$$D(I_q, I_m) = \sum_{i=1}^{N} |q_i - m_i|^p \quad (5)$$

$$= 2 + \sum_{i|q_i \neq 0, m_i \neq 0} (|q_i - m_i|^p - q_i^p - m_i^p) \quad (6)$$

However, the dissimilarity measure by the L_p-distance is not optimal. As revealed in [182], there exists the neighborhood reversibility issue, which means that an image is usually not the k-nearest neighbor of its k-nearest neighbor images. Such issue causes that problem that some images are frequently returned while others are rarely returned when submitting query images. To address this problem, Jegou *et al.* proposed a novel contextual dissimilarity measure to refine the Euclidean distance based distance [182]. It modifies the neighborhood structure in the BoW space by iteratively estimating distance update terms in the spirit of Sinkhorns scaling algorithm. Alternatively, in [183], a probabilistic framework is proposed to model the feature to feature similarity measure and a query adaptive similarity is derived. Different from the above approaches, in [184], the similarity metric is implicitly learnt with diffusion processes by exploring the affinity graphs to capture the intrinsic manifold of database images.

In [138], Jegou *et al.* investigated the phenomenon of co-missing and co-occurrence in the regular BoW vector representation. The co-missing phenomenon denotes a negative evidence, *i.e.*, a visual word is jointly missing from two BoW vectors. To include the under-estimated evidence for similarity measurement refinement, vectors of images are centered by mean substraction [138]. On the other hand, the co-occurrence of visual words across BoW vectors will cause over-counting of some visual patterns. To limit this impact, a whitening operation is introduced to the BoW vector to generate a new representation [138]. Such preprocessing also applies to the VLAD vector [116]. Considerable accuracy gain has been demonstrated with the above operations.

inconsistency of the matching between two bundled features. In [185][186], Zheng *et al* propose a novel Lp-norm IDF to extend the classic IDF weighting scheme.

The context clues in the descriptor space and the spatial domain are important to contribute the similarity score when comparing images. In [123], a contextual weighting scheme is introduced to enhance the original IDF-based voting so as to improve the classic vocabulary tree approach. Two kinds of weighting scheme, *i.e.*, descriptor contextual weighting (DCW) and spatial contextual weighting, are formulated to multiply the basic IDF weight as a new weighting scheme for image scoring. In [187], Shen *et al.* proposed a spatially-constrained similarity measure based on a certain transformation to formulate voting score. The transformation space is discretized and a voting map is generated based on the relative locations of matched features to determine the optimal transformation.

In [179], each indexed feature is embedded with a binary signature and the image distance is defined as a summation of the hamming distance between matched features, of which the distance for the unobserved match is set as statistical expectation of the distance. Similar scoring scheme for the unobserved match is also adopted by Liu *et al.* [157]. In [63], to tolerate the correspondences of multiple visual objects with different transformations, local similarity of deformations is derived from the peak value in the histogram of pairwise geometric consistency [188]. This similarity score is used as a weighting term to the general voting scores from local correspondences.

In image retrieval with visual word representation, similar to text-based information retrieval [189], there is a phenomenon of visual word burstiness, *i.e.*, some visual element appears much more frequently in an image than the statistically expectation, which undermines the visual similarity measure. To address this problem, Jegou *et al* proposed three strategies to penalize the voting scores from the bursting visual words by removing multiple local matches and weaken the influence of intra- and inter-images bursts [190] [191].

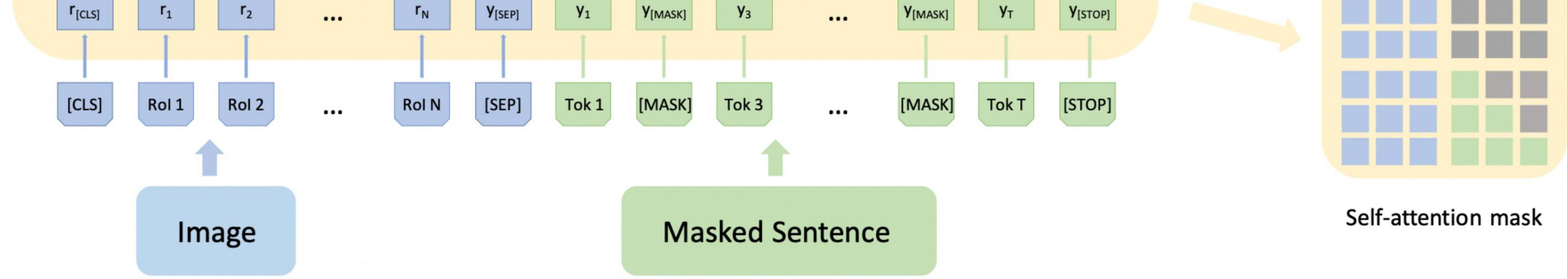

Figure 2: Model architecture for pre-training. The input comprises of image input, sentence input, and three special tokens ([CLS], [SEP], [STOP]). The image is processed as N Region of Interests (RoIs) and region features are extracted according to Eq. 1. The sentence is tokenized and masked with [MASK] 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