Text and Image Multi-modal Fusion Embedding Classification Model based on BERT and ResNet

ZhangQi	Gao Renfei	Deng Jianing	Dai Zehua
23065105g	23053538g	23059218g	23045206g

Data Science & Analytics

Date: June 16, 2024

Abstract

Learned representations of social images have recently achieved remarkable success in many tasks such as cross-modal retrieval and multi-label classification. However, since social images contain multi-modal content (e.g.visual images and textual descriptions) as well as social relationship images, simply modeling the content information may lead to suboptimal embeddings. This paper proposes a multi-modal fusion embedding classification model for text and images based on BERT and ResNet, which achieves multi-modal fusion and classification of images and image descriptions at the feature layer. BERT technology is used to process the text to obtain the text embedding vector, and the residual network (ResNet) is used to preprocess the image to obtain the image word embedding vector. The two types of embedding vectors are sent to the MLP model to complete the final classification. In addition, in order to adapt to the final multi-modal fusion model and obtain better classification performance, the effects of the BERT fine-tuning model and the BGE fine-tuning model on text classification were compared, and the BERT model with higher accuracy was finally selected. After obtaining the embeddings of the two modal data, we used the concat and attention methods to fuse the embeddings respectively. The final results show that the application of attention in multi-modal fusion enables the model to better understand and utilize information from different modalities. Finally, the experimental results on the Microsoft COCO Caption data set show that the overall accuracy of the MLP multi-modal model is 9.8 percentage points and 5.6 percentage points higher than the single text classification model and the single image classification model respectively. The model after PCA dimensionality reduction The overall accuracy is 2 percentage points and 6 percentage points higher than the model before dimensionality reduction respectively. This proves that fusion classification of social image multi-modal data will be more accurate after dimensionality reduction.

Keywords: High dimentional data, Multimodality, text embedding, image embedding, BERT, ResNet, Attention, MLP

1 Introduction

With the rapid development of social networks, social images are becoming increasingly available in social networking sites such as Xiaohongshu, Tik Tok, and Instagram. Social images are often associated with multiple modalities (such as visual content and textual descriptions) and social relationships, as shown in Figure 1[1]the various links between images. Learning efficient multimodal representations of social images helps to efficiently process and analyze these data. The learned representations have achieved great success in various applications, such as image classification[2,3]; dimensionality reduction[4]; and cross-modal retrieval[5,6].

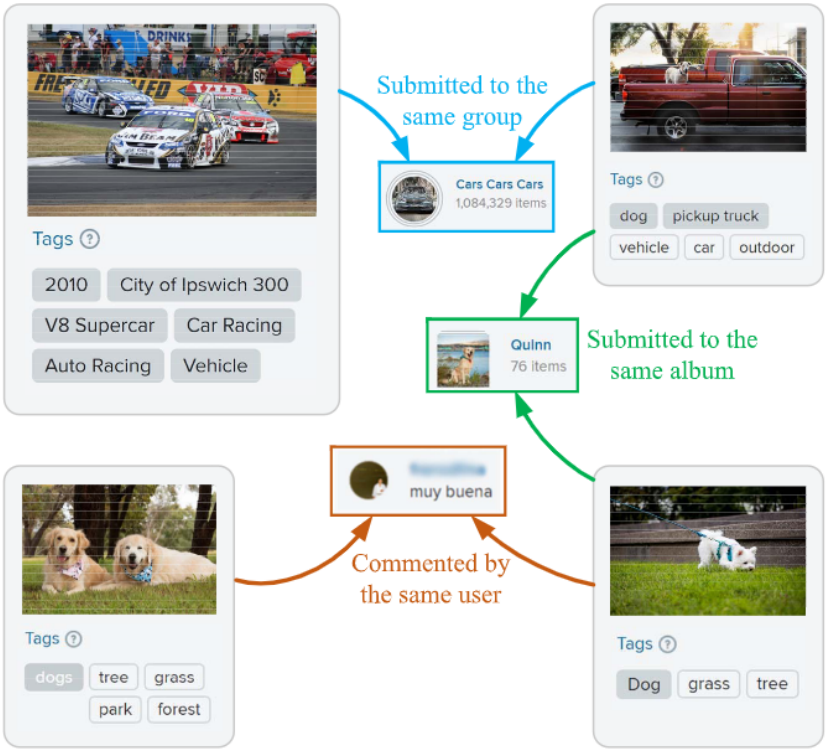

Figure 1: Illustration of the multiple types of links among social images,for example, submitted to the same album or submitted to the same group.

In recent years, how to embed multimodal data into low-dimensional vectorized space has attracted increasing attention from industry and academia. With the advent of deep learning, we can leverage textual and visual descriptions to effectively classify different images. Given an image and text, we generate image embedding and text embedding, and input the joint embedding of the two into our trained model, and then classify the image.

In summary, the contributions of our work are:

• Proposed mlp multi-modal weighted fusion model

• Encode contextual embeddings in latent space based on image descriptions and images.

• Perform downstream tasks, including image tag matching and image category mapping.

• Conduct multiple experiments on single-modal, multi-modal models to verify and validate performance and demonstrate better performance

2 Architecture

2.1 Overall architecture

By using BERT to generate text embeddings and ResNet to generate image embeddings, and fine-tuning the model, we investigated the following methods. The intuitive structure is shown in Figure2.

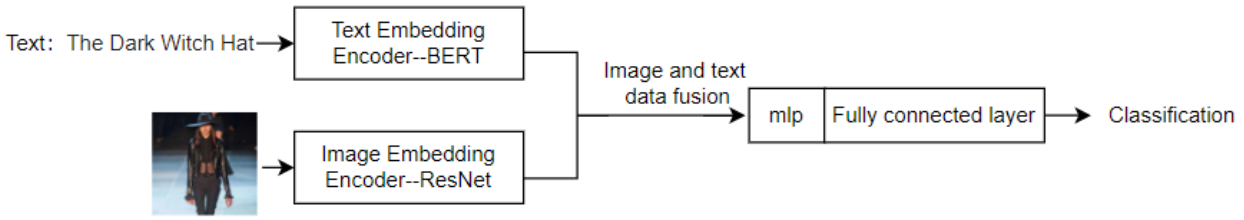

Figure 2: Model structure.

2.2 Feature layer multi-modal data fusion

According to the time of data fusion, the fusion method can be divided into input layer, feature layer and decision-making layer fusion.

The input layer is fused after data input and before feature extraction. Fusion at the input layer retains the originality of the data and emphasizes the correlation between data Since raw data is used, there is a lot of irrelevant data, which leads to increased uncertainty and instability. It not only increases the amount of calculations, but also increases the processing difficulty of the system and reduces the real-time performance of the system. When processing images and text Poor performance when fusing models.

The decision-making layer is integrated after feature extraction and recognition but before decision-making voting. Each part of the sensor in the decision-making layer information fusion has independent data processing capabilities for the same observation target. When applied to the mlp multi-modal model, the performance is as follows: after processing the image and text separately and making judgments, then merging the judgment results, and finally deciding on the required results. Since each model has processed multi-modal data separately, the coupling between data is reduced and the correlation of results is reduced. Therefore, decision-making layer fusion loses the general significance of multi-modal fusion.

Feature layer fusion occurs after data feature extraction and before decision-making. Compared with input layer fusion, feature layer fusion removes interference data in the original data, effectively organizes

the data, appropriately reduces the amount of calculation, and improves the processing speed of the system; compared with decision-making layer fusion, feature layer fusion It retains the effective connections of the data and improves the relevance of the results.

Comparing the three fusion methods, the loss values of the input layer and decision-making layer cannot converge stably[9]. To sum up, the mlp multi-modal fusion model chooses to fuse at the feature layer. This fusion method is more suitable for image-text multi-modal social image data processing. The principle of feature layer fusion is shown in Figure3:

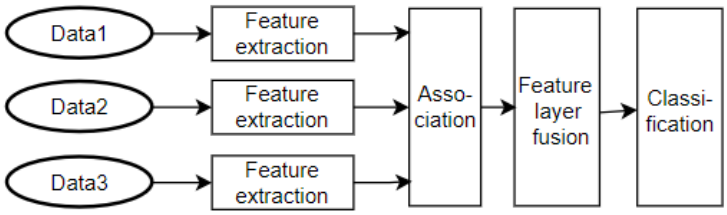

Figure 3: Feature layer fusion

2.3 Comparison and selection of text feature extraction models

2.3.1 BERT

BERT (Bidirectional Encoder Representations from Transformers)[7]is a bidirectional Transformer model used to generate context-sensitive word embeddings. BERT performs well in various natural language processing tasks. Since the advent of Bert, since the native Bert cannot directly obtain high-quality sentence vectors, many related methods have been derived. The more iconic ones are sentence-Bert and simcse. The former has verified that the language model has passed the text correlation task finetune. Afterwards, higher quality sentence vectors can be generated, and the latter verifies the benefits of contrastive learning for this type of task. The model structure of BERT is shown in Figure4:

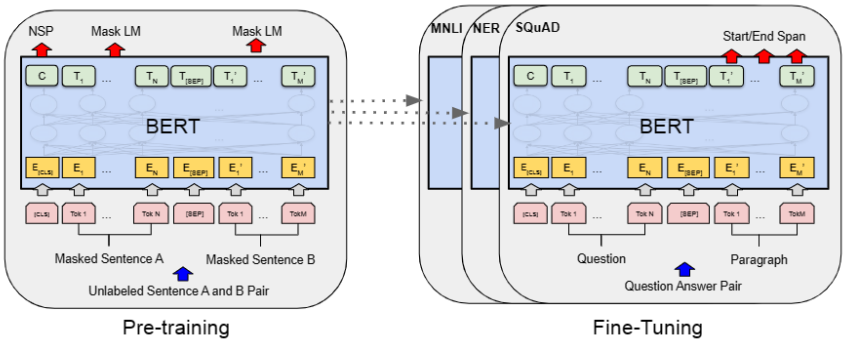

Figure 4: Overall pre-training and fine-tuning procedures for BERT.

2.3.2 BGE

BGE appeared in August 2023, adopting the route of automatic encoding + comparative learning. BGE pre-training adopts the RetroMAE[8] scheme, which includes an Encoder based on Bert and a Decoder with only one layer. During training, the Encoder side masks the original text at a ratio of 30%, and finally obtains the last layer. The vector representation of the [CLS] position is used as a sentence vector, and the Decoder side masks the original text at a ratio of 50%, and combines the sentence vectors on the Encoder side to reconstruct the original text. The model structure of RetroMAE is shown in Figure5:

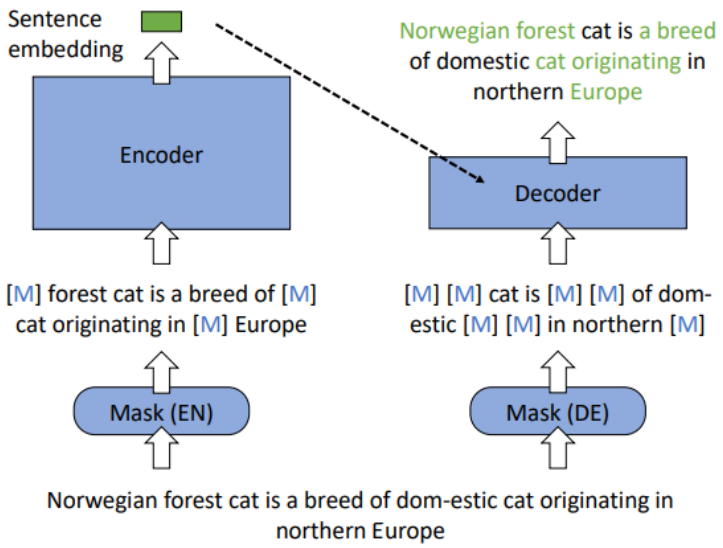

Figure 5: RetroMAE

2.3.3 Model comparison and selection

Both the BERT model and the BGE model are committed to learning meaningful low-dimensional representations from original data, and both utilize the idea of bidirectional modeling to enhance representation learning capabilities by capturing the bidirectional dependencies of data. It can be applied to various downstream tasks, such as Classification, Clustering, Pair Classification, Reranking, Retrieval, STS, Summarization, Bitext Mining, etc. But their modeling objects are different. BERT models text sequences and captures the contextual relationships between words; BGE models graph structures, capturing the relationships between nodes, features, and nodes. In addition, the training methods of the two are also different. BERT adopts an unsupervised pre-training method and fine-tunes after pre-training on large-scale corpus. BGE uses a supervised training method to directly perform end-to-end training on specific tasks. The comparison of accuracy and F1score after fine-tuning of BERT and BGE is shown in the figure6:

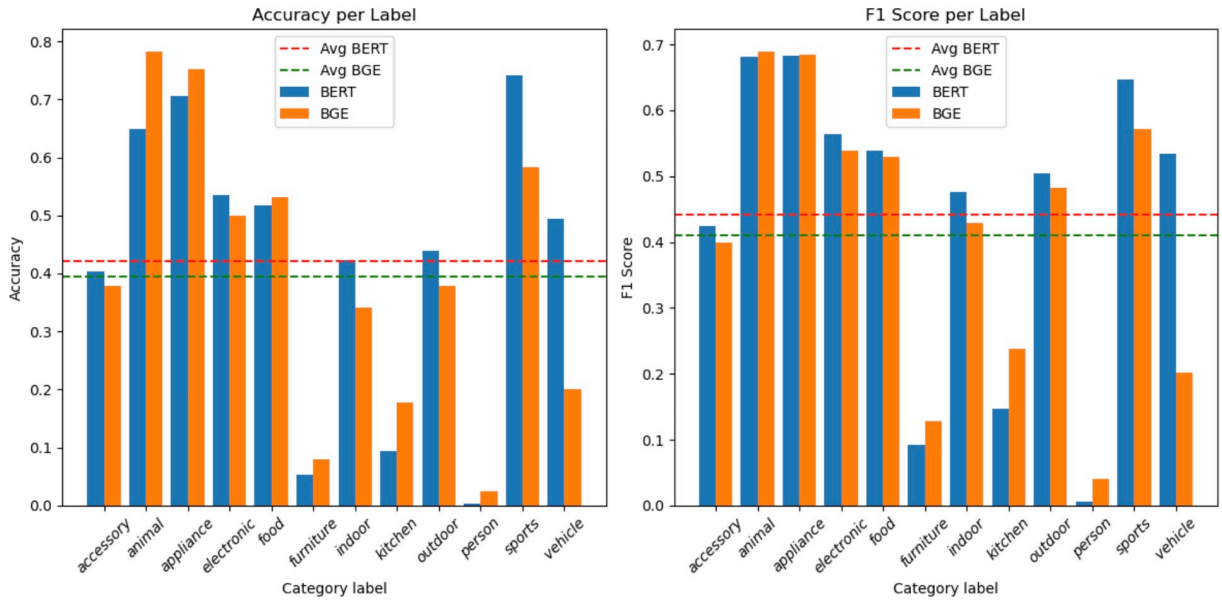

Figure 6: Comparison of accuracy and F1score after fine-tuning of BERT and BGE.

From the results:

• In terms of Accuary indicators, the overall accuracy of BERT is 0.423, while BGE is 0.39, and BERT is slightly better.

• In the accuracy and F1 score indicators on a single label, BGE performs better on some labels, which may be because BGE can better mine the potential correlation between tasks.

In summary, although BGE performs better on certain tags, it can better utilize the semantic correlation between tasks. But in terms of overall accuracy index, BERT is still slightly better. Considering that the goal of our project is to classify multi-modal images, we choose BERT as the feature extraction model for the text embedding encoder.

2.4 Image and text embedding encoder construction and image-text fusion process

2.4.1 Image and text embedding encoder construction

The function of the image embedding encoder is to convert the image into the input form required by the MLP multi-modal fusion model and obtain the embedded word vector. The image embedding encoder uses ResNet-152 as the image feature extractor to obtain word embeddings for a single image. ResNet uses two types of residual modules: the first residual module is composed of two 3×3 convolution kernels connected in series; the second residual module is composed of 1×1, 3×3 and 1×1 Convolution kernels are connected in series. In ResNet, the first residual learning module of each network layer needs to adjust the size and depth of the input feature matrix, so the two residual module structures need to be modified. Therefore, a 1×1 convolution kernel is added in parallel to the branch of the first residual module; a 3×3 convolution kernel

is added in parallel to the branch of the second residual module. The processing flow is shown in Figure 6 . During the entire model adjustment process, keep the weight parameters from being frozen. After the original image is processed as above, the image tag sequence is obtained, and the sequence is expanded into a 2048-dimensional vector. These image embedding vectors will be merged with text embedding vectors in the future and then entered into the mlp model for classification. The structure of the image embedding encoder is shown in Figure7:

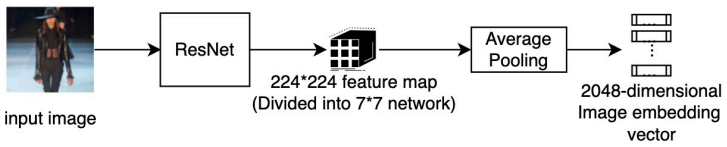

Figure 7: Image Embedding Encoder

The text embedding encoder is shown in Figure8. The text data of the image is processed by the encoder to obtain the text embedding vector. The first is the BertEmbeddings layer, which converts input words into 768-dimensional word embedding vectors. Then there is the BertEncoder layer composed of 12 BertLayer layers. BertEncoder is responsible for encoding word embeddings. Each BertLayer layer contains modules such as self-attention mechanism, feed-forward neural network, and layer normalization. Next is the Bert-Pooler layer, which pools the matrix output by BertEncoder to obtain a 768-dimensional vector representing the entire input sequence.

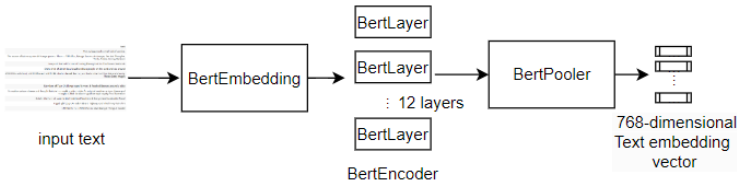

Figure 8: Text Embedding Encoder

The image word embedding vector and the text word embedding vector are fused together (the two vectors are spliced) and then input into the mlp model for training and learning.

2.4.2 Concat

Since both text and images contain rich information, by fusing them, we can obtain a more comprehensive and rich expression, which helps to improve the performance of the model. A common way to fuse text embedding and image embedding is to directly concatenate them together to form a multimodal representation. Specifically, for each sample, we concatenate the text embedding vector and the image embedding vector in a certain order to form a larger vector that contains the information of both text and image.

2.4.3 Attention

The second method is to use the attention mechanism (Attention) to fuse text and image embeddings. We use an attention mechanism to dynamically decide how much each modality contributes to the fused representation. First, we use the similarity between text embeddings and image embeddings to calculate attention weights. Then based on the calculated attention weights, we perform a weighted sum of text embeddings and image embeddings to obtain a fused representation. Finally, we use the resulting weighted sum as the fused representation. This fused representation incorporates information from text and images while taking into account the relationship and importance between them, thus better capturing the diversity and complexity of the data.The structure of the attention is shown in Figure9:

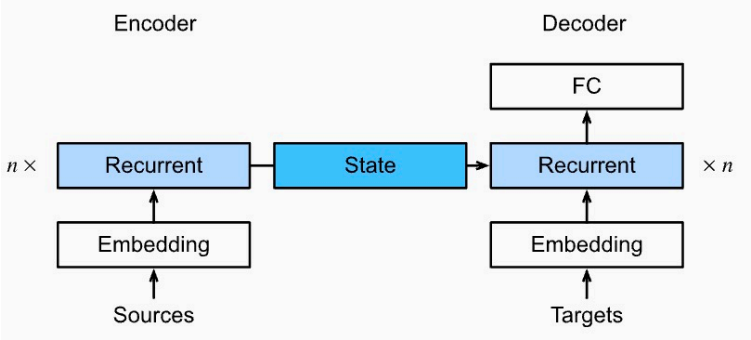

Figure 9: Image Embedding Encoder

3 Experiment and result analysis

3.1 Experimental datasets

This article uses the public dataset on Microsoft COCO Caption dataset[10]. The Microsoft Common Objects in Context (MSCOCO) dataset contains 91 common object categories, 82 of which have more than 5,000 labeled instances. The dataset has a total of 2,500,000 labeled instances in 328,000 images. We perform preprocessing and feature extraction on the image data to obtain a 2048-dimensional image embedding. The text data is preprocessed and feature extracted to obtain a 768-dimensional text embedding. The label data is converted into a multi-label form through one-hot encoding, and each sample corresponds to multiple binary labels. Then the data set is divided into training set and validation set according to 8:2. The data format is as shown in the table1:

Table 1: Image Data

image_id	file_name	category_id	category	super_category	caption
408201	COCO_val2014_408201.jpg	1	person	person	A man on a cellphone near two other men. a gro...
56592	COCO_val2014_56592.jpg	1	person	person	A young man in a white shirt is playing tennis...
93329	COCO_val2014_93329.jpg	1	person	person	A man standing in a living room with a woman a...
3761	COCO_val2014_3761.jpg	1	person	person	a woman wearing a cowboy hat face to face with...
242610	COCO_val2014_242610.jpg	1	person	person	A man standing next to a blue and brown bed. A...

3.2 Data dimensionality reduction

PCA is an unsupervised learning algorithm that aims to map high-dimensional data to a low-dimensional space while retaining the main information and changing characteristics of the original data as much as possible. After encoding the text data and image input embeddings, we use principal component analysis (PCA) dimensionality reduction technique to reduce the embedding vectors from the original high-dimensional space to lower dimensions for further processing. In this process, we reduce the dimensionality of the embedding vectors to 3 dimensions, then fuse text and image embeddings and use them as input to the mlp multimodal model.

3.3 Analysis of results

3.3.1 Experimental results from MLP multimodal model by concat

The MLP multimodal model results are shown in Figure10:

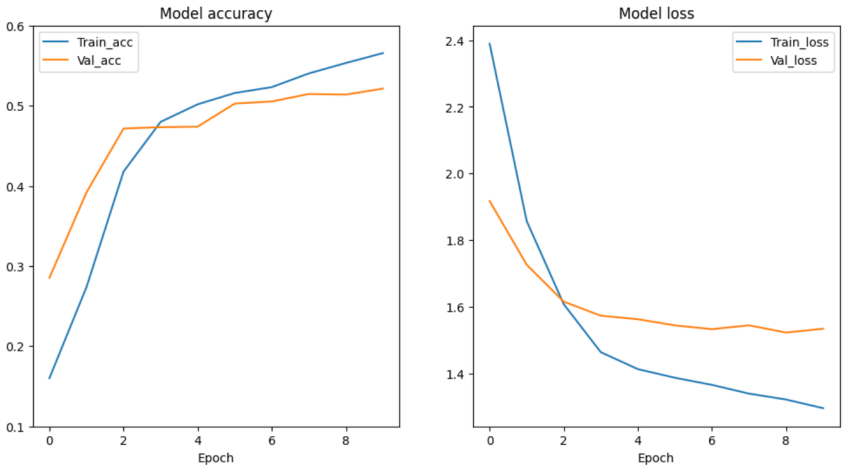

Figure 10: Model accuary and loss

As can be seen from the loss and accuracy curves, the loss on the first few epoch verification sets is smaller than that on the training set, and the accuracy on the verification set is higher than that on the training

set. It indicates that the previously obtained image-text embedding already contains more information, and the model has not yet extracted more effective features.

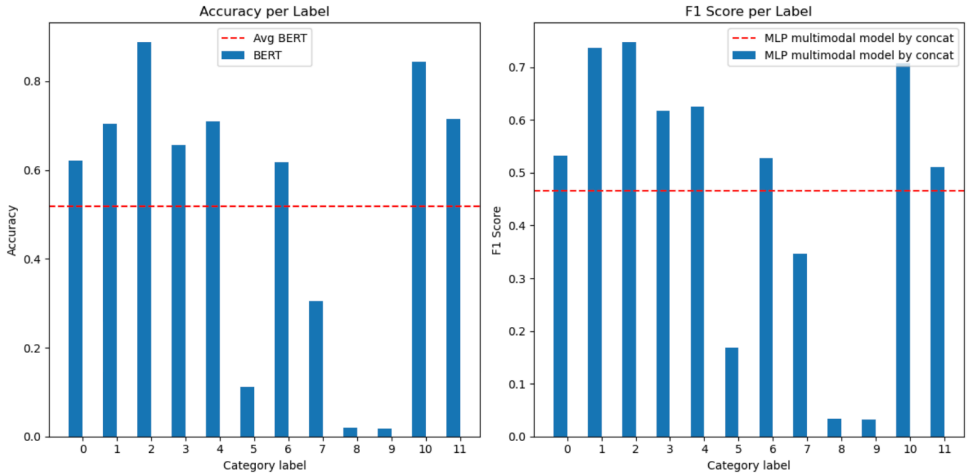

Figure 11: Accuracy and F1score for each category

It can be seen from the figure that although the overall accuracy and f1 score are improved compared with that of a single image and text embedding, the effect is still not very good for difficult to judge classes such as person.

3.3.2 Experimental results from MLP multimodal model by attention

The MLP multimodal model results are shown in Figure12:

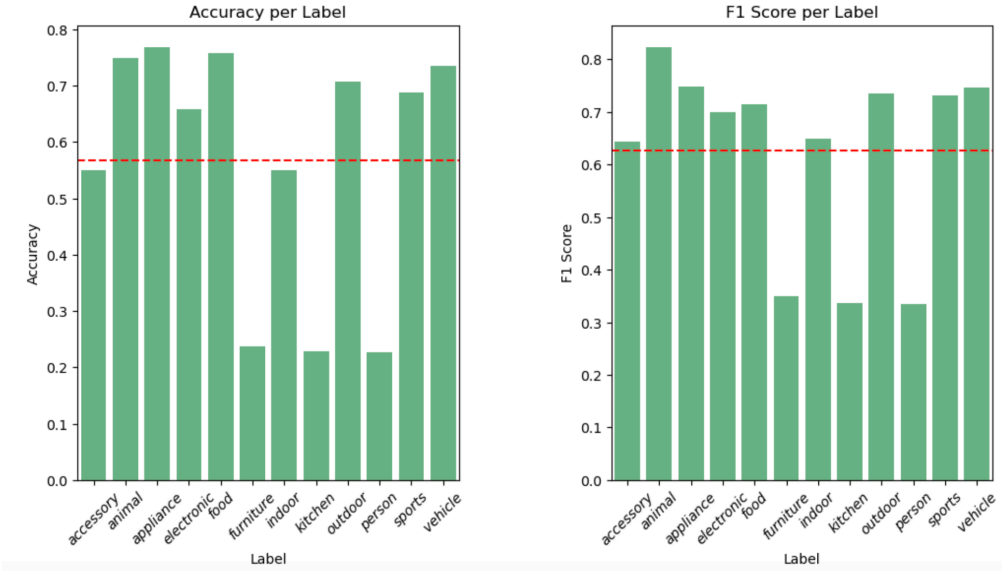

Figure 12: Comparison of accuracy and F1score from MLP multimodal model by attention

By connecting the text embeddings and image embeddings separately to a hidden layer, combining them through an attention layer, and then passing them through three additional hidden layers, we obtained our attention fusion model. After training the model, it is evident that there is a significant improvement in the classification performance.

Furthermore, it is evident that features such as ”person” that had poor classification performance when using only text information or multimodal by concat showed noticeable improvement after incorporating image information through the attention mechanism (although there is still room for further improvement).

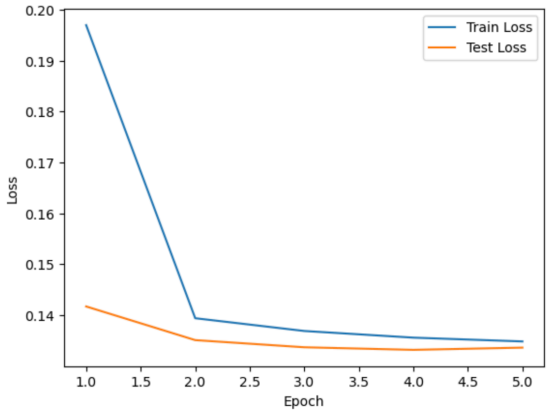

Figure 13: loss from MLP multimodal model by attention training

In addition, we also observed that during the training process of embedding fusion, there was not a significant decrease in the validation loss. Right from the beginning, the model exhibited good classification performance. We believe this is due to the fact that our inputs (text and image embeddings) were obtained from pre-trained models that were fine-tuned for the same downstream task (multi-label classification). As a result, the initial training already provided a good level of generalization performance.

3.3.3 Model comparison

Table 2: Model performance comparison

model	accuracy	F1 score
Single text classification model	0.4231	0.4331
Single image classification model	0.4646	0.4424
MLP multimodal model by concat	0.5213	0.4654
Multimodal model after PCA	0.5464	0.5290
MLP multimodal model by attention	0.5668	0.6261

The mlp multi-modal fusion human model, pure image classification model and pure text classification model were compared on the same training set of MSCOCO. The comparison results are shown in table2. All parameters are the same, the batch size is 4, the dropout rate is set to 0.1, and the initial learning rate is 5×10^{-5}.

As can be seen from Table 2, the final MLP multi-modal classification model accuracy and average F1 achieved the optimal results. Compared with the pure image classification model and the pure text classification model, the accuracy of MLP increased by 5.6 and 9.8 percentage points, and the average F1 increased by 2.3 and 3.2 percentage points. Experimental results show that the accuracy and F1 value of the MLP model are better than those of a simple image or text classification model, indicating that the MLP classification model fully integrates image information and text information, realizes the complementarity of the two types of information, and improves classification capabilities.

In addition, it can be found that after performing PCA dimensionality reduction on the fused embedding and classifying it through the same MLP, the accuracy and F1 have improved. It is speculated that PCA extracts the principal components and filters out the images and text embeddings. The interference information improves the classification effect of the model.

Last but not the least,using attention mechanism to fuse text embeddings with image embeddings yields a significant improvement compared to simply concatenating text embeddings and image embeddings and training with a neural network: both accuracy and F1 score show a clear enhancement. This undoubtedly demonstrates that employing attention mechanism for fusion is a more effective approach.We believe that using attention allows the model to dynamically capture the importance distribution of the two embeddings in the hidden layers. As a result, this approach enables a better integration of information from both embeddings, leading to the best classification results in our experimental process.

Besides, we also believe that combining the BERT model for obtaining text embeddings and the ResNet model for obtaining image embeddings within a larger model framework that incorporates attention mechanism (i.e., directly predicting based on input of both image-text descriptions and the images themselves) would yield better results compared to using fixed embeddings obtained from two separate models. This is because during the model training process, the fusion model can dynamically adjust the way both types of information are embedded and integrated, allowing for more effective utilization of the available information.

4 Summary

This paper explores learning multimodal data representation by integrating text content and image content to complete social image label classification tasks. Major contributions include:

1. Two multimodal fusion methods are tried to combine text and image information on the feature layer to improve the accuracy of classification.

2. By comparing BERT and BGE fine-tuning models, the BERT model was chosen as the baseline for text processing because it showed higher accuracy on text classification tasks.

3. In the experiment, concat and attention methods are used to fuse the embedding vector of text and image, and it is found that the attention mechanism can better utilize the information of different modes, thus improving the performance of the model.

4. PCA dimensionality reduction technology is introduced into the model to reduce the computational complexity and improve the classification effect. The experimental results show that the accuracy and F1 value of the model after dimensionality reduction are significantly improved.

5. Compared with the single text classification model and the single image classification model, it is proved that the multi-modal fusion model has obvious advantages in the classification task, in which the accuracy and average F1 value are improved.

Overall, the study demonstrates the effectiveness of multimodal fusion models in processing social image data, especially in improving classification task performance. By combining advanced BERT and ResNet technologies, as well as attention mechanisms and PCA dimensionality reduction, the model is able to capture and integrate image and text information more accurately, resulting in significant performance improvements in real-world applications. As for future work, we hope to use global network structure as well as local network information in social images for embedding. In addition, other heterogeneous information, such as user networks and user profiles, can be combined to better learn multimodal representations.

5 Contribution

Table 3: Team Member Contributions

Name	**Contributions**
Deng Jianing	1. Participate in topic selection for the project 2. Responsible for the image-text embedding fusion part 3. Participate in report writing
Zhang Qi	1. Participate in topic selection for the project 2. Responsible for data collection and preprocessing 3. Responsible for reporting (mainly)
Gao Renfei	1. Participate in topic selection for the project 2. Responsible for data collection and preprocessing 3. Responsible for text embedding 4. Participate in the image-text embedding fusion part 5. Participate in the report
Dai Zehua	1. Participate in topic selection for the project 2. Responsible for image embedding 3. Participate in the report

6 References

[1] Feiran Huang et al. “Multimodal learning of social image representation by exploiting social relations”. In: IEEE transactions on cybernetics 51.3 (2019), pp. 1506–1518.

[2] Xiaodong Yang, Pavlo Molchanov, and Jan Kautz. “Multilayer and multimodal fusion of deep neural networks for video classification”. In: (2016), pp. 978–987.

[3] Hanwang Zhang et al. “Online collaborative learning for open-vocabulary visual classifiers”. In: (2016), pp. 2809–2817.

[4] Jian Zhang, Jun Yu, and Dacheng Tao. “Local deep-feature alignment for unsupervised dimension reduction”. In: IEEE trans-actions on image processing 27.5 (2018), pp. 2420–2432.

[5] A Frome, GS Corrado, J Shlens, et al. “A deep visual-semantic embedding model”. In: Proceedings of the Advances in Neural Information Processing Systems (), pp. 2121–2129.

[6] Yuxin Peng, Xin Huang, and Jinwei Qi. “Cross-media shared representation by hierarchical learning with multiple deep net-works.” In: IJCAI. Vol. 3846. 2016, p. 3853.

[7] Nils Reimers and Iryna Gurevych. “Sentence-bert: Sentence embeddings using siamese bert-networks”. In: arXiv preprint arXiv:1908.10084 (2019).

[8] Xiao S, Liu Z, Shao Y, et al. RetroMAE: Pre-training retrieval-oriented language models via masked auto-encoder[J]. arXiv preprint arXiv:2205.12035, 2022.

[9] Shimaa El-Bana, Ahmad Al-Kabbany, and Maha Sharkas. “A multi-task pipeline with specialized streams for classification and segmentation of infection manifestations in COVID-19 scans”. In: PeerJ Computer Science 6 (2020), e303.

[10] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.