LXMERT:
Learning Cross-Modality Encoder Representations from Transformers

Hao Tan, Mohit Bansal

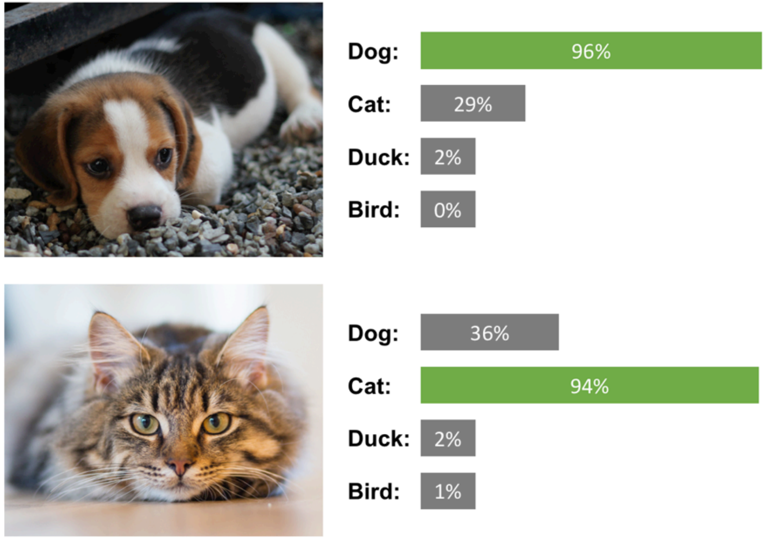
haotan, mbansal@cs.unc.edu

Vision Tasks

Understand Visual Concepts

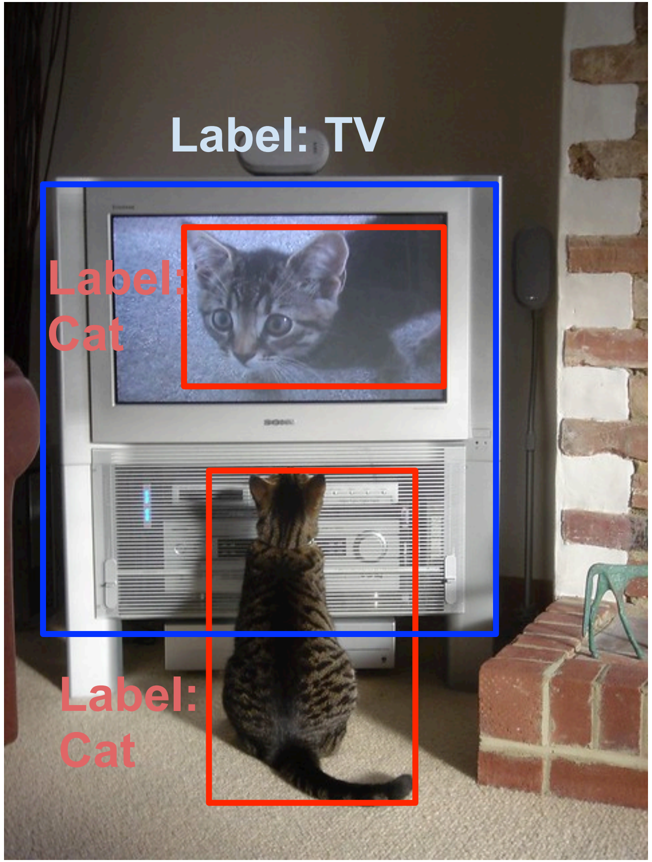
Image Classification: What is it?

Predict the **class label** of the image.



Object Detection: What and Where are They?

Find the objects and then predict their regions and labels.



Vision-and-Language Tasks

A series of tasks that require both
vision and language information to complete.

Image Captioning: Describe the Image

Use one natural-language sentence to describe the content in the image.

An orange cat sits in the suitcase ready to be packed.

-- One of my favorite examples in MS COCO [Lin, ECCV 2009]

Visual Question Answering

Answer a question about the image.

What color are her eyes? Answer: Black
What is the mustache made of? Answer: Bananas

Pre-training → Fine-tuning

A general methodology to solve [vision tasks] and [language tasks].

Pre-training → Fine-tuning

Pre-training:

Large-scale Datasets →

General Tasks →

Backbone Model

Fine-tuning:

Relative Small Datasets →

Specific Tasks →

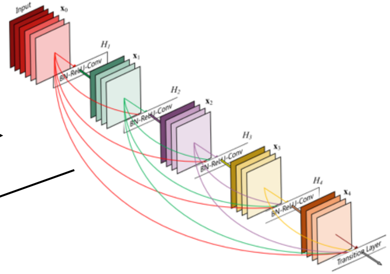
Task-dependent Model

Vision: Pre-training → Fine-tuning

Visual Pre-training:

ImageNet
[Deng, CVPR 2009]

Image Classification

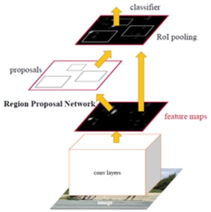


DenseNet
[Huang,CVPR 2017]

Visual Fine-tuning:

MS COCO
[Lin, ECCV 2009]

Object Detection

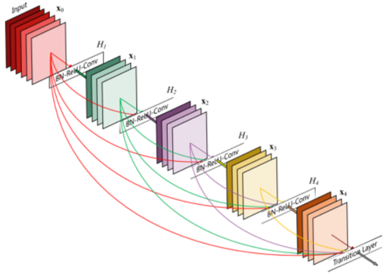


Faster RCNN
[Ren, NeurIPS 2015]

Language: Pre-training

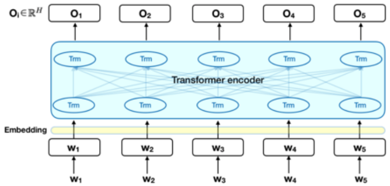
Visual Pre-training:

Image Classification



Language Pre-training:

Language Model



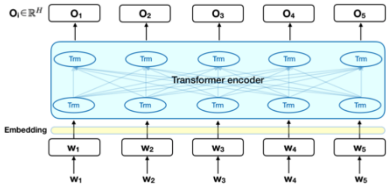
Language: Pre-training

ELMo
[Peters, NAACL 2018]

BERT
[Devlin, NAACL 2019]

Language
Pre-training:

Language
Model

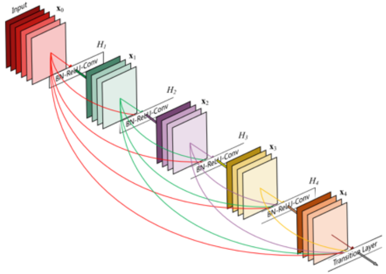


Fine-tuning

Detection, Segmentation, Identification, ...

Visual Pre-training:

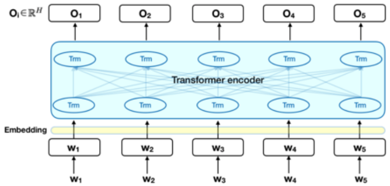
Image Classification



Question Answering, Sentiment Analysis, ...

Language Pre-training:

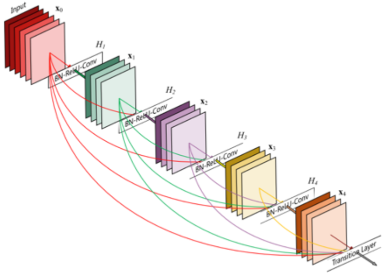
Language Model



Fine-tuning on Vision and Language Tasks?

Visual Pre-training:

→ Image Classification →

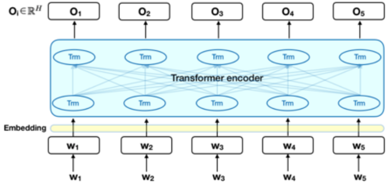


Fusion Module

Visual Question Answering, Navigation, Grounding, ...

Language Pre-training:

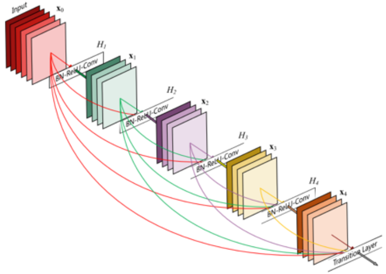
→ Language Model →



Fine-tuning on Vision and Language Tasks?

Visual Pre-training:

Image Classification

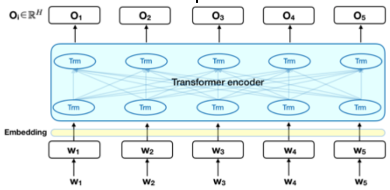


Fusion Module

Visual Question Answering, Navigation, Grounding, ...

Language Pre-training:

Language Model

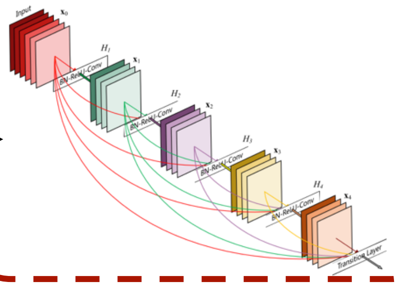


Pre-trained language modules do not help.

Fine-tuning on Vision and Language Tasks?

Visual Pre-training:

Image Classification

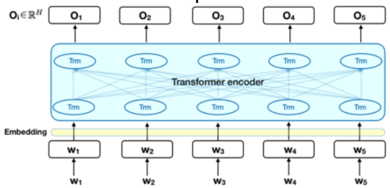


Fusion Module

Visual Question Answering, Navigation, Grounding, ...

Language Pre-training:

Language Model

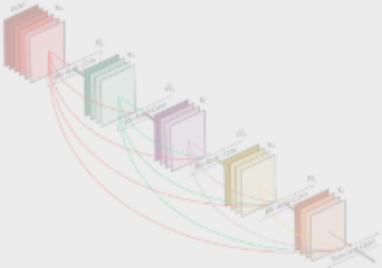


Single-modality pre-training is not aware of cross-modality relationships.

Pre-train for Vision and Language jointly?

Visual Pre-training:

Image Classification

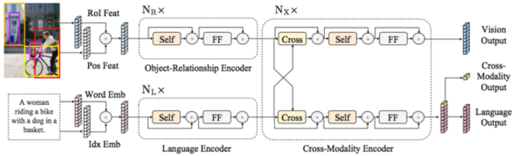


Data? → Pre-training Method? → Model?

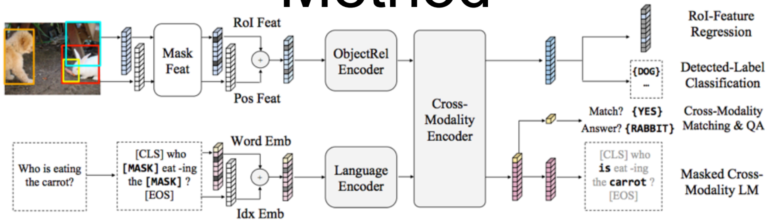
Language Pre-training:

Language Model

LXMERT (Learning Cross-Modality Encoder Representations from Transformers)



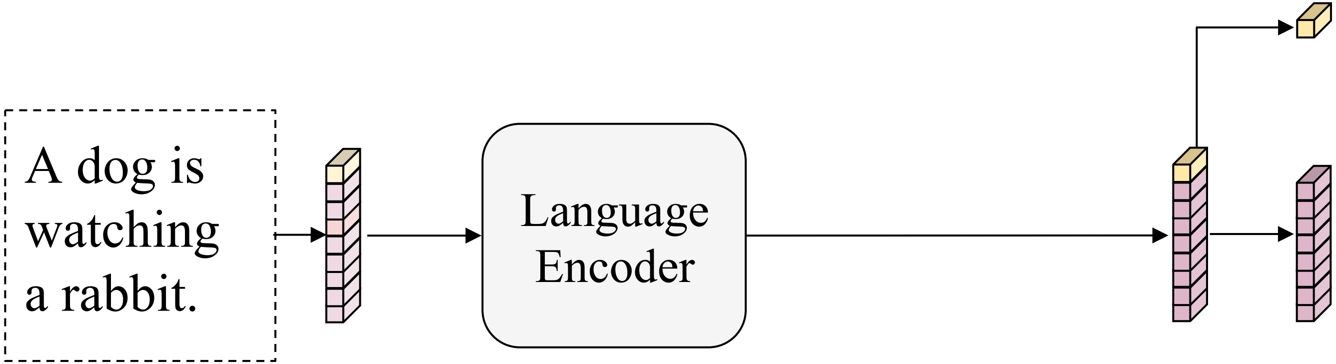
Data → Pre-training Method → Model



LXMERT

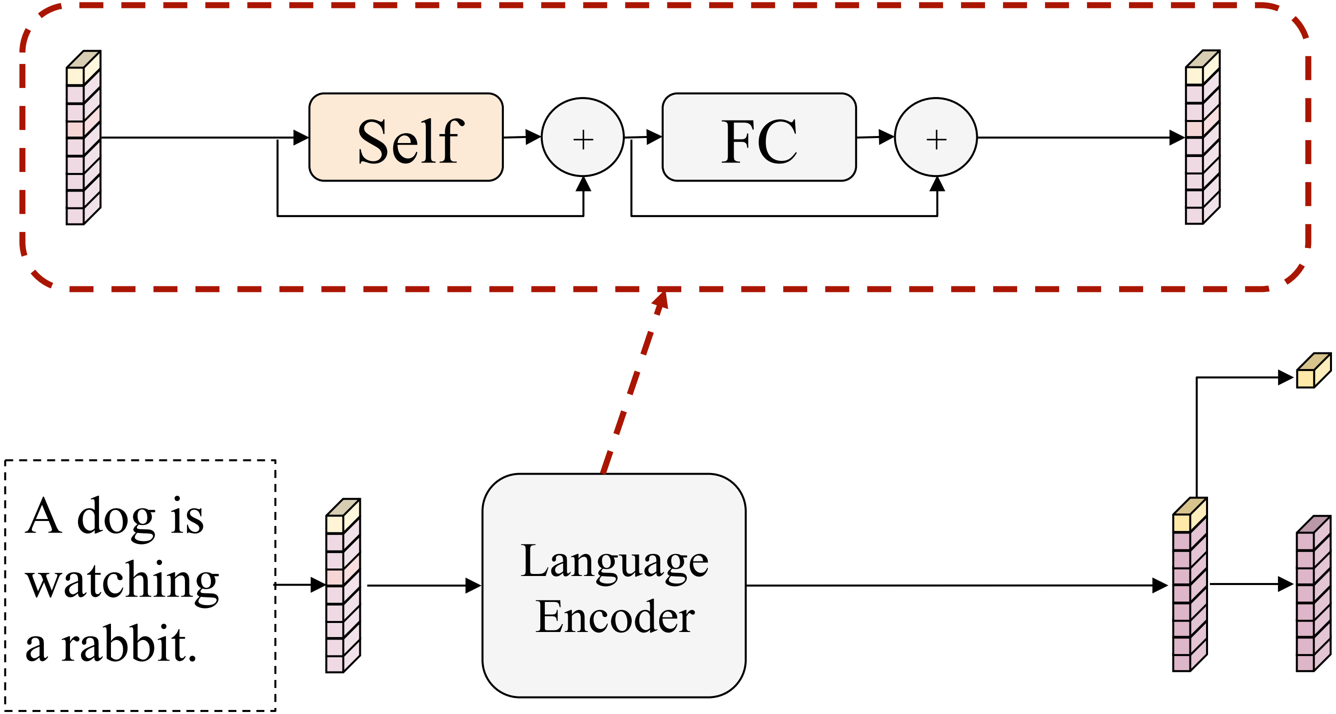
A pre-training and fine-tuning framework
for vision-and-language tasks

Model: BERT

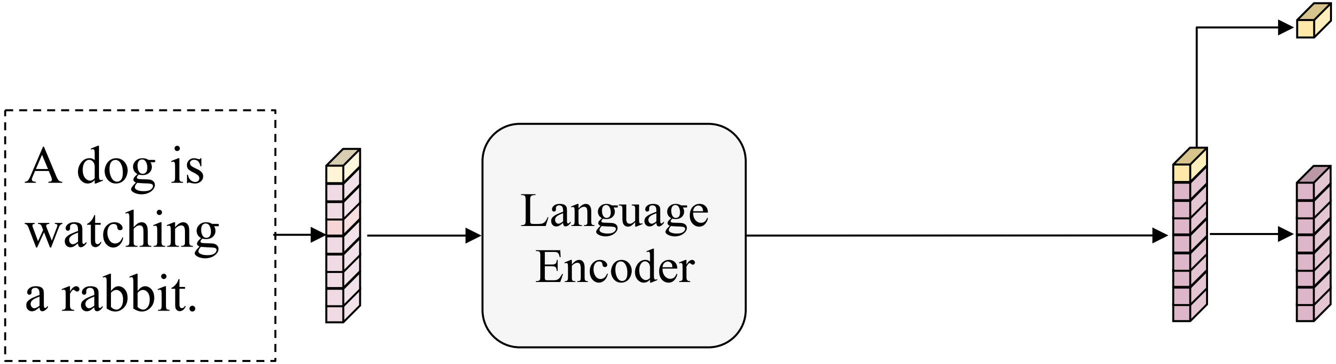


Model: BERT

BERT's language encoder is a stack of self attention layers.

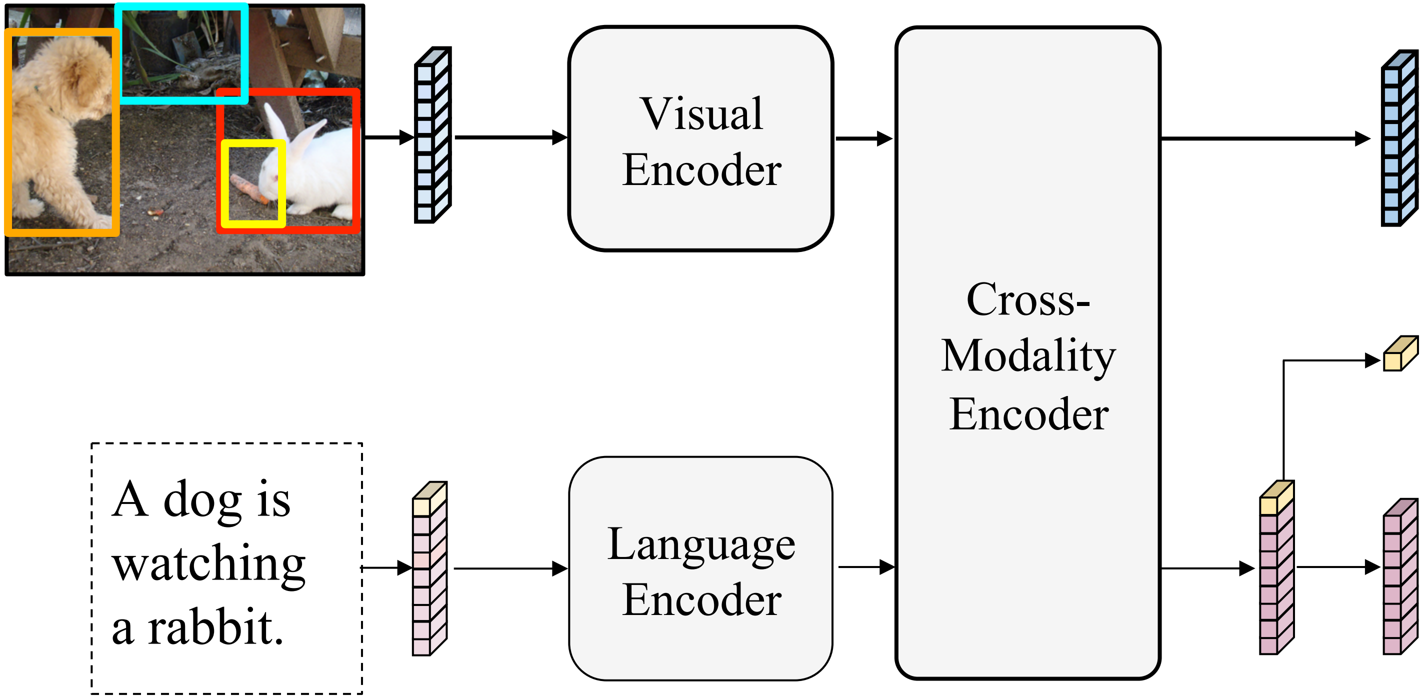


Model: BERT



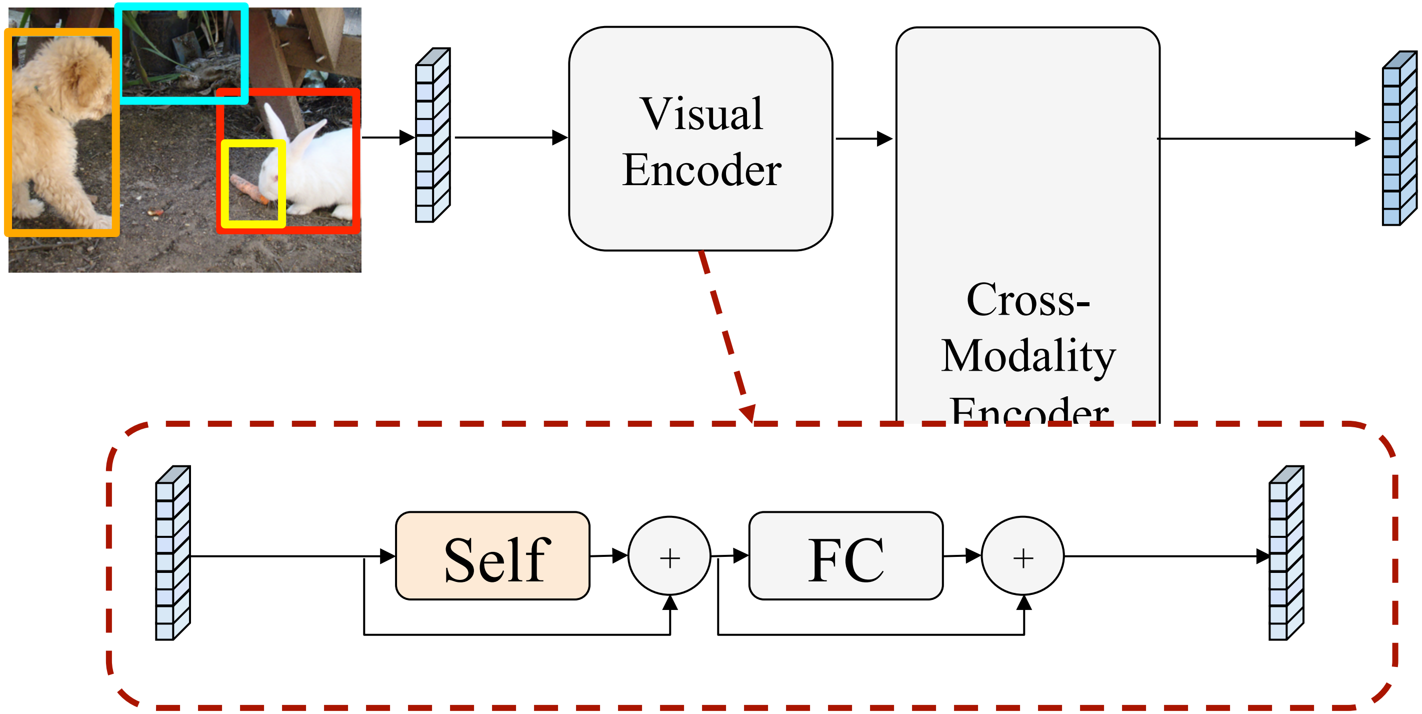
Model: LXMERT

LXMERT adds a new branch for the visual modality.



Model: LXMERT

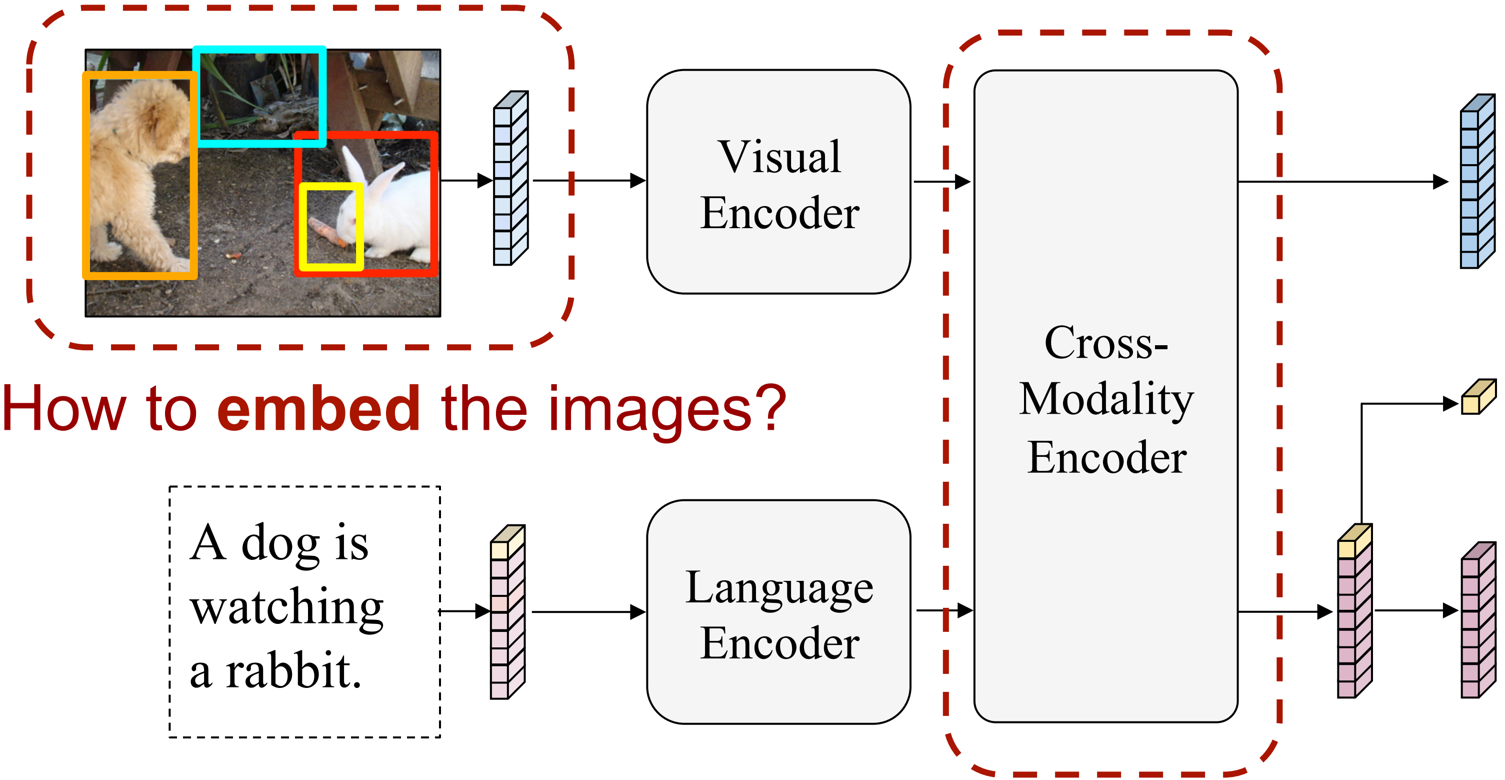
Visual encoder is similar to language encoder (with different weights).



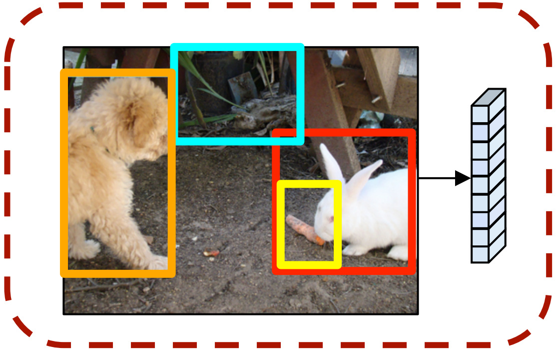
Model: LXMERT

2. How to build **connections** between modalities?

1. How to **embed** the images?

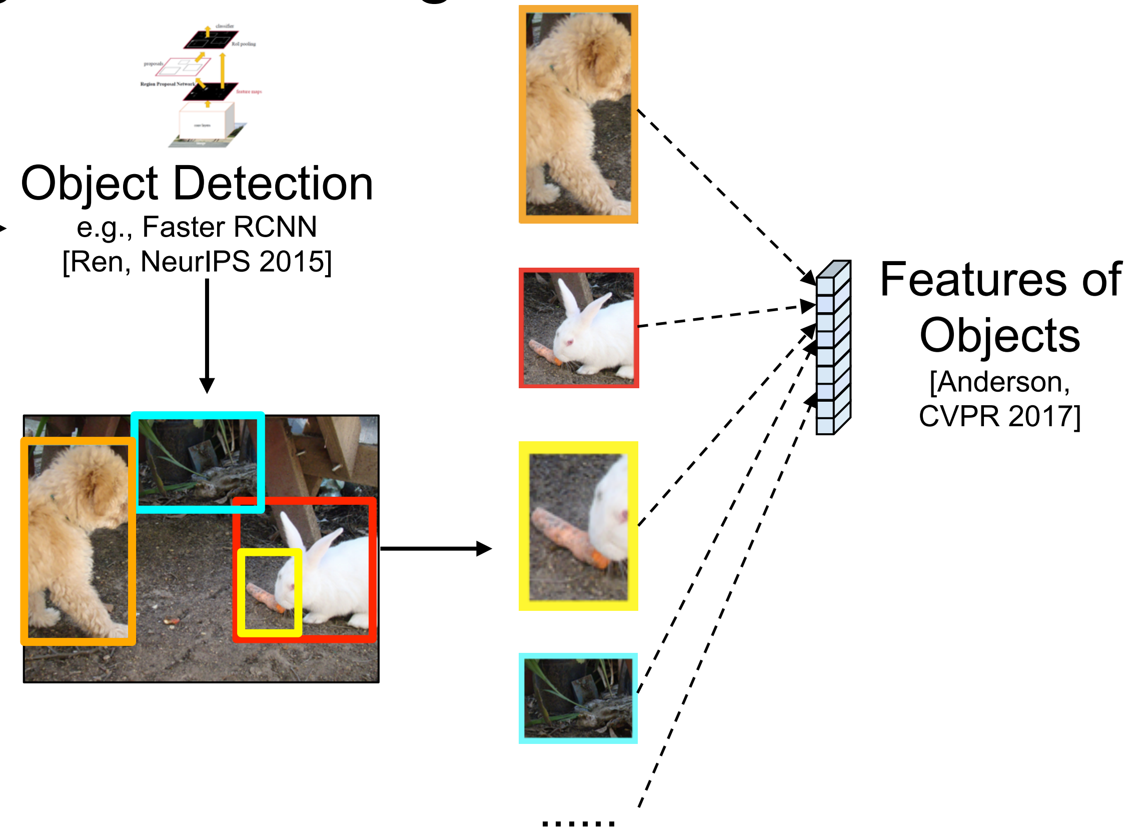


Model: LXMERT



1. How to **embed** the images?

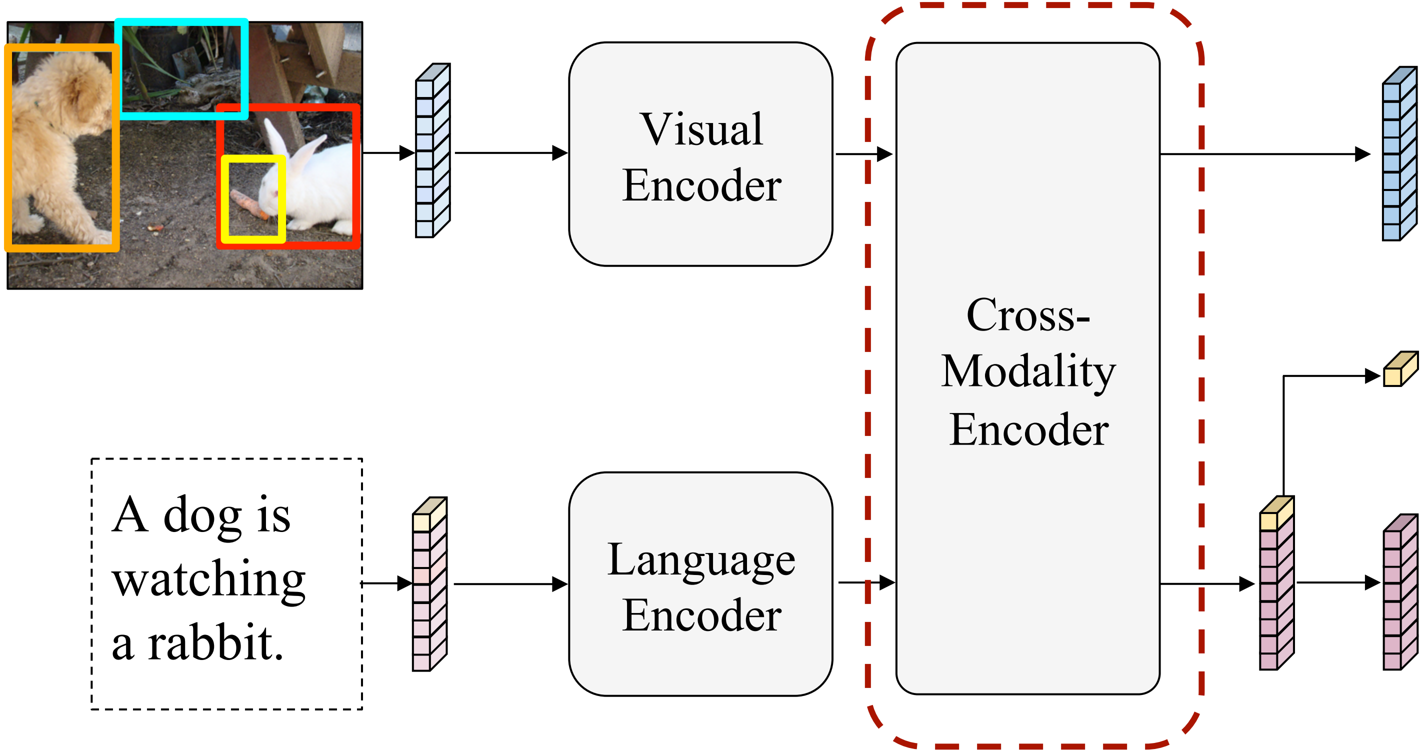
Object-Level Image Embedding



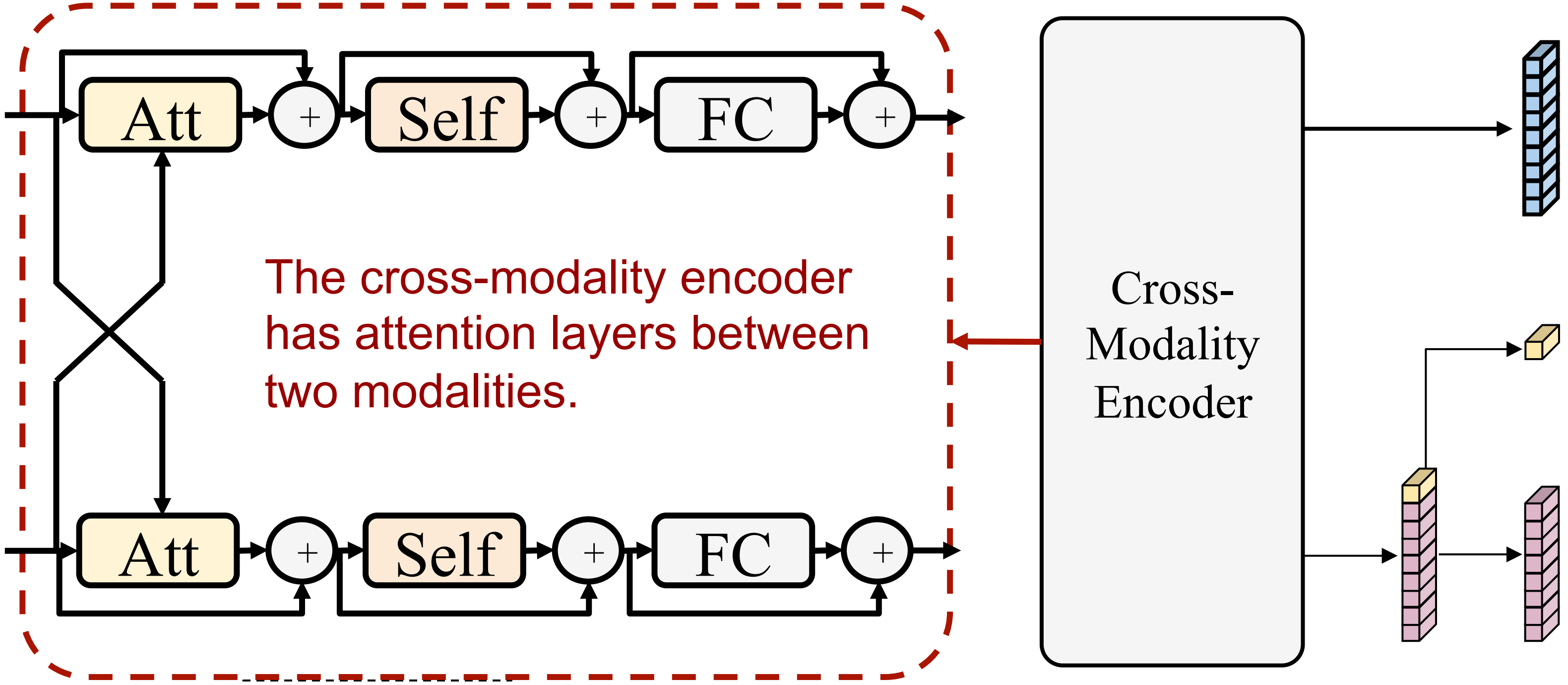
Images are embedded with object-level Image representation.

Model: LXMERT

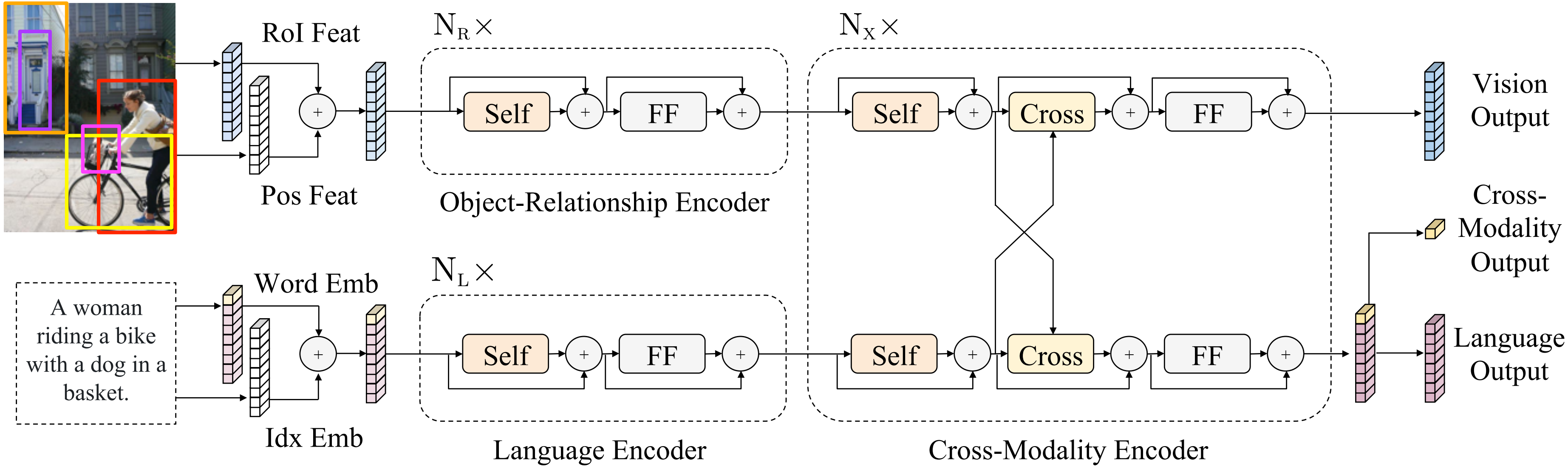
2. How to build **connections** between modalities?



Cross-Modality Attention Layers



LXMERT Full Model



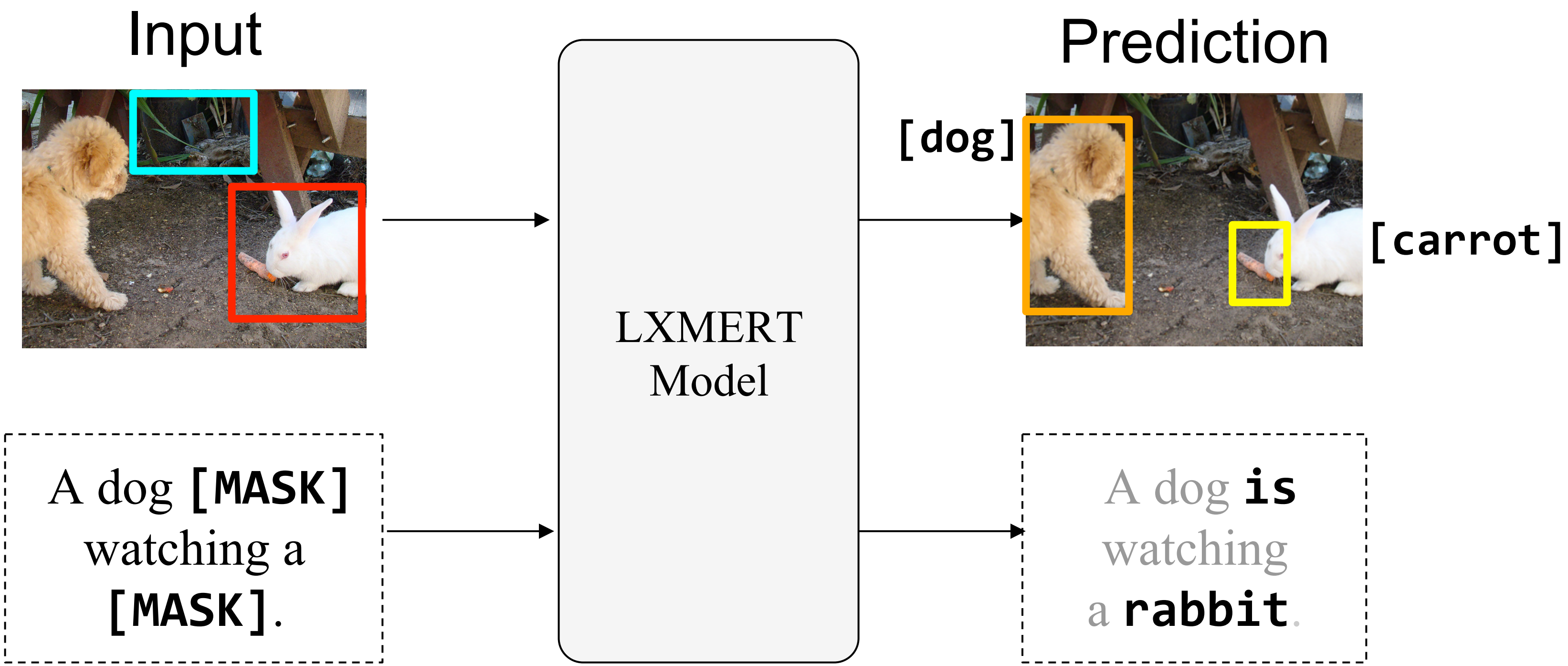
BERT Pre-training: Mask and Predict

Input

Prediction

A dog **[MASK]** watching a **[MASK]**. → BERT Model → A dog **is** watching a **rabbit**.

LXMERT Pre-training: Mask and Predict

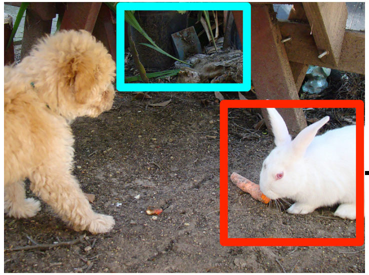


LXMERT Pre-training: Mask and Predict

Input

Single-Modality Contextualized Learning

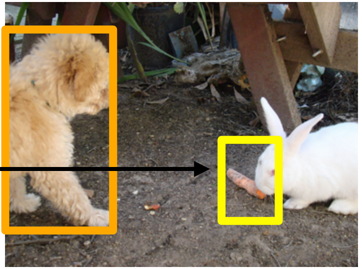
Prediction



[dog]

Rabbits like eating carrots.

Visual Relationship



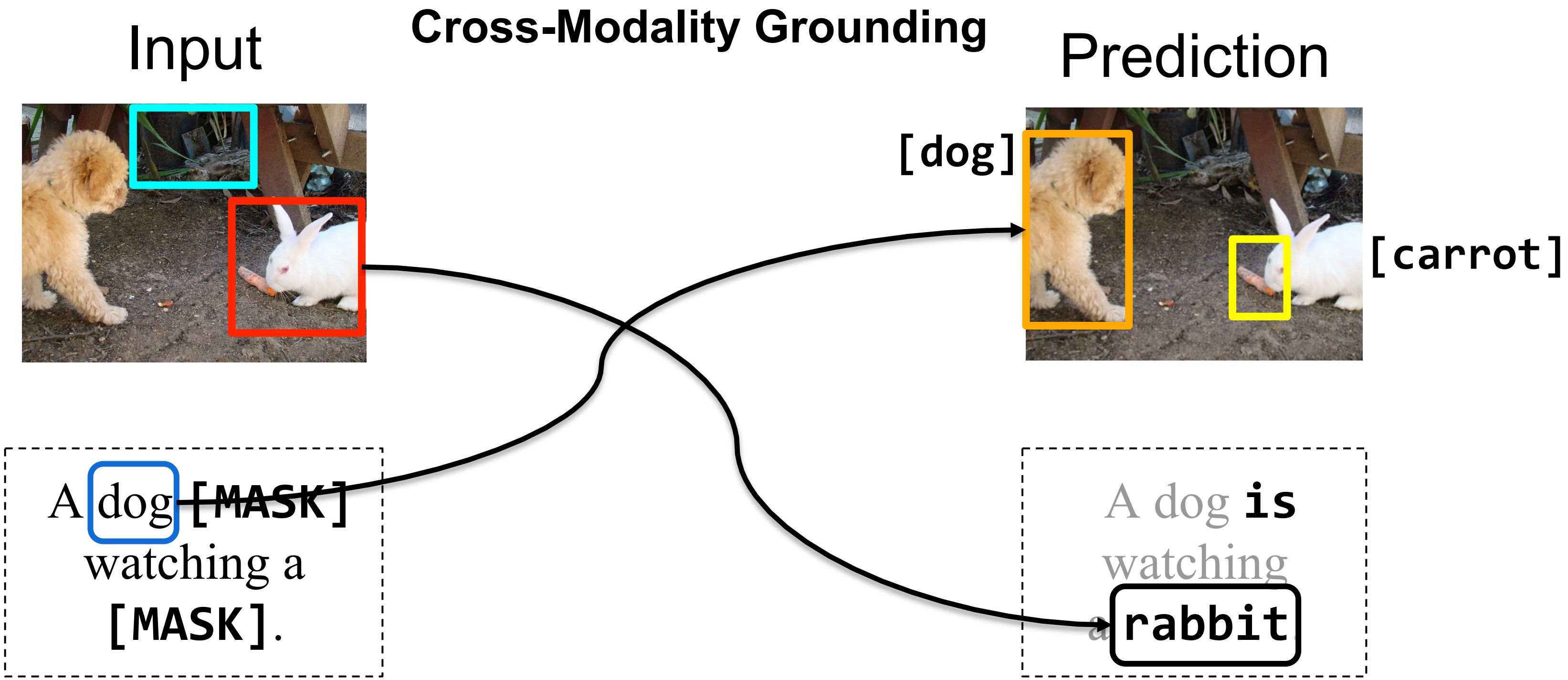
[carrot]

A dog **[MASK]** watching a **[MASK]**.

Grammar

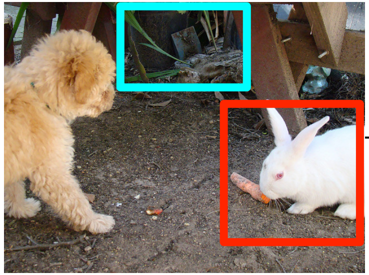
A dog **is** watching a **rabbit**.

LXMERT Pre-training: Mask and Predict



LXMERT Pre-training: Cross-Modality Matching

Input



A dog **[MASK]** watching a **[MASK]**.

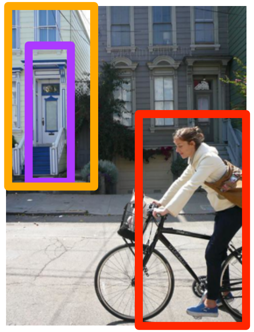
LXMERT Model

Prediction

Match or not?
Yes

LXMERT Pre-training: Cross-Modality Matching

Input



A dog **[MASK]** watching a **[MASK]**.

LXMERT Model

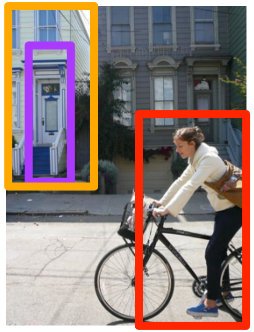
Prediction

Match or not?

No

LXMERT Pre-training: Image-Related Questions

Input



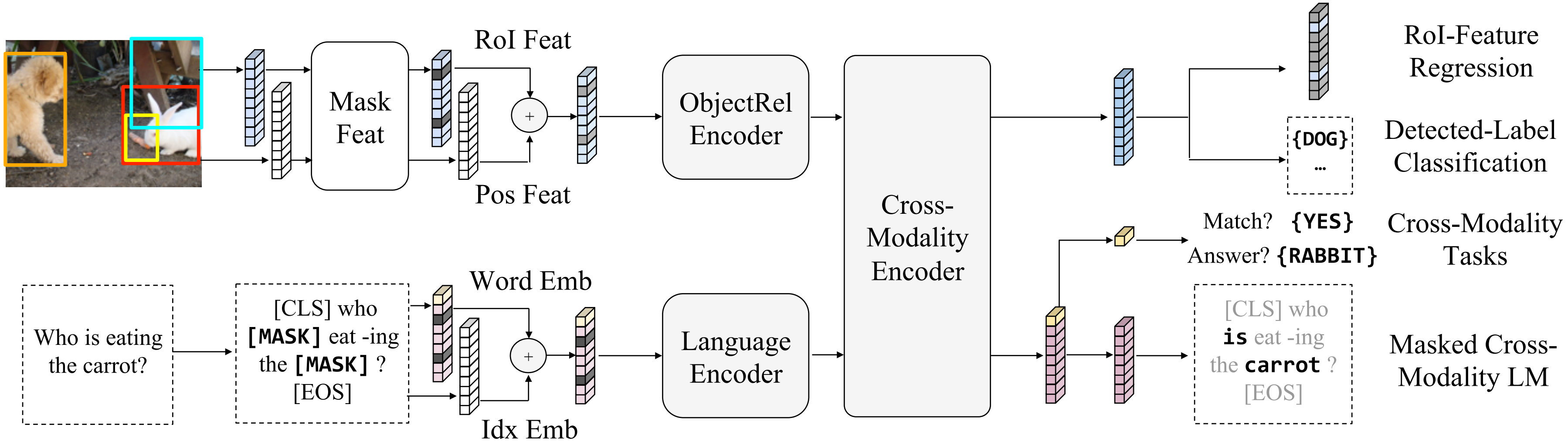
What is [mask] color of the bike?

LXMERT Model

Prediction

Black

LXMERT Pre-training Method



LXMERT Aggregated Data

Image

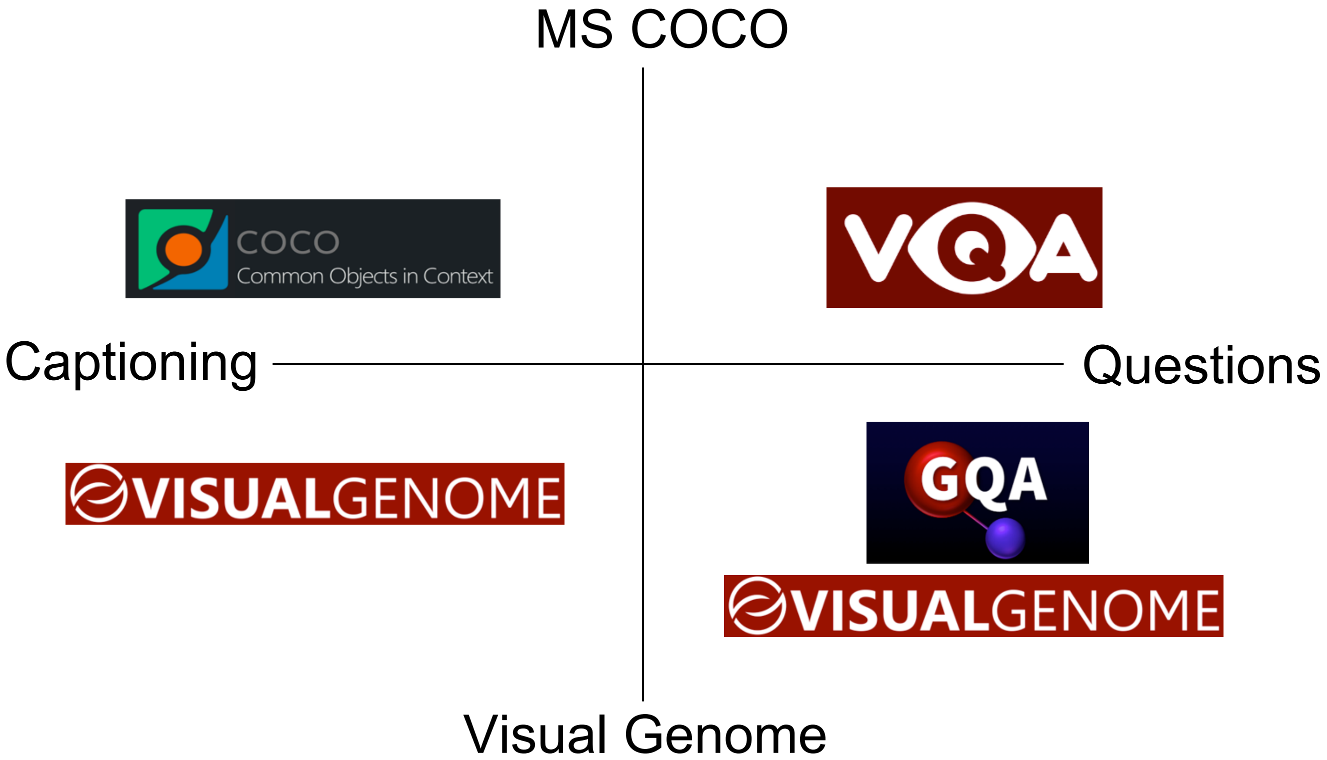
Captioning

A classic car sitting beside the road with a surfboard on top.

Related Questions

What is the horizontal bar fixed across the front of the car?

LXMERT Aggregated Data



LXMERT Aggregated Data: Amount

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

Number of Images

Number of Sentences

LXMERT Aggregated Data: Comparison

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	$\boxed{180K}$	617K	5.39M	658K	1.07M	1.44M	$\boxed{9.18M}$

Number of Images

ImageNet (ILSVRC2012): 1.2 M Images

Number of Sentences

BERT: ~3000M Sentences.

Results

Comparing LXMERT to previous works on multiple datasets.

Antol et al., ICCV 2015.

Dataset: Visual Question Answering

Answer a question about the image.

What color are her eyes? Answer: Black
What is the mustache made of? Answer: Bananas

Hudson et al., CVPR 2019.

Dataset: GQA

Focus on multi-hop reasoning.

Does the vehicle near the palms look red or blue?

Suhr et al., ACL 2019

Dataset: Natural Language for Visual Reasoning

The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

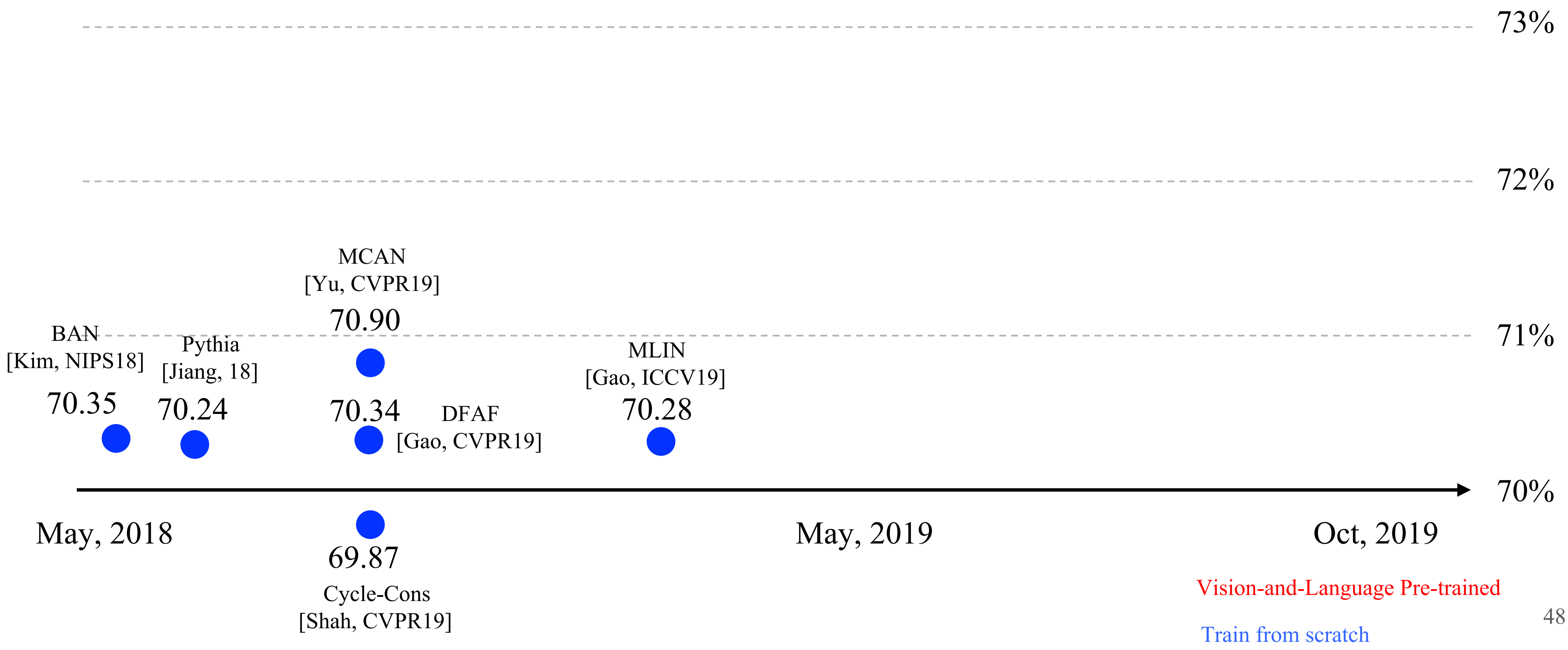
Answer: True

LXMERT Results

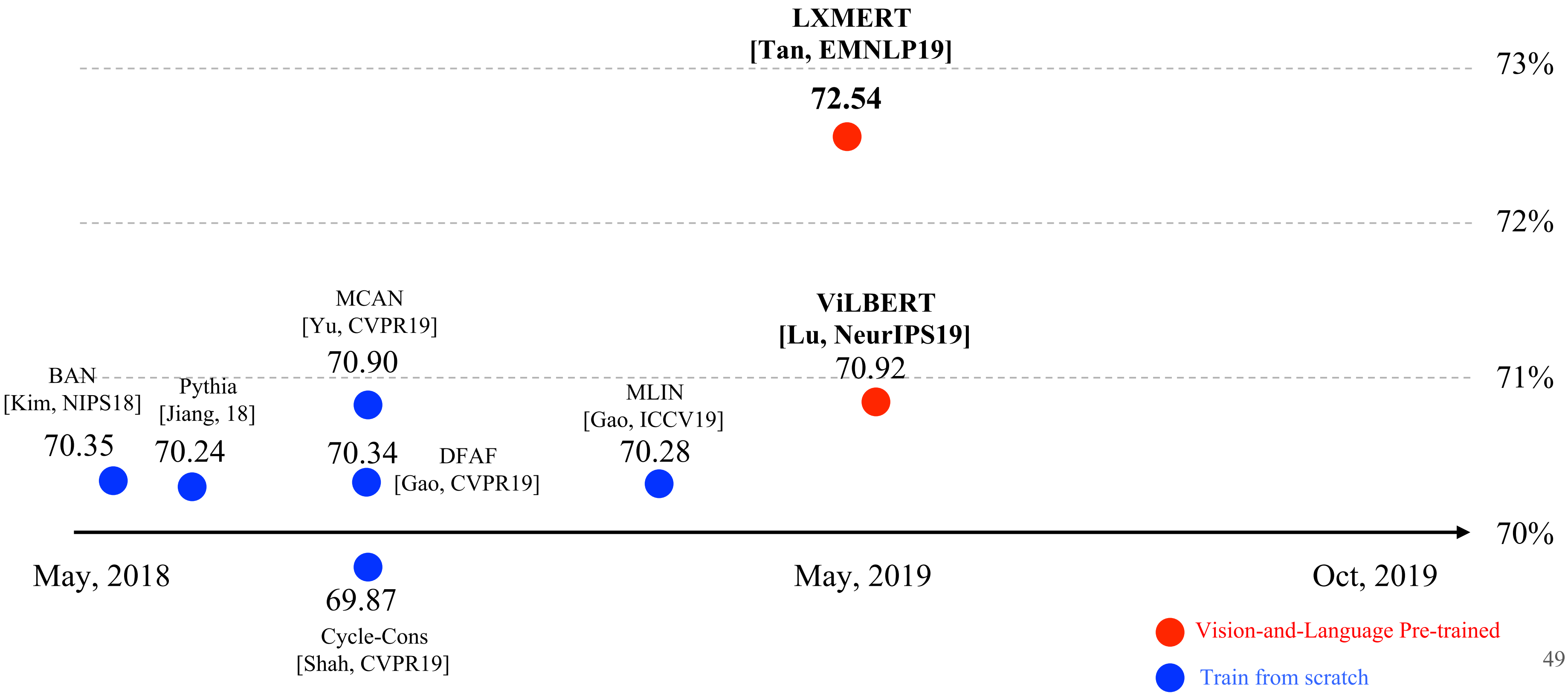
Method	VQA				GQA			$NLVR^2$	
	Binary	Number	Other	**Accu**	Binary	Open	**Accu**	Cons	**Accu**
Human	-	-	-	-	91.2	87.4	89.3	-	96.3
Image Only	-	-	-	-	36.1	1.74	17.8	7.40	51.9
Language Only	66.8	31.8	27.6	44.3	61.9	22.7	41.1	4.20	51.1
State-of-the-Art	85.8	53.7	60.7	70.4	76.0	40.4	57.1	12.0	53.5
LXMERT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

+ **2.1%** on VQA + **3.2%** on GQA + **22.7%** on NLVR2

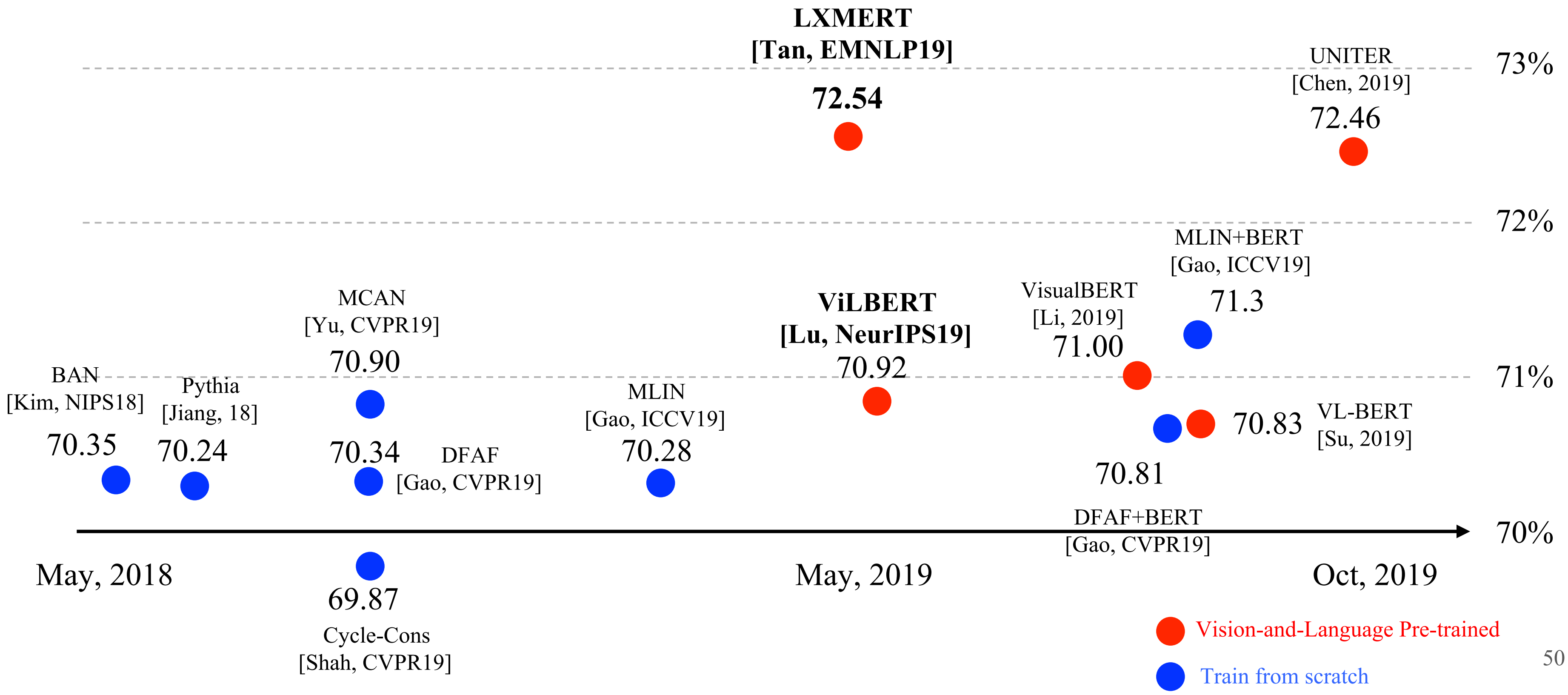
Recent Progress on Visual Question Answering



Recent Progress on Visual Question Answering



Recent Progress on Visual Question Answering



LXMERT Results

NLVR^2 Leaderboard

NLVR^2 presents the task of determining whether a natural language sentence is true about a pair of photographs.

Rank	Model	Dev. (Acc)	Test-P (Acc)	Test-U (Acc)	Test-U (Cons)
	Human Performance *Cornell University* (Suhr et al. 2019)	96.2	96.3	96.1	-
1 Aug 20, 2019	LXMERT *UNC* (Tan and Bansal 2019)	**74.9**	**74.5**	**76.2**	**42.1**
2 Aug 11, 2019	VisualBERT *UCLA & AI2 & PKU* (Li et al. 2019)	67.4	67.0	67.3	26.9
3 Nov 1, 2018	MaxEnt *Cornell University* (Suhr et al. 2019)	54.1	54.8	53.5	12.0

Top-1 on Natural Language and Visual Reasoning task.

LXMERT Results

NLVR² Leade

NLVR² presents
about a pair of ph

Rank
1 Aug 20, 2019
2 Aug 11, 2019
3 Nov 1, 2018

Cornell University
(Suhr et al. 2019)

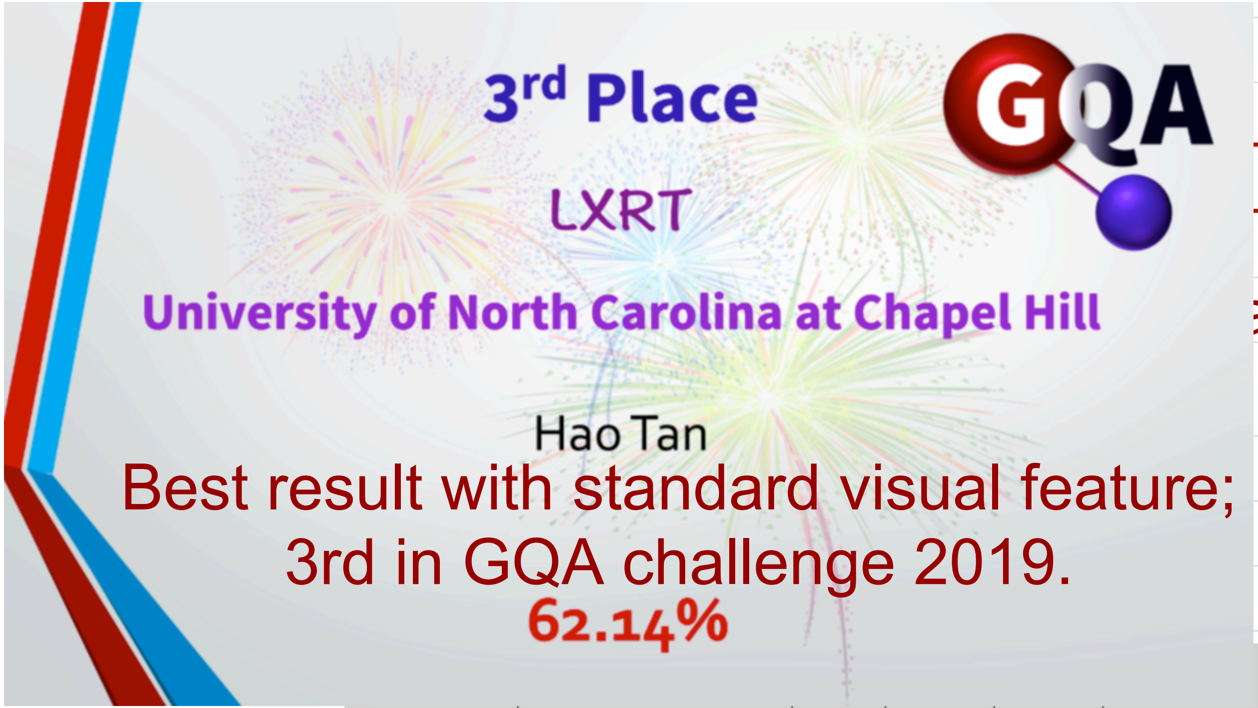
Top-1 on Reasoning task.

Rank	Participant team	yes/no	number	other	overall	Last submission at
1	MIL@HDU (MCAN)	90.36	59.17	65.75	75.23	3 months ago
2	MSM@MSRA	89.81	58.36	65.69	74.89	3 months ago
3	LXMERT (LXR955, Ensemble)	89.45	56.69	65.22	74.34	3 months ago
4	AIOZ (AIOZ-QTA)	88.26	55.22	63.63	72.93	3 months ago
5	[illegible] vqa	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
6	LXMERT github version (LXR955, Single Model)	[illegible]	[illegible]	[illegible]	[illegible]	7 days ago
7	HappyTeam (A-18)	88.24	54.15	62.11	72.08	3 months ago
8	Dream	87.95	54.17	62.05	71.93	7 days ago
9	BAN (Bilinear Attention Networks (B)	87.22	54.37	62.45	71.84	4 months ago

Best result with standard visual feature; 3rd in VQA challenge 2019.

LXMERT Results

Rank	Participant team	yes/no	number	other	overall	Last submission at
				65.75	75.23	3 months ago
				65.69	74.89	3 months ago
				65.22	74.34	3 months ago
				63.63	72.93	3 months ago
				63.80	72.92	3 months ago
				63.13	72.54	7 days ago
				62.11	72.08	3 months ago
				62.05	71.93	7 days ago
				62.45	71.84	4 months ago



Best result with standard visual feature;
3rd in GQA challenge 2019.

ard visual feature;
enge 2019.

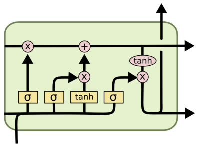
Analysis

Ablation studies and attention graphs.

https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Analysis: LXMERT Ablation Results

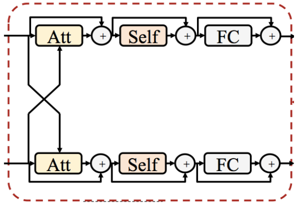
Results of BERT encoder are similar to LSTM for the baseline model.



Method	VQA	GQA	$NLVR^2$
LSTM + BUTD	63.1	50.0	52.6
BERT + BUTD	62.8	52.1	51.9
BERT + 1 CrossAtt	64.6	55.5	52.4
BERT + 2 CrossAtt	65.8	56.1	50.9
BERT + 3 CrossAtt	66.4	56.6	50.9
BERT + 4 CrossAtt	66.4	56.0	50.9
BERT + 5 CrossAtt	66.5	56.3	50.9
Train + BERT	65.5	56.2	50.9
Train + scratch	65.1	50.0	50.9
Pre-train + BERT	68.8	58.3	70.1
Pre-train + scratch	**69.9**	**60.0**	**74.9**

Table 3: Dev-set accuracy of using BERT.

Analysis: Ablation Results



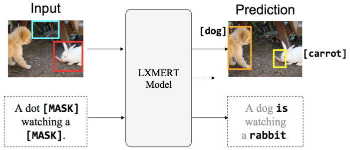
Stacking cross-modality layers helps.

Method	VQA	GQA	NLVR^2
LSTM + BUTD	63.1	50.0	52.6
BERT + BUTD	62.8	52.1	51.9
BERT + 1 CrossAtt	64.6	55.5	52.4
BERT + 2 CrossAtt	65.8	56.1	50.9
BERT + 3 CrossAtt	66.4	56.6	50.9
BERT + 4 CrossAtt	66.4	56.0	50.9
BERT + 5 CrossAtt	66.5	56.3	50.9
Train + BERT	65.5	56.2	50.9
Train + scratch	65.1	50.0	50.9
Pre-train + BERT	68.8	58.3	70.1
Pre-train + scratch	**69.9**	**60.0**	**74.9**

Table 3: Dev-set accuracy of using BERT.

- 1.0% on NLVR2
+ 4.5% on GQA
+ 3.6% on VQA

Analysis: Ablation Results

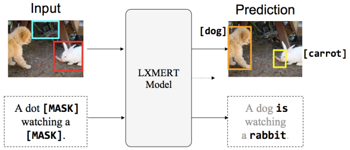


Pre-training boosts the performance.

Method	VQA	GQA	NLVR^2
LSTM + BUTD	63.1	50.0	52.6
BERT + BUTD	62.8	52.1	51.9
BERT + 1 CrossAtt	64.6	55.5	52.4
BERT + 2 CrossAtt	65.8	56.1	50.9
BERT + 3 CrossAtt	66.4	56.6	50.9
BERT + 4 CrossAtt	66.4	56.0	50.9
BERT + 5 CrossAtt	66.5	56.3	50.9
Train + BERT	65.5	56.2	50.9
Train + scratch	65.1	50.0	50.9
Pre-train + BERT	68.8	58.3	70.1
Pre-train + scratch	**69.9**	**60.0**	**74.9**

Table 3: Dev-set accuracy of using BERT.

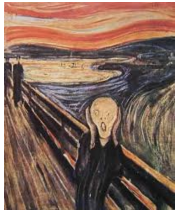
Analysis: Ablation Results



Pre-training boosts the performance.

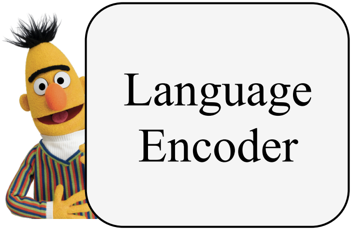
Method	VQA	GQA	NLVR^2
LSTM + BUTD	63.1	50.0	52.6
BERT + BUTD	62.8	52.1	51.9
BERT + 1 CrossAtt	64.6	55.5	52.4
BERT + 2 CrossAtt	65.8	56.1	50.9
BERT + 3 CrossAtt	66.4	56.6	50.9
BERT + 4 CrossAtt	66.4	56.0	50.9
BERT + 5 CrossAtt	66.5	56.3	50.9
Train + BERT	65.5	56.2	50.9
Train + scratch	65.1	50.0	50.9
Pre-train + BERT	68.8	58.3	70.1
Pre-train + scratch	**69.9**	**60.0**	**74.9**

Table 3: Dev-set accuracy of using BERT.



+ **24.0%** on NLVR2
+ **10.0%** on GQA
+ **4.4%** on VQA

Analysis: Ablation Results

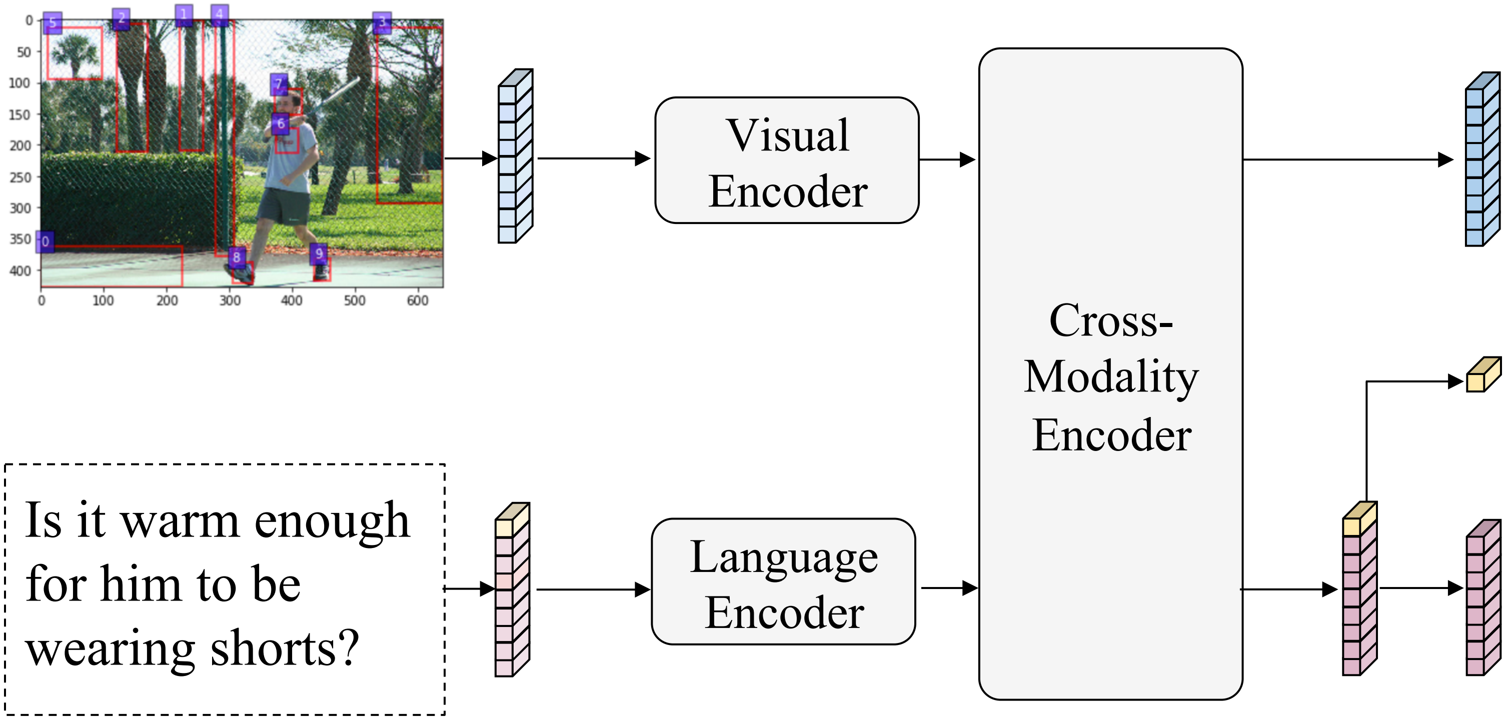


Method	VQA	GQA	NLVR^2
LSTM + BUTD	63.1	50.0	52.6
BERT + BUTD	62.8	52.1	51.9
BERT + 1 CrossAtt	64.6	55.5	52.4
BERT + 2 CrossAtt	65.8	56.1	50.9
BERT + 3 CrossAtt	66.4	56.6	50.9
BERT + 4 CrossAtt	66.4	56.0	50.9
BERT + 5 CrossAtt	66.5	56.3	50.9
Train + BERT	65.5	56.2	50.9
Train + scratch	65.1	50.0	50.9
Pre-train + BERT	68.8	58.3	70.1
Pre-train + scratch	**69.9**	**60.0**	**74.9**

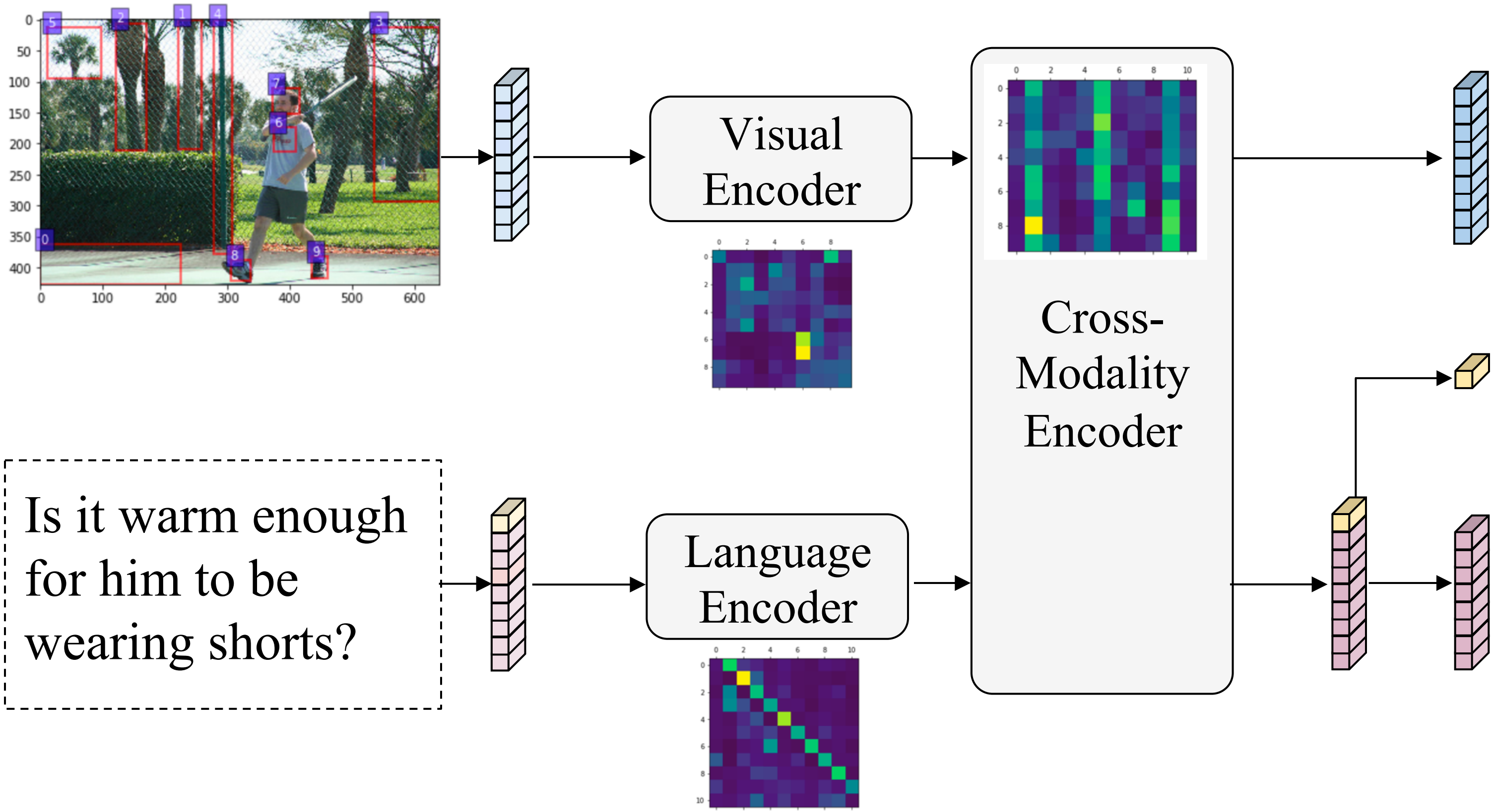
Table 3: Dev-set accuracy of using BERT.

Loading pre-trained BERT weights into LXMERT pre-training **does not help**.

Analysis: Visualizing Attention Graphs

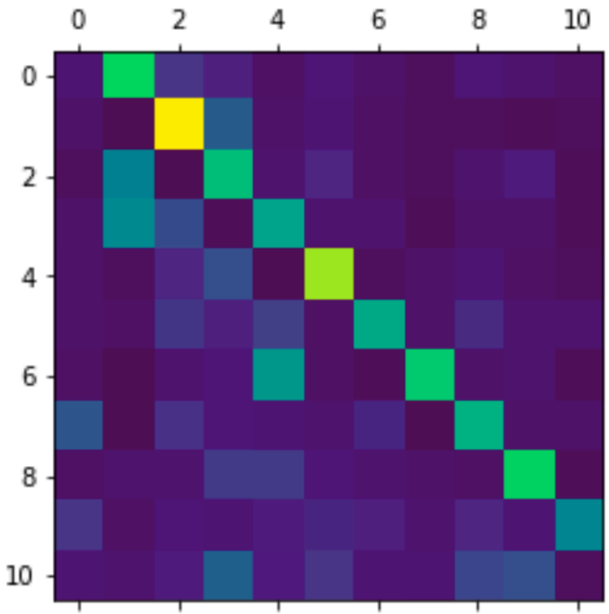


Analysis: Visualizing Attention Graphs

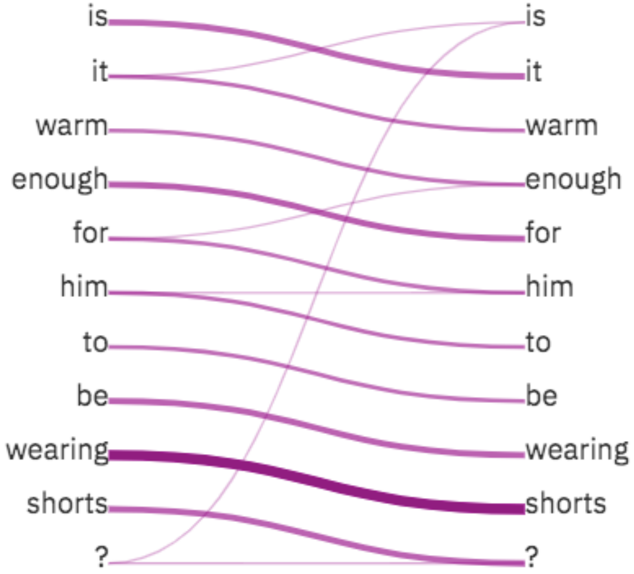


Attention Graphs: Language Encoder

Example: Is it warm enough for him to be wearing shorts?



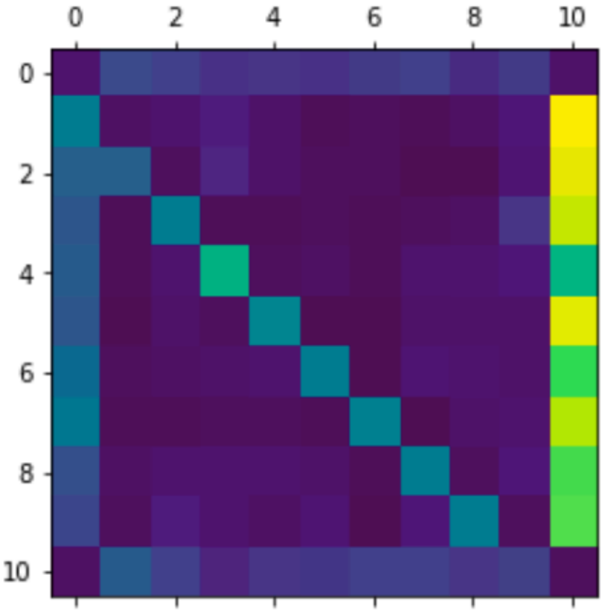
LXMERT Lang Layer 2: Attend to the **next** words.



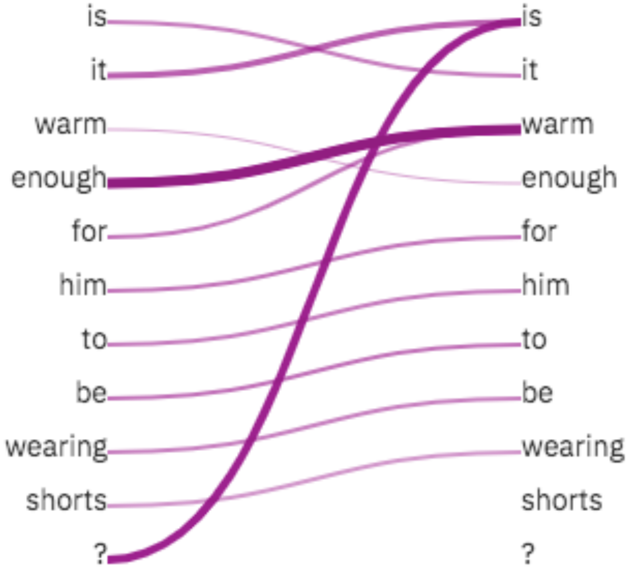
BERT Layer 3: Attend to the **next** words.

Attention Graphs: Language Encoder

Example: Is it warm enough for him to be wearing shorts?



LXMERT Lang Layer 4: Attend to the **previous** words.

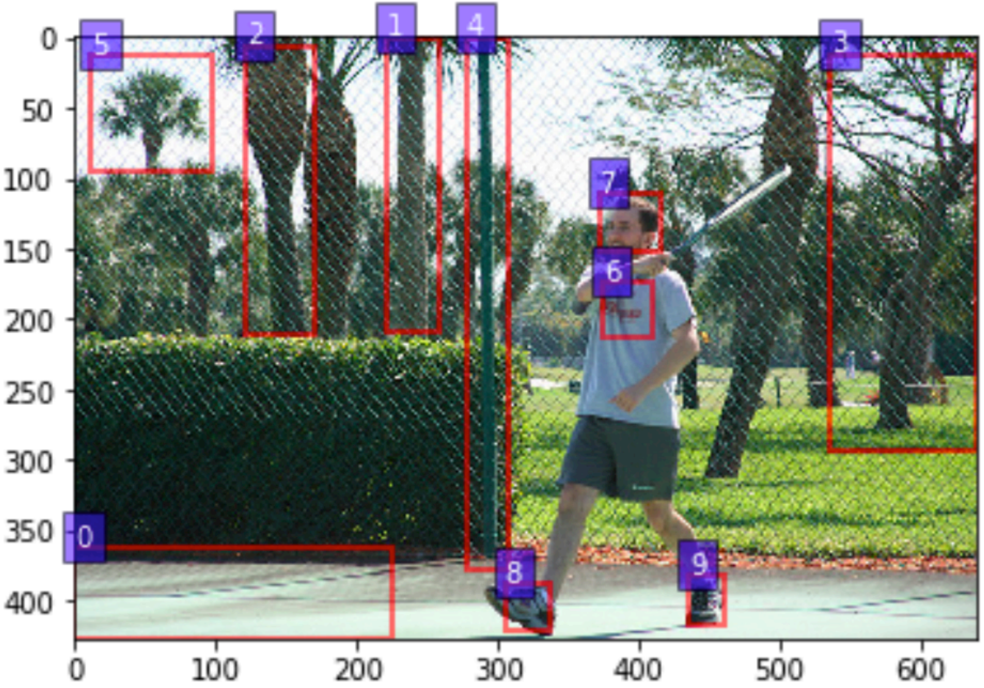


BERT Layer 4: Attend to the **previous** words.

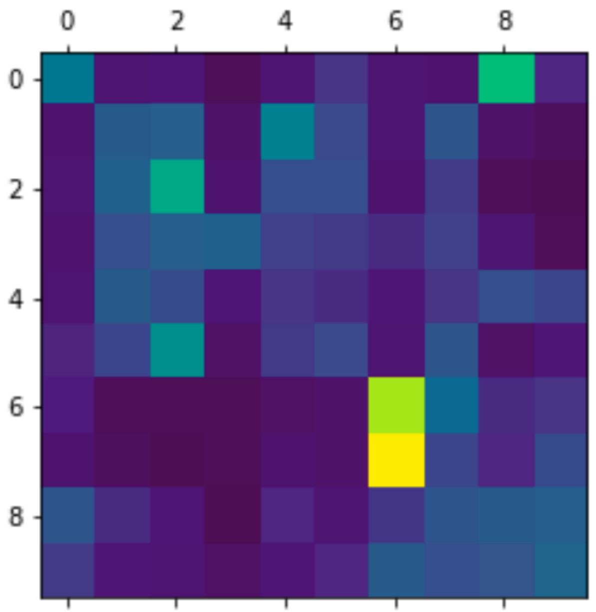
Attention Graphs: Visual Encoder

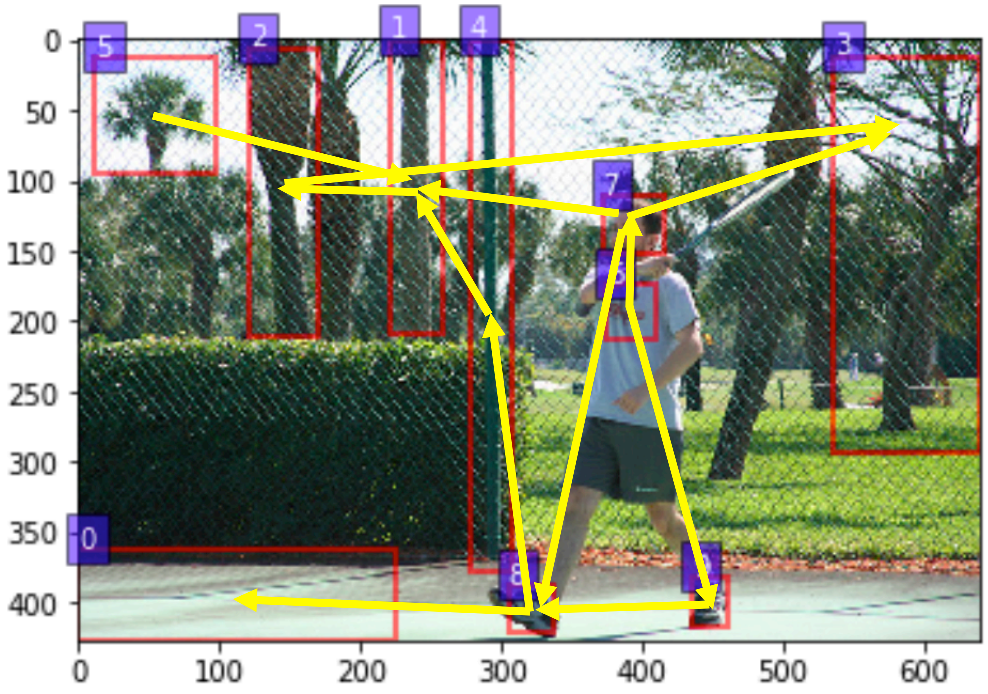
The most attended visual objects are:

A. Separated.
B. Lied at the center of semantic regions.

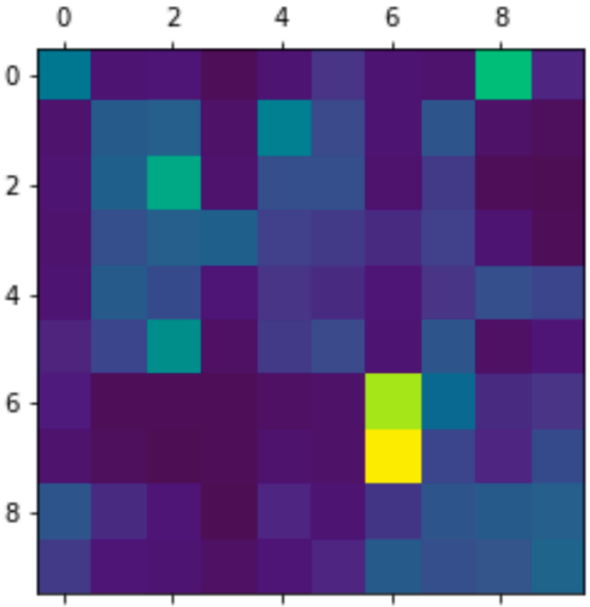


Attention Graphs: Visual Encoder

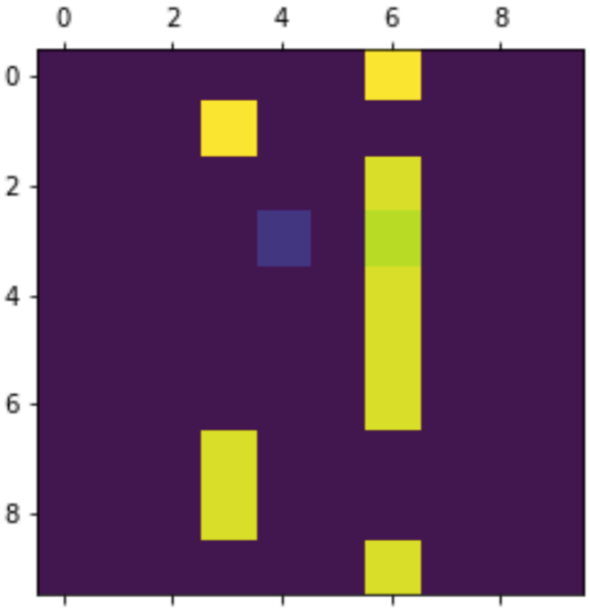




Attention Graphs: Visual Encoder



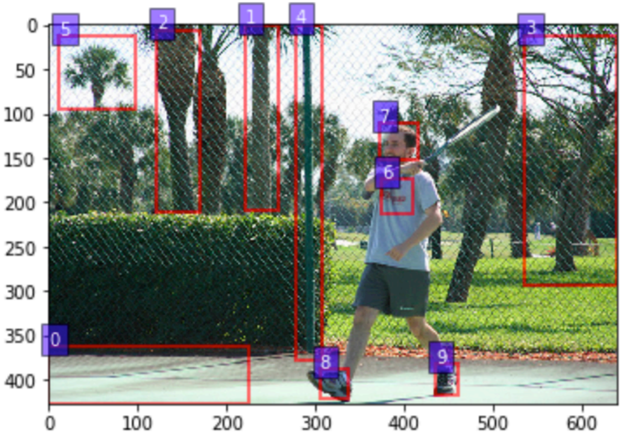
LXMERT has less of this issue.



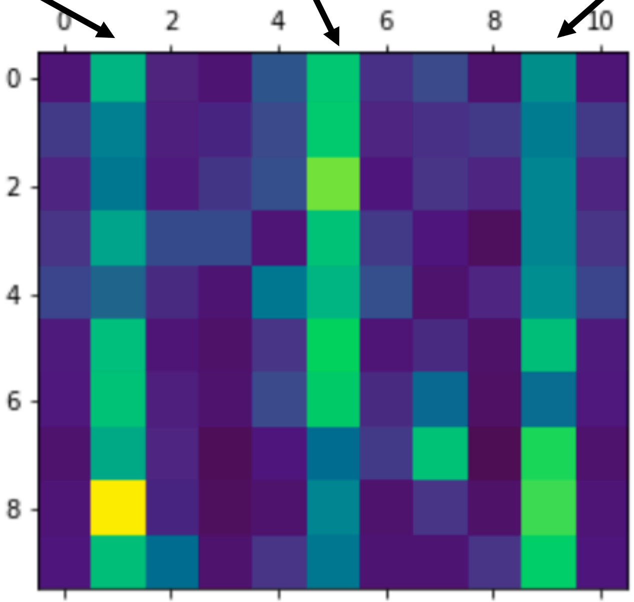
Directly applying self-attention on object sequences would lead to a one-hot attention. [Jinwon An]

Attention Graphs: Cross-Modality Encoder

Attention are focusing on **Nouns** and **Pronouns**.



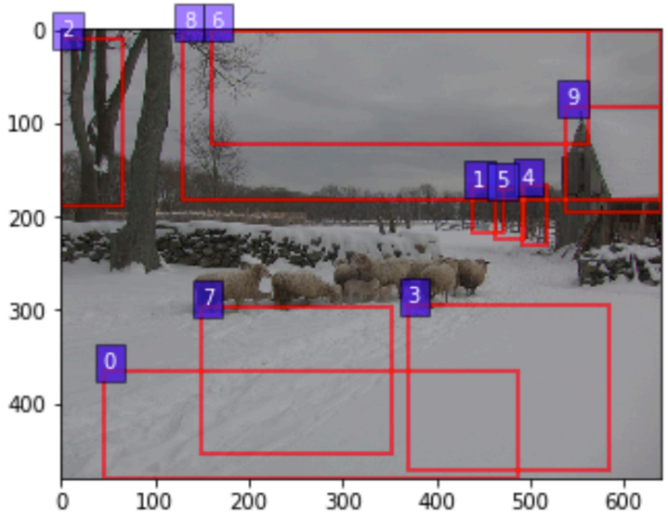
Is **it** warm enough for **him** to be wearing **shorts** ?

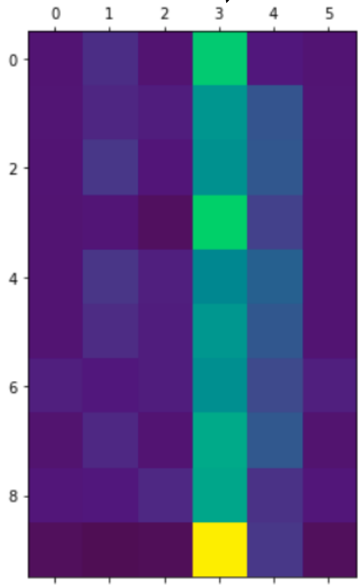


Attention Graphs: Cross-Modality Encoder

For **Non-plural Nouns**, the attention will focus on the **Articles**!

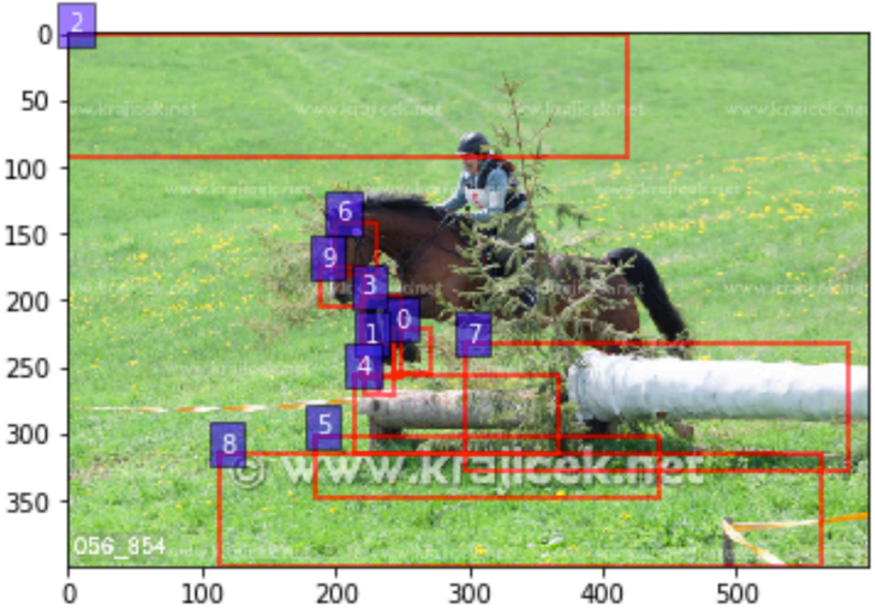
Can you see **the grass** ?



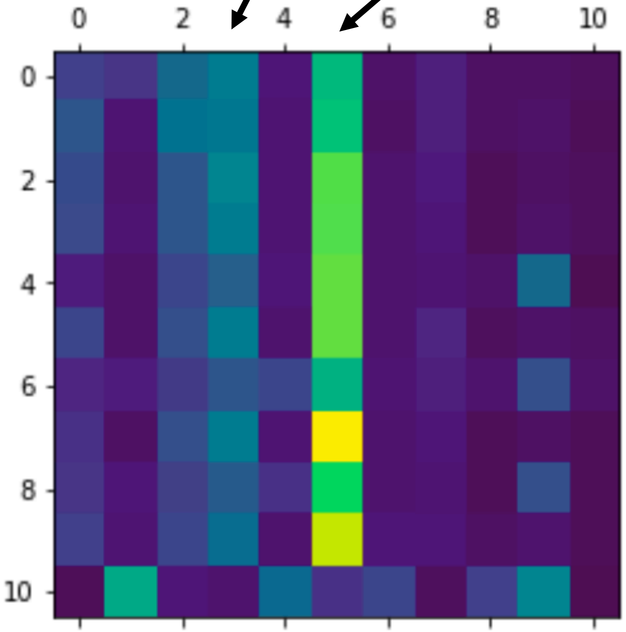


Attention Graphs: Cross-Modality Encoder

For **Non-plural Nouns**, the attention will focus on the **Articles**!



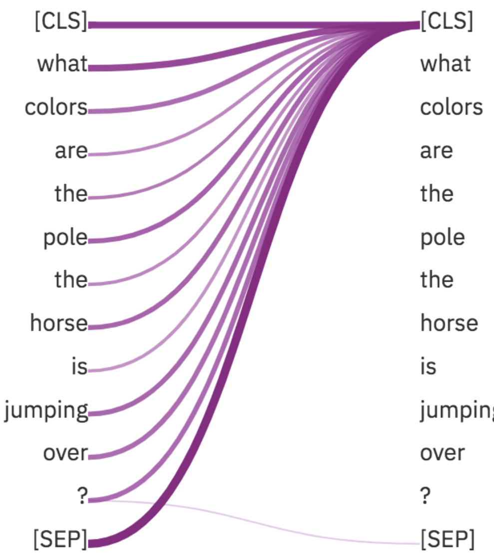
What colors are **the pole the horse** is jumping over?

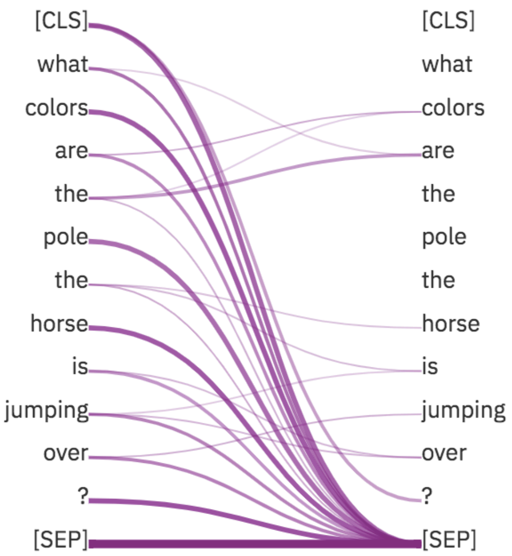


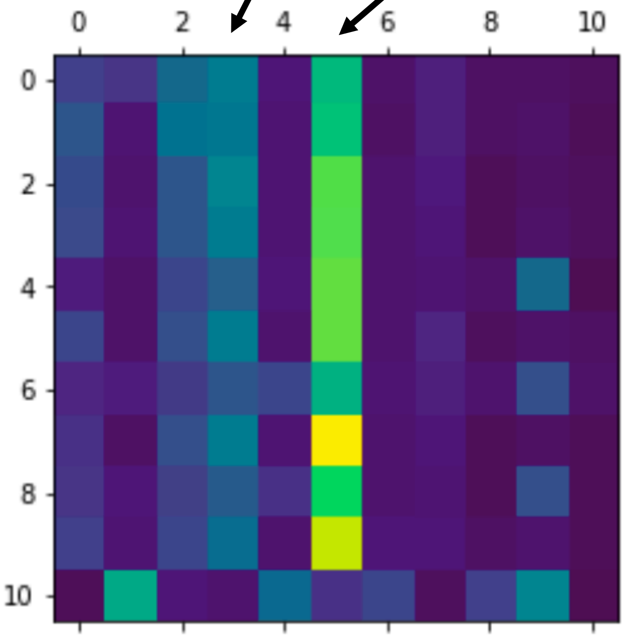
Attention Graphs: Cross-Modality Encoder

Articles are possibly serving as special tokens (e.g., [CLS], [SEP]).

What colors are **the pole the horse** is jumping over?







What’s Next?

The future of vision-and-language pre-training.

Data

Short Sentence

Long Paragraph

Caption, Question, Instruction, …….

$\longrightarrow$

News, Books, Tutorial, …….

Data

Balanced Data

An orange cat sits in the suitcase ready to be packed.

Unbalanced Data

Sociability

Social grooming

The social behavior of the domestic cat ranges from widely dispersed individuals to feral cat colonies that gather around a food source, based on groups of co-operating females.[94][95] Within such groups, one cat is usually dominant over the others.[96] Each cat in a colony holds a distinct territory, with sexually active males having the largest territories, which are about 10 times larger than those of female cats and may overlap with several females' territories. These territories are marked by urine spraying, by rubbing objects at head height with secretions from facial glands, and by defecation.[76] Between these territories are neutral areas where cats watch and greet one another without territorial conflicts. Outside these neutral areas, territory holders usually chase away stranger cats, at first by staring, hissing, and growling and, if that does not work, by short but noisy and violent attacks. Despite some cats cohabiting in colonies, they do not have a social survival strategy, or a pack mentality and always hunt alone.[97]

In wiki/news/tutorial, they usually have long text and only one image.

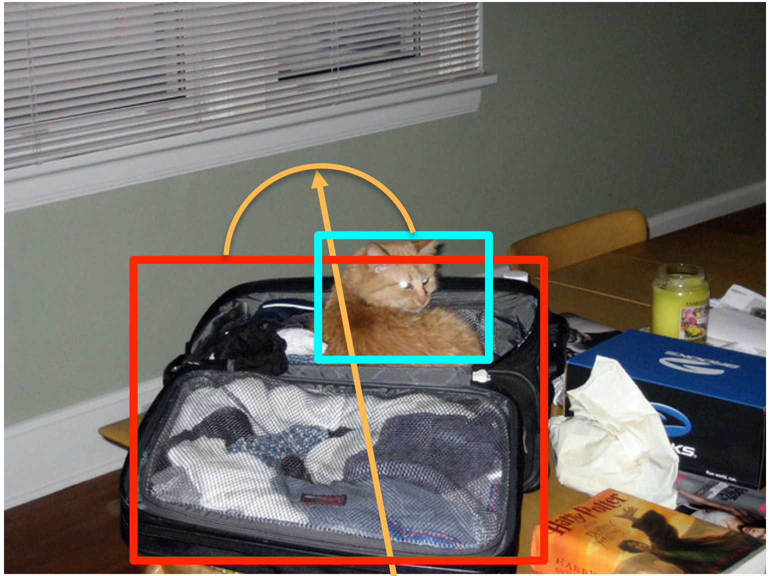
Data

Limited Aligned Data

(Nearly) Unlimited Unaligned Data

Tasks

Pre-training tasks which capture pairwise noun-noun and noun-verb relationships.



An orange cat sits in the suitcase ready to be packed.

LXR Thanks!!

Code available at: github.com/airsplay/lxmert

Hao Tan, Mohit Bansal
haotan, mbansal@cs.unc.edu