Switch-BERT: Learning to Model Multimodal Interactions by Switching Attention and Input Supplementary material

Qingpei Guo\textsuperscript{1}, Kaisheng Yao\textsuperscript{2}, and Wei Chu\textsuperscript{1}

\textsuperscript{1} Ant Group \{qingpei.gqp, weichu.cw\}@antgroup.com
\textsuperscript{2} Amazon AWS AI kaishey@amazon.com

1 An overview of Switch-BERT

We provide further details about Switch-BERT. Fig. 1 illustrates the overall architecture of the proposed model. The visual and text embeddings are fed into Switch-Encoder that consists of a stack of Switch-BERT layers, with Switch-Input Block inserted between consecutive Switch-BERT layers. Three proxy tasks are used for pre-training, including Masked language modeling with visual clues (MLM) \cite{3}, Masked region classification with KL-divergence (MRC-KL) \cite{2} and Image-Text matching (ITM).

2 Implementation details and hyper-parameter tuning

For datasets used in both pretraining and finetuning phase except RefCOCO+, we obtain regional bounding boxes and features from the Faster R-CNN object detector \cite{7} with ResNet101 \cite{4} as the backbone that is well trained on Visual Genome\cite{5}. For the RefCOCO+ dataset, we directly extract the mean-pooled RoI features for bounding boxes provided by \cite{8} from Faster R-CNN. Following \cite{1}, we select the top 36 regional features in each image for training and share this setting across all downstream tasks and model architectures. The maximum number of word tokens varies with different tasks, which are set as 30, 20, and 23, respectively, for image-text retrieval, referring expression comprehension and visual question answering. Switch-BERT is pre-trained with the AdamW \cite{6} optimizer with the following settings: initial\_lr=1e-4, \(\beta_1=0.9\), \(\beta_2=0.999\), \(\epsilon=1e-6\), weight\_decay=0.01, warmup\_ratio=0.1. The initial temperature is set as 5.0 and gradually decay to 0.2 with anneal\_rate=1e-6. For finetuning, the initial temperature is set as 1.0. Switch-BERT is finetuned on 4 Nvidia P100 GPUs for 20 epochs for each downstream task with initial\_lr=1e-4, weight\_decay=1e-4.
Fig. 1: An Overview of the Switch-BERT Architecture. LN, EL, RL each denote layers that perform layer-normalization, embedding lookup and reverse embedding lookup, respectively. (Best viewed in color)
References