Multi-label Multi-modal Classification Competition: Advanced Technologies and Analysis

Linear Classifiers: Yanming Guo 520531581, Jiacheng Zhang 52056442, Yongjiang Shi 480406628*

COMP5329 Deep Learning

Overview

Our method utilises cutting-edge technology, Contrastive Language-Image Pre-training (CLIP) [1] as the feature extractor with various fusion methods, classification heads and loss functions. We **ranked third** in the public leaderboard, achieving an F-1 score of **0.90114**.

Big Data or Big Model?

Intivitly, big data leads to better model performance. However, training a sizeable dataset requires a more complex model with more parameters. Therefore, instead of searching external data learning from scratch, we use a powerful pre-trained model trained with 400 million images as the feature extractor and apply transfer learning by fine-tuning with the classification head (see Figure 1).

Fusion Methods

Fusion is a method to combine the knowledge from each modality to enhance the model performance, which provides a more comprehensive understanding of the sentiment being expressed . We explore three fusion methods:

- **Concatenation Fusion:** a early fusion method that combines image and text features before feeding into the classifier.
- **Sum Fusion:** a late fusion method that adds the output of image classifier and text classifier.
- **Mixed Fusion:** a mixed fusion method that combines the concentration fusion and sum fusion.

Sum fusion is chosen as our final fusion method because it is the most stable one and less prone to overfitting (see Figure 2).

Classification Heads

The goal of classification heads is to map the features to the output. We explore two types of classification heads:

- **Multi-Layer Perceptron (MLP):** MLP is a forward-structured artificial neural network that can be considered as a directed graph consisting of multiple layers of nodes, each layer fully connected to the next.
- **Gated Multi-Layer Perceptron (gMLP) [2]:** gMLP is a gating mechanism introduced on top of the traditional MLP. It consists of two main components: a Spatial Gating Unit (SGU) and a fully connected feedforward layer, which controls the flow of information by element-wise multiplication, thereby increasing the expressiveness of the model.

Loss Functions

The choice of loss function is crucial for evaluating the difference between predicted and actual outputs and updating model parameters during training. We explore three loss functions:

- **Binary Cross Entropy**: A common loss function for multilabel classification tasks that simplifies the problem into binary classification tasks and calculates the overall classification loss.
- **Focal Loss:** A specialized loss function for imbalanced multilabel classification problems that down-weights easy examples and emphasizes harder examples during training.
- **Asymmetric Loss [3]:** A loss function for multi-label classification that addresses label imbalance by modulating positive and negative samples separately and incorporating probability shifting.

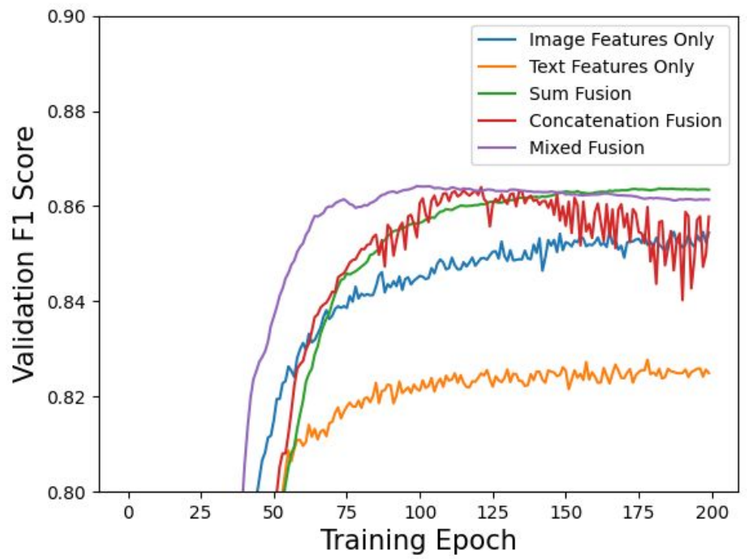

Figure 2: Validation F-1 score of different fusion methods.

Reference

[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

[2] Liu, Hanxiao, et al. "Pay attention to mlps." *Advances in Neural Information Processing Systems* 34 (2021): 9204-9215.

[3] Ridnik, Tal, et al. "Asymmetric loss for multi-label classification." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

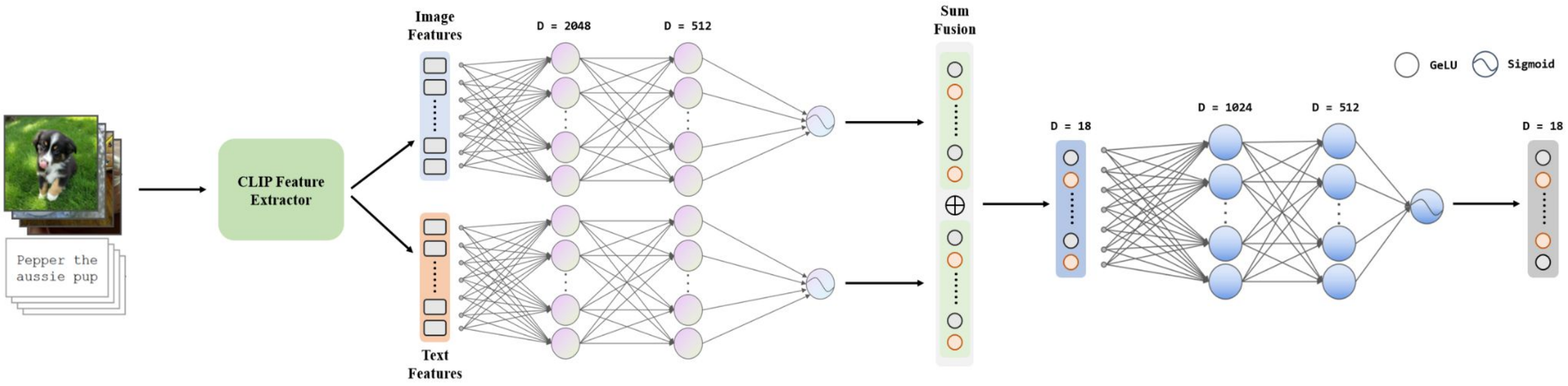

Figure 1: Overview of the pipeline.

*All presenters contribute equally.