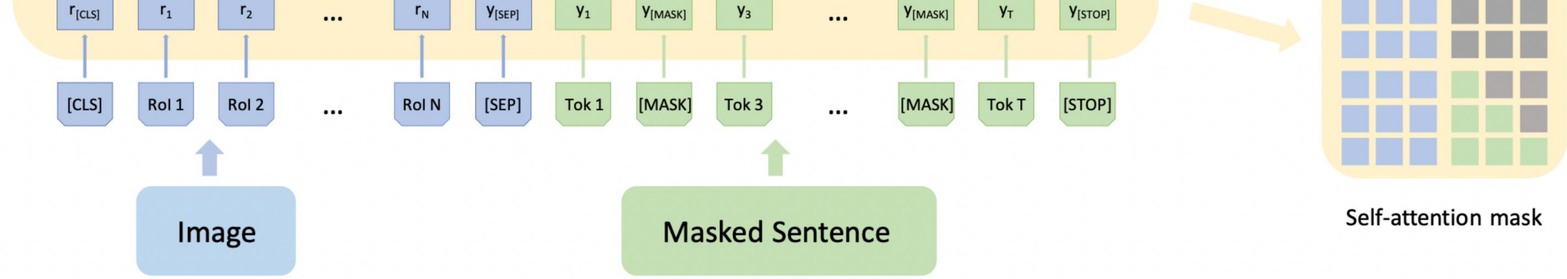

Figure 2: Model architecture for pre-training. The input comprises of image input, sentence input, and three special tokens ([CLS], [SEP], [STOP]). The image is processed as N Region of Interests (RoIs) and region features are extracted according to Eq. 1. The sentence is tokenized and masked with [MASK] tokens for the later masked language modeling task. Our Unified Encoder-Decoder consists of 12 layers of Transformer blocks, each having a masked self-attention layer and feed-forward module, where the self-attention mask controls what input context the prediction conditions on. We implemented two self-attention masks depending on whether the objective is bidirectional or seq2seq. Better viewed in color.

can be formulated as:

$$A^l = \text{softmax}(\frac{Q^\top K}{\sqrt{d}} + M)V^\top, \quad (3)$$

$$V = W_V^l H^{l-1},\ Q = W_Q^l H^{l-1},\ K = W_K^l H^{l-1}, \quad (4)$$

where W_V^l, W_Q^l, and W_K^l are the embedding weights (the bias terms are omitted). The intermediate variables V, Q, and K indicate values, queries and keys, respectively, as in the self-attention module (Vaswani et al. 2017). A^l is further encoded by a feed-forward layer with a residual connection to form the output H^l. During the pre-training, we alternate per-batch between the two objectives and the proportions of seq2seq and bidirectional are determined by hyper-parameters λ and $1-\lambda$, respectively.

It is worth noting that in our experiments we find that incorporating the region class probabilities (C_i) into region feature (r_i) leads to better performance than having a masked region classification pretext as in (Lu et al. 2019; Tan and Bansal 2019). Therefore, differing from existing works where masked region prediction tasks are used to refine the visual representation, we indirectly refine the visual representation by utilizing it for masked language reconstruction. We also choose not to use the Next Sentence Prediction task as in BERT, or in our context predicting the correspondence between image and text, because the task is not only weaker than seq2seq or bidirectional but also computationally expensive. This coincidentally agrees with a concurrent work of RoBERTa (Liu et al. 2019b).

More details follow next in the Image Captioning section.

Fine-Tuning for Downstream Tasks

Image Captioning

We fine-tune the pre-trained VLP model on the target dataset using the seq2seq objective. During inference, we first encode the image regions along with the special [CLS] and [SEP] tokens and then start the generation by feeding in a [MASK] token and sampling a word from the word likelihood output (*e.g.*, greedy sampling). Then, the [MASK] token in the previous input sequence is replaced by the sampled word and a new [MASK] token is appended to the input sequence to trigger the next prediction. The generation terminates when the [STOP] token is chosen. Other inference approaches like beam search could apply as well.

Visual Question Answering

We frame VQA as a multi-label classification problem. In this work we focus on open domain VQA where top k most frequent answers are selected as answer vocabulary and used as class labels. Following (Anderson et al. 2018) we set k to 3129.

During the fine-tuning, a multi-layer Perceptron (Linear+ReLU+Linear+Sigmoid) on top of the element-wise product of the last hidden states of [CLS] and [SEP] is learned, similar to (Lu et al. 2019). We optimize the model output scores with respect to the soft answer labels using