Vision-Language Tasks: Image Captioning & Visual Question Answering

Danna Gurari

University of Colorado Boulder

Spring 2025

https://dannagurari.colorado.edu/course/neural-networks-and-deep-learning-spring-2025/

Review

- Last lecture
  - Explosion of transformers
  - GPT
  - BERT
  - ViT
  - Programming tutorial
- Assignments (Canvas)
  - Problem set 4 due Tuesday
- Questions?

Today’s Topics

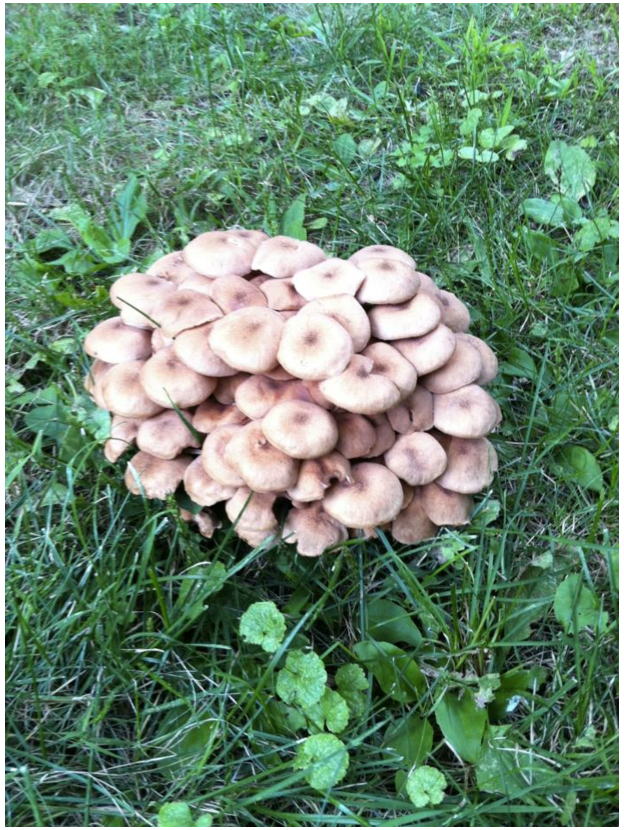
- Motivating applications
- Image captioning: pioneering dataset and model
- Visual question answering: pioneering dataset and model
- LXMERT: multimodal representations
- Programming tutorial

Today’s Topics

- Motivating applications
- Image captioning: pioneering dataset and model
- Visual question answering: pioneering dataset and model
- LXMERT: multimodal representations
- Programming tutorial

Multimodal Tasks: Uses 2+ Modalities

e.g., computer vision + natural language processing tasks



Caption:

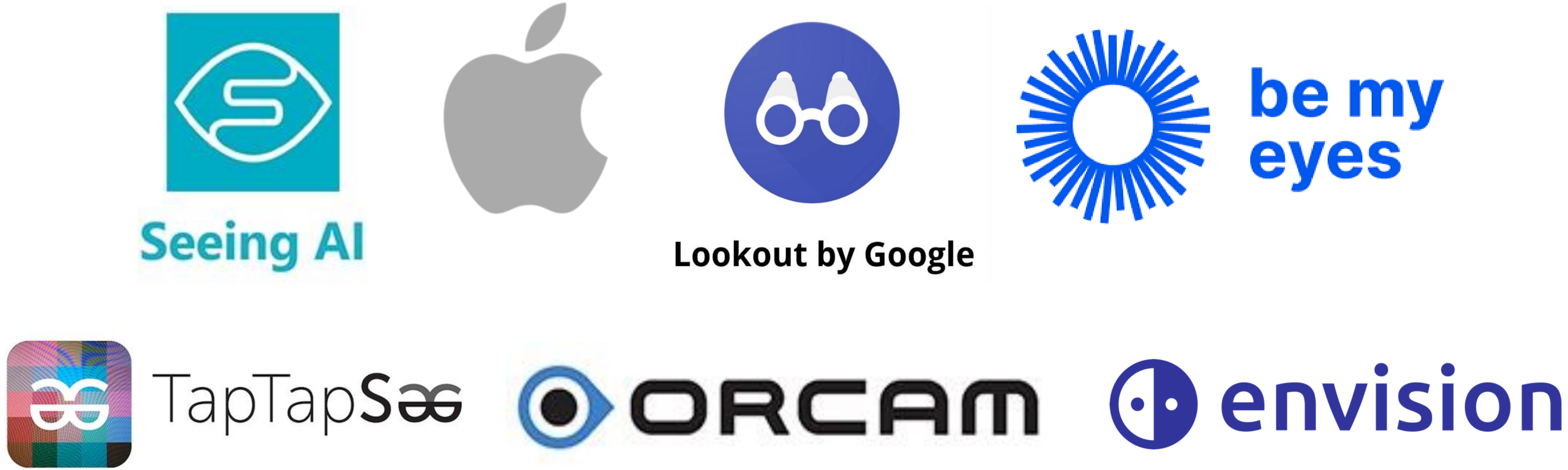
A bunch of small light brown mushrooms in a green field.

Answer Visual Question:

Q: Is it edible or poisonous?

A: Poisonous

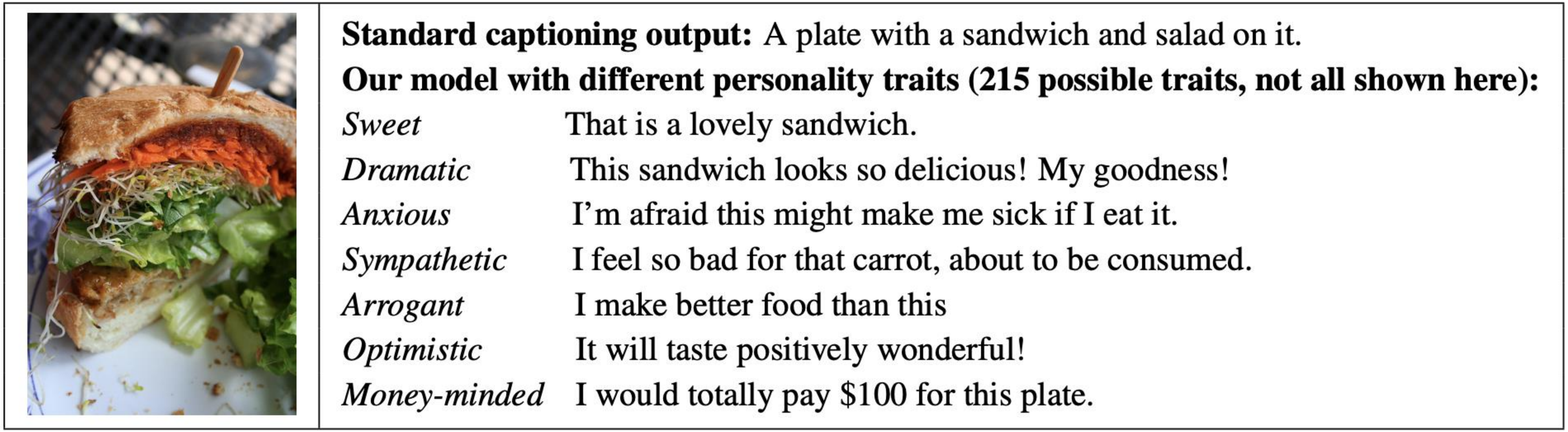
Applications: Visual Interpretation for People with Vision Loss; e.g.,



Applications: Visual Interpretation for People with Vision Loss; e.g.,

https://www.youtube.com/watch?v=cUSeFnZGIzY

Describing and Responding to Images Posted to Social Media with “Personality”



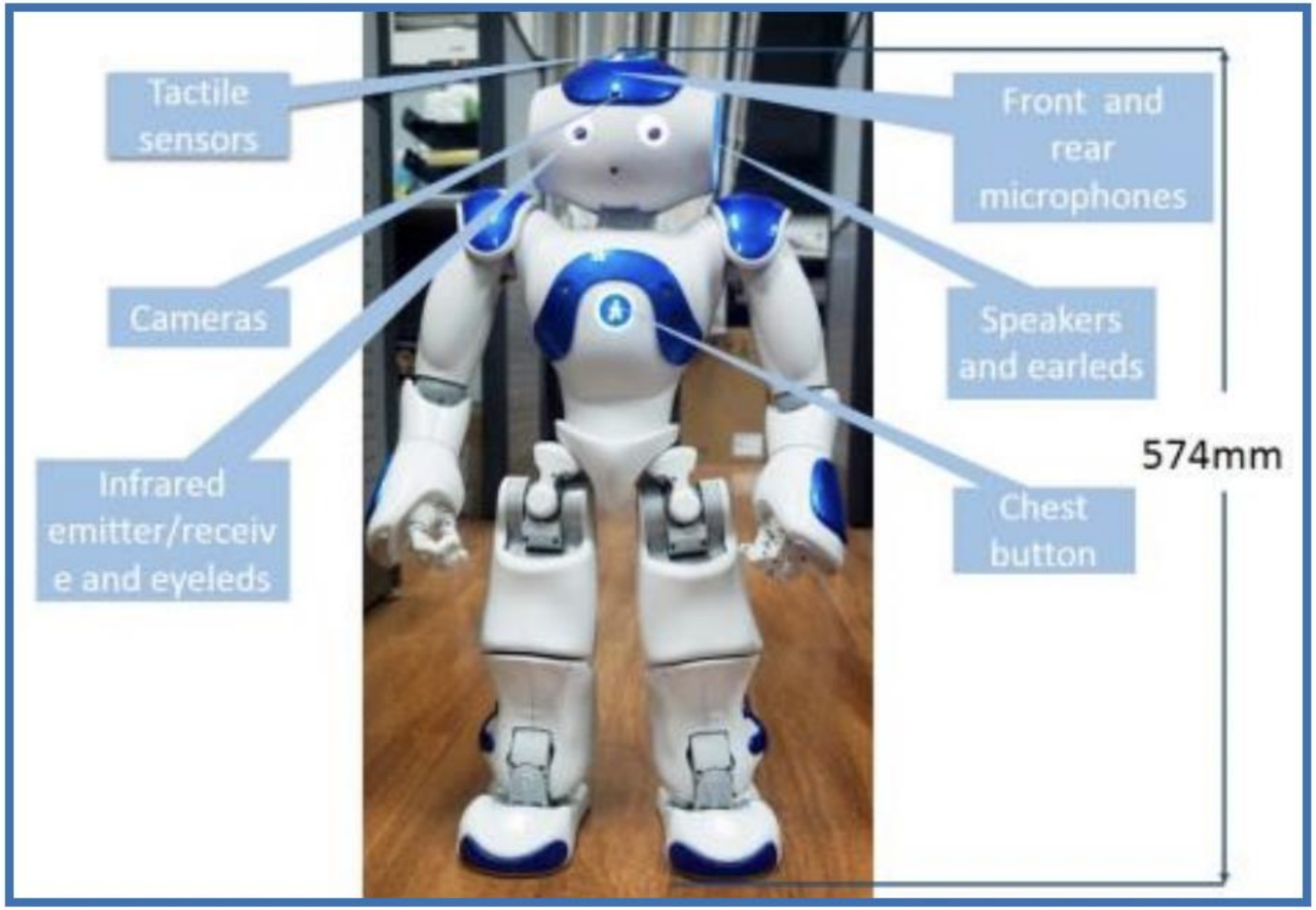
Standard captioning output: A plate with a sandwich and salad on it.

Our model with different personality traits (215 possible traits, not all shown here):

Sweet	That is a lovely sandwich.
Dramatic	This sandwich looks so delicious! My goodness!
Anxious	I’m afraid this might make me sick if I eat it.
Sympathetic	I feel so bad for that carrot, about to be consumed.
Arrogant	I make better food than this
Optimistic	It will taste positively wonderful!
Money-minded	I would totally pay $100 for this plate.

Shuster et al. Engaging Image Captioning via Personality. 2019

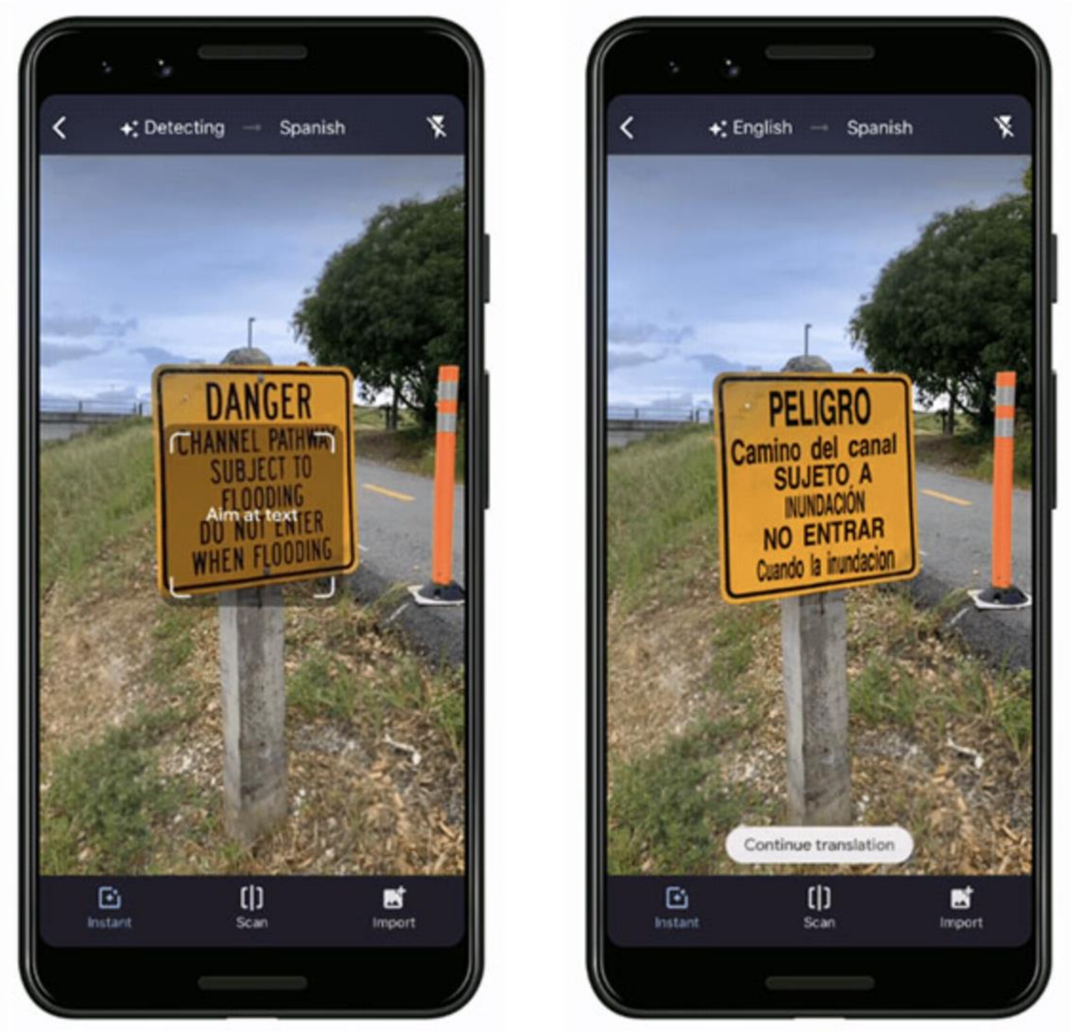
Education (e.g., for Preschoolers)


Tactile sensors
Front and rear microphones
Cameras
Speakers and earleds
574mm
Infrared emitter/receive and eyeleds
Chest button

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

He et al. An Educational Robot System of Visual Question Answering for Preschoolers. 2017.

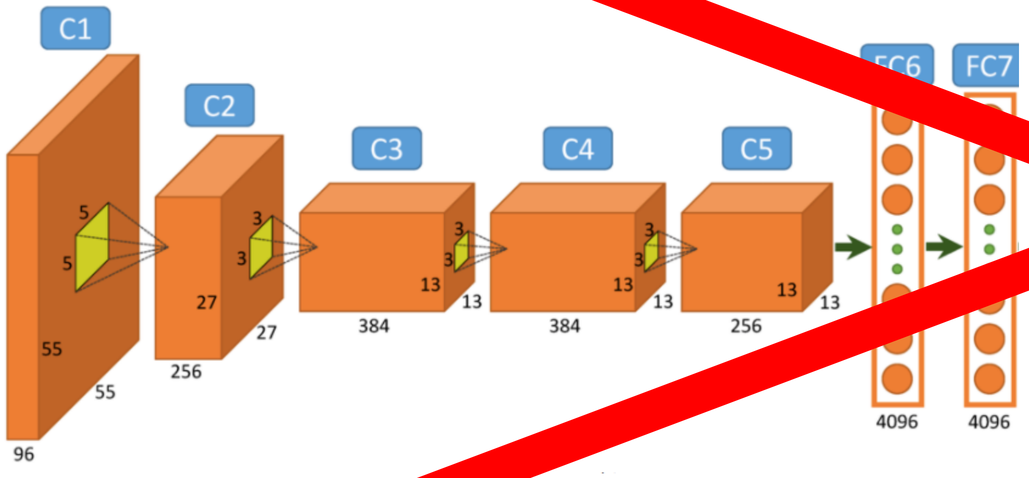
Education (e.g., Learning Foreign Languages)



https://hexus.net/tech/news/software/132620-google-translate-apps-camera-mode-auto-detects-languages/

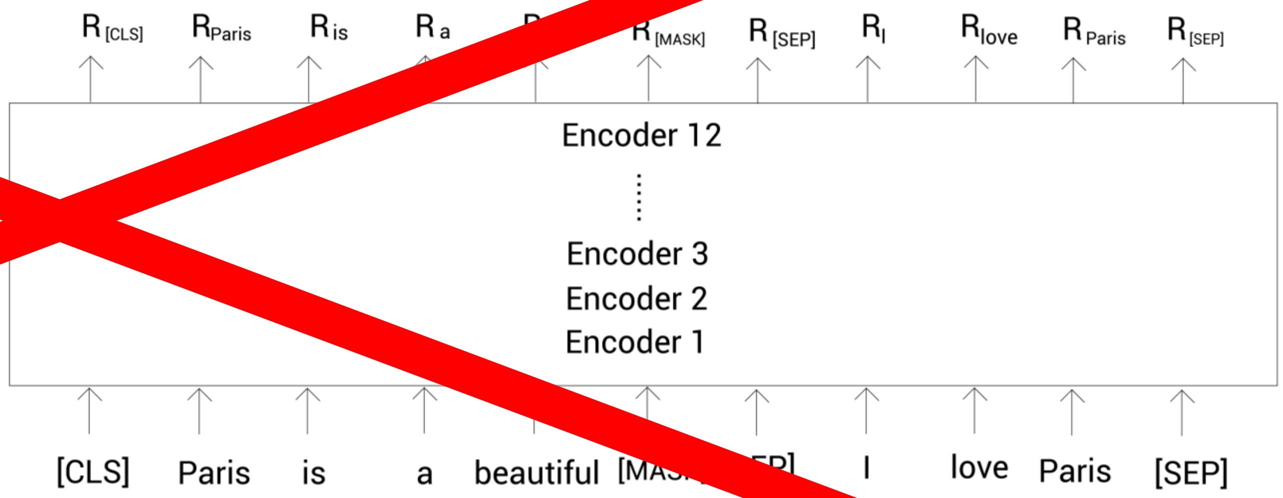
Challenge: Two Modalities in One Framework

e.g., vision representation with AlexNet



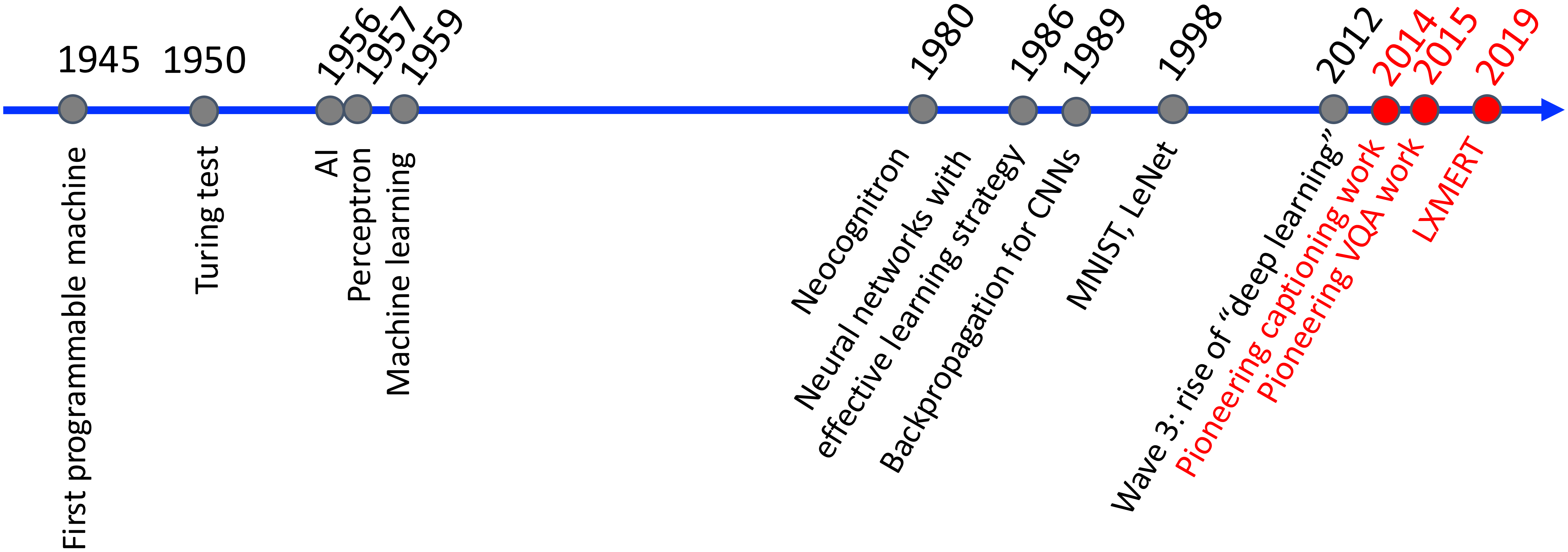
https://www.researchgate.net/figure/Architecture-of-Alexnet-From-left-to-right-input-to-output-five-convolutional-layers_fig2_312303454

e.g., language representation with BERT



https://static.packt-cdn.com/downloads/9781838821593_ColorImages.pdf

Historical Context



1945 First programmable machine

1950 Turing test

1956 AI

1957 Perceptron

1959 Machine learning

1980 Neocognitron

1986 Neural networks with effective learning strategy

1989 Backpropagation for CNNs

1998 MNIST, LeNet

2012 Wave 3: rise of “deep learning”

2014 Pioneering captioning work

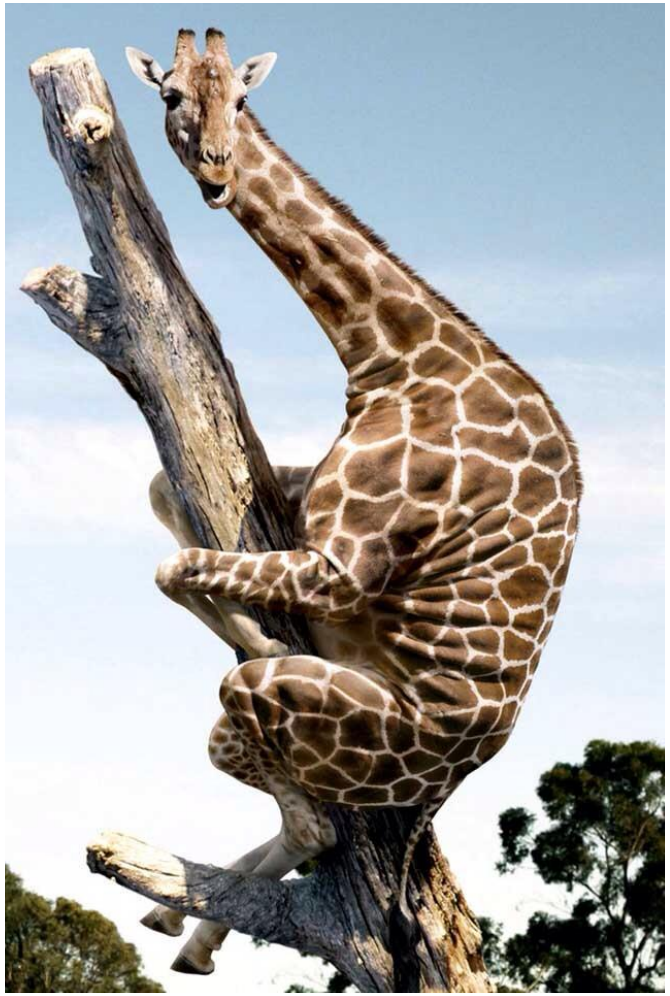
2015 Pioneering VQA work

2019 LXMERT

Today's Topics

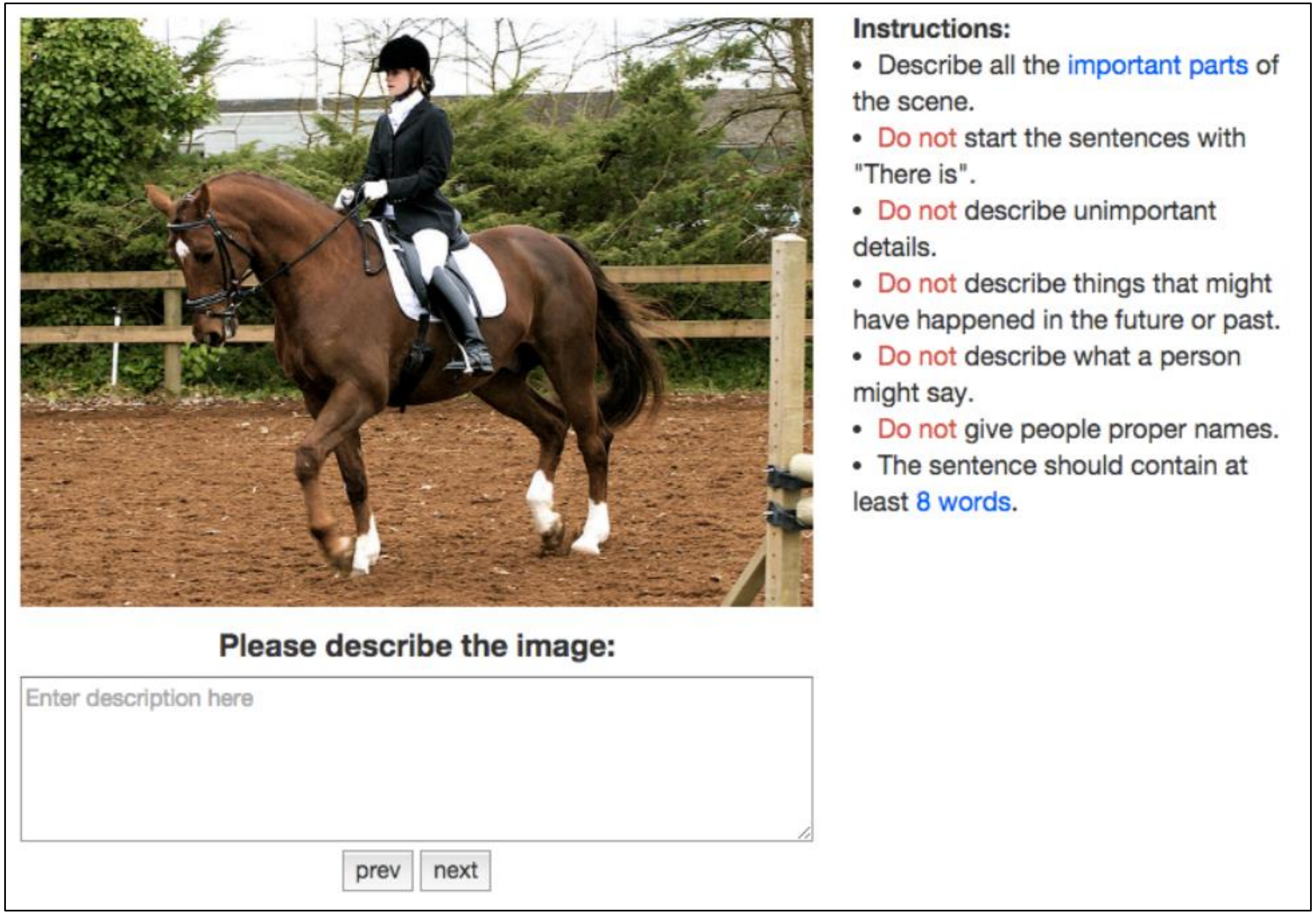
- Motivating applications
- Image captioning: pioneering dataset and model
- Visual question answering: pioneering dataset and model
- LXMERT: multimodal representations
- Programming tutorial

Class Task: How Would You Describe This Image?



Form:

MSCOCO: Annotation Instructions



Instructions:

- Describe all the important parts of the scene.
- Do not start the sentences with "There is".
- Do not describe unimportant details.
- Do not describe things that might have happened in the future or past.
- Do not describe what a person might say.
- Do not give people proper names.
- The sentence should contain at least 8 words.

Please describe the image:

Chen et al. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv 2015.

MSCOCO Dataset

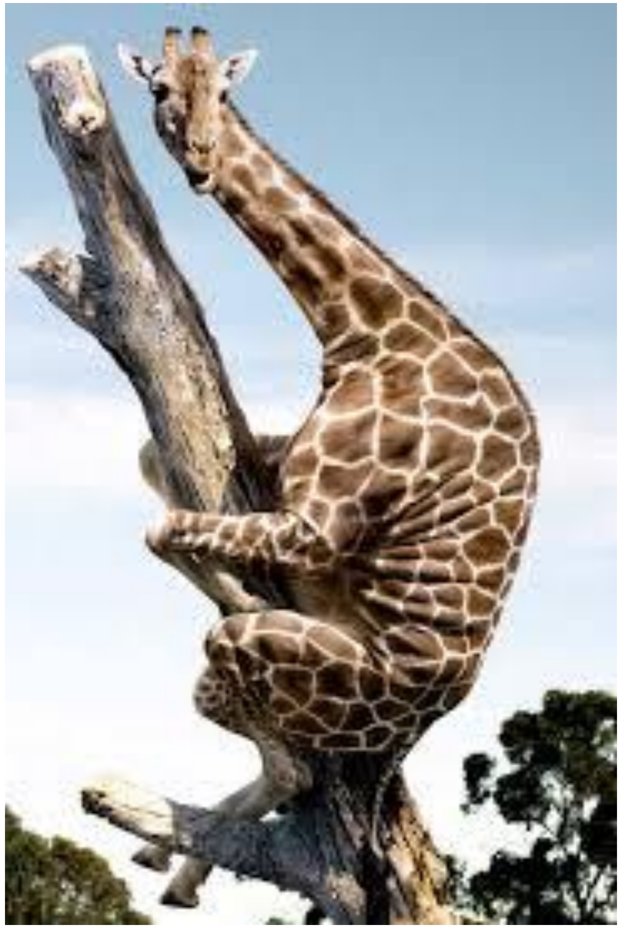
- 1,026,459 captions collected for 164,062 images from AMT workers

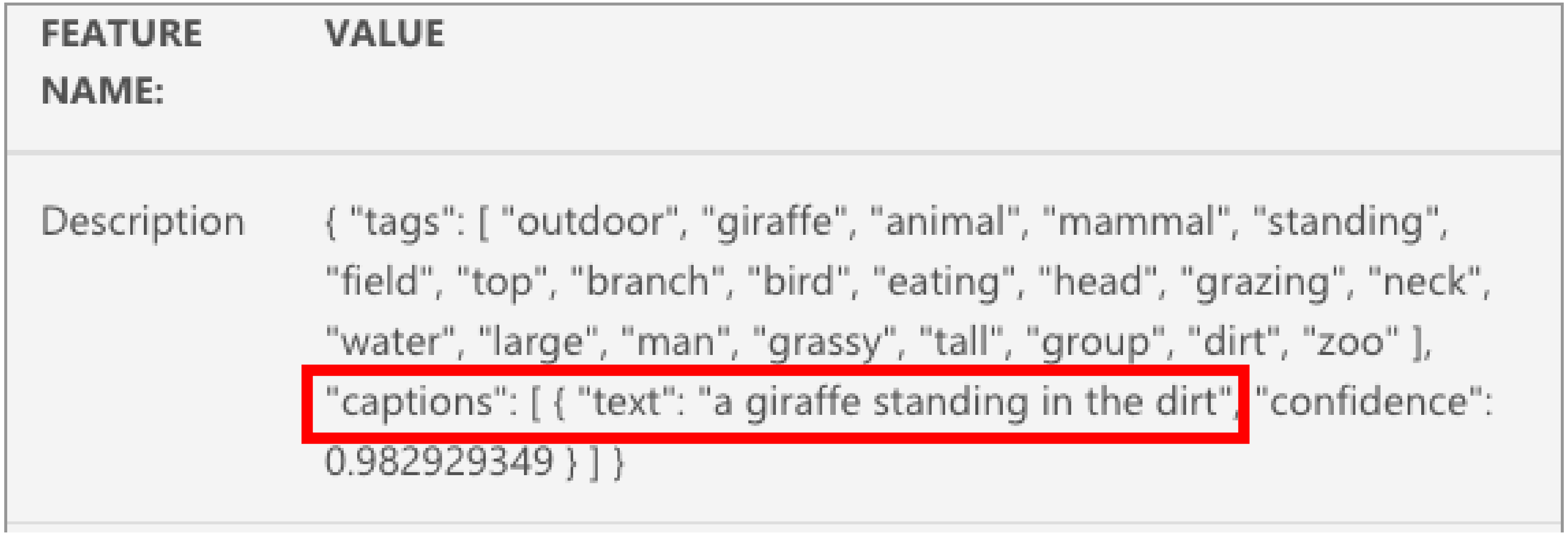
- How long do you think data collection took?
  - ~4 years (40 hrs per week, 52 weeks per year) or ~8,500 hours

- How much do you think it cost?
  - ~$128k (assumes 30 sec per caption and $15/hour; would yield for one person $32,000 per year)

Chen et al. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv 2015.

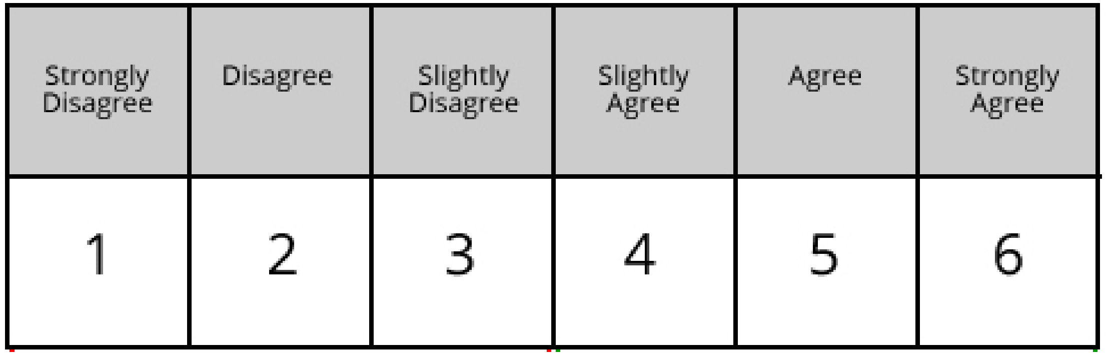
How Would You Evaluate Predicted Captions?





FEATURE NAME:	VALUE
Description	{ "tags": ["outdoor", "giraffe", "animal", "mammal", "standing", "field", "top", "branch", "bird", "eating", "head", "grazing", "neck", "water", "large", "man", "grassy", "tall", "group", "dirt", "zoo"], "captions": [{ "text": "a giraffe standing in the dirt", "confidence": 0.982929349 }] }

Evaluation: Human Judgments



Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
1	2	3	4	5	6

- The description accurately describes the image (Kulkarni et al., 2011; Li et al., 2011; Mitchell et al., 2012; Kuznetsova et al., 2012; Elliott & Keller, 2013; Hodosh et al., 2013).
- The description is grammatically correct (Yang et al., 2011; Mitchell et al., 2012; Kuznetsova et al., 2012; Elliott & Keller, 2013).
- The description has no incorrect information (Mitchell et al., 2012).
- The description is relevant for this image (Li et al., 2011; Yang et al., 2011).
- The description is creatively constructed (Li et al., 2011).
- The description is human-like (Mitchell et al., 2012).

Bernardi et al. Automatic Description from Images: A Survey of Models, Datasets, and Measures. JAIR 2016.

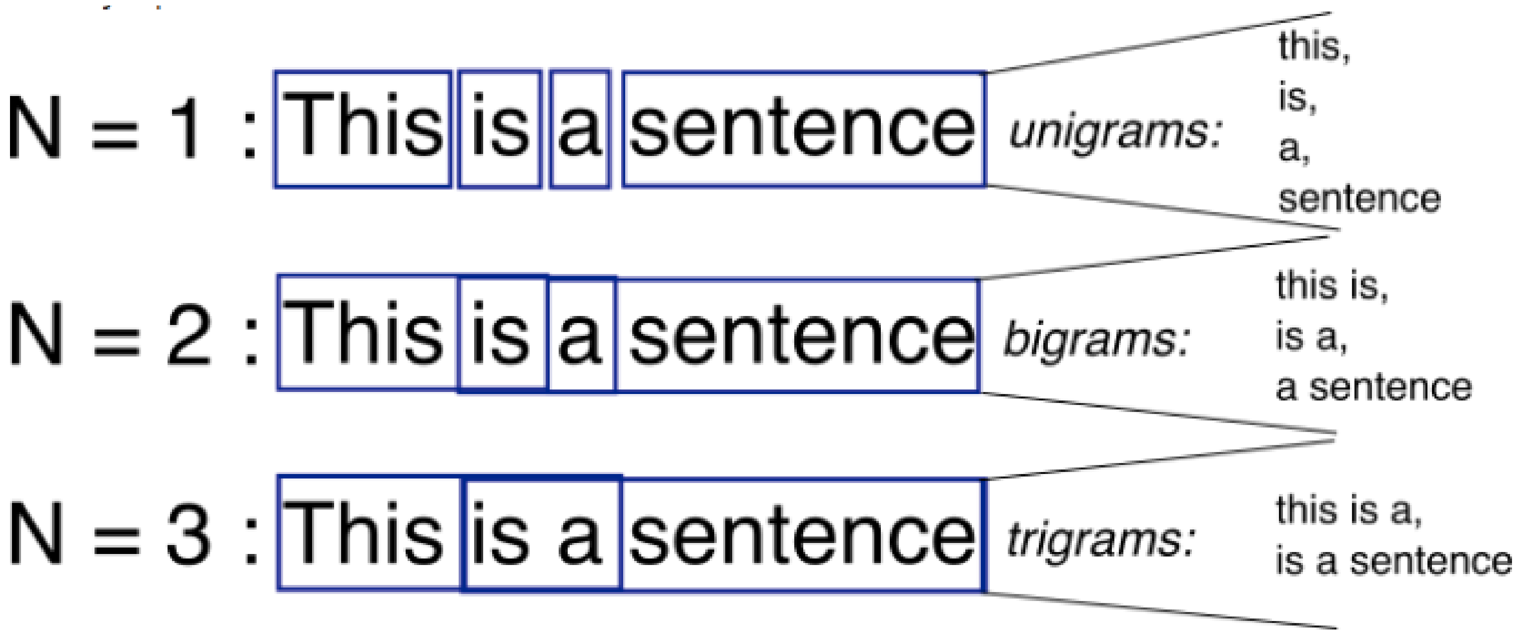
Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Idea: compute similarities of n-grams between a predicted caption and each ground truth caption



N = 1 : This is a sentence *unigrams:* this, is, a, sentence

N = 2 : This is a sentence *bigrams:* this is, is a, a sentence

N = 3 : This is a sentence *trigrams:* this is a, is a sentence

http://recognize-speech.com/language-model/n-gram-model/comparison

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Idea: measure similarity of a predicted caption to how most people describe an image based on *n*-grams unique to the image

A cow is standing in a field.

A cow with horns and long hair covering its face stands in a field.

A cow with hair over its eyes stands in a field.

This horned creature is getting his picture taken.

A furry animal with horns roams on the range.

Vedantam et al. CIDEr: Consensus-based Image Description Evaluation. CVPR 2015.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

What content do most people describe in this image?

A cow is standing in a field.

A cow with horns and long hair covering its face stands in a field.

A cow with hair over its eyes stands in a field.

This horned creature is getting his picture taken.

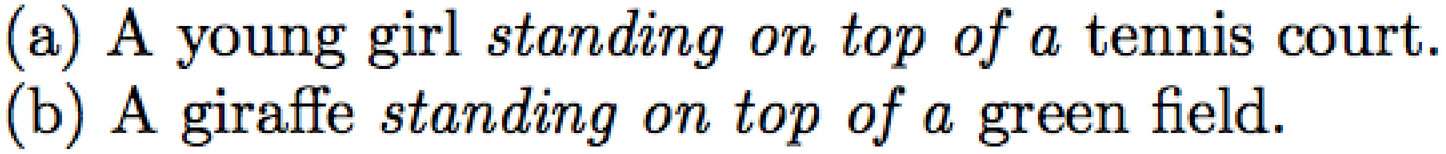
A furry animal with horns roams on the range.

Vedantam et al. CIDEr: Consensus-based Image Description Evaluation. CVPR 2015.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Do you think these two captions describe the same image?



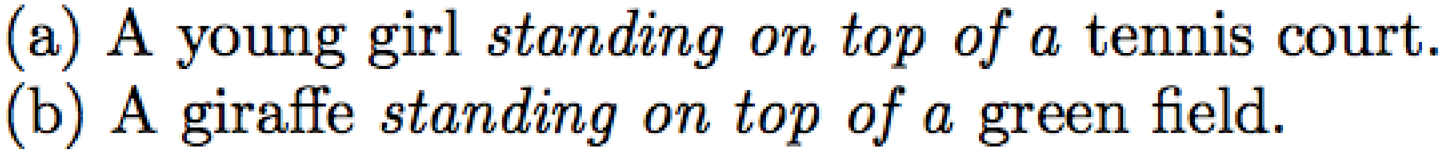
(a) A young girl *standing on top of a* tennis court.
(b) A giraffe *standing on top of a* green field.

Anderson et al. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Problem: n-gram methods scores these as very similar



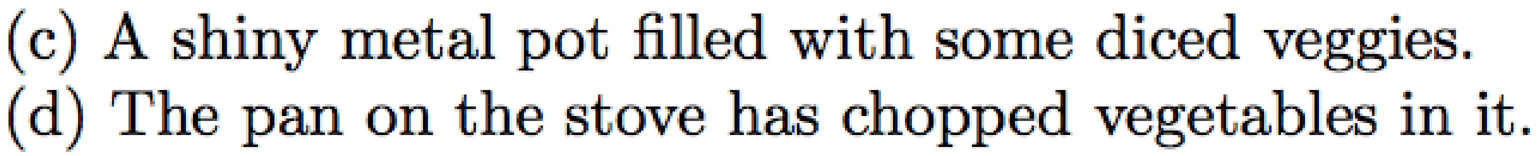
(a) A young girl *standing on top of a* tennis court.
(b) A giraffe *standing on top of a* green field.

Anderson et al. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Do you think these two captions describe the same image?



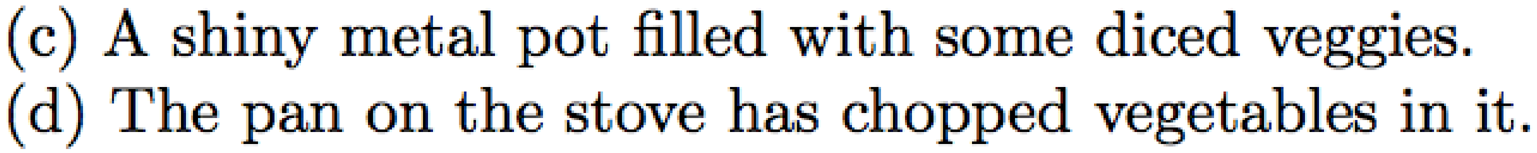
(c) A shiny metal pot filled with some diced veggies.
(d) The pan on the stove has chopped vegetables in it.

Anderson et al. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Problem: n-gram methods scores these as very different



(c) A shiny metal pot filled with some diced veggies.
(d) The pan on the stove has chopped vegetables in it.

Anderson et al. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

Idea: compare scene graph of prediction to scene graph of ground truth

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

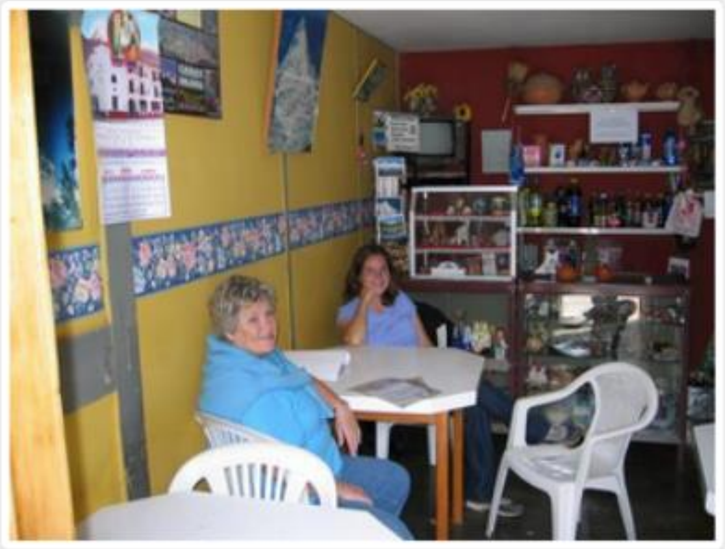
"two woman are sitting at a table"

Anderson et al. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

What is the meaningful semantic content in these captions?



"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

"two woman are sitting at a table"

Anderson et al. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

Evaluation: Automated

- BLEU
- METEOR
- Rouge
- CIDEr
- SPICE

Meaningful semantic content in these captions:

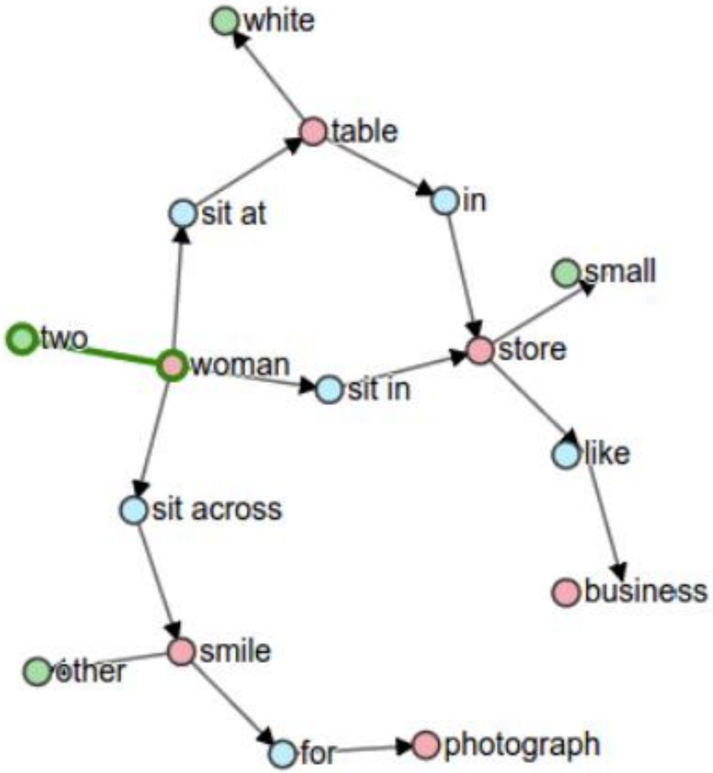
"two women are sitting at a white table"

"two women sit at a table in a small store"

"two women sit across each other at a table smile for the photograph"

"two women sitting in a small store like business"

"two woman are sitting at a table"



Anderson et al. SPICE: Semantic Propositional Image Caption Evaluation. ECCV 2016.

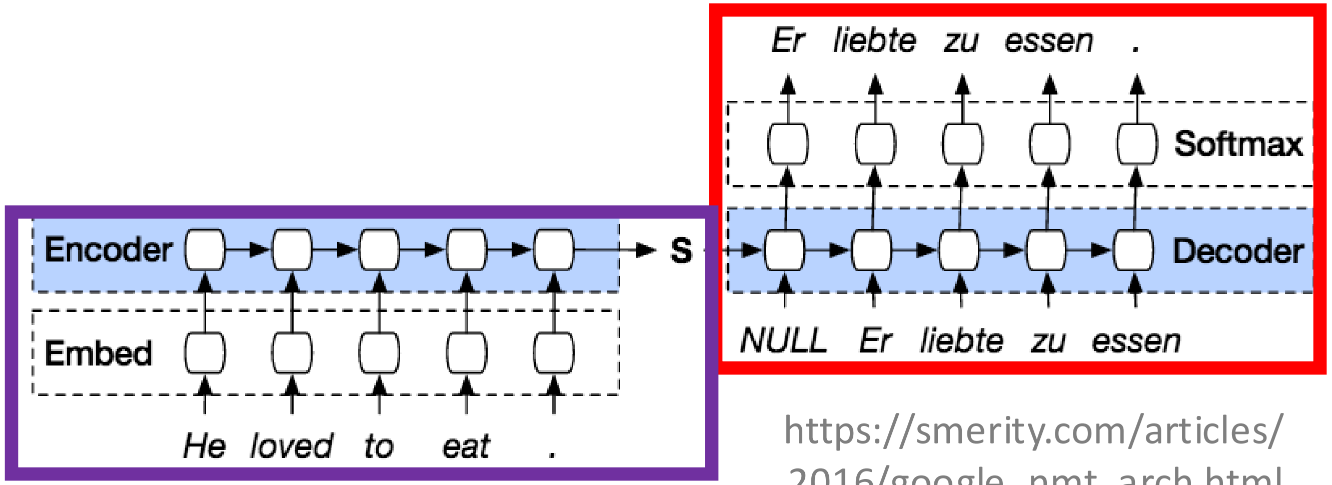
Evaluation: Implementation Detail

- Text pre-processing; e.g., for COCO-Captions
  - Captions tokenized: word-based Stanford PTBTokenizer
  - Punctuation removed

Chen et al. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv 2015.

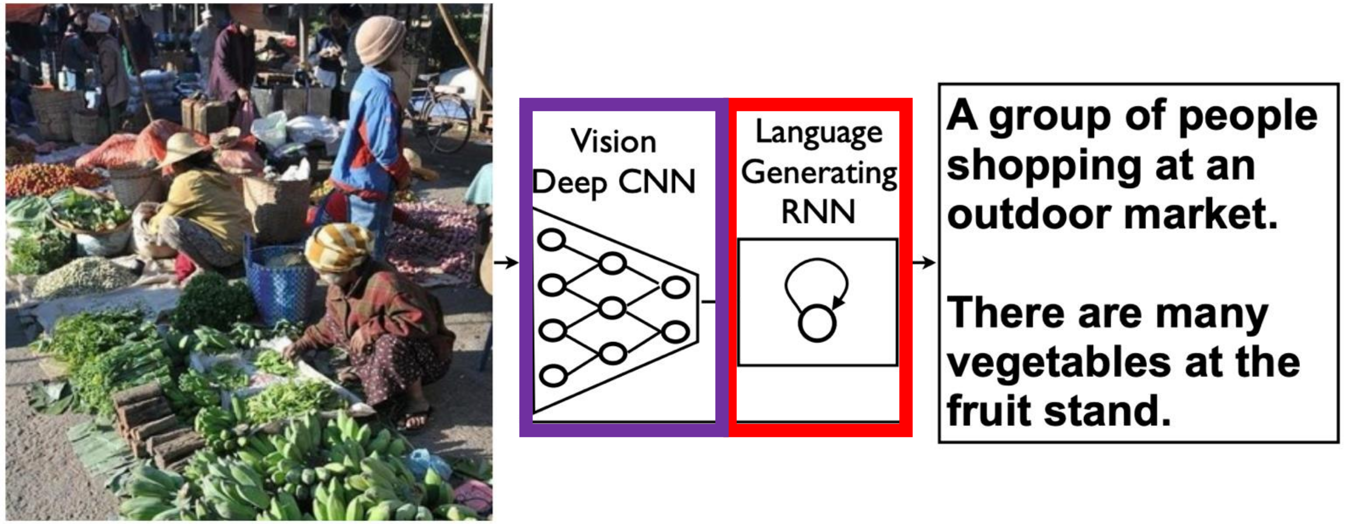
Pioneering Work: “Show and Tell”

Inspiration is machine translation:



https://smerity.com/articles/2016/google_nmt_arch.html

Idea is to translate image to text in an end-to-end trained model:



A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.

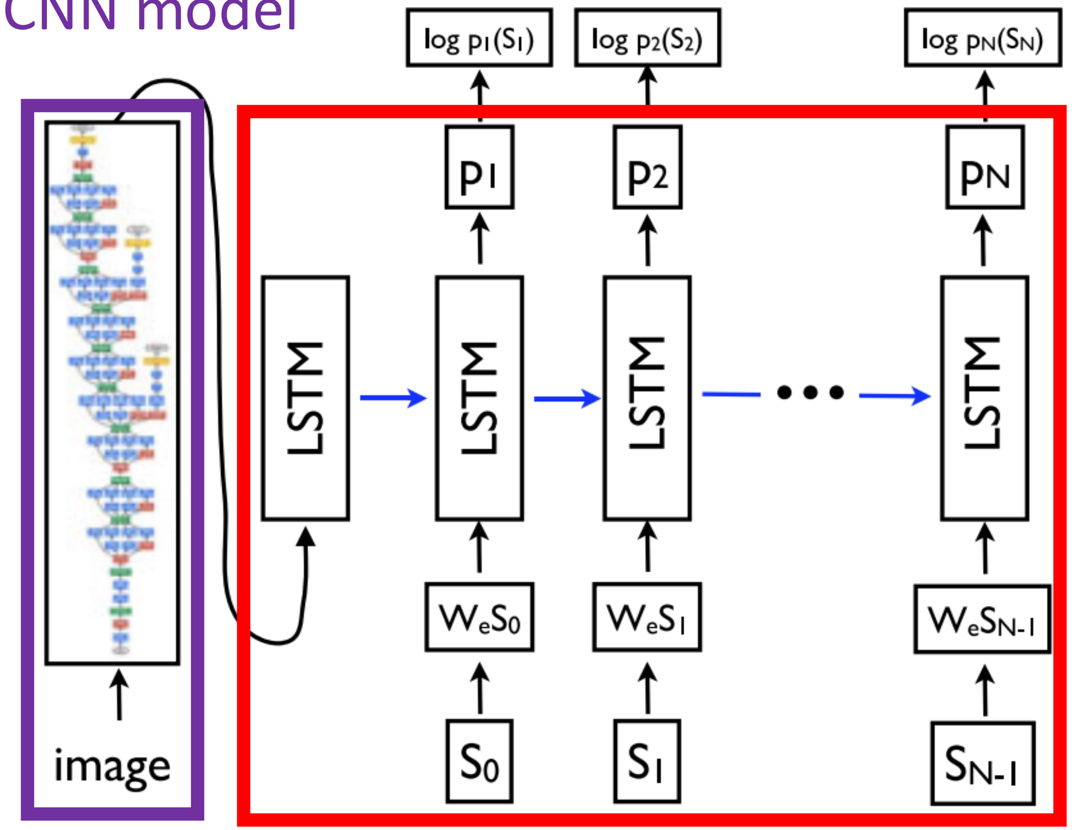
Vinyals et al. Show and Tell: A Neural Image Caption Generator. CVPR 2015.

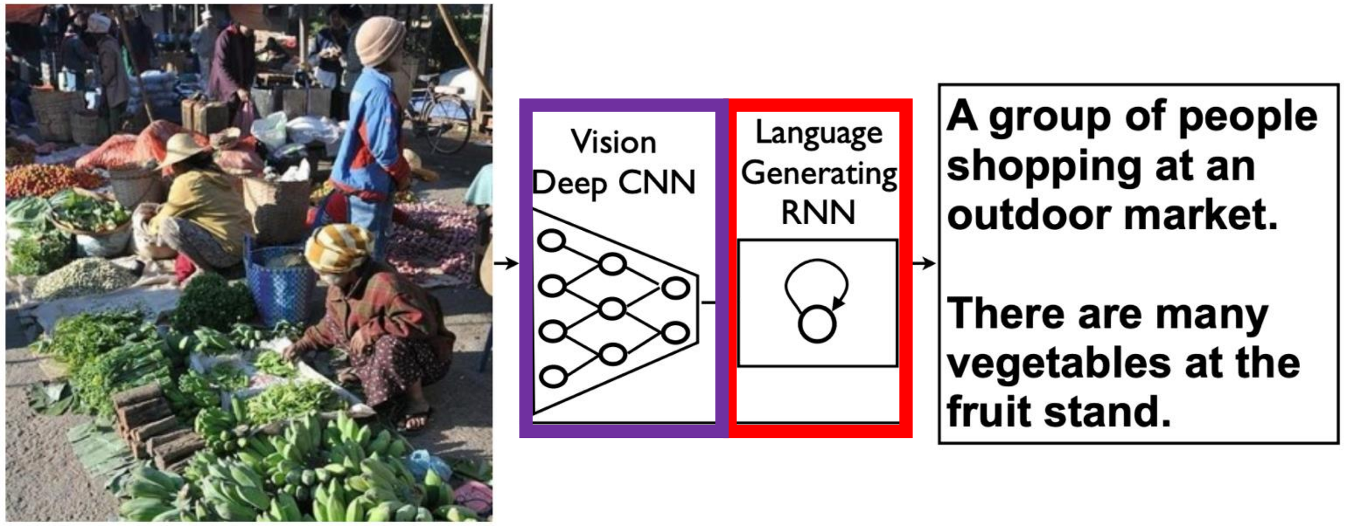
Pioneering Work: “Show and Tell”

Initialized to ImageNet 2014 winner’s CNN model

LSTM, predicts next word

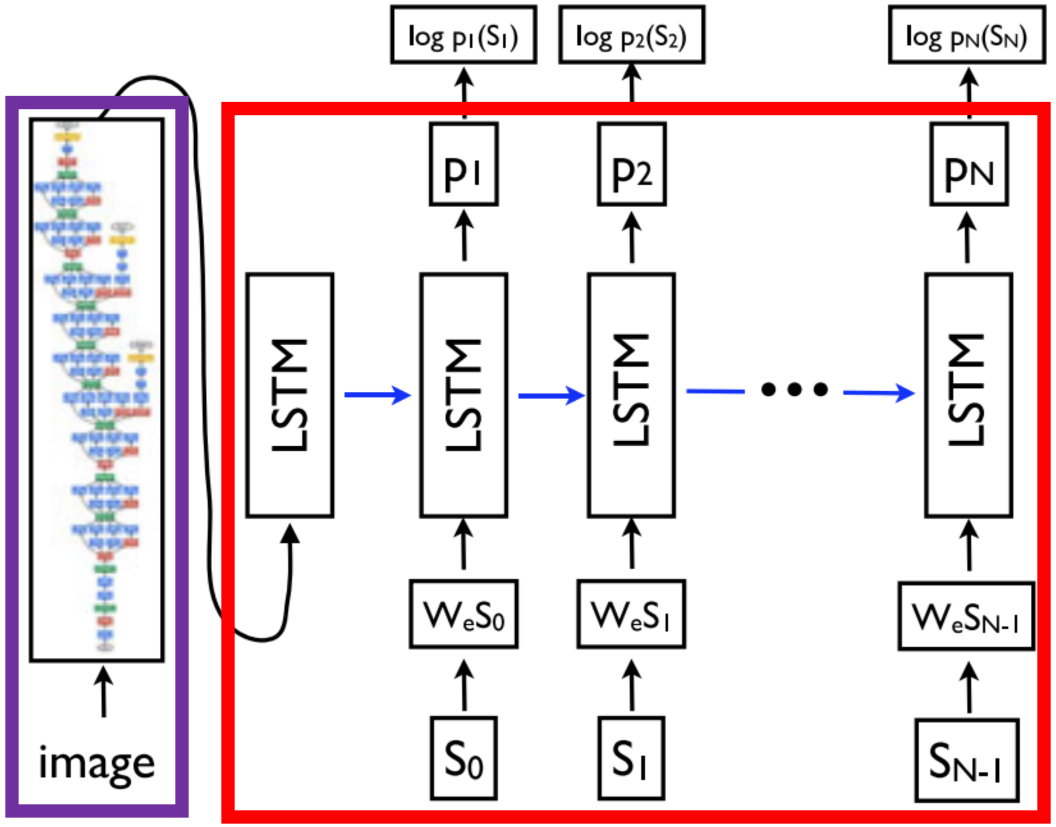
(words tokenized, kept if seen <4 times in training data, and then converted to learned 512-d word embedding)



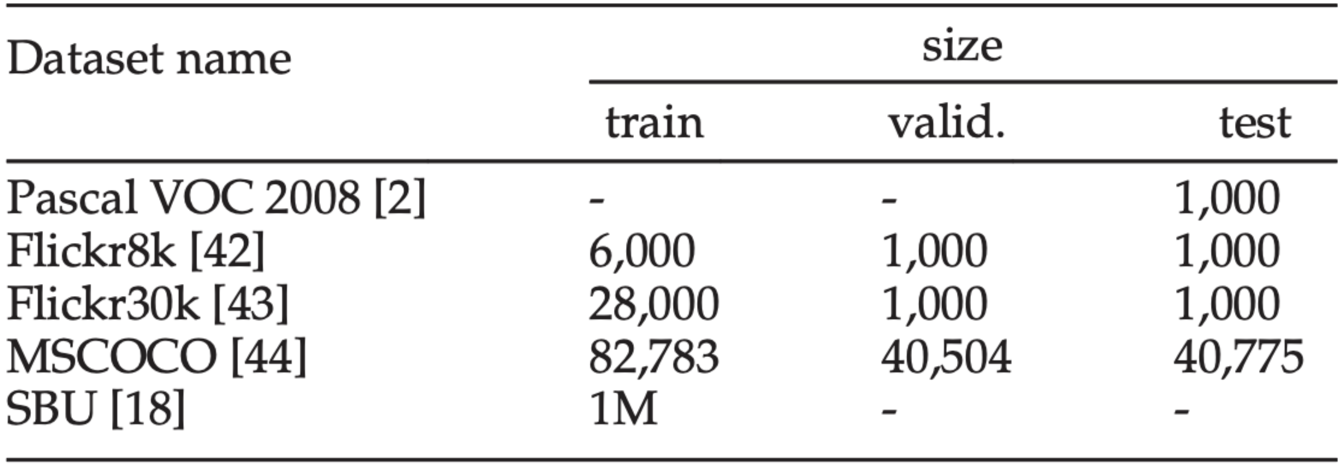

Vision Deep CNN
Language Generating RNN
A group of people shopping at an outdoor market.
There are many vegetables at the fruit stand.

Vinyals et al. Show and Tell: A Neural Image Caption Generator. CVPR 2015.

Pioneering Work: “Show and Tell”



- Trained with CNN parameters frozen for 500K steps, and then all parameters for 100K steps;
- Training took over 3 weeks on a K20 GPU
- Training and evaluation datasets:

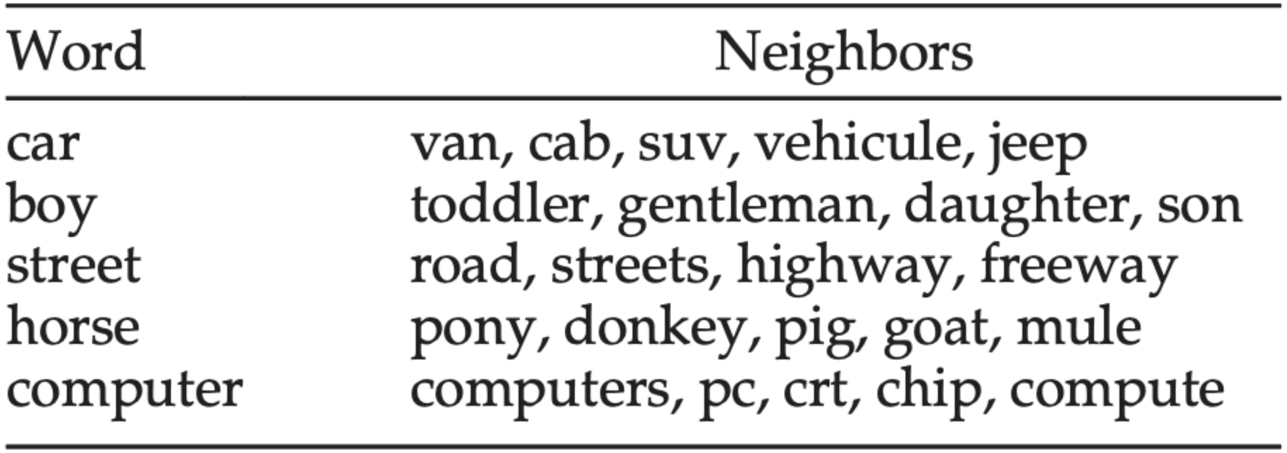


Dataset name	size		
	train	valid.	test
Pascal VOC 2008 [2]	-	-	1,000
Flickr8k [42]	6,000	1,000	1,000
Flickr30k [43]	28,000	1,000	1,000
MSCOCO [44]	82,783	40,504	40,775
SBU [18]	1M	-	-

Vinyals et al. Show and Tell: A Neural Image Caption Generator. CVPR 2015.

Pioneering Work: “Show and Tell”

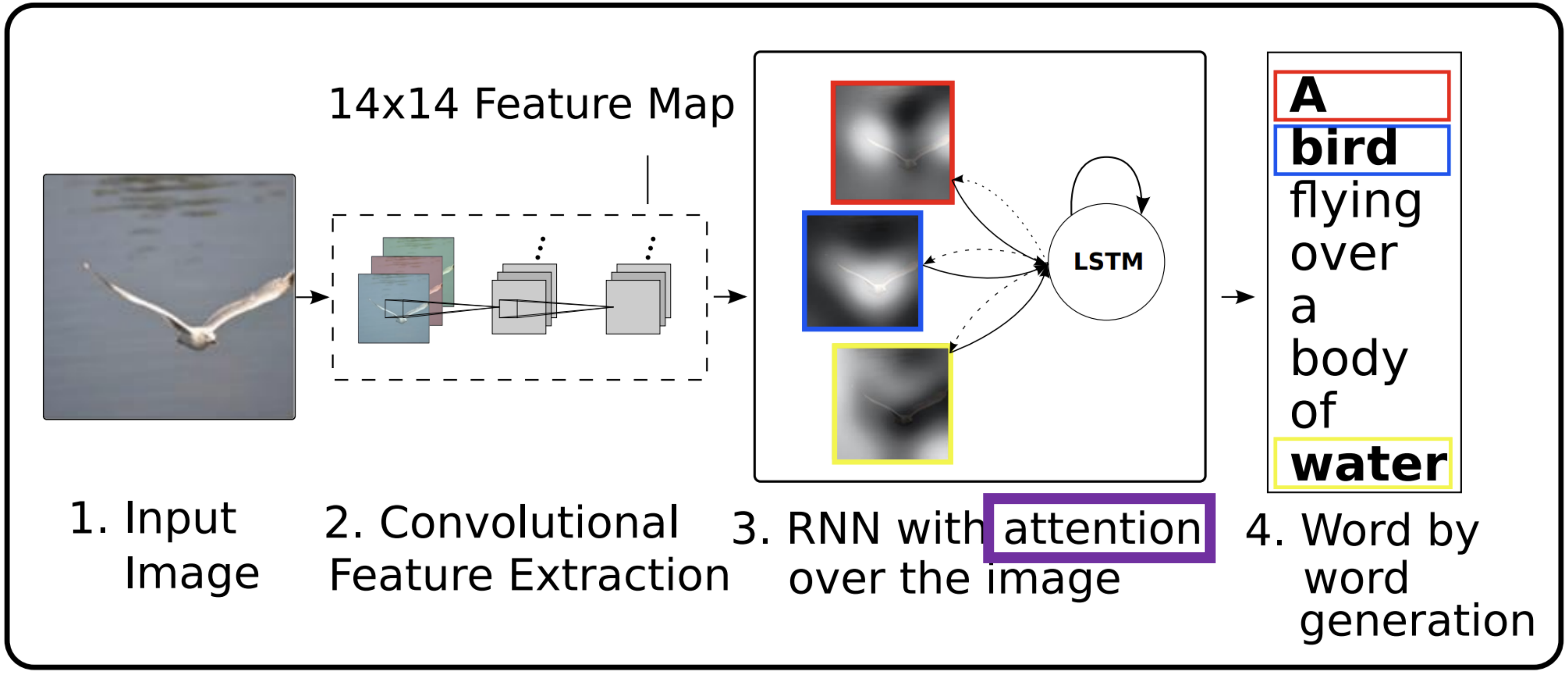
- Achieved state-of-the-art performance
- Hypothesis: more training data would boost performance
- Exemplar nearest neighbor words in learned word embedding space:



Word	Neighbors
car	van, cab, suv, vehicule, jeep
boy	toddler, gentleman, daughter, son
street	road, streets, highway, freeway
horse	pony, donkey, pig, goat, mule
computer	computers, pc, crt, chip, compute

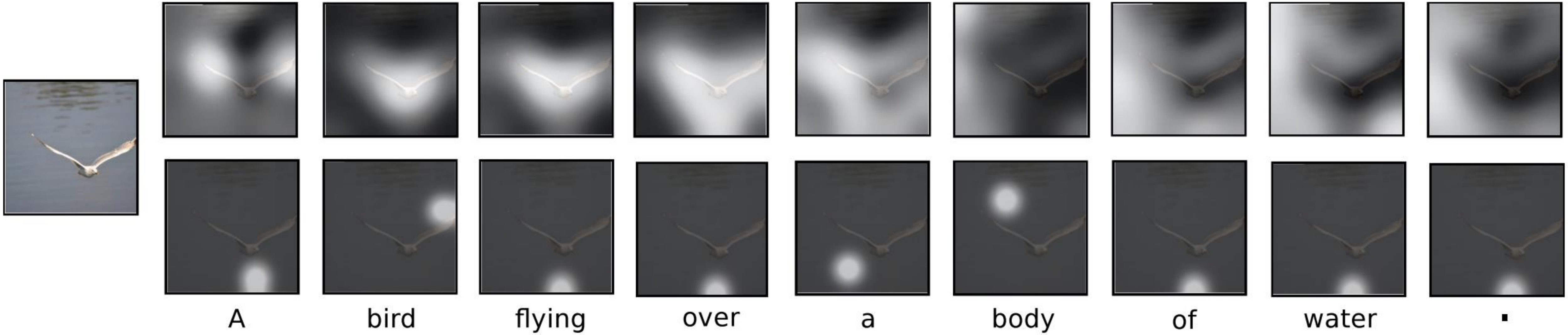
Vinyals et al. Show and Tell: A Neural Image Caption Generator. CVPR 2015.

Subsequent Work: “Show, Attend, and Tell”


14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

Xu et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML 2015.

Subsequent Work: “Show, Attend, and Tell”



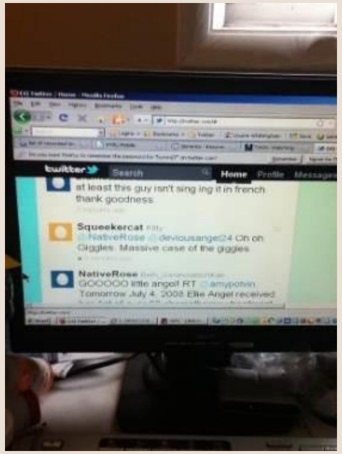
Xu et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML 2015.

Today's Topics

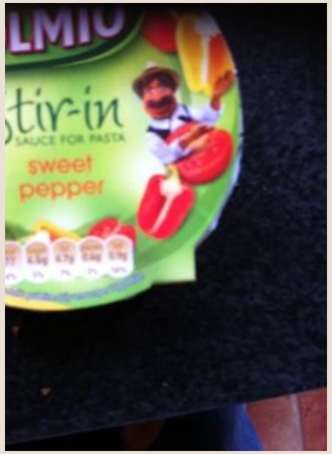
- Motivating applications
- Image captioning: pioneering dataset and model
- Visual question answering: pioneering dataset and model
- LXMERT: multimodal representations
- Programming tutorial

Class Task: Answer Visual Question

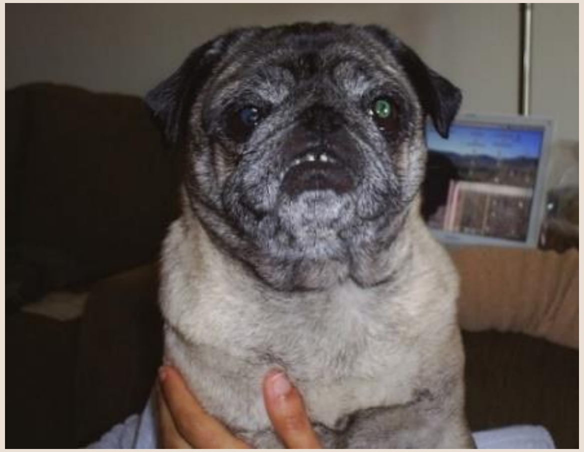
Fill out Google form:


twitter
Home Profile Messages
at least this guy isn't sing ing it in french thank goodness
Squeekercat
@NativeRose @deviousangel24 Oh on Giggles. Massive case of the giggles
NativeRose
GOOOOO little angel! RT @amypotvin Tomorrow July 4, 2008 Elle Angel received

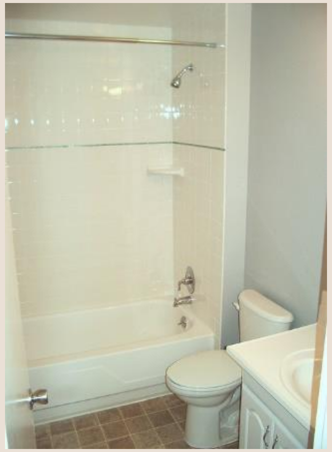
Is my monitor on?



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

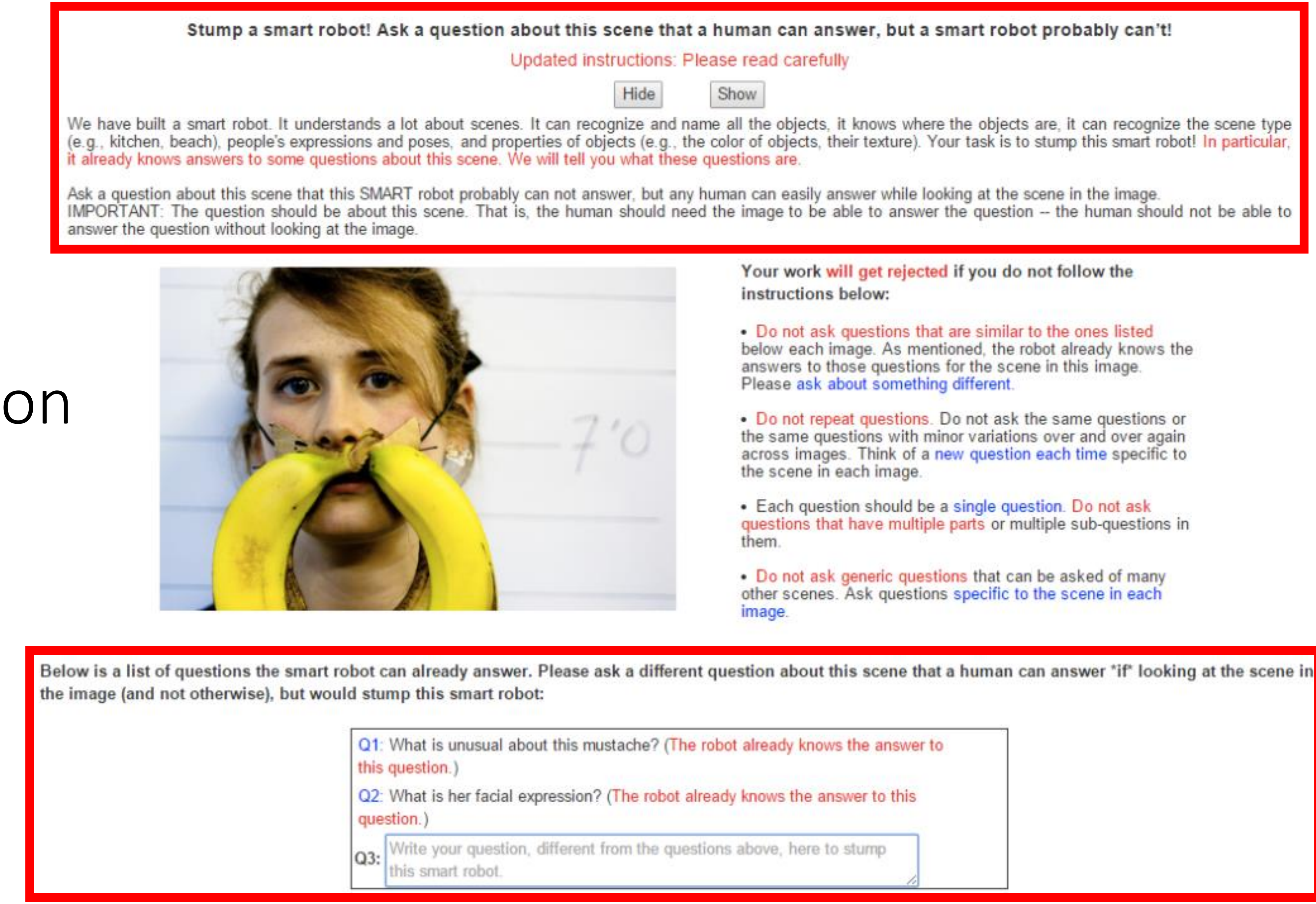


Does this picture look scary?



Which side of the room is the toilet on?

e.g., Question Generation



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

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.

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:

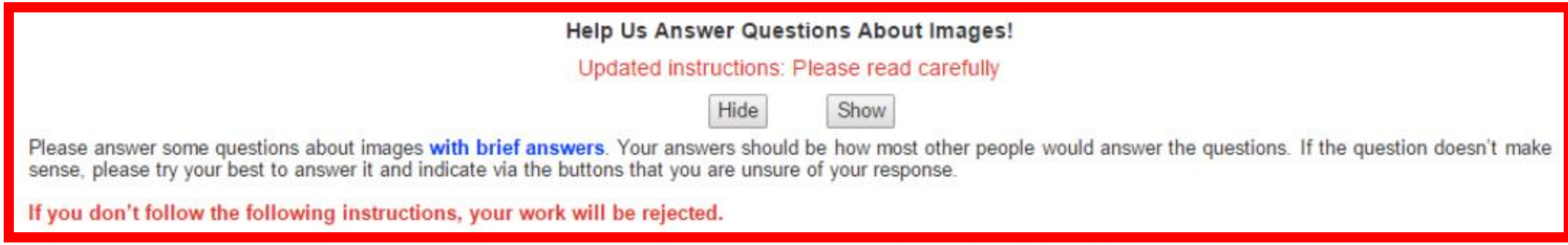
Q1: What is unusual about this mustache? (The robot already knows the answer to this question.)

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.

Agrawal et al. VQA: Visual Question Answering. ICCV 2015.

e.g., Answer Generation

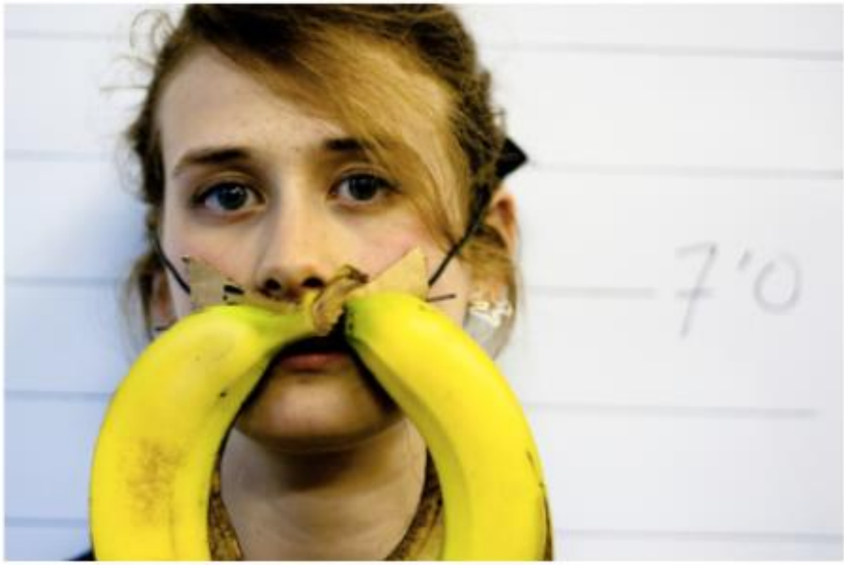


Help Us Answer Questions About Images!

Updated instructions: Please read carefully

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.

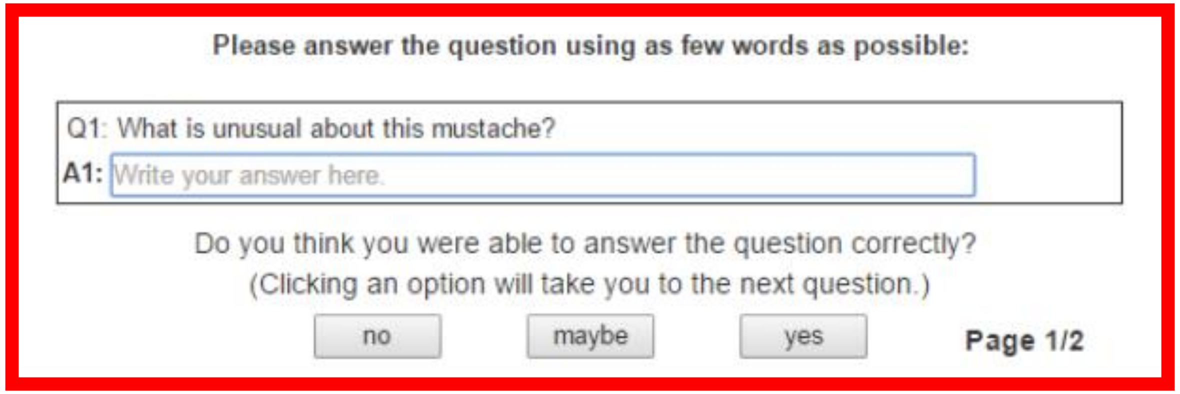
If you don't follow the following instructions, your work will be rejected.



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.

10 answers collected from 10 crowdworkers



Please answer the question using as few words as possible:

Q1: What is unusual about this mustache?

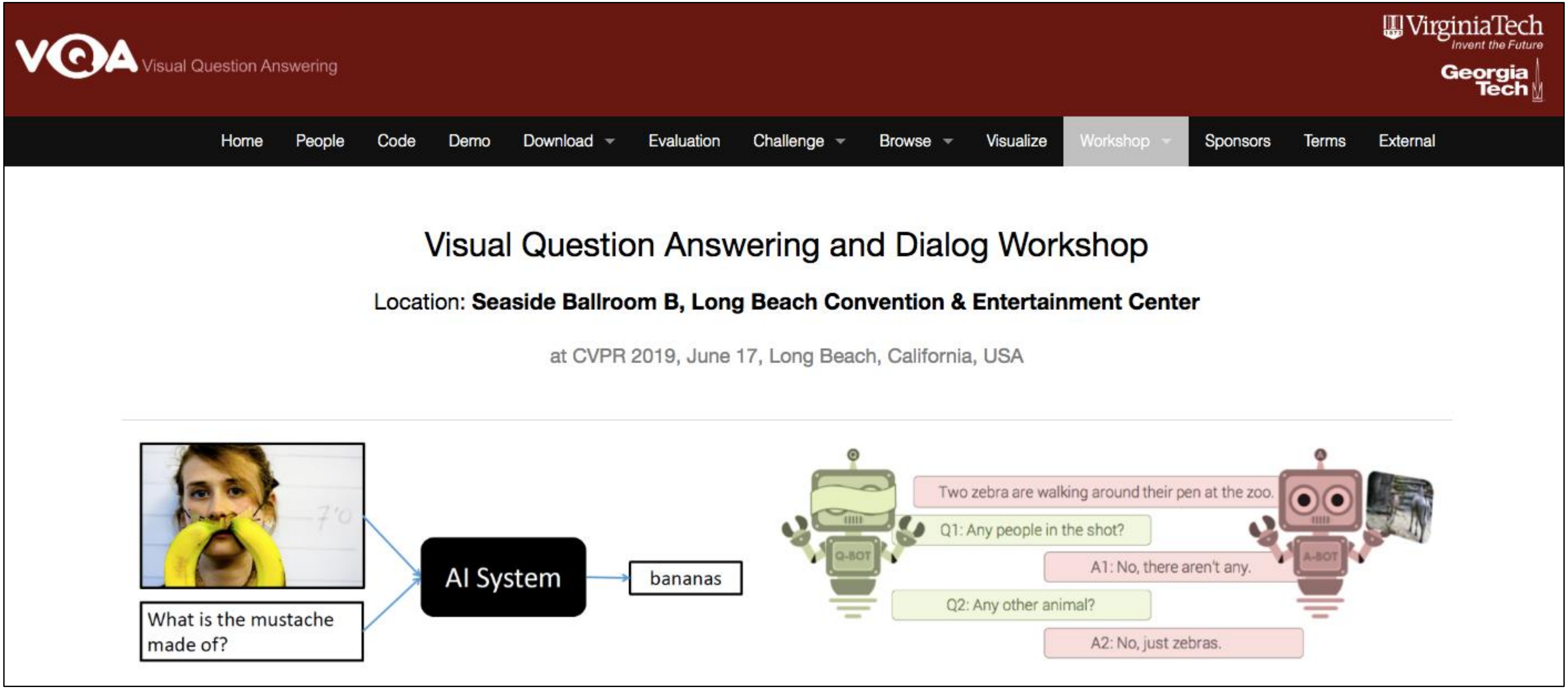
A1: Write your answer here.

Do you think you were able to answer the question correctly?
(Clicking an option will take you to the next question.)

Page 1/2

Agrawal et al. VQA: Visual Question Answering. ICCV 2015.

Mainstream VQA Challenge (held for 6 years)



Visual Question Answering and Dialog Workshop

Location: **Seaside Ballroom B, Long Beach Convention & Entertainment Center**

at CVPR 2019, June 17, Long Beach, California, USA

What is the mustache made of?

Two zebra are walking around their pen at the zoo.

Q1: Any people in the shot?

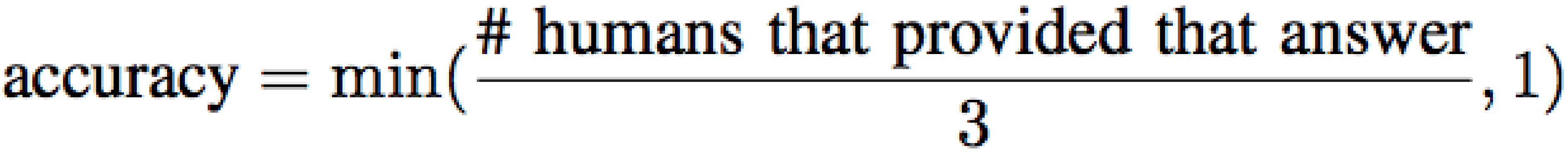
A1: No, there aren't any.

Q2: Any other animal?

A2: No, just zebras.

https://visualqa.org/workshop.html

Evaluating Automated Predictions: Basic Equation



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

Agrawal et al. VQA: Visual Question Answering. ICCV 2015.

Evaluating Automated Predictions: Example

Does this picture look scary?

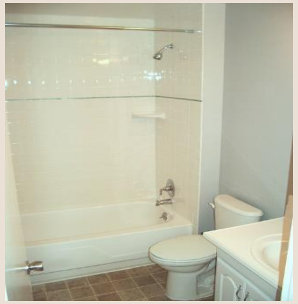
(1) yes
(2) no
(3) no
(4) yes
(5) no
(6) yes
(7) yes
(8) no
(9) no
(10) no

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

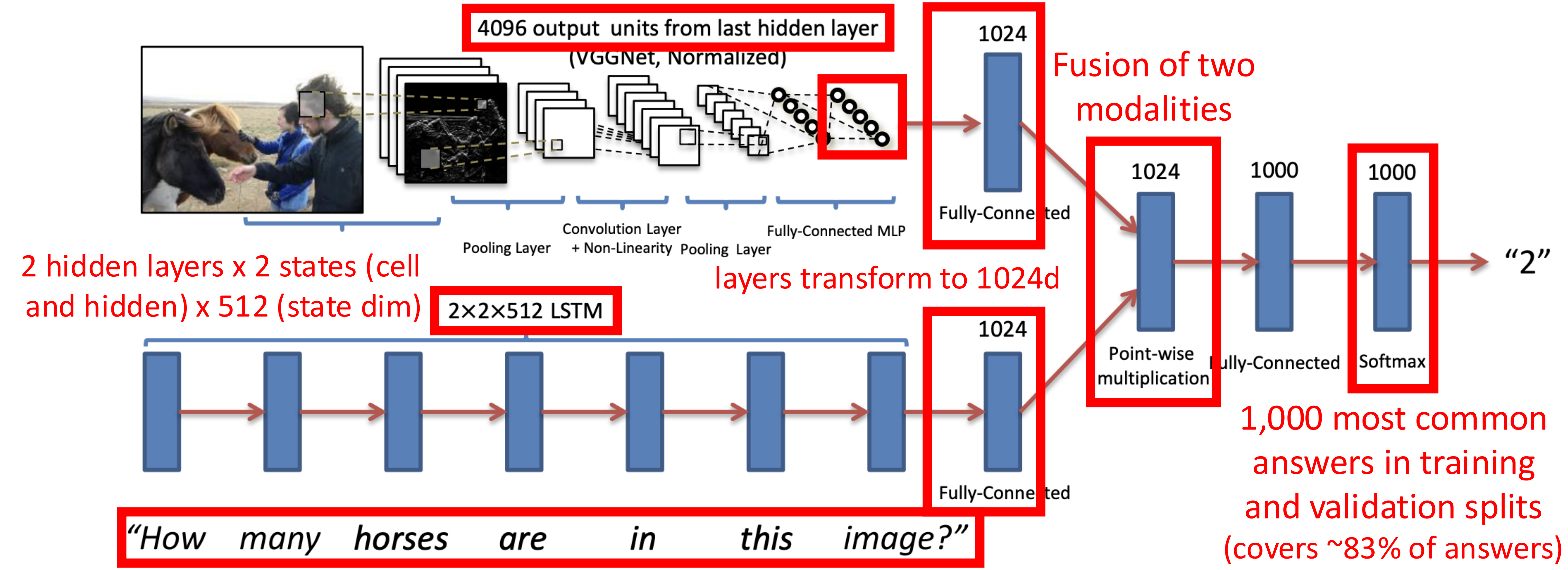
What is the accuracy of an algorithm prediction of

- "right"?
- "left"?
- "right side"?
- "center"?
- "bottom"?

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

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

Proposed Model



Fusion of two modalities

2 hidden layers x 2 states (cell and hidden) x 512 (state dim)

layers transform to 1024d

1,000 most common answers in training and validation splits (covers ~83% of answers)

Words encoded as 300-d embeddings, by fully-connected layer + tanh

Trained end-to-end with cross-entropy loss and VGGNet parameters frozen

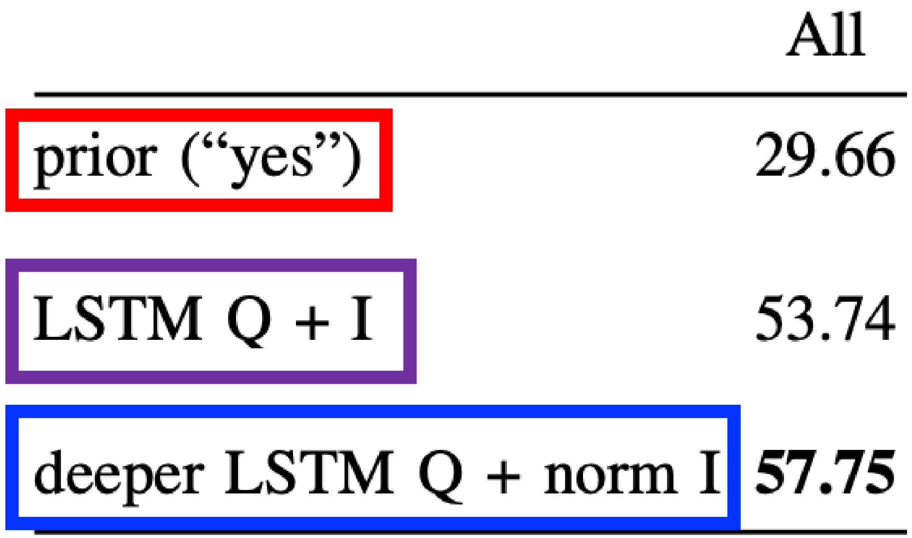
Agrawal et al. VQA: Visual Question Answering. ICCV 2015.

Model Analysis

Baseline: most popular answer from train/val splits

Baseline: 1-hidden layer LSTM and image activations without normalization

Proposed model



	All
prior ("yes")	29.66
LSTM Q + I	53.74
deeper LSTM Q + norm I	**57.75**

How does the model's performance compare to the simpler baselines?

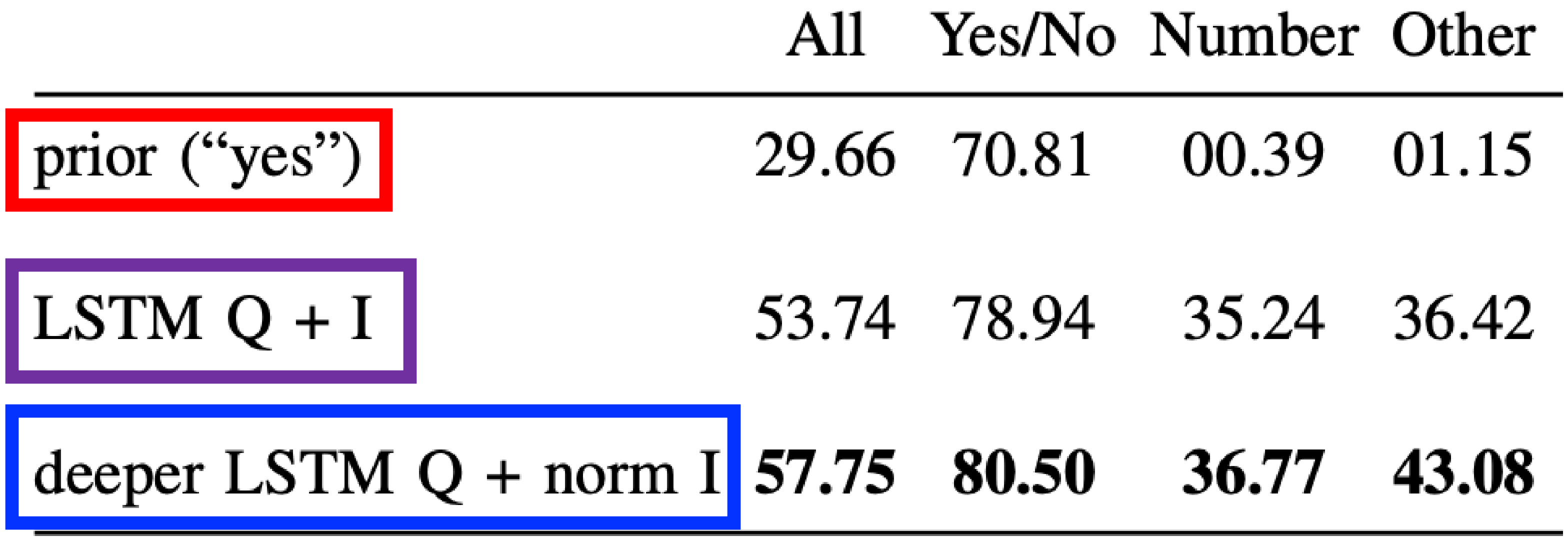
Agrawal et al. VQA: Visual Question Answering. ICCV 2015.

Model Analysis

Baseline: most popular answer from train/val splits

Baseline: 1-hidden layer LSTM and image activations without normalization

Proposed model

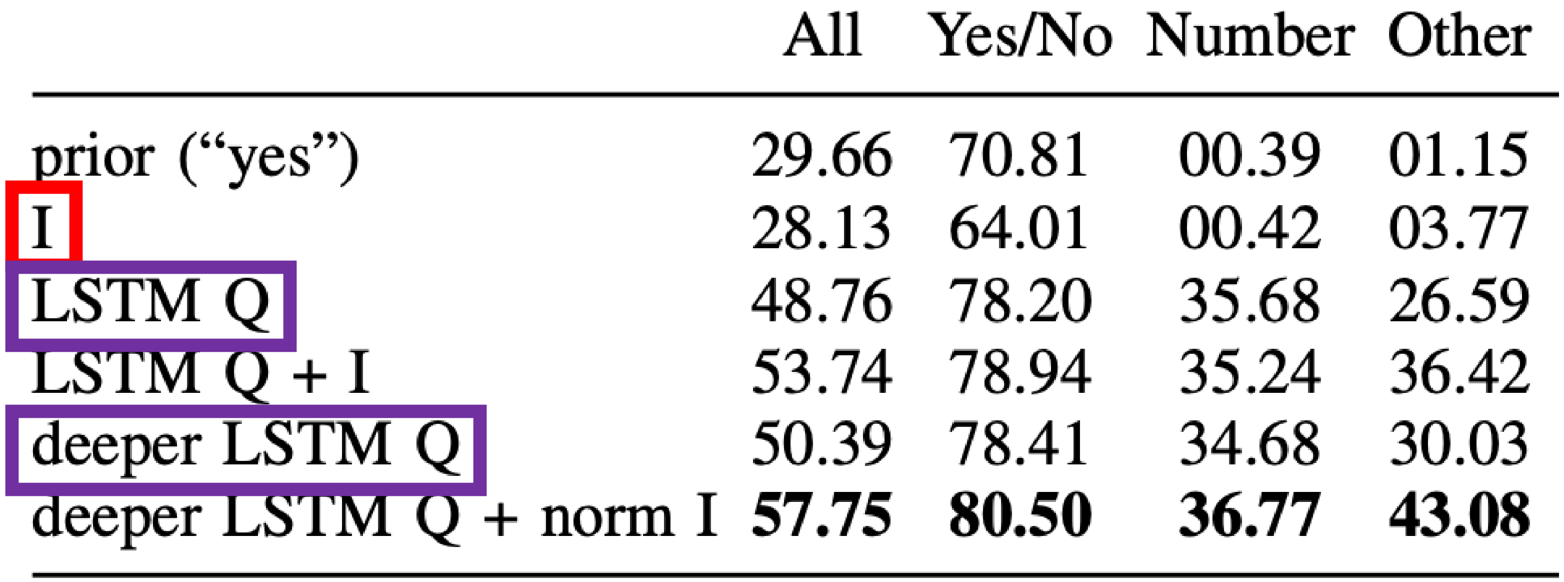


	All	Yes/No	Number	Other
prior (“yes”)	29.66	70.81	00.39	01.15
LSTM Q + I	53.74	78.94	35.24	36.42
deeper LSTM Q + norm I	**57.75**	**80.50**	**36.77**	**43.08**

What trends are observed across different question types?

Agrawal et al. VQA: Visual Question Answering. ICCV 2015.

Model Analysis: Group Discussion



		All	Yes/No	Number	Other
	prior ("yes")	29.66	70.81	00.39	01.15
Only image	I	28.13	64.01	00.42	03.77
Only question	LSTM Q	48.76	78.20	35.68	26.59
	LSTM Q + I	53.74	78.94	35.24	36.42
Only question	deeper LSTM Q	50.39	78.41	34.68	30.03
	deeper LSTM Q + norm I	**57.75**	**80.50**	**36.77**	**43.08**

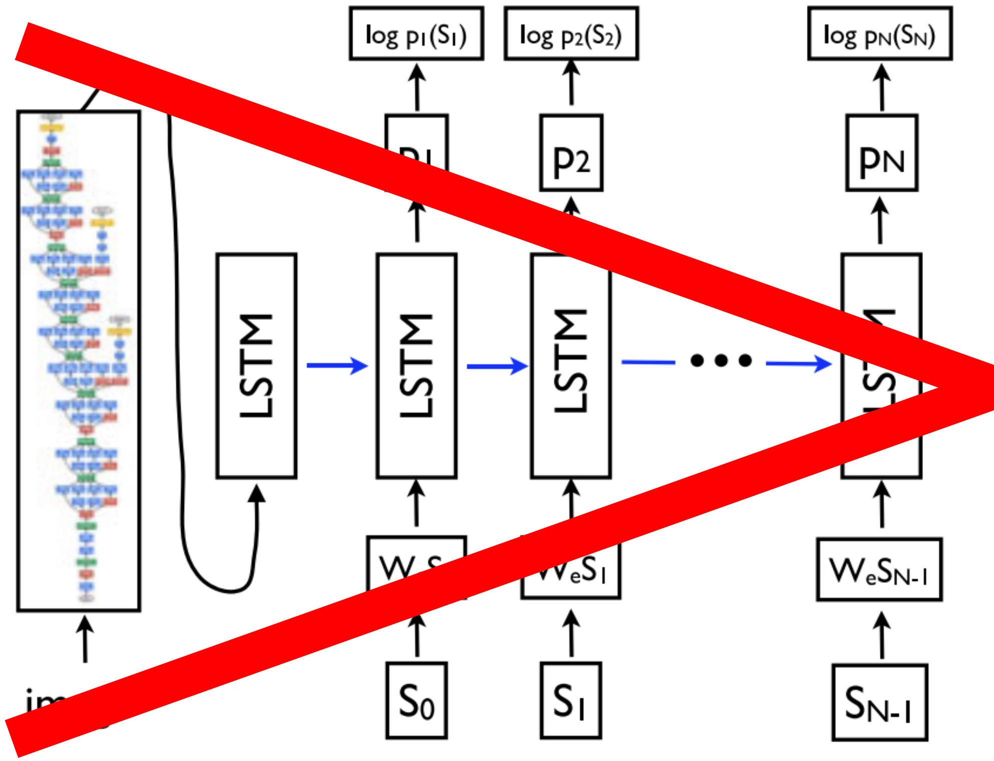
- How does each modality, vision and language, influence performance?
- Which modality is most predictive?

Agrawal et al. VQA: Visual Question Answering. ICCV 2015.

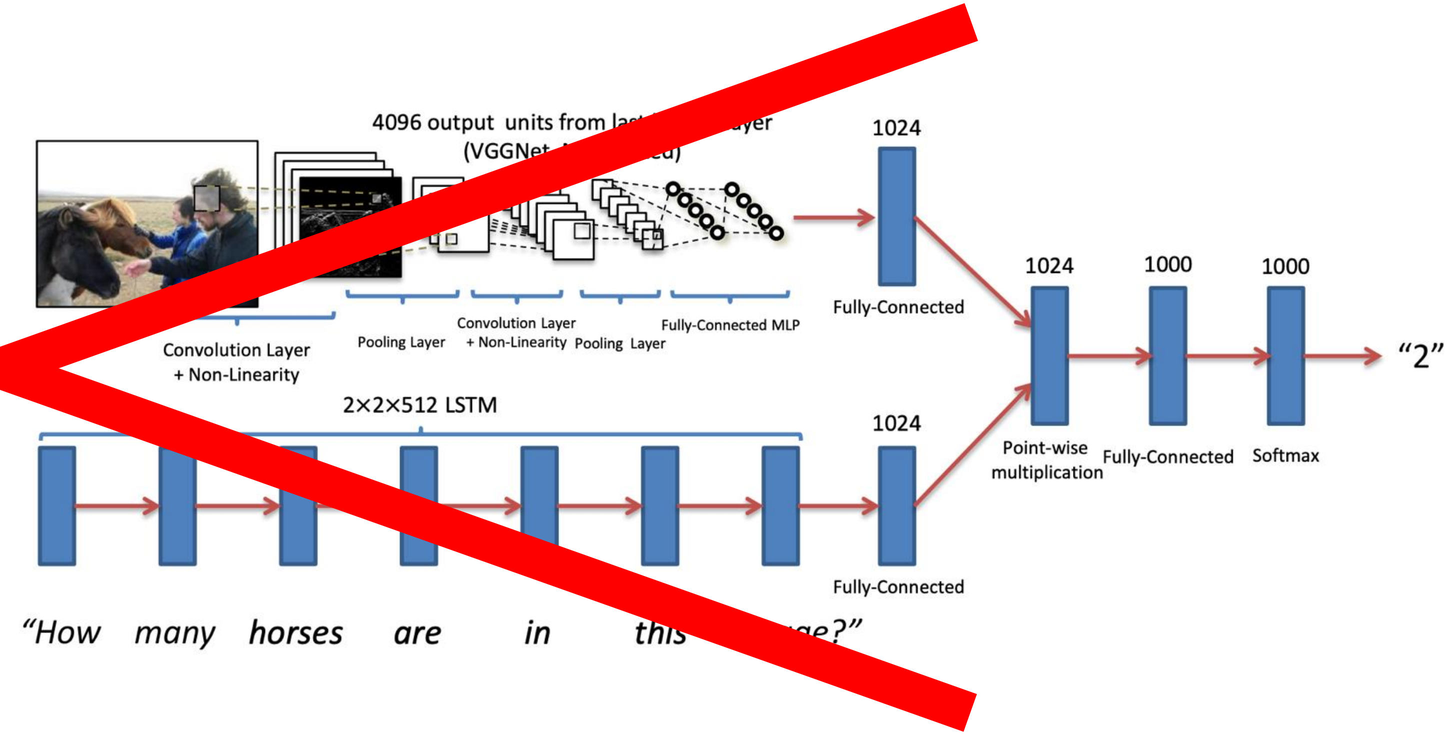
Today’s Topics

- Motivating applications
- Image captioning: pioneering dataset and model
- Visual question answering: pioneering dataset and model
- LXMERT: multimodal representations
- Programming tutorial

Idea: Pre-trained **Multimodal** Representation



Vinyals et al. Show and Tell: A Neural Image Caption Generator. CVPR 2015.

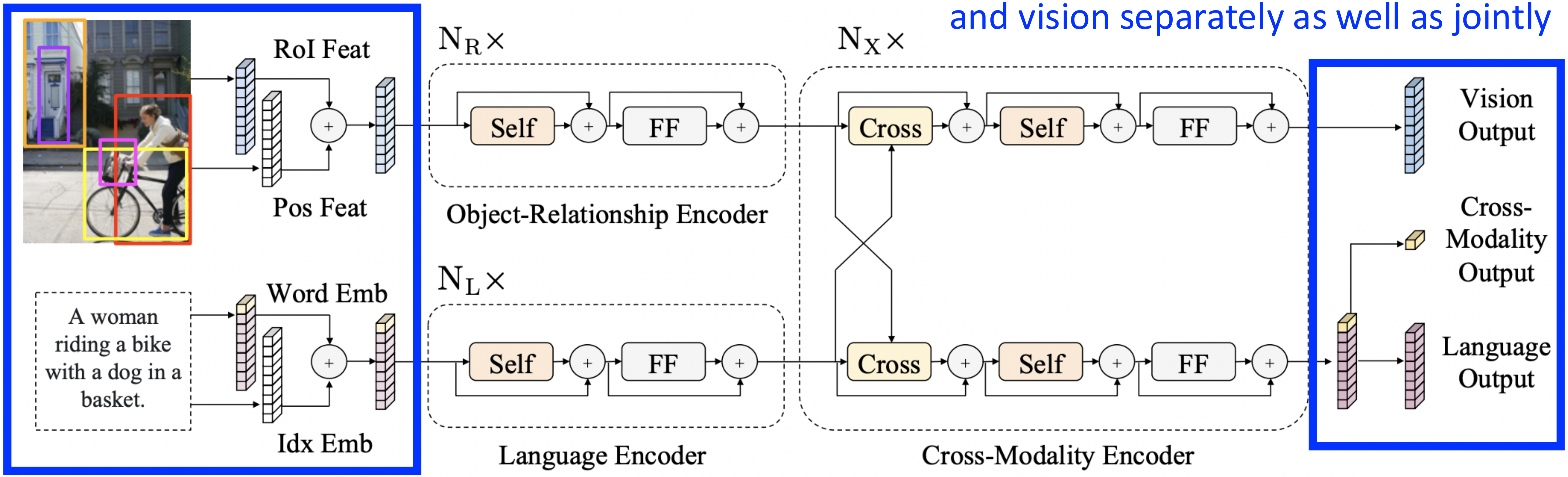


4096 output units from las
(VGGNet
Convolution Layer + Non-Linearity
Pooling Layer
Convolution Layer + Non-Linearity
Pooling Layer
Fully-Connected MLP
1024
Fully-Connected
2×2×512 LSTM
1024
Fully-Connected
1024
Point-wise multiplication
1000
Fully-Connected
1000
Softmax
"2"
"How many horses are in this
ge?"

Agrawal et al. VQA: Visual Question Answering. ICCV 2015.

LXMERT: Learning Cross-Modality Encoder Representations from Transformers

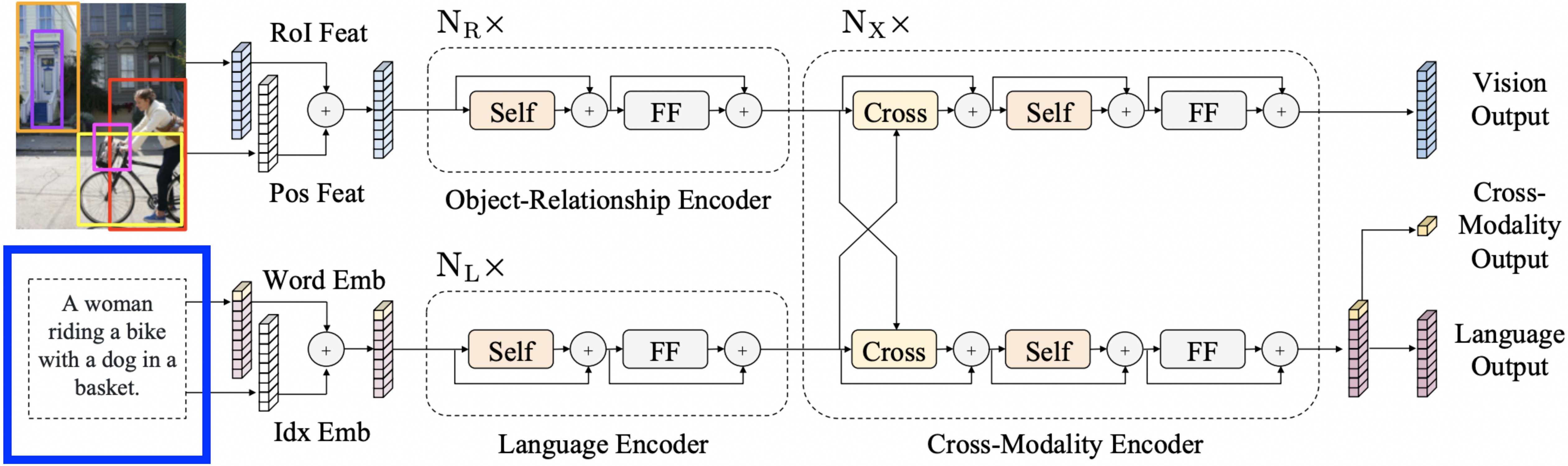
Generates representations for image and vision separately as well as jointly



Pretrains using language and vision input

Tan and Bansal, EMNLP 2019.

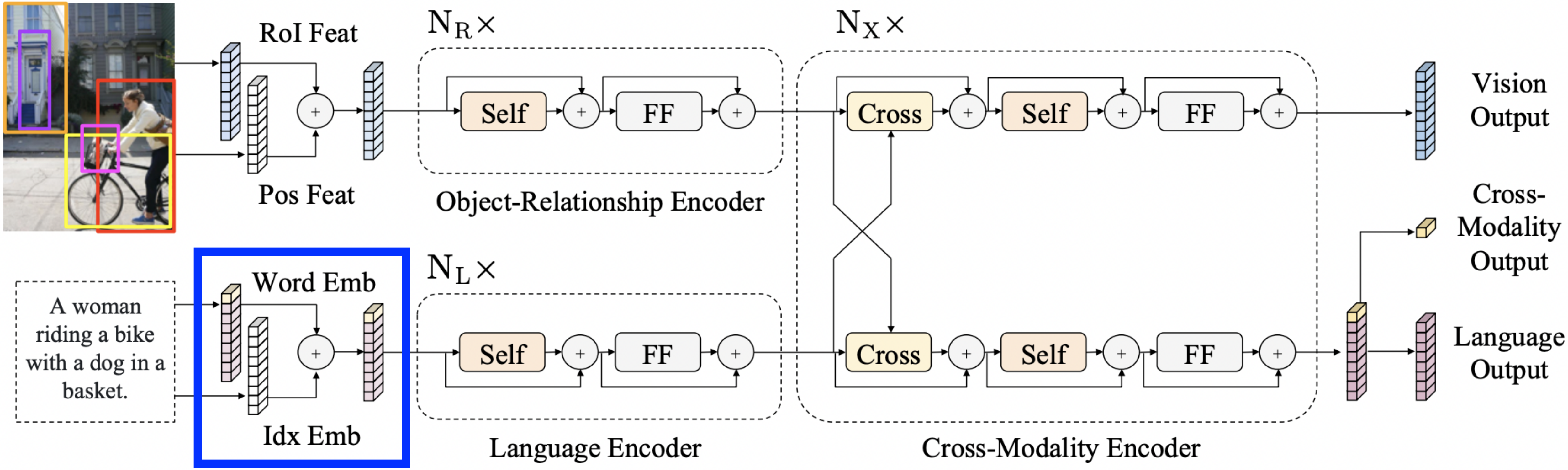
LXMERT: Language Input



[CLS] is added to the start of the sequence

Tan and Bansal, EMNLP 2019.

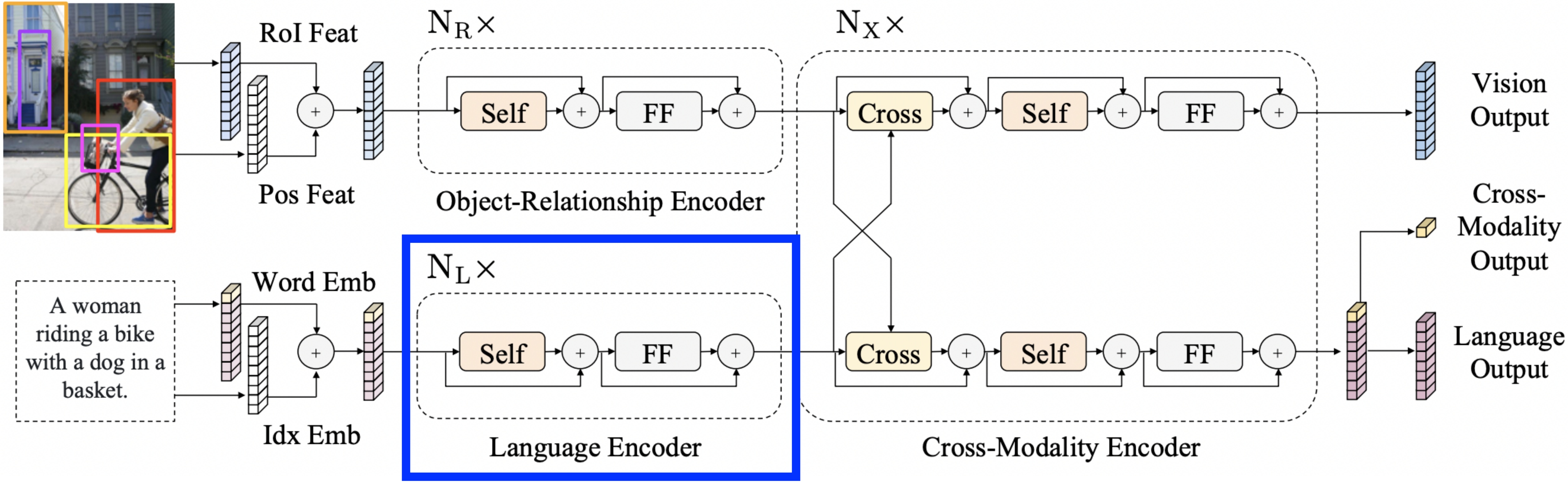
LXMERT: Language Input



Subword tokenization with WordPiece followed by representing each token as sum of its word embedding and position encoding

Tan and Bansal, EMNLP 2019.

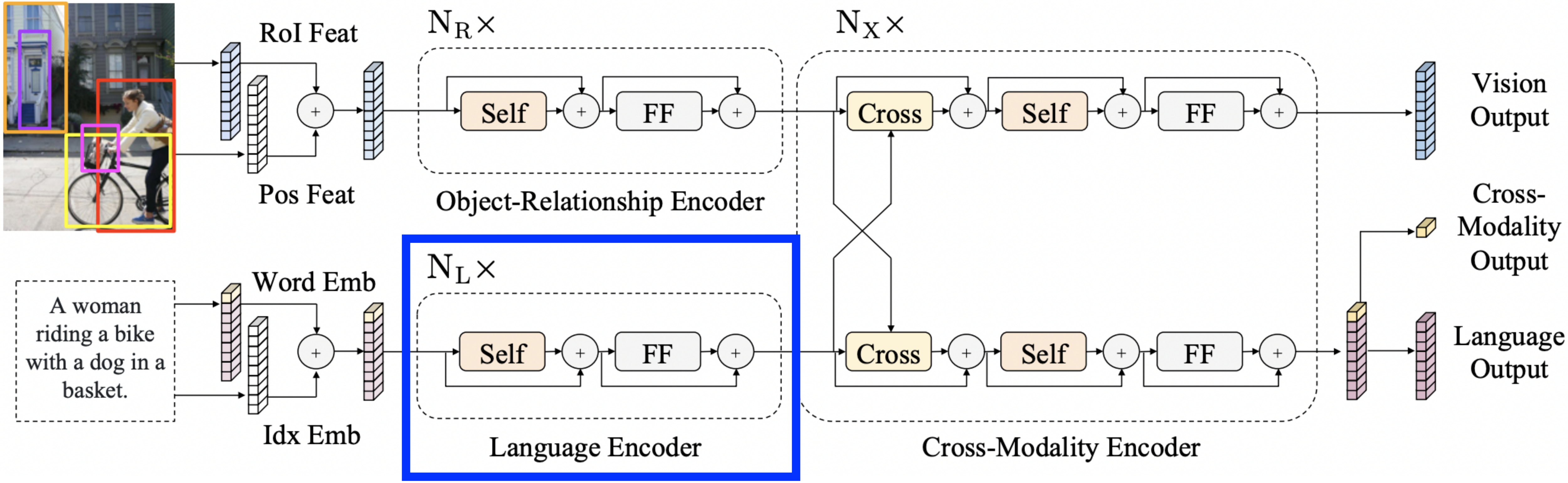
LXMERT: Language Input



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

Tan and Bansal, EMNLP 2019.

LXMERT: Language Input

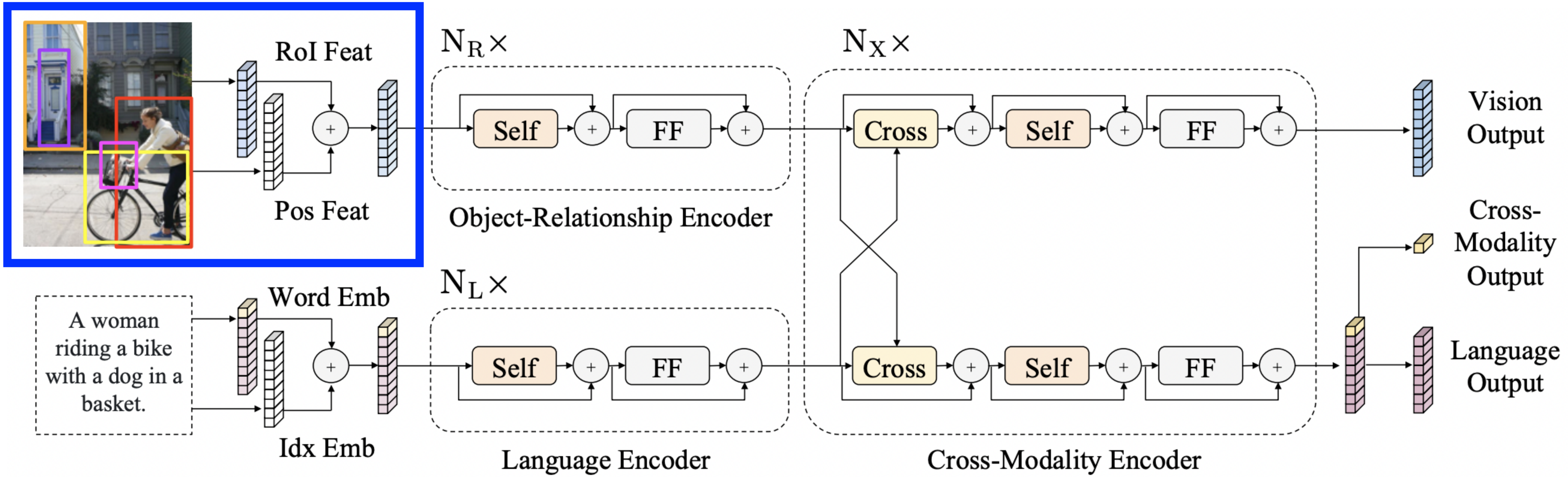


Transformer encoder (i.e., BERT); represents words with their relationships to all words

Tan and Bansal, EMNLP 2019.

LXMERT: Vision Input

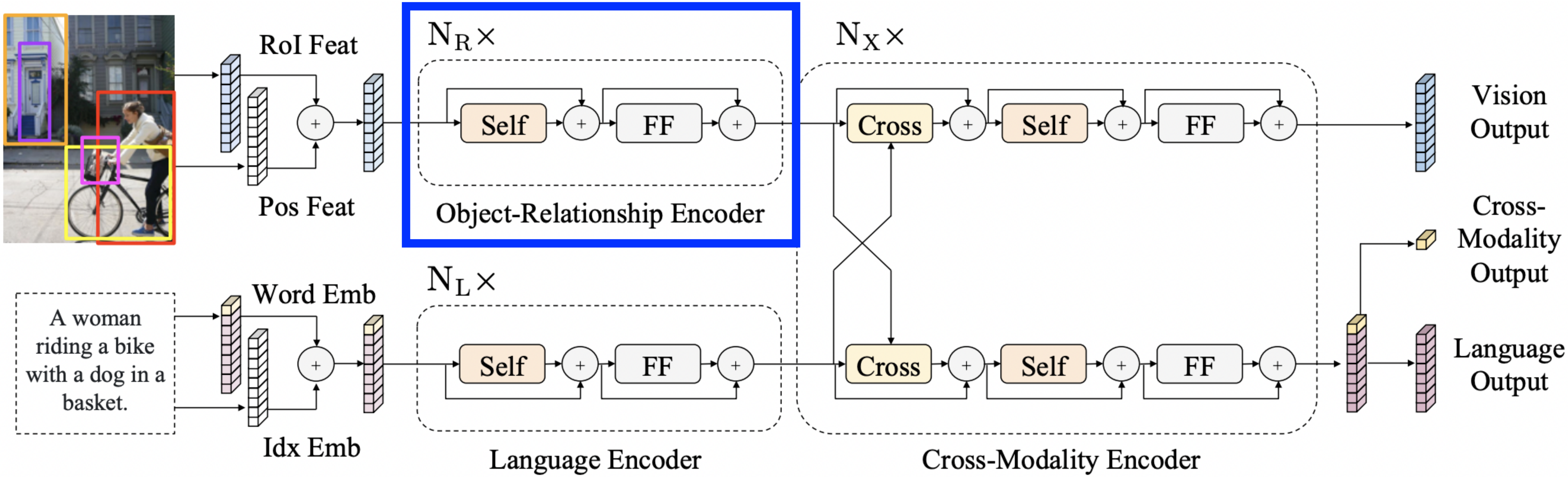
Each image is represented as a description of *m* objects detected with Faster R-CNN using features from Faster R-CNN and position encodings



Tan and Bansal, EMNLP 2019.

LXMERT: Architecture

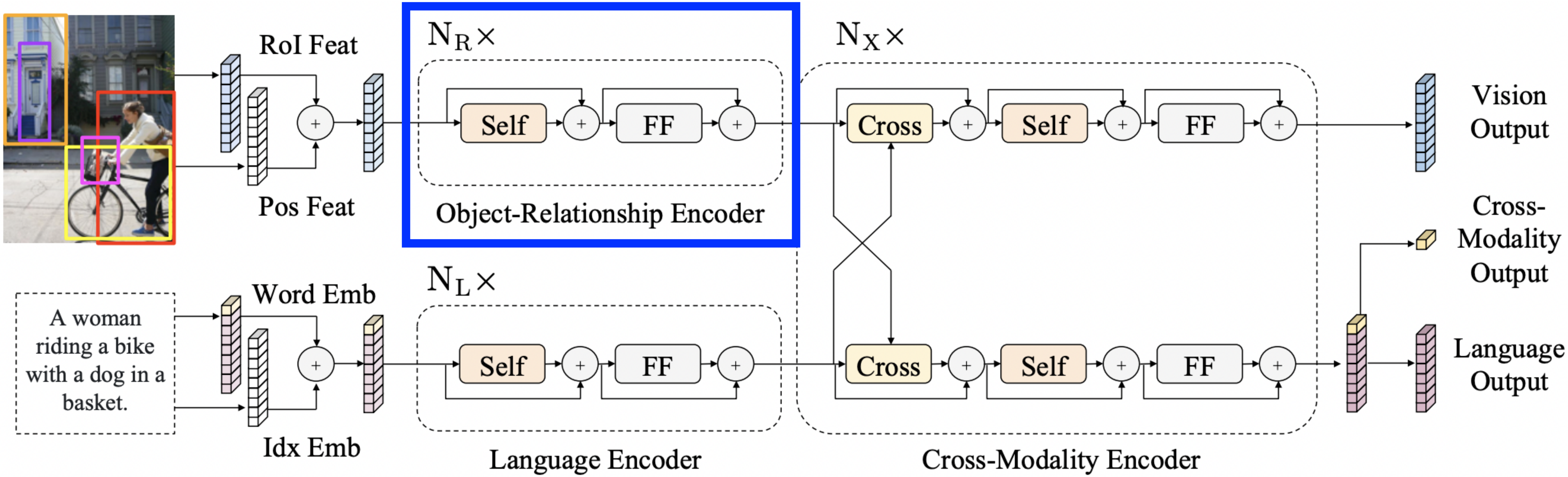
Transformer encoder (i.e., BERT);
what does its output represent?



Tan and Bansal, EMNLP 2019.

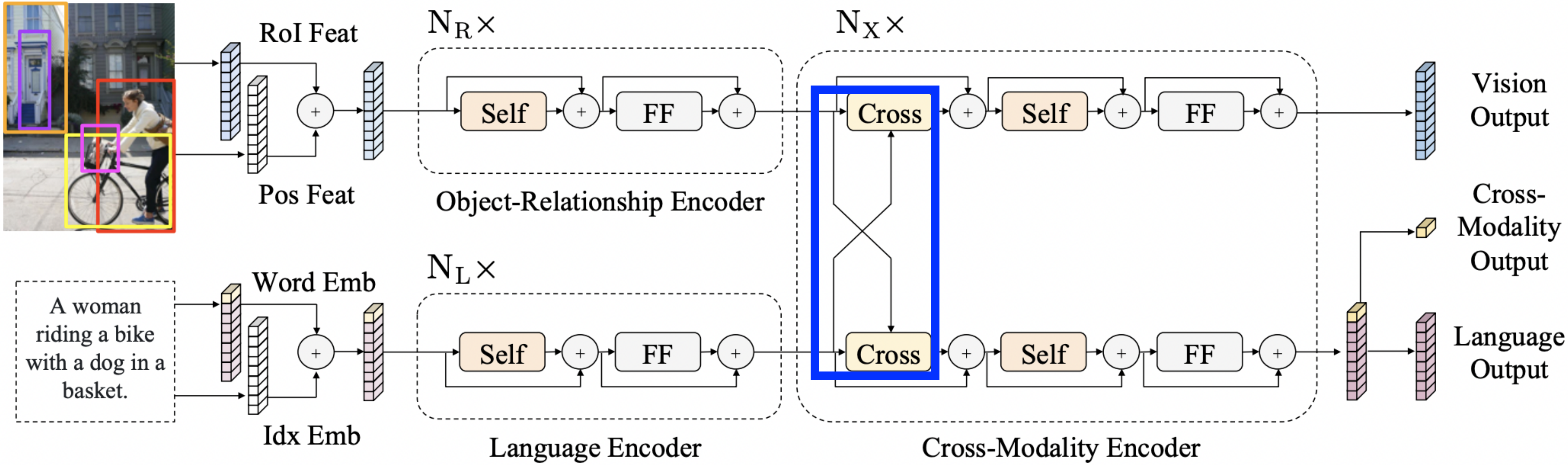
LXMERT: Architecture

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



Tan and Bansal, EMNLP 2019.

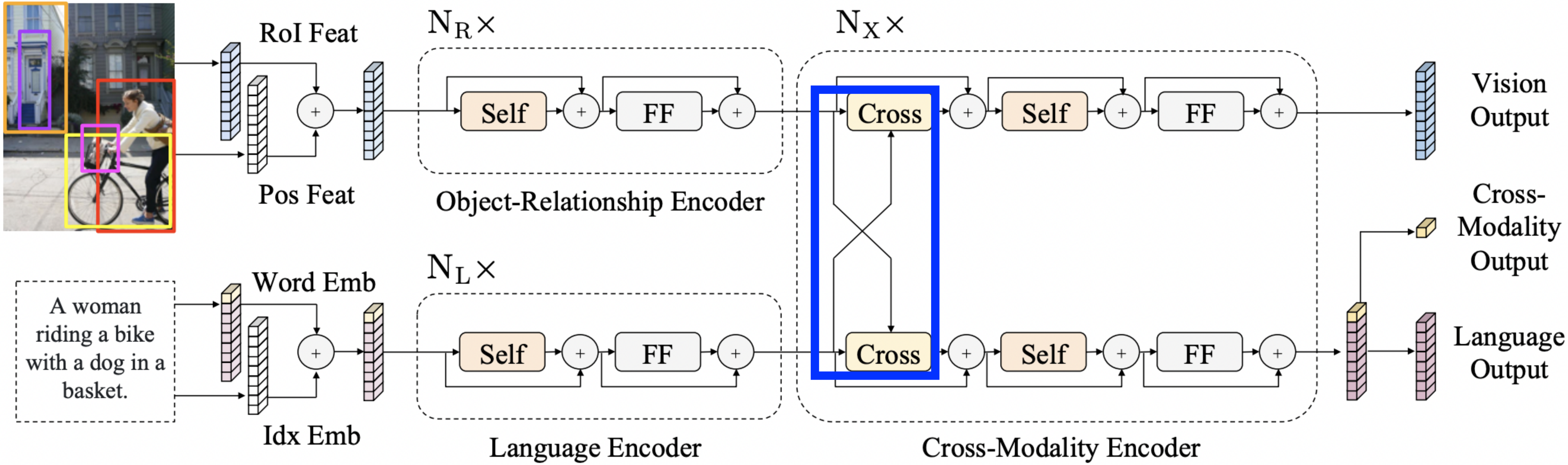
LXMERT: Architecture



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

Tan and Bansal, EMNLP 2019.

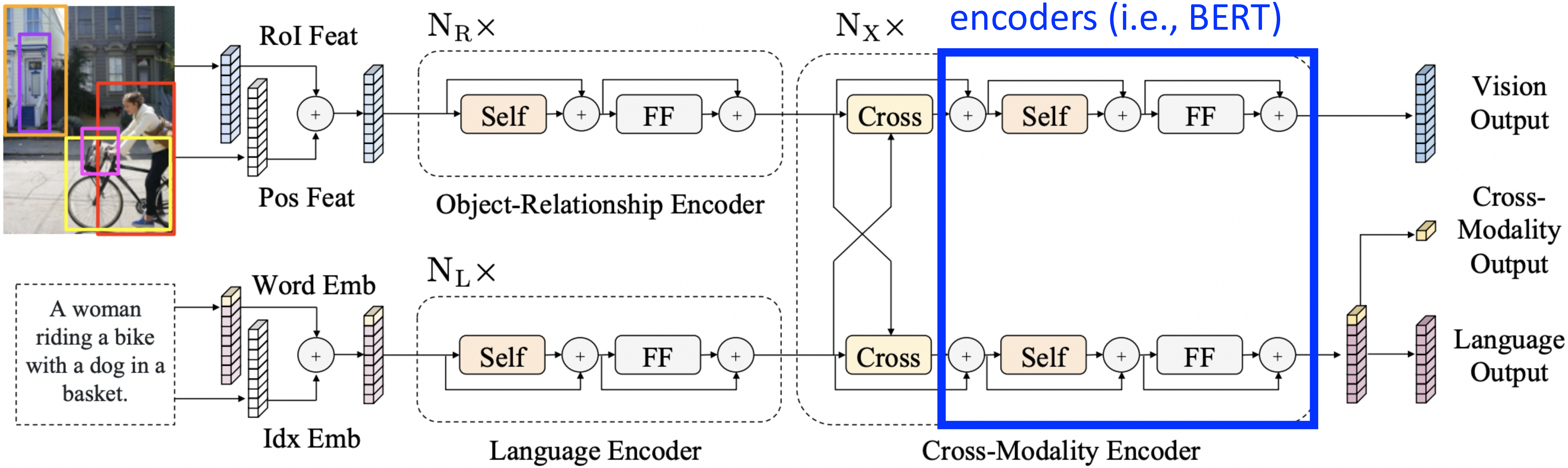
LXMERT: Architecture



Two cross-attention layers are functions of the "query" with "keys" and "values" from the opposite modalities' features

Tan and Bansal, EMNLP 2019.

LXMERT: Architecture

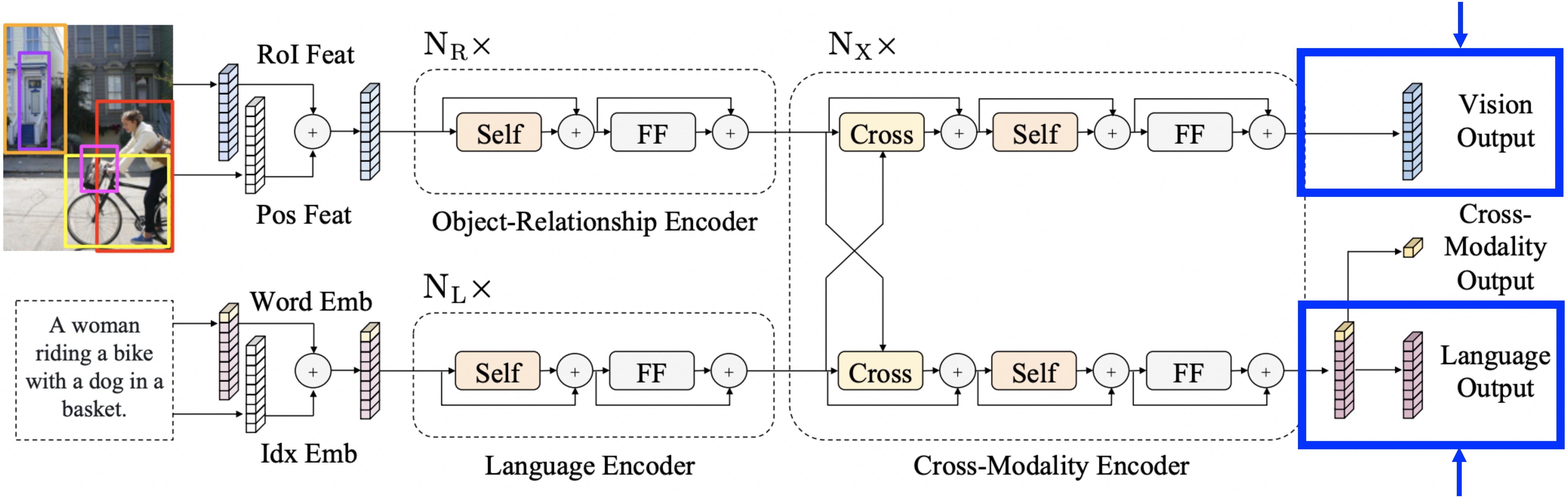


Two more layers of transformer encoders (i.e., BERT)

Tan and Bansal, EMNLP 2019.

LXMERT: Output

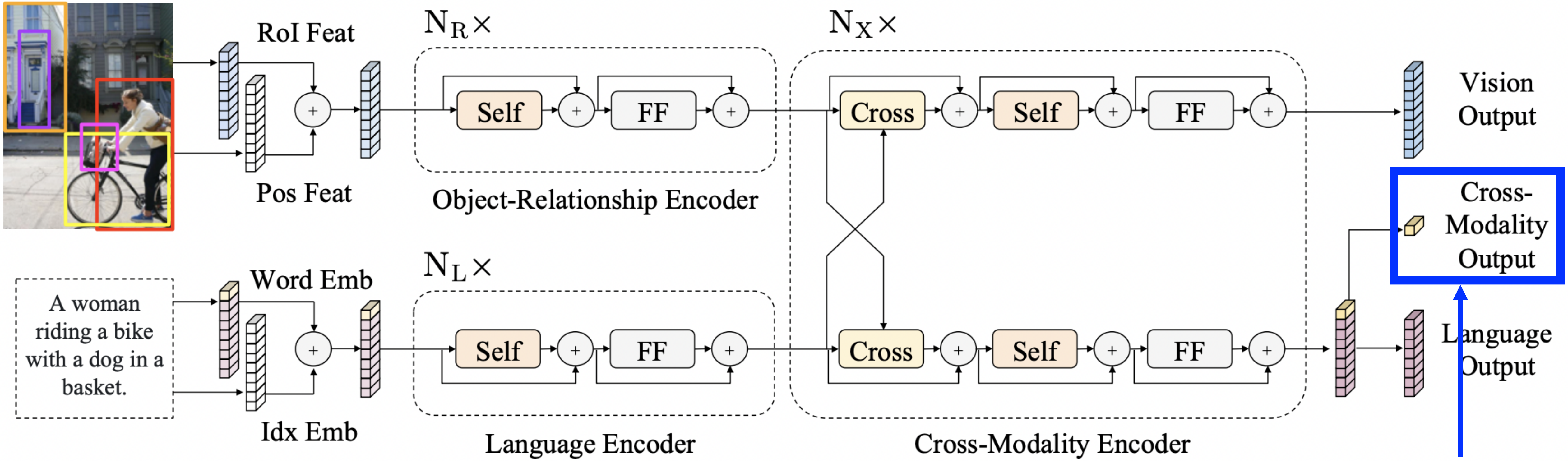
New representation of input detected objects



New representation of input language sequence

Tan and Bansal, EMNLP 2019.

LXMERT: Output

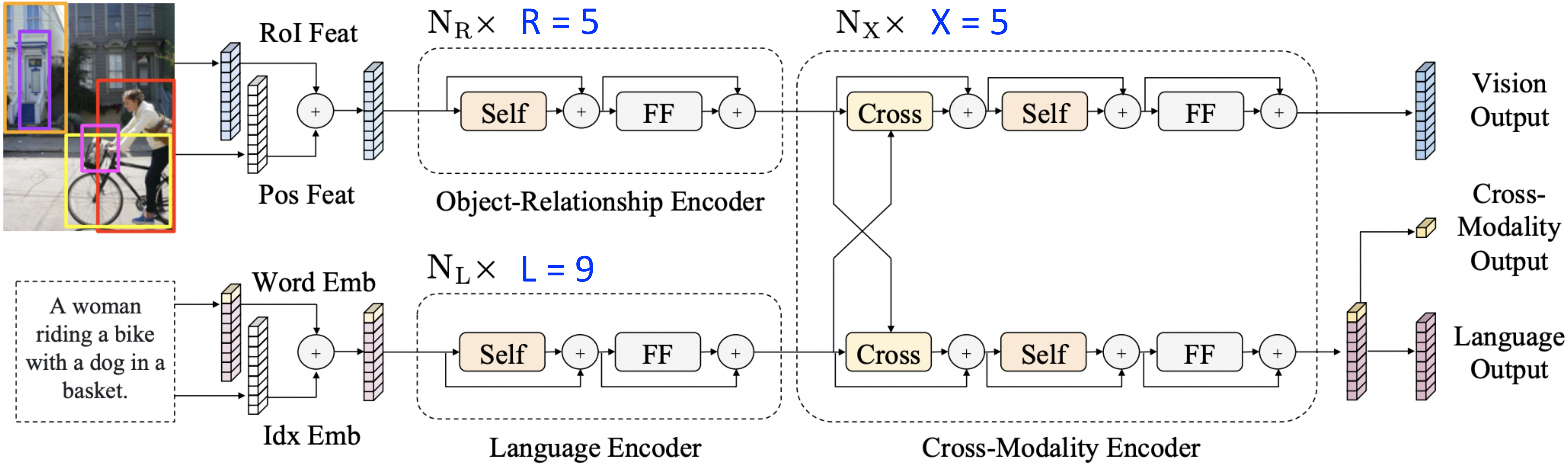


Cross-modality representation is the [CLS] token appended at the start of the sentence

Tan and Bansal, EMNLP 2019.

LXMERT: Implementation Details

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

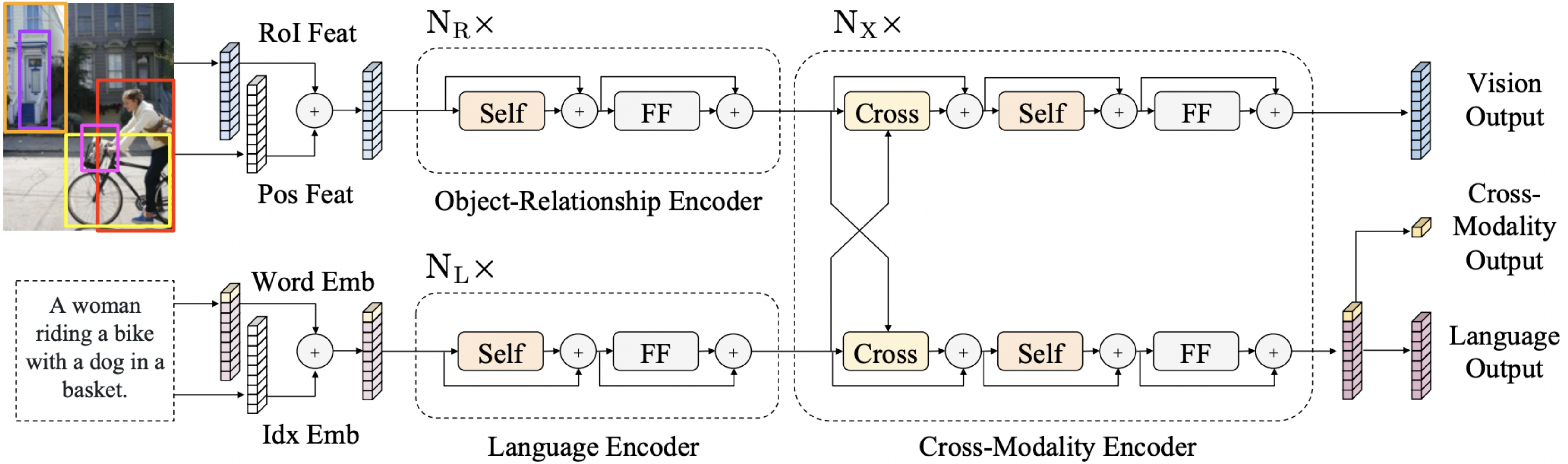


Number of layers mimics the size of BERT (base), with 12 layers;

i.e., (5+9)/2 + 5

Tan and Bansal, EMNLP 2019.

LXMERT: Implementation Details

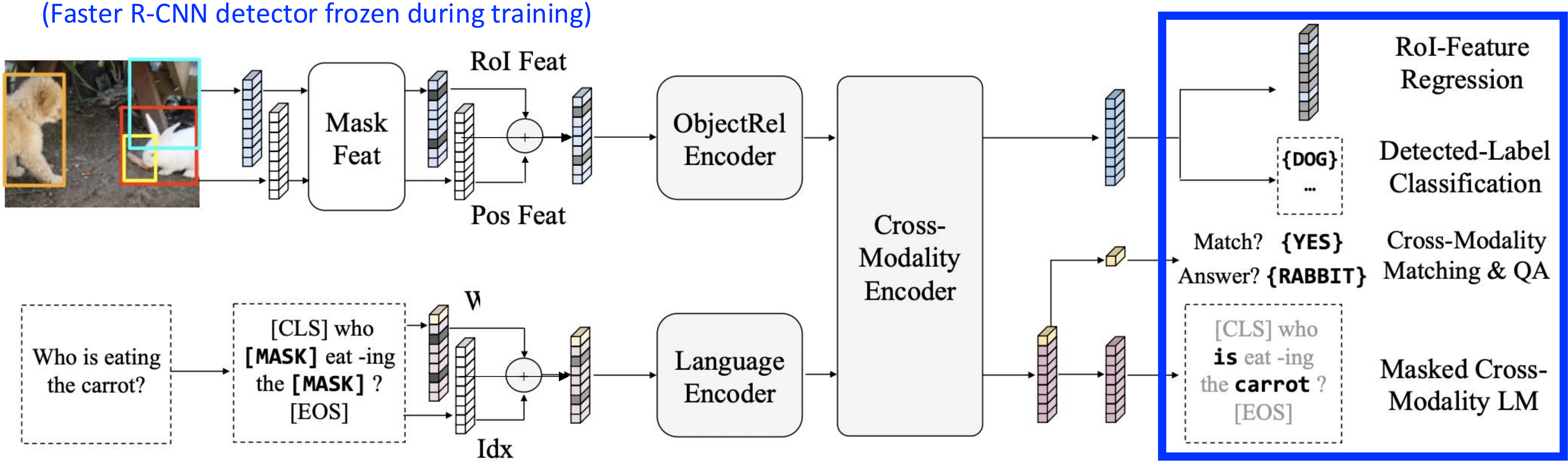


What might be strengths and limitations of the resulting feature representations based on the architecture used?

Tan and Bansal, EMNLP 2019.

LXMERT: Pretraining Uses Sum of 5 Task Losses

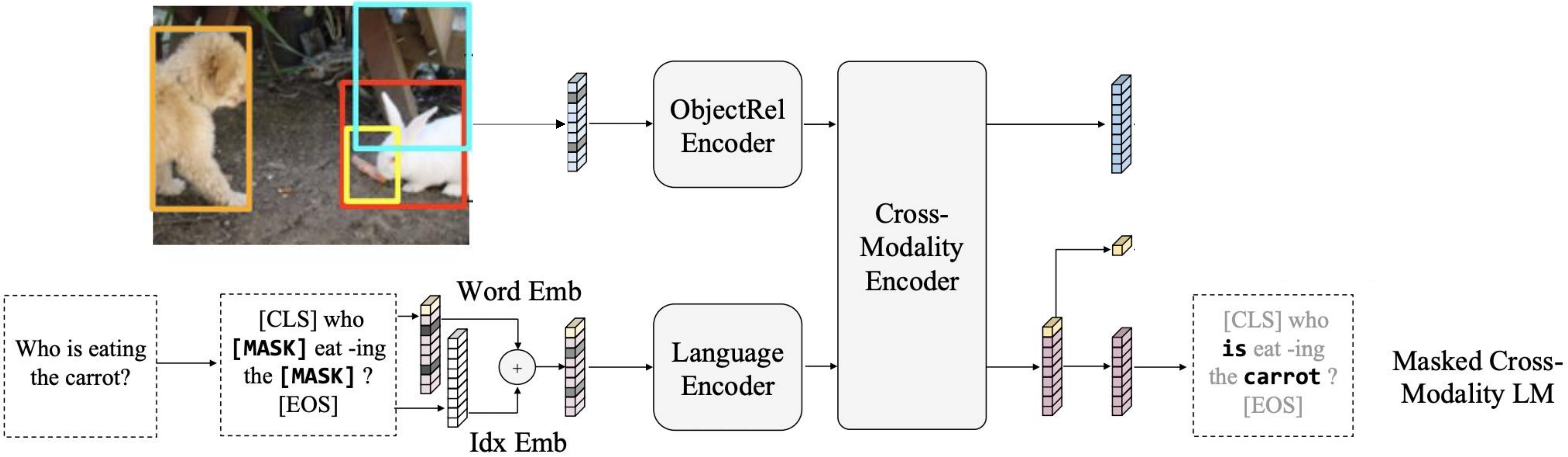
(Faster R-CNN detector frozen during training)


RoI Feat
Mask Feat
Pos Feat
ObjectRel Encoder
Cross-Modality Encoder
Who is eating the carrot?
[CLS] who [MASK] eat -ing the [MASK] ? [EOS]
W
Idx
Language Encoder
RoI-Feature Regression
{DOG}
...
Detected-Label Classification
Match? {YES}
Answer? {RABBIT}
Cross-Modality Matching & QA
[CLS] who is eat -ing the carrot ? [EOS]
Masked Cross-Modality LM

Took 10 days on 4 Titan Xp GPUs

Tan and Bansal, EMNLP 2019.

LXMERT: Pretraining Task 1 (Language)



ObjectRel Encoder
Cross-Modality Encoder
Language Encoder
Word Emb
Idx Emb
+
Who is eating the carrot?
[CLS] who [MASK] eat -ing the [MASK] ? [EOS]
[CLS] who is eat -ing the carrot ? [EOS]
Masked Cross-Modality LM

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

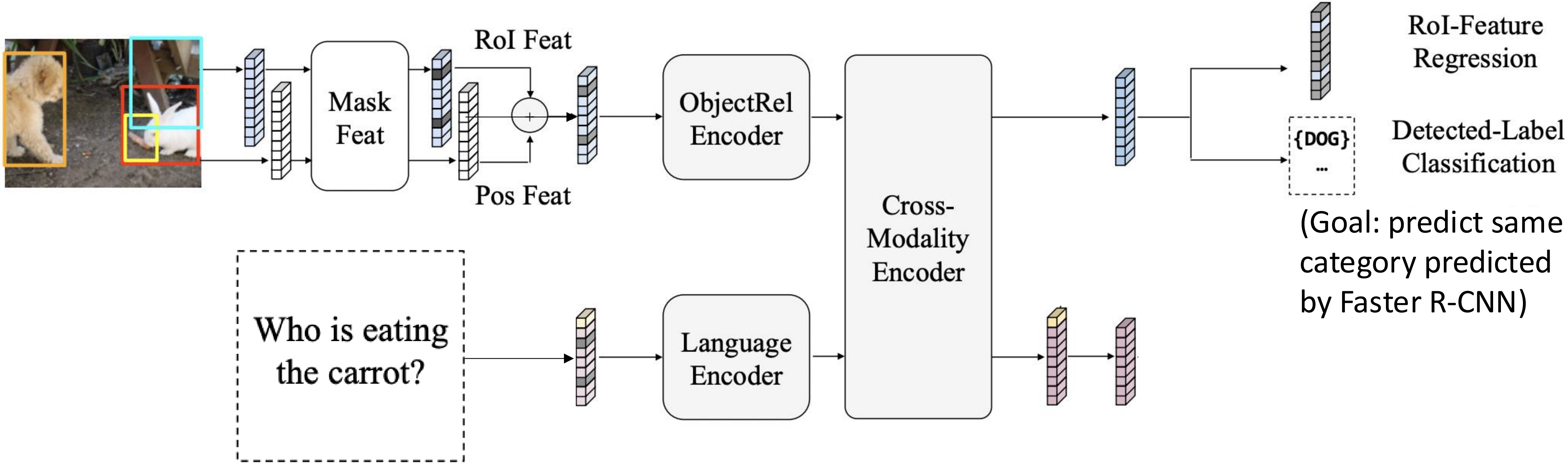
Unlike BERT, vision modality can resolve language ambiguity; e.g., shows what is being eaten

Tan and Bansal, EMNLP 2019.

LXMERT: Pretraining Tasks 2 & 3 (Vision)

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

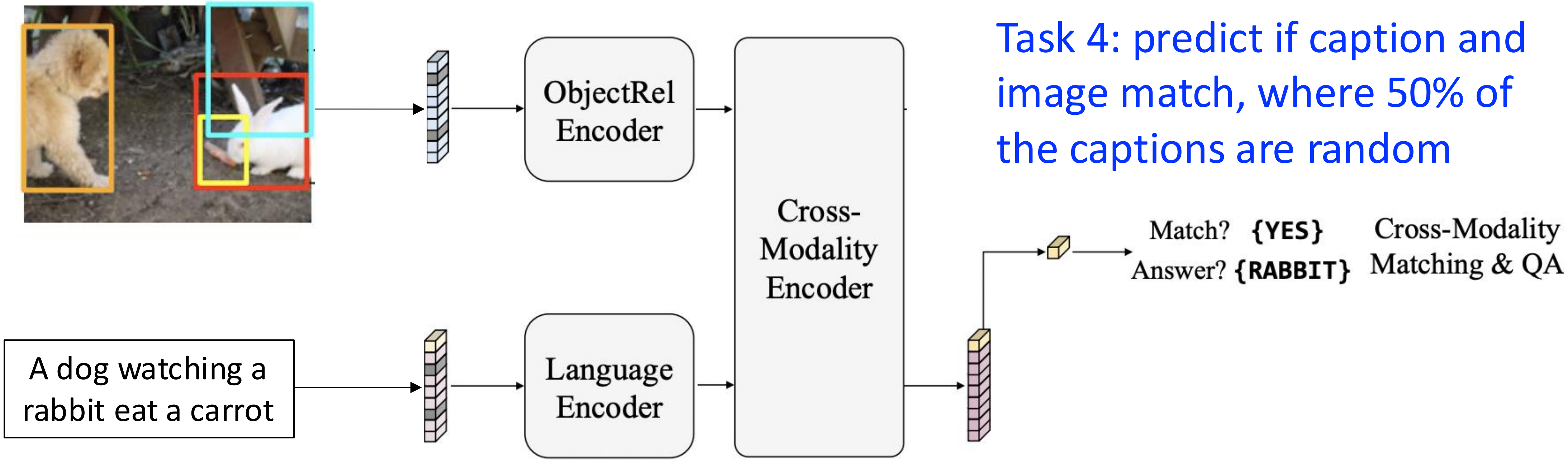
Knowledge about other objects and the language should help predict masked objects



(Goal: predict same category predicted by Faster R-CNN)

Tan and Bansal, EMNLP 2019.

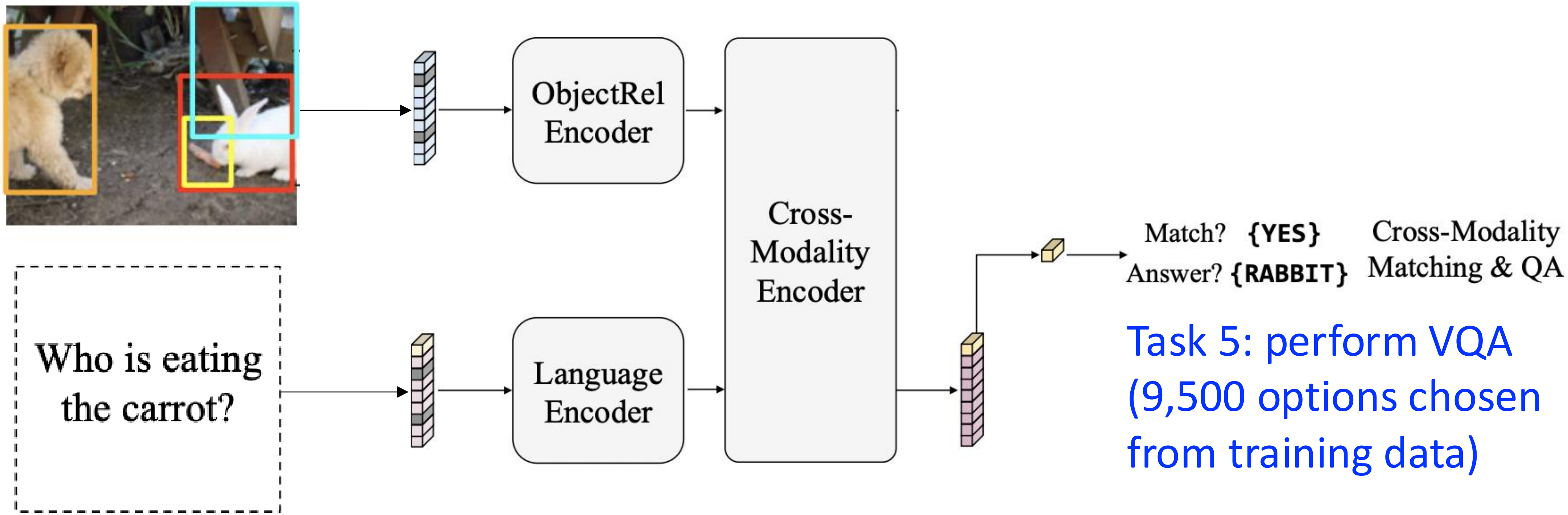
LXMERT: Pretraining Tasks 4 & 5 (Both Modalities)



Task 4: predict if caption and image match, where 50% of the captions are random

Tan and Bansal, EMNLP 2019.

LXMERT: Pretraining Tasks 4 & 5 (Both Modalities)



Task 5: perform VQA (9,500 options chosen from training data)

Tan and Bansal, EMNLP 2019.

LXMERT: Pretraining Data

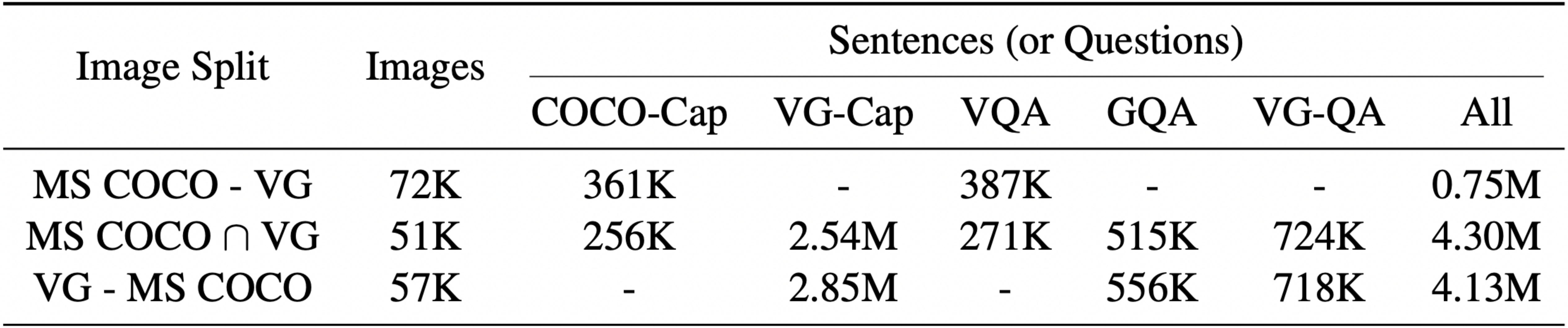


Image Split	Images	Sentences (or Questions)					
		COCO-Cap	VG-Cap	VQA	GQA	VG-QA	All
MS COCO - VG	72K	361K	-	387K	-	-	0.75M
MS COCO ∩ VG	51K	256K	2.54M	271K	515K	724K	4.30M
VG - MS COCO	57K	-	2.85M	-	556K	718K	4.13M

All images from MS COCO and Visual Genome, which were collected by scraping images from the photo-sharing website Flickr

(Visual Genome includes the MS COCO images)

Tan and Bansal, EMNLP 2019.

LXMERT: Pretraining Data

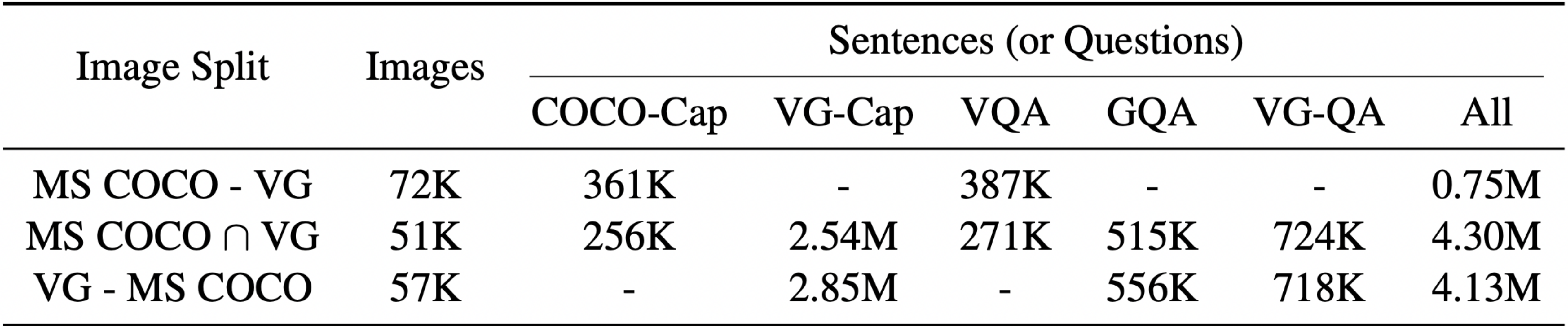


Image Split	Images	Sentences (or Questions)					
		COCO-Cap	VG-Cap	VQA	GQA	VG-QA	All
MS COCO - VG	72K	361K	-	387K	-	-	0.75M
MS COCO ∩ VG	51K	256K	2.54M	271K	515K	724K	4.30M
VG - MS COCO	57K	-	2.85M	-	556K	718K	4.13M

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

Tan and Bansal, EMNLP 2019.

LXMERT: Pretraining Data

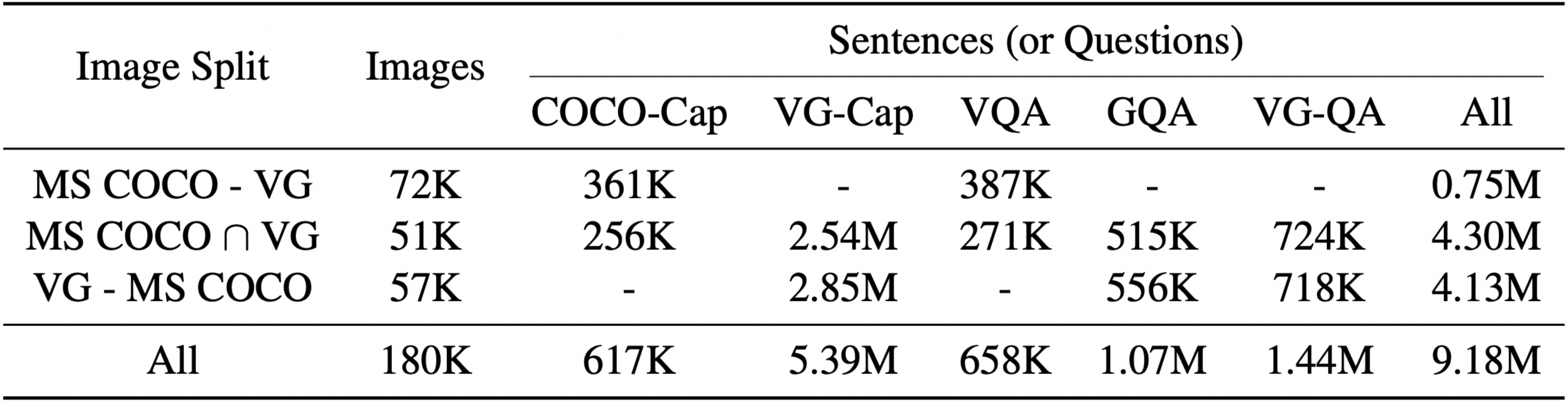
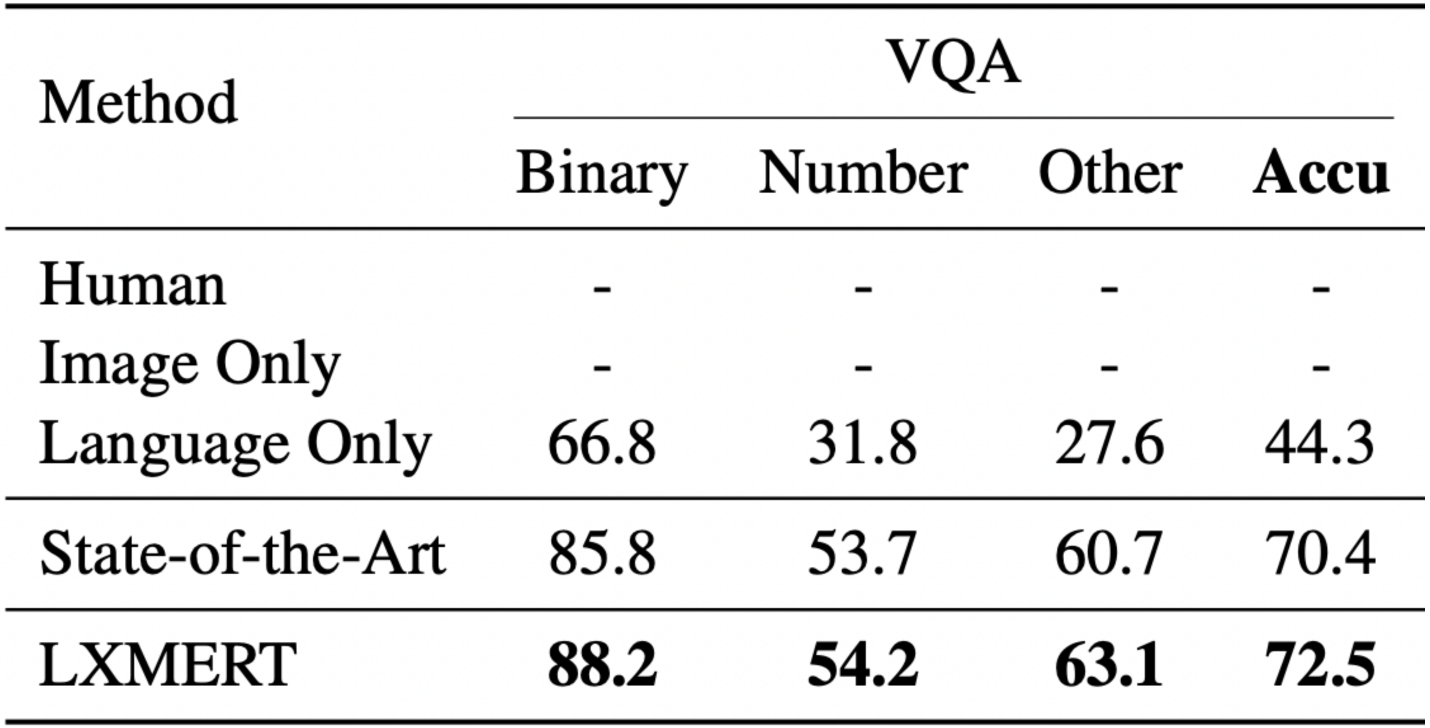


Image Split	Images	Sentences (or Questions)					
		COCO-Cap	VG-Cap	VQA	GQA	VG-QA	All
MS COCO - VG	72K	361K	-	387K	-	-	0.75M
MS COCO ∩ VG	51K	256K	2.54M	271K	515K	724K	4.30M
VG - MS COCO	57K	-	2.85M	-	556K	718K	4.13M
All	180K	617K	5.39M	658K	1.07M	1.44M	9.18M

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)

Tan and Bansal, EMNLP 2019.

LXMERT: Fine-Tuning Experimental Results

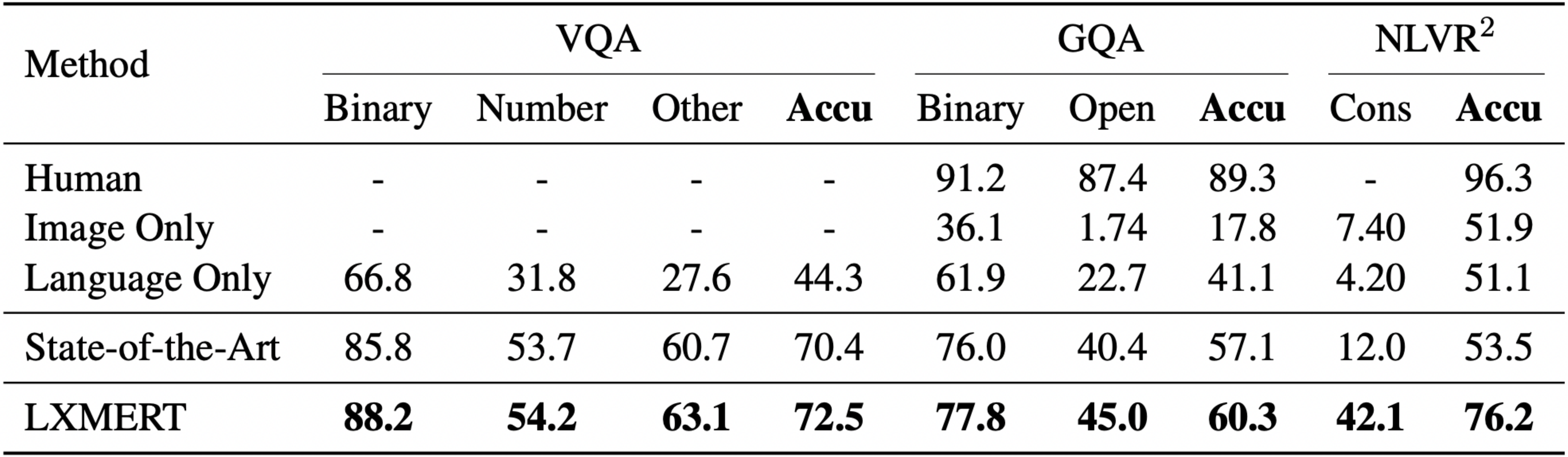


Method	VQA			
	Binary	Number	Other	**Accu**
Human	-	-	-	-
Image Only	-	-	-	-
Language Only	66.8	31.8	27.6	44.3
State-of-the-Art	85.8	53.7	60.7	70.4
LXMERT	**88.2**	**54.2**	**63.1**	**72.5**

State-of-the-art performance, with stronger gains over prior work for questions leading to "binary" and "other" answers

Tan and Bansal, EMNLP 2019.

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**

State-of-the-art performance for an additional VQA dataset and a visual reasoning task (i.e., does statement describe two images or not?)

Tan and Bansal, EMNLP 2019.

Today’s Topics

- Motivating applications
- Image captioning: pioneering dataset and model
- Visual question answering: pioneering dataset and model
- LXMERT: multimodal representations
- Programming tutorial

Today’s Topics

- Motivating applications
- Image captioning: pioneering dataset and model
- Visual question answering: pioneering dataset and model
- LXMERT: multimodal representations
- Programming tutorial

The End