Assignment 2

Jielin Feng:480019484, Yifeng Gao:500083754, Zedong Wen:500334863

May 22, 2022

1 Introduction

1.1 Aim

This study aims to complete a multi-label and multi-class classification. The training dataset in this study contains 30,000 images of different sizes. Each image corresponds to one or more classes and a caption that matches the image. There are 18 different classes in this dataset. Through this study, firstly, it is aiming to experiment with single modal models. by using Efficient net and Clip(ViT) for image, and Bert and Clip(Transformer) for text respectively. Secondly, it is aiming in finding the proper Multimodal Fusion methods to process and combine the data of the two modalities, image, and text, to obtain richer feature information and make the model more effective. Thirdly, it is expected to design the structures of the models and tune the models with different parameters and find suitable parameters. Fourthly, the performance of every single modal model and multimodal fusion methods will be compared and discussed. It is worth mentioning that the main performance metric of this study is the F1 score. Finally, a deep analysis and justification of the multimodal model will be evaluated, and further future work will be discussed.

1.2 Why is the study important

In today's society, people live in a multimedia society composed of a large number of different modal content. The content of different modalities includes text, image, audio, video, 3D, etc. In different tasks, different modal content with high correlation will likely appear at the same time. In this case, using multimodal retrieval will perform better. There are various applications related to multimodal. Firstly, "Multimodal Representation", which learns better feature representations by taking advantage of the complementarity between multiple modalities and eliminating the redundancy between modalities. Secondly, "Translation", it can convert the information of one modality into another modality. Such as speech synthesis as specific applications, which automatically convert the input text into a piece of speech. Thirdly, "Alignment", can find correspondences between different modal information in the same instance. Practical applications such as automatically matching voice and subtitles to movies. fourthly, "Multimodal Fusion", which can combine information from multiple modalities and perform target classification or regression. Practical applications such as synthesizing the voice and video information of the same instance to identify the target. In addition, there is Co-learning, which refers to the use of one resource-rich modality to assist another resource-poor modality in learning. The most typical representative is transfer learning. In general, multimodal retrieval can use the content of multiple modalities to assist in processing tasks and enable models to perform better to a large extent. In addition, in today's society, with the development of media, more and more multimodal forms are produced. Therefore, using multimodal content to process tasks or construct models may become the mainstream of future society. Which is the reason why this study is very important to us.

1.3 General introduction and Motivation

The main task of this study is to classify multi-label data containing two modal content of image and text. We will first build separate models for image and text. For image, Clip(ViT) and Efficientnet will be applied. For text, Clip(Transformer) and Bert will be applied. It is expected to find the best model performance in the single modality model and further analysis. Secondly, we will try to achieve the multimodal fusion of image and text modalities from different angles according to the fusion methods of "Features Fusion", "Decisions Fusion" and "Mixed Fusion". Finally, fusion methods will be compared with models that only use the image or text modalities. After the results are obtained, more analysis will be performed on the results to discuss whether the multimodal fusion method performs better for our dataset than the single modal model. If so, what type of fusion method can play the biggest role in improvement. Our motivation comes from first, in the case of including both caption and image modal content, we are curious about the effect of using a single modal content on the overall classification performance, and try a variety of different single-modal models in this study. Secondly, based on the particularity of multi-modality and multi-label in this data, our curiosity about whether using two modalities at the same time will improve the overall performance led us to try a variety of multimodal fusion methods.

2 Related works

In this study, we would firstly discovered single-modal models that handles image and text separately. For the image processing model, Tan and Le has mentioned in "EffificientNet: Rethinking Model Scaling for Convolutional Neural Networks" that EfficientNet which use the compound scaling method has improved the accuracy on ImageNet, with only uses the most original baseline model EfficientNet-B0 The accuracy is as high as 77.3. [1] Dosovitsky and Beyer et al found that ViT (Vision Transformer) is currently the best model for image classification in year 2020, even surpassing the best CNN network. ViT outperforms the best Resnet on all public datasets. On larger datasets, ViT can exert its advantages more.[2]

For the text model, Vaswani,Shazeer et val proposed the transformer in which is a model used in natural language processing, in year 2017. Compared with the slow training RNN, the transformer uses the self-attention mechanism to achieve fast operation. And the transformer can increase a very deep depth and improve the accuracy of the model.[3] Devlin, Chang et val proposed that Bert showed good results in 11 natural language processing tasks. The GLUE benchmark was advanced to 80.4 percentage, and the MultiNLI accuracy was improved to 86.7 percentage. It is gradually used in the multimodal field.[4]

Even with the help of multi-modal fusion method, the performance of each single modality will still have a decisive impact on the performance after multi-modal fusion. Which is also used to decide the baseline model required for modal fusion.

We would secondly focused in Multimodal fusion. Castellano, Kessous et al. used Bayesian classifier in 2008 to compare feature fusion and decision fusion, and gave the reasons and analysis why feature fusion is better than decision fusion.[5] Lan, Bao and Yu et al. combined feature fusion and decision fusion to propose a hybrid fusion method in 2014. It takes into account both the correlation of features and the fusion method when modal fusion of data with very low correlation is carried out.[6]

In 2019, Sahu and Vechtomova introduced how to change the previous blunt fusion method and find a better and more natural fusion method. Therefore, the work adopts the fusion method of Autoencoder approach. It passes all the modal fc through the encoder, then uses the decode to restore the features, and finally calculates the loss between the features. In this way, the output of the encoder is optimized and used as the extracted fusion feature for further training.[7] At the same time, the article proposes how to use GAN network for automatic search fusion, and further optimizes the fusion method of Autoencoder. The collocation of GAN achieved neural network learning for the first time in "Mfas: multimodal fusion architecture search. IEEE." in 2019, and achieved good results. [8]

In "Deep Multimodal Multilinear Fusion with High-order Polynomial Pooling" published in 2019 by Ming, Jiajia et al., consider that the full expressiveness and restricted interaction order of multilinear fusion cannot be released due to simple linear feature fusion. More importantly, simply fusing features simultaneously ignores complex local interrelationships. So upgrade to a Polynomial Tensor Pooling (PTP) block that integrates multimodal features by considering higher-order moments.[9]

In 2017, Zadeah, Chen et al. proved that TFN is a typical multimodal network that fuses features through matrix operations. This work improves the problem of feature correlation between different modes by changing the linear feature connection to the matrix. But the shortcomings are also very obvious. Because the generated matrix is too large, the training is very difficult, which will be mentioned in the following chapters.[10] In order to optimize the TFN network, in 2018, Liu and Shen et al. used LMF to perform low-rank matrix decomposition on the weights, and changed the process of TFN first tensor outer product and then FC into each mode first linearly transformed and then more Dimensional dot product, which can be viewed as the sum of the results of multiple low-rank vectors, reducing the number of parameters in the model.[11]

The traditional multimodal fusion methods are divided into early fusion, mid-term fusion, late fusion and other fusion methods. In our work, with the introduction of networks such as Transformer and VIT, it is difficult to define the classification method of early, medium fusion. Therefore, we conduct experiments based on feature, decision, and mixed fusion methods. At the same time, we also propose a new feature fusion method based on the early feature fusion model. The specific content will be introduced in the Method in the following part of this paper.

3 Techniques

3.1 principle of models

3.1.1 Image Model

Clip(ViT)

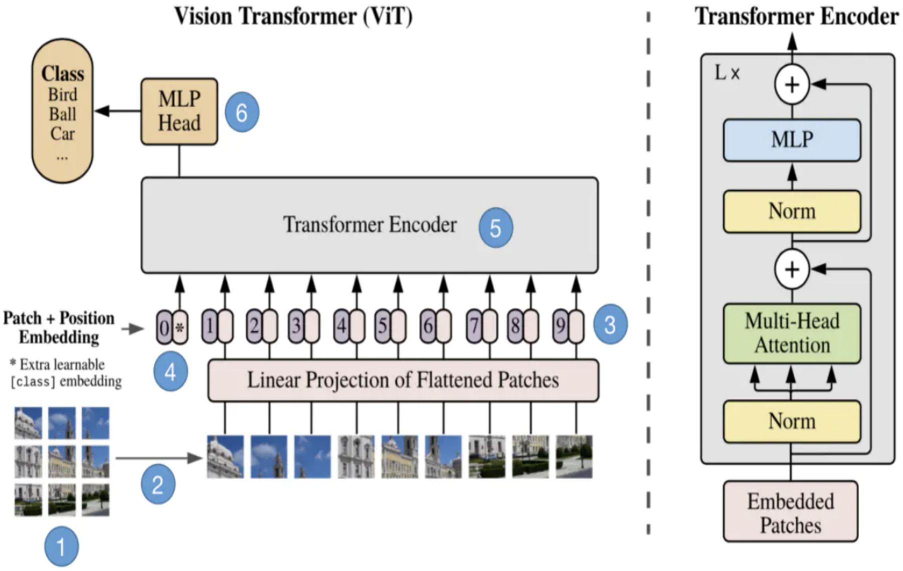

Figure 1: ViT

As shown in Figure1, firstly, we tried the ViT (Visual Transform) model in Clip. The working principle of ViT is as follows: First, the input image is divided into patches of equal size. Secondly, flatten the divided patches, because the dimension after flattening is easy to be too large, so it needs to be linearly mapped to a lower-dimensional space. The third step is to add position embedding to the patched to better track the position of the flattened patches. In the fourth step, add the class token to obtain the encoded features of the input image. For the fifth step, input to the transformer encoder to get the features of the image. Finally, the features are fed into the MLP architecture to get the final classification effect.

EfficientNet

The modal's performance can be enhanced by increase the network parameters under the condition that the data is sufficient and will not cause overfitting problems. There are three ways to increase network parameters. The first way is to add the number of layers of the network, that is, to deepen. The second way is to increase the number of channels in the network, that is, to widen. The third way is to increase the resolution of the network, which is the network input size. These three ways of increasing network parameters are related and dependent on each other. Therefore, EfficientNet introduced the compound scaling method, which was improved on the basis of the original only amplifying one of the three methods alone. EfficientNet uniformly scales network depth, width and resolution. This allows EfficientNet to achieve better performance than upscaling alone.

3.1.2 Text Model

Clip(Transformer)

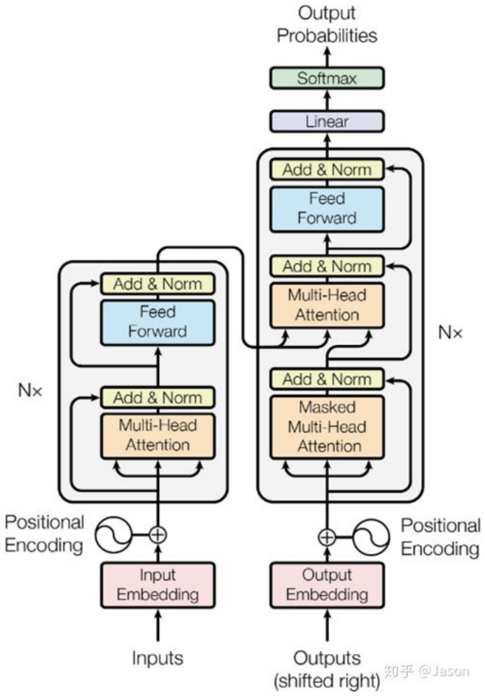

Figure 2: Transformer

As shown in Figure2,Transformer consists of two parts, an Encoders and a Decoders. Each Encoder is composed of six Encoders, and each Decoders is also composed of six Encoders. For each Encoder, their structure is the same, but they don't share the weight. The input of each Encoder will first pass through a self-attention layer, which will help Endcoder check other words in the input sequence during the process of encoding words. The output of Self-attention will be fed into a fully connected feed-forward neural network. The number of feedforward neural network parameters of each encoder is the same, but their functions are independent. Each Decoder also has such a hierarchical structure, but there is an Attention layer in between, which helps the Decoder focus on the word corresponding to the input sentence.

Bert

BERT is a transformer-based bidirectional encoding representation. It is a pre-training model. The two tasks of the model training are to predict the masked words in the sentence and to determine whether the two input sentences are upper and lower sentences. After the pre-trained BERT model is added to the corresponding network according to the specific task, the downstream tasks of NLP can be completed, such as text classification, machine translation, etc. Although BERT is based on the transformer, it only uses the encoder part of the transformer, and its overall framework is formed by stacking the encoders of multiple layers of transformers. The encoder of each layer is composed of a layer of muti-head-attention and a layer of feed-forword. The large model has 24 layers, each layer has 16 attention, and the small model has 12 layers, each layer has 12 attention. The main role of each attention is to re-encode the target word through the relevance of the target word to all words in the sentence. Therefore, the calculation of each attention includes three steps: calculating the correlation between words, normalizing the correlation, and obtaining the encoding of the target word through the weighted summation of the correlation and the encoding of all words. When calculating the correlation between words through attention, firstly, the input sequence vector (512*768) is linearly transformed through three weight matrices, and three new sequence vectors of query, key and value are generated respectively. The query vector is multiplied with the key vector of all words in the sequence to obtain the correlation between words, and then this correlation is normalized by softmax, and the normalized weight is summed with the value weight, Get a new encoding for each word.

3.1.3 DNN

Feature-Decision MLP

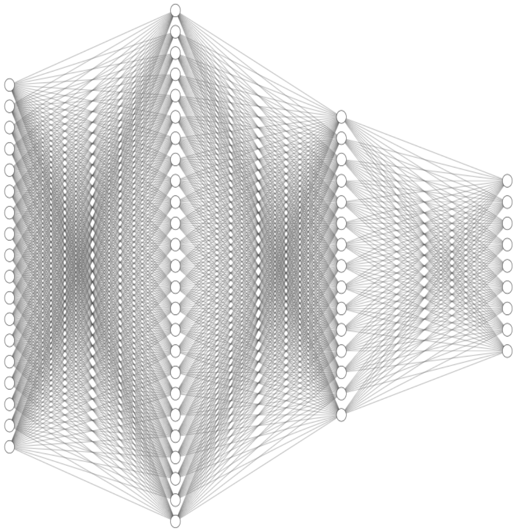

Figure 3: Feature-Decision MLP

As shown in Figure3, "Feature-Decision MLP" network contains four layers, in which the nodes of the input layer are the corresponding number of original features size 768, and the activation function is "gelu". The nodes of the two hidden layers are increased first and then decreased, with 1024 and 512 nodes respectively. The activation methods are also "gelu". Dropout regularization is added in hidden layers with a dropout rate of 0.5 to effectively prevent overfitting. The last layer is the output layer, the nodes of the output layer are 18, corresponding to 18 classes, and the activation function of the output layer is "sigmoid", which regulates the output to the range of [0, 1]. When the output result is ¿ 0.5, it is determined to belong to specific classes and vice versa.

Decision-Decision MLP

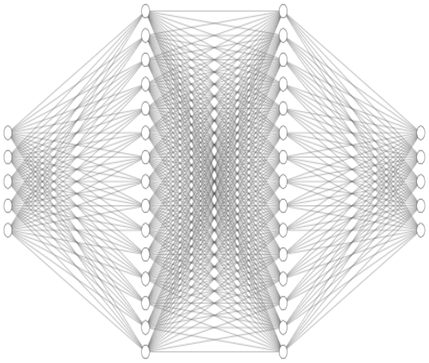

Figure 4: Decision-Decision MLP

As shown in Figure4, "Decision-Decision MLP" architecture also includes four layers, the input layer contains 18 nodes, and the activation function is "gelu". The two hidden layers contain 512 nodes respectively, and the activation function is also "gelu". The final output layer contains 18 nodes corresponding to 18 categories, and the activation function of the final output layer is still set to "sigmoid", so that the output is still range in [0, 1] and the final probability result is obtained.

HLO MLP

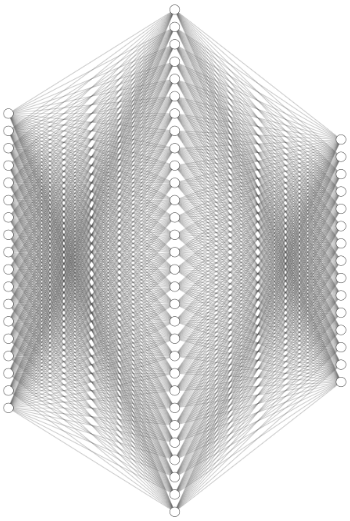

Figure 5: HLO MLP

As shown in Figure5, "HLO MLP" architecture is used specifically for HLO approch in feature fusion. This architecture original contains4 layers. The first input layers contains 768 nodes, and the activation function is "gelu". The hidden layers contain

512 nodes. In HLO, we directly get the output from the last hidden layer instead of going through output layer. Thus the final shape of output after going through HLO structure is (30000,18).

3.2 Principle of method

3.2.1 Features Fusion

Flatten concatenation

Flatten concatenation is a features model fusion approach. The principle and implementation method are as follows: the features of the image and caption are extracted as a matrix of (30000, x). Then flatten the extracted text and image information horizontally to form a matrix with a final size of (30000, 2x). At this time, the feature matrix has the same amount of feature information of the image and caption at the same time. After using the flatten concatenation approach, it is expected to have more accurate feature information than single text or image features. The 30000 datasets is then divided into training and validation sets, and the final features are input into "Feature-Decision MLP" to get the decision result. The final output shape is (30000, 18).

PCA

PCA method is also an approach to features fusion, which can be regarded as an improved version of Flatten concatenation. Flatten concatenation only initially tiled the features of the image and caption and merged them horizontally, which may lack a certain fusion effect. PCA combines and extracted information based on flatten concatenation and then removes redundant information, which is expected to be superior to Flatten. The principle of PCA is to remap n-dimensional features to k-dimensions and generate new features in k-dimensions. The advantage of PCA is that firstly, it can reduce the feature dimension after flattening, thereby reducing the computational cost. Second, it can effectively remove noise. The way to accomplish PCA is that firstly, extracted features information of image and caption individually, and flatten them horizontally. Then use PCA on the matrix of (30000, 2x) dimension to extract new fusion features. The new features are then put into the "Feature-Decision MLP" architecture, to get the decision result. The final output shape is (30000, 18).

AutoEncoder

AutoEncoder approach is also a features fusion method. Just like PCA, AutoEncoder can also be regarded as an improved and enhanced version of Flatten concatenation. AutoEncoder can better combine features based on flattened concatenation. AutoEncoder can be split into two steps, Encoder, and Decoder. The principle is to first extract features of image and text individually and flatten them horizontally to form a matrix with shape (30000,2x), then compress the (30000, 2x) flatten feature into a low-dimensional vector during the Encoder to achieve nonlinear dimension reduction. The Decoder will restore the low-dimensional vector back, and the reduced-dimensional feature decoder will convert it back to the high-dimensional feature. The further implementation of the AutoEncoder approach is the same as PCA. Put the new features after Autoencoder as input into the "Feature-Decision MLP" architecture, to get the decision result. The final output shape is (30000, 18).

TFN 2D

The principle of TFN 2D is to multiply the extracted image feature (30000, x) and caption feature (30000, x) by matrix. After matrix multiplication, a fusion feature of shape (30000, 1, x, x) will be obtained. The "1" represented the added channel. Then downsample the fused features of shape (30000, 1, x, x) by max pooling to form a new matrix (30000, 1, y, y). Finally, flatten the matrix after max pooling, and finally form the final features after fusion with shape (30000, 1*y*y), which can be represented by (30000, z). Put the features after fusion of shape (30000, z) into "Feature-Decision MLP", to get the decision result. The final output shape is (30000, 18).

HLO(Hidden Layer Output) concatenation

The principle of HLO concatenation, which is Hidden Layer Output concatenation, is to first extract the (30000, x) feature information of the image and caption respectively. Then put the (30000, x) feature information of the image and caption into "HLO MLP" respectively, and take the output of the last hidden layer which has a shape of (30000, 512). After getting the (30000, 512) hidden layer output features, They are concat by adding these two matrices together and get a (30000,512) matrix. Finally put the features(30000, 512) as input into the "Feature-Decision MLP" architecture, to get the decision result. The final output shape is (30000, 18).

3.2.2 Decisions Fusion

Max Fusion

The Max fusion approach is a decisions fusion method. Different from the above features fusion, the decisions fusion's goal is to let the (30000, x) features extracted from the image and caption go through the "Feature-Decision MLP" network respectively. Firstly make the image and caption produce the resulting output of (30000, 18) respectively. After that, we fuse the output decisions of these two different modalities. Max Fusion approach will first align the output of the image and caption (30000, 18), and then compare, leaving the modal result with a larger value. For example, when the output of an image is determined to be 0.7 for a certain class, and the output of text is determined to be 0.4, then max-fusion will keep the output of the image at 0.7 and discard the output of text at 0.4. The max-fusion method does its best to preserve the maximum likelihood of each class, thereby might improve the final performance. After the maximum fusion processing of the output results of the two modalities, we put the retained probability of (30000, 18) back into the "Decision-Decision MLP" to get the final decision result. The final output shape is (30000, 18).

Sum Fusion

The sum fusion approach is also a decision fusion method. Similar to max-fusion, the Max fusion method also lets the (30000, x) features extracted from the caption and the image respectively pass through the "Feature-Decision MLP" network and produce the shape (30000, 18) output results respectively. Then the output decisions of the two modalities are fused. Unlike Max fusion, Sum fusion is less "extreme", and this fusion method tends to comprehensively consider the results of both modalities and perform a final fusion. This is achieved by adding the (30000, 18) result matrices of the two modalities, and producing the final result matrix of shape (30000, 18) by fusing the decision output of the two modalities. Finally, just like Max fusion, put the input (30000,18) to get the final decision result. The final output shape is (30000, 18).

Weighted Average Fusion

Just like Sum fusion, Weighted Average Fusion also considers the performance of both modalities, but in a more detailed way. based on Sum fusion, in Weighted Average Fusion, different modalities are weighted when fused according to the original F1 score of different modalities models result. Finally, the weighted decision fusion result goes through the ”Decision-Decision MLP”, and the final result with shape (30000, 18) is obtained.

3.2.3 Mixed Fusion

Basic Mixed Fusion

The Basic Mixed Fusion, which combines features fusion and decisions fusion, is expected to produce outperforming results. The principle is that first, the features of the image and caption are merged horizontally with a shape of (30000,2x), put the (30000,2x) features as the input of “Feature-Decision MLP” architecture, which produces the first output result of (30000, 18). Second, extract the (30000,x) features of the image individually, put the (30000,x) features as the input of the “Feature-Decision MLP” architecture, and generate the second output of (30000, 18). Third, extract the (30000,x) features of the caption individually, put the (30000,x) features as the input of the “Feature-Decision MLP” architecture, and generate the third output of (30000, 18). Finally, the above three (30000, 18) outputs are used for Sum fusion, the three results are summed, produced a final shape of (30000,18), then put into the ”Decision-Decision MLP” architecture, and get the final output.

HLO Mixed Fusion

HLO Mixed Fusion is also a combinations of features fusion and decisions fusion. The only difference between HLO Mixed Fusion and Basic Mixed Fusion is that HLO Mixed fusion firstly produced the features fusion output by using the ”HLO concatenation” method as mentioned above, then sum the output(30000,18) after applied HLO, the output(30000,18) of using only image modality and the output(30000,18) of using only text modality together to produce a final result of (30000,18) after all decisions are mused. Finally, put the (30000,18) as input into the ”Decision-Decision MLP” architecture, and get the final output.

3.3 Justify of method

It is expected that after applying the Multimodal fusion method, it will eventually produce a better performance on multi-labels and multi-classes classification tasks. In order to verify our idea and the rationality of this method, we will first observe the result of the method using only the image and the method of using only the caption, and then observe the result of the method of combining the two modalities of caption and text. We will compare the performance of the multi-modal fusion method with the performance of the single-modal model method. If the performance of the multimodal fusion is better than that of the two single-modal models, it is proved that in the classification task with multiple modal contents, it is better to use the Multimodal fusion method than any single-modal model which can improve the classification effect. The above-mentioned features fusion methods, decisions fusion methods, and mixed fusion methods in the method section will also be compared, find out which fusion method is the best method for this study, and discuss it. To ensure the performance of the experiments, we will set the random state when separating the training set and the validation set, to ensure that each set of experiments has the same training and validation set, and there will be no leakage problems. At the same time, the architecture of different types of mlps will remain the same in all experiments.

3.4 Advantage and Novelty

The Advantage and Novelty are that first, we found that it is very hard to extract image feature information in this task. Images come in different sizes and seem to be difficult to classify accurately. Basic CNN networks such as Vgg and Resnet have been tried, but the effect is not ideal, and can not classify images with good performance. Therefore, we effectively solve this problem by using Efficient net and Clip (ViT). Among them, Efficientnet is a very efficient network. It produces better performance for image content with a very small model size. Compared with Vgg, Resnet, etc., Efficient net has significant performance improvement on our image task. And Clip (ViT) contains 400 million image data, which contains a lot of information. We boldly tried this very informative model to maximize the performance of the image task. The second novelty is our use of multimodal fusion. After a lot of understanding and attempts of the existing fusion methods, we not only experimented with and verified the existing fusion methods but also innovated new multi-modal fusion methods such as ”TFN 2D”, ”HLO”, ”HLO mixed”, ”weighted Average”, etc. These fusion methods created by our own extensions even outperform some of the existing fusion methods.

4 Experiments and results

4.1 Results

Model Type	Model	F1 Score
Image	Clip(ViT)	87.76
Image	EfficientNet	73.48
Text	Clip(Transforms)	83.24
Text	Bert	83.05

Table 1: Single Modal Models Performance Table.

Fusion Type	Fusion Method	F1 Score
Baseline Model	Single Image	87.76
Baseline Model	Single Text	83.24
Features Fusion	Flatten concatenation	88.13
Features Fusion	PCA	87.74
Features Fusion	AutoEncoder	84.83
Features Fusion	TFN2D	-
Features Fusion	HLO	87.78
Decisions Fusion	MaxFusion	88.11
Decisions Fusion	SumFusion	88.65
Decisions Fusion	WeightedAvgFusion	88.54
Mixed Fusion	BasicMixedFusion	87.79
Mixed Fusion	HLOMixedFusion	88.36

Table 2: Fusion Methods Performance Table.

As can be seen from Table1, for single modal models, Clip(ViT) has the best performance of 87.76 F1 score for Image in validation set, which is much higher then Efficient Net which only has 73.48 F1 score. Clip(Transforms) has the best performance of 83.24 F1 score for text in validation set, which is slightly higher than Bert which only has 83.05. This might because of the large training datas for image and text in Clip models, which produce the overall best result for both image and text for our datasets. Thus, Clip(ViT) and Clip(Transform) is decided to be the final model to do further fusions for image and text, respectively. It is also decided to be the baseline model for single image and text as shown in Table2

As can be seen from Table2,Figure6 and Figure8 the results show that the Sum Fusion method has the best overall performance with an F1 score of 88.65. From the Baseline Models, the F1 score of the Single Image Classifier is 87.76, and the F1 of the Single Text Classifier is 83.24. All fusion methods outperform single text classifiers, but not all fusion methods outperform single image classifiers. All feature fusion methods perform poorly, except that Flatten can reach 88.13, which is slightly better than Single Image Classifier, the performance of all remaining feature fusions is similar to or even worse than Single Image classifier.(It is worth mentioning that TFN2D cannot successfully running out the result due to its extremely large parameter size and extremely slow speed.) Among them, AutoEncoder performed the worst, with only an F1 score of 84.83. Both PCA and AutoEncoder have further combined features after Flatten, and it is expected to have better performance than the Flatten method. However, the reality is not ideal. It can be seen that mixed fusion methods and decision fusion methods have better effects. In the mixed fusion methods, Basic Mix fusion and HLO Mix fusion achieved f1 scores of 87.79 and 88.36, respectively, which were better than the performance of the single image classifier itself. Among them, HLOmix fusion is improved by about 0.6 compared with pure feature fusion HLO. Among all fusion methods, decision fusion has the best performance, Max fusion, Sum fusion, and Weighted Average fusion all have significantly better performance than the single-modal model.

As can be seen from Figure7 and Figure9, First of all, for fusion methods flatten concatenation, PCA, basic mixed fusion, single image and single text, loss are still going down in the training set. However, their loss are all in upward trend after a turning point in Validation set, which indicates that these methods have relatively serious overfitting problems. Secondly, As can be shown in Figure7 that AutoEncoder approach has a weak learning ability, its loss convergence speed is slow, and there is a certain underfitting problem. In general, the fusion methods of Max fusion, weighted Average fusion and HLO mixed fusion have relatively the most stable performance, and there is no obvious over-fitting or under-fitting problem in these method. This is also consistent with our conclusion comes from F1 score that overall, all decisions fusions have relatively better and more stable results.

Features Fusion Evaluation

Features Fusion methods which include Flatten concatenation, PCA, AutoEncoder, TFN2D and HLO does not seemed to perform well in this task. this may be due to the high correlation between modalities, it is more difficult to extract features or data dimensions in this situation. Feature extraction cannot fully reflect the complementarity between modalities. And although the features extracted from the text and image modalities before fusion have the same dimensions, their distributions are quite different, so they cannot be well fused at the feature level.

Decisions Fusion Evaluation

Decisions Fusion methods which include Max Fusion, Sum Fusion and Weighted Average Fusion has the significant best overall result. all three methods achieved a F1 score that higher than 88 in validation set. The better performance of Decisions Fusion methods may be due to the fact that, first, the fusion process of Decisions Fusion methods is completely feature-independent. And the error results in different modalities are also irrelevant. This shows that the decision fusion has better robustness. Secondly, in this task, the correlation between the two modalities is not large, and the corresponding relationship between the two modalities of pictures and text is not large. So using Decision Fusion methods in this study has better performance.

Mixed Fusion Evaluation

Mixed Fusion methods which include Basic Mixed Fusion and HLO Mixed Fusion has slightly better performance than Features Fusion methods, but not as good as Decisions Fusion methods. Mixed Fusion methods combine features fusion and decisions fusion. The model structure is more complex and the training process is harder. The most important factor affecting the performance of Mixed Fusion is the combination method and combination strategy, and its rationality will have a decisive impact on the final performance of Mixed Fusion. In this study, HLO mixed fusion has a good performance, but the performance of basic mixed fusion is average. In general, the performance cannot surpass Decision fusion, which may be because the most reasonable Mixed Fusion is not arranged, and the greatest advantages of feature fusion and decision fusion are not extracted to the greatest extent.

5 Conclusion and Discussion

In conclusion, It is found that not all multimodal fusion methods can achieve better performance than single-modality models. In this study, the performance of feature fusion is far inferior to decision fusion, and its effect cannot even surpass a single image modality model. Decision fusion achieves significantly better performance.

Among them, the Sum fusion method has the best performance with a f1 score of 88.65 in validation set. The reason for the good performance of decision fusion may be related to the low correlation between modalities, the good robustness and the high tolerance of errors in single modality model.

In future work, firstly, we would like to make further improvement in TFN, make the process of TFN first linearly transformed and then multi-dimensional dot product, which can be regarded as the sum of the results of multiple low-rank vectors, thereby reducing the number of parameters in the model. After reducing parameters, it is expected TFN can be realized in future. Secondly, we would like to search further resolutions for fusions of not highly correlated modalities. Besides, it is also expected to deal with better combination strategy for Mixed fusion methods to make better performances.

A Appendix

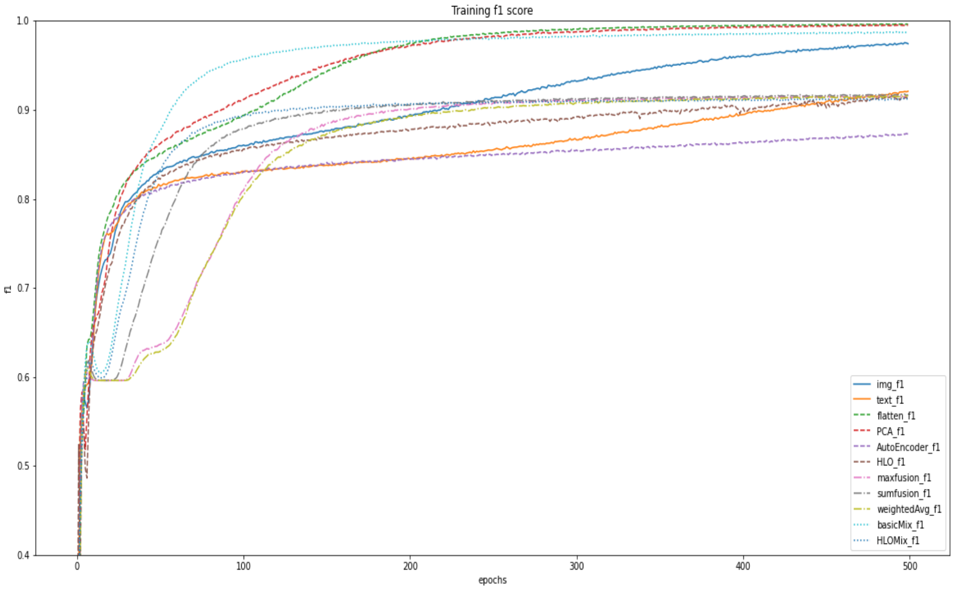

Figure 6: F1 score in Training set

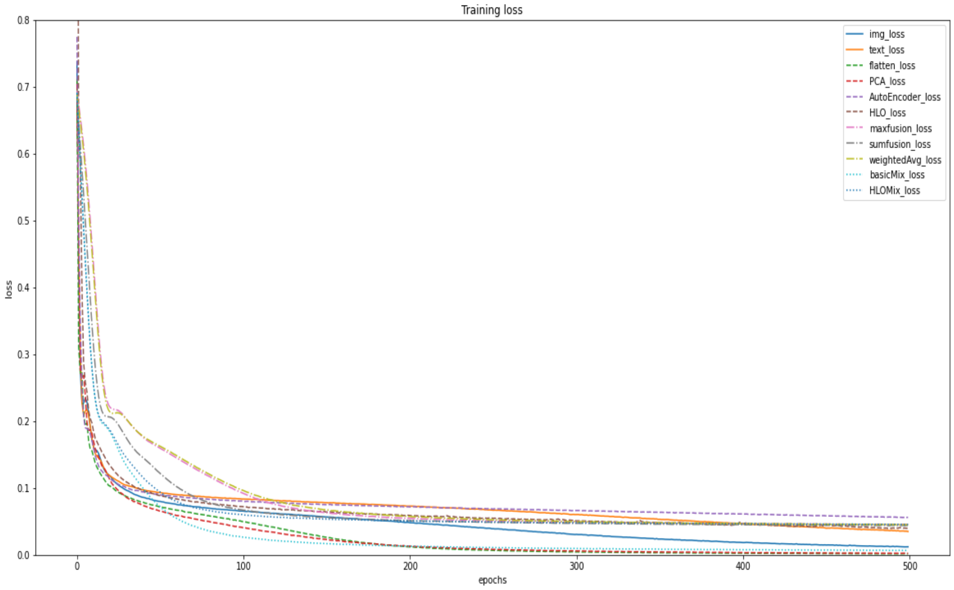

Figure 7: Loss in Training set

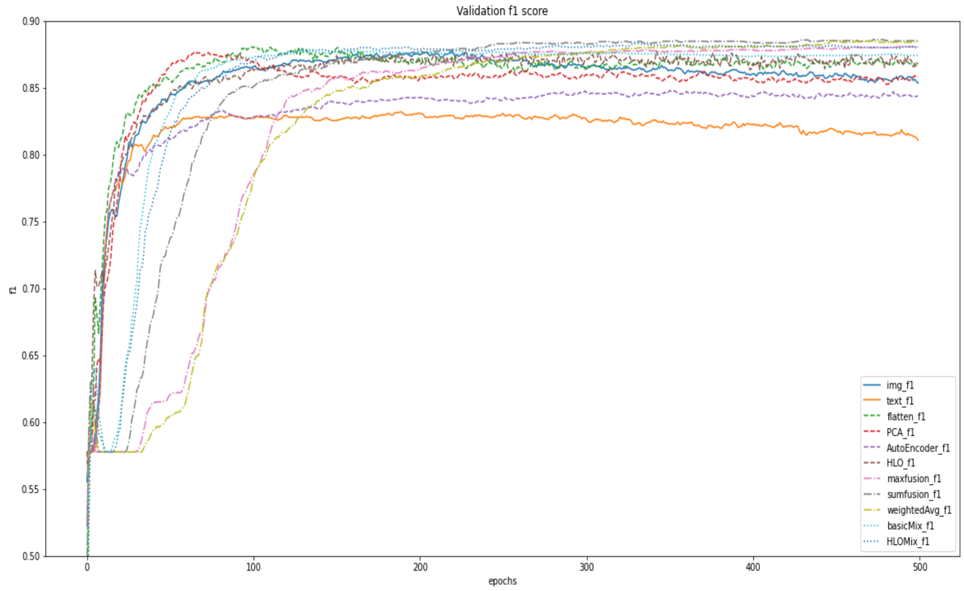

Figure 8: F1 score in Validation set

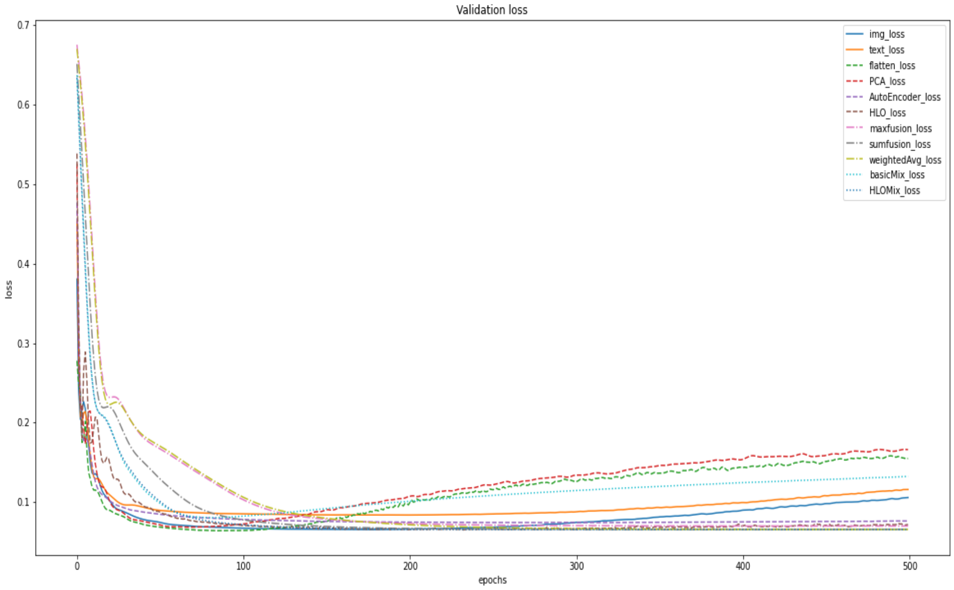

Figure 9: Loss in Validation set

Run Time
45 mins
hardware and software specifications
colab and Tesla P-100
Predicted result on test examples
0.89269 based on Kaggle
read me
To run our best model, we need the following steps:
1. Clip (ViT and Transformer all included)
2. THIS IS THE MODEL! ! ! ! ! ! ! ! ! ! ! ! LOOK HERE

Reference

[1] Mingxing Tan, Quoc V. Le, EffificientNet: Rethinking Model Scaling for Convolutional Neural Networks

[2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[4] Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Internet]. arXiv.org. 2022 [cited 22 May 2022]. Available from: https://arxiv.org/abs/1810.04805

[5]Castellano, G. , Kessous, L. , & Caridakis, G. . (2008). Emotion recognition through multiple modalities: face, body gesture, speech. Springer Berlin Heidelberg.

[6] Z. Z. Lan, L. Bao, S. I. Yu, W. Liu, and A. G. Hauptmann, “Multimedia classification and event detection using double fusion,” Multimedia Tools and Applications, 2014.

[7] Sahu, G. , & Vechtomova, O. . (2019). Dynamic fusion for multimodal data.

[8] Perez-Rua, J. M. , Vielzeuf, V. , Pateux, S. , Baccouche, M. , & Jurie, F. . (2019). Mfas: multimodal fusion architecture search. IEEE.

[9]Ming H., Jiajia T , Jianhai Z, Wanzeng K, Qibin Z. (2019). Deep Multimodal Multilinear Fusion with High-order Polynomial Pooling

[10]]Zadeh, A. , Chen, M. , Poria, S. , Cambria, E. , & Morency, L. P. . (2017). Tensor fusion network for multimodal sentiment analysis. arXiv.

[11] Liu, Z. , Shen, Y. , Lakshminarasimhan, V. B. , Liang, P. P. , Zadeh, A. , & Morency, L. P. . (2018). Efficient low-rank multimodal fusion with modality-specific factors.