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

Qingpei Guo[1], Kaisheng Yao[2], and Wei Chu[1]

[1] Ant Group {qingpei.gqp, weichu.cw}@antgroup.com
[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) [3], Masked region classification with KL-divergence (MRC-KL) [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 [7] with ResNet101 [4] as the backbone that is well trained on Visual Genome[5]. For the RefCOCO+ dataset, we directly extract the mean-pooled RoI features for bounding boxes provided by [8] from Faster R-CNN. Following [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 [6] optimizer with the following settings: initial_lr=1e-4, β_1=0.9, β_2=0.999, ϵ=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 annel_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.

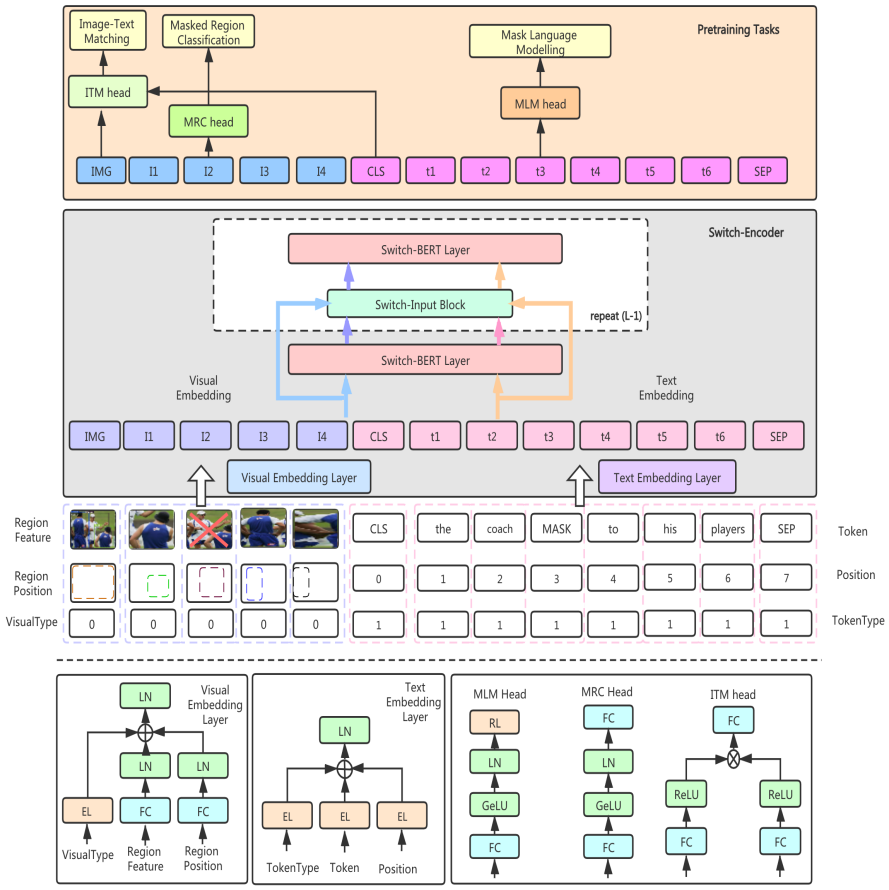

Fig. 1: An Overview of the Switch-BERT Architecture. LN, EL, RL each denote layers that perform layer-normalization, embedding lookup and reverse embedding lookup, respectively. (Best viewed in color)

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
2. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European Conference on Computer Vision. pp. 104–120. Springer (2020)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: The North American Chapter of the Association for Computational Linguistics (2019)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
5. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)
6. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
7. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
8. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1307–1315 (2018)