LXMERT: Learning Cross-Modality Encoder Representations from Transformers

Hao Tan, Mohit Bansal

Published at EMNLP 2019

Keypoints / Takeaways

Three types of encoders used: object relationship, language and cross modal encoder.

Five types of pre-training tasks are involved: masked language modeling, masked object prediction (feature regression and label classification), cross-modality matching, and image question answering.

Input representation

Word-Level Sentence Embeddings: Word embeddings plus index embeddings.

$$\begin{aligned}
\hat{w}_i &= \mathrm{WordEmbed}\left(w_i\right) \\
\hat{u}_i &= \mathrm{IdxEmbed}\left(i\right) \\
h_i &= \mathrm{LayerNorm}\left(\hat{w}_i + \hat{u}_i\right)
\end{aligned}$$

Object-Level Image Embeddings: Object detected from image along with the positional embeddings(bounding box coordinates)

$$\begin{aligned}
\hat{f}_j &= \mathrm{LayerNorm}\left(W_{\mathrm{F}} f_j + b_{\mathrm{F}}\right) \\
\hat{p}_j &= \mathrm{LayerNorm}\left(W_{\mathrm{P}} p_j + b_{\mathrm{P}}\right) \\
v_j &= \left(\hat{f}_j + \hat{p}_j\right)/2
\end{aligned}$$

Encoders

Three types of encoders used:

1. Language encoder (Single modality)
2. Object relationship encoder (Single modality)
3. Cross modality encoder

Single modality encoders

- Language encoder and an object-relationship encoder.
- Each layer contains a self-attention (‘Self’) sub-layer and a feed-forward (‘FF’) sub-layer.
- Feed-forward sub-layer is further composed of two fully-connected sub-layers.

Cross modality encoders

Each layer of encoder consists of two self-attention sub-layers, one bi-directional cross-attention sublayer, and two feed-forward sub-layers.

The cross modality layers are stacked.

Inside the k-th 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 (k-1)-th layer $\{h_i^{k-1}\}$ and vision features $\{v_j^{k-1}\})$

$$
\begin{aligned}
\hat{h}_i^k &= \mathrm{CrossAtt}_{\mathrm{L}\rightarrow\mathrm{R}}\left(h_i^{k-1},\{v_1^{k-1},\ldots,v_m^{k-1}\}\right)\\
\hat{v}_j^k &= \mathrm{CrossAtt}_{\mathrm{R}\rightarrow\mathrm{L}}\left(v_j^{k-1},\{h_1^{k-1},\ldots,h_n^{k-1}\}\right)
\end{aligned}
$$

Cross modality encoders (contd.)

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.

For further building internal connections, the self-attention sub-layers (‘Self’) are then applied to the output of the cross attention sub-layer

$$\tilde{h}_i^k = \mathrm{SelfAtt}_{\mathrm{L}\rightarrow\mathrm{L}}\left(\hat{h}_i^k, \{\hat{h}_1^k, \ldots, \hat{h}_n^k\}\right)$$

$$\tilde{v}_j^k = \mathrm{SelfAtt}_{\mathrm{R}\rightarrow\mathrm{R}}\left(\hat{v}_j^k, \{\hat{v}_1^k, \ldots, \hat{v}_m^k\}\right)$$

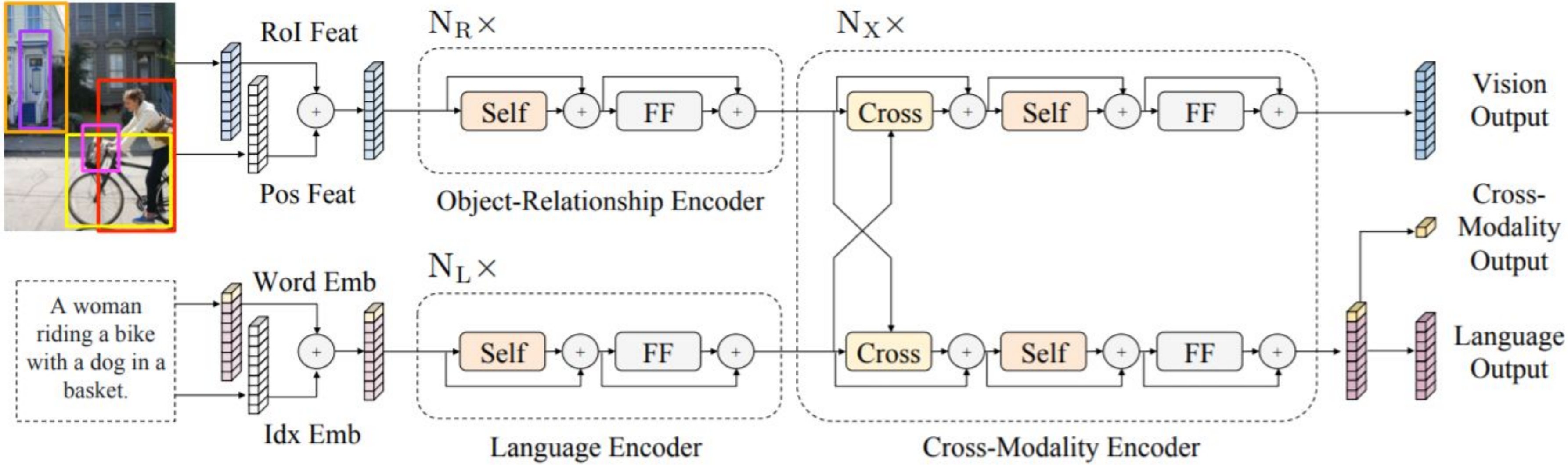


Figure 1: The LXMERT model for learning vision-and-language cross-modality representations. ‘Self’ and ‘Cross’ are abbreviations for self-attention sub-layers and cross-attention sub-layers, respectively. ‘FF’ denotes a feed-forward sub-layer.

Pre-training tasks

Five tasks are used:

- Language Task: Masked Cross-Modality LM
- Vision Task: Masked Object Prediction
  - RoI-Feature Regression
  - Detected Label Classification
- Cross-Modality Tasks
  - Cross-Modality Matching
  - Image Question Answering (QA)

Language Task: Masked Cross-Modality LM

- Words are randomly masked 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.
- For example, as shown in Fig. 2, it is hard to determine the masked word ‘carrot’ from its language context but the word choice is clear if the visual information is considered.

Vision Task: Masked Object Prediction

- Pretrain the vision side by randomly masking objects (i.e., masking RoI features with zeros) and predicting properties of these masked objects.
- Similar to the masked cross-modality LM, the model can infer the masked objects either from visible objects or from the language modality.
- **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.

Cross-Modality Tasks (Cross-Modality Matching)

- For each sentence, with a probability of 0.5, replace it with a mismatched sentence.
- Then train a classifier to predict whether an image and a sentence match each other.
- Similar to ‘Next Sentence Prediction’ in BERT.

Cross-Modality Tasks (Image Question Answering)

- Around 1/3 sentences in the pre-training data are questions about the images.
- Model tries to predict the answer to these image related questions when the image and the question are matched.

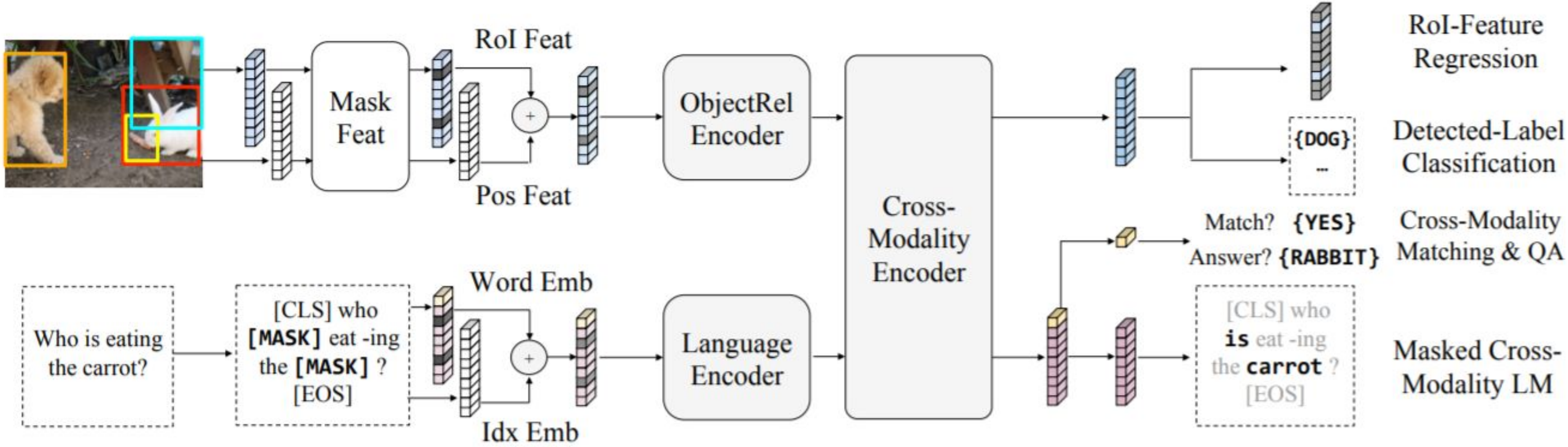
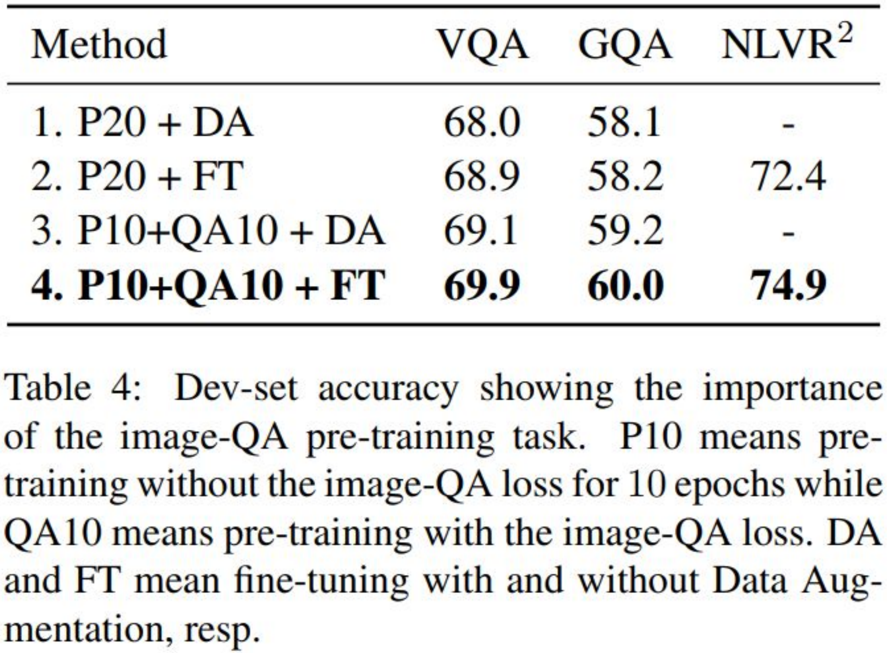


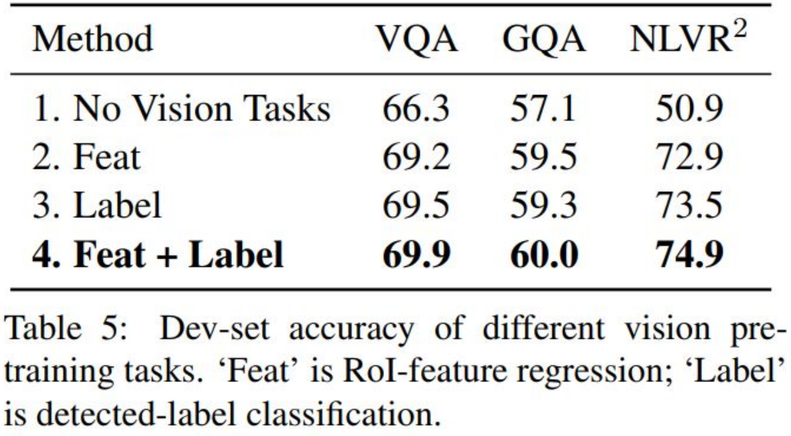
Figure 2: Pre-training in LXMERT. The object RoI features and word tokens are masked. Our five pre-training tasks learn the feature representations based on these masked inputs. Special tokens are in brackets and classification labels are in braces.

Importance of some pre-training tasks



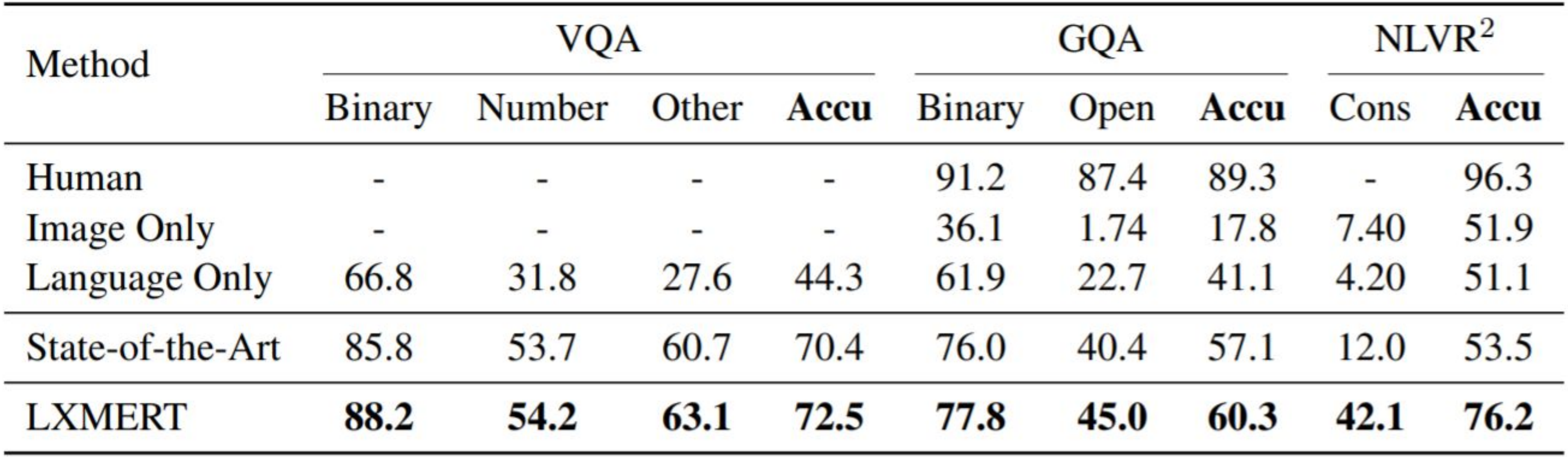
Method	VQA	GQA	NLVR^2
1. P20 + DA	68.0	58.1	-
2. P20 + FT	68.9	58.2	72.4
3. P10+QA10 + DA	69.1	59.2	-
4. P10+QA10 + FT	**69.9**	**60.0**	**74.9**

Table 4: Dev-set accuracy showing the importance of the image-QA pre-training task. P10 means pre-training without the image-QA loss for 10 epochs while QA10 means pre-training with the image-QA loss. DA and FT mean fine-tuning with and without Data Augmentation, resp.



Method	VQA	GQA	NLVR^2
1. No Vision Tasks	66.3	57.1	50.9
2. Feat	69.2	59.5	72.9
3. Label	69.5	59.3	73.5
4. Feat + Label	**69.9**	**60.0**	**74.9**

Table 5: Dev-set accuracy of different vision pre-training tasks. ‘Feat’ is RoI-feature regression; ‘Label’ is detected-label classification.



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**

Table 2: Test-set results. VQA/GQA results are reported on the 'test-standard' splits and NLVR^2 results are reported on the unreleased test set ('Test-U'). The highest method results are in bold. Our LXMERT framework outperforms previous (comparable) state-of-the-art methods on all three datasets w.r.t. all metrics.