Journal of Economics and Technology Research
ISSN 2690-3695 (Print) ISSN 2690-3709 (Online)
Vol. 4, No. 1, 2023
www.scholink.org/ojs/index.php/jetr

Original Paper

DXVNet-ViT-Huge (JFT) Multimode Classification Network Based on Vision Transformer

Haoran Li[1], Daiwei Li[1], Haiqing Zhang[1*], Xincheng Luo[1], Lang Xu[1], & LuLu Qu[1]

[1] Chengdu University of Information Technology, Chengdu City, Sichuan Province, China

[*] Haiqing Zhang, Chengdu University of Information Technology, Chengdu City, Sichuan Province, China

Received: May 17, 2023 Accepted: May 20, 2023 Online Published: May 23, 2023

doi:10.22158/jetr.v4n1p59 URL: http://dx.doi.org/10.22158/jetr.v4n1p59

Abstract

Aiming at the problem that traditional CNN network is not good at extracting global features of images, Based on DXVNet network, Conditional Random Fields (CRF) component and pre-trained ViT-Huge (Vision Transformer) are adopted in this paper Transformer model expands and builds a brand new DXVNet-ViT-Huge (JFT) network. CRF component can help the network learn the constraint conditions of each word corresponding prediction label, improve the D-GRU method based word label prediction errors, and improve the accuracy of sequence annotation. The Transformer architecture of the ViT (Huge) model can extract the global feature information of the image, while CNN is better at extracting the local features of the image. Therefore, the ViT (Huge) Huge pre-training model and CNN pre-training model adopt the multi-modal feature fusion technology. Two complementary image feature information is fused by Bi-GRU to improve the performance of network classification. The experimental results show that the newly constructed Dxvnet-Vit-Huge (JFT) model achieves good performance, and the F1 values in the two real public data sets are 6.03% and 7.11% higher than the original DXVNet model, respectively.

Keywords

DXVNet-ViT-Huge (JFT), Pre-training ViT-Huge, CRF

1. Introduction

With the development of science and technology and the popularization of the Internet, the information processing based on single mode has been very mature at present, but there is still a broad research prospect for the multimodal fusion analysis of text and picture. The basic idea of multimodal fusion is to integrate data from multiple sources, so that we can have a more complete understanding of the

collected data, so as to obtain better performance in a series of tasks. At present, the existing multimodal fusion technologies include: (1) multimodal fusion based on explicit fusion, including feature layer fusion, decision layer fusion and hybrid layer fusion. (2) Multimodal fusion based on implicit fusion, including bilinear fusion (Lin, 2015), conditional random field (Baltrušaitis, 2013), tensory-based (Zadeh, 2017) fusion and attention-based fusion(Tsai, 2019; Wang, 2019). Fusing features from different modes in different layers of a neural network has several advantages. First, it allows the modeling of complex relationships between different modes, where text and images may have potential connections. Second, complementary features from different modes can be combined to reduce noise and improve overall quality.

The fusion of multimodal feature layer makes use of the connection between features for fusion, aiming to integrate the complementarity of multimodal data features into the same model, so as to improve the classification performance of the model. This approach is usually to express multimodal data in a unified form and fuse it into a larger feature representation. Multiple feature data are connected in series to form a complete feature vector, which is then input into the classifier for classification. The recognition algorithm based on feature layer fusion has been applied in a large number of early studies (Mansoorizadeh, 2010; Schuller, 2009; Wang, 2008; Wang, 2012).

The multimodal classification model chosen to be constructed in this paper utilizes the method of feature layer fusion. On the one hand, in the early fusion, data layer fusion would produce more redundant information, and this fusion mode did not carry out information interaction between modes, so it could not capture the correlation between different features. On the other hand, the importance of graphic modal information in multimodal classification will vary according to different situations. If late fusion is adopted, it will not only fail to carry out cross-modal information interaction, but also fixed the weight of a modal feature. Implicit fusion lacks popularity and generalization due to the need to build complex models and powerful equipment support. Therefore, in order to fully explore the internal features of multi-modes and model the information interaction between cross-modes, this paper chooses the way of feature layer fusion to combine different modal features.

This paper introduces the VIT-Huge (JFT) pre-training model on the basis of the existing DXVNet network. Vision Transformer (ViT) model is expanded and constructed from the famous Transformer model in the field of NLP. No internal changes have been made to Transformer. The biggest change in ViT is the chunking and dimensionality reduction of the image in the input side, which can be converted into a lexics-like expression for easy subsequent processing. The difference between ViT and traditional CNN classification network lies in the global feature brought by long range. Since the perception field of CNN is relatively limited, a multi-level structure of convolution, pooling and multilayer stacking must be adopted to expand its perception range. In this way, an obstacle is that the “effective/real” perception field is outwards from the center point and Gaussian attenuates, so the effective attention of CNN is only one or two more important components. The long range

characteristic of Transformer enables it to make better use of all useful information from shallow to deep, and the multi-head mechanism can ensure that the model can complete multiple discriminative parts at once. Therefore, each head can be regarded as an independent attention, which differentiates Transformer from CNN and expands another aspect of multi-mode fusion capability.

2. Related Work

At present, variants of Convolutional Neural Network (CNN) based models are widely used in image feature extraction, such as ResNet, VGG, Xception, etc. However, CNN also has a defect that cannot be ignored, that is, in the large model, it is likely to appear the phenomenon of overfitting. To solve this problem, ViT (Dosovitskiy, 2020) is the first Transformer model for image classification proposed by Google team in 2020.

Different from traditional CNN, ViT is a method of self-attentional learning of images, requiring each pixel to pay attention to another pixel. However, due to the square of the number of pixels, this method cannot be generalized to the real input size. For this reason, people have tried to apply the approximate method in the field of image processing (Parmar, 2018). It is only valid for local areas, but not for the whole area. This local multipoint product automatic recognition module can be a good substitute for convolution (Hu, 2019; Ramachandran, 2019; Zhao, 2020).

First, ViT splits the image into pieces and then uses the linear embedding sequence of the pieces as input to transformer. These fragments are handled similarly to tokens in NLP. The supervised method is used to classify the images. Vits perform well when pre-trained on a sufficient scale and migrated to downstream tasks with less data.

There is a special module in Transformer called “self-attention module”, in which every element will be associated with other elements when output. Therefore, the computational complexity of Transformer is $O(n^2)$, but it is this characteristic that makes Transformer have certain limitations. Only sequences with a length that cannot be too long can be calculated.

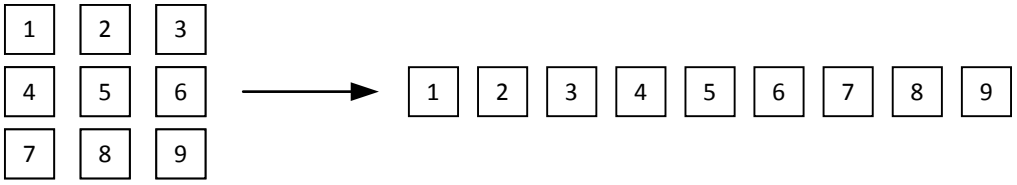

Figure 1. Image Sequence Segmentation

Therefore, a position embedding can be added to the image sequence segmentation strategy and the image patch embedding process. This allows space/location information to be stored globally using a variety of different strategies, as shown in Figure 1. In the process of preprocessing, the image should first be divided into patches, and each patch should be input into transformer as a token, because

attention will be paid to each token in transformer. Therefore, there is nothing inherently wrong with the order of the inputs. However, in the image, each connection point is arranged in a certain order. Therefore, similar to bert, a position embedding is added to each patch embedding. The final result refers to bert and uses 0 and cls to replace the entire result. The embedding corresponding to the result is the final output.

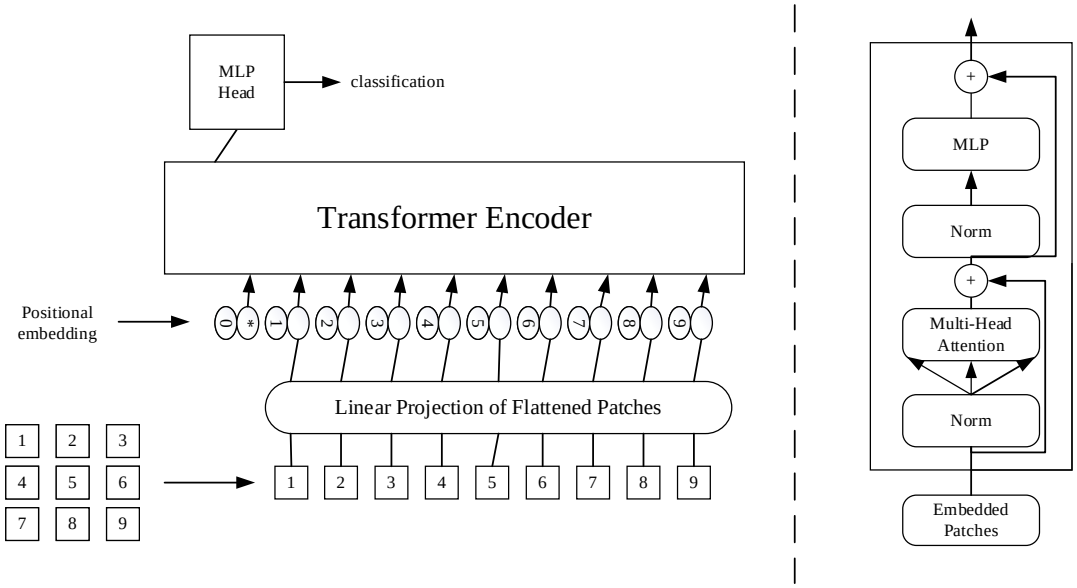

Figure 2. Overall ViT Network Structure

Firstly, the whole photo is segmented in Patchs, and then the features of the photo are extracted in Linear. On this basis, class labels are added, and the location is encoded, and then input into Transformer Encoder for calculation. This step is the same with Transformer model, without any changes. The first output corresponding to the class tag is then put into the MLP header to get the result. Finally, it is placed in the Softmax layer to obtain the final classification result. As shown in Figure 2.

From another perspective, Vision Transformer has less image-specific induction bias than CNN. Each level of the network includes locality, two - dimensional neighborhood and transformation equivalence. In ViT, only MLPS are local and translation-equivalent structures, while the self-attention layer is global. Among them, two-dimensional domain construction is rarely used, that is, at the beginning, the image is first divided into several small blocks, and then adjusted according to the resolution of the image, and then the experimental parameters are fine-adjusted. In addition, the initial position embedding of the algorithm does not include the two-dimensional position positioning of the fragments, but needs to recalculate the spatial relationship between the fragments.

3. The Proposed Method

DXVNet-ViT-Huge (JFT) network is expanded on the basis of the existing DXVNet, introducing CRF component and the pre-trained ViT (Huge) model. CRF can help the model learn the constraint conditions of each word corresponding to prediction tags, and improve the prediction error of word tags based on D-GRU method. Improve the accuracy of sequence labeling. The long range characteristics of ViT (Huge) model can bring global features and form a complementary connection with CNN in feature information extraction, because one problem of CNN's conv operator is that the field of perception is relatively limited, and a structure such as convolution-pooling-multilayer stacking is needed to increase the attention range of the network. Therefore, CNN is not good at extracting global features, but better at extracting local features of images. The overall architecture of the proposed approach is shown in Figure 3.

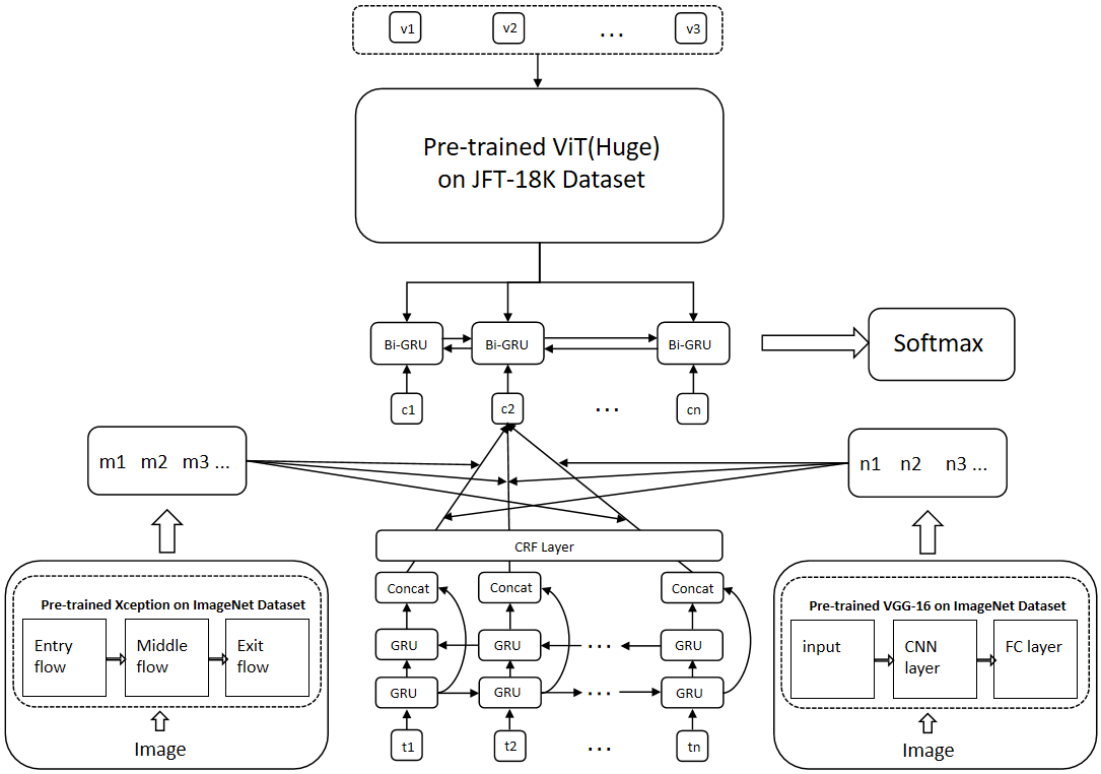

Figure 3. DXVNet-ViT-Huge (JFT) Network Architecture

The advantage of Convolutional Neural Network (CNN) is that it can provide two types of inductive bias: locality/two-dimensional neighborhood structure, which means that adjacent areas in the image have similar characteristics. One is called translation equivariance. When CNN has both the above two induction biases, it has a lot of prior information, and only needs less data to learn a better model, and can better extract the local features of the image that ViT model is not good at.

The main idea of DXVNet-ViT-Huge (JFT) model is to use the pre-trained Xception and VGG16 models to extract the local features of images, and then connect the text features extracted by D-GRUs introduced into CRF components. The connected multi-mode features and the global image information extracted from the pre-trained ViT (Huge) model are input into the Bi-GRU fusion module, and the Softmax layer carries out multi-classification at last.

3.1 CRF Component Setup and Detail

CRF (Huang, 2015), Conditional Random Fields, also called conditional random fields. For a set of input sequences, a conditional probability model is established, which can be used to describe another set of output sequences. A set of multiple random variables is called a random process. Random field is a random process whose index is spatial variable. A series of random parameters are distributed at a certain point according to a certain probability, the point on this point is called random field. The CRF structure diagram is shown in 4.

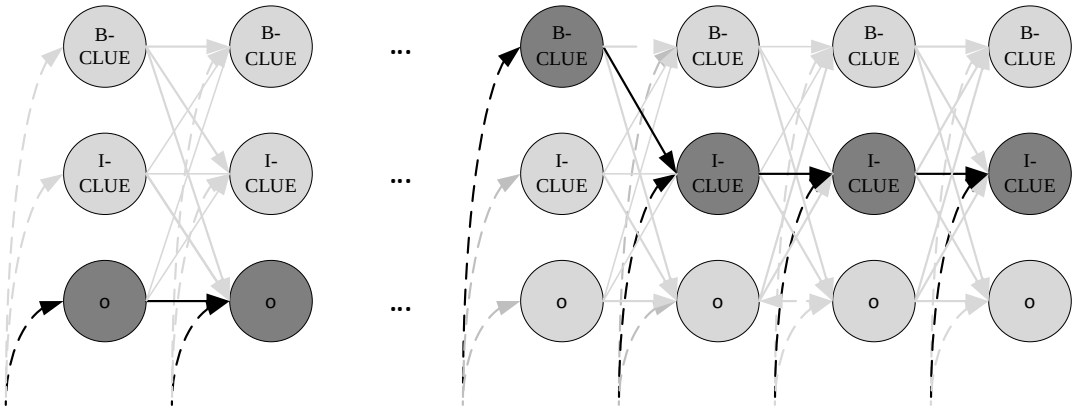

Figure 4. CRF Structure Diagram

In this paper, text modal information extraction is regarded as a sequence annotation task. Currently, sequence annotation is based on BiGRU-CRF. On the BiGRU layer, CRF layer is superimposed on BiGRU, so that the constraint conditions of prediction labels corresponding to each word can be learned through CRF layer. Thus, the error of word label prediction in BiGRU method is improved, so as to improve the accuracy of sequence labeling. Therefore, on this basis, this paper introduces the method based on D-GRU-CRF to extract text modal information.

3.2 Setup and Analysis of ViT Pre-training Model

At present, the ViT pre-training model can be divided into Base, Large and Huge variants according to the number of stacked layers. Like another commonly used Transformer model (GPT, BERT and RoBERTa), the size and number of layers of ViT are also different. This is reflected in the transformer layer and heads. Vit-l /16, for example, can be understood as a large (24-layer) ViT model with input image patch of 16x16. The different variant structures are shown in Table 1. In this paper, ViT-Huge

pre-training model is selected to form a part of the whole model.

Table 1. The Structural Composition of the Different Variants of ViT

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

The ViT pre-training model realizes fine adjustment of downstream tasks (small samples) by training a large amount of data. Among them:

The ImageNet dataset of ILSVRC-2012 contains 1000 categories and 1.3 million pictures.

ImageNet-21k has 14 million pictures and 21k classes.

JFT has 303 million high-resolution pictures and 18k class.

In VIT, it is necessary to have more samples in order to obtain better results, so the selection of ViT pre-training model is particularly important. In this model, VIT-Huge (JFT) pre-training model is selected, because when the amount of pre-training data is small (ImageNet), the performance of ViT is poor. However, with a large amount of pre-training data (JFT-300M), the ViT performed well. Transformers lacks the inductive biases of CNN, that is, preset prior knowledge, so there are certain limitations in the training process. When trained on sufficiently large sample sets, the performance of ViT will be better than that of convolutional neural network, and the limitation of Transformer lacking inductive bias will be broken, thus achieving better downstream task migration. The performance comparison of different pre-training data sets with different ViT variants is shown in Table 2.

Table 2. Performance Comparison of Different Pretraining Data Sets with Different ViT Variants

	JFT (ViT-H/14)	JFT(ViT-L/16)	21k(ViT-L/16)
ImageNet	**88.14**	87.23	85.03
ImageNet ReaL	**90.47**	90.11	88.24
CIFAR-10	**98.84**	98.41	98.16
CIFAR-100	**94.27**	93.58	93.02
Oxford-IIIT Pets	**97.65**	97.41	94.72
Oxford Flowers-102	**99.42**	99.28	99.16
VTAB(19 tasks)	**77.39**	76.19	71.89
TPUv3-core-days	**2.5k**	0.68k	0.23k

Multiple data sets were used to pre-train the ViT model on ImageNet-21k and JFT300M. For ImageNet-21k with small samples, the performance of both ViT-Large and ViT-Base models is not

ideal. It is at the JFT-300M that the benefits of the larger model are fully realized. At the same time, it can be concluded that with the growth of data set, the performance of the larger ViT variant exceeds that of the smaller ViT variant, as well as many traditional CNN image classification models.

4. Experimental Standard and Result Analysis

4.1 Experiment Preparation and Dataset Description

This section describes the data preparation and experiment environment. All the experiments were completed on a computer, running on Windows10 system, with AMD Ryzen 7 4800H CPU, NVIDIA RTX 2060 GPU and 1.5tb hard disk. The software environments are TensorFlow_gpu (1.14) and Keras (2.2.5) and Python (3.7).

In this paper, we use the data of MUTLA multimodal network teaching system to do this research. The data source includes the user's records, as well as the images taken by the webcam. There are 29996 user records and 6510 pre-processed emotions. In this dataset, user records are obtained from the user record storage area. Through web-based teaching, students are asked to slove problems the accuracy of current courses (physics, mathematics, English, language, chemistry), chapters, and exercises. The facial expressions of each student were photographed through a webcam installed on the computer. This data is time-stamped so that it can be synchronized properly. Table 3 shows an example dataset.

Table 3. MUTLA Multimodal online Learning Dataset

Subject	User record	Number of pictures	Total answer time(s)
phy	3780	820	50790
math	8271	1795	17012
en	15071	3271	25991
cn	1565	340	3835
chem	1309	284	3809
Total	29996	6510	101437

The article also uses evaluation information from online restaurants classified by food and restaurants on the Yelp.com website. Its business scope includes: Boston, Chicago, Los Angeles, New York, San Francisco, five major cities in the United States. The detailed data are shown in Table 4. Los Angeles has the most information, while Boston has the least. The results show that this method can effectively reduce the influence of uneven sample distribution on sample distribution. There are 44305 messages and 245569 photos in this data set. Because the number of pictures involved in each review is different, each review only selects 3 pictures, a total of 132915 pictures, in order to ensure the balance of the review data.

Table 4. Multimodal Restaurant Review Dataset

City	Document	Average sentence	Maximum sentence	Average vocabulary	Largest vocabulary	Picture
BO	2080	13.4	85	222.3	1115	10743
CH	2165	13.5	96	219.0	1107	12360
LA	24860	14.4	104	227.2	1134	137920
NY	11425	13.4	95	217.5	1129	61474
SF	3775	14.8	98	237.3	1145	22072
Total	44305	14.8	104	237.3	1145	244569

4.2 Dataset Preprocessing

Based on the fact that the goal of this study is to classify and identify the two groups of samples, the two groups of samples are divided into five categories.and the number of samples in each category is the same.The dataset will be 80% of the training set, 5% of the verification set and 15% of the test set. Considering the huge cross-discipline, cross-city and cross- cross-domain data scale of the two types of data sets, this project intends to use a unified input method for modeling and testing, and the samples to be tested are tested respectively. in order to better evaluate the specific performance of the model.

4.3 Experimental Results and Analysis

Experiments are carried out to prove the performance of DXVNet-ViT-Huge (JFT) model in this chapter. First of all, three variants of the ViT pre-training model are selected to train different pre-training datasets respectively for the test of real data, which reflects that the bigger the model and dataset, the better the pre-training can be transferred to its own dataset, the better the effect, more scalability and computational efficiency. Then, the DXVNet-ViT-Huge (JFT) model was tested by ablation experiments to verify its effectiveness. Finally, the DXVNet-ViT-Huge (JFT) model is used to compare the multi-modal classification models with different construction ideas of other authors, and the experimental results are demonstrated and analyzed.

Firstly, the setting of super-parameters of ViT-Huge (JFT) model was pre-trained: Adam optimizer was used for training, smoothing parameter was set to (0.9, 0.999), weight attenuation rate was set to 0.1, batch_size=4096. These parameters have good performance. Use the SGD optimizer for fine tuning with momentum set to 0.9 and batch_size=32. On this basis, cosine attenuation method is used, and the initial value is set to 0.03.

4.3.1 Performance comparison of three variants of the ViT pre-training model after training different pre-training data sets

Table 5. DXVNet Fusion of Different ViT Pretraining Model Performance Comparison

	MUTLA learning dataset			Restaurant reviews dataset		
Model	Precision	Recall	F1	Precision	Recall	F1
DXVNet	84.32	83.74	84.03	64.48	62.73	63.59
DXVNet-ViT-Base (ImageNet)	85.45	85.31	85.38	65.63	64.86	65.24
DXVNet-ViT-Base (JFT)	88.63	88.84	88.73	69.03	68.41	68.72
DXVNet-ViT-Large (ImageNet)	86.92	86.33	86.62	67.32	66.72	67.02
DXVNet-ViT-Large (JFT)	89.42	89.12	89.27	70.67	69.50	70.08
DXVNet-ViT-Huge (ImageNet)	87.34	86.45	86.90	68.52	68.81	68.66
DXVNet-ViT-Huge (JFT)	**90.92**	**89.23**	**90.06**	**71.64**	**69.78**	**70.70**

This experiment tests and analyzes the pre trained models of the three variants of the current ViT pre training model, Base, Large, and Huge, in ImageNet and JFT datasets.The results are shown in Table 5 above. The first line of results shows that the original DXVNet network without the introduction of ViT pre-training model has the worst effect, and the model only extracts the image features with local information, while the image features without global information are fused for discrimination, which proves that the new architecture introduced in this paper can improve the overall classification accuracy of the model. The results of lines 2, 3, 4, 5 and 6 and 7 in the table show that, in the case of the same variation of ViT pre-training model, the larger the data amount, the better the pre-training data set is required for the model to achieve better results. According to the F1 value of online education data set, it is found that the difference between the two types of pre-training data sets will cause about 3% difference in results. The comparison between the results in line 2 and 6 and line 3 and 7 shows that, with the same pre-training data set, different variants of the ViT pre-training model will also affect the classification accuracy, which proves that stacking the internal structure of the ViT can improve the model performance within a certain range. According to the observation of the two data sets, the difference of the ViT variants can cause a maximum difference of 2%. The DXVNet-ViT-Huge (JFT) model constructed in this paper achieved the best performance, with F1 values 6.03% and 7.11% higher than the original DXVNet model in the two real public data sets, respectively.

To sum up, the three variant structures of the ViT pre-training model and the difference in the amount of data in the pre-training data set will affect the model classification effect to some extent. In order to

improve the performance of the model, the selection of large data sets is prioritized over the selection of ViT variant structures.

4.3.2 Ablation analysis

According to the contribution degree of different components in the DXVNet-ViT-Huge (JFT) network, the ablation experiment is carried out in this section. The ablation experiment starts with the most basic configuration as the benchmark model, and gradually builds a complete DXVNet-ViT-Huge (JFT) network. The ablation experiment structure is shown in Table 6. This paper starts with Bi-GRU as the basic model, and the first line shows that the average test accuracy of the basic model in the online learning data set is 78.15%. In the second line, by constructing the DXVNet network in Chapter 3, the experimental results are significantly improved by 6.31%, indicating that DXVNet network has substantial innovation and improvement compared with the traditional single-mode classification network. The third line indicates that the CRF component introduced into the D-GRU structure of DXVNet network improves the performance of the original DXVNet network by 1.3%, indicating that to a certain extent, the CRF component can add constraints to the prediction label more accurately and obtain better auxiliary information extraction results through screening. The ViT-Huge (JFT) pre-training model is introduced in the last line to form the complete DXVNet-ViT-Huge (JFT) network in this paper, which is finally improved by 11.79% compared with the basic model. It is proved that DXVNet-ViT-Huge (JFT) network can indeed bring performance improvement compared with some traditional models.

Table 6. DXVNet-ViT-Huge(JFT) Network Ablation Experiment

DXVNet	CRF	ViT-Huge	PHY	MATH	EN	CN	CHEM	Avg
×	×	×	76.12	77.45	77.83	74.52	73.34	78.15
√	×	×	83.48	84.35	85.27	81.73	81.04	84.46
√	√	×	84.62	85.23	86.72	83.13	82.77	85.76
√	√	√	**89.24**	**89.81**	**90.54**	**87.39**	**86.26**	**89.94**

4.3.3 Comparison of the performance of multimodal fusion models with other construction concepts

In this section, the DXVNet-ViT-Huge (JFT) model is used to compare the multi-mode classification models of other authors in the field of multi-mode fusion, and the specific results are shown in Table 7.

Table 7. Performance Comparison of Different Multimode Fusion Models

Model	TFN-mVGG	BiGRU-mVGG	HAN-mVGG	DXVNet	DXVNet-ViT-Huge(JFT)
Text feature fusion	√	√	√	√	√
Image feature fusion	√	√	√	√	√
Hierarchical feature fusion	√	×	√	×	√
ViT feature fusion	×	×	×	×	√
Phy	71.83	77.05	80.62	83.48	**89.24**
Math	72.36	77.72	81.39	84.35	**89.81**
En	74.74	79.46	82.62	85.27	**90.54**
Cn	78.83	75.63	77.24	81.73	**87.39**
Chem	78.34	74.27	77.91	81.04	**86.26**
Avg	73.32	79.15	83.25	84.46	**89.94**

TFN-mVGG is the core part of the international advanced multi-mode tensor fusion network (Zadeh, 2017). By using the tensor fusion layer, the characters in HAN-ATT and the images in VGG are effectively fused, thus realizing the effective identification of the inferred subnet. Although TFN model can provide rich interaction between text and visual features, its performance is the worst among comparison methods, and the accuracy of TFN-MVgg is 73.32%. This result supports the hypothesis in this paper that the classification ability of models with only local image feature information is weak. Due to the incompleteness of feature information, there are great differences between image and text. Therefore, it is difficult to achieve accurate alignment between image and text in the model by combining features through complex fusion matrix.

Bigru-mvgg combines the representations BiGRU gets from the text with the representations VGG gets from the image, and then feeds these representations back to the classifier. BiGRU has been proven to be useful for sequential data, such as text (Tang, 2015). In the aspect of image coding, VGG-16 framework is used, and Imagenet-based data set is used for pre-training. In the FC7 layer, image features are extracted before the classification layer. The accuracy of BiGRU-mVGG reached 79.15%, which was 5.83% higher than that of TFN-mVGG. These models combine the characteristics of the two modes by concatenation.

HAN-mVGG is a combination of VGG targeting text and pictures. HAN-ATT(Yang, 2016) method adopted word and sentence encoding methods to realize the hierarchical structure of text. Compared

with HAN-ATT's plain text soft attention layer, the main difference lies in the modeling of global image feature information extraction by DXVNet-ViT-Huge (JFT). The performance of layered HAN-mVGG is superior to that of BiGRU-mVGG, with a performance gap of 4.1%. In contrast to BiGRU's single-level text model (word level only), the former introduces hierarchical modeling into the model. This model is based on text elements and has soft attention support.

DXVNet model is a multimodal classification network proposed and constructed in my previous research. Compared with the previous three methods, this model constructs a brand new asymmetric bidirectional D-GRU structure, which converts the simple series output information in the past and future directions into the output that can use the past information in the backward transmission. Thus, the important features of the text can be better extracted, so the classification performance is slightly better than the previous three methods. Finally, the latest DXVNet-ViT-Huge (JFT) model proposed in this paper has the best performance in all data sets, with an average accuracy of 89.94%, which is 16.62% higher than the basic model TFN-mVGG and 5.48% higher than the original DXVNet model.

5. Conclusion

In this paper, CRF mechanism and ViT-Huge (JFT) pre-training model are introduced based on DXVNet network to form a brand-new Dxvnet-Vit-Huge (JFT) network. The new network model also uses D-GRU structure and Bi-GRU structure. It can better combine the fusion modes of the two modes in different extraction stages and enhance the classification performance of the network. DXVNet-ViT-Huge (JFT) network combines the fusion mechanism from low-level features to high-level features and from local features to global features, and finally gets the fused features for classification. Through several experiments, the results show that the DXVNet-ViT-Huge (JFT) network is very efficient for multimodal text classification, and good performance indexes are obtained. In future work, this study will further utilize deep learning and other technologies to improve the classification accuracy of multimodal data fusion.

References

Baltrušaitis, T., Banda, N., & Robinson, P. (2013, April). Dimensional affect recognition using continuous conditional random fields. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (pp. 1-8). IEEE.

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*.

Hu, H., Zhang, Z., Xie, Z., & Lin, S. (2019). Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3464-3473).

Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Lin, T. Y., RoyChowdhury, A., & Maji, S. (2015). Bilinear CNN models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 1449-1457).

Mansoorizadeh, M., & Moghaddam Charkari, N. (2010). Multimodal information fusion application to human emotion recognition from face and speech. *Multimedia Tools and Applications*, *49*(2), 277-297.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A. & Tran, D. (2018). Image Transformer. *Proceedings of the 35th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* 80:4055-4064. Retrieved from https://proceedings.mlr.press/v80/parmar18a.html

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., & Shlens, J. (2019). Stand-alone self-attention in vision models. *Advances in neural information processing systems*, *32*.

Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., & Wendemuth, A. (2009, November). Acoustic emotion recognition: A benchmark comparison of performances. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding* (pp. 552-557). IEEE.

Tang, D., Qin, B., & Liu, T. (2015, September). Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1422-1432).

Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019, July). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting* (Vol. 2019, p. 6558). NIH Public Access.

Wang, Y., & Guan, L. (2008). Recognizing human emotional state from audiovisual signals. *IEEE transactions on multimedia*, *10*(5), 936-946.

Wang, Y., Guan, L., & Venetsanopoulos, A. N. (2012). Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. *IEEE Transactions on Multimedia*, *14*(3), 597-607.

Wang, Y., Shen, Y., Liu, Z., Liang, P. P., Zadeh, A., & Morency, L. P. (2019, July). Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 7216-7223).

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016, June). Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 1480-1489).

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Zhao, H., Jia, J., & Koltun, V. (2020). Exploring self-attention for image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10076-10085).