Finger language recognition algorithm based on ATTAlexnet

Qina Xie¹, Alifu•Kuerban ²*, Minawaer •Abula³ and Dingjie Han⁴

¹ College of Software, Xinjiang University, Urumqi 830046, China
² Key Laboratory of Software Engineering, Xinjiang University, Urumqi 830046, China

*Corresponding author. Email: 506470001@qq.com

Abstract. The mainstream algorithm model structure based on deep neural networks is relatively simple, and it is easy to lose effective information in the pooling layer, the recognition accuracy is not high, and the recognition speed is slow. To solve this problem, a fusion convolutional attention mechanism and AlexNet's Finger language recognition algorithm (ATTAlexNet). By introducing the convolutional attention mechanism in the AlexNet network, it can effectively perform feature learning, realize feature screening, enhance the characterization ability of the network, and introduce the AdderNet to replace the multiplication operation of the convolution layer in the AlexNet network, enhance the robustness of the network, and improve the network calculation speed. The experimental results show that ATTAlexNet is superior to other comparison algorithms, and under the same experimental conditions, the recognition rate of ATTAlexNet is increased by 2.0%, which proves that the ATTAlexNet algorithm can effectively realize finger language recognition, has fast calculation speed, and good robustness.

1. Introduction

Finger language, also known as "fingerspelling", uses finger changes to represent letters, and spells out the expressions of words in sequence according to the phonetic order [¹]. On November 1, 2019, the "Chinese Finger Letter Program" [²] was implemented as the language specification of the National Language Commission, which stipulated a total of 30 basic letter finger language(A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z, ZH,CH,SH,NG). Through this finger language, according to the logic of written expression, the words can be expressed in order to express the meaning of the sentence. Hao Ziyu et al. [³] believe that finger language can not only serve as a single morpheme, but also as a basic word, and divide finger language into three main forms: single letter gestures, letter variant gestures, and letter syllable gestures. The use of finger language in the hearing impaired has increased the expression of sign language, enriched the sign language, and made it more accurate, providing a better way for the hearing impaired to communicate with the hearing healthy. Therefore, finger language recognition has more extensive application value to society. With the development of deep neural networks, sign language recognition as a research task of computer vision has once again ushered in development opportunities.

Finger language is an indispensable part of Chinese sign language. It can not only express sign language alone but also express sign language information with several gestures. Sign language recognition aims to develop algorithms that can correctly classify signal sequences to understand their specific meanings [⁴]. Up to now, traditional computer vision-based sign language recognition methods have certain limitations in the recognition of complex backgrounds and require high environmental
backgrounds, such as Convolutional Neural Networks (CNN) that have outstanding performance in image classification tasks[5] to a certain extent guarantees the invariance of the displacement and rotation of the image, and this performance will gradually increase with the increase of the number of network layers, ensuring the robustness of the classification model. Although CNN's performance on image classification tasks is very prominent, as the number of network layers increases, CNN needs to constantly add layers for creating deep networks. In such an architecture, the positional relationship between high-level features and low-level features will gradually become blurred. Therefore, Hinton et al.[6] proposed a new method in 2012-AlexNet applied to image classification, and achieved good results. Compared with many current networks, it has fewer levels and is simple and easy understanding. However, the introduction of AlexNet has brought new problems, such as slow training speed, low recognition rate, long training time, and too many weights that the file size is large.

In response to the above problems, this paper proposes an AlexNet network model (Attention-based AlexNet, ATTAlexNet) based on the attention mechanism. The introduction of the Convolutional Block Attention Module (CBAM) between the convolutional layers[7] makes the network pay more attention to the spatiotemporal information related to the target, focusing on extracting the effective feature information of the target, and improving the network’s ability to target a complex background. Feature extraction capabilities. The AdderNet[8] is used in the convolutional layer to replace the traditional multiplication rules, which can reduce the accuracy loss in the binary network while ensuring that the calculation does not include the multiplication operation, thereby improving the calculation speed of the network. Experiments show that ATTAlexNet has good recognition performance.

2. Related work

In the early development of sign language recognition technology, hardware devices mainly used sensors to obtain gesture information as input. The computer uses the information provided by the data glove (such as hand shape and degree of bending) to recognize gestures and feed them back to the user. Since the 1920s and 1970s, many research institutions have devoted themselves to the research of data gloves, such as the Sayre Glove studied by Zimmerman in 1978 and the Cyber Glove studied by Kramer in 1989[9]. With the development of computer vision, researchers have found that vision-based gesture information capture can more accurately complete gesture detection and recognition. However, vision-based recognition algorithms are very complex, require extremely high image clarity, are easily affected by the external environment, and input a large number of gesture samples in advance as a data set for gesture detection. Gestures have rich meanings and randomness, and are easily affected by temperature, background, and operation speed; the accuracy and robustness of sign language recognition based on machine learning models have entered a bottleneck period. People gradually turn their attention to deep learning and try to use deep learning methods for sign language recognition. In 2020, Zhou Lei et al.[10] proposed a finger language recognition technology based on R-FCN. By detecting and locating finger language, use online hard case mining technology to learn difficult cases that are easy to identify and improve. Improve the recognition rate of finger language in complex backgrounds.

3. ATTAlexNet model

This section introduces the ATTAlexNet network structure of this article in detail. By introducing a convolutional attention mechanism into the network structure to perform feature screening, enhance the characterization ability of the target feature, and use AdderNet to replace the multiplication operation in the convolution layer without reducing the network level to mine features and improve the calculation speed of the network. The network model is shown in Figure 1. The network model has five convolutional layers and three fully connected layers. The convolutional layer is used for feature extraction, and the number of channels is 3, 96, 192, 384, and 256. The CBAM module combines the space and channel attention mechanism module. Its function is effective channel attention, and the space attention module focuses on which functions are meaningful. The fully connected layer is used to distinguish the background and the target. In this network structure, in order to solve the problem that a large number
of neurons cannot be activated due to sparseness, LeakyReLU is used instead of the ReLu activation function to ensure that the value range of the feature map is within a reasonable range and to enhance the robustness of the network.

Figure 1. Attention based AlexNet.

On the basis of the AlexNet model, the attention mechanism is added after the convolutional layer 1, 2, 3, 4, and 5 respectively, and 5 sets of comparison experiments are performed, and the running time comparison is shown in table 1 and the models in Figure 2 are obtained. The comparison curve of accuracy rate changing with the number of iterations.

Table 1. Different convolutional layer running time.

| convolutional layers | Conv1 | Conv2 | Conv3 | Conv4 | Conv5 |
|----------------------|-------|-------|-------|-------|-------|
| Running times/h      | 2.6   | 4.5   | 2.8   | 3.4   | 3.8   |

Figure 2. Numbers of iterations for different convolutional layers accuracy.

It can be seen intuitively from Table 1 that the running time of the Conv2 layer is close to twice the number of other layers. The main reason is that after the convolutional layer "flattens" the features, there are too many parameters that the calculation is too large. In combination with Figure 2, it can be seen that Conv3 is the number of network layers suitable for the data set in this paper, and the accuracy rate of higher or lower than this layer will decrease. Although the accuracy of the Conv2 layer is almost the same as that of Conv3, and because of the shallow number of layers, the extracted features are simple, and the accuracy converges to 64.83% in 10000 steps, but its convergence time is twice that of other layers, which is too expensive. In contrast, choosing the features of the Conv3 layer is more
representative. In order to fully extract the features, this paper adds an attention mechanism behind
the Conv3 layer and inputs the 13*13*192 feature map data extracted by Conv3 into the attention module.
First, the input feature maps are passed through the global width and height respectively. After the
maximum pooling operation and the global average operation, it passes through the multilayer
perceptron (MLP). The features output by the MLP are subjected to an element-based weighting
operation and then activated through the sigmoid function to generate the final channel attention feature
map. The channel attention feature map and the input feature map input by this module is subjected to
element multiplication operations to generate the input features required by the spatial attention module.
The channel attention mechanism expression is:

\[ M_c(F) = \sigma \left( W_0 \left( \frac{F}{W_{\text{avg}}} \right) + W_6 \left( \frac{F}{W_{\text{max}}} \right) \right) \]

Among them, \( \sigma \) is the Sigmoid operation, \( r \) is the reduction rate, and \( W_0 \) is followed by ReLU
activation.

The feature map output by the channel attention module is used as the input of the spatial attention
module. Similar to the spatial attention module, first do two pooling operations, then combine the results
based on channels, then go through a convolution operation to reduce to one channel, and then generate
a spatial attention module through Sigmoid. Finally, the feature is multiplied with the input feature of
the module to generate the final feature.

\[ M_s = \sigma \left( f^{7*7}[\frac{\text{AvgPool}(F)}{W_{\text{avg}}}, \frac{\text{MaxPool}(F)}{W_{\text{max}}}] \right) \]

Among them, \( 7*7 \) represents the size of the convolution kernel.

Compared with the traditional multiplication-based convolution operation, the network proposed in
this paper uses an AdderNet to replace the multiplication operation of the convolutional layer, so that
the network convolutional layer can obtain a network with the same structure with a multiplication
operation with almost no multiplication. Almost the same accuracy rate can effectively solve the
problems of high energy consumption and slow calculation speed caused by the traditional convolutional
layer requiring a large number of multiplications.

4. Experiment

4.1. Experiment preparation

This paper uses self-collected finger language data set based on portable devices(Figure 3) and finger
language data set based on virtual humans(Figure 4). This data set contains 10 finger languages
(respectively A, B, C, D, E, F, G, H, I, J), and each finger language is collected from three directions
(left view, front view, Right view) and mark the finger language. At the same time, in order to make the
scale of the data set larger and increase the versatility of the network model, the data set is also expanded
using data enhancement methods.

![Figure 3. Finger language data set based on portable devices.](image)

![Figure 4. Finger language data set based on virtual humans.](image)

The experiment in this article is running under the environment of the Pytorch platform built on the
CPU @3.20GHz, 16GB memory, and Windows10. The experimental environment configuration
parameters are shown in table 2.
Table 2. Configuration parameters of the experimental environment.

| parameter name | Version or value |
|----------------|------------------|
| operating system | Windows 10 |
| CPU | CPU @3.20GHz |
| GPU | NVIDIA GeForce GTX 1080Ti/16G |
| CUDA | CUDA 10.1 |
| Pytorch | Pytorch1.4 |

Table 3. Comparison with existing models under the same number of experimental steps.

| Network model | Accuracy /% | Running time/h |
|---------------|-------------|---------------|
| ResNet101     | 93.03       | 3.38          |
| VGG19_bn      | 93.89       | 4.03          |
| Ours          | **96.32**   | **2.45**      |

Note: Bold font is the best result

4.2. Experimental comparative analysis

First, conduct a comparative experiment under the same number of experimental steps. On the finger language data set of this article, the model of this article is compared with ResNet101 and VGG19_bn, and the accuracy and running time of each model after 10,000 runs are analyzed. The comparison results are shown in the table 3.

It can be seen from table 3 that the running time of this method when the number of running steps reaches 10000 times is about 2.5h, which is 3/4 and 3/5 of ResNet101 and VGGNet19_bn. The main reason is that this paper uses addition to replace the multiplication operation of the convolutional layer. The calculation speed can be improved. The accuracy rate reaches 96.32%, which is 3.29% and 2.43% higher than ResNet101 and VGGNet19_bn. The main reason is that the latter are large deep network models with large training parameters and time-consuming. Secondly, these large networks require data quantity. Strictly, when the amount of data does not reach a certain scale, over-fitting is prone to occur, and the extracted features cannot express finger language features well.

In order to verify the applicability of the method in this paper, add salt and pepper noise to the data, and then use three methods to classify and recognize it. The experimental results show that the recognition rate after adding noise to the data does not decrease significantly, which proves that the model in this paper has good robustness. Awesome. The experimental results are shown in table 4. It can be seen that the recognition rates of ResNet101 and VGG19_bn are slightly lower than those of the ATTAlexnet network model on both noisy and no-noise recognition tasks. The highest recognition rate for single finger language reached 100%.

In order to ensure that the Attention method proposed in this article is universal for improving the recognition rate of experimental results, the traditional AlexNet method and the method of this article are applied to the dataset dogs vs cats and the sign language dataset. The experimental results obtained are shown in the following table:

As can be seen from the above table, on the two data sets, the recognition rate based on the ATTAlexNet network structure has improved to a certain extent, both by about 5%. The experimental results prove that adding an attention mechanism to the traditional AlexNet network and using the AdderNet to replace the multiplication operation of the convolutional layer can emphasize the effective features on the basis of appropriate feature Dropout and ignore irrelevant information, reducing data the discarding of information improves the accuracy of experimental results.
Table 4. Comparison of experimental results of three methods.

| Network model | Average accuracy | Single highest accuracy rate |
|---------------|-----------------|-----------------------------|
|               | No noise        | noise                       |
| ResNet101     | 93.03%          | 92.89%                      | 100%                          |
| VGG19_bn      | 93.89%          | 93.78%                      | 100%                          |
| AttAlexNet    | 96.32%          | 96.22                       | 100%                          |

Table 5. Recognition rate of two methods on three data sets.

| Datasets     | Experimental method |
|--------------|---------------------|
|              | AlexNet | AttAlexNet |
| Cats vs dogs | 81.67%  | 91.40%     |
| figure       | 94.29%  | 97.43%     |
| figures      | 90.67%  | 96.22%     |

5. Conclusion

In this article, we merge the AlexNet and Attention models, use convolutional structure, overlapped maximum pooling, LRN, data enhancement and attention mechanism algorithms, and replace the multiplication operation of the convolutional layer in AlexNet with the addition operation to achieve static finger language Recognition. Experimental results show that based on the ATTAlexNet network structure, good results can be obtained on finger language recognition tasks. The highest average recognition rate can reach 95.78% under the condition of adding noise, especially when some finger language is similar, it still achieves high accuracy. Rate shows that the algorithm that adds attention mechanism to the AlexNet network has better performance. Sign language includes four characteristics: hand shape, position, direction, and trajectory. Finger language is only part of the hand shape of sign language. In future research work, we will use attention mechanism to integrate AlexNet to realize complete sign language recognition based on multiple features of dynamic sign language.

References

[1] Tang Lingyan. Research on the selection and conversion of language codes in the communication of the hearing impaired [D]. Shanghai: Special Education of East China Normal University(2007).
[2] Ministry of Education of the People’s Republic of China, National Language Working Committee, China Disabled Persons’ Federation. Chinese Finger Language Alphabet Project [M]. Beijing: Huaxia Publishing House(2018).
[3] Hao Ziyu, Alifu·Kurban, Li Xiaohong, et al. Chinese finger language recognition based on CapsNet[J]. Computer Application Research. (2019), (10): 3157-3159.
[4] NETO G M R , JUNIOR G B , JO ÑAODALLYSON SOUSA DE ALMEIDA, et al. Sign Language Recognition Based on 3D Convolutional Neural Networks[M]/Image Analysis and Recognition. Springer, Cham, (2018).
[5] Yi Jingguo,Cheng Jianghua,Ku Xishu.A review of visual gesture recognition[J].Computer Science.2016,(S1):103-108.
[6] Krizhevsky, Alex & Sutskever, Ilya & Hinton, Geoffrey. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems. 25. 10.1145/3065386.
[7] Woo, S. et al. “CBAM: Convolutional Block Attention Module.” ECCV (2018).
[8] H. Chen et al., "AdderNet: Do We Really Need Multiplications in Deep Learning?," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, (2020).pp.1465-1474.doi:10.1109/CVPR42600.2020.00154.
[9] Dai Yukun. Research on Gesture Recognition Method Based on Wearable Data Gloves [Master's Thesis]. South China University of Technology, Guangzhou(2019).
[10] Zhou Lei, Alifu·Kurban, Lu Qingshen, et al. Chinese finger language recognition based on R-
FCN[J]. Journal of Xinjiang University: Natural Science Edition (Chinese and English). (2020), (2):170-176.