LXMERT - Learning Cross Modal Encoder Representation from Transformers

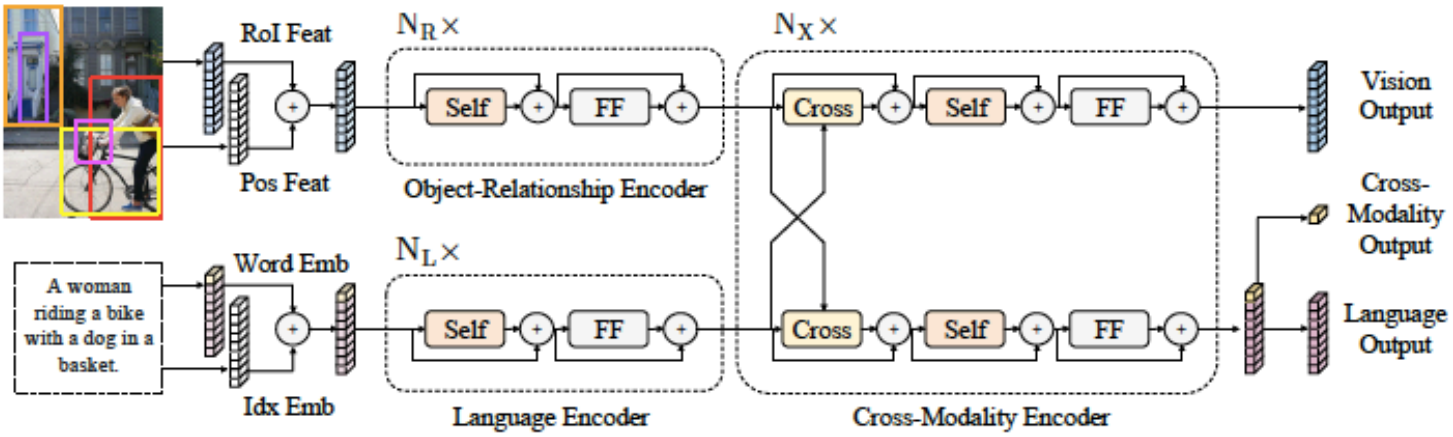

Figure 1: The LXMERT model for learning vision-and-language cross-modality representations. 'Self' and 'Cross' are abbreviations for self-attention sub-layers and cross-attention sub-layers, respectively. 'FF' denotes a feed-forward sub-layer.

- LXMERT tries to make a bert style cross modal transformer for VQA tasks. The model is shown above and consists of 3 encoders
- The image encoder is a self attention encoder----> encodes the objects detected through bounding boxes and also the positional encoding of the bounding boxes
- The question encoder is self attention encoder ----> encodes the word embeddings along with the positional embeddings of the words

- Later we a stack of cross modal encoders. The cross modal transformer is the one that learns the alignment between the image and text modalities.
    - Self attention follows the crossmodal attention and feed forward layers follow that

- LXMERT is pretrained on mutiple tasks - like part of image prediction(4 more such tasks). Here the modal learns to use not just the image but also information from the sentence to complete the image.
    - This helps the model to learn alignment between the image and the words.

- LXMERT is pre-trained on many tasks - check the paper for further details.

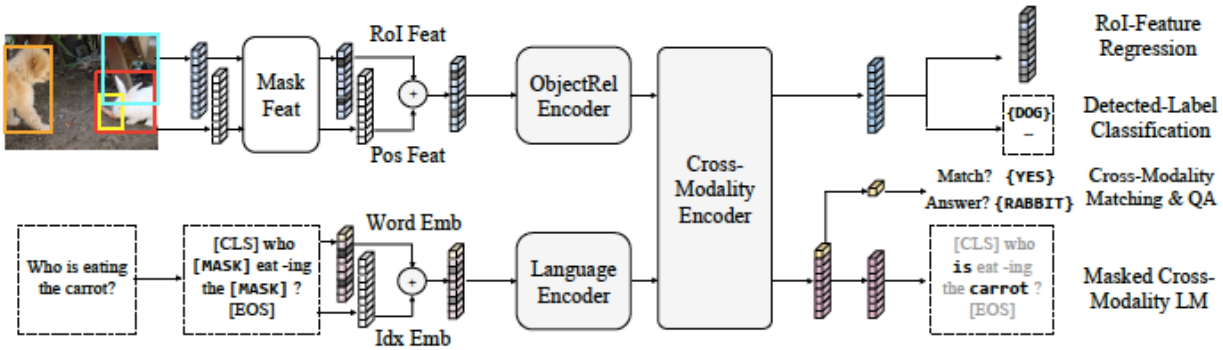

Figure 2: Pre-training in LXMERT. The object RoI features and word tokens are masked. Our five pre-training tasks learn the feature representations based on these masked inputs. Special tokens are in brackets and classification labels are in braces.