

Visual Question Answering

Danna Gurari

University of Colorado Boulder

Fall 2022



<https://home.cs.colorado.edu/~DrG/Courses/NeuralNetworksAndDeepLearning/AboutCourse.html>

Review

- Last week:
 - Explosion of transformers
 - GPT
 - BERT
 - ViT
 - Limitations of transformer models
- Assignments (Canvas)
 - Problem set 3 due earlier today
 - Lab assignment 4 due in 1.5 weeks
 - Final project proposal due in 3 weeks
- Questions?

Today's Topics

- Visual question answering applications
- Visual question answering datasets
- Visual question answering evaluation
- Mainstream challenge 2015 winner: baseline approach
- Mainstream challenge 2019 winner: transformer-based approach
- Programming tutorial

Today's Topics

- Visual question answering applications
- Visual question answering datasets
- Visual question answering evaluation
- Mainstream challenge 2015 winner: baseline approach
- Mainstream challenge 2019 winner: transformer-based approach
- Programming tutorial

Task: Answer Visual Questions (VQs)



Is my monitor on?



Hi there can you please tell me what flavor this is?



Does this picture look scary?



Which side of the room is the toilet on?

Visual Assistance for People with Vision Loss; e.g.,



Be My Eyes

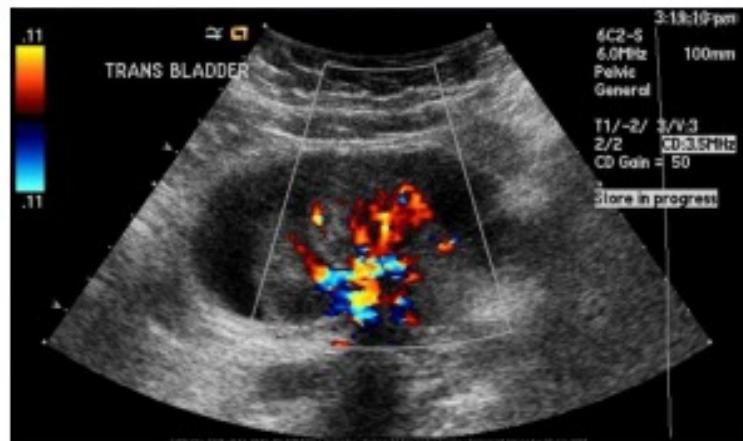
Visual Assistance for People with Vision Loss; e.g.,

The screenshot shows a YouTube search results page with the query "vizwiz app". The results are filtered by "VizWiz app". There are three video thumbnails displayed:

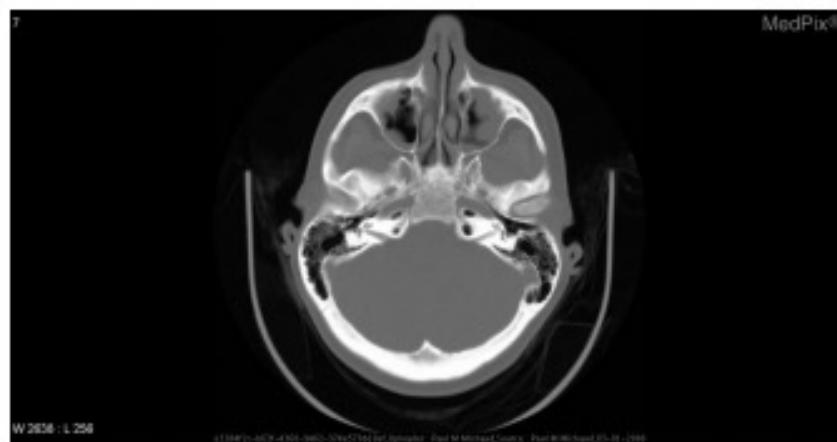
- VizWiz Tutorial** by UMBCHCC (11K views, 7 years ago). Description: "In this video, we show you how to submit a question using VizWiz and the VoiceOver accessibility features on your Apple iOS ...".
- Identify Anything with VizWiz** by The Carroll Center for the Blind (723 views, 4 years ago). Description: "Today I'm going to show you a variety of applications I refer to as identification applications. On my Apple iPhone 4S I am going to ...".
- VizWiz Demo - Tom Dekker** by WoofManTom (1.4K views, 7 years ago). Description: "Using VizWiz app on iPhone."

The left sidebar of the YouTube interface is visible, showing navigation links like Home, Trending, Subscriptions, Library, and History, along with a "SIGN IN" button.

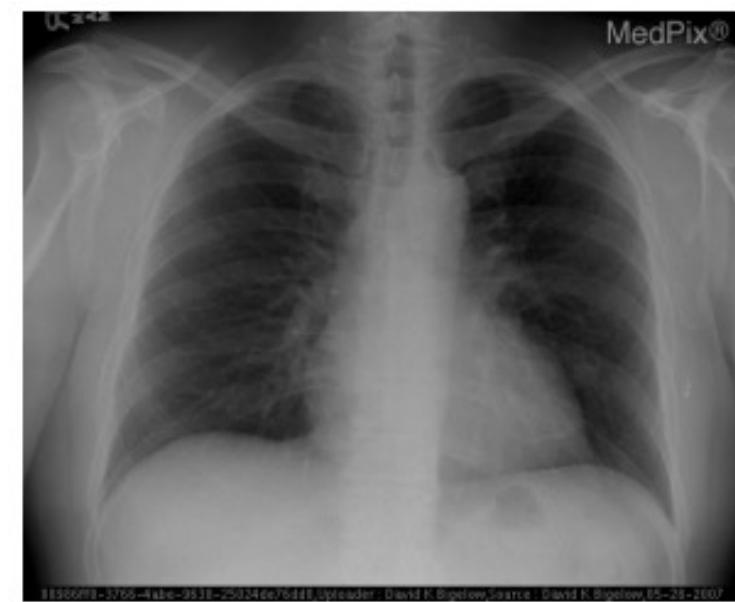
Medical VQA



(a) **Q:** what imaging method was used? **A:** us-d - doppler ultrasound



(b) **Q:** which plane is the image shown in? **A:** axial



(e) **Q:** what abnormality is seen in the image? **A:** nodular opacity on the left#metastatic melanoma

Video Surveillance



Attribute-based Query:

Q: Is it a person in the green bounding box?

(Define the person as P1)

A: Yes

Q: Is P1 female?

A: Yes

Q: Does P1 hold a bag?

A: Yes

Q: Does P1 has long hair and wear leather shoes?

A: Yes

Q: Is P1 in padded jacket and skirt?

A: No

Q: ...

A: ...

Relationship-based Query:

Q: Are they persons in both of the two red bounding boxes?

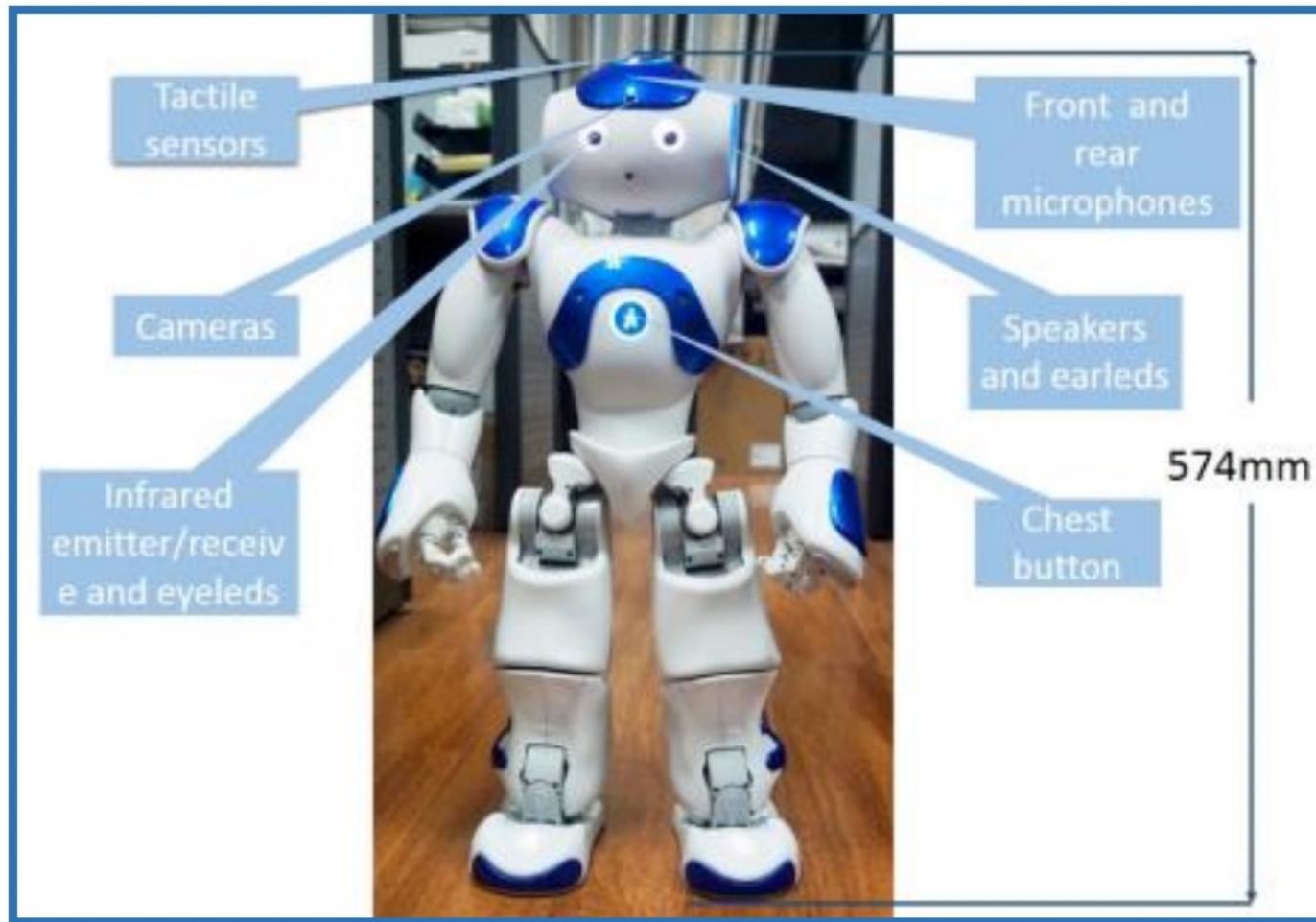
A: Yes

(Define the upper one as P2, and define the lower one as P3)

Q: Are P2 and P3 the same person?

A: Yes

Education (e.g., for Preschoolers)



Answers questions about
quantity and **colors** of
detected objects

Audio Guide for Museums and Art Galleries

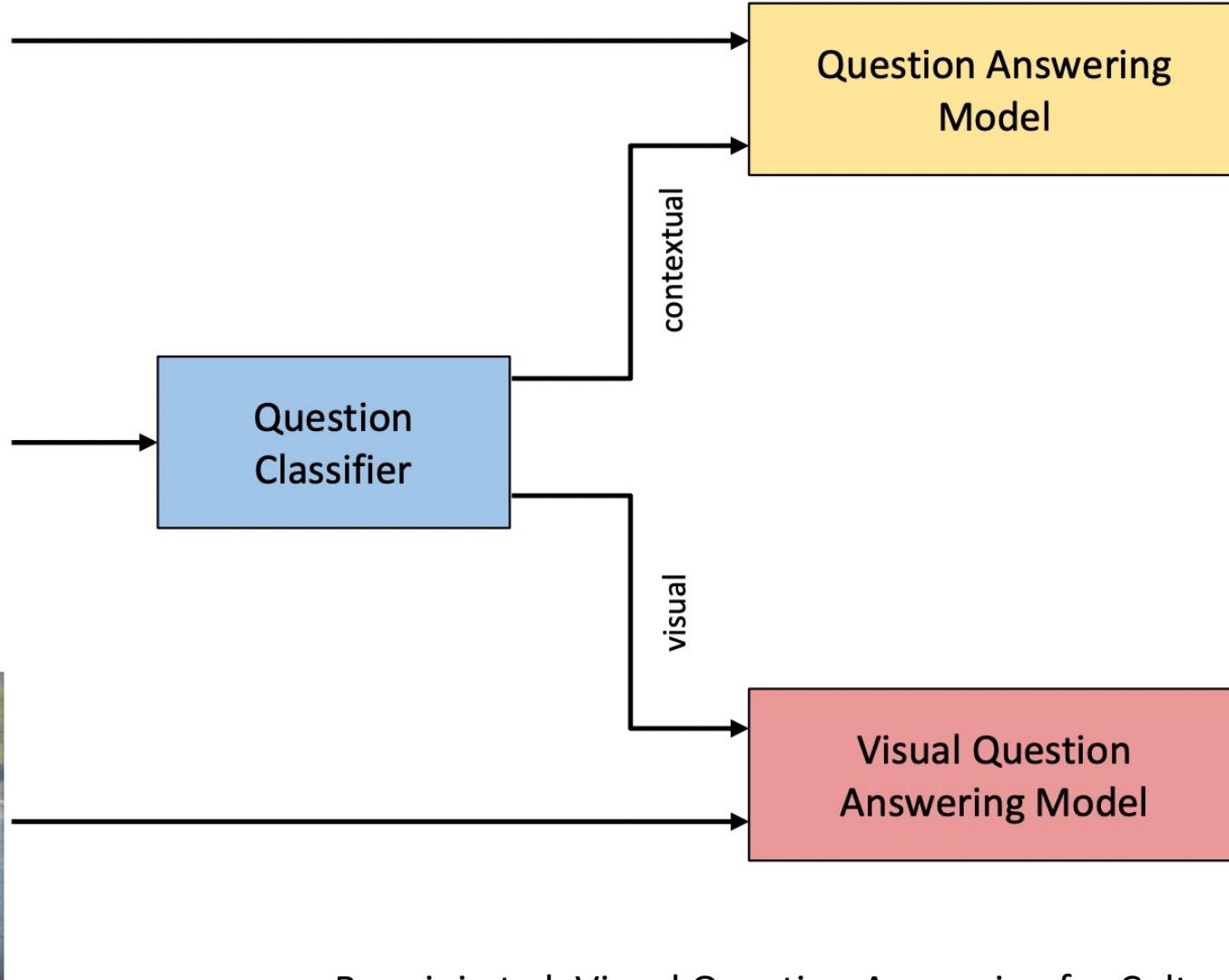
TEXTUAL DESCRIPTION

Bathers at Asnieres is an oil-on-canvas painting by the French artist Georges Pierre Seurat, the first of his two masterpieces on the monumental scale...

QUESTION

Who is the author of this painting?

IMAGE



Advertising: Understanding the Messaging and Identifying Effective Persuasion Strategies



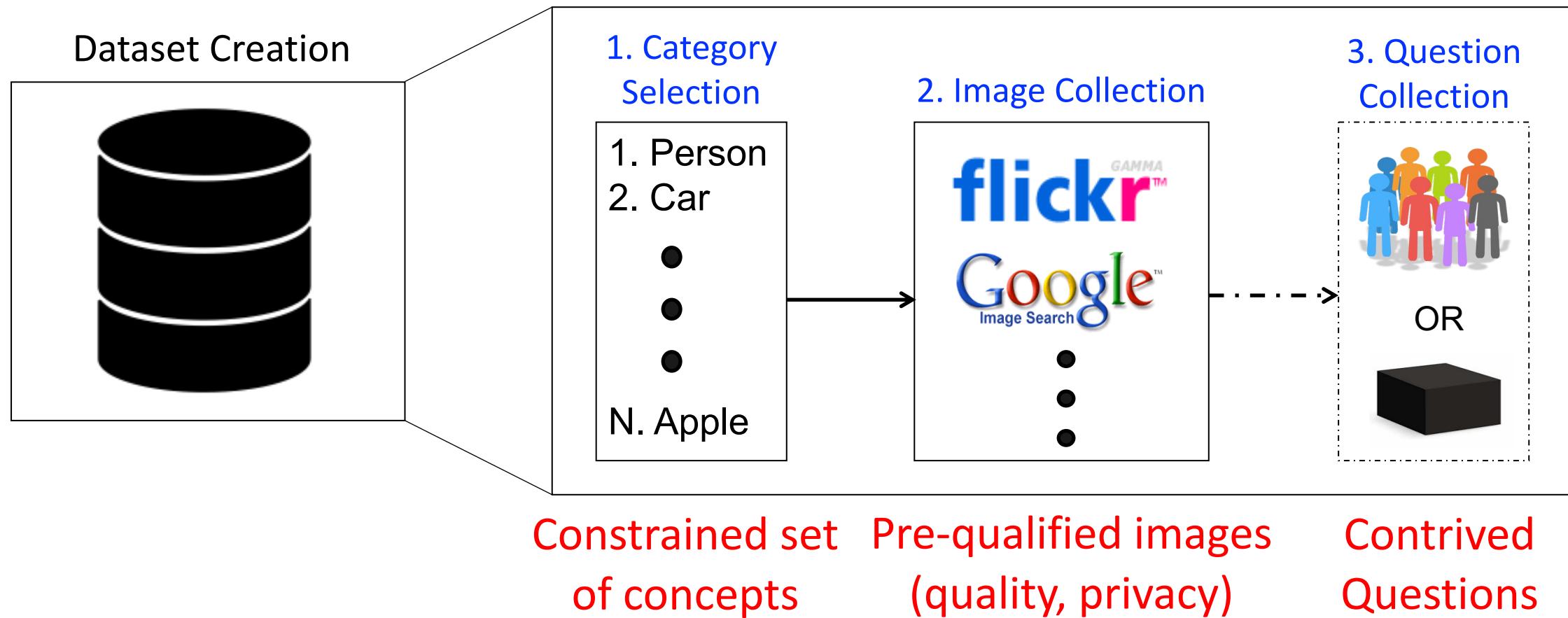
What should I do, according to this ad?	Why, according to this ad, should I take this action?	What should I do, according to this ad?	Why, according to this ad, should I take this action?	What should I do, according to this ad?	Why, according to this ad, should I take this action?
I should try this lipstick	Because it is better than the brand Mac	I should sign up for the Asus promo	Because there are free gifts	I should prevent verbal abuse	Because its as bad as physical abuse

For what other applications might visual question answering systems be useful?

Today's Topics

- Visual question answering applications
- **Visual question answering datasets**
- Visual question answering evaluation
- Mainstream challenge 2015 winner: baseline approach
- Mainstream challenge 2019 winner: transformer-based approach
- Programming tutorial

Status Quo (Approach to Create 14+ Datasets)



Stump a smart robot! Ask a question about this scene that a human can answer, but a smart robot probably can't!

Updated instructions: Please read carefully

Hide

Show

We have built a smart robot. It understands a lot about scenes. It can recognize and name all the objects, it knows where the objects are, it can recognize the scene type (e.g., kitchen, beach), people's expressions and poses, and properties of objects (e.g., the color of objects, their texture). Your task is to stump this smart robot! In particular, it already knows answers to some questions about this scene. We will tell you what these questions are.

Ask a question about this scene that this SMART robot probably can not answer, but any human can easily answer while looking at the scene in the image.

IMPORTANT: The question should be about this scene. That is, the human should need the image to be able to answer the question – the human should not be able to answer the question without looking at the image.

e.g., Question Generation



Your work will get rejected if you do not follow the instructions below:

- Do not ask questions that are similar to the ones listed below each image. As mentioned, the robot already knows the answers to those questions for the scene in this image. Please ask about something different.
- Do not repeat questions. Do not ask the same questions or the same questions with minor variations over and over again across images. Think of a new question each time specific to the scene in each image.
- Each question should be a single question. Do not ask questions that have multiple parts or multiple sub-questions in them.
- Do not ask generic questions that can be asked of many other scenes. Ask questions specific to the scene in each image.

Below is a list of questions the smart robot can already answer. Please ask a different question about this scene that a human can answer "if" looking at the scene in the image (and not otherwise), but would stump this smart robot:

Q1: What is unusual about this mustache? (The robot already knows the answer to this question.)

Q2: What is her facial expression? (The robot already knows the answer to this question.)

Q3: Write your question, different from the questions above, here to stump this smart robot.

e.g., Answer Generation

10 answers
collected from
10 crowdworkers



Help Us Answer Questions About Images!

Updated instructions: Please read carefully

Hide

Show

Please answer some questions about images **with brief answers**. Your answers should be how most other people would answer the questions. If the question doesn't make sense, please try your best to answer it and indicate via the buttons that you are unsure of your response.

If you don't follow the following instructions, your work will be rejected.

Your work **will get rejected** if you do not follow the instructions below:

- Answer the question based on what is going on in **the scene depicted in the image**.
- Your answer should be a **brief phrase** (not a complete sentence).
 - "It is a kitchen." -> "kitchen"
- For yes/no questions, please **just say yes/no**.
 - "You bet it is!" -> "yes"
- For numerical answers, please use **digits**.
 - "Ten." -> "10"
- If you need to speculate (e.g., "What just happened?"), provide an answer **that most people would agree on**.
- If you don't know the answer (e.g., specific dog breed), provide **your best guess**.
- Respond matter-of-factly and **avoid using conversational language or inserting your opinion**.

Please answer the question using as few words as possible:

Q1: What is unusual about this mustache?

A1: Write your answer here.

Do you think you were able to answer the question correctly?

(Clicking an option will take you to the next question.)

no

maybe

yes

Page 1/2

Mainstream VQA Challenge (held for 6 years)

The screenshot shows the VQA workshop page from CVPR 2019. At the top, there's a red header bar with the VQA logo and "Visual Question Answering". To the right are logos for Virginia Tech and Georgia Tech. Below the header is a black navigation bar with links: Home, People, Code, Demo, Download, Evaluation, Challenge, Browse, Visualize, Workshop (which is highlighted in grey), Sponsors, Terms, and External.

The main content area has a white background. It features a large title "Visual Question Answering and Dialog Workshop" in bold black font. Below it is the location information: "Location: Seaside Ballroom B, Long Beach Convention & Entertainment Center" and "at CVPR 2019, June 17, Long Beach, California, USA".

On the left side, there's a diagram illustrating the AI System. It shows a photo of a woman with a yellow mustache, a question "What is the mustache made of?", and the AI System output "bananas".

On the right side, there's a cartoon illustration of two robots, Q-BOT and A-BOT, having a conversation. Q-BOT asks, "Two zebra are walking around their pen at the zoo." A-BOT responds, "A1: No, there aren't any." Q-BOT then asks, "Q2: Any other animal?" A-BOT responds, "A2: No, just zebras."

At the bottom, there's a blue link: <https://visualqa.org/workshop.html>.

VizWiz: Authentic Use Case



VizWiz: Authentic Use Case



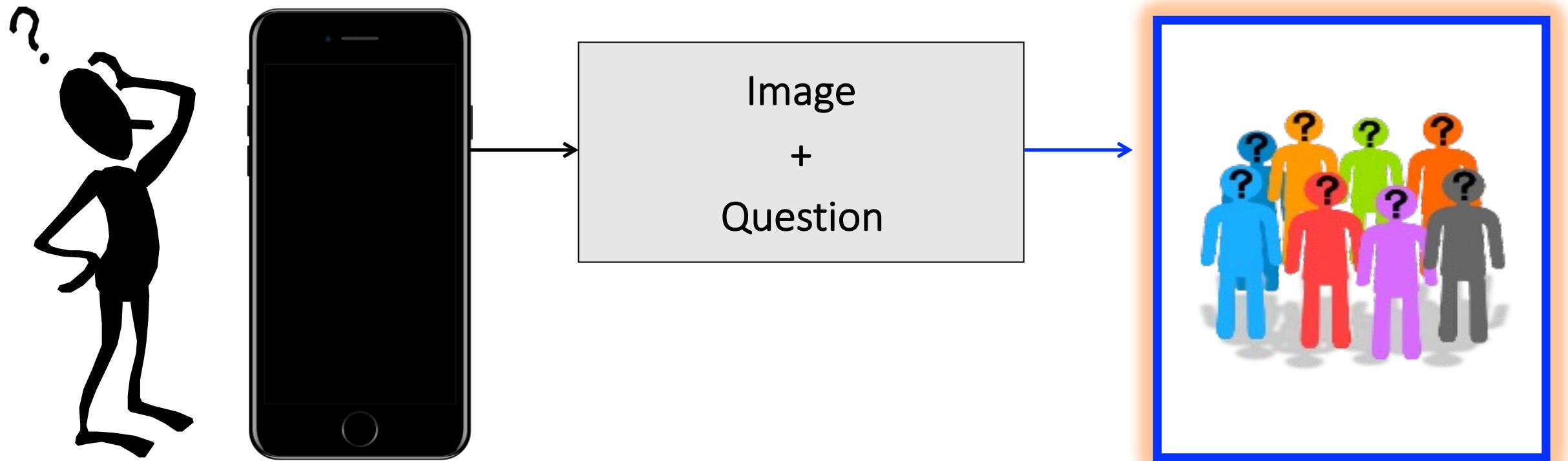
Image

VizWiz: Authentic Use Case

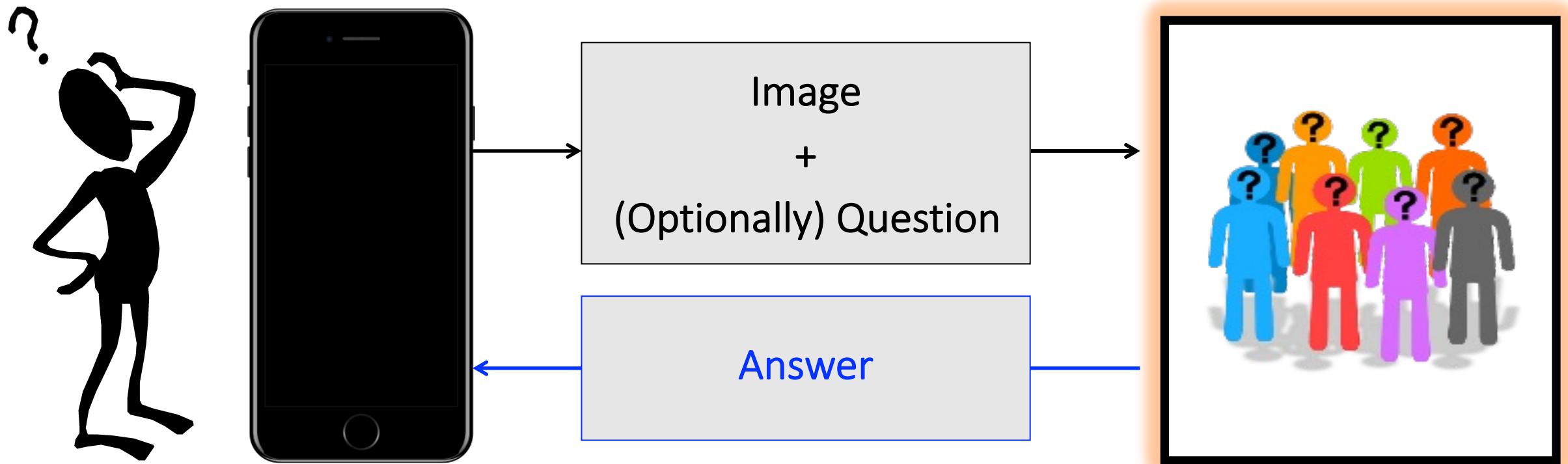


Image
+
Question

VizWiz: Authentic Use Case



VizWiz: Authentic Use Case



VizWiz: Authentic Use Case

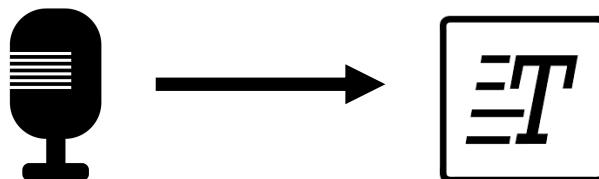


Users agreed to share **44,799 (62%)** of requests for dataset creation

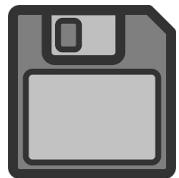
VizWiz: Authentic Use Case

Anonymization

1. Transcribe questions (removes voice)



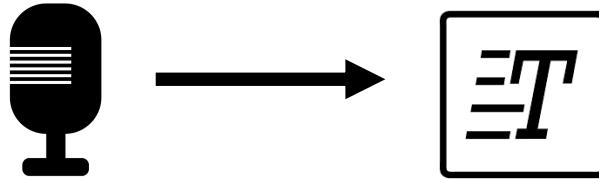
2. Re-save images (removes metadata)



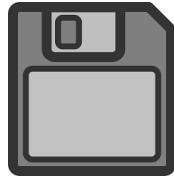
VizWiz: Authentic Use Case

Anonymization

1. Transcribe questions



2. Re-save images



In-House Filtering

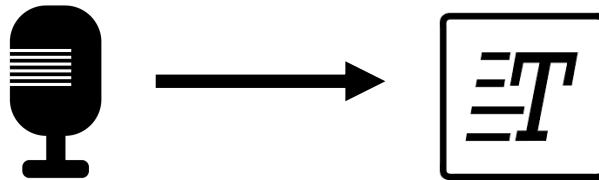
(personally identifying information)



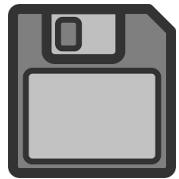
VizWiz: Authentic Use Case

Anonymization

1. Transcribe questions



2. Re-save images



In-House Filtering

Data Labeling (high quality answers)



VizWiz: Authentic Use Case

VQA: 32,842 image/question pairs → 328,420 answers

VizWiz-VQA Grand Challenge (4th year in 2022)



Home Browse Dataset Tasks & Datasets ▾ Workshops ▾ Acknowledgments

2022 VizWiz Grand Challenge Workshop

Visual Question Answering



Q: Does this foundation have any sunscreen?
A: yes



Q: What is this?
A: 10 euros



Q: What color is this?
A: green



Q: What type of pills are these?
A: unsuitable image



Q: What type of soup is this?
A: unsuitable image



Q: Who is this mail for?
A: unanswerable

Answering Grounding



Q: What is this?
A: dog



Q: What does the package say?
A: burrito

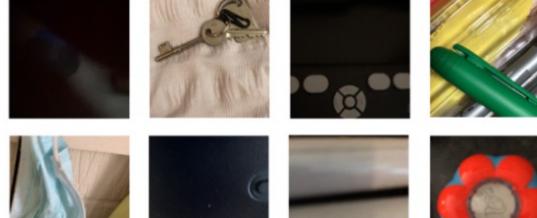


Q: What is this?
A: crystal



Q: How many tablets in this box?
A: 8

Few-Shot Object Recognition



67
Blind and low-vision collectors

486
Objects

3,822
Videos

2,687,934
Frames

Difference of Status Quo from a Real-World Use Case

Widely-Used Application

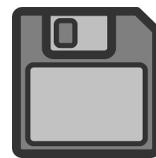


Anonymization

1. Transcribe questions



2. Re-save images



In-House Filtering



Data Labeling

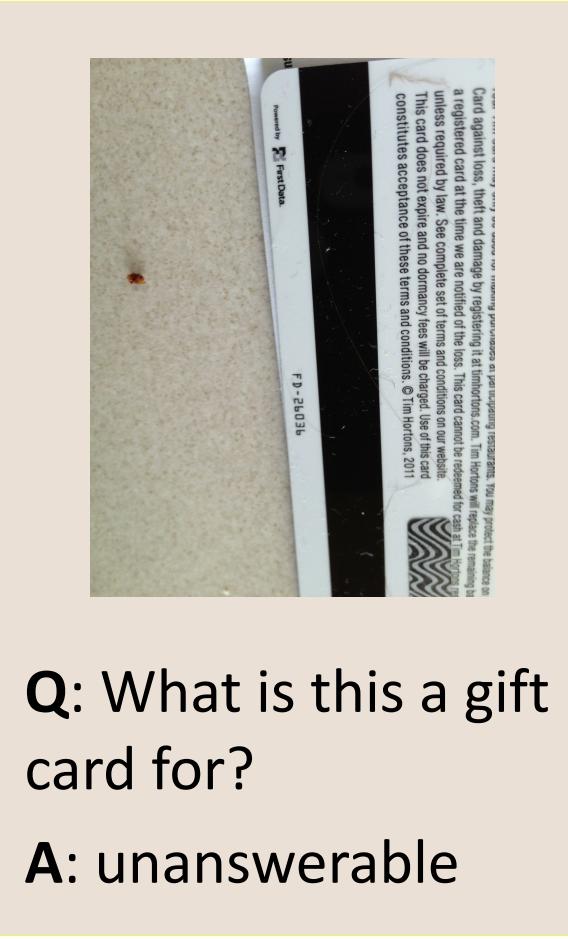


amazon mechanical turk
beta

Difference of a Real-World Use Case from Status Quo: VQs Can Be Unanswerable!



Q: What is the expiration date?
A: unanswerable



Q: What is this a gift card for?
A: unanswerable



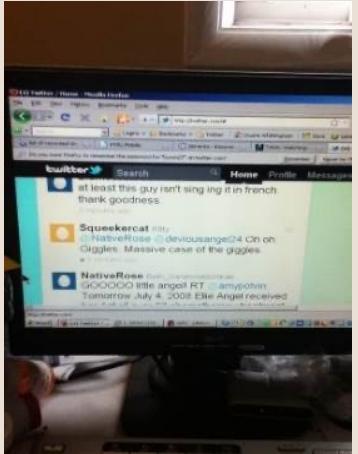
Q: What temperature is the dial set to?
A: unanswerable

Today's Topics

- Visual question answering applications
- Visual question answering datasets
- **Visual question answering evaluation**
- Mainstream challenge 2015 winner: baseline approach
- Mainstream challenge 2019 winner: transformer-based approach
- Programming tutorial

Class Task: Answer Visual Question

(<https://forms.gle/A58M1gm23QMckZBq6>)



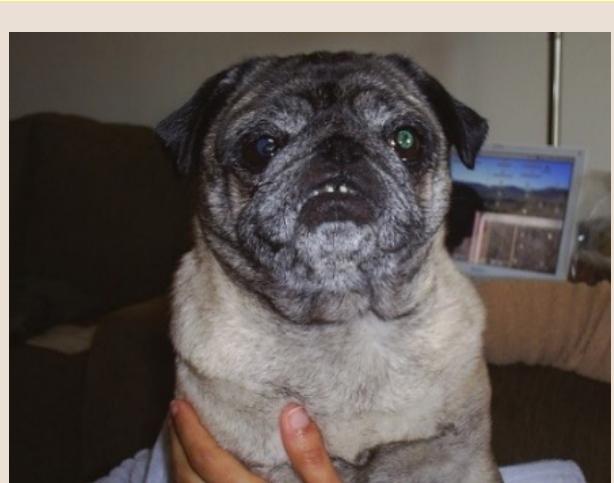
Is my
monitor on?

(1)



Hi there can you
please tell me
what flavor this is?

(2)



Does this picture
look scary?

(3)



Which side of
the room is the
toilet on?

(4)

Crowdsourced Answers



Is my monitor on?

- (1) yes
- (2) yes
- (3) yes
- (4) yes
- (5) yes
- (6) yes
- (7) yes
- (8) yes
- (9) yes
- (10) yes



Hi there can you please tell me what flavor this is?

- (1) sweet pepper
- (2) sweet pepper
- (3) sweet pepper
- (4) sweet pepper
- (5) sweet pepper
- (6) sweet pepper
- (7) sweet pepper
- (8) sweet pepper
- (9) sweet pepper
- (10) sweet pepper



Does this picture look scary?

- (1) yes
- (2) no
- (3) no
- (4) yes
- (5) no
- (6) yes
- (7) yes
- (8) no
- (9) no
- (10) no



Which side of the room is the toilet on?

- (1) right
- (2) left
- (3) right
- (4) right
- (5) right
- (6) right
- (7) right side
- (8) right
- (9) center
- (10) right

Class Discussion

1. Why do different answers arise for a visual question?
2. All crowdworkers were restricted to US locations for many datasets. How might involving individuals from other backgrounds outside the US affect VQA datasets?
3. How would you decide what answer you use when different answers arise? Of note, a method must scale to efficiently support large datasets.

Automated Predictions

VQA: Ask any question about this image



Is this man thirsty?

Answer

Answer

Confidence

yes

0.8778

no

0.1211

6

0.0001

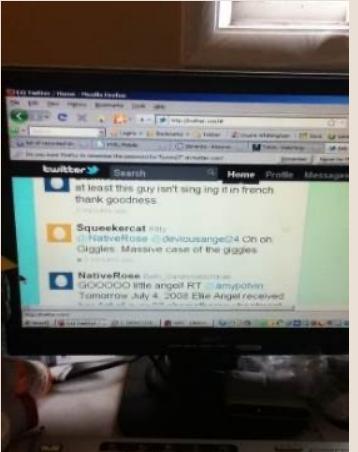
5

0.0001

pink

0.0001

Automated Predictions



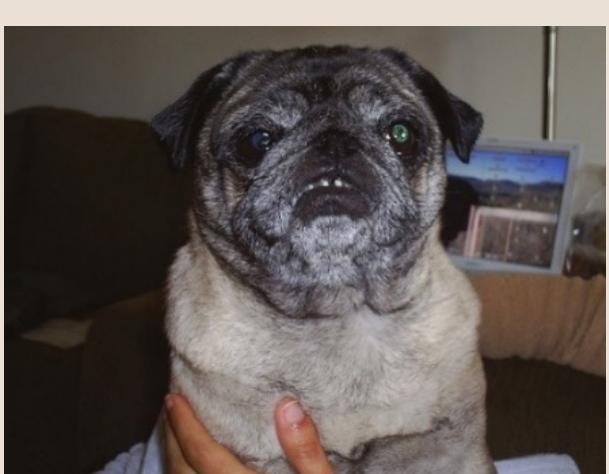
Is my
monitor on?

(1) yes



Hi there can you
please tell me
what flavor this is?

(2) chocolate



Does this picture
look scary?

(3) yes



Which side of
the room is the
toilet on?

(4) right

Evaluating Automated Predictions: Basic Equation

$$\text{accuracy} = \min\left(\frac{\# \text{ humans that provided that answer}}{3}, 1\right)$$

Evaluating Automated Predictions: Example



Does this picture
look scary?

- (1) yes
- (2) no
- (3) no
- (4) yes
- (5) no
- (6) yes
- (7) yes
- (8) no
- (9) no
- (10) no

What is the accuracy of an algorithm prediction of

- “yes”?
- “no”?
- “maybe”?

$$\text{accuracy} = \min\left(\frac{\# \text{ humans that provided that answer}}{3}, 1\right)$$

Evaluating Automated Predictions: Example



Which side of the room is the toilet on?

- (1) right
- (2) left
- (3) right
- (4) right
- (5) right
- (6) right
- (7) right side
- (8) right
- (9) center
- (10) right

What is the accuracy of an algorithm prediction of

- “right”?
- “left”?
- “right side”?
- “center”?
- “bottom”?

$$\text{accuracy} = \min\left(\frac{\# \text{ humans that provided that answer}}{3}, 1\right)$$

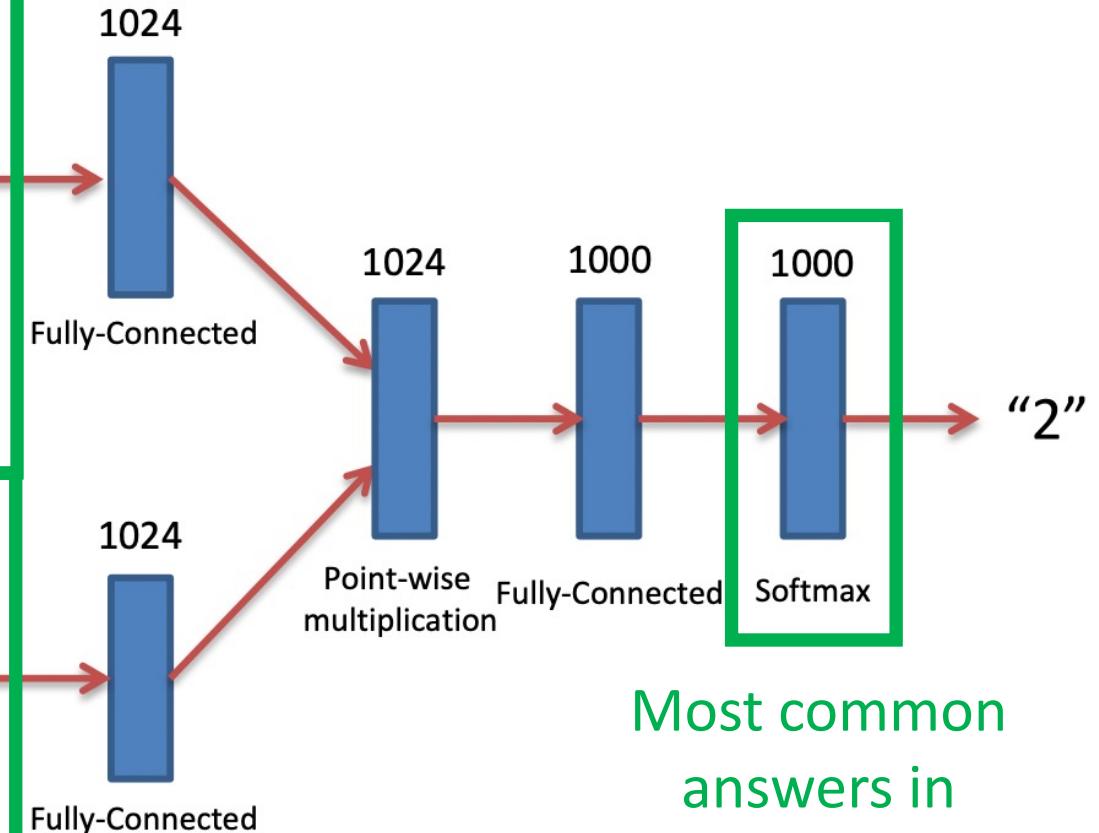
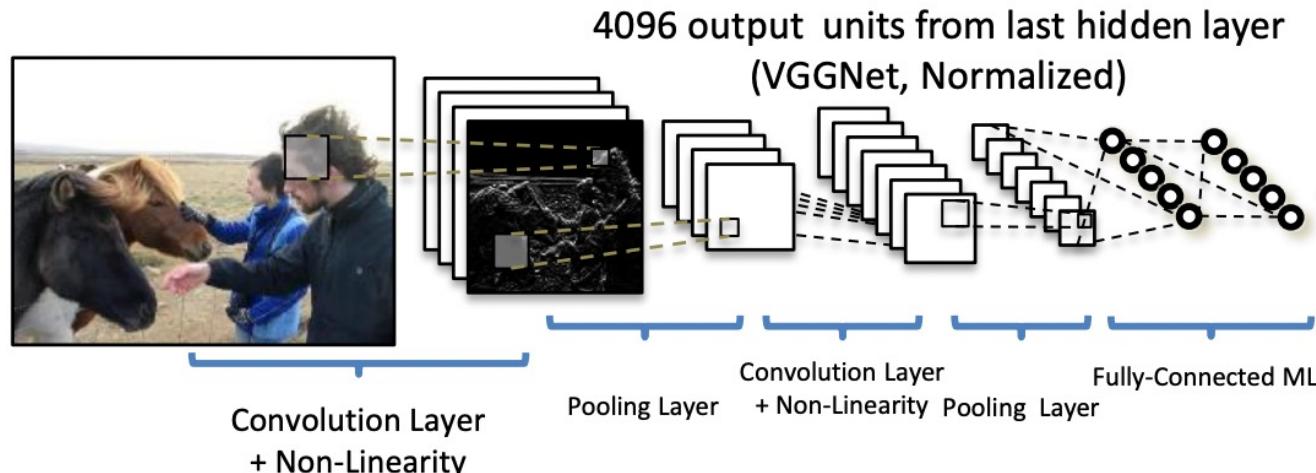
Implementation detail: to enable fair comparison to humans, compute 10 rounds of comparing a prediction with each possible set of 9 human-supplied answers

Today's Topics

- Visual question answering applications
- Visual question answering datasets
- Visual question answering evaluation
- **Mainstream challenge 2015 winner: baseline approach**
- Mainstream challenge 2019 winner: transformer-based approach
- Programming tutorial

Architecture

Image representation



Most common
answers in
train+val splits

"How many horses are in this image?"

Question representation: concatenates cell and hidden states of last hidden layer

Experimental Results (Fine-Grained Analysis with Respect to Answer Type)

All	Yes/No	Number	Other
57.75	80.50	36.77	43.08

On which answer type, does the model achieve the best performance?

Experimental Results (Fine-Grained Analysis with Respect to Answer Type)

All	Yes/No	Number	Other
57.75	80.50	36.77	43.08

On which answer type, does the model achieve the worst performance?

Experimental Results (Fine-Grained Analysis with Respect to Answer Type)

All	Yes/No	Number	Other
57.75	80.50	36.77	43.08

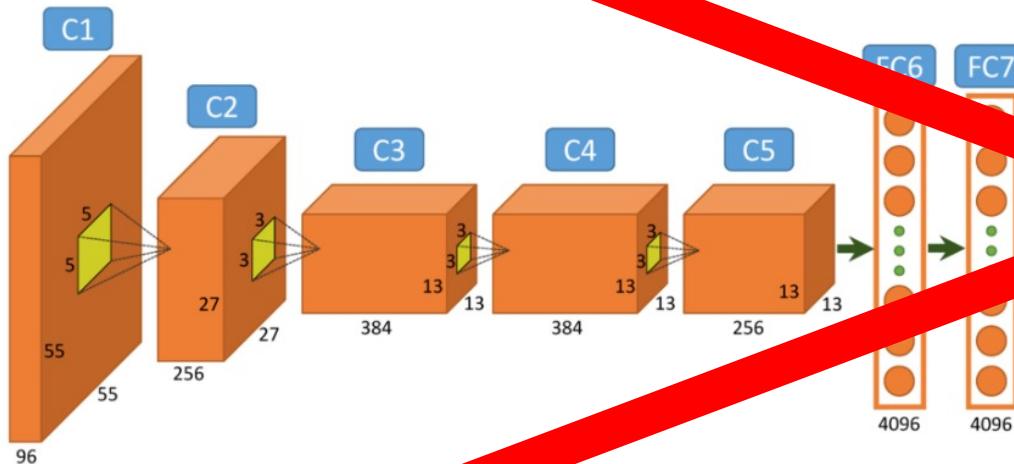
Why might we observe the above performance trends for answer types?

Today's Topics

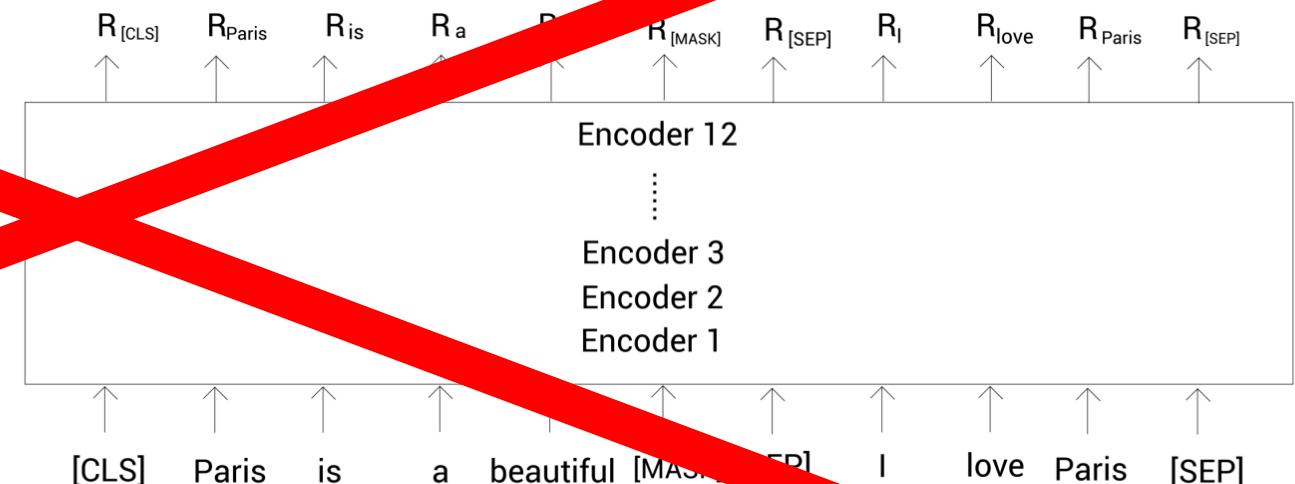
- Visual question answering applications
- Visual question answering datasets
- Visual question answering evaluation
- Mainstream challenge 2015 winner: baseline approach
- Mainstream challenge 2019 winner: transformer-based approach
- Programming tutorial

Key Idea: Multimodal Representation Rather Than Single Modality Representations

e.g., visual representation with AlexNet



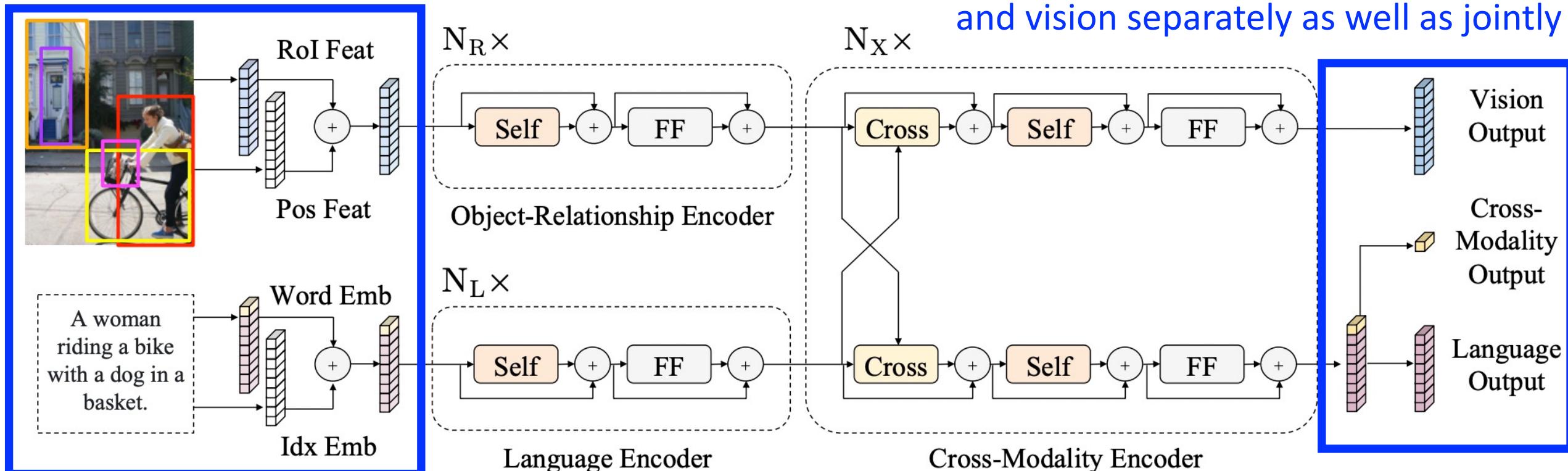
e.g., language representation with BERT



https://www.researchgate.net/figure/Architecture-of-Alexnet-From-left-to-right-left-to-output-five-convolutional-layers_fig2_312303454

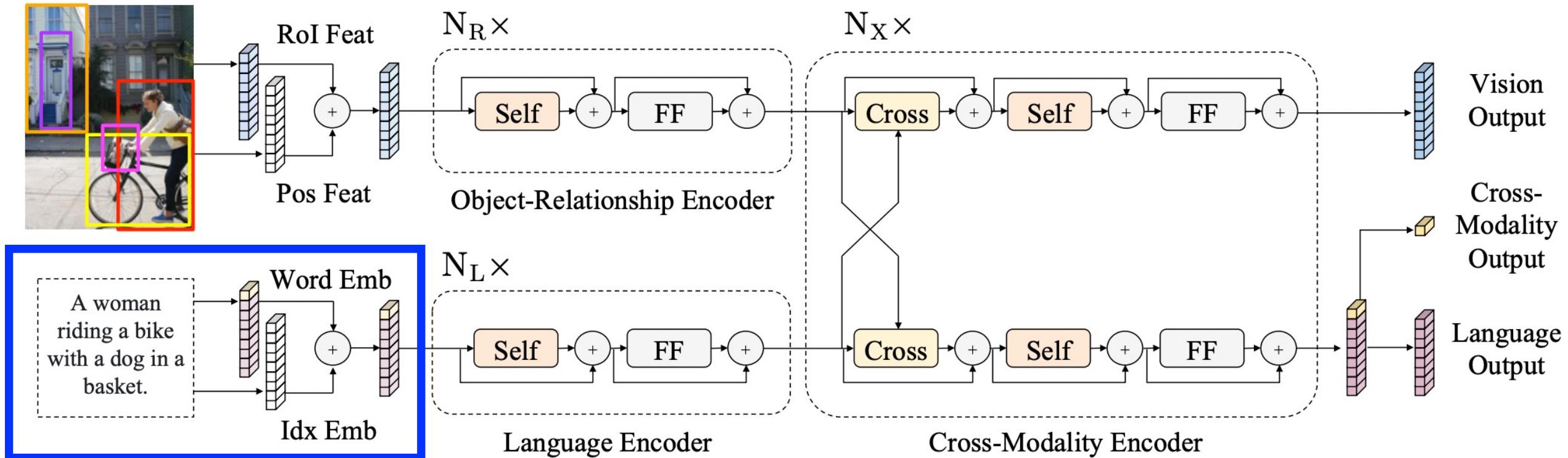
<https://static.packt-cdn.com/downloads/9781838821537-ColorImages.pdf>

LXMERT: Learning Cross-Modality Encoder Representations from Transformers



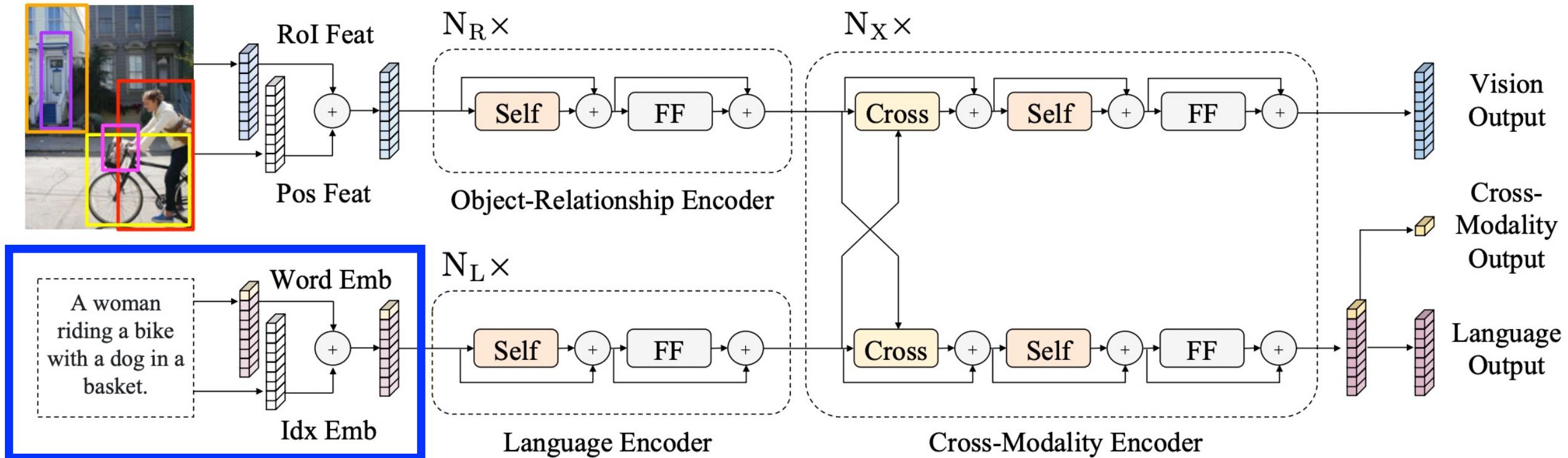
Pretrains using language and vision input

LXMERT: Language Input



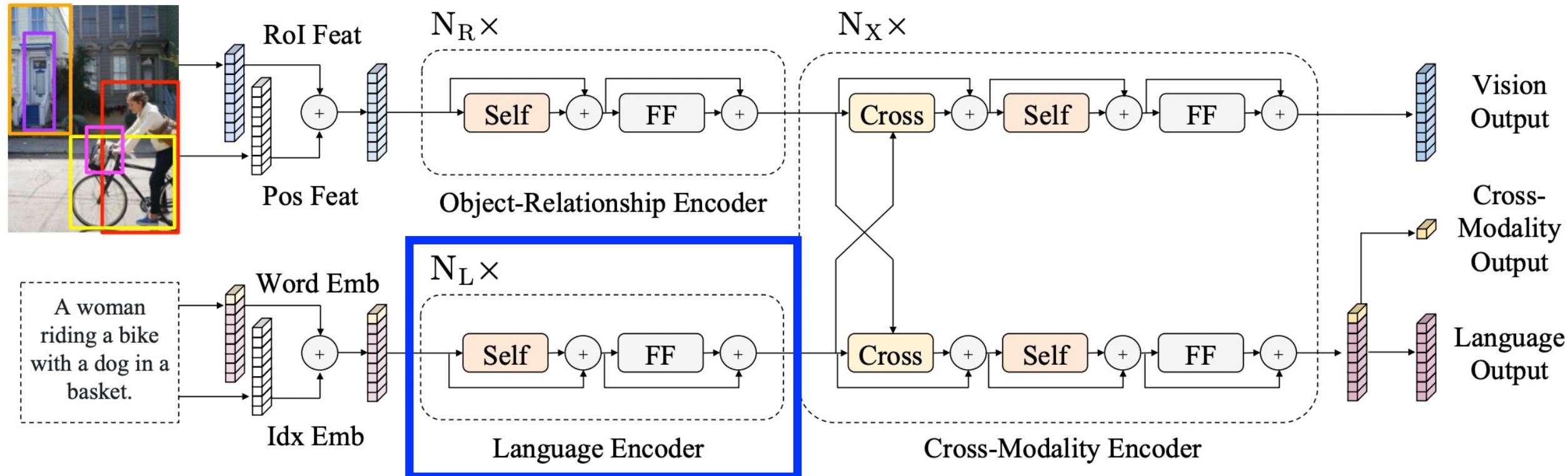
[CLS] is added to the start of the sequence

LXMERT: Language Input



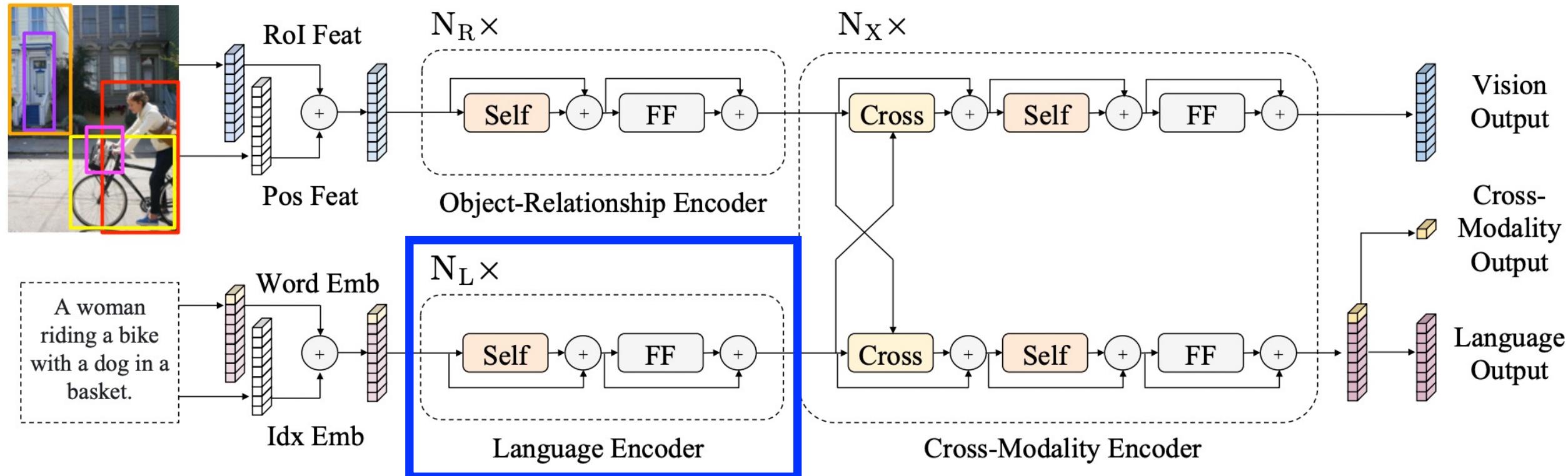
Each word is represented as sum of its
word embedding and position encoding

LXMERT: Language Input



Transformer encoder (i.e., BERT);
what does its output represent?

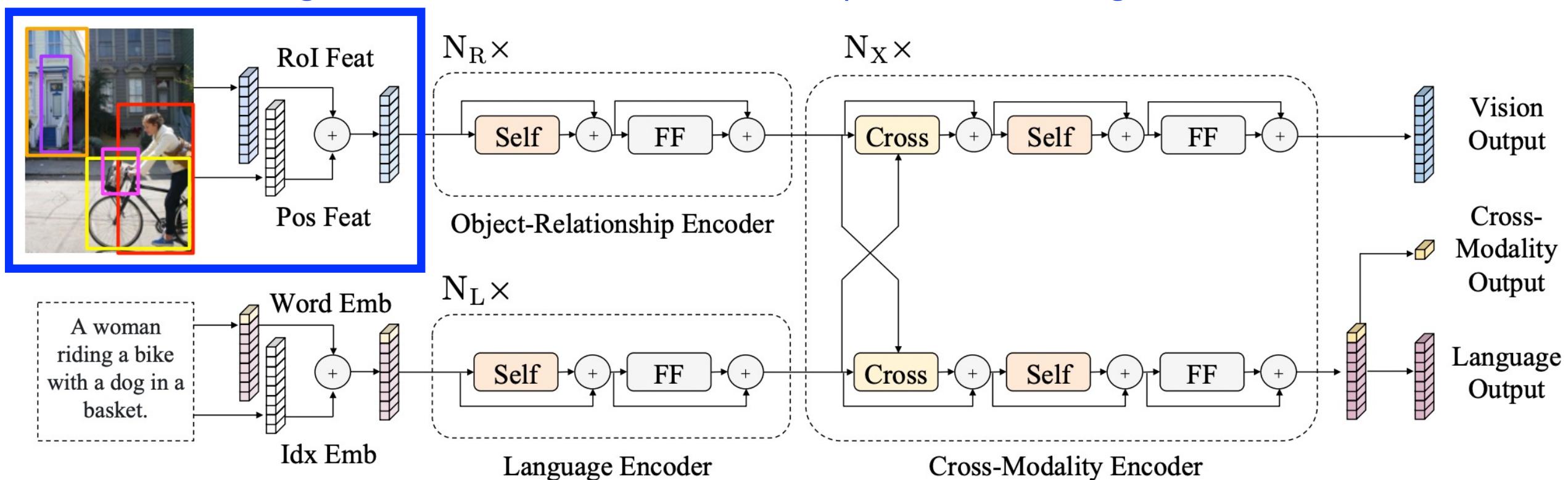
LXMERT: Language Input



Transformer encoder (i.e., BERT); represents words with their relationships to all words

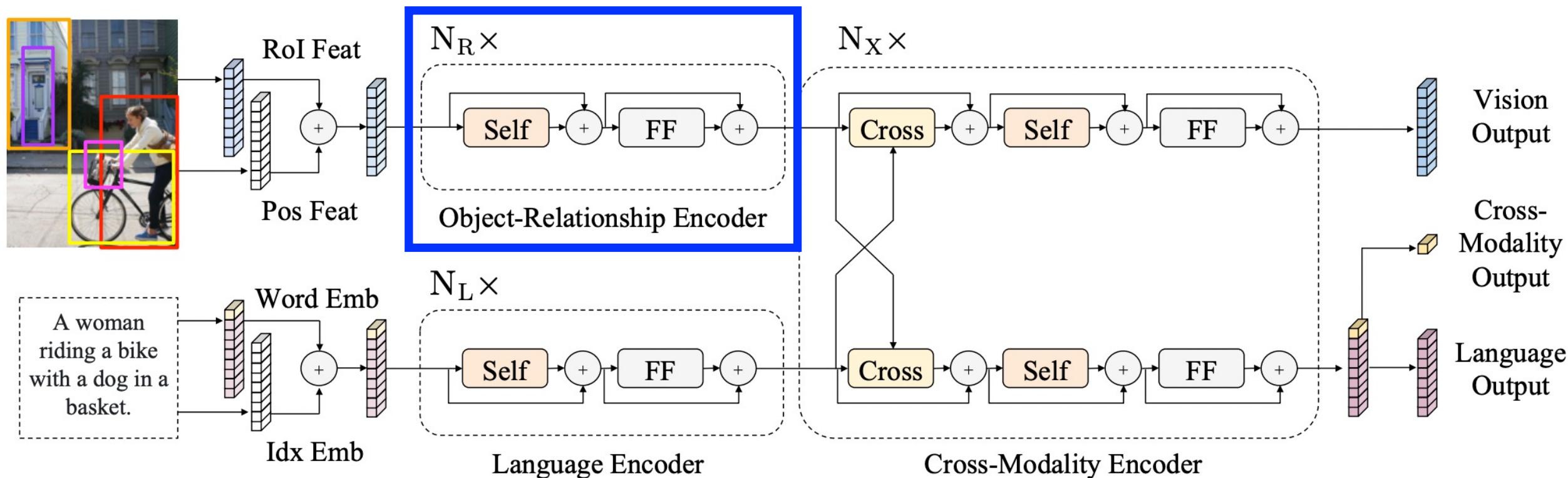
LXMERT: Vision Input

Each image is represented as a description of m objects detected with Faster R-CNN using features from Faster R-CNN and position encodings



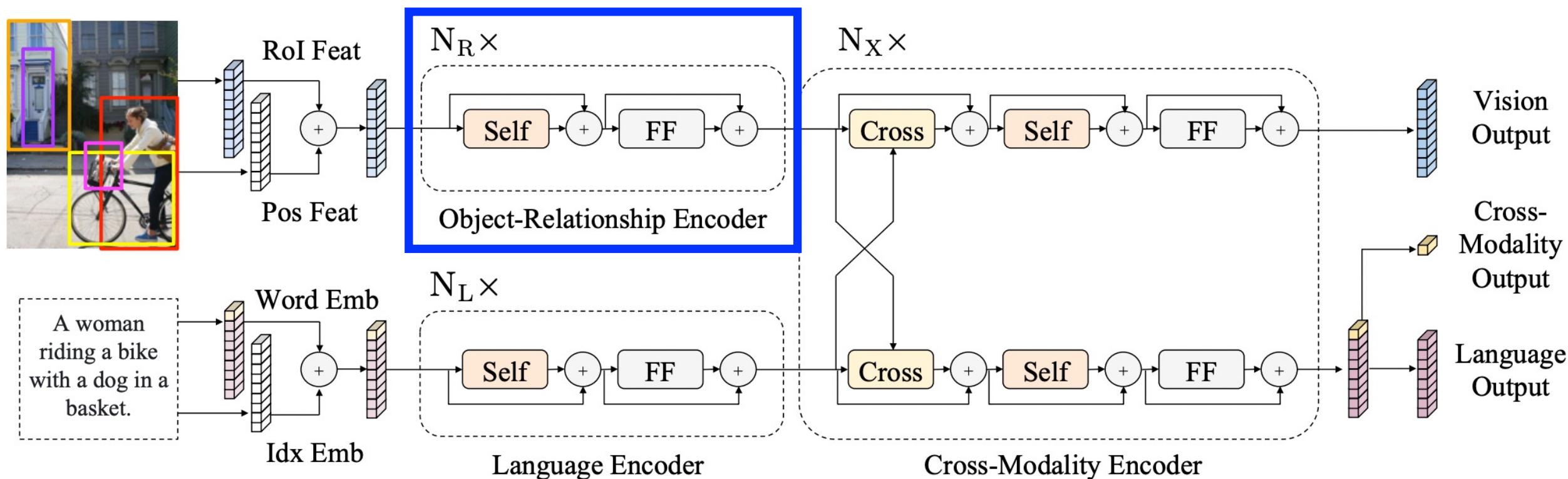
LXMERT: Architecture

Transformer encoder (i.e., BERT);
what does its output represent?

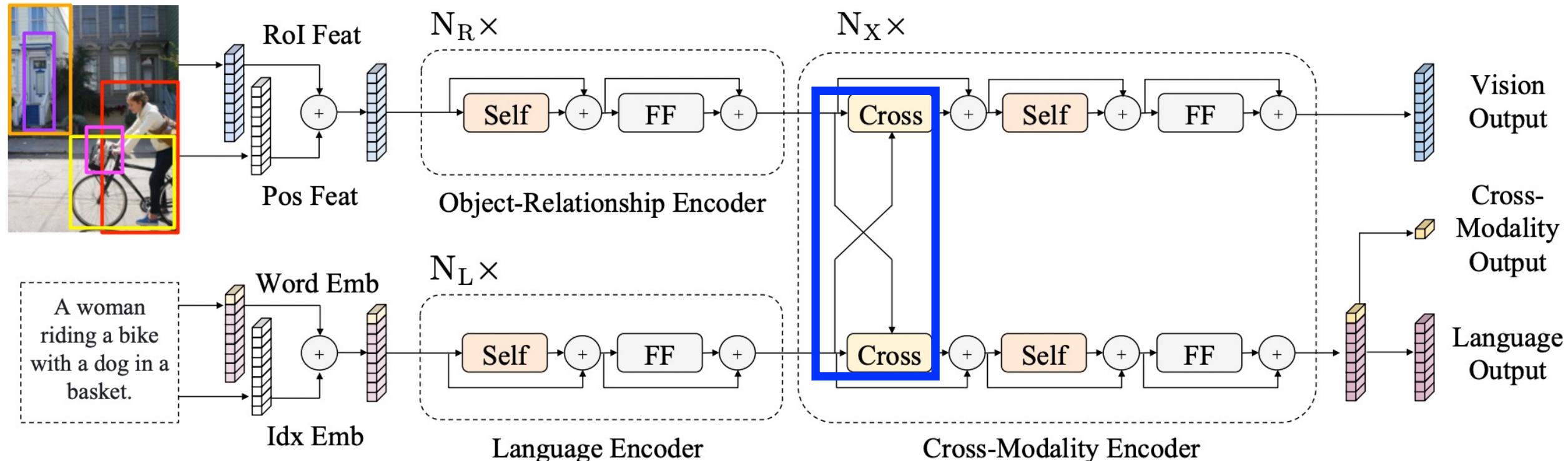


LXMERT: Architecture

Transformer encoder (i.e., BERT); represents objects with their relationships to all objects

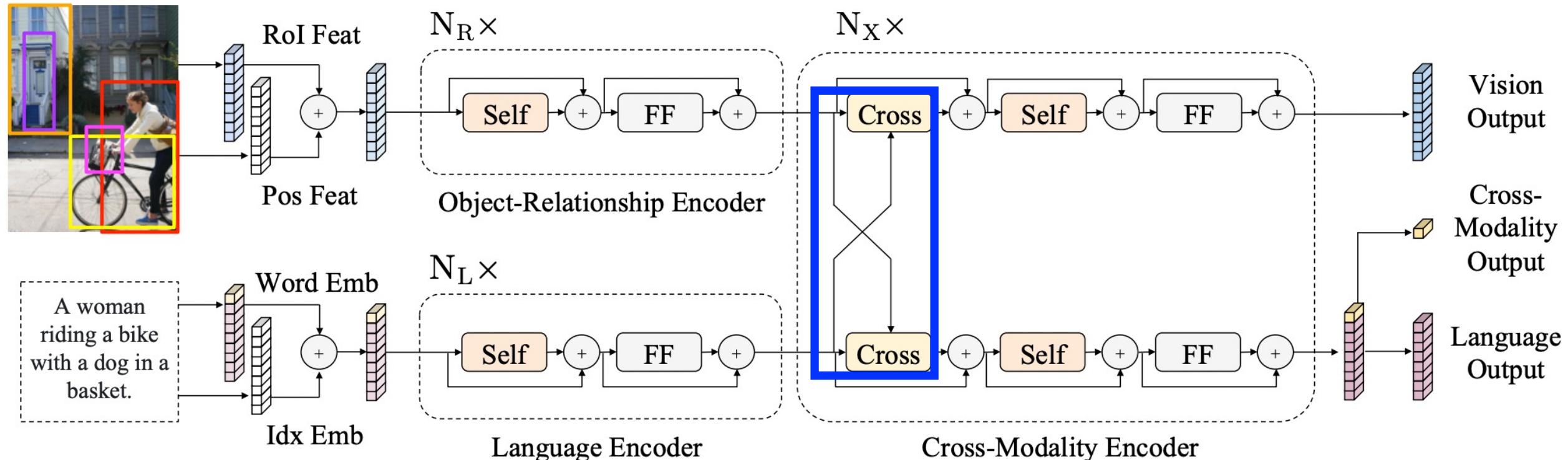


LXMERT: Architecture



Learns cross-modality representations
by aligning entities in the two modalities

LXMERT: Architecture

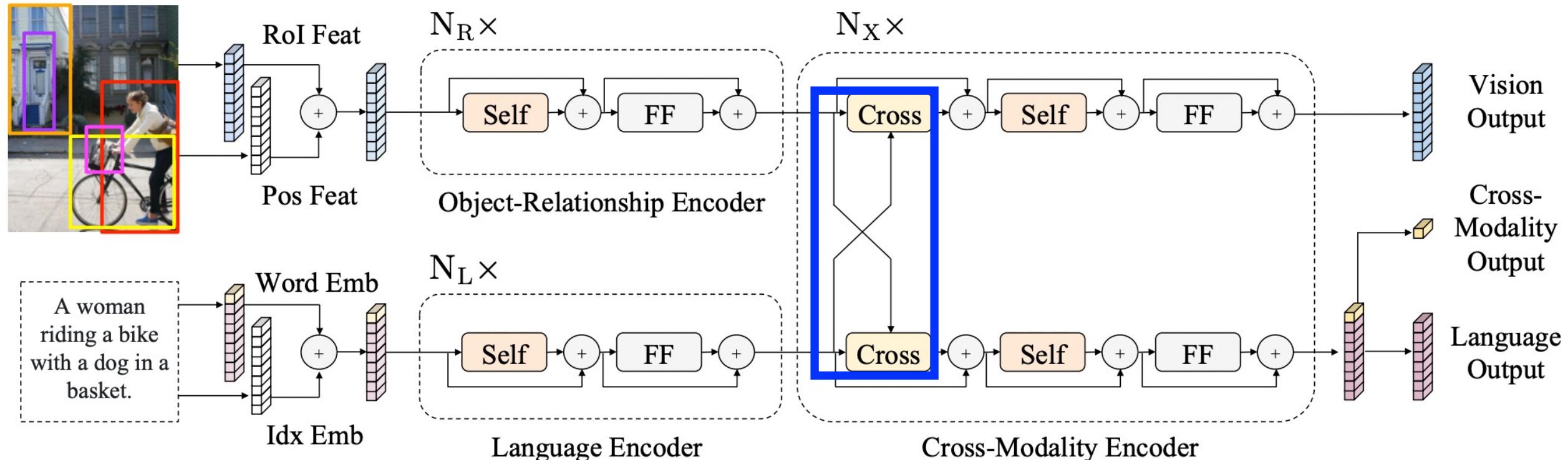


$$a_j = \text{score}(x, y_j)$$

$$\alpha_j = \exp(a_j) / \sum_k \exp(a_k)$$

Two cross-attention layers are functions of the “query” and “context” vectors which are language and vision features

LXMERT: Architecture



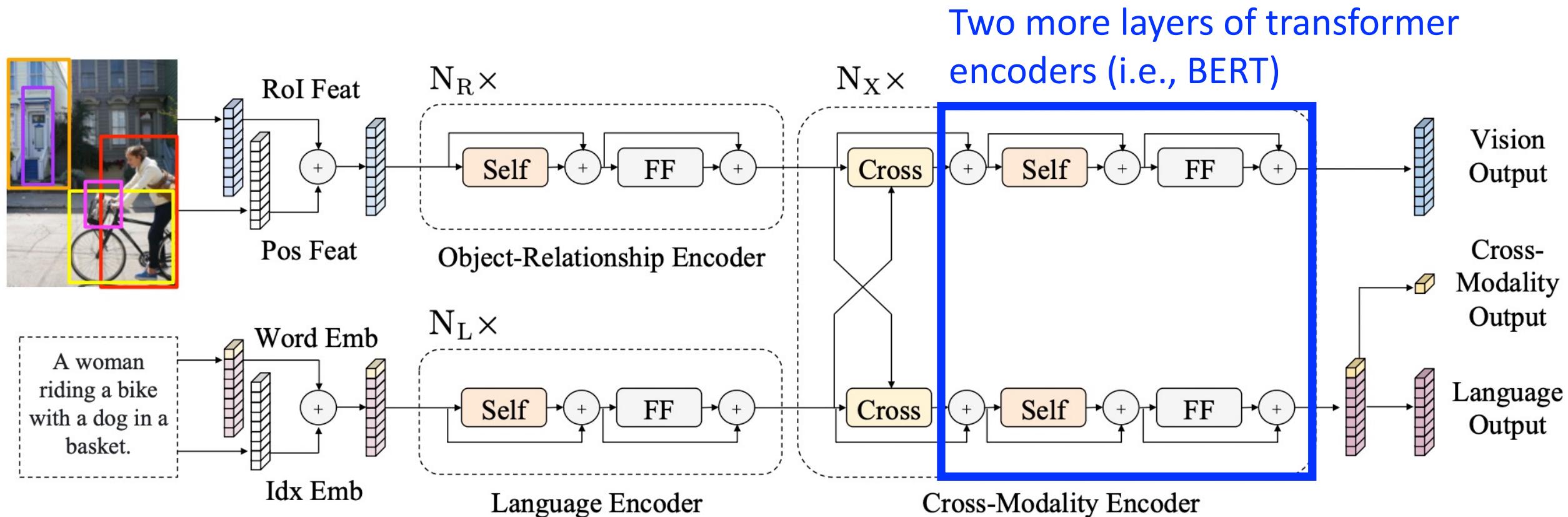
$$a_j = \text{score}(x, y_j)$$

$$\alpha_j = \exp(a_j) / \sum_k \exp(a_k)$$

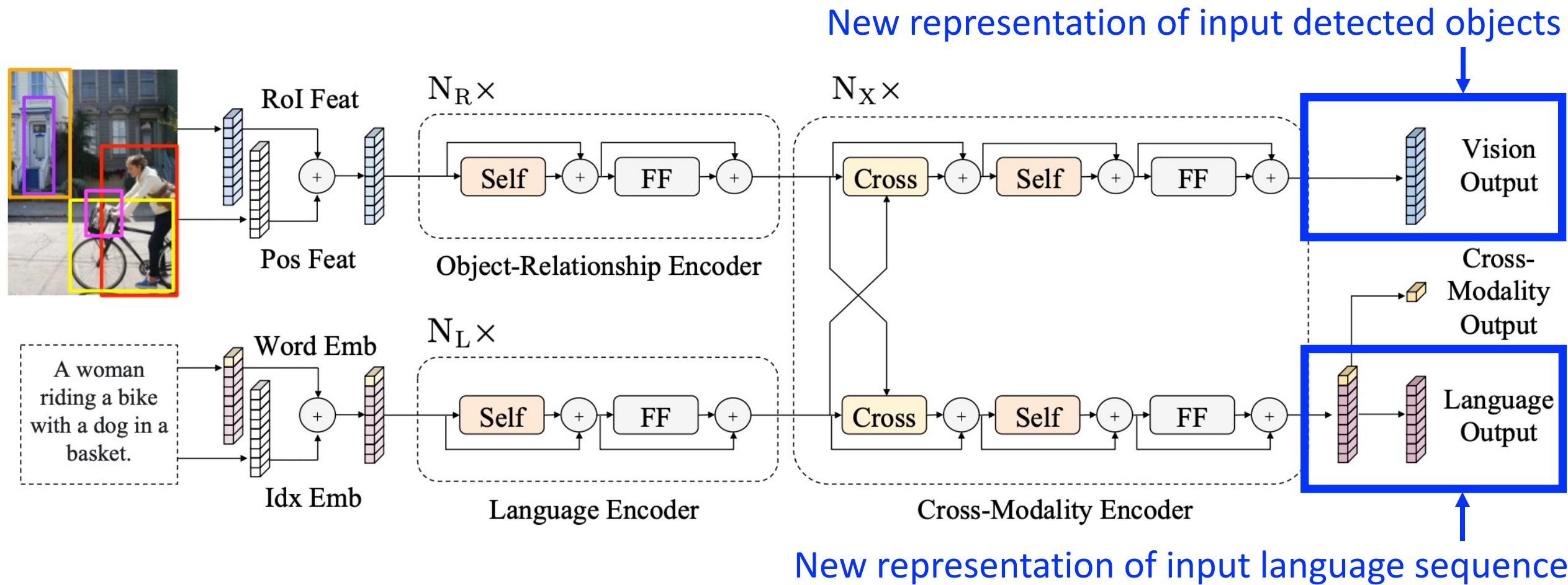
Output is weighted sum of context

vectors, weighted by the attention scores

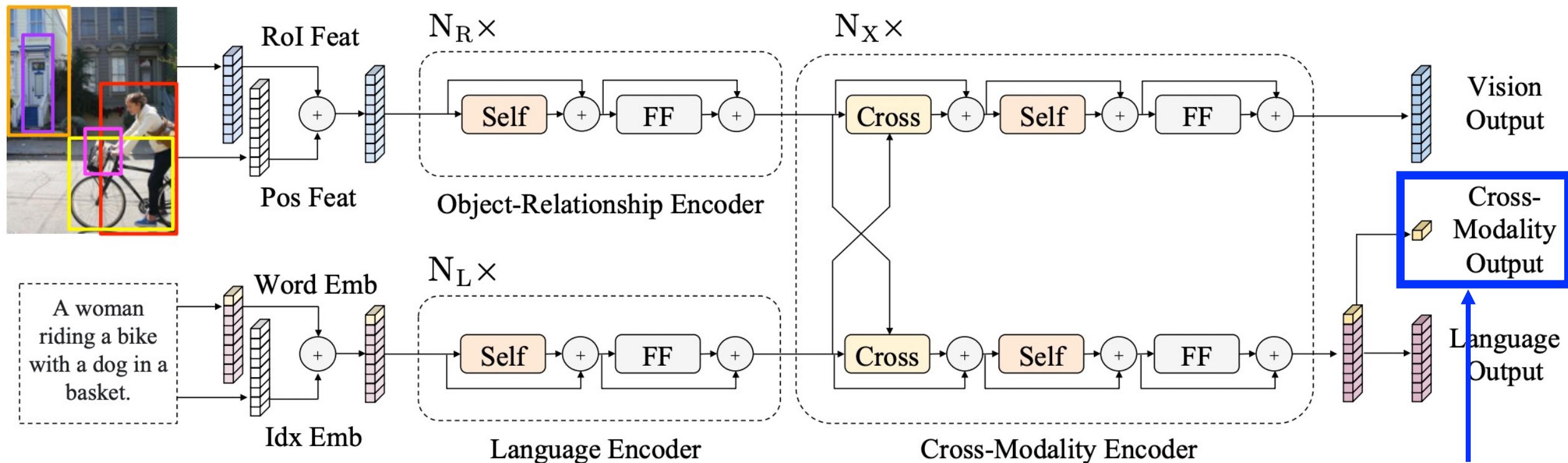
LXMERT: Architecture



LXMERT: Output



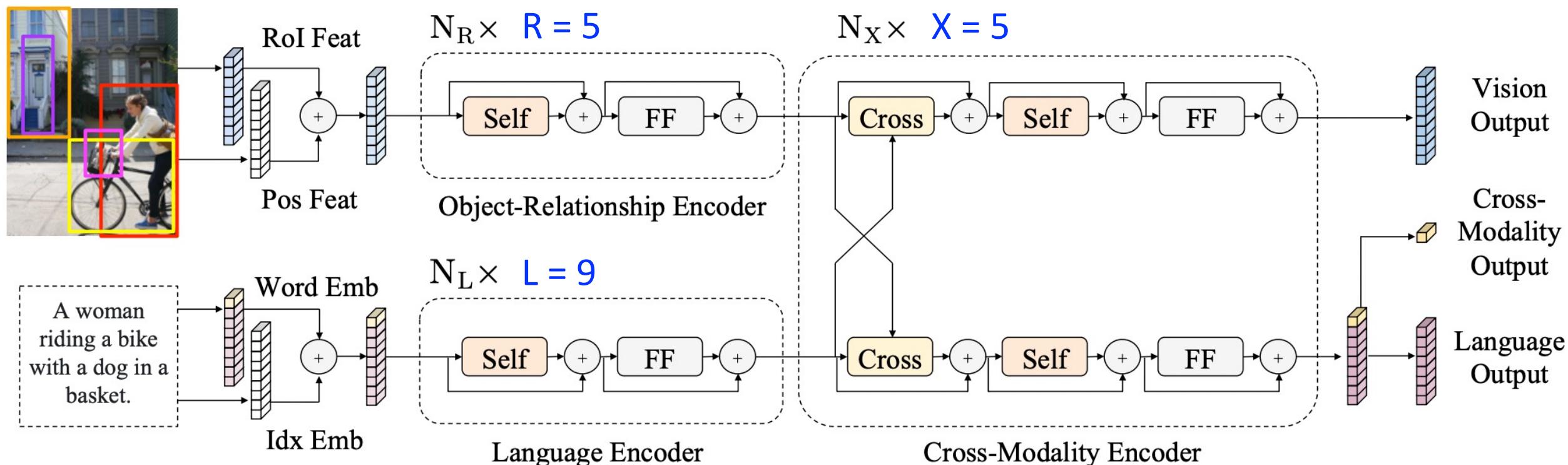
LXMERT: Output



Cross-modality representation is the [CLS] token appended at the start of the sentence

LXMERT: Implementation Details

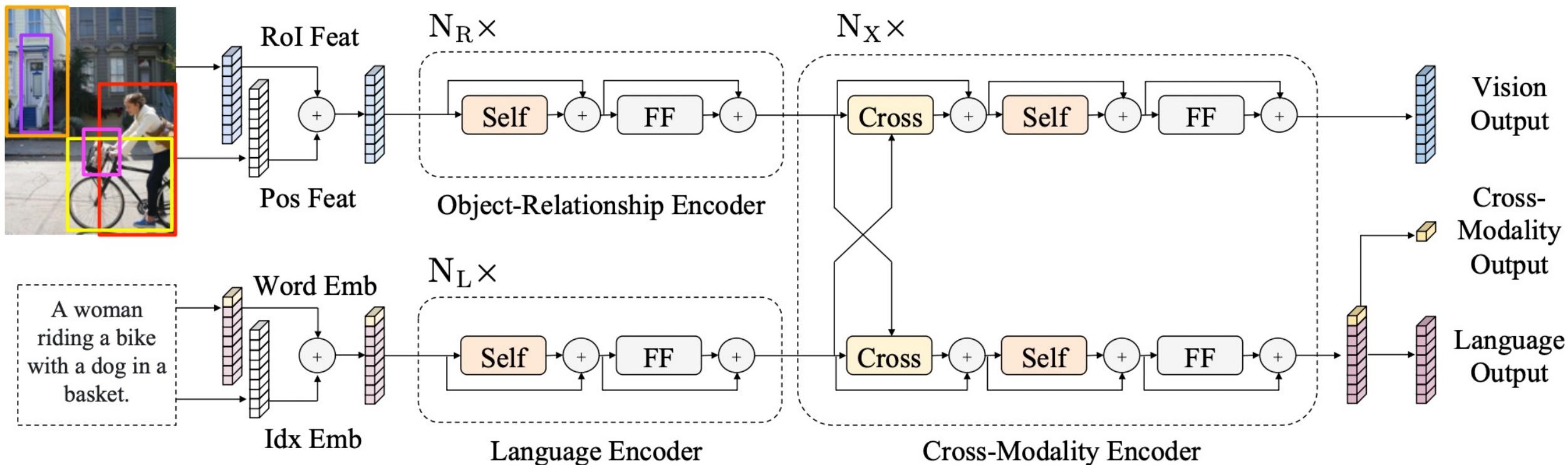
Pretrained Faster R-CNN can locate 1,600 categories
and only 36 object detections are kept per image



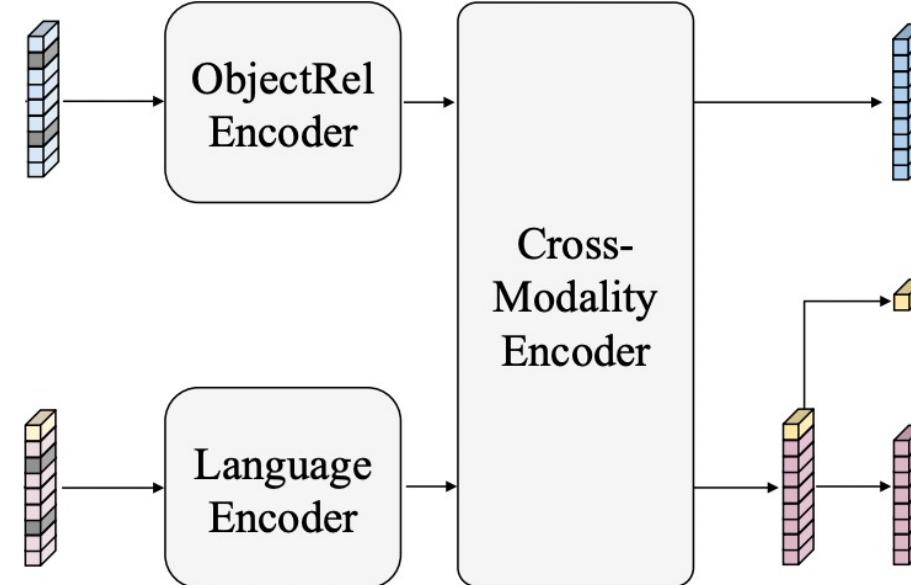
Number of layers mimics the size of BERT base of 12 layers;
i.e., $(5+9)/2 + 5$

LXMERT: Architecture

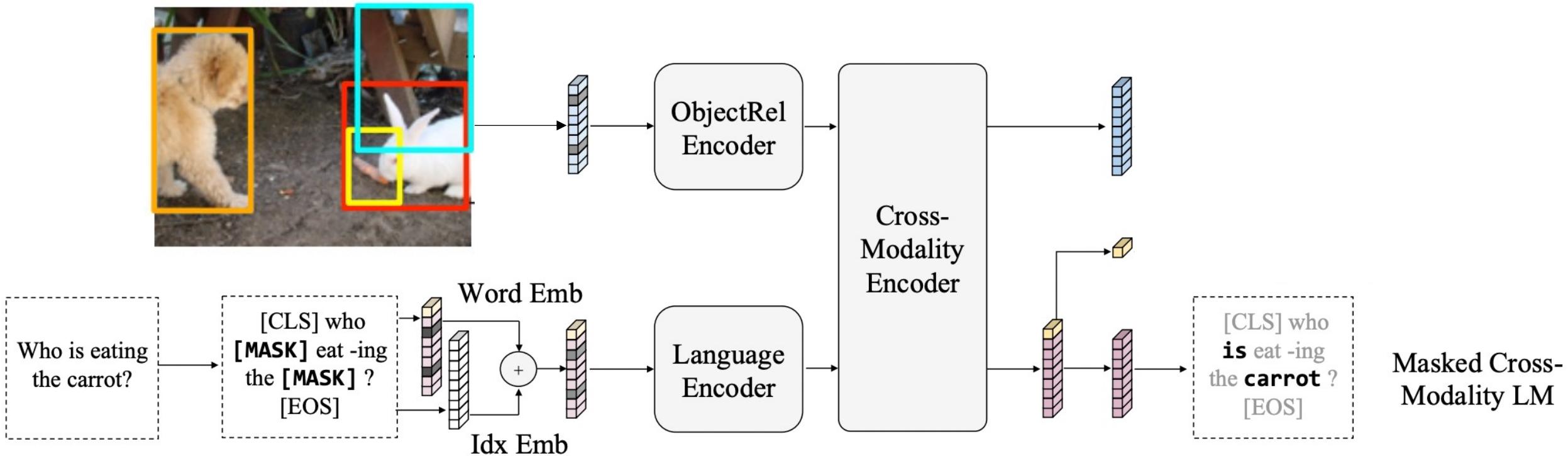
What might be strengths and limitations of the resulting feature representations based on the architecture used?



LXMERT: Summary of Architecture

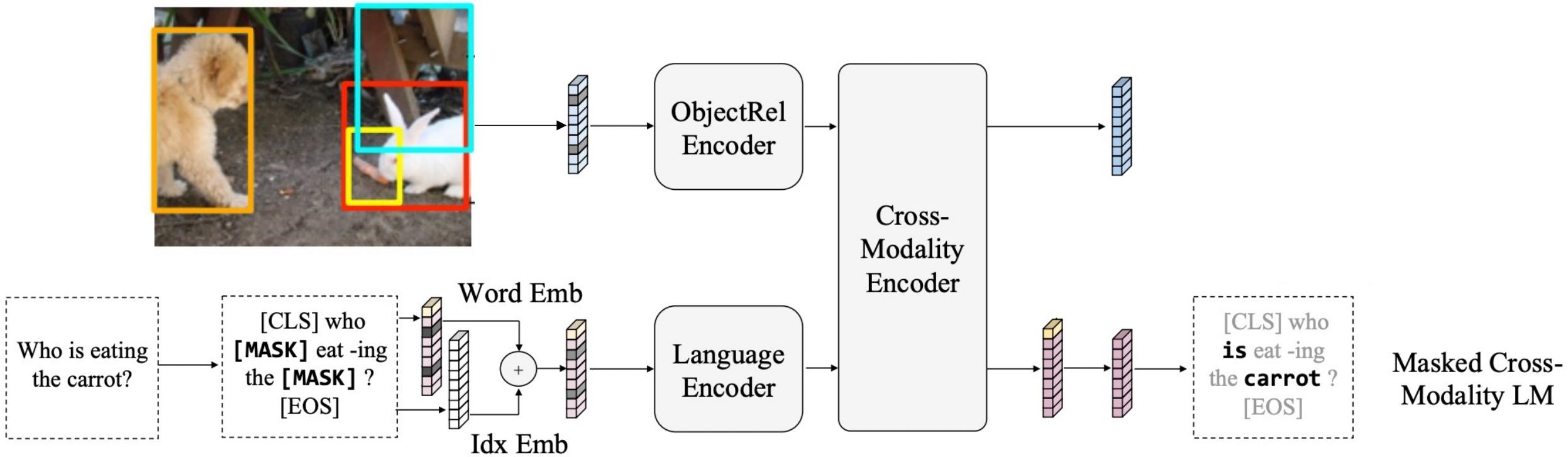


LXMERT: Pretraining Task 1 (Language)



Task used for BERT: mask 15% of input words and then predict them

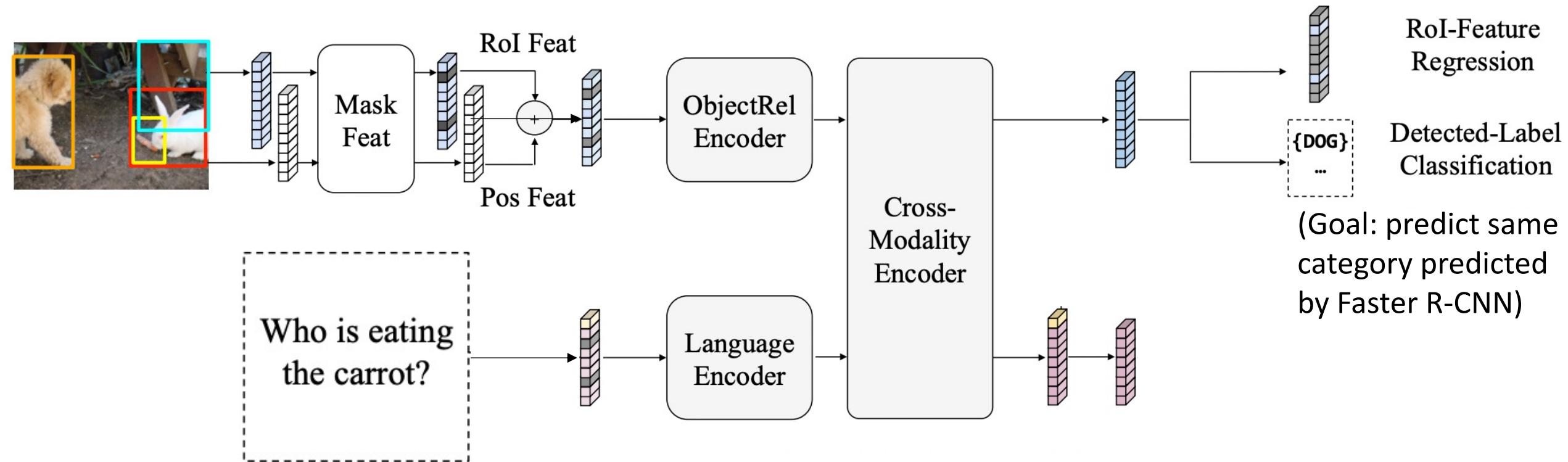
LXMERT: Pretraining Task 1 (Language)



Unlike BERT, vision modality can resolve language ambiguity; e.g., shows what is being eaten

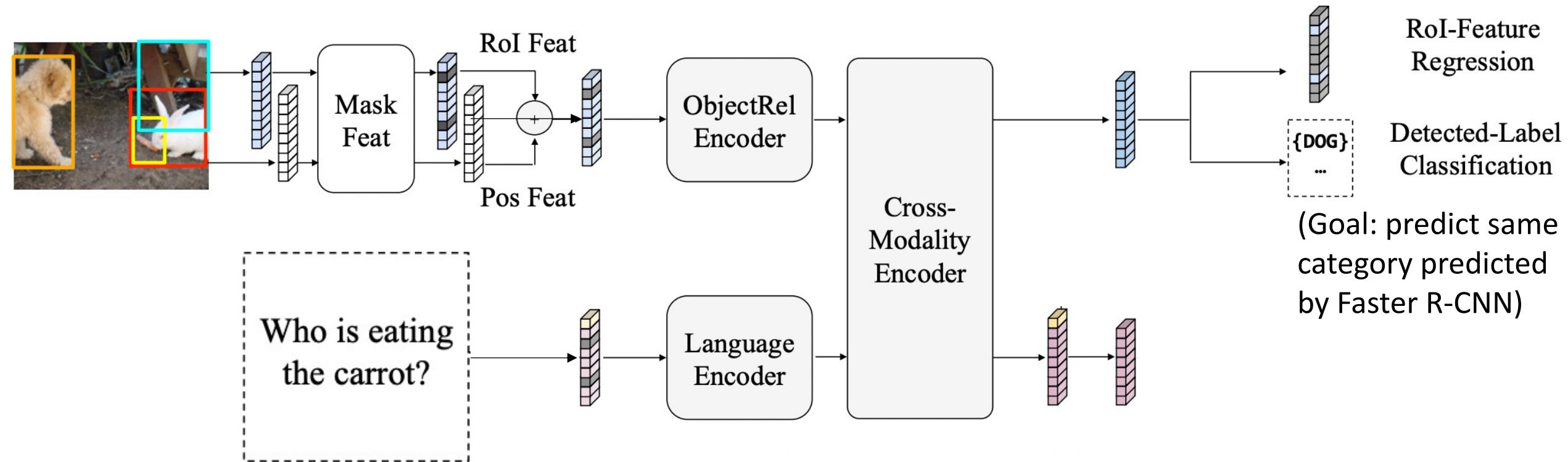
LXMERT: Pretraining Tasks 2 & 3 (Vision)

Mask 15% of input objects and then predict their original feature values and categories

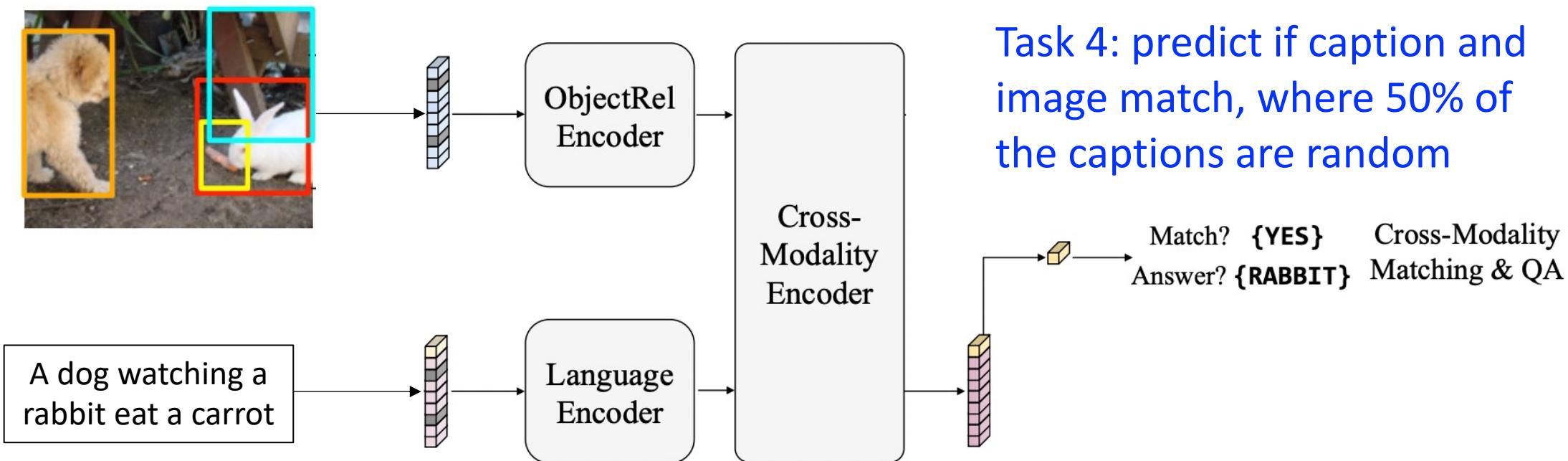


LXMERT: Pretraining Tasks 2 & 3 (Vision)

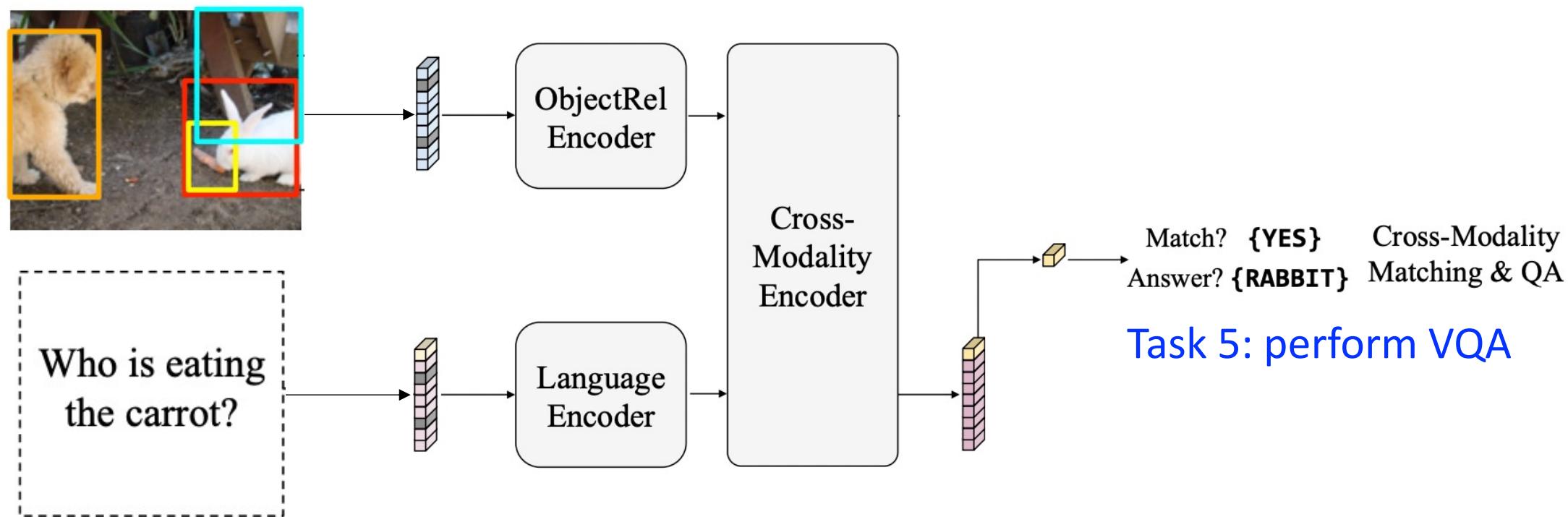
Knowledge about other objects and the language should help predict masked objects



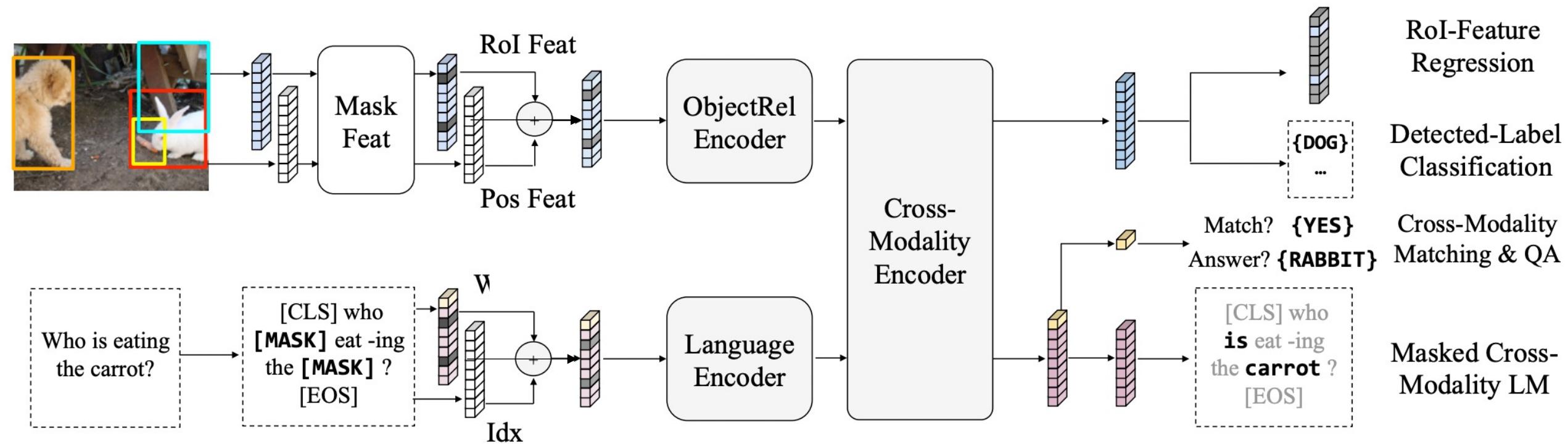
LXMERT: Pretraining Tasks 4 & 5 (Both Modalities)



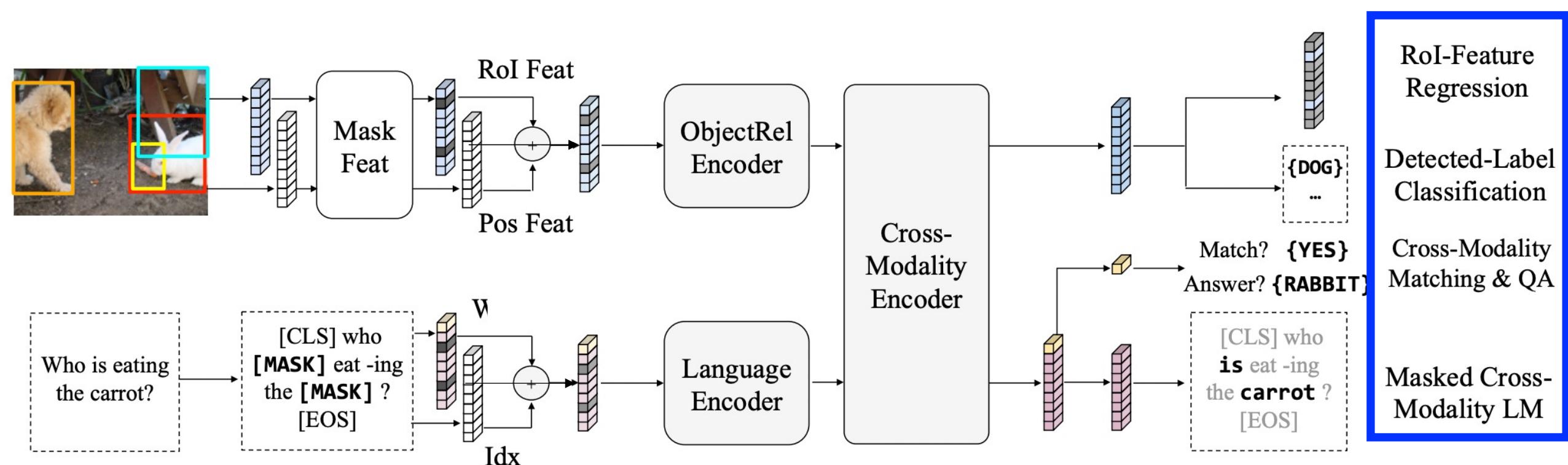
LXMERT: Pretraining Tasks 4 & 5 (Both Modalities)



LXMERT: 5 Pretraining Tasks



LXMERT: All Pretraining Task Losses Are Summed During Training



What might be strengths and limitations of the resulting feature representations based on the type of pretraining tasks used?

LXMERT: Training Data for Pretraining

Image Split	Images	Sentences (or Questions)					
		COCO-Cap	VG-Cap	VQA	GQA	VG-QA	All
MS COCO - VG	72K	361K	-	387K	-	-	0.75M
MS COCO ∩ VG	51K	256K	2.54M	271K	515K	724K	4.30M
VG - MS COCO	57K	-	2.85M	-	556K	718K	4.13M

All images are from two image sets, MS COCO and Visual Genome, which were collected by scraping images from the photo-sharing website Flickr
(Visual Genome includes the MS COCO images)

LXMERT: Training Data for Pretraining

Image Split	Images	Sentences (or Questions)					
		COCO-Cap	VG-Cap	VQA	GQA	VG-QA	All
MS COCO - VG	72K	361K	-	387K	-	-	0.75M
MS COCO \cap VG	51K	256K	2.54M	271K	515K	724K	4.30M
VG - MS COCO	57K	-	2.85M	-	556K	718K	4.13M

Language annotations came from 2 image captioning and 3 VQA datasets,
authored by crowdworkers paid to create captions, questions, and answers

LXMERT: Training Data for Pretraining

Image Split	Images	Sentences (or Questions)					
		COCO-Cap	VG-Cap	VQA	GQA	VG-QA	All
MS COCO - VG	72K	361K	-	387K	-	-	0.75M
MS COCO \cap VG	51K	256K	2.54M	271K	515K	724K	4.30M
VG - MS COCO	57K	-	2.85M	-	556K	718K	4.13M
All	180K	617K	5.39M	658K	1.07M	1.44M	9.18M

A total of 9.18M image-sentence pairs are included for 180,000 images
(questions in VQA datasets are used for the image-sentence pairs)

LXMERT: Training Data for Pretraining

Image Split	Images	Sentences (or Questions)					
		COCO-Cap	VG-Cap	VQA	GQA	VG-QA	All
MS COCO - VG	72K	361K	-	387K	-	-	0.75M
MS COCO \cap VG	51K	256K	2.54M	271K	515K	724K	4.30M
VG - MS COCO	57K	-	2.85M	-	556K	718K	4.13M
All	180K	617K	5.39M	658K	1.07M	1.44M	9.18M

What might be strengths and limitations of the resulting feature representations based on the type of training data that is used?

LXMERT: Fine-Tuning Experimental Results

Method	VQA				Accu
	Binary	Number	Other		
Human	-	-	-	-	-
Image Only	-	-	-	-	-
Language Only	66.8	31.8	27.6	44.3	
State-of-the-Art	85.8	53.7	60.7	70.4	
LXMERT	88.2	54.2	63.1	72.5	

Achieved the best performance, with stronger gains over prior work for questions that lead to “binary” and “other” answers

LXMERT: Fine-Tuning Experimental Results

Method	VQA				GQA			NLVR ²	
	Binary	Number	Other	Accu	Binary	Open	Accu	Cons	Accu
Human	-	-	-	-	91.2	87.4	89.3	-	96.3
Image Only	-	-	-	-	36.1	1.74	17.8	7.40	51.9
Language Only	66.8	31.8	27.6	44.3	61.9	22.7	41.1	4.20	51.1
State-of-the-Art	85.8	53.7	60.7	70.4	76.0	40.4	57.1	12.0	53.5
LXMERT	88.2	54.2	63.1	72.5	77.8	45.0	60.3	42.1	76.2

The representations also led to the best performance for an additional VQA dataset and a visual reasoning task (i.e., does statement describe two images or not)

Today's Topics

- Visual question answering applications
- Visual question answering datasets
- Visual question answering evaluation
- Mainstream challenge 2015 winner: baseline approach
- Mainstream challenge 2019 winner: transformer-based approach
- Programming tutorial

Today's Topics

- Visual question answering applications
- Visual question answering datasets
- Visual question answering evaluation
- Mainstream challenge 2015 winner: baseline approach
- Mainstream challenge 2019 winner: transformer-based approach
- Programming tutorial

The End