Video-Based Recognition of Online Learning Behaviors Using Attention Mechanisms

Bingchao Huang
Sino-French Engineer School
Beihang University
Beijing, China
huangbingchao@buaa.edu.cn

Chuantao Yin*
Sino-French Engineer School
Beihang University
Beijing, China
chuantao.yin@buaa.edu.cn

Chao Wang
EURECOM
Sorbonne University
Biot, France
chao.wang@eurecom.fr

Hui Chen
Department of Planning and Finance
Beihang University
Beijing, China
chenhui@buaa.edu.cn

Yanmei Chai
School of Information Central
University of Finance and Economics
Beijing, China
ymchai@cufe.edu.cn

Yuanxin Ouyang
School of Computer Science and Engineering
Beihang University
Beijing, China
oyyx @buaa.edu.cn

***Abstract*—In the field of education, identifying students' online learning behavior is a very effective means to understand students' learning status and improve teaching efficiency. However, previous research has mostly been based on older models. The shortage of datasets in this task can also be regarded as a problem. Therefore, this study first constructed a video dataset consisting of 10 types of students' online-learning behaviors (SOLB), and then proposed a Neural Network model based on Attention mechanism for identifying student online learning behaviors (CNN-Swin). The network is inspired by Swin Transformer and Convolutional Neural Network(CNN) at the same time. It takes a single frame of image as input, and first uses a series of convolutional layers to efficiently extract the primary spatial features of the image and reduce the spatial size of the feature map. Then, it uses a local Self-Attention mechanism with window translation to extract deep spatial features of the image. The network has a high prediction speed due to its low complexity and compression of inputs. The study also adds the popular ImageNet dataset as pre-training to demonstrate the effectiveness and out-performing of this proposed model, which finally approach accuracy of 90.42% for classification of students' behavior. In comparison with SOTA models, the outstanding perform of CNN-Swin with pre-trained methods is also be proved in many benchmarks.**

Keywords—Attention mechanism, Online Learning, Neural Network, Image Classification, Behavior Recognition

I. INTRODUCTION

The COVID-19 pandemic in recent years has made the use of online learning and distance learning more common. However, one problem that online learning may face is that teachers may have difficulty observing students' learning status in real time, finding students with learning problems quickly and adjusting methods of teaching. This issue will influence negatively the effects of students' participation in class. Therefore, it is essential to provide teachers with real-time feedback on student learning status in online courses. On the other hand, in the fields of face detection and human behavior classification, relevant research has already been emerging in an endless stream. These Neural Network models have stronger robustness and higher accuracy compared to traditional mathematical methods.

In this study, a new Neural Network model based on Attention mechanism[1] will be applied to identify student behavior in online learning. Due to the lack of relevant datasets, the study will define a classification method for students' online learning behavior and construct a dataset for it. Then, a Neural Network model and a complete data dealing process will be proposed, as well as a method of pre-training and transfer learning. The model (structure presented in Fig. 1) proposed is based on Attention mechanism[1] - one of the most popular algorithms in recent years. It uses 2 stage of blocks in the model of Swin Transformer[2], and at the same time, reduce the size of feature maps by adding covolutional layers[3]. It has an extraordinary innovative design to make the model be lightweight while maintaining accuracy of classification. Effectiveness and advancement of model will be verified through a series of experiments.

As a result, the model and dataset proposed in this article, after rigorous demonstration, will wonderfully improve the classification effect of students' movements and expressions, and overcome the problems faced by this sort of task in the past - the antiquity of models and the lack of datasets. This research can be applied in real-time online courses, in order to provide effective assistance for teachers in tracking student learning status.

II. RELATED WORK

A. Previous study of student behavior recognition

In recent years, Neural Network-based image classification methods have gradually been applied to student behavior recognition in the classroom. For example, Abdallah et al.[5] pre-trained the VGG16 model[6] on a facial expression dataset, and then transferred learning on their own constructed student offline classroom behavior dataset. Wang et al.[7] also used the VGG16 network, but constructed a dataset of cameras recording videos in remote classrooms in primary schools. Liu et al.[8] aimed to identify abnormal behavior among students in offline classrooms, and improved the YOLO v3[9] network to increase the receptive field.

In addition, there are some related studies that choose to use dynamic frame image sequences as inputs to the model. For example, Liu et al. [10] applied Residual Connections to a 3D Convolutional Neural Network(CNN) and used the network to classify student behavior datasets. Lin et al. [11] extracted the data of human joint landmarks in each frame in the data preprocessing and used the data of consecutive multiple frames to synthesize the prediction. Xie et al. [12] chose to use an edge detection algorithm to extract the edges

*Corresponding author

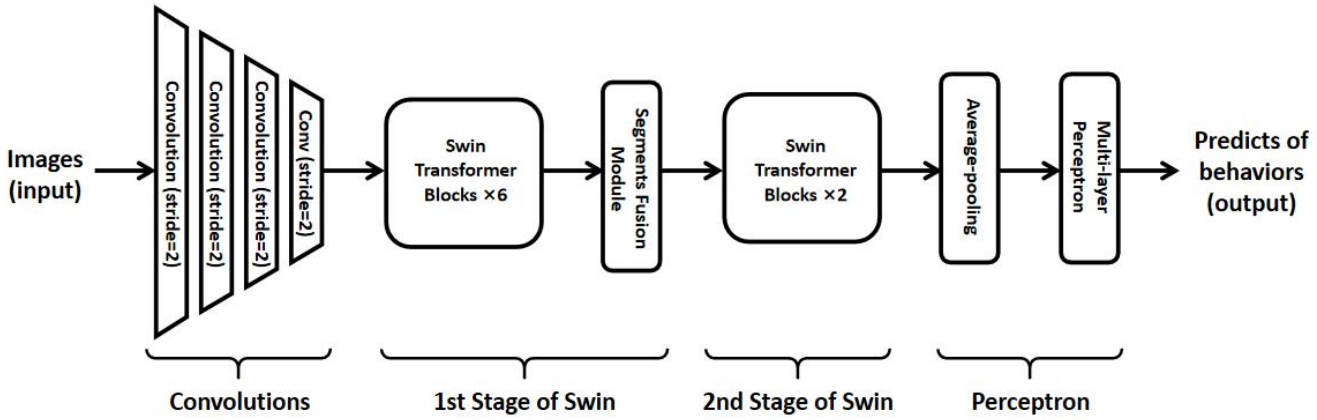

Fig. 1. Structure of the model CNN-Swin

of each frame in the frame sequence and then import them into a CNN. In general, training a model with a series of dynamic frames (equivalent to video) as input generally has a slight improvement in accuracy compared with using every single frame image separately as input. However, the computational cost of models with dynamic frames tends to increase significantly. In this study, the authors hope to explore a lightweight method to achieve accurate sample classification with minor cost of computation, so dynamic frames cannot be used as input.

In the field of education, there is still a significant lack of student behavior datasets. In the behavioral dataset constructed for traditional offline teaching scenarios, Bo et al.[13] used high-definition cameras in university classrooms to monitor videos and constructed a dataset of student offline classroom behavior videos. This dataset contains 11 typical behaviors of students in offline classrooms, but it is currently not publicly available. DAiSEE[14] is a video dataset about the emotional states of students in online learning. But its data imbalance problem poses challenges for model training.

Overall, the current research on student classroom behavior recognition mainly faces the following issues: a lack of publicly available datasets, and most studies focus on offline classrooms, with little research on student online learning behavior recognition. In terms of classification models with single frame input, the Neural Network models used are more traditional, such as VGG16, YOLO v3, etc.. But the recent research hot-spots, such as the Transformer model or other Attention mechanism models, have not yet been applied to student classroom behavior recognition tasks. These are the issues that need to be addressed in this study.

B. Face detection model

The face bounding box detection model BlazeFace[15] is a lightweight and effective face detection model proposed by Google in 2019. In order to achieve lightweight network for real-time detection of targets, the model adopts deep separable convolutions instead of conventional convolutions, and increases the size of the convolution kernel to reduce model depth in order to expand the receptive field. It also adds an additional layer of deep convolution to accelerate the process of reducing the spatial size of the feature map.

C. Transformer[1] and ViT[16]

The Transformer model is designed based on the Multi-Head Self-Attention(MSA) mechanism[1]. The Google team first proposed it in 2017, which is used to calculate the association between each word and all other words in the field of Natural Language Processing(NLP). Transformer has been widely adopted and achieved excellent performance in tasks such as sequence annotation, classification, sentence relationship judgment, and generative tasks in the field of NLP. The first model to successfully migrate it to the field of Computer Vision(CV) was Vision Transformer(ViT)[16]. The overall structure of ViT consists of three parts: sequential construction of image embedding vectors, Transformer Encoder, and classifier of Multi-Layer Perceptron(MLP). Only the first step is constructed as an innovative design for CV. Although the ViT model is a breakthrough attempt to apply Transformer, it still has two major drawbacks due to the nature of tasks in the field of CV: insufficient capture of image spatial information and high complexity.

Due to the commonality of Transformer and ViT, their improved models have been a hot topic. In the last year or two, many new research has emerged that has given us more inspiration. For example, the ViT model with Registers proposed by Darcet, et al.[21] proves that the training effect can be greatly improved in some aspects by transforming the structure of the ViT or similar model. In addition, Chen, et al.[22] proposed a multi-scale Transformer, and Jain, et al.[23] proposed an improved Transformer model that dominates the field of universal image segmentation. They all prove that there is still a lot of room for improvement in Transformer and ViT in the fields of time series images, image segmentation and classification. All of this has encouraged us to explore this topic.

D. Swin Transformer[2]

In response to the shortcomings of ViT pointed out in the previous section, recent research has proposed many variants of ViT for improvement, the most famous being the Swin Transformer model[2]. Due to the high complexity of the Multi-Head Self-Attention mechanism in the ViT model, the Swin Transformer module replaces the global MSA mechanism with local MSA based on window translation to make the computation more efficient, while other layers in the Transformer encoder remain unchanged. Each Swin Transformer block first includes a local MSA part based on window translation (Window MSA, W-MSA), followed by a two-layer feed-forward network MLP. The nonlinear activation function used between the two layers of MLP is GELU. Before each Self-Attention part and MLP part, there must be a Layer Normalization(LN).

E. ImageNet-1k dataset[4]

The public dataset used in this study is mainly ImageNet-1k[4]. It is a subset of the ImageNet dataset used for Large-Scale Visual Recognition Challenges(ILSVRC) and is one of the most famous benchmark datasets in image classification tasks. It contains 1000 item categories of items, including 1 281 167 training images, 50 000 validation images, and 100 000 test images.

III. METHOD

A. Face Cropping

Considering that most of the behaviors in online learning for students will include student faces, before inputting images into model, it is necessary to first attempt to detect the region where the face is located and remove background regions to reduce irrelevant information on behavior recognition. This will be beneficial for the training. We plan to mainly use BlazeFace[15] - the facial recognition technologies mentioned earlier, first in the data preprocessing stage, to complete the recognition of faces and the cropping of facial parts. It is embedded into a framework called Mediapipe[17], which is also developed by Google. In the following part, we'll use this framework for body cropping too.

B. Pre-training and Transfer Learning

So far, ViT model[16] and their variants typically require pre-training on larger datasets to achieve good performance. However, our dataset has the limitation of having a smaller data volume. According to Steiner et al.[18] found through experiments in their article, for training of ViT, it is more cost-effective and efficient to pre-train on other large datasets and then transfer learning to train a model on their own dataset. Therefore, we chose to pre-train the CNN-Swin model on a larger ImageNet-1k dataset for the task of image classification, and then transfer learning on our student online learning behavior dataset SOLB.

C. Student Online Learning Behavior(SOLB) dataset

The dataset proposed in this paper (SOLB) mainly refers to some previous highly relevant datasets as the basis for action classification. For example, Bo et al. [13] used a video dataset of students' offline classroom behaviors to include 11 typical behaviors: listening carefully, taking notes, using mobile phones, and so on. Although the dataset is not publicly available, the division of these behaviors is instructive for this study. In addition to this, the dataset constructed by Lin et al. [11] contains 4 behaviors; Wang's [7] dataset contains 8 behaviors, and so on. In the process of constructing our dataset (SOLB), this paper synthesizes the typical behaviors contained in the dataset in each literature, and also asks teachers who use real-time online classroom for instruction. In addition, the research also consider the characteristics of students facing computers during online learning. Ultimately, the authors defined 10 typical online learning behaviors of students, which are showed in the TABLE I.

After clarifying the content of behavioral data to be collected, we invited 58 students as participants to use their phones or computers in their daily environment of online classes (i.e. home or dormitory), and record each behavior for 5 to 10 seconds. The color of the recorded video is in color (RGB), with a resolution of not less than 448 × 448 pixel array. Fig. 2 shows a series of examples of student videos (each one is a frame capturing in the video, corresponding with the behavior in TABLE I). The dataset is abbreviated as SOLB(Student Online Learning Behavior).

TABLE I. LIST OF STUDENT BEHAVIOR IN DATASET SOLB

No.	Student learning behavior
1	Listen normally
2	Take notes with the head down
3	Frown and confuse
4	Turn head to the left
5	Lower the head to play smartphone
6	Speak
7	Yawn
8	Sleep on the table
9	Sleep with hand on the head
10	Drink water

D. CNN-Swin Model

By comparing Transformer[1] and Convolutional Neural Networks(CNN)[3], we constructed a lightweight online learning behavior recognition Neural Network model based on inputs of single-frame images: CNN-Swin.

In the model (shown in Fig. 1), we chose to first use CNN to efficiently extract spatial primary information, and reduce the complexity for the Transformer Encoder by reducing the size of feature maps. Then, the main structure of Swin Transformer[2] is used in the second part of the CNN-Swin model, to obtain deep Self-Attention information of feature maps. The overall structure of the Swin Transformer have been already introduced earlier. This spatial deep feature self-attention mechanism is formed by a group of Swin Transformer blocks. It consists of two stages: the first stage includes six Swin Transformer blocks and one block of patch merging, and the second stage includes two Swin Transformer blocks.

Among them, the Swin Transformer blocks is based on Transformer Encoder of ViT model, with modifications to the attention mechanism. It spatially divides the feature map into multiple patches as tokens by using a predetermined size window, and then calculates the self-attention information for each patch. In addition, the windows of different Swin Transformer blocks have been translated to establish connections between different windows.

Besides the Swin Transformer blocks, to generate a multi-level representation that mimics the pyramid structure commonly found in CNN, at the end of stage 1 in our model, we use the patch merging method to reduce the number of patch tokens and thereby reduce the size of the feature map. The patch merging part first uses 2×2 window to divide the input feature map without overlapping, and then feature vectors of these 4 adjacent patchs, which are in the same window, are concatenated along the channel dimension to obtain a 4C-dimensional vector (the original dimension is C). Then, there's a linear layer to reduce the channel dimension from 4C to 2C. As a result, the number of input tokens for the 2nd stage become ($H/2 \times W/2$). These two stages together generate a multi-level representation.

IV. EXPERIMENT

A. Preprocessing of SOLB dataset

The dataset preprocessing can be divided into five steps, and the flowchart of each step is shown in the Fig. 3.

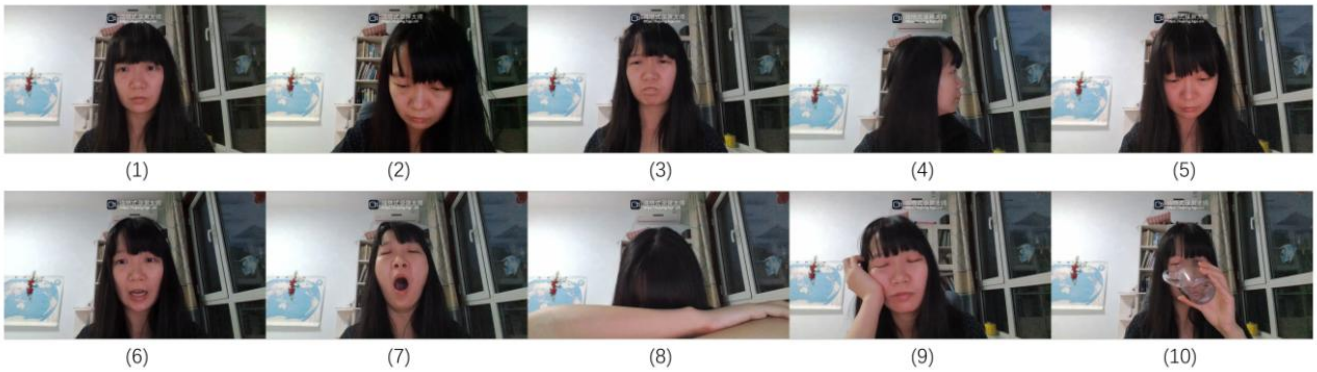

Fig. 2. Examples of images cut in videos of dataset SOLB

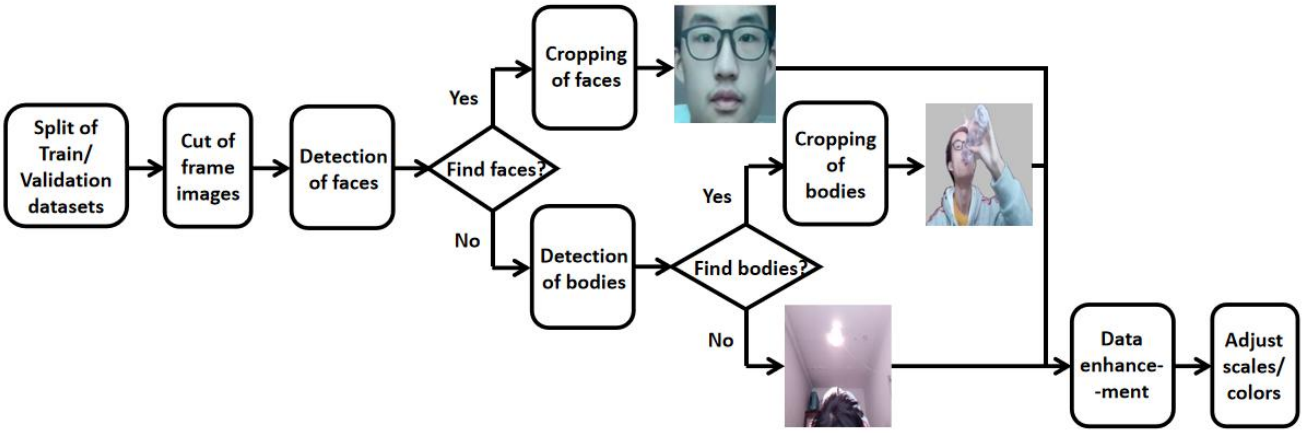

Fig. 3. Flow chart of preprocessing of SOLB

- Divide the 58 subjects in the online learning behavior dataset SOLB into training set subjects and validation set subjects in a ratio of 8.5:1.5.
- Capture images from each video frames. Due to the small differences between some behaviors, we removed the behaviors of "Take notes with the head down" and "Frown and confuse", and only retained the remaining 8 categories.
- Conduct face detection on frame images of each sample in BlazeFace[15]. Then the pre-trained human body model in MediaPipe[17] is also used to recognize human bodies in the images. All faces and bodies will be cropped. In summary, all these images of faces and bodies, and the original frame images which have neither the face nor the body, constitute the image dataset of the input model.
- Enhance the data of the image samples to improve the robustness of the model, including: random color jitter with a certain probability, rotation, affine transformation or translation along the X/Y axis, clarity adjustment, contrast adjustment, exposure adjustment, and brightness adjustment. We only choose one or two methods but not all of them, for saving some time of training.
- Scale each image to 256×256 pixels, then cut a 224×224 pixel area in the center, and standardize RGB 3 channels to an average value of [0.485, 0.456, 0.406], with a standard deviation of [0.229, 0.224, 0.225]. As a result, the input scale of images are 224×224 pixels.

B. Model pre-training

We originally planned to use the ImageNet-1k dataset[4] for pre-training of image classification tasks. But due to hardware limitations, in order to shorten training time, we randomly selected 200 categories of objects from the 1000 categories in ImageNet-1k as the dataset for model pre-training. We call this dataset ImageNet-200 and pre-train the CNN-Swin model on it for image classification tasks. In pre-training, we use AdamW as the optimization algorithm. The AdamW optimization improves the model's generalization ability and decouples the learning rate hyper-parameters and weight decay hyper-parameters, making hyper-parameter optimization easier.

C. Formal training with transfer learning

After pre-traing, we conduct transfer learning on our student online learning behavior dataset SOLB. In transfer learning, we use the same loss function as in pre-training, which is Cross-Entropy. The optimization algorithm is Stochastic Gradient Descent (SGD). The research designs several sets of comparative experiments. The first is to verify the effectiveness of the model and pre-training. Then the research will try to adjust and optimize the hyper-parameters to achieve the best training effect. At the same time, the cost and efficiency of computation will also be put in an important position to measure the model. We can see the detail settings of hyper-parameters in the following of article.

V. Result of Experiment

For the classification effectiveness evaluation of the model, we use the evaluation metrics of Accuracy(ACC), Macro Precision(MAP), Macro Recall (MAR), and Macro F1 Score(MAF).

A. Validation of model and pre-training

According to the plan, we first conducted a test of the effectiveness of CNN-Swin itself and the pre-training of ImageNet-200. By the way, we make general adjustments by modifying hyper-parameters. In TABLE II and TABLE III, we have temporarily changed parameters of epochs, batch size, and initial learning rate. These parameters are independent of the model structure. In the following, we will see adjustments and experiments on the model structure itself.

TABLE II. ACCURACY RESULTS OF DIFFERENT HYPER-PARAMETERS

No.	Epochs	Batch size	lr_{init}[a]	ACC
1	50	16	0.02	43.04%
2	100	16	0.02	44.36%
3	100	32	0.02	36.38%
4	100	16	0.05	40.40%

[a] lr_{init}: initiative learning rate

TABLE III. COMPARISON OF ACCURACY BETWEEN MODEL WITH PRE-TRAINING AND NO PRE-TRAINING

Model	ImageNet-200	SOLB
No pre-training	-	44.36%
With pre-training	80.88%	90.42%

It is not difficult to see that after the epochs reach more than 50 rounds, the accuracy of the model will not change significantly (±2.98%). Similarly, making minor changes to the initial learning rate will not have a significant impact on the accuracy of the model (±8.93%). However, if the batch size is set too large, it will lead to a decrease in accuracy (-17.99%). Therefore, the parameter setting of group 1, 2 and 4 is basically reasonable.

For the test of pre-training, we've gotten the accuracy of pre-training on Imagenet-200 dataset and transfer learning on SOLB. It can be seen that pre-training significantly improved the classification performance of the model, with an accuracy of about 2 times (+103.83%) compared to no pre-training case. In Fig. 4, we can see a student yawning (left) and a student turning his head to the left (right). Both images are from the video of the SOLB dataset. To show the success of the training, we mark the detected faces in the recognition process with squares. In this example, we mark the probability of detecting a face is greater than or equal to 80% with a green square, while the probability is less than 80% with a red square. (Of course, we are more concerned about the accuracy of classification, but not positioning the face, so these examples are only for showing our training results.)

Fig. 4. Examples of detection results after training

B. Discussion of structural hyper-parameters

During this part, we adjusted the hyper-parameters which are related to the model structure, and conducted more detailed statistics. The TABLE IV shows the ACC, MAP, MAR, and MAF of the CNN-Swin model with different hyper-parameters. The first three groups of models did not undergo pre-training, while the latter three groups did. An obvious conclusion is that adding pre-training significantly improves the performance.

Comparing respectively both inside the first three groups of data and the last three groups of data, it can be seen that reducing one layer of convolution in the early stages of the model has a significant impact on the classification performance, as the spatial size of the feature map input to the Swin Transformer block is still large, making it difficult in extraction of features. At the same time, increasing the number of Swin blocks improves the classification performance, but due to limitations in the dataset and batch size, the improvement is limited. However, it cause a huge augmentation of calculation cost - we'll see that in the following of article.

The TABLE V shows the number of parameters and computation every block in different hyper-parameter CNN-Swin models. The "3" and "4" in the header represent the number of convolutional layers, and the "[6,2]" and "[18,2]" represent the number of Swin blocks in each stage.

By comparison of groups above, it can be seen that in the operation process, the number of parameters of the 3-layer convolutional model is almost the same as that of the 4-layer convolutional model, but the computation amount is about 3.8 times that of the 4-layer convolutional model. Therefore, it can be seen that the 3-layer convolutional model is too heavy to be used. At the same time, compared to using 18 Swin blocks in the first stage, using only 6 blocks is more economical, as the number of parameters and computation of 18 blocks far exceeds that of 6 blocks. The lower part of the TABLE V shows the Throughput with different hyper-parameters on 1 GPU or 1 CPU single thread, with batch size of 1 or 64. It's clear to see that the 2nd group (4, [6,2]) has the fastest prediction speed among the three models.

TABLE IV. STATISTICS OF DIFFERENT STRUCTURAL HYPER-PARAMETERS

Pre-training dataset	N_{conv}[a]	N_{swin}[b]	ACC	MAP	MAR	MAF
-	3	[6,2]	37.84%	40.71%	39.17%	36.32%
-	4	[18,2]	44.70%	54.17%	52.44%	50.13%
-	4	[6,2]	44.36%	52.12%	48.53%	47.47%
ImageNet-200	3	[6,2]	79.32%	80.75%	81.12%	79.19%
ImageNet-200	4	[18,2]	90.56%	91.49%	90.82%	90.67%
ImageNet-200	**4**	**[6,2]**	**90.42%**	**91.40%**	**90.19%**	**90.59%**

[a] N_{conv}: Number of convolution layers
[b] N_{swin}: Number of Swin in each stage

TABLE V. COMPUTING RATE AND COST OF DIFFERENT STRUCTURAL HYPER-PARAMETERS

Group No.		1	**2**	3
Model setting		3, [6,2]	**4, [6,2]**	4, [18,2]
Parameter amount (M)		6.75	**6.76**	12.12
Computing amount (GFLOPs)		3.21	**0.84**	1.93
Throughput (GPU) (s-1)[b]	***BS*[a]*=1***	100.44	**96.43**	42.08
	BS=64	1405.89	**4686.69**	1934.32
Throughput (CPU) (s-1)	***BS=1***	5.31	**16.18**	6.67
	BS=64	3.51	**16.47**	6.39

[a] BS: Batch sizes for each GPU or CPU
[b] Unit of Throughput: number of samples / sec

Overall, the hyper-parameter setting of the 6th group in TABLE IV (that is the 2nd group in TABLE V, both in bold letters) is the most reasonable one, with 4 convolutional layers, 6 Swin blocks in the first stage, and 2 Swin blocks in the second stage. It has a wonderful balance between accuracy and cost, and is finally adopted in this paper.

C. Comparison with SOTA models

We finally conduct experiments to verify the out-performing of CNN-Swin by comparing with other SOTA models in the field of CV.

In order to evaluate the performance of the CNN-Swin model proposed in this paper, the researchers first selected two representative lightweight models in the image classification task as the baseline models for comparison. They are both one of the commonly used neural networks, MobileNet V3[19] and LVT[20], respectively. MobileNet V3 is a lightweight neural network proposed by the Google team in 2019, which is optimized compared to MobileNet V2 and proposes several innovations. Its principle is similar to that of CNN, which uses multiple convolutional layers for feature extraction and model training, and researchers believe that it can be used as an important lightweight network reference group for CNN. The network has achieved good performance in image classification, object detection and semantic segmentation. LVT is a lightweight ViT variant proposed by Johns Hopkins University in 2021, in order to reduce the computational complexity of the Transformer mechanism and make it more efficient to run on the mobile side. So the author bilieve it is a good reference group in the category of Transformer.

In addition, a network with different characteristics, RegionViT[24], is selected as the third baseline model. The network is also a variant of ViT, but of the heavyweight type. It achieves higher accuracy by sacrificing computational cost and time, as will be reflected in the experimental results. This network will serve as another counterpoint to the other three, especially the model we propose, to demonstrate the efficiency of our proposed model.

All the baselines are pre-trained on ImageNet-200 and trained on SOLB. The followings are their results.

The training effects of CNN-Swin, MobileNet v3, LVT and RegionViT models are shown in Fig. 5. By comparing the Accuracy, MAP, MAR, and Macro F1 values of different network models, we can find that CNN-Swin outperforms the other two lightweight benchmark models without significant shortcomings. In this regard, among the lightweight models, the CNN-Swin proposed in this paper has advantages over the two benchmark models. For the heavyweight model RegionViT, we can see that its prediction accuracy is slightly higher than that of the other three models, including CNN-Swin.

The four histograms in Fig. 6 are models of CNN-Swin, MobileNet v3, LVT and RegionViT parameter amount, computing amount, and the throughput(number of predicted samples per second) at a batch size(bs) of 1 and 64 on one CPU. The smaller the first 2 values, the better, and the larger the throughput, the better. Among the three lightweight networks, the parameter amount of the CNN-Swin model proposed in this paper is greater than that of the two lightweight benchmark models. However, CNN-Swin has less computing amount than LVT, and therefore CNN-Swin is faster on the CPU than LVT. MobileNet has the least computation among the three lightweight networks, but when the batch size is 1 on the CPU, the CNN-Swin model still predicts faster. Here, we can aslo see the shortcoming of heavyweight models (like RegionViT), i.e., the computational cost is too high. RegionViT is an order of magnitude higher than the lightweight model in the first two metrics, and much lower than the lightweight model in terms of throughput. By comparing it with the heavyweight model, we can also better understand the superiority of CNN-Swin, which greatly reduces the cost of computation with little reduction in accuracy.

In summary, compared with other benchmark models, experiments show that the proposed CNN-Swin model has efficient classification performance on SOLB datasets.

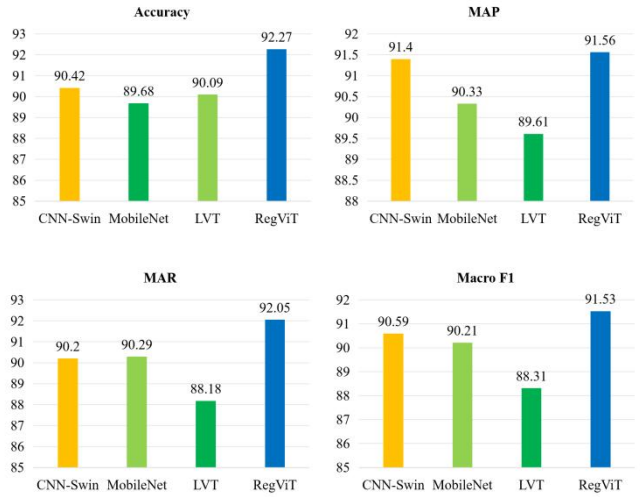

Fig. 5. Comparison of classification performing among models

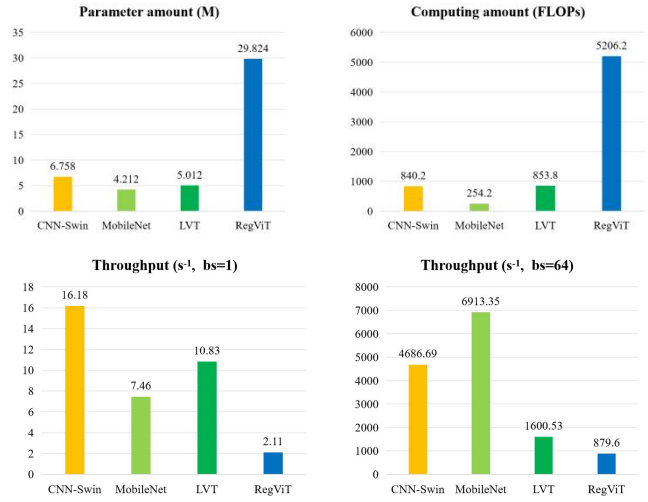

Fig. 6. Comparison of computation cost among models

D. More discussion of the SOLB dataset

The data collected by the dataset (SOLB) implemented in this study has obvious group characteristics. The sample is mainly from undergraduate and graduate students aged 18-25, all from Beihang University. The gender ratio of the sample is relatively balanced, but the ethnic group is entirely East Asian.

Because of the above sample characteristics, the researchers think that collected samples may have a few limitations, such as the function of facial recognition, which may be more suitable for students in East Asia. Similarly, since the sample is all from university students, there may be a bias in face or human's body recognition for younger people (middle or elementary school students). These possible biases are the direction in which this dataset can be improved in the future. From this, we can propose more possibilities for further exploration - to make the learning of the model more accurate through a wider collection of learning videos from groups of students of different races and ages.

VI. Conclusion

This research work is based on the understanding of the needs of the real society and the investigation of existing research results. It aims to build a student behavior category video dataset SOLB in the actual online learning environment for classification of students' behaviors in online learning. A deep Neural Network model, namely CNN-Swin, is designed. It takes single-frame images as input and can provide the discrimination results of student expressions and actions. It first uses a series of convolutional layers[3] to efficiently extract spatial primary feature, and then uses the Swin Transformer[2] based on local Multi-Head Self-Attention mechanism[1] to calculate spatial deep information. Through pre-training using the ImageNet-200 dataset, the accuracy and various statistical data of the model can reach a high level. In the experimental verification on the SOLB dataset, the Accuracy of classification of students' behavior of 8 categories is 90.42% with pre-training, while other benchmarks such as MAP, MAR and MAF maintain out-performing. Compared with the existing research on classroom behavior status, this article is the first time to carry out recognition task on online learning, and the Neural Network model designed is the first time to use Self-Attention mechanism in this kind of task, and achieved good classification results.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 72134001, No. 62377002).

References

[1] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[2] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[3] Gu J, Wang Z, Kuen J, et al. Recent advances in convolutional neural networks[J]. Pattern recognition, 2018, 77: 354-377.

[4] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International journal of computer vision, 2015, 115: 211-252.

[5] Abdallah T B, Elleuch I, Guermazi R. Student behavior recognition in classroom using deep transfer learning with vgg-16[J]. Procedia Computer Science, 2021, 192: 951–960.

[6] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[C]//Bengio Y, LeCun Y. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. 2015.

[7] Wang R, Zhang G, Zhang F, et al. Student behavior recognition in remote video classrooms [M]. 2021: 496–504.

[8] Liu H, Ao W, Hong J. Student abnormal behavior recognition in classroom video based on deep learning[C]//EITCE 2021: Proceedings of the 2021 5th International Conference on Electronic Information Technology and Computer Engineering. New York, NY, USA: Association for Computing Machinery, 2021: 664–671.

[9] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arxiv preprint arxiv:1804.02767, 2018.

[10] Liu H, Liu Y, Zhang R, et al. Student behavior recognition from heterogeneous view perception in class based on 3-d multiscale residual dense network for the analysis of case teaching[J]. Frontiers in Neurorobotics, 2021, 15.

[11] Lin F C, Ngo H H, Dow C R, et al. Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection[J]. Sensors, 2021, 21(16).

[12] Xie Y, Zhang S, Liu Y. Abnormal behavior recognition in classroom pose estimation of college students based on spatiotemporal representation learning[J]. Traitement du Signal, 2021, 38: 89–95.

[13] Sun B, Wu Y, Zhao K, et al. Student class behavior dataset: a video dataset for recognizing, detecting, and captioning students'behaviors in classroom scenes[J]. Neural Computing and Applications, 2021, 33(8): 8335–8354.

[14] Gupta A, D'Cunha A, Awasthi K, et al. Daisee: Towards user engagement recognition in the wild[J]. arxiv preprint arxiv: 1609.01885, 2016.

[15] Bazarevsky V, Kartynnik Y, Vakunov A, et al. Blazeface: Sub-millisecond neural face detection on mobile gpus[J]. arxiv preprint arxiv: 1907.05047, 2019.

[16] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. ICLR, 2021.

[17] Lugaresi, Camillo, et al. "Mediapipe: A framework for building perception pipelines." arxiv preprint arxiv:1906.08172 (2019).

[18] Steiner A, Kolesnikov A, Zhai X, et al. How to train your vit? data, augmentation, and regularization in vision transformers[J]. arxiv preprint arxiv: 2106.10270, 2021.

[19] Howard A, Sandler M, Chen B, et al. Searching for mobilenetv3[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 1314–1324.

[20] Yang C, Wang Y, Zhang J, et al. Lite vision transformer with enhanced self-attention[J]. arxiv preprint arxiv:2112.10809, 2021.

[21] Darcet, Timothée, et al. "Vision transformers need registers." arxiv preprint arxiv:2309.16588 (2023).

[22] Chen, Peng, et al. "Multi-scale Transformers with Adaptive Pathways for Time Series Forecasting." The Twelfth International Conference on Learning Representations. 2023.

[23] Jain, Jitesh, et al. "Oneformer: One transformer to rule universal image segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[24] Chen C F R, Panda R, Fan Q. Regionvit: Regional-to-local attention for vision transformers[C]//arxiv: 2106.02689. 2021.