LXMERT: Learning Cross-Modality Encoder Representations from Transformers

----------**ABSTRACT**----------

Vision-language reasoning requires a certain understanding of visual concepts and language semantics, especially it needs to be able to align and find relationships between these two modalities. The authors proposed the LXMERT framework to learn the connection between these languages and vision. It contains three encoders: an **object-relational encoder**, **a language encoder**, and a **cross-modal encoder**. In order to make the model have the ability to link vision and language semantics, 5 different representative pre-training tasks are used:
(1) Mask cross-modal language modeling
(2) Mask target prediction through ROI feature regression
(3) Masking target prediction by detected label classification
(4) Cross-modal matching
(5) Image problem solving.
These multi-modal pre-training can not only help to learn the connections within the same modal, but also help to learn the cross-modal connections.

---

1. Introduction

For visual-content understanding, people have developed several backbone models shown their effectiveness on large vision datasets also show the generalizability of these pre-trained (especially on ImageNet) backbone models by fine-tuning them on different tasks. In terms of language understanding, last year, we witnessed strong progress towards building a universal backbone model with large-scale contextualized language model pre-training .Despite these influential single modality works, large-scale pretraining and fine-tuning studies for the modality-pair of vision and language are still under-developed. In order to better learn the cross-modal alignments between vision and language, we next pre-train our model with five diverse representative tasks:
(1) masked cross-modality language modeling,
(2) masked object prediction via RoI-feature regression,
(3) masked
object prediction via detected-label classification,
(4) cross-modality matching, and
(5) image question answering.
Further, to show the generalizability of our pre-trained model, we fine-tune LXMERT on a challenging visual reasoning task, Natural Language for Visual Reasoning for Real,where we do not use the natural images in their dataset for our pre-training, but fine-tune and evaluate on these challenging, real-world images.

2. Model Architecture

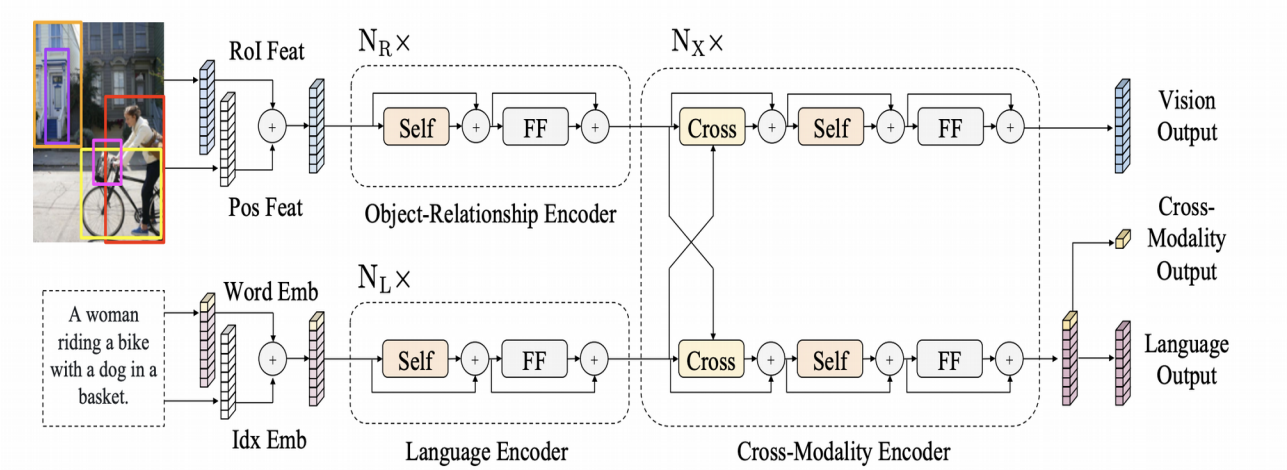


Figure 1: The LXMERT model for learning vision-and-language cross-modality representations. 'Self' and 'Cross' are abbreviations for self-attention sub-layers and cross-attention sub-layers, respectively. 'FF' denotes a feed-forward sub-layer.

We build our cross-modality model with self attention and cross-attention layers following the recent progress in designing natural language processing models. As shown in Fig, our model takes two inputs: an image and its related sentence (e.g., a caption or a question). Via careful design and combination of these self-attention and cross-attention layers, our model is able to generate language representations, image representations, and cross-modality representations from the inputs.

2.1 Input Embeddings

The input embedding layers in LXMERT convert the inputs (i.e., an image and a sentence) into two sequences of features: **word-level sentence embeddings** and **object-level image embeddings**. These embedding features will be further processed by the latter encoding layers

Word-Level Sentence Embeddings :

A sentence is first split into words(w1,....,wn) with lenght of n by the same WordPiece Tokenizer. Next as shown in the figure, the word wi and its index are projected to vectors by embeddings sub-layers and then added to the index-aware word embeddings

Object-Level Image Embeddings :

Instead of using the feature map output by a convolutional neural network in taking the features of detected objects as the embeddings of images. Each object is represented by its position feature (i.e., bounding box coordinates) and its 2048 dimensional region-of-interest (RoI) feature.

In

$$\hat{f}_j = \text{LayerNorm}\left(W_{\text{F}} f_j + b_{\text{F}}\right)$$
$$\hat{p}_j = \text{LayerNorm}\left(W_{\text{P}} p_j + b_{\text{P}}\right)$$
$$v_j = \left(\hat{f}_j + \hat{p}_j\right)/2 \qquad (1)$$

addition to providing spatial information in visual reasoning, the inclusion of positional information is necessary for our masked object prediction pre-training task. Since the image embedding layer and the following attention layers are agnostic to the absolute indices of their inputs, the order of the object is notspecified.

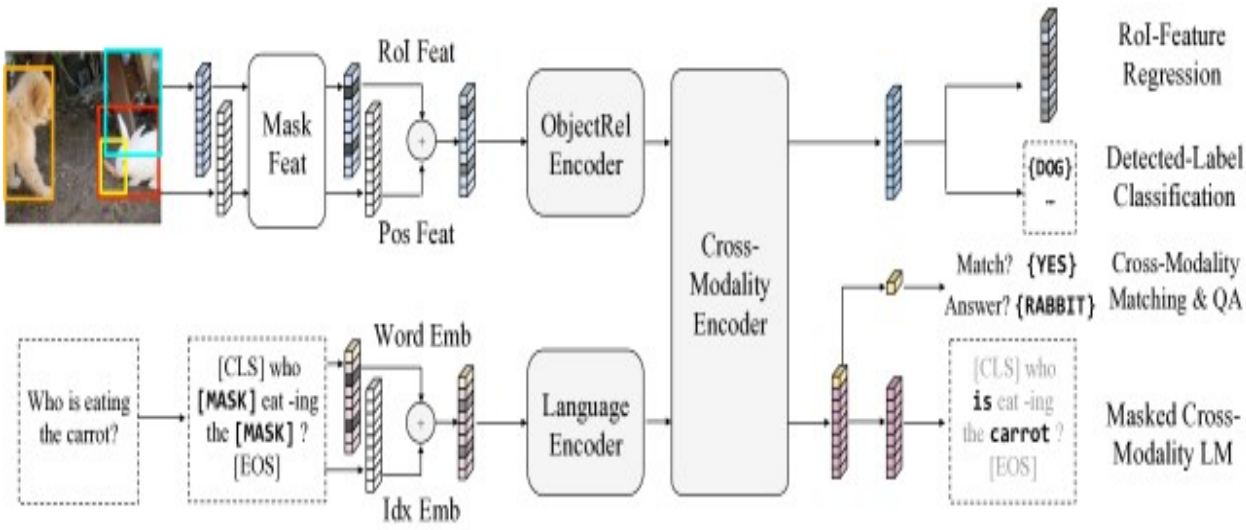
2.2 Encoders

We build our encoders, i.e., the language encoder, the object-relationship encoder, and the cross-modality encoder, mostly on the basis of two kinds of attention layers: **self-attention** layers and **cross-attention layers**.

Single-Modality Encoders

After the embedding layers, we first apply two transformer encoders , i.e., a language encoder and an object-relationship encoder, and each of them only focuses on a single modality (i.e., language or vision). Different from BERT, which applies the transformer encoder only to language inputs, we apply it to vision inputs as well (and to cross modality inputs as described later below).

Cross-Modality Encoder



Each cross-modality layer in the cross-modality encoder consists of two self-attention sub-layers, one bi-directional cross-attention sub-layer, and two feed-forward sub-layers. Inside the kth layer, the bi-directional cross-attention sub-layer (‘Cross’) is first applied, which contains two unidirectional cross-attention sub-layers: one from language to vision and one from vision to language. The query and context vectors are the outputs of the layer. The cross-attention sub-layer is used to exchange the

information and align the entities between the two modalities in order to learn joint cross-modality representations.

2.3 Output Report

LXMERT cross-modality model has three outputs for language, vision, and cross-modality, respectively.

3. Pre-Training

In order to learn a better initialization which understands connections between vision and language, we pre-train our model with different modality pre-training tasks on a large aggregated dataset.

Language Task: Masked Cross-Modality LM

On the language side, we take the masked cross-modality language model (LM) task. 2, the task setup is almost same to BERT words are randomly masked with a probability of 0.15 and the model is asked to predict
these masked words. In addition to BERT where masked words are predicted from the non-masked words in the language modality, LXMERT, with its cross-modality model architecture, could predict masked words from the vision modality as well, so as to resolve ambiguity. Hence, it helps building connections from the vision modality to the language modality, and we refer to this task as masked cross-modality LM to emphasize this difference.

Vision Task: Masked Object Prediction

As shown in the top branch of Fig, we pre-train the vision side by randomly masking objects (i.e., masking RoI features with zeros) with a probability of 0.15 and asking the model to predict proprieties of these masked objects. Similar to the language task (i.e., masked cross-modality, the model can infer the masked objects either from visible objects or from the language modality. Therefore, we perform two sub-tasks: RoI-Feature Regression regresses the object RoI feature with L2 loss, and Detected Label Classification learns the labels of masked
objects with cross-entropy loss.

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

4. Fine-tuning

Fine-tuning is fast and robust. We only perform necessary modification to our model with respect to different tasks. We use a learning rate of 1e−5 or 5e−5, a batch size of 32, and fine-tune the model from our pre-trained parameters for 4 epochs.

5. Datasets

We used three datasets for evaluating our LXMERT framework: VQA v2.0 dataset , GQA and NLVR2

6. Empirical Comparison 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

We compare our single-model results with pre-vious best published results on VQA/ GQA test standard sets and NLVR2 public test set. Our LXMERT model improves consistency (‘Cons’) to 42.1% (i.e., by 3.5 times)

7. Analysis

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**

We analyze our LXMERT framework by comparing it with some alternative choices or by excluding certain model components/pre-training strategies.

8. Limitations of LXMERT

- Intractability
- Reliance on background knowledge
- Failure to process noise and uncertainty.

9. Applications

- Image question answering.
- Visual Reasoning.

10. Conclusion

We presented a cross-modality framework, LXMERT, for learning the connections between vision and language. We build the model based on Transformer encoders and our novel cross-modality encoder. This model is then pre-trained with diverse pre-training tasks on a large-scale dataset of image-and-sentence pairs. Empirically, we show state-of-the-art results on two image QA datasets (i.e., VQA and GQA) and show the model generalizability with a 22% improvement on the challenging visual reasoning dataset of NLVR2. We also show the effectiveness of several model components and training methods via detailed analysis and ablation studies.

You'll find the demo code here

12. References:

- Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh, Hierarchical Question-Image Co-Attention for Visual Question Answering (2016)
- https://github.com/ritvikshrivastava/ADL_VQA_Tensorflow2
- https://www.appliedaicourse.com/
- https://arxiv.org/abs/1908.07490
- https://huggingface.co/transformers/model_doc/lxmert.html
- Hao Tan and Mohit Bansal. *Lxmert: Learning cross-modality encoder representations from transformers.* Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International JointConference on Natural Language Processing (EMNLP-IJCNLP), 2019.
- Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. *Unicoder-vl: A universal encoder forvision and language by cross-modal pre-training*, arxiv 1908.06066, 2019.
- Guillaume Lample and Alexis Conneau. *Cross-lingual language model* pretraining, arxiv 1901.07291, 2019.