

VQA(Visual Question Answering)



QA vs VQA



- What information source does a system build on?
 - A text passage, all Web documents, knowledge bases, tables, images..
- Question type
 - Factoid vs non-factoid, open-domain vs closed-domain, simple vs compositional, ..
- Answer type
 - A short segment of text, a paragraph, a list, yes/no, ...

Beyond textual QA problems

Today, we will mostly focus on how to answer questions based on **unstructured text**.

Visual QA



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



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

2. Reading comprehension

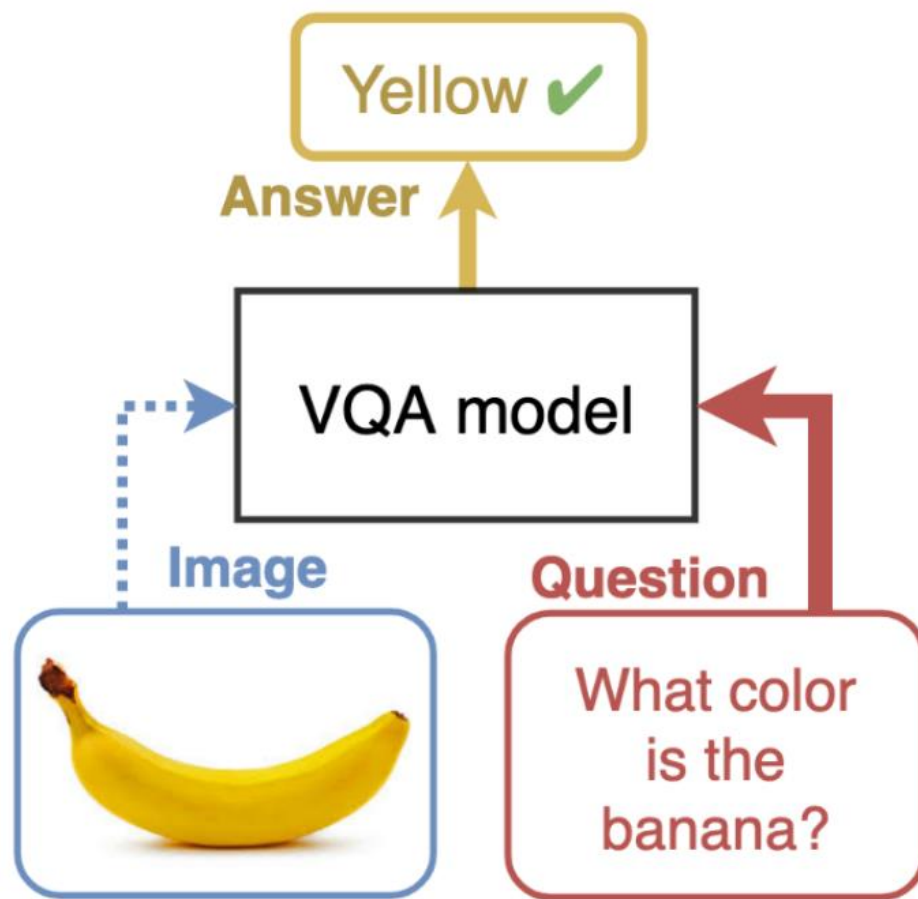
Reading comprehension = comprehend a passage of text and answer questions about its content (P, Q) \longrightarrow A

Tesla was the fourth of five children. He had an older brother named Dane and three sisters, Milka, Angelina and Marica. Dane was killed in a horse-riding accident when Nikola was five. In 1861, Tesla attended the "Lower" or "Primary" School in Smiljan where he studied German, arithmetic, and religion. In 1862, the Tesla family moved to Gospić, Austrian Empire, where Tesla's father worked as a pastor. Nikola completed "Lower" or "Primary" School, followed by the "Lower Real Gymnasium" or "Normal School."

Q: What language did Tesla study while in school?

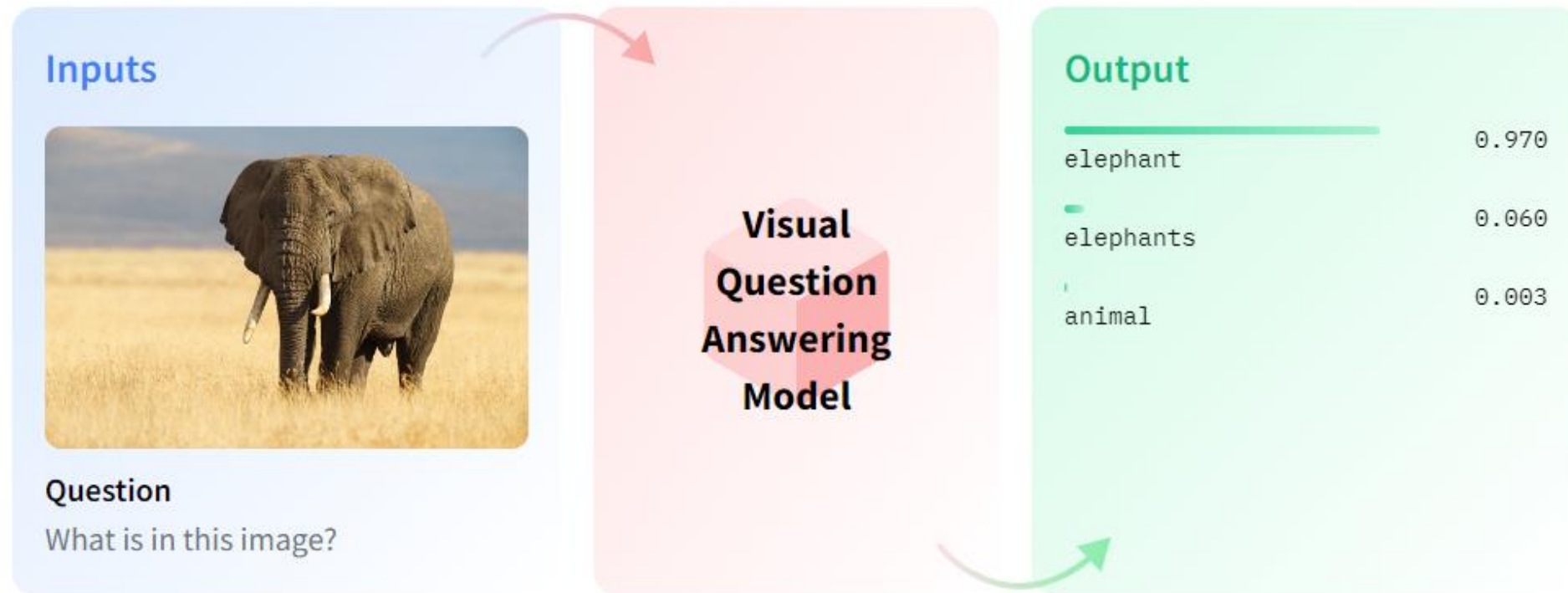
A: German

VQA system



VQA

- Visual Question Answering is the task of answering open-ended questions based on an image. They output natural language responses to natural language questions.



GPT4

User What is funny about this image? Describe it panel by panel.



Source: [hmmm \(Reddit\)](#)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

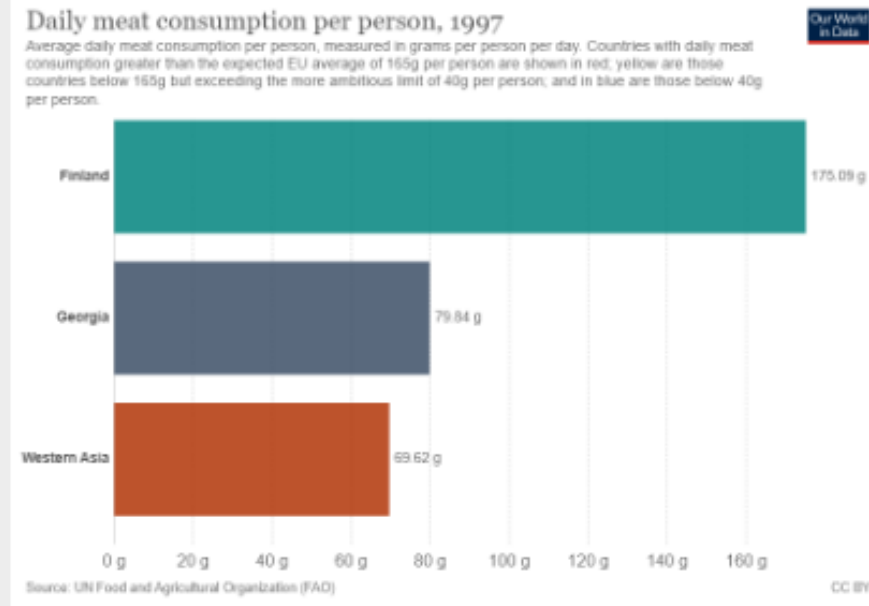
Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

GPT4

User What is the sum of average daily meat consumption for Georgia and Western Asia?
Provide a step-by-step reasoning before providing your answer.



GPT-4 To find the sum of average daily meat consumption for Georgia and Western Asia, we will follow these steps:

1. Identify the average daily meat consumption for Georgia.
2. Identify the average daily meat consumption for Western Asia.
3. Add the values from steps 1 and 2.

Step 1: From the provided chart, we see that the average daily meat consumption for Georgia is 79.84 grams per person per day.

Step 2: From the chart, we see that the average daily meat consumption for Western Asia is 69.62 grams per person per day.

Step 3: Add the values from steps 1 and 2.

$$79.84 \text{ g (Georgia)} + 69.62 \text{ g (Western Asia)} = 149.46 \text{ g}$$

So, the sum of average daily meat consumption for Georgia and Western Asia is 149.46 grams per person per day.

GPT4

User Answer question I.1.a. Think step-by-step.

I. Principe de la détection de rayonnement avec un bolomètre

Comme illustré sur la figure 1 un bolomètre est constitué d'un absorbeur qui reçoit le rayonnement que l'on désire détecter. Sa température T , supposée uniforme, est mesurée à l'aide d'un thermomètre incorporé, constitué d'un matériau conducteur dont la résistance $R(T)$ varie avec la température T ; cette variation est caractérisée par le coefficient $\alpha = \frac{1}{R} \frac{dR}{dT}$. L'ensemble possède la capacité thermique C_{th} .

Un barreau, conducteur thermique, homogène, de longueur L , de section S et de conductivité thermique λ et sous échanges thermiques latéraux, relie le bolomètre à un thermostat de température T_b fixe.

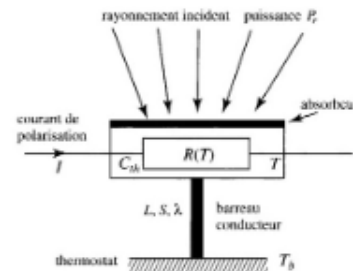


Figure 1 - Schéma d'un bolomètre

I.1. Dans cette première partie, on étudie le comportement du bolomètre en régime stationnaire, à la température de fonctionnement $T = T_b$.

I.1.a) Exprimer la température $T(x)$ en tout point du barreau conducteur repéré par son abscisse x ; on prendra $T(0) = T_0$ et $T(L) = T_b$.

GPT-4 To answer question I.1.a, we need to express the temperature $T(x)$ in every point of the conductive bar, which is characterized by the abscissa x . We are given that $T(0) = T_0$ and $T(L) = T_b$.

For a stationary regime, the heat equation for the one-dimensional heat conduction in the bar is:

$$d^2T/dx^2 = 0$$

Integrating once, we get:

$$dT/dx = C_1$$

Integrating a second time:

$$T(x) = C_1x + C_2$$

To determine the constants C_1 and C_2 , we use the boundary conditions:

$$T(0) = T_0, \text{ which gives } C_2 = T_0$$

$$T(L) = T_b, \text{ which gives } C_1 = (T_b - T_0) / L$$

Finally, the expression for $T(x)$ is:

$$T(x) = (T_b - T_0) * (x / L) + T_0$$

This represents a linear temperature profile along the conductive bar, with a slope of $(T_b - T_0) / L$.

GPT4

User What is unusual about this image?



Source: [Barricade](#)

GPT-4 The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

GPT4

User Below is part of the InstructGPT paper. Could you read and summarize it to me?



GPT-4 The InstructGPT paper focuses on training large language models to follow instructions with human feedback. The authors note that making language models larger doesn't inherently make them better at following a user's intent. Large models can generate outputs that are untruthful, toxic, or simply unhelpful.

To address this issue, the authors fine-tune language models on a wide range of tasks using human feedback. They start with a set of labeler-written prompts and responses, then collect a dataset of labeler demonstrations of the desired model behavior. They fine-tune GPT-3 using supervised learning and then use reinforcement learning from human feedback to further fine-tune the model. The resulting model, called InstructGPT, shows improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets.

The authors conclude that fine-tuning with human feedback is a promising direction for aligning language models with human intent.

User Could you explain the process described in Figure 2?

GPT-4 Figure 2 in the InstructGPT paper illustrates the three steps of their method:

1. Supervised fine-tuning (SFT): They collect demonstration data, which consists of labeler-written prompts and desired output behavior. This data is used to fine-tune GPT-3 with supervised learning.
2. Reward model (RM) training: They collect comparison data, where a labeler ranks the outputs from several model samples (from best to worst). This data is used to train the reward model.
3. Reinforcement learning using Proximal Policy Optimization (PPO): They optimize a policy against the reward model using reinforcement learning. For each new prompt sampled from the dataset, the policy generates an output. The reward model calculates a reward for the output, and the reward is used to update the policy using the PPO algorithm.

These three steps are iteratively performed to train the InstructGPT model, which aims to generate outputs that align better with human preferences and follow instructions more effectively.

GPT4

User Can you explain this meme?

Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.



GPT-4 This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world. The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.

GPT4

User Can you explain why this is funny. Think about it step-by-step.



GPT-4 The comic is satirizing the difference in approaches to improving model performance between statistical learning and neural networks.

In statistical learning, the character is shown to be concerned with overfitting and suggests a series of complex and technical solutions, such as minimizing structural risk, reworking the loss function, and using a soft margin.

In contrast, the neural networks character simply suggests adding more layers to the model. This is often seen as a common solution to improving performance in neural networks, but it's also considered a simplistic and brute-force approach.

The humor comes from the contrast between the complexity and specificity of the statistical learning approach and the simplicity and generality of the neural network approach. The "But unironically" comment adds to the humor by implying that, despite being simplistic, the "stack more layers" approach is often effective in practice.

Dataset



Is something under the sink broken?	yes	no
	yes	no
	yes	no
What number do you see?	33	5
	33	6
	33	7



Does this man have children?	yes	yes
	yes	yes
	yes	yes
Is this man crying?	no	no
	no	yes
	no	yes



How many glasses are on the table?	3	2
	3	2
	3	6
What is the woman reaching for?	door handle	fruit
	glass	glass
	wine	remote



Can you park here?	no	no
	no	no
	no	yes
What color is the hydrant?	white and orange	red
	white and orange	red
	white and orange	yellow



Has the pizza been baked?	yes	yes
	yes	yes
	yes	yes
What kind of cheese is topped on this pizza?	feta	mozzarella
	feta	mozzarella
	ricotta	mozzarella



Do you think the boy on the ground has broken legs?	yes	no
	yes	no
	yes	yes
Why is the boy on the right freaking out?	his friend is hurt	ghost
	other boy fell down	lightning
	someone fell	sprayed by hose



What kind of store is this?	bakery	art supplies
	bakery	grocery
	pastry	grocery
Is the display case as full as it could be?	no	no
	no	yes
	no	yes



How many pickles are on the plate?	1	1
	1	1
	1	1
What is the shape of the plate?	circle	circle
	round	round
	round	round



Are the kids in the room the grandchildren of the adults?	probably	yes
	yes	yes
	yes	yes
What is on the bookshelf?	nothing	books
	nothing	books
	nothing	books



How many bikes are there?	2	3
	2	4
	2	12
What number is the bus?	48	4
	48	46
	48	number 6



What does the sign say?	stop	stop
	stop	stop
	stop	yield
What shape is this sign?	octagon	diamond
	octagon	octagon
	octagon	round



How many balls are there?	2	1
	2	2
	2	3
What side of the teeter totter is on the ground?	right	left
	right	left
	right side	right side

Steps to solve VQA

- Input Processing
- Feature Extraction
- Alignment of Features
- Multimodal Fusion
- Answer Prediction

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.

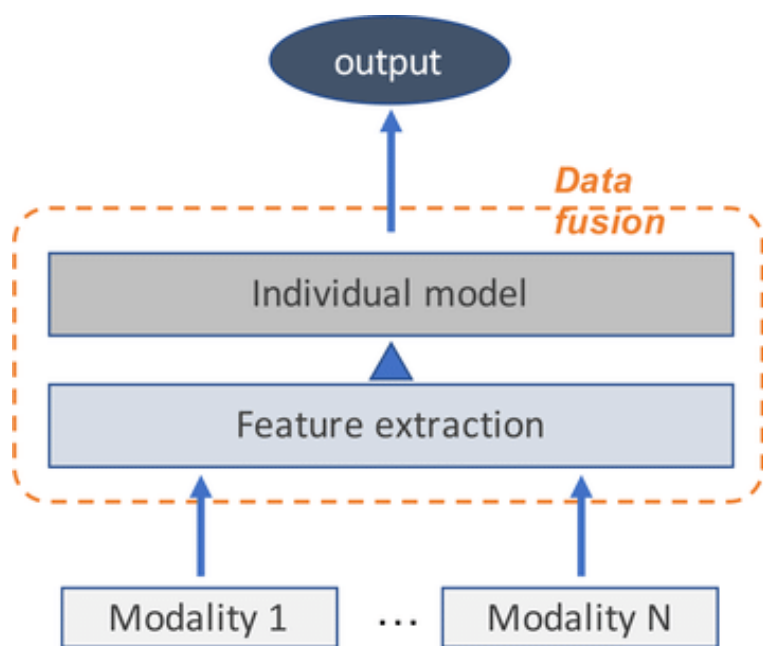


A group of people sitting on a boat in the water.

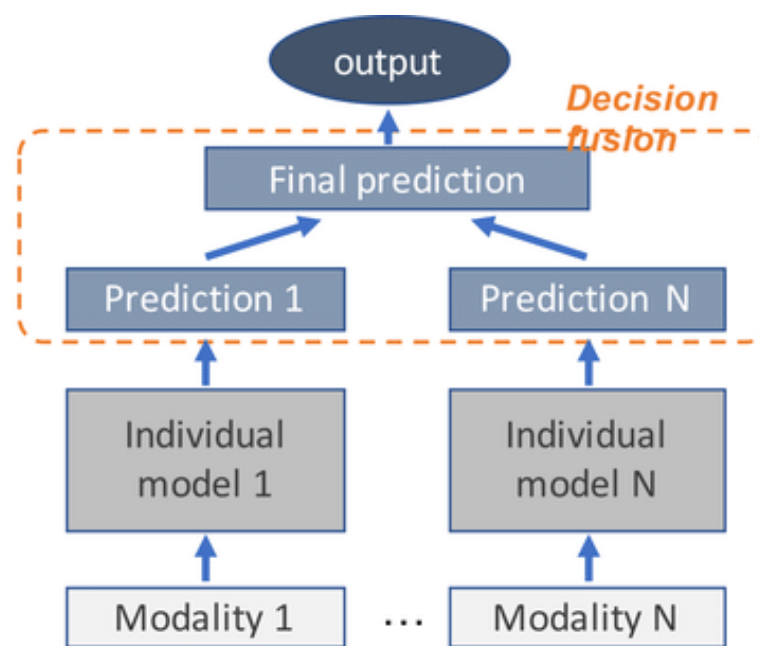


A giraffe standing in a forest with trees in the background.

Early Fusion and Late Fusion

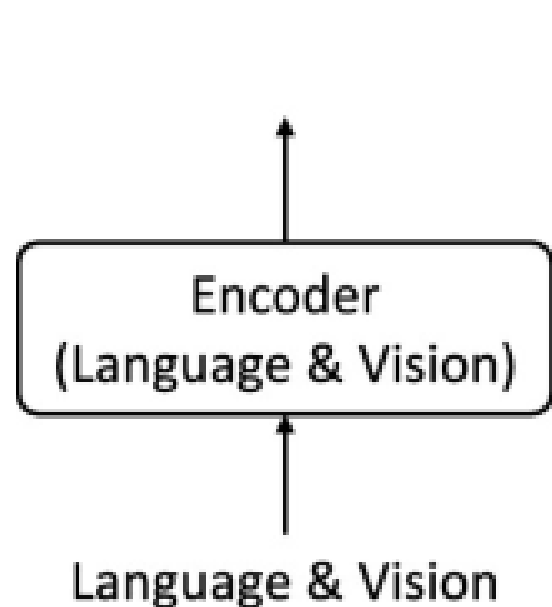


(a) Early fusion

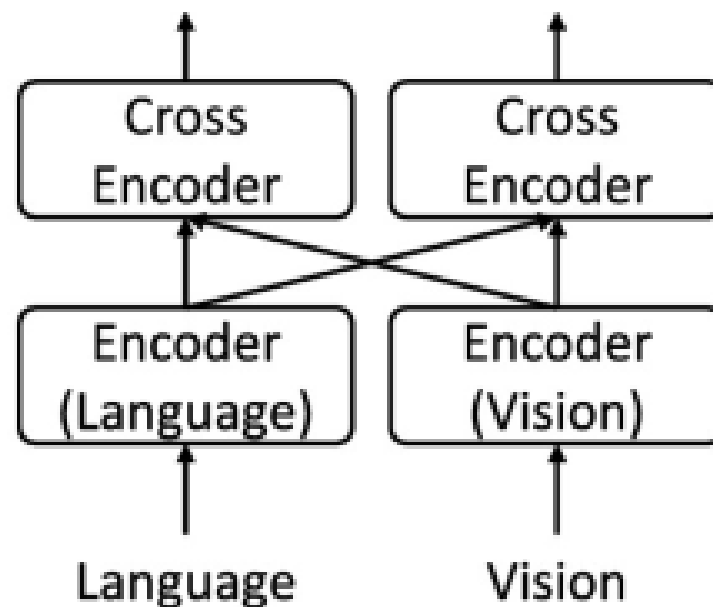


(b) Late fusion

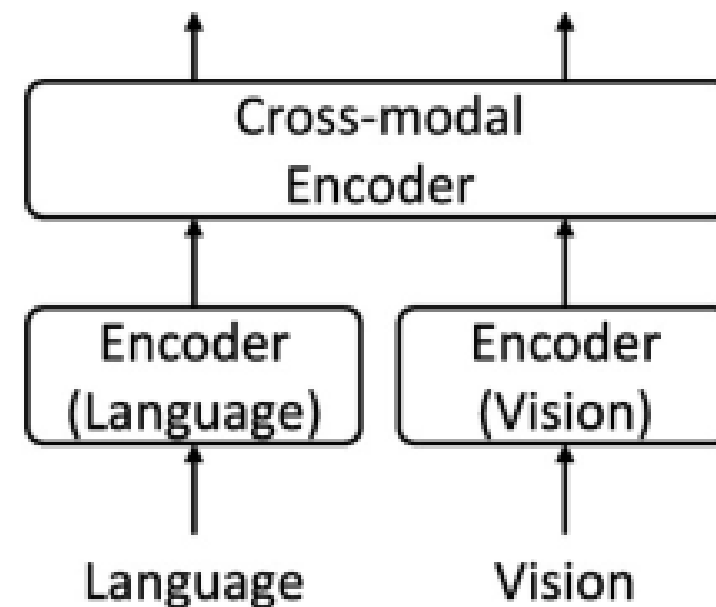
Architectures



(a) Single Stream



Cross Type

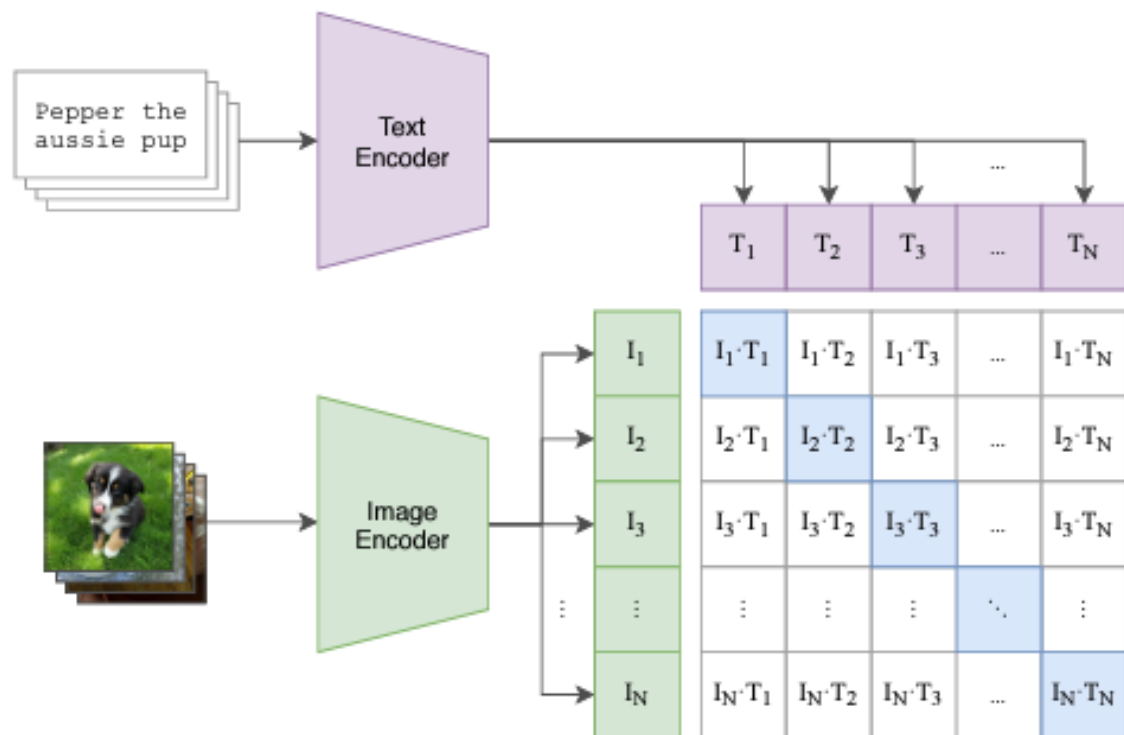


Joint Type

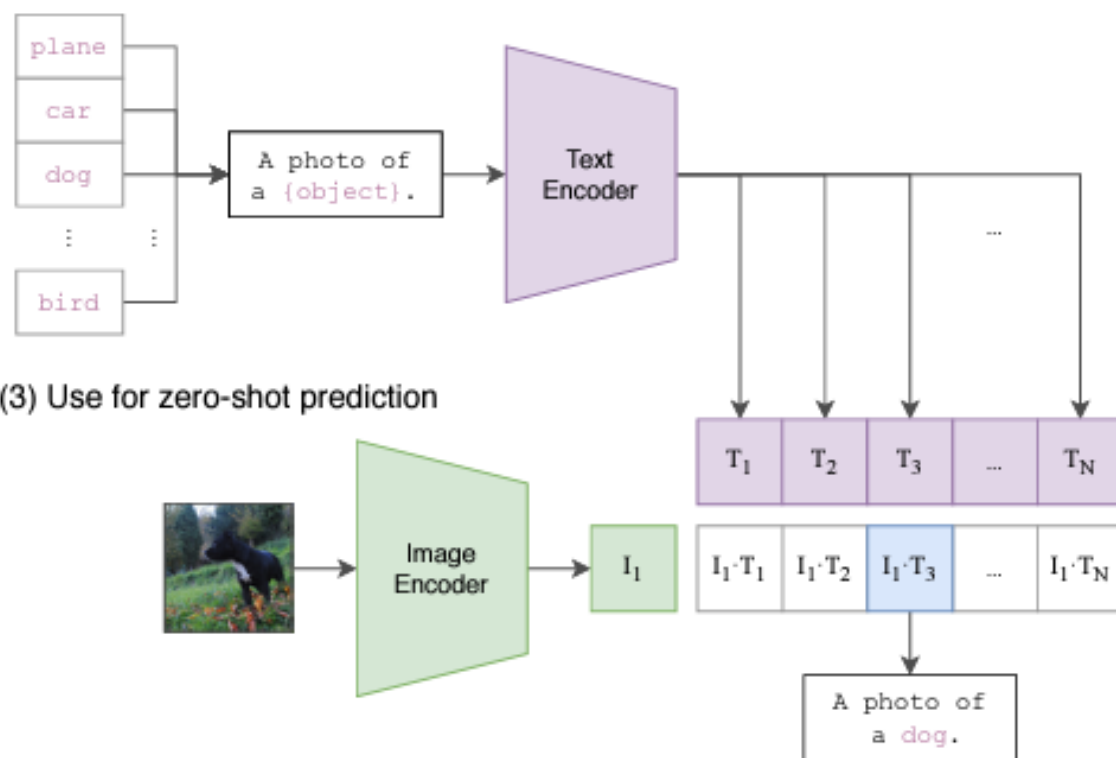
(b) Two Stream

CLIP

(1) Contrastive pre-training



(2) Create dataset classifier from label text

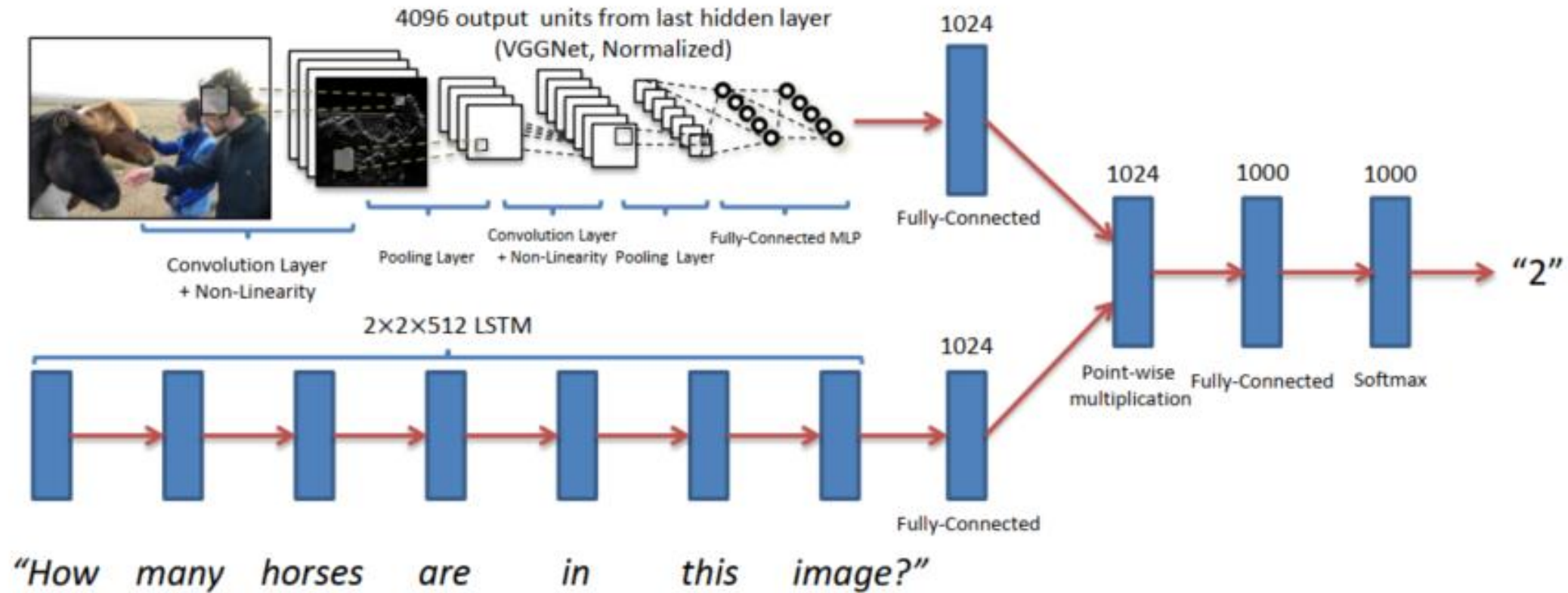


(3) Use for zero-shot prediction

Evaluation

- Classification:
 - Accuracy
 - Top-k Accuracy
- Open-ended
 - VQA Accuracy
 - BLEU
 - METEOR
 - ROUGE
 - CIDEr
 - SPICE

First Deep learning VQA model



How to Combine Image Representation and Question Representation?

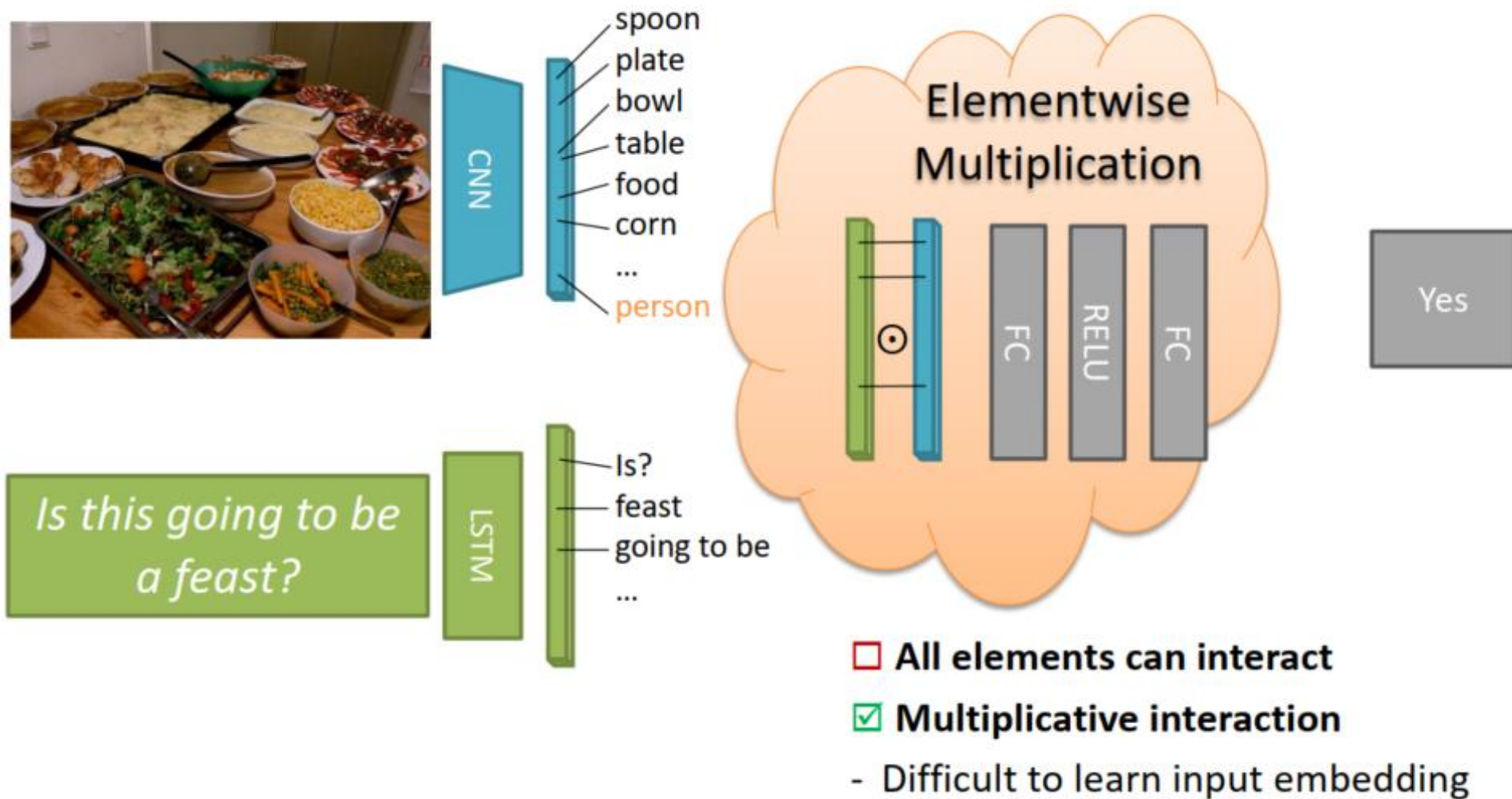
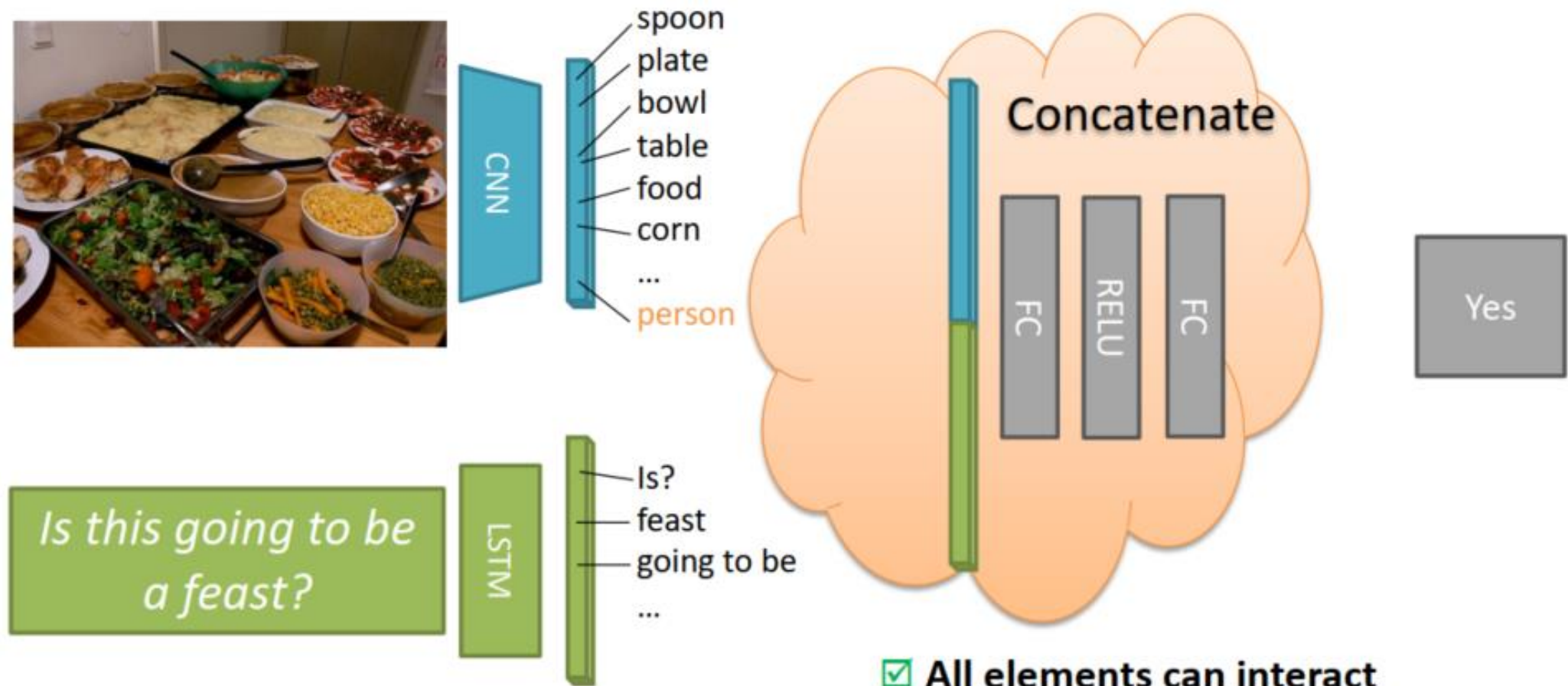


Image Representation and Question Representation?

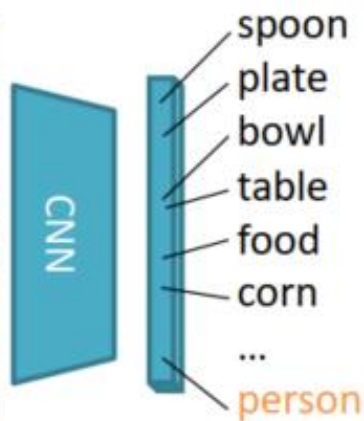


✓ **All elements can interact**

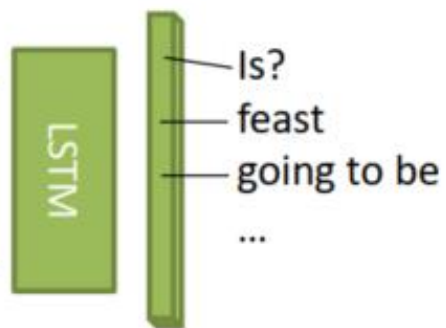
□ **Multiplicative interaction**

- Difficult to learn output classification

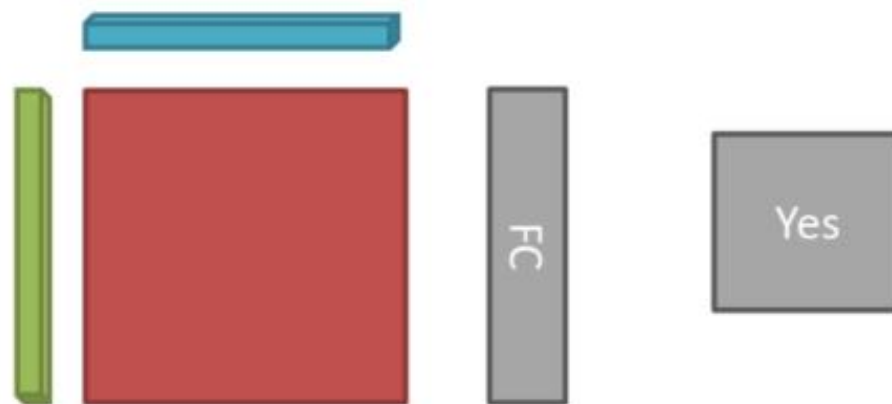
How to Combine Image Representation and Question Representation?



*Is this going to be
a feast?*



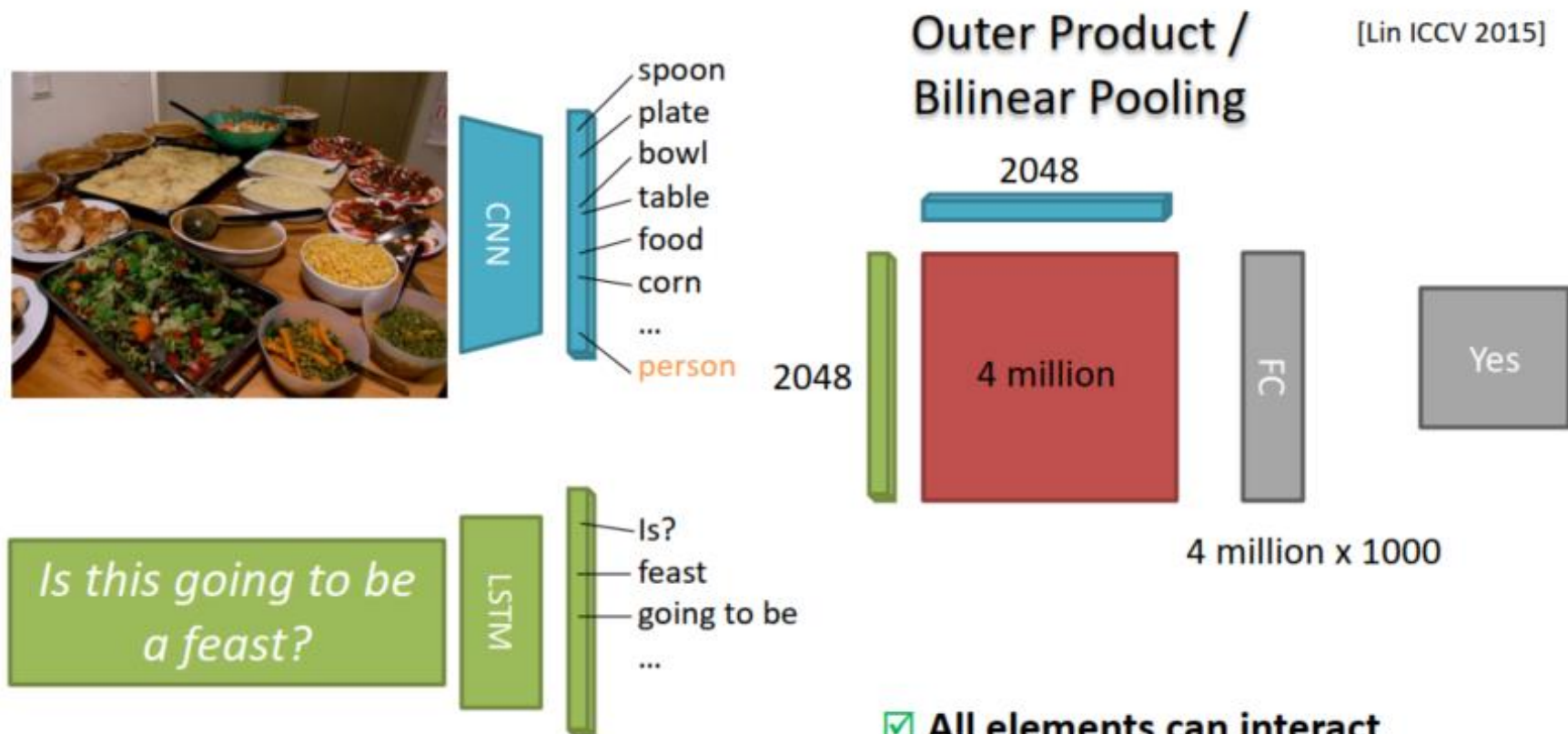
Outer Product /
Bilinear Pooling [Lin ICCV 2015]



✓ All elements can interact

✓ Multiplicative interaction

How to Combine Image Representation and Question Representation?



- ✓ All elements can interact
- ✓ Multiplicative interaction
- High #activations & computation
- High #parameters

[Lin ICCV 2015] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji.
Bilinear CNN models for fine-grained visual recognition. ICCV 2015

Slide credit: Akira Fukui and Marcus Rohrbach

Multimodal Fusion

Similarity

- Inner product: $\mathbf{u}\mathbf{v}$

Linear / sum

- Concat: $W[\mathbf{u}, \mathbf{v}]$
- Sum: $W\mathbf{u} + V\mathbf{v}$
- Max: $\max(W\mathbf{u}, V\mathbf{v})$

Multiplicative

- Multiplicative: $W\mathbf{u} \odot V\mathbf{v}$
- Gating: $\sigma(W\mathbf{u}) \odot V\mathbf{v}$
- LSTM-style: $\tanh(W\mathbf{u}) \odot V\mathbf{v}$

Attention

- Attention: $\alpha W\mathbf{u} + \beta V\mathbf{v}$
- Modulation: $[\alpha\mathbf{u}, (1-\alpha)\mathbf{v}]$

Bilinear

- Bilinear: $\mathbf{u}W\mathbf{v}$
- Bilinear gated: $\mathbf{u}W\sigma(\mathbf{v})$
- Low-rank bilinear: $\mathbf{u}U^T V\mathbf{v} = P(U\mathbf{u} \odot V\mathbf{v})$
- Compact bilinear: $\text{FFT}^{-1}(\text{FFT}(\Psi(\mathbf{x}, \mathbf{h}_1, \mathbf{s}_1)) \odot \text{FFT}(\Psi(\mathbf{x}, \mathbf{h}_2, \mathbf{s}_2)))$

Pre-training Objectives

- Image-Text matching
- Masked language modeling
- Masked region modeling (masked image modeling)
- Image-text alignment (image-text contrastive learning)

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 + EfficientNet	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	EfficientNet		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

Codes

- <https://colab.research.google.com/drive/1YN0fVQnu5mzsUlfjdMFqh-HNCcjy-nC2?usp=sharing>