
Multimodal Multilabel Classification by CLIP

Yanming Guo
University of Sydney*

Abstract

Multimodal multilabel classification (MMC) is a challenging task that aims to design a learning algorithm to handle two data sources, the image and text, and learn a comprehensive semantic feature presentation across the modalities. In this task, we review the extensive number of state-of-the-art approaches in MMC and leverage a novel technique that utilises the Contrastive Language-Image Pre-training (CLIP) as the feature extractor and fine-tune the model by exploring different classification heads, fusion methods and loss functions. Finally, our best result achieved more than 90% F_1 score in the public Kaggle competition leaderboard. This paper provides detailed descriptions of novel training methods and quantitative analysis through the experimental results.

1 Introduction

Multimodal Multi-label Classification (MMC) refers to a category of learning tasks that involve predicting one or more labels based on two modalities of information: images and text. MMC tasks present a greater complexity level than traditional single-modality, single-label classification tasks. This is primarily due to the requirement for models to process and interpret data from two distinct sources simultaneously. Designing an effective MMC system presents significant challenges, as it necessitates the creation of a learning structure capable of handling and fusing disparate input sources. This fusion must produce a comprehensive feature representation that integrates visual information with semantic content. The design of such a structure requires a deep understanding of both the visual and textual modalities. Furthermore, multi-label tasks inherently have more difficulty than single-label. In single-label classification tasks, the prediction with the highest posterior probability is typically selected as the output. Conversely, in multi-label classification tasks, the model is tasked with making predictions for each label, thereby increasing the complexity of the task. An additional layer of complexity in multi-label classification tasks arises from the intricate underlying dependencies that exist across labels. For instance, the label 'boat' is often associated with the label 'water'. Learning such relationships and incorporating them into the model further increases the complexity of the task. Therefore, the multimodal nature and the multi-label aspect of MMC tasks add to the complexity, making them a challenging yet intriguing area for machine learning research and development.

This study aims to classify 30000 multi-label images with textual descriptions of various sizes comprising 18 distinct classes. Each image is associated with one or more classes and a corresponding caption. The study has two primary objectives: Firstly, the study aims to experiment with different techniques that fuse the extracted features from the image and caption together. Secondly, the study plans to design and optimize the classification heads' structure and find a suitable design to enhance the models' performance. The performance of each technique and design choice will be compared and discussed, with the F_1 score serving as the primary performance metric. After obtaining the

*This work partially fulfils the requirements for the Bachelor of Science (Honours) degree at the University of Sydney 2023. The GitHub repository is: <https://github.com/Yvnminc/Multimodal-Multilabel-Classification-by-CLIP>

experiment’s performance measures, the study will thoroughly analyze and justify the multimodal model’s performance.

In today’s society, people consume vast amounts of multimedia content across various modalities, such as text, image, audio, video, and 3D. In many applications, utilising multiple modalities is extremely important, as it enhances performance by leveraging complementary information from different modalities, resulting in improved accuracy and robustness. It enables a richer representation of data by integrating distinct aspects conveyed by different modalities, facilitating more nuanced and accurate classification. Multimodal classification also helps handle ambiguity in classification problems by overcoming the limitations of individual modalities. The additional modalities essentially serve as extra opinions, effectively disambiguating data. Furthermore, studying the techniques and methods that deal with real-world data, which are usually multimodal, is crucial, enabling effective handling of diverse data types. Studies like this enable the development of models capable of cross-modal understanding and reasoning with practical implications in various domains.

Learning a good feature presentation from a multimodal is a challenging task, as one must aim to learn robust and generalized features. Therefore, a straightforward solution is to train the model with many parameters on the sizable dataset. However, due to the limited training data, the increase in model parameters will lead to an overfitting problem. Moreover, the model size is constrained to 100 MB. Therefore, instead of training the model from scratch, we leverage the pre-trained model training on the data set and fine-tune it with the downstream dataset. We found that CLIP is the model that satisfies this objective well, as CLIP is trained on 400 million images and shows an impressive transferable ability. Therefore, our method is inspired by the motivation and fine-tuning of the classification heads. This work explores different classifiers, loss functions and fusion methods.

2 Related Works

2.1 Multimodal Learning

Multimodal learning, a burgeoning branch of machine learning, aims to amalgamate information from disparate modalities into a unified model structure. The input data could span across various formats, including but not limited to images, texts, audio, and video. Fusing this multimodal information, achieved by a learning algorithm, results in a comprehensive and precise understanding encapsulated within a unified latent feature space. In the realm of image-language learning, the input data generally consists of two components: the visual content (images) and the corresponding textual information that describes the visual content. This form of multimodal learning has a broad spectrum of real-world applications. A notable instance is Visual Question Answering (VQA), where algorithms are designed to interpret images and comprehend text, thereby enabling human-like interaction [1]. Other applications include text grounding, which aims to identify objects based on textual descriptions. [2] proposed a method to localize segments within a video based on the classification sense. Multimodal learning holds immense potential and is a promising direction for future research.

The multimodal learning framework could be summarised as consisting of three parts: the text encoder, the image encoder and the modality fusion module. The major difference between each method could be the different backbone networks, fusion methods and the size of two encoders. The multimodal networks could be categorised into four classes [3], shown in Fig.(1). Specifically, many models fall in the first category, exploiting a larger image encoder than the text encoder. For instance, the VSE++ [4] is an improved visual semantic embedding (VSE) model architecture utilising the hinge-based triplet ranking loss to better compare the positive and negative samples. Moreover, SCAN [5] leverages the cross-attention mechanism to understand better the highlighted part between the image and text feature.

As for the second class, where the text encoder and image encoder have a close computational size, Contrastive Language-Image Pre-training (CLIP [6]) is an example that leverages the contrastive learning paradigm to jointly train image-text pairs in a self-supervised fashion, guided by language-based supervision. The fundamental premise of CLIP is the shared semantic information between images and language. The model performance can be significantly enhanced by learning from images and text. Moreover, the availability of vast quantities of unlabelled images accompanied by

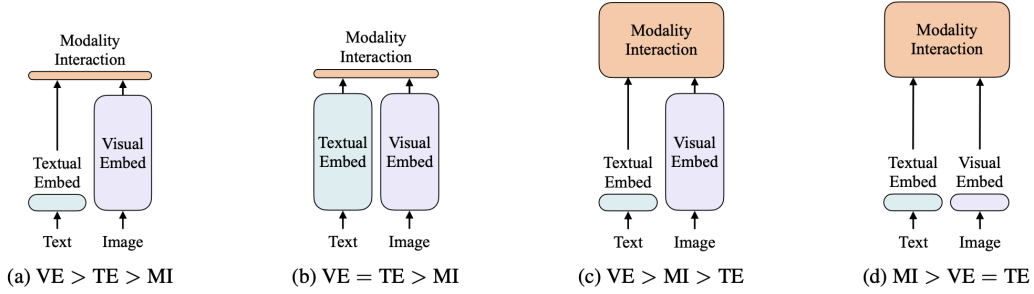


Figure 1: Four classes of the proposed methods according to the size of each module [3].

captions facilitates the training process. CLIP has been trained on a dataset encompassing 400 million image-text pairs, thereby extending its transferability capabilities.

The architecture of CLIP is illustrated in Fig.(2). It comprises two encoders, one for text and image, extracting features from the respective modalities. These features are subsequently fused through a joint multimodal embedding process, which involves computing the scaled pairwise cosine similarities after L2 normalization. The predicted embeddings are then contrasted with the positive samples, with the choice of positive and negative samples determined by their location on or off the diagonal, respectively (see Fig.(2)). This process encourages the model to maximize the similarity of matched image-text pairs while minimizing the similarity of unmatched pairs. This leads to a better understanding of the shared semantic space between images and text. In terms of applications, CLIP has demonstrated remarkable performance across various tasks, such as image classification and object detection, exhibiting its broad applicability and effectiveness in diverse scenarios.

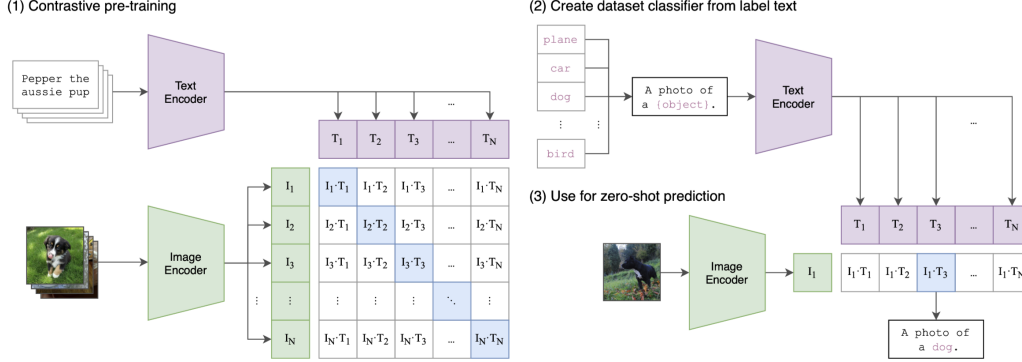


Figure 2: Structure of the CLIP network [6].

Feature-wise Linear Modulation (FiLM [7]) introduces a linear projection layer to transfer the text feature into a visual model. Therefore, FiLM could be classified into the third category with the higher computational cost for the fusion of two modality information.

The last class where has a significant computational size of the modal interaction but less for the text and image encoder. Vision-and-Language Transformer (ViLT [3]) is one example that is a new novel multimodal learning framework. The significant contribution of ViLT is simplifying and optimising the training speed for the multimodal learning framework. The input for the ViLT is the direct concatenation of the text and the image data. However, most of the computational cost is made for the fusion of two sources, which can be visually illustrated in Fig.(3).

2.2 Backbones Networks

In developing deep learning networks, there are three major backbones for multimodal learning: CNNs, RNNs and Transformers.

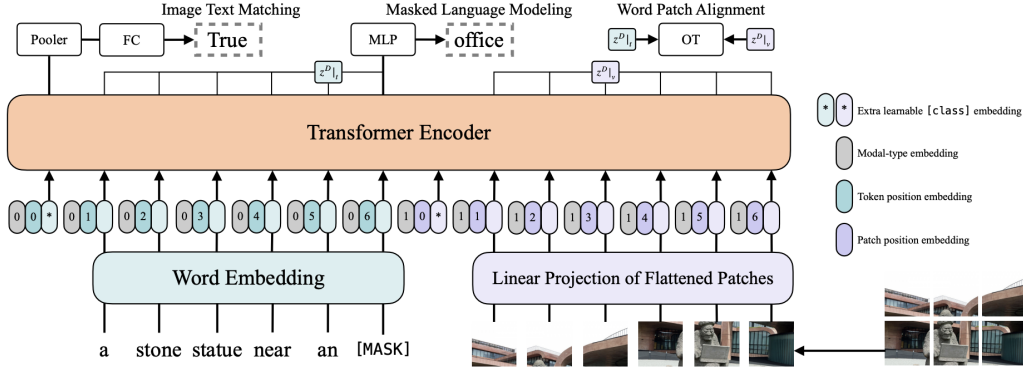


Figure 3: Structure of the ViLT network [3].

Convolutional Neural Networks (CNNs) CNNs are the conventional choice for the backbone for extracting the image feature in multimodal learning, which is optimised for the vision tasks. There are many existing CNN architectures. AlexNet [8] achieves impressive results in the ImageNet, which makes CNNs a popular choice for vision tasks. After that, VGGNet [9] and GoogleNet [10] were proposed and made CNNs deeper. However, with the depth increase, the gradient becomes smaller during the chain rule production. Therefore, ResNet [11] was postponed to address the problem. Another problem in deep neural networks is the high computational cost. Then, depthwise separable convolution [12] is proposed to reduce the model parameters. Another solution to improve efficiency is utilising Neural Architecture Search (NAS) to find the network structure automatically. Therefore, EfficientNet [13] compound scaling factor to find a high-performance network with less computational resource requirements.

Recurrent Neural Networks (RNNs) As for the text module, the conventional backbone network is RNNs [14] to deal with sequential input such as texts or audio. RNNs could extract extended dependency features through sequential data. Long Short-Term Memory (LSTM [15]) is a variant of RNNs that introduces the gates to balance the long and short dependency. Bidirectional Long Short-Term Memory (BiLSTM [16]) consists of two LSTMs to deal with the sequential input from left to right and right to left, respectively. However, the structure of LSTM is complicated both in implementation and computation. Therefore, a Gated Recurrent Unit (GRU [17]) is proposed consisting of two gates, and it achieves similar performance with LSTM but higher efficiency.

Transformer The Transformer [18] is the third neural network architecture after CNNs and RNNs. Compared with RNNs sequentially processing with text input. Transformers could deal with this parallelly by leveraging the attention mechanism. Bidirectional Encoder Representations from Transformers (BERT [19]) is a thriving network structure in Natural Language Processing through a directional encoder to better learn the context information. Opposite to BERT, GPT [20, 21, 22] utilizes a decoder for the text generation task. The parameters of GPT3 increase to 175 billion, proving the Transformer’s scalability. Moreover, the Transformer becomes a uniform architecture for vision and text modules. Vision Transformer [23] is proposed to show that self-attention could also benefit vision tasks. To better improve the performance of ViT, SWin Transformer [24] was proposed with a CNN-like hierarchical structure.

3 Method

3.1 Feature Extraction

Given the dataset contains both image and textual information, it is natural to ask how much content we should leverage for the multilabel classification task. Intuitively, natural languages, especially image captions, can better supervise image classification tasks because they usually contain the concepts of images. Therefore, we decided to use both of them in our project. Specifically, we harness the power of the CLIP model, which learns patterns from supervision contained in language information [6].

The philosophy of CLIP can be explained into two downstream tasks: image feature extraction and text feature extraction.

Image Feature Extraction For image feature extraction, a Vision Transformer (ViT) [23] is used. Initially, the input image is partitioned into equal-sized patches. These patches are then flattened for ease of processing. However, given the propensity for the dimensionality to become excessively large post-flattening, linear mapping is employed to reduce the data to a more manageable, lower-dimensional space. Subsequently, positional embeddings are appended to these patches, which allows the model to effectively keep track of the location of each patch within the original image. As a fourth step, a class token is added, which aids in deriving the encoded characteristics of the input image. This data is then supplied to the encoder for further refinement and extraction.

Text Feature Extraction On the other hand, a language Transformer [18] is used to extract text features. Specifically, as a sequence of text enters into an encoder, it first traverses a self-attention layer that allows the encoder to evaluate the rest of the words in the input sequence during the word encoding process. A fully connected feed-forward neural network then processes the product of this self-attention operation. Although the quantity of parameters in each encoder’s feed-forward neural network is identical, their functions operate independently. The decoder component mirrors this layered structure but incorporates an additional Attention layer. This layer aids the decoder in focusing on the word that corresponds to the input sentence, further enhancing the feature extraction process.

By contrastively combining these two processes during the training, CLIP can learn a shared semantic space between images and text. According to [6], CLIP is much more robust to distribution shift than standard ImageNet models such as ResNet families [25]. As shown in Fig.(4), Zero-Shot CLIP demonstrates its dominated generalization ability to unseen data. This supports that CLIP’s capability to learn joint representations for images and text from the internet-scale dataset significantly improves the quality of the extracted features, thereby enhancing the accuracy of our task.







	Dataset Examples	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet		76.2	76.2	0%
ImageNetV2		64.3	70.1	+5.8%
ImageNet-R		37.7	88.9	+51.2%
ObjectNet		32.6	72.3	+39.7%
ImageNet Sketch		25.2	60.2	+35.0%
ImageNet-A		2.7	77.1	+74.4%

Figure 4: Generalization ability of ResNet-101 vs. Zero-Shot CLIP [6].

3.2 Fusion

Since it is a multi-modality task, we aim to find a way to combine the knowledge from each modality to enhance the model performance. Fusion is the answer. Fusion can be implemented at different levels, such as feature and decision levels, depending on where the fusion occurs. In this project, we

implement three different fusion methods: one feature-level fusion (i.e., concatenation fusion), one decision-level fusion (i.e., sum fusion) and one mixed fusion.

Concatenation Fusion Concatenation fusion combines image and text features before feeding into the classification head. It first extracts the features of both the images and captions as two matrices with the same dimension. Then, merge the text and image information by flattening them horizontally, resulting in a combined matrix with twice as many dimensions. This simple technique can guide the model to learn more accurate feature data than relying on text or image features alone.

Sum Fusion Similar to concatenation fusion, the philosophy of sum fusion is addition. However, different to concatenation fusion, instead of fusing features, sum fusion adds the output of image features and text features after being fed into the classification head. As a result, this approach will treat the output of the two modalities as an input to the meta-classifier. That is to say. The final decision comprehensively considers the results of both image and text features.

Mixed Fusion Mixed fusion combines the idea of concatenation fusion and sum fusion. The procedure begins with concatenation fusion to merge the image and text features. Then, train the concatenated attributes to obtain the output. Next, feed the image features and text features separately into the classification head to obtain the second and the third outputs. Finally, sum fusion is used to sum up three output features and retrain them with a meta-classifier to produce the final outcome.

Using fusion methods has several potential benefits in general. Firstly, fusion helps improve the robustness of a model because other modalities can compensate for the unreliability of one modality. Secondly, fusion provides a more comprehensive understanding of the sentiment being expressed because different data sources can provide different perspectives on a problem. For example, the image data could provide information about what the objects look like, and the text data could provide information about the image’s content. Furthermore, fusion methods often have more excellent generalization capabilities, especially when they are designed to handle input from multiple modalities. This can lead to better performance on unseen data.

Specifically, mixed fusion is a novel fusion approach with multiple advantages. Firstly, combining feature fusion and decision fusion leverages the strengths of both methods, potentially leading to more accurate and robust predictions. Secondly, processing image and caption features both together and separately ensures a comprehensive understanding of the data. This could be particularly beneficial in tasks where the images and captions provide complementary information.

3.3 Loss Function

The loss function is a fundamental aspect of deep learning that evaluates the difference between the predicted and actual output and is employed to update the model parameters during training. The choice of loss function is influenced by the nature of the task and the data being analyzed. For instance, the cross-entropy loss function is commonly employed in binary classification tasks, while the mean squared error (MSE) or mean absolute error (MAE) loss functions are prevalent in regression tasks. Recently, a rising interest has been in designing new loss functions that embed domain-specific expertise, such as domain adaptation and resilience to adversarial attacks. Selecting and creating a loss function is a crucial factor in developing deep learning models, as it directly influences their performance and generalization capacity.

In developing our model, we experimented with three different types of loss functions: binary cross-entropy, focal loss, and asymmetric loss.

Binary Cross Entropy Binary cross-entropy is a common loss function in deep learning for multilabel classification tasks. In the context of multi-label classification, it is a common practice to simplify the problem by transforming it into a sequence of binary classification tasks. Specifically, in this approach, the base network generates a single logit, z_k , for each of the K labels. The activation of each logit is carried out independently, using a sigmoid function $\delta(z_k)$. The ground truth for class k is denoted as y_k . The overall classification loss, L_{tot} , is derived by summing up the binary loss for all K labels:

$$L_{tot} = \sum_{k=1}^K L(\delta(z_k), y_k), \quad (1)$$

where the binary loss per label, L , is given by:

$$L = -yL_+ - (1 - y)L_-, \quad (2)$$

where y is the ground-truth label, and L_+ and L_- are the positive and negative loss parts, respectively. Binary cross-entropy loss has several advantages: it is differentiable, computationally efficient, and easy to interpret.

Focal Loss The Focal loss function is a specialized loss function tailored for multilabel classification problems characterized by imbalanced class distributions [26]. In such scenarios, the distribution of samples may be skewed towards one or a few dominant classes, whereas the minority classes may have an insufficient number of instances. As a result, the model’s predictions may be biased toward the dominant classes, thereby leading to suboptimal performance on the minority classes. Focal loss addresses this issue by selectively down-weighting the contribution of the easy examples (i.e., well-classified negative samples) and emphasizing the contribution of the harder examples (i.e., misclassified rare positive samples) during training. Specifically, the Focal loss function introduces a modulating factor that magnifies the impact of the loss incurred by the hard-to-classify samples and diminishes the impact of the loss incurred by the easy-to-classify samples. This modulating factor is governed by a user-specified parameter known as the focal parameter, which governs the degree of emphasis on the hard examples. Focal loss is obtained by further setting L_+ and L_- as:

$$\begin{cases} L_+ = (1 - p)^{\gamma} \log(p) \\ L_- = p^{\gamma} \log(1 - p) \end{cases}, \quad (3)$$

where $p = \delta(z)$ is the network’s output probability and γ is the focusing parameter ($\gamma = 0$ yields binary cross-entropy). By setting $\gamma > 0$ in Eq.(3), the contribution of easy negatives (having low probability, $p \ll 0.5$) can be down-weighted in the loss function, enabling to focus more on harder samples during training. Focal loss has been shown to improve the performance of deep learning models on imbalanced multilabel classification tasks and has become a popular choice in the research community.

Asymmetric Loss In multi-label classification, there is often an imbalance between the number of positive and negative labels due to the large number of possible labels. Typically, an image will have only a few positive labels and many negative ones. This imbalance can dominate the optimization process during training and lead to insufficient emphasis on gradients from positive labels, resulting in lower accuracy. To address this issue, a new asymmetric loss function called ASL has been introduced [27]. ASL operates differently on positive and negative samples, allowing for dynamic down-weighting and hard-thresholding of easy negative samples and the removal of possibly mislabelled samples. ASL has two main features: First, the positive and negative samples are modulated separately using different exponential decay factors to focus on hard negatives while maintaining the contribution of positive samples. Second, the probabilities of negative samples are shifted to discard very easy negatives using hard thresholding, enabling the removal of challenging negative samples suspected to be mislabelled.

To define Asymmetric Loss (ASL), the two mechanisms, asymmetric focusing and probability shifting, are integrated into a unified formula:

$$\begin{cases} L_+ = (1 - p)^{\gamma} \log(p) \\ L_- = (p_m)^{\gamma} \log(1 - p_m) \end{cases}, \quad (4)$$

where the shifted probability p_m is defined as:

$$p_m = \max(p - m, 0). \quad (5)$$

ASL allows us to apply two types of asymmetries for reducing the contribution of easy negative samples to the loss function: soft thresholding via the focusing parameters $\gamma_- > \gamma_+$ and hard thresholding via the probability m .

Experimenting with different types of loss functions is essential when developing deep learning models because the choice of loss function significantly impacts the model’s performance. Different loss functions are designed to optimize for different objectives, such as classification accuracy or robustness to outliers. By trying out different loss functions, we can identify which loss function is best suited to the task at hand and can fine-tune the model to achieve better performance. Besides the common choice of binary cross-entropy and focal loss, we implemented the novel asymmetric loss (ASL), which addresses the inherent imbalance nature of multi-label datasets.

3.4 Exponential Moving Average (EMA)

The exponential moving average (EMA) represents a technique utilized in neural network training that may enhance model accuracy. Specifically, rather than adopting the optimized parameters derived from the final training iteration as the ultimate parameters for the model, the EMA of the parameters obtained throughout all the training iterations is employed. In supervised learning, a neural network is trained on a labelled dataset D consisting of N instances to optimize the network parameters Θ by minimizing the associated loss function $L(D; \Theta)$. When the neural network does not overfit the training dataset, and the validation dataset is representative of the training dataset, the parameters that minimize the training loss also minimize the validation loss. These parameters are called Θ^* , corresponding to a local optimum.

However, optimizing the parameters using the entire training dataset can be computationally expensive, particularly with standard optimization techniques such as gradient descent, given the large size of N . Therefore, training is often performed on small-sized samples randomly extracted from the training dataset, referred to as mini-batches. Although mini-batches introduce noise into the training process, this noisy gradient may sometimes be advantageous to the optimization process, resulting in a better local optimum than the one obtained with the entire dataset. Nonetheless, as parameter optimization approaches convergence, the optimized parameters may not represent a local minimum. Still, they may fluctuate around it, representing a disadvantage of the training noise introduced by mini-batches.

To mitigate this problem, smoothing techniques, such as the exponential moving average (EMA), can be employed to reduce the optimized parameter fluctuation noise. The parameter EMAs are more likely to be closer to a local minimum. Specifically, the optimized parameters after each update step are $\Theta_1, \Theta_2, \dots, \Theta_t, \dots, \Theta_n$, forming a sequence of time-series data that contains noise. EMA can be applied to this sequence to improve model generalization.

The principle behind EMA is to smooth out the noise in the training process and reduce the variance of the model. To remove the noise, the EMA for the optimized parameter is computed as:

$$EMA_t = \begin{cases} \Theta_1 & t = 1 \\ \alpha EMA_{t-1} + (1 - \alpha)\Theta_t & else \end{cases} \quad (6)$$

As shown in Eq.(6), EMA is implemented by taking a weighted average of the current and previous model parameters. The weight given to the current parameters is higher than the previous ones. This ensures that the current parameters have a larger influence on the new parameter values while still retaining some of the information from the previous parameters. The weight given to the current parameters is controlled by a hyperparameter α called the decay rate. The higher the decay rate, the more weight is given to the current parameters and the more the model is adapted to the current data.

EMA is typically used in the context of stochastic gradient descent optimization, where the model parameters are updated after each mini-batch of data. Using EMA, the model can adapt to the current mini-batch of data while retaining some information from the previous mini-batches, resulting in more stable and better-performing models.

The exponential moving average (EMA) provides several benefits when used in deep learning models. First, it stabilizes the training process by reducing the variance of the gradients during training. This helps to smooth out fluctuations in the gradient updates and provides a more stable estimate of the gradient. Second, EMA improves generalization performance by preventing overfitting and reducing the impact of outliers and noise in the training data. Third, it increases the model's robustness to changes in the training data, preventing overreaction to small changes. Finally, EMA can improve the convergence speed of the training process by reducing oscillations in the loss function and helping the model to reach the optimal solution faster.

The developmental trajectory of Model EMA was initially paved with Snapshot Ensembles [28]. This approach entailed utilizing a cyclic learning rate to obtain multiple model checkpoints from a single training session, followed by the ensemble averaging of predictions. This technique represented a favourable alternative to the erstwhile norm of training multiple models from scratch to create the ensemble. Subsequently, the logical progression from prediction averaging was to incorporate weight averaging, leading to the advent of Stochastic Weight Averaging (SWE) [29]. Before

SWE, Mean Teacher employed a running average of model weights [30]. Today, EMA remains a prevalent tool in recent semi-supervised and unsupervised methodologies, such as BYOL [31].

3.5 Classification Head

Given that our task is image classification, the goal is to map the high-level features learned during the feature extraction phase to the classes the model is trying to predict. Since the feature extraction and classification process is separated into two parts, we have to design a classification head to make final predictions. Specifically, the classification head is usually a more straightforward structure, often consisting of several fully connected layers. In this project, we explore two types of classification heads, the vanilla Multi-Layer Perceptron (MLP) and Gated Multi-Layer Perceptron (gMLP) [32].

Multi-Layer Perceptron (MLP) MLP is a fully connected neural network consisting of three layers: an input layer, a hidden layer, and an output layer. Intuitively, the input layer receives the data, the hidden layers process the data, and the output layer produces the final result.

The learning process in MLPs is achieved through backpropagation. During the backpropagation, the model predicts based on the input data. The output is compared to the desired output (i.e., the target), and the difference (i.e., the error) is calculated. This error is then propagated back through the network, adjusting the weights of the connections between the neurons in the process. This process guides the model to learn the underlying structure of the data.

Gated Multi-Layer Perceptron (gMLP) gMLP is an improved version of MLP. The basic principle of gMLP is the same as that of MLP. The difference is that gMLP incorporates the attention mechanism into the original structure. The main idea of gMLP is leveraging the gate to decide the block or accretion of the information. The gate mechanism makes gMLP capable of dealing with sequential data by extracting long-dependent information, as illustrated in Fig.(5).

A standard MLP layer is mathematically defined as $f_{W,b}(Z)$, where W is the weight and b is the bias term. Z is input as follows:

$$f_{W,b}(Z) = WZ + b. \quad (7)$$

The gate $s(\cdot)$ is defined in Eq.(8):

$$s(Z) = Z \odot f_{W,b}(Z). \quad (8)$$

Therefore, as stated in Fig.(5), the input Z is divided into two independent parts, Z_1 and Z_2 , in the channel dimension. Then the gate operation is illustrated in Eq.(9):

$$s(Z) = Z_1 \odot f_{W,b}(Z_2). \quad (9)$$

The standard MLPs could learn non-linear features by utilising the activation functions [33, 34, 35, 36, 37] since the CLIP already extracts the feature representation with high transferability. Therefore, training an MLP classification should be lightweight and not very complicated. By incorporating the attention mechanism into the MLP structure, gMLP allows the model to focus selectively on parts of the input sequence that are more relevant to a given task. This can be particularly beneficial for tasks involving long sequences where vital information can be located anywhere.

From another perspective, using the classification heads alone is advantageous. Specifically, separating feature extraction and classification allows for modularity and flexibility in designing machine learning models. For example, it's possible to use the exact feature extraction mechanism with different classification heads, depending on the specific task at hand. Conversely, it's also possible to experiment with different feature extraction mechanisms while keeping the same classification head.

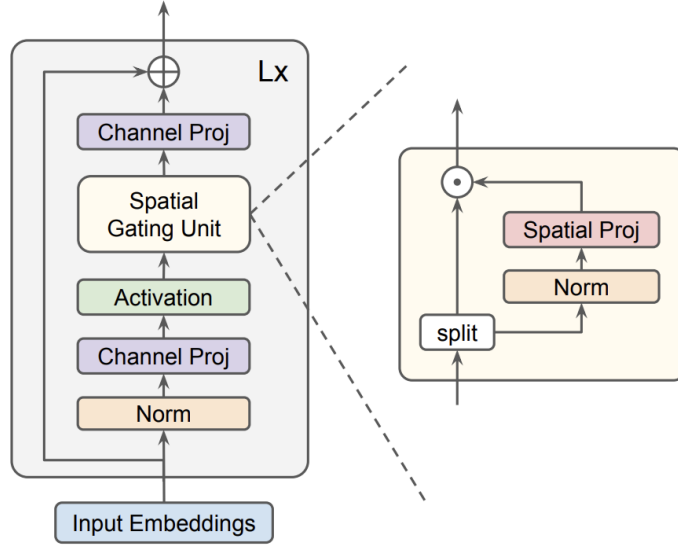


Figure 5: Strucuture of gMLP [32].

4 Experiments and Results

We aim to design our experiments into three main parts. The first part aims to choose the best pipeline of our methods by parallely comparing different techniques such as classification heads, fusion methods and loss functions. Then the second part is to fine-tune the best pipeline by modifying hyperparameters such as the layer structure of the classification heads, learning rates, epoch numbers, etc. Finally, there will be an ablation study to investigate the effect of a specific parameter.

4.1 Experiment Setups

4.1.1 Evaluation Metric

This project uses the mean F_1 score as the evaluation metric. The F_1 score is a widely accepted measure in the field of information retrieval, which calculates accuracy based on the statistical values of precision and recall. Precision refers to the proportion of true positives (i.e., tp) out of all predicted positives (i.e., $tp + fp$). In contrast, recall refers to the proportion of true positives (i.e., tp) compared to all actual positives (i.e., $tp + fn$). The F_1 score can be formulated as:

$$F_1 = 2 \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \text{ where Precision} = \frac{tp}{tp + fp}, \text{ Recall} = \frac{tp}{tp + fn}. \quad (10)$$

4.1.2 Hyperparameter Initialization

For fair comparisons, we set the training settings to be the same for all the methods. We summarized our experiment setups in Tab.(1). Also, we split the given data into 25,000 training data and 5,000 validation data and used a training batch size equivalent to the training data size. Note that in Sec.(4.3), we will tune some of these hyperparameters to achieve the best F_1 score.

Learning Rate	Activation Function	Optimizer	Dropout Rate	Epoch Number
0.001	GeLU	Adam	0.5	200

Table 1: Experiment Setups.

Classification Head	Layer Number	Validation F_1 score (%)
MLP	3	85.533 \pm 0.001
gMLP	3	83.506 \pm 0.000

Table 2: Comparison of the validation F_1 score of different classification heads for 200 epochs. We use **bold** to represent the best result.

4.2 Comparison Methods

4.2.1 Classification Head

We compare the performance of two classification heads (i.e., MLP and gMLP) and report the averaged F_1 score with standard deviations in Tab.(2). Both models have the same layer structure and were trained over 200 epochs with the same training settings. For simplicity, we use the image features only. The MLP had achieved a F_1 score of 85.533% on the validation set. In contrast, the gMLP head achieved a slightly lower F_1 score of 83.506%, representing MLP outperforms gMLP with an improvement of 2%. The result seems counterintuitive to the theoretical analysis of gMLP as it is an improved version of MLP that incorporates the attention mechanism.

To further validate the performance of MLP and gMLP, we plot the line graph for the validation loss of these two methods (see Fig.(6)). It could be found that initially, the validation loss of the gMLP is higher. This is probably because the gMLP is more complex than MLP. At around the 100th epoch, they can achieve almost the same performance. However, possibly due to the complexity of gMLP, the loss begins to increase after around 125 epochs, indicating an overfitting problem occurs for gMLP. In contrast, the loss drops faster for MLP with a more stable training process afterwards. This result suggests that a more complex model structure may not perform better in every case. **Therefore, we choose to use the MLP as the classification head for future experiments.**

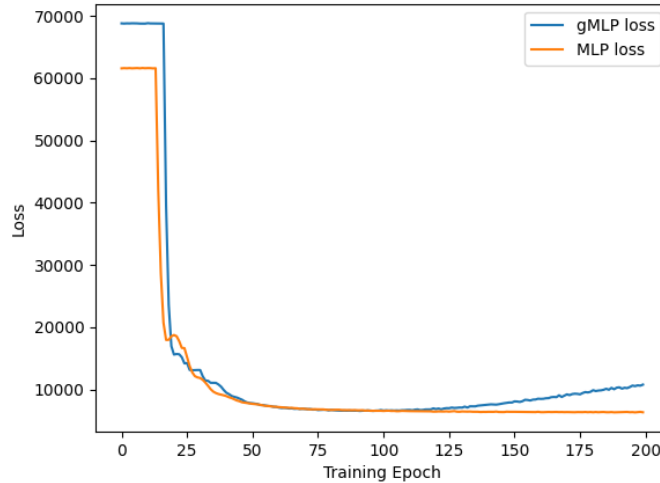


Figure 6: Validation loss for MLP and gMLP.

4.2.2 Fusion

We compare the performance of different fusion methods to the baseline models that use only image features or text features. The averaged F_1 scores with standard deviations are summarized in Tab.(3). Note that we report the last F_1 scores for a fair comparison instead of the best F_1 scores. To reflect the actual effectiveness of each fusion method, we control all the hyperparameters in the same way, which includes hyperparameters of the training procedure and the structure of the classification head. As illustrated by Tab.(3), fusion methods outperform the baseline, supporting our previous

Fusion Methods	Validation F_1 score (%)
Image Features Only	85.375 ± 0.051
Text Features Only	82.493 ± 0.000
Sum Fusion	86.338 ± 0.005
Concatenation Fusion	85.779 ± 0.000
Mixed Fusion	86.120 ± 0.011

Table 3: Comparison of the validation F_1 score of different fusion methods. We use **bold** to represent the best result.

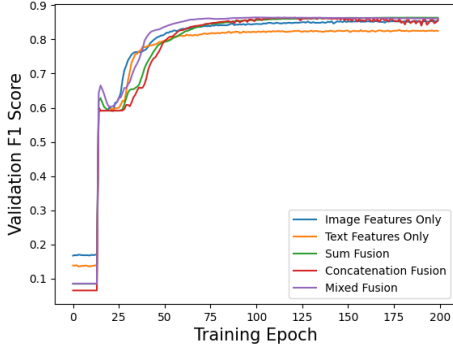


Figure 7: An overall look at validation F_1 score of different fusion methods

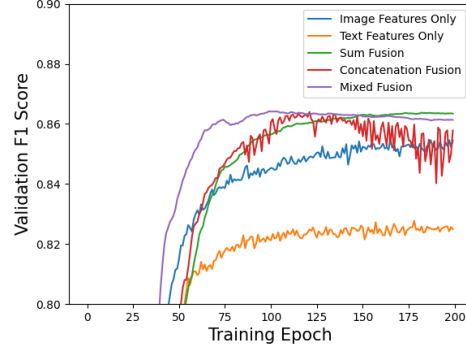


Figure 8: A closer look at validation F_1 score of different fusion methods

statements that fusion leads to better performance in multi-modality tasks. Among all the fusion methods, sum fusion achieves the best performance with a low variance, followed by mixed fusion.

Theoretically, mixed fusion should perform best since it combines concatenation and sum fusion. Empirically, by plotting the change in validation F_1 scores against training epochs (see Fig.(7) and Fig.(8)), we find that mixed fusion outperforms other methods in the first 100 epochs, but begins to decrease afterwards. This is probably attributed to the fluctuations in concatenation fusion (i.e., as part of the mixed fusion), where it appears to be unstable after around the 150th epoch. Sum fusion, on the other hand, still has a trend to further improve the validation F_1 score after the 200th epoch. **Therefore, we choose to apply sum fusion for future experiments** since we believe it will further boost the performance if we carefully tune the training epochs and other hyperparameters.

4.2.3 Loss Functions

Loss Functions	Validation F_1 score (%)
Binary Cross-Entropy	86.443 ± 0.005
Focal Loss	86.432 ± 0.005
Asymmetric Loss	85.051 ± 0.005

Table 4: Comparison of the validation F_1 score of different loss fuction. We use **bold** to represent the best result.

We compared the performance of three different loss parameter settings. The first setting sets the negative gamma = 0, positive gamma = 0, and clipping = 0. This is the equivalent of essential Binary Cross-Entropy Loss. The second setting sets the negative gamma = 3, positive gamma = 3, and clipping = 0. This is the equivalent of focal loss with a focusing parameter of 3. The third setting sets the negative gamma = 4, positive gamma = 1, and clipping = 0.05, which is asymmetric loss with different focusing factors for negative and positive instances and a probability shift of 0.05. **We increased the number of epochs to 300 to allow each loss function to converge sufficiently and**

chose the best-performing loss function, binary cross-entropy, as the loss function for future experiments.

4.3 Hyperparameter Analysis

Name	Initialization	Range for Tuning
Layer Structure	[768, 2048, 18]	[768, 2048, 18], [768, 2048, 512, 18], [768, 2048, 1024, 512, 18]
Training Batch Size	25000	25000, 10000, 5000
Learning Rate	0.001	0.001, 0.01, 0.1
Epoch Number	200	100, 200, 300, 400
Activation Function	GeLU	GeLU, ReLU, Leaky ReLU
Dropout Rate	0.5	0.2, 0.5, 0.6, 0.8

Table 5: Hyperparameter Analysis.

The greedy approach illustrates the pipeline of model tuning in Tab.(5). We first designed the layer structure of the classification head. After that, we simultaneously designed the training batch size and learning rate, followed by the choice of epoch numbers. Then is the selection of activations. Finally, we test the effect of different dropout rates to avoid overfitting.

4.3.1 Layer Structure

Layer Structure	Validation F_1 score (%)
3 layers	86.295 ± 0.000
4 layers	86.526 ± 0.000
5 layers	86.258 ± 0.000

Table 6: Comparison of the validation F_1 score of different layer structures. We use **bold** to represent the best result.

We test how different layer structures of the classification head (i.e., MLP) could affect the final result. Specifically, we use a 3-layer MLP with the layer structure of [768, 2048, 18], a 4-layer MLP with the layer structure of [768, 2048, 512, 18], and a 5-layer MLP with the layer structure of [768, 2048, 1024, 512, 18]. We report the averaged F_1 score with the standard deviation in Tab.(6). Although different layer structures have little effect on the final performance, the 4-layer MLP outperforms others with a 0 variance. Also, it is reasonable to assume that increasing or decreasing the layer numbers will not boost the performance further, as the 4-layer MLP is better than both the 3-layer MLP and the 5-layer MLP. **Therefore, we decided to use the 4-layer MLP as the classification head for future experiments.**

4.3.2 Batch Size & Learning Rate

The batch size and learning rate significantly determine the efficacy of gradient descent-based optimizers. A small learning rate may prolong the convergence of the training process, while a large one can accelerate it but at the risk of causing the loss function to diverge. In contrast, the batch size dictates the volume of data fed into the system. Notably, there is a correlation between the learning rate and batch size. Consequently, finding the optimal combination of batch size and learning rate through experimentation could potentially improve model performance.

There are three candidate values for both batch size and learning rate: [25000, 10000, 5000] and [0.001, 0.01, 0.1], respectively, creating nine distinct combinations. The mean F_1 score, along with the standard deviation, is reported in Tab.(7), and the validation loss for each combination is visually depicted in Fig.(9). From the perspective of validation loss, most combinations demonstrate a consistent decrease during the training. However, it is observed that with a relatively high learning rate (e.g., 0.1), the training process can abruptly diverge, leading to an undefined validation F_1 score (i.e., when the batch size equals 5000 and the learning rate is 0.1, the validation F_1 score becomes

Batch Size & Learning Rate	Validation F_1 score (%)
25000, 0.001	86.531 \pm 0.001
25000, 0.01	80.136 \pm 0.003
25000, 0.1	59.163 \pm 0.001
10000, 0.001	85.890 \pm 0.000
10000, 0.01	80.292 \pm 0.002
10000, 0.1	59.163 \pm 0.000
5000, 0.001	84.721 \pm 0.003
5000, 0.01	80.160 \pm 0.002
5000, 0.1	0.000 \pm 0.000

Table 7: Comparison of the validation F_1 score of different combinations of batch size and learning rate. We use **bold** to represent the best result.

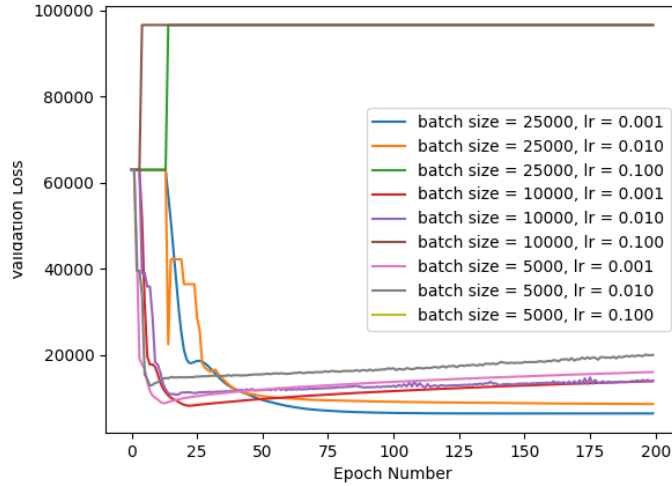


Figure 9: Validation loss with different batch sizes and learning rates.

0). We also discovered that a large batch size can lead to a stable decrease in validation loss. Among all the combinations, the model performs best when the batch size = 25,000 and the learning rate = 0.001. **As a result, we decided to use the best-performing combination in the future.**

4.3.3 Epoch Number

Epoch Number	Validation F_1 score (%)
100	84.586 \pm 0.000
200	86.496 \pm 0.000
300	86.554 \pm 0.001
400	85.770 \pm 0.000

Table 8: Comparison of the validation F_1 score of different training epochs. We use **bold** to represent the best result.

We are also curious about how different epoch numbers affect the model’s performance. Therefore, we test the validation F_1 score under 4 different epoch numbers ranging from 100 to 400 (see Tab.(8)). When the epoch size is 100, the validation F_1 is the lowest. This is probably due to the underfitting of the model, as the learning process may not fully converge. When we increase the epoch number to 300, the validation performance is further improved by approximately 2%. If we

further increase the epoch number, an overfitting problem occurs as the validation F_1 score drops. **As a result, 300 might be a suitable choice for the epoch number**, where the algorithm can fully converge without raising an overfitting issue.

4.3.4 Activation Functions

Activation Function	Validation F_1 score (%)
GeLU	86.554 \pm 0.001
ReLU	86.050 \pm 0.000
Leaky ReLU	86.267 \pm 0.000

Table 9: Comparison of the validation F_1 score of different activation functions. We use **bold** to represent the best result.

Activation functions are important in the design of deep neural networks. We investigate the effect of activation functions on the classification head by testing the model performance under GeLU, ReLU, and Leaky ReLU, respectively. These three activation functions have some fundamental differences. Specifically, ReLU is the simplest and most widely used activation function. While it's computationally efficient and helps with the vanishing gradient problem, it suffers from a problem known as "dying ReLU," where neurons can sometimes get stuck in the state where they output zero for all inputs. Leaky ReLU is a variant of ReLU designed to fix the "dying ReLU" problem. Instead of outputting zero for negative inputs, it assigns a small negative slope. This way, the neuron can still learn and adjust its weights during backpropagation, even if the initial output for a given input is negative. GeLU is more complex than the previous two. It is a smooth, differentiable function approximating a Gaussian distribution's cumulative distribution function. GeLU has been shown to perform better than ReLU in some cases, but it is more computationally expensive. From our experiment results (see Tab.(9)), GeLU outperforms the others, but the gap between GeLU and Leaky ReLU is small. **Therefore, we may try both GeLU and Leaky ReLU in future experiments to see which one performs better.**

4.3.5 Dropout Rate

The dropout rate can effectively control the model's ability to generalize. If the dropout rate is too low, the neural network might have too many parameters, which could cause overfitting. In this experiment section, we are interested in determining the dropout rate parameter for our development data to perform best and adapt the same dropout rate in the following experiments and the best model. We have four candidate values for the dropout rate: [0.2, 0.5, 0.6, 0.8], representing the probability of the node being dropped after activation.

From the perspective of training loss, as shown in Fig.(10), all dropout rate values decreased training loss with more epochs, each resulting in about 50000 in total training loss after 50 epochs. We then evaluate the model on development data using the F_1 score to investigate which dropout rate parameter will produce the best performance while also experimenting on two different activation functions from the previous experiments: GeLU and Leaky ReLU.

Dropout Rate	Validation F_1 score (%)
0.2, GeLU	86.297 \pm 0.001
0.5, GeLU	86.554 \pm 0.000
0.6, GeLU	86.606 \pm 0.000
0.8, GeLU	86.202 \pm 0.000
0.2, Leaky ReLU	85.927 \pm 0.000
0.5, Leaky ReLU	86.267 \pm 0.001
0.6, Leaky ReLU	86.349 \pm 0.000
0.8, Leaky ReLU	86.516 \pm 0.001

Table 10: Comparison of the validation F_1 score of different dropout rate. We use **bold** to represent the best result.

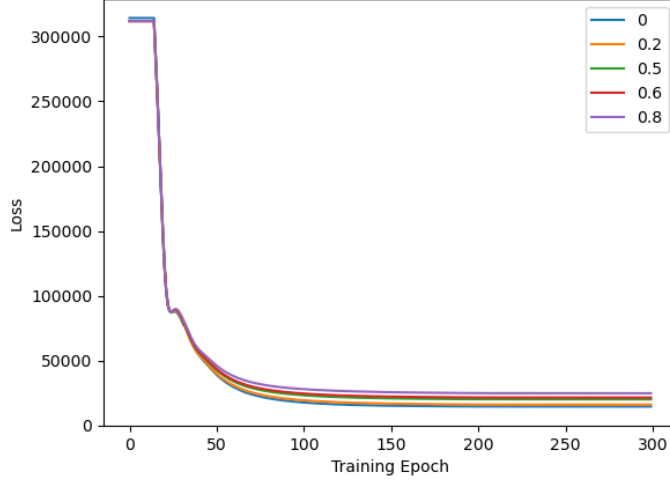


Figure 10: Training loss with different dropout rates on GeLU activation.

After comparing the model performance on development data from Tab.(10), we noticed the highest validation performance when the dropout rate = 0.6 and using the GeLU activation function. **Therefore, in all subsequent experiments and the best model architecture, we set the dropout rate to 0.6 and use GeLU as the activation function.**

4.4 Ablation Study

4.4.1 Dropout Rate

We can see in the dropout experiments that the difference between keeping all nodes (i.e., dropout rate = 0) and dropping out a portion of nodes (i.e., dropout rate = 0.6) is considerable, with the dropout rate set to 0.6, and the validation F_1 score (%) is 86.606. When the dropout rate is set to 0, the validation F_1 score (%) is 85.955, resulting in a 0.65 difference (see Tab.(11)). This effect is even more prominent when using the Leaky ReLU activation function.

This shows dropout does help with the overfitting problem. By leaving out a portion of the nodes in a layer, the model is less likely to rely on specific inputs and is forced to learn more robust features, increasing its generalization ability. **With this observation, dropout is kept in our best model, set to the best value found in parameter analysis with dropout rate = 0.6.**

Dropout Rate	Validation F_1 score (%)
0, GeLU	85.955 \pm 0.001
0.6, GeLU	86.606 \pm 0.000
0, Leaky ReLU	85.269 \pm 0.000
0.6, Leaky ReLU	86.349 \pm 0.001

Table 11: Comparison of the validation F_1 score of dropout and no dropout. We use **bold** to represent the best result.

4.5 Best Pipeline Results and Justification

Best Pipeline To summarise the experimental result, the model obtains the best performance by leveraging 4 layers of MLP with GeLU activation function and 0.6 dropout rate to prevent overfitting. Moreover, the sum fusion is trained with the BCE loss. The implementation details of the hyperparameters are summarised in Tab.(12). The model is trained for 300 epochs with a learning

rate of $1e-4$ and batch size of 30000 (i.e., train with no validation data). The performance of the best pipeline achieves a 90.114% public F_1 score on the Kaggle competition leaderboard. Since we only fine-tune with the classification head, our model is very lightweight; it could be trained within 4 minutes, and the model size is less than 25 MB. The device and software details are summarised in Tab.(??).

Justification Previous studies have shown that gMLP performs better than standard MLP. However, our experiments found that the extra gMLP layer will lead to the overfitting problem after 125 training epochs. This result infers that the feature extracted by the CLIP had already learned a good representation. The more complicated classification head may lead to a performance drop. Therefore, as for fine-tuning downstream tasks, the model complexity should be controlled. Moreover, we explored different fusion methods. Although the sum fusion by training two extra MLPs breaks the end-to-end training framework, it achieves a better result. Therefore, the conclusion is that fusion by a learning model is necessary, but the fusion part could not be complex.

Parameters	Value
Classification Head	MLP
Activation Function	GeLU
Dropout Rate	0.6
Fusion Method	Sum Fusion
Loss Function	BCE
Epochs	300
Learning Rate	0.001
Batch Size	30000

Table 12: Parameters details of the best pipeline.

5 Conclusions

In conclusion, we implemented the state-of-the-art (SOTA) techniques by utilising CLIP trained on a large dataset as the feature extractor for image and text modalities. Then, we fine-tuned the classification head and explored different classification heads, such as MLP and gMLP, with the composition of various feature fusion methods and loss functions. Finally, we obtained the model with the best performance of 90.114% F_1 score in the Kaggle completion leaderboard due to the impressive transferability of CLIP and the fusion method. It is well worth noting that our approach is lightweight for both time and space complexity. The model could be trained in RTX3080 GPU in 4 minutes. The total model size is no more than 25 MB.

Although our model achieves the top 5 performance in the Kaggle leaderboard, some limitations remain. The first improvement could be the development of a unified learning architecture. The current training approach is not in the end-to-end schema, which results in a more complex hyperparameter tuning and deployment for real-world applications. Moreover, two-step learning can not be trained parallelly. The second training stage requires the output for the first step. Therefore, a possible improvement could be designing a model with two input sources simultaneously into a unified structure. However, due to the time limitation of the Kaggle completion, more experiments for the new model structures are required.

Moreover, we find the distribution of the data set is skewed. Further data preprocessing may lead to higher performance. To be specific, the majority of the distribution of label class is class 1. Therefore, it results in a shortcut solution for the model to predict class 1 with a low confidence level. The possible solution for addressing this is oversampling with the fewer classes or downsampling with the majority classes. Since our model consumes low training time, it is possible to train multiple models with different subsets of training data and ensemble the models for a more robust prediction. Finally, the shifting threshold may also be helpful by reducing the threshold value for the fewer classes. However, these methods introduce extra hyperparameters and training times, which is a trade-off between the performance and the computational cost.

Overall, the development of multimodal learning provides a solution to deal with the natural world by learning the comprehensive feature dependency across different data sources. The conventional research directions could be summarised as studying different interactions between different modalities, such as calculating the cosine similarity and consistency [6], or using complex fusion method [3]. Therefore, there are a few suggestions and insights:

Dynamic Fusion Methods Traditional fusion methods, such as decision fusion and feature fusion, often lack flexibility. Recent advancements suggest that dynamic fusion methods, which allow the model to combine different modalities based on the input data adaptively, could provide more accurate and robust results [38].

Domain Adaptation and Transfer Learning With the increasing complexity of multimodal tasks, leveraging knowledge from pre-existing models or similar domains is vital. Domain adaptation and transfer learning could help reduce the demand for labelled data and improve the models' generalisation ability [39].

Robustness Multimodal models should be robust to the absence or degradation of one or more modalities. This requires further research into methods for improving the resilience of such models to missing or noisy data.

Scalability and Efficiency As the data volume and model complexity increase, the demand for computational resources also increases. Hence, it is essential to explore methods that can improve the scalability and efficiency of multimodal learning, such as lightweight model architectures and efficient training algorithms.

References

- [1] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. Van Den Hengel, “Visual question answering: A survey of methods and datasets,” *Computer Vision and Image Understanding*, vol. 163, pp. 21–40, 2017.
- [2] J. Mun, M. Cho, and B. Han, “Local-global video-text interactions for temporal grounding,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [3] W. Kim, B. Son, and I. Kim, “Vilt: Vision-and-language transformer without convolution or region supervision,” in *International Conference on Machine Learning*, 2021.
- [4] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, “Vse++: Improving visual-semantic embeddings with hard negatives,” 2018.
- [5] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, “Stacked cross attention for image-text matching,” 2018.
- [6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” 2021.
- [7] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, “Film: Visual reasoning with a general conditioning layer,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.
- [8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, pp. 84 – 90, 2012.
- [9] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *CoRR*, vol. abs/1409.1556, 2014.
- [10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2014.
- [11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.
- [12] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” *ArXiv*, vol. abs/1704.04861, 2017.
- [13] M. Tan and Q. V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” *ArXiv*, vol. abs/1905.11946, 2019.
- [14] L. R. Medsker and L. Jain, “Recurrent neural networks,” *Design and Applications*, vol. 5, pp. 64–67, 2001.
- [15] A. Graves and A. Graves, “Long short-term memory,” *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.
- [16] S. Cornegruta, R. Bakewell, S. Withey, and G. Montana, “Modelling radiological language with bidirectional long short-term memory networks,” *arXiv preprint arXiv:1609.08409*, 2016.
- [17] R. Dey and F. M. Salem, “Gate-variants of gated recurrent unit (gru) neural networks,” in *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pp. 1597–1600, IEEE, 2017.
- [18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *ArXiv*, vol. abs/1810.04805, 2019.
- [20] A. Radford and K. Narasimhan, “Improving language understanding by generative pre-training,” 2018.
- [21] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” 2019.

- [22] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. J. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” *ArXiv*, vol. abs/2005.14165, 2020.
- [23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” *ArXiv*, vol. abs/2010.11929, 2020.
- [24] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, 2021.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015.
- [26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.
- [27] E. Ben-Baruch, T. Ridnik, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, “Asymmetric loss for multi-label classification,” *arXiv preprint arXiv:2009.14119*, 2020.
- [28] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, “Snapshot ensembles: Train 1, get m for free,” *arXiv preprint arXiv:1704.00109*, 2017.
- [29] P. Izmailov, D. Podoprikin, T. Garipov, D. Vetrov, and A. G. Wilson, “Averaging weights leads to wider optima and better generalization,” *arXiv preprint arXiv:1803.05407*, 2018.
- [30] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” *Advances in neural information processing systems*, vol. 30, 2017.
- [31] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, “Bootstrap your own latent-a new approach to self-supervised learning,” *Advances in neural information processing systems*, vol. 33, pp. 21271–21284, 2020.
- [32] H. Liu, Z. Dai, D. So, and Q. V. Le, “Pay attention to mlps,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 9204–9215, 2021.
- [33] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.
- [34] B. Xu, N. Wang, T. Chen, and M. Li, “Empirical evaluation of rectified activations in convolutional network,” 2015.
- [35] P. Ramachandran, B. Zoph, and Q. V. Le, “Searching for activation functions,” *arXiv preprint arXiv:1710.05941*, 2017.
- [36] D. Hendrycks and K. Gimpel, “Gaussian error linear units (gelus),” *arXiv preprint arXiv:1606.08415*, 2016.
- [37] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.
- [38] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, “Multi-modal transformer for unaligned multimodal language sequences,” 2019.
- [39] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, “A survey on deep transfer learning,” 2018.