Multilabel Classification

Wenquan Lu (500443280), Enqi Liu (500518773), Siqi Gao (480365749)

Presented by Team 3DUCKS

COMP5329 Deep Learning

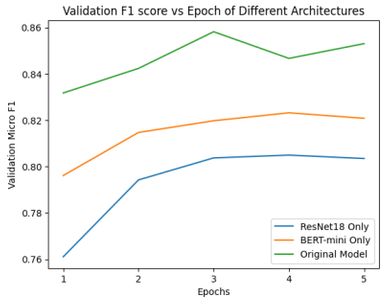
Aim

The aim of the study is to design and implement a deep learning model to perform multi-label classification using image and text data; the goal is to achieve a F1-score as high as possible. We also aim to make the model as lightweight as possible, keeping the size under 100MB, so it is both efficient and accurate.

Methodology

ResNet18 + BERT-mini

We combine a vision model ResNet18 and a language model BERT-mini to make predictions. The outputs of the models are concatenated then passed through two fully-connected layers. The graph below shows removing either model will have a detrimental effect on the accuracy.



ResNet18 (He et al., 2015) is a convolutional neural network that extracts visual features, it introduces skip connections to perform identity mapping between non-consecutive layers. This effectively reduces gradient vanishing problem and allows deeper architecture.

BERT-mini (Turc et al., 2019) is a variant of BERT. It consists of 4 transformer encoder blocks. To leverage the power of BERT-mini, we used the output $h_{[CLS]}$ in the last hidden state that corresponds to the classification token [CLS] for classification. This output summarises and aggregates the entire sequence's representation, which is useful for classification task.

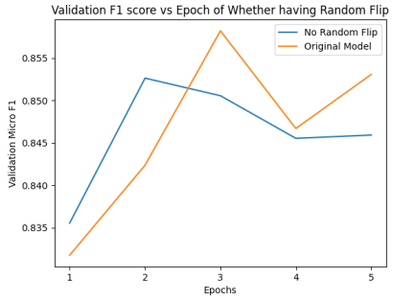
BCEWithLogitsLoss

$$L = -\frac{1}{n}\sum_{i=1}^{n} t_i \log(\sigma(x_i)) + (1 - t_i)\log(1 - \sigma(x_i))$$

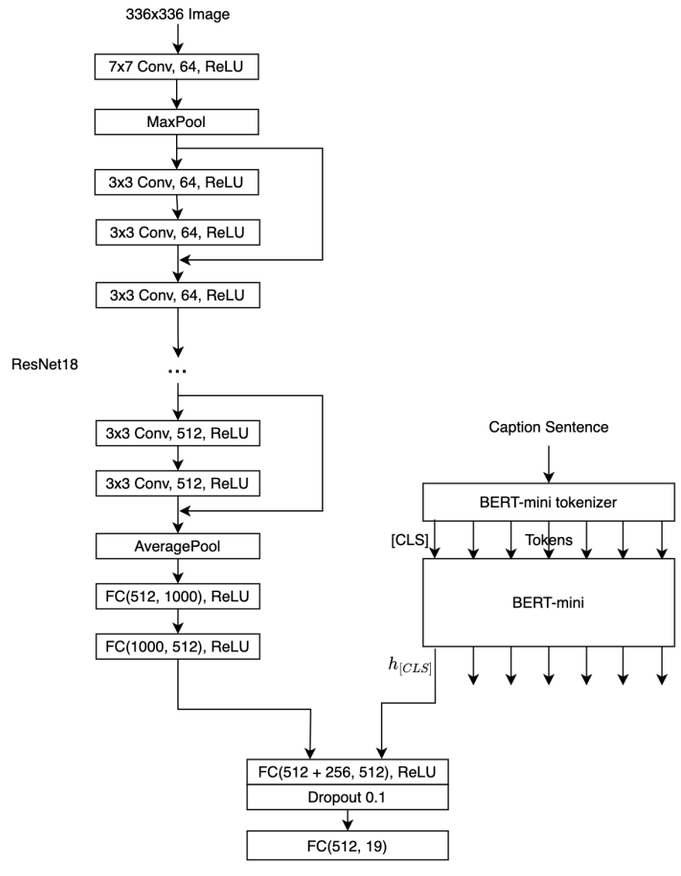
t_i is the target value for label i, $t_i \in \{0, 1\}$, $t_i = 1$ if the label i belongs
to the instance. x_i is the output of the neural network, σ is the sigmoid activation function that transforms x_i into range $(0,1)$. Different to cross entropy loss + softmax, the prediction for each label in this case is independent of each other.

Data Augmentation: Random Flip

Horizontal random flip acts as an effective data augmentation technique, creating a more diverse dataset, and mitigating the overfitting issue.



Architecture



Input Size Adjustment

As shown in the architecture diagram, we resized the input image size to 336x336. Although the standard input size of ResNet is 224x224, we found using higher resolution can improve the classification accuracy by a large margin, as it reveals more details of the visual features.

Hyperparameters

Hyperparameter	Value
Epoch	3
Learning rate	0.00002
Dropout rate	0.1
Batch size	8
Optimizer	ADAM

Result

Metric	Value
Test set F1 score (20%)	0.88651
Test set F1 score (100%)	0.88350
Training time	3 min 54 s
Inference time	1 min 6 s
Size	90.8 MB

The final model achieves a F1 score of 0.88651 on the 20% test set, which is mildly accurate. However, a highlight of our model is that it is very efficient in both model size and training time. The model is lightweight with a size of only 90.8 MB. The training takes under 4 minutes, and the convergence is very fast happening in 3 epochs. The efficiency can be justified by the use of pre-trained models, as many features could be transferred with a little fine-tuning required.

Possible areas for improvement and exploration include incorporating attention mechanisms into the vision model and using a better way to combine the two modules.

Reference

He, K., Zhang, X., Ren, S., & Sun, J. (2015, December 10). Deep Residual Learning for Image Recognition. ArXiv.org. https://arxiv.org/abs/1512.03385

Turc, I., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. ArXiv:1908.08962 [Cs]. https://arxiv.org/abs/1908.08962