Enhancing Multi-label Classification with Multimodal Deep Learning

Junzhi Ning (490059823) and Yuhao Li (520576076)

Presented by Team jnin9950_yuli6273

COMP5329 Deep Learning

Introduction:
We designed a multimodal model for a 19-class multilabel image classification competition.. The dataset contains 30,000 training and 10,000 test samples of images and captions. Our goal was to build a high mean F1 Score performance of model using both visual and text features.

Data Preprocessing:
For images, we resized them to 224x224 and standardized pixel values. For captions, we tokenized the text, padded sequences and added attention masks. We also performed stratification to balance class labels.

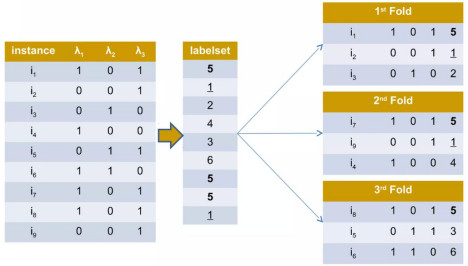

Figure 13: Example of the Data Stratification over a small dataset in accordance to the label distribution

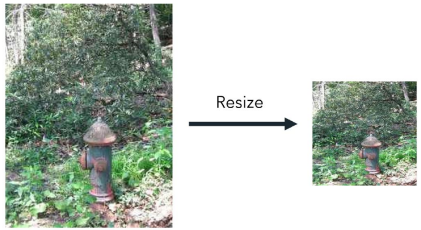

Figure 7: Example of resizing an image from 240 × 320 to 224 × 224.

Features:
We extracted the following features:
Visual features: We used an EfficientNet model as the vision backbone due to its high performance and efficiency. The output from the pretrained model represented the visual embeddings.
Text features: We used a BERT model to obtain contextualized word embeddings from the captions. The output from BERT formed the text embeddings.
Combined features: We concatenated the visual and text embeddings to form the input to our multimodal model. The combined embeddings captured both visual and semantic information.

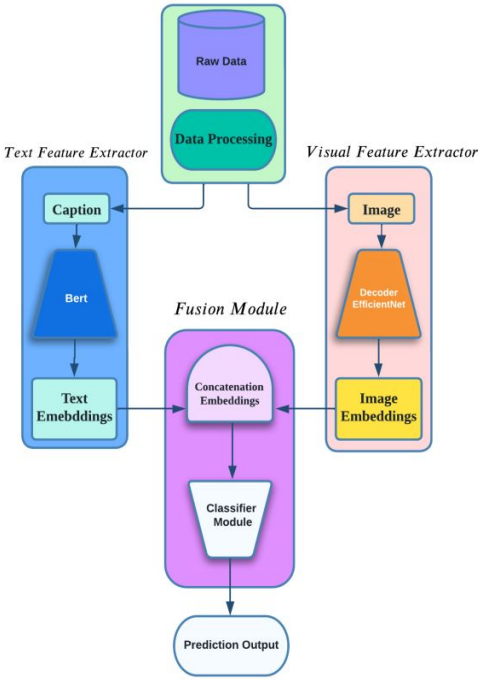

Figure 12: The design of our best Text-Vision Model

Best Model Setting:

Hyper-parameters Name	Setting
Epoch	10
Optimizer	Adam
Loss function	Weighted Binary Cross Entropy
Output threshold	0.45
Dropout probability	0.2
Batch size	256
Learning Rate	2e-3
Default Model:	BERT-Vision
Text Model:	BERT
Vision model:	EfficientNetV2-S
Data Augmentation:	Random Crop & Hori and Vert Flip

Machine Learning Models:
We experimented with different vision backbones (MobileNet, ResNet, EfficientNet), text models (DistilBERT, MobileBERT) and loss functions (BCE, weighted BCE). Hyperparameters like learning rate, dropout and threshold were tuned.
Final Model:
Our final multimodal model uses:
1. EfficientNet-V2 as the vision backbone due to its high performance and efficiency compared to other tested models like MobileNet and ResNet.
2.The full BERT model as the text backbone since it achieved the strongest performance compared to DistilBERT and MobileBERT for our task.
3.A weighted BCE loss function to optimize performance for our imbalanced dataset.

Models	Train Loss	Val Loss	Train Marco F1 (%)	Train Micro F1 (%)	Train Marco F1 (%)	Val Micro F1 (%)	Test F1 (%)	Running Time(s)
Default	0.0664	0.080	74.74	84.83	67.83	83.27	87.041	4383.23
Improved	0.653	0.022	80.82	90.85	79.90	90.37	89.195	1770.84