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.
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

Antol et al., ICCV 2015.
Pre-training $\rightarrow$ 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 $\rightarrow$ Fine-tuning

Visual Pre-training:
- ImageNet
  - [Deng, CVPR 2009]

Visual Fine-tuning:
- MS COCO
  - [Lin, ECCV 2009]

Image Classification
- DenseNet
  - [Huang, CVPR 2017]

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]
Fine-tuning

Visual Pre-training: Image Classification

Language Pre-training: Language Model

Detection, Segmentation, Identification, ...

Question Answering, Sentiment Analysis, ...
Fine-tuning on Vision and Language Tasks?

Visual Pre-training:

Image Classification

Language Pre-training:

Visual Question Answering, Navigation, Grounding, ...

Fusion Module

Language Model
Fine-tuning on Vision and Language Tasks?

Visual Pre-training: Image Classification

Language Pre-training: Pre-trained language modules do not help.

Visual Question Answering, Navigation, Grounding, ...

Fusion Module
Fine-tuning on Vision and Language Tasks?

Visual Pre-training: Image Classification

Language Pre-training: Language Model

Fusion Module

Visual Question Answering, Navigation, Grounding, ...

Single-modality pre-training is not aware of cross-modality relationships.
Pre-train for Vision and Language jointly?

Visual Pre-training: Image Classification → Model?

Language Pre-training: Language Model → Model?

Pre-training Method?

Data?
LXMERT (Learning Cross-Modality Encoder Representations from Transformers)
LXMERT

A pre-training and fine-tuning framework for vision-and-language tasks
A dog is watching a rabbit.
Model: BERT

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

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

Model: BERT
Model: LXMERT

LXMERT adds a new branch for the visual modality.

A dog is watching a rabbit.
Model: LXMERT

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

1. How to embed the images?

A dog is watching a rabbit.

2. How to build connections between modalities?
Model: LXMERT

1. How to embed the images?
Object-Level Image Embedding

Images are embedded with object-level Image representation.

Object Detection
e.g., Faster RCNN [Ren, NeurIPS 2015]

Features of Objects
[Anderson, CVPR 2017]
Model: LXMERT

2. How to build connections between modalities?

A dog is watching a rabbit.
The cross-modality encoder has attention layers between two modalities.
LXMERT Full Model

A woman riding a bike with a dog in a basket.

Diagram showing the components of the LXMERT Full Model:
- **RoI Feat**
- **Pos Feat**
- **Word Emb**
- **Idx Emb**

The model consists of:
- **Object-Relationship Encoder**
  - Self + FF
  - \( N_R \times \)

- **Language Encoder**
  - Self + FF
  - \( N_L \times \)

- **Cross-Modality Encoder**
  - Self + Cross + FF
  - \( N_X \times \)

Outputs:
- **Vision Output**
- **Cross-Modality Output**
- **Language Output**
BERT Pre-training: Mask and Predict

Input

A dog \([\text{MASK}]\)
watching a \([\text{MASK}]\).

Prediction

A dog is
watching
a \text{rabbit}. 

BERT
Model
LXMERT Pre-training: Mask and Predict

A dog [MASK] watching a [MASK].

A dog is watching a rabbit.
LXMERT Pre-training: Mask and Predict

Input

Single-Modality Contextualized Learning

Prediction

Rabbits like eating carrots.

Visual Relationship

A dog [MASK] watching a [MASK].

A dog is watching a rabbit.
LXMERT Pre-training: Mask and Predict

A dog [MASK] watching a [MASK].

A dog is watching a rabbit.
LXMERT Pre-training: Cross-Modality Matching

A dog [MASK] watching a [MASK].
LXMERT Pre-training: Cross-Modality Matching

A dog [MASK] watching a [MASK].

Match or not? No
LXMERT Pre-training: Image-Related Questions

What is [mask] color of the bike?

Input

LXMERT Model

Prediction

Black
LXMERT Pre-training Method

Who is eating the carrot?


Word Emb

[CLS] who is eat-ing the carrot ? [EOS]

Mask Feat

RoI Feat

Pos Feat

ObjectRel Encoder

Language Encoder

Cross-Modality Encoder

Detected-Label Classification

Cross-Modality Tasks

Masked Cross-Modality LM

RoI-Feature Regression

Match? {YES} {RABBIT}

Answer? {DOG} …
LXMERT Aggregated Data

Image

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

Captioning

Related Questions

What is the horizontal bar fixed across the front of the car?
LXMERT Aggregated Data

MS COCO

Captioning

Questions

Visual Genome

Captioning

Questions

Visual Genome
### LXMERT Aggregated Data: Amount

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

<table>
<thead>
<tr>
<th>Number of Images</th>
<th>Number of Sentences</th>
</tr>
</thead>
</table>
**LXMERT Aggregated Data: Comparison**

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

**Number of Images**

ImageNet (ILSVRC2012): 1.2 M Images

**Number of Sentences**

BERT: ~3000M Sentences.
Results

Comparing LXMERT to previous works on multiple datasets.
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

Antol et al., ICCV 2015.
Dataset: GQA

Focus on multi-hop reasoning.

Does the vehicle near the palms look red or blue?
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

<table>
<thead>
<tr>
<th>Method</th>
<th>VQA</th>
<th>GQA</th>
<th>NLVR²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Binary</td>
<td>Number</td>
<td>Other</td>
</tr>
<tr>
<td>Human</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Image Only</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Language Only</td>
<td>66.8</td>
<td>31.8</td>
<td>27.6</td>
</tr>
<tr>
<td>State-of-the-Art</td>
<td>85.8</td>
<td>53.7</td>
<td>60.7</td>
</tr>
<tr>
<td>LXMERT</td>
<td>88.2</td>
<td>54.2</td>
<td>63.1</td>
</tr>
</tbody>
</table>

+ 2.1% on VQA         + 3.2% on GQA         + 22.7% on NLVR²
Recent Progress on Visual Question Answering

- BAN [Kim, NIPS18] 70.35
- Pythia [Jiang, 18] 70.24
- MCAN [Yu, CVPR19] 70.90
- Cycle-Cons [Shah, CVPR19] 69.87
- DFAF [Gao, CVPR19] 70.34
- MLIN [Gao, ICCV19] 70.28

Vision-and-Language Pre-trained
Train from scratch
Recent Progress on Visual Question Answering

**LXMERT**
[Tan, EMNLP19]
72.54
73%

**ViLBERT**
[Lu, NeurIPS19]
71%

- **BAN**
[Kim, NIPS18]
70.35
- **Pythia**
[Jiang, 18]
70.24
- **MCAN**
[Yu, CVPR19]
70.90
- **DFAF**
[Gao, CVPR19]
70.34
- **MLIN**
[Gao, ICCV19]
70.28
- **Cycle-Cons**
[Shah, CVPR19]
69.87

**Train from scratch**

**Vision-and-Language Pre-trained**

May, 2018
May, 2019
Oct, 2019

70%
71%
72%
73%
Recent Progress on Visual Question Answering

<table>
<thead>
<tr>
<th>Model</th>
<th>Year</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>LXMERT</td>
<td>2019</td>
<td>72.54</td>
</tr>
<tr>
<td>UNITER</td>
<td>2019</td>
<td>72.46</td>
</tr>
<tr>
<td>ViLBERT</td>
<td>2019</td>
<td>71.00</td>
</tr>
<tr>
<td>MLIN+BERT</td>
<td>2019</td>
<td>71.3</td>
</tr>
<tr>
<td>VL-BERT</td>
<td>2019</td>
<td>70.83</td>
</tr>
<tr>
<td>DFAF+BERT</td>
<td>2019</td>
<td>70.81</td>
</tr>
<tr>
<td>DFAF</td>
<td>2019</td>
<td>70.28</td>
</tr>
<tr>
<td>LXMERT</td>
<td>2018</td>
<td>70.92</td>
</tr>
<tr>
<td>ViLBERT</td>
<td>2019</td>
<td>70.90</td>
</tr>
<tr>
<td>Pythia</td>
<td>2018</td>
<td>70.90</td>
</tr>
<tr>
<td>BAN</td>
<td>2018</td>
<td>70.35</td>
</tr>
<tr>
<td>MLIN</td>
<td>2019</td>
<td>70.28</td>
</tr>
<tr>
<td>Cycle-Cons</td>
<td>2018</td>
<td>69.87</td>
</tr>
<tr>
<td>Vision-and-Language Pre-trained</td>
<td></td>
<td>73%</td>
</tr>
<tr>
<td>Train from scratch</td>
<td></td>
<td>72%</td>
</tr>
<tr>
<td>Train from scratch</td>
<td></td>
<td>71%</td>
</tr>
<tr>
<td>Train from scratch</td>
<td></td>
<td>70%</td>
</tr>
</tbody>
</table>
LXMERT Results

Top-1 on Natural Language and Visual Reasoning task.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>Dev. (Acc)</th>
<th>Test-P (Acc)</th>
<th>Test-U (Acc)</th>
<th>Test-U (Cons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Human Performance Cornell University (Suhr et al. 2019)</td>
<td>96.2</td>
<td>96.3</td>
<td>96.1</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>LXMERT UNC (Tan and Bansal 2019)</td>
<td>74.9</td>
<td>74.5</td>
<td>76.2</td>
<td>42.1</td>
</tr>
<tr>
<td>2</td>
<td>VisualBERT UCLA &amp; AI2 &amp; PKU (Li et al. 2019)</td>
<td>67.4</td>
<td>67.0</td>
<td>67.3</td>
<td>26.9</td>
</tr>
<tr>
<td>3</td>
<td>MaxEnt Cornell University (Suhr et al. 2019)</td>
<td>54.1</td>
<td>54.8</td>
<td>53.5</td>
<td>12.0</td>
</tr>
</tbody>
</table>
LXMERT Results

<table>
<thead>
<tr>
<th>Rank</th>
<th>Participant team</th>
<th>yes/no</th>
<th>number</th>
<th>other</th>
<th>overall</th>
<th>Last submission at</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MIL@HDU (MCAN)</td>
<td>90.36</td>
<td>59.17</td>
<td>65.75</td>
<td>75.23</td>
<td>3 months ago</td>
</tr>
<tr>
<td>2</td>
<td>MSM@MSRA</td>
<td>89.81</td>
<td>58.36</td>
<td>65.69</td>
<td>74.89</td>
<td>3 months ago</td>
</tr>
<tr>
<td>3</td>
<td><strong>LXMERT (LXR955, Ensemble)</strong></td>
<td><strong>89.45</strong></td>
<td><strong>56.69</strong></td>
<td><strong>65.22</strong></td>
<td><strong>74.34</strong></td>
<td><strong>3 months ago</strong></td>
</tr>
<tr>
<td>4</td>
<td>AIOZ (AIOZ-QTA)</td>
<td>88.26</td>
<td>55.22</td>
<td>63.63</td>
<td>72.93</td>
<td>3 months ago</td>
</tr>
<tr>
<td>5</td>
<td>Dream</td>
<td>87.69</td>
<td>54.69</td>
<td>65.80</td>
<td>73.12</td>
<td>3 months ago</td>
</tr>
<tr>
<td>6</td>
<td>LXMERT github model (LXR955, Single Model)</td>
<td>87.07</td>
<td>55.80</td>
<td>65.80</td>
<td>72.81</td>
<td>7 days ago</td>
</tr>
<tr>
<td>7</td>
<td>HappyTeam (A-18)</td>
<td>88.24</td>
<td>54.15</td>
<td>62.11</td>
<td>72.08</td>
<td>3 months ago</td>
</tr>
<tr>
<td>8</td>
<td>Dream</td>
<td>87.95</td>
<td>54.17</td>
<td>62.05</td>
<td>71.93</td>
<td>7 days ago</td>
</tr>
<tr>
<td>9</td>
<td>BAN (Bilinear Attention Networks (B))</td>
<td>87.22</td>
<td>54.37</td>
<td>62.45</td>
<td>71.84</td>
<td>4 months ago</td>
</tr>
</tbody>
</table>

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

Top-1 on Natural Language and Visual Reasoning task.
LXMERT Results

**Top-1 on Natural Language and Visual Reasoning task.**

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

<table>
<thead>
<tr>
<th>Rank</th>
<th>Participant team</th>
<th>yes/no</th>
<th>number</th>
<th>other</th>
<th>overall</th>
<th>Last submission at</th>
</tr>
</thead>
<tbody>
<tr>
<td>65.75</td>
<td>75.23</td>
<td>3 months ago</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65.69</td>
<td>74.89</td>
<td>3 months ago</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65.22</td>
<td>74.34</td>
<td>3 months ago</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>63.63</td>
<td>72.93</td>
<td>3 months ago</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>65.80</td>
<td>75.32</td>
<td>3 months ago</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>63.91</td>
<td>72.51</td>
<td>3 days ago</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>62.11</td>
<td>72.08</td>
<td>3 months ago</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>62.05</td>
<td>71.93</td>
<td>7 days ago</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>62.45</td>
<td>71.84</td>
<td>4 months ago</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analysis

Ablation studies and attention graphs.
Results of BERT encoder are similar to LSTM for the baseline model.

### Analysis: LXMERT Ablation Results

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

Table 3: Dev-set accuracy of using BERT.
## Analysis: Ablation Results

Stacking cross-modality layers helps.

### Table 3: Dev-set accuracy of using BERT.

<table>
<thead>
<tr>
<th>Method</th>
<th>VQA</th>
<th>GQA</th>
<th>NLVR²</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM + BUTD</td>
<td>63.1</td>
<td>50.0</td>
<td>52.6</td>
</tr>
<tr>
<td>BERT + BUTD</td>
<td>62.8</td>
<td>52.1</td>
<td>51.9</td>
</tr>
<tr>
<td>BERT + 1 CrossAtt</td>
<td>64.6</td>
<td>55.5</td>
<td>52.4</td>
</tr>
<tr>
<td>BERT + 2 CrossAtt</td>
<td>65.8</td>
<td>56.1</td>
<td>50.9</td>
</tr>
<tr>
<td>BERT + 3 CrossAtt</td>
<td>66.4</td>
<td>56.6</td>
<td>50.9</td>
</tr>
<tr>
<td>BERT + 4 CrossAtt</td>
<td>66.4</td>
<td>56.0</td>
<td>50.9</td>
</tr>
<tr>
<td>BERT + 5 CrossAtt</td>
<td>66.5</td>
<td>56.3</td>
<td>50.9</td>
</tr>
<tr>
<td>Train + BERT</td>
<td>65.5</td>
<td>56.2</td>
<td>50.9</td>
</tr>
<tr>
<td>Train + scratch</td>
<td>65.1</td>
<td>50.0</td>
<td>50.9</td>
</tr>
<tr>
<td>Pre-train + BERT</td>
<td>68.8</td>
<td>58.3</td>
<td>70.1</td>
</tr>
<tr>
<td>Pre-train + scratch</td>
<td><strong>69.9</strong></td>
<td><strong>60.0</strong></td>
<td><strong>74.9</strong></td>
</tr>
</tbody>
</table>

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

Note: The diagram illustrates the flow of information through the network, with arrows representing the direction of data processing. The table shows the accuracy improvements for different configurations of BERT, with the best results highlighted.
Analysis: Ablation Results

<table>
<thead>
<tr>
<th>Method</th>
<th>VQA</th>
<th>GQA</th>
<th>NLVR²</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM + BUTD</td>
<td>63.1</td>
<td>50.0</td>
<td>52.6</td>
</tr>
<tr>
<td>BERT + BUTD</td>
<td>62.8</td>
<td>52.1</td>
<td>51.9</td>
</tr>
<tr>
<td>BERT + 1 CrossAtt</td>
<td>64.6</td>
<td>55.5</td>
<td>52.4</td>
</tr>
<tr>
<td>BERT + 2 CrossAtt</td>
<td>65.8</td>
<td>56.1</td>
<td>50.9</td>
</tr>
<tr>
<td>BERT + 3 CrossAtt</td>
<td>66.4</td>
<td>56.6</td>
<td>50.9</td>
</tr>
<tr>
<td>BERT + 4 CrossAtt</td>
<td>66.4</td>
<td>56.0</td>
<td>50.9</td>
</tr>
<tr>
<td>BERT + 5 CrossAtt</td>
<td>66.5</td>
<td>56.3</td>
<td>50.9</td>
</tr>
<tr>
<td>Train + BERT</td>
<td>65.5</td>
<td>56.2</td>
<td>50.9</td>
</tr>
<tr>
<td>Train + scratch</td>
<td>65.1</td>
<td>50.0</td>
<td>50.9</td>
</tr>
<tr>
<td>Pre-train + BERT</td>
<td>68.8</td>
<td>58.3</td>
<td>70.1</td>
</tr>
<tr>
<td><strong>Pre-train + scratch</strong></td>
<td><strong>69.9</strong></td>
<td><strong>60.0</strong></td>
<td><strong>74.9</strong></td>
</tr>
</tbody>
</table>

Table 3: Dev-set accuracy of using BERT.

Pre-training boosts the performance.
### Analysis: Ablation Results

<table>
<thead>
<tr>
<th>Method</th>
<th>VQA</th>
<th>GQA</th>
<th>NLVR²</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM + BUTD</td>
<td>63.1</td>
<td>50.0</td>
<td>52.6</td>
</tr>
<tr>
<td>BERT + BUTD</td>
<td>62.8</td>
<td>52.1</td>
<td>51.9</td>
</tr>
<tr>
<td>BERT + 1 CrossAtt</td>
<td>64.6</td>
<td>55.5</td>
<td>52.4</td>
</tr>
<tr>
<td>BERT + 2 CrossAtt</td>
<td>65.8</td>
<td>56.1</td>
<td>50.9</td>
</tr>
<tr>
<td>BERT + 3 CrossAtt</td>
<td>66.4</td>
<td>56.6</td>
<td>50.9</td>
</tr>
<tr>
<td>BERT + 4 CrossAtt</td>
<td>66.4</td>
<td>56.0</td>
<td>50.9</td>
</tr>
<tr>
<td>BERT + 5 CrossAtt</td>
<td>66.5</td>
<td>56.3</td>
<td>50.9</td>
</tr>
<tr>
<td>Train + BERT</td>
<td>65.5</td>
<td>56.2</td>
<td>50.9</td>
</tr>
<tr>
<td>Train + scratch</td>
<td>65.1</td>
<td>50.0</td>
<td>50.9</td>
</tr>
<tr>
<td>Pre-train + BERT</td>
<td>68.8</td>
<td>58.3</td>
<td>70.1</td>
</tr>
<tr>
<td>Pre-train + scratch</td>
<td>69.9</td>
<td>60.0</td>
<td>74.9</td>
</tr>
</tbody>
</table>

Table 3: Dev-set accuracy of using BERT.

- **Pre-training boosts the performance.**
- + **24.0%** on NLVR²
- + **10.0%** on GQA
- + **4.4%** on VQA
Analysis: Ablation Results

<table>
<thead>
<tr>
<th>Method</th>
<th>VQA</th>
<th>GQA</th>
<th>NLVR²</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM + BUTD</td>
<td>63.1</td>
<td>50.0</td>
<td>52.6</td>
</tr>
<tr>
<td>BERT + BUTD</td>
<td>62.8</td>
<td>52.1</td>
<td>51.9</td>
</tr>
<tr>
<td>BERT + 1 CrossAtt</td>
<td>64.6</td>
<td>55.5</td>
<td>52.4</td>
</tr>
<tr>
<td>BERT + 2 CrossAtt</td>
<td>65.8</td>
<td>56.1</td>
<td>50.9</td>
</tr>
<tr>
<td>BERT + 3 CrossAtt</td>
<td>66.4</td>
<td>56.6</td>
<td>50.9</td>
</tr>
<tr>
<td>BERT + 4 CrossAtt</td>
<td>66.4</td>
<td>56.0</td>
<td>50.9</td>
</tr>
<tr>
<td>BERT + 5 CrossAtt</td>
<td>66.5</td>
<td>56.3</td>
<td>50.9</td>
</tr>
<tr>
<td>Train + BERT</td>
<td>65.5</td>
<td>56.2</td>
<td>50.9</td>
</tr>
<tr>
<td>Train + scratch</td>
<td>65.1</td>
<td>50.0</td>
<td>50.9</td>
</tr>
<tr>
<td>Pre-train + BERT</td>
<td>68.8</td>
<td>58.3</td>
<td>70.1</td>
</tr>
<tr>
<td>Pre-train + scratch</td>
<td><strong>69.9</strong></td>
<td><strong>60.0</strong></td>
<td><strong>74.9</strong></td>
</tr>
</tbody>
</table>

Table 3: Dev-set accuracy of using BERT.

Loading pre-trained BERT weights into LXMERT pre-training does not help.
Analysis: Visualizing Attention Graphs

Is it warm enough for him to be wearing shorts?
Analysis: Visualizing Attention Graphs

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

http://exbert.net/
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 \[\rightarrow\] Long Paragraph

Caption, Question, Instruction, \[\rightarrow\] News, Books, Tutorial, \[\rightarrow\] .......
An orange cat sits in the suitcase ready to be packed.

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