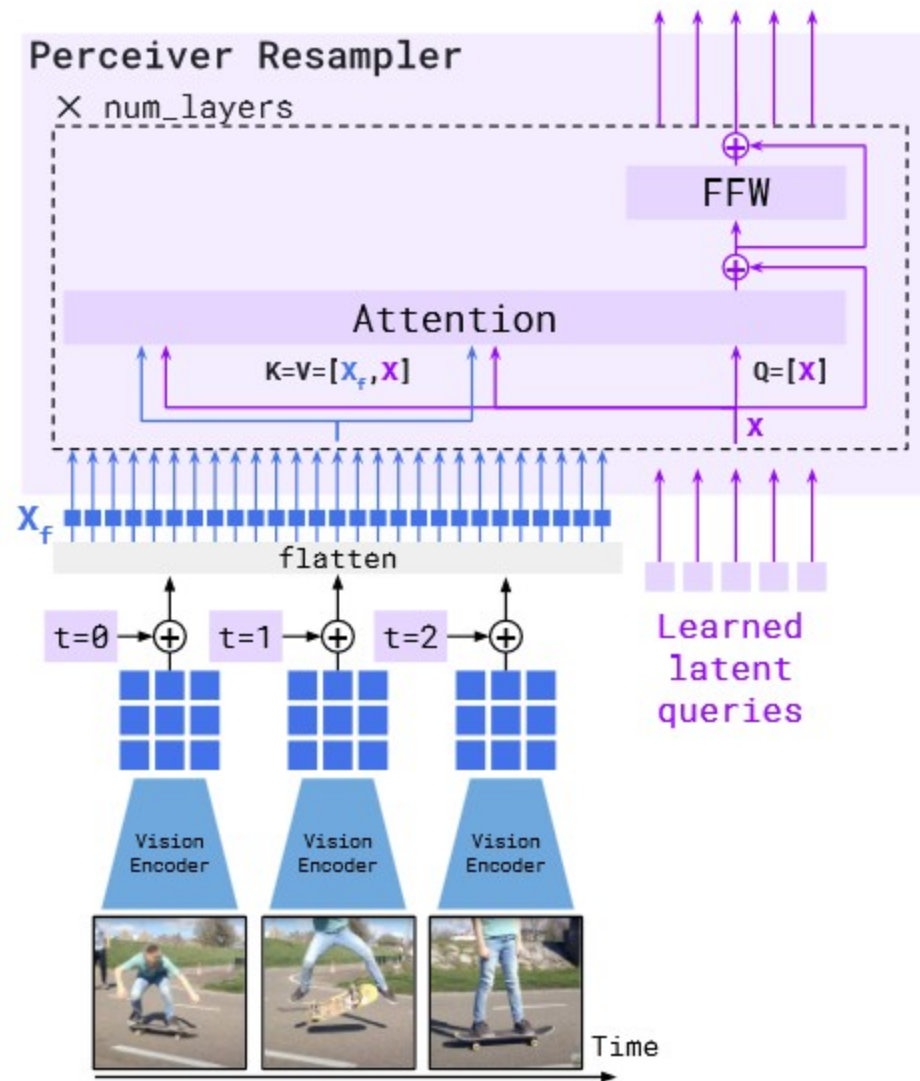
Trainables



Perceiver/Resampler

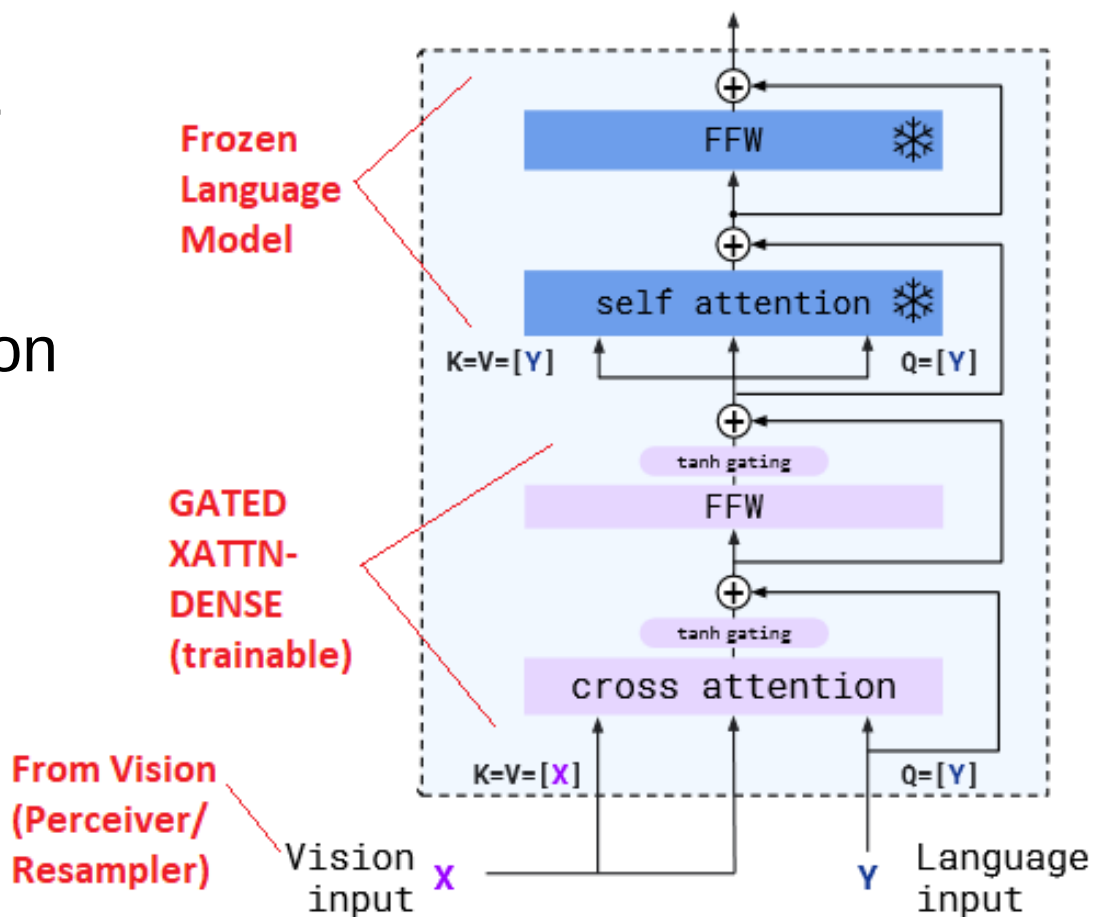
- Perceiver (Jaegle et al. 2021)⁽¹⁾ maps features to latent query units, similar to decoder-only transformer
- Transforms variable-size grid/time to fixed number of tokens
- Key, Value from data

(1) Jaegle, A., Gimeno, F., Brock, A., Zisserman, A., Vinyals, O., & Carreira, J. (2021). Perceiver: General Perception with Iterative Attention (arXiv:2103.03206).

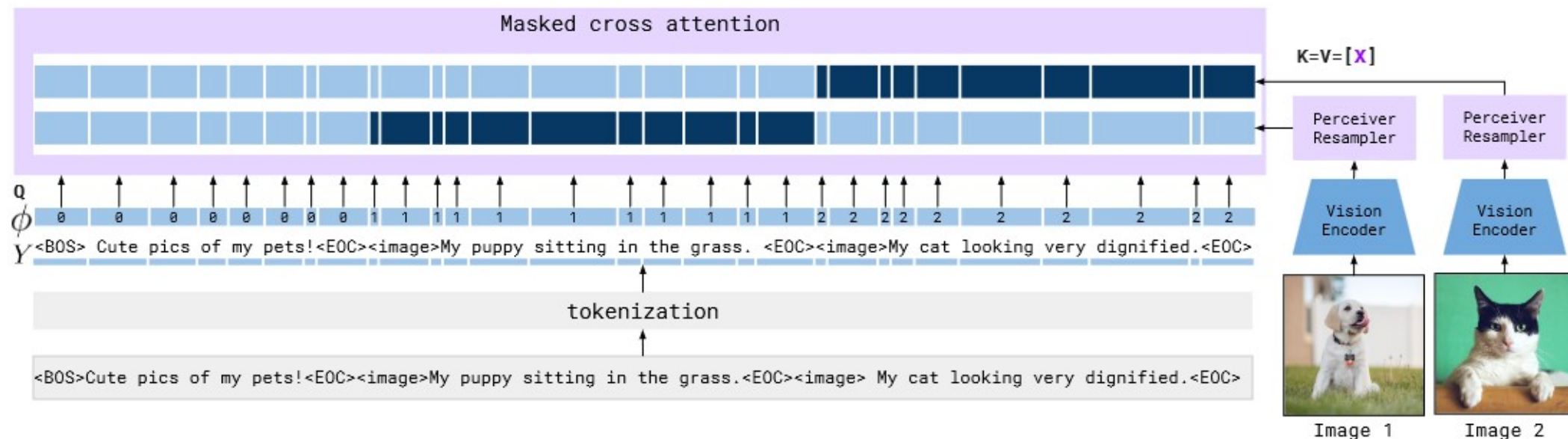


GATED XATTN-DENSE

- A cross-attention layer inserted into frozen Language Model
- Used to condition LM on Vision inputs
- K,V come from Vision while Q comes from Language



Interleaving Visual Data and Text



Training Datasets and Baseline Multi-Objective Task

- MultiModal Massive Web (M3W)
 - *Interleaved* Image and Text
 - Scraped from 43M web page (text + image)
 - 256 tokens of text, up to 5 images
- ALIGN dataset
 - Paired Image/Video and Text
 - 1.5B images paired w/ text (+ an additional 350M in-house)

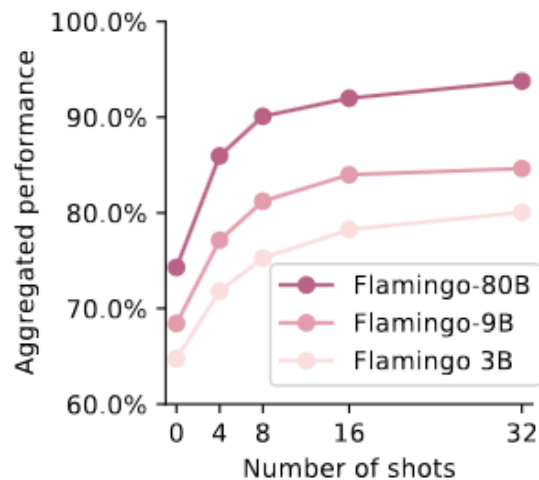
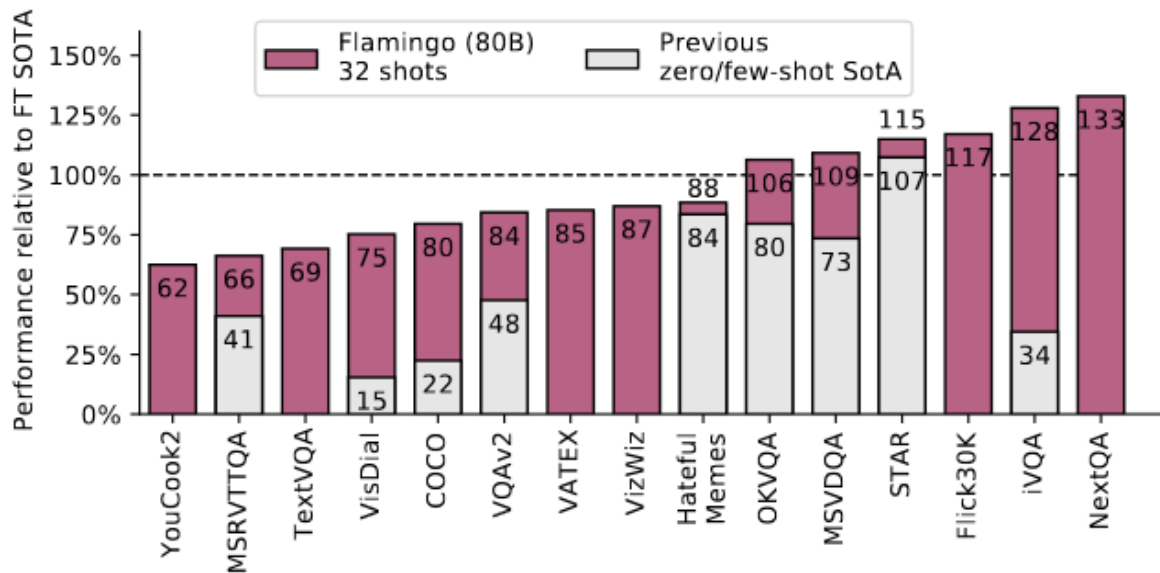
Multi-objective training and optimisation strategy. We train our models by minimizing a weighted sum of per-dataset expected negative log-likelihoods of text, given the visual inputs:

$$\sum_{m=1}^M \lambda_m \cdot \mathbb{E}_{(x,y) \sim \mathcal{D}_m} \left[- \sum_{\ell=1}^L \log p(y_\ell | y_{<\ell}, x_{\leq \ell}) \right], \quad (2)$$

where \mathcal{D}_m and λ_m are the m -th dataset and its weighting, respectively. Tuning the per-dataset weights λ_m is key to performance. We accumulate gradients over all datasets, which we found outperforms a “round-robin” approach [17]. We provide further training details and ablations in Appendix B.1.2.

Results

- Out-of-the-box beats fine-tuned SOTA (6/16 cases)
- On few-shot it outperforms all 9 peers



3 Model Sizes

	Requires model sharding	Frozen		Trainable		Total count
		Language	Vision	GATED XATTN-DENSE	Resampler	
<i>Flamingo-3B</i>	✗	1.4B	435M	1.2B (every)	194M	3.2B
<i>Flamingo-9B</i>	✗	7.1B	435M	1.6B (every 4th)	194M	9.3B
<i>Flamingo</i>	✓	70B	435M	10B (every 7th)	194M	80B

Task Adaption

- (to be provided)

This slide needs work

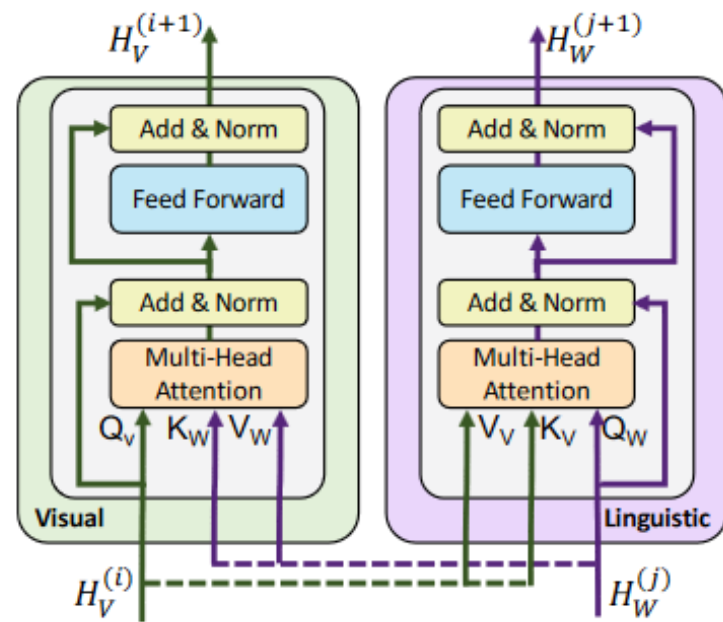
VILBERT

ViLBERT

- “Vision and Language BERT” - Facebook Research 2019
- “Learning a joint representation of language and visual content from paired data (...) specifically static images and corresponding descriptive text” (1)
- Applies BERT architecture to both NL and Vision data, adding a co-attention bridge between them

ViLBERT Approach

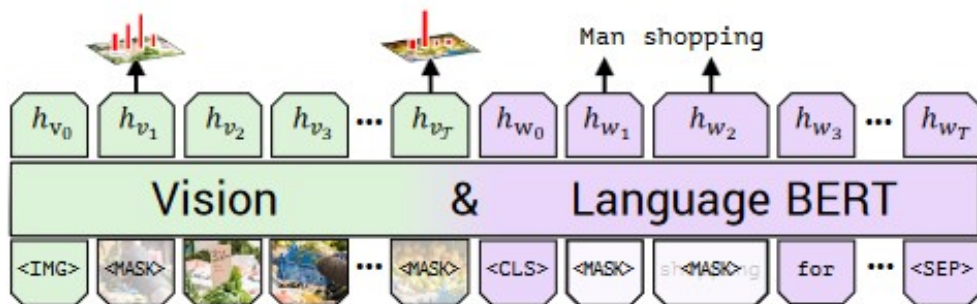
- Parallel transformers (modalities mostly independent)
- “Co-attention” layers pass keys and values across modalities
- Depth is different for each modality



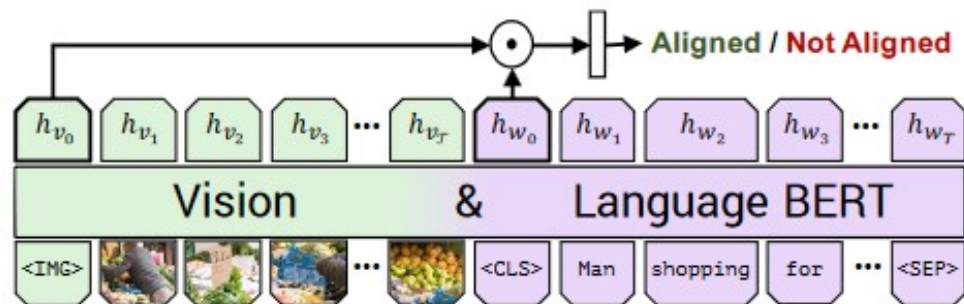
Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks (arXiv:1908.02265). arXiv.

ViLBERT Training Tasks

- Two tasks: a) Masked multi-modal learning, b) Multi-modal alignment prediction
 - a) reconstruct masked image region categories, masked words
 - b) predict whether caption describes the image content



(a) Masked multi-modal learning



(b) Multi-modal alignment prediction

ViLBERT Pre-training Dataset

- Conceptual Captions dataset (3.3M images w/ captions)
 - Language stream: Start with BERT-Base pretrained on BookCorpus and Wikipedia.
 - Vision stream: start with Faster R-CNN pretrained on VisualGenome

This slide needs work

ViLBERT Pre-Training

- Masked word, masked image (predict the masked word/image)
 - 15% masked (90/10 zero, unaltered)
- Predicting text segment / image correlation (does the text go with the image)

This slide needs work

ViLBERT Downstream tasks: VQA, VCR

- Visual Question Answering:
 - Choosing single word answer from dictionary, given image and NL question (a multi-class classification problem)
 - VQA 2.0 (Virginia Tech, Georgia Tech) is 260K images, 5 questions and 10 answers per image.
- Visual Commonsense Reasoning
 - Given image and NL question, select answer and rationale where A and R are NL (a pair of multi-choice selection problems)
 - VCR dataset (U.Washington et al) is 110K images, 290K questions, answers and rationales. Q,A,R are natural language.

3.2 – ViLBERT Downstream tasks: Referring Expressions, Caption-based Retrieval

- Referring expressions
 - Localize an image region based on natural language expression
 - RefCOCO dataset is 142k referring expressions for 50k objects in 20k images
 - ViLBERT fine tuning uses IoU on the area, 20 epochs
- Caption-based image retrieval
 - Choosing an image from a pool based on caption
 - Flickr30k dataset is a 31K image dataset with 5 human annotated natural language captions per image (150K captions)
 - Fine tuning uses 4-way multi choice (3 distractors) per image (1 pos, 3 negative)

Results

- ViLBERT is better than single-stream models (i.e. ViLBERT vs BERT)
- Pre-training improves performance (ViLBERT vs ViLBERT*)
- Fine tuned ViLBERT extends state of art on all 4 target tasks

CLIP

CLIP

- "Contrastive Language – Image Pre-training"
- OpenAI 2021

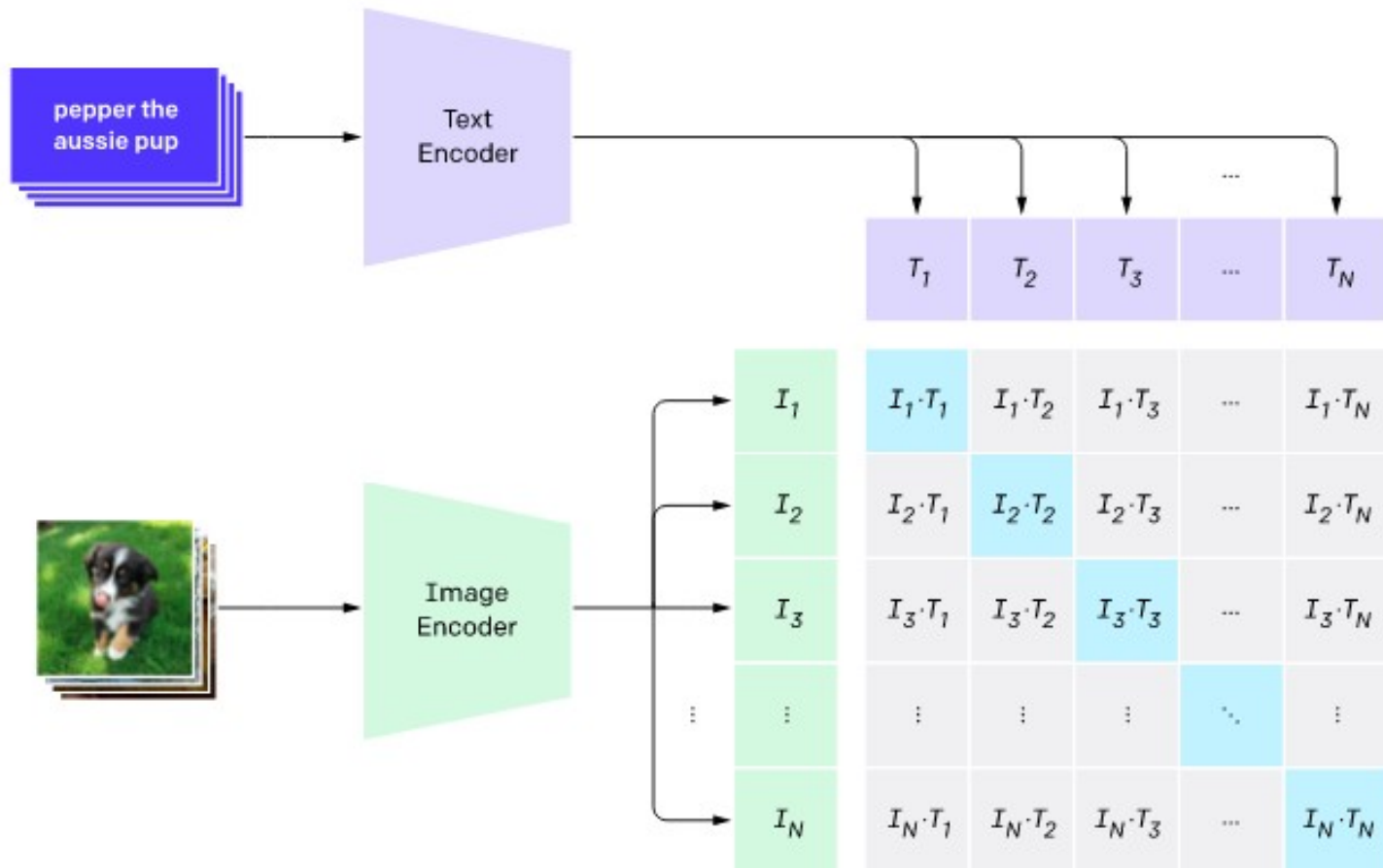
Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Aspell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision (arXiv:2103.00020).

Dataset and Training

- Ground truth data:
 - Start with a paired image and text (found on internet)
- Create a pretext task:
 - Given 32,768 randomly sampled text snippets ask which was actually paired with it
- To solve this task CLIP models need to learn to recognize a wide variety of visual concepts in images and associate them with their names.

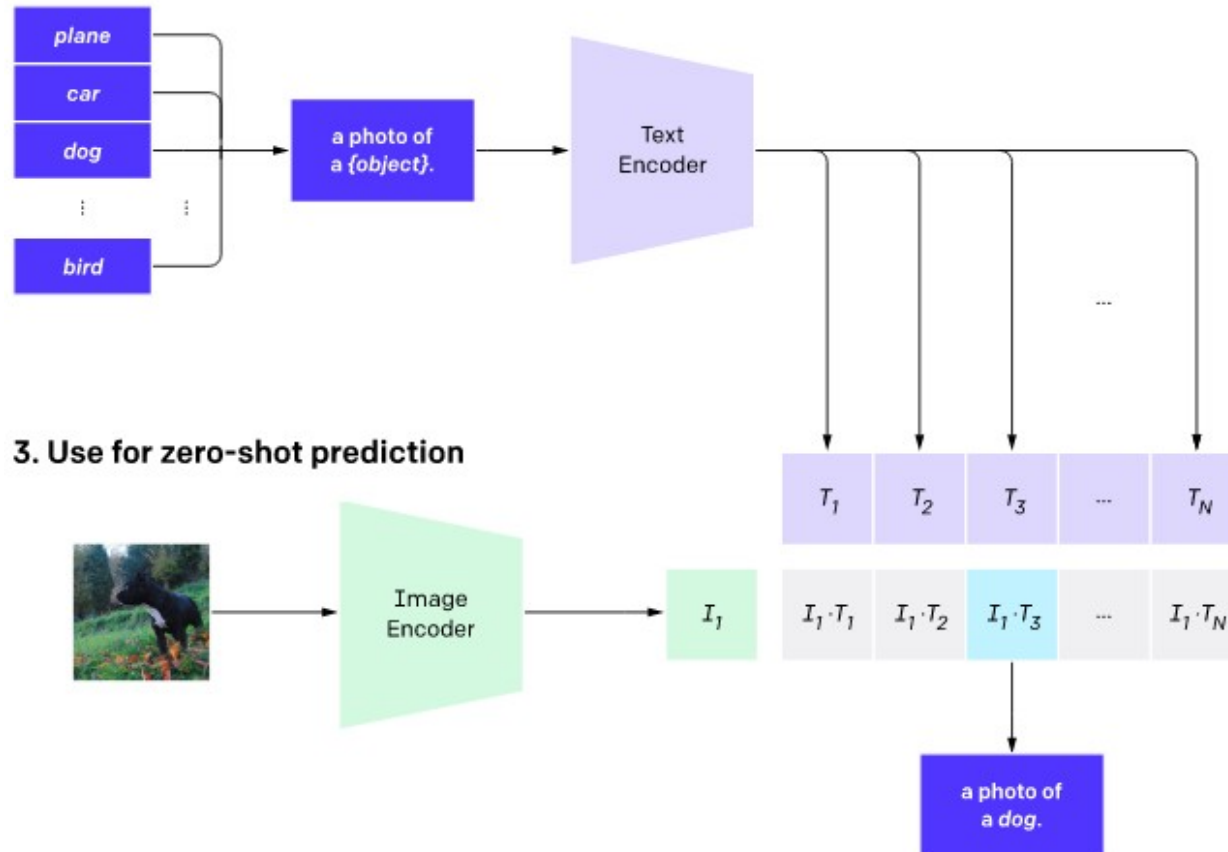
Contrastive Pre-training

(image and text encoders + cross-correlation)



Zero-shot Prediction

Each class converted to a caption, model predicts caption



Thank you!