Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/15/20230847

Cross-data recognition on sign language letters based on convolutional neural networks

Zhixi Zhu

The Affiliated High School of South China Normal University, Guangzhou, China

zhuzx.cicely2021@gdhfi.com

Abstract. With an estimated 1.5 billion hearing-impaired people globally, sign language is a vital means of communication among them. Meanwhile, the complexity and diversity of sign languages across different regions bring challenges to researchers. As existing studies in this field usually lack the implementation of recognition models with multiple datasets, which limits their practical value in real-life scenarios, this study intends to get around this constraint. This work constructs a baseline American Sign Language Letter Recognition model using Convolutional Neural Network (CNN) and then optimizes it to enhance its ability. Finally, cross-data recognition is carried out. In order to train and evaluate the model, this study gathers data from several sources, including the MNIST sign language set and real-life photos. It also investigates how data augmentation affects recognition ability. Consequently, the CNN model is able to recognize hand gestures for different alphabetic letters in solid backgrounds. Its accurate rate of it reaches about 99.83%. For the extra real-life dataset, which contains 96 images, the model could still detect the majority of the images with sign language equivalents, despite some accuracy loss in more crowded backgrounds. This work, in general, concentrates on the potential of CNNs for sign language identification and highlights the significance of cross-data recognition in creating useful recognition models.

Keywords: computer vision, convolutional neural networks, sign language recognition.

1. Introduction

Sign language is a form of communication that utilizes hand gestures, having practically the same properties and functions just as spoken languages. It is widely used within the hearing-impaired community. Globally, there are currently more than 1.5 billion people suffering from hearing loss, which is predicted to reach 2.5 billion in 2050 [1]. Among all those individuals, 430 million are confronted with hearing disabilities, according to the World Health Organization [1]. Given these statistics, it is plausible to assert that there is presently a substantial population of hearing-impaired persons, who are likely to encounter a multitude of communication barriers in their daily lives since individuals who hear typically do not acquire the ability to comprehend sign languages. To overcome the obstacles in communication between deaf and hearing people, sign language recognition, with the help of technologies, has become a research focus among researchers. Due to an increasing need for convenience in real-time communication, explorations into this field can be of high practical value on various occasions, especially for sign language users to convey their messages more efficiently.

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/15/20230847

Studies in this field are important for bridging the communication gap by helping hearing people understand sign language, whereas challenging due to sign languages' complexity and diversity [2]. Similar to spoken languages, sign languages differ completely across countries, and even among distinct regions within the same nation. Moreover, many sign language categories, like Chinese sign language, involve dynamic gesture expressions, which make recognition tasks harder. As a result, to develop practical and accurate methods to recognize and translate sign languages, researchers need not only to collect a large amount of data but also to design and keep optimizing algorithms.

In terms of different methods used, there are currently two main categories of sign language recognition: data-glove based, and computer vision based. As the latter better simulates daily scenarios in which people are not wearing special gloves for communication, a computer vision-based data source was chosen for this study. Computer vision-based recognition is based on video or image data collected by input devices such as cameras. Then the recognition model is constructed mainly using deep learning, which has experienced a surge in recent years due to its prominent capabilities in handling complicated data from different sources [3]. Convolutional Neural Networks (CNNs) were deemed suitable for constructing a sign language recognition model due to their robust feature extraction capabilities. Other algorithms, like 3-Dimensional Convolutional Neural Networks, Recurrent Neural Networks, and YOLO are also introduced later by many researchers to carry out recognition tasks [4]. However, it is worth noting that sign language recognition has not received as much attention as other deep learning subjects. Moreover, existing studies usually lack the implementation of recognition models with multiple datasets, which means researchers often focus on the accuracy of their models in predicting the same dataset; the pictures in that dataset are likely to be taken under rather ideal circumstances: a neat background and high resolution. To enable practical sign language recognition, it may eventually be necessary to use data from other sources and test the model in a more realistic setting, while the main limitation of existing studies is their failure to do cross-data recognition.

To tackle the problem mentioned above, this study uses Convolutional Neural Network to construct a baseline American Sign Language Letter Recognition model and then carries out optimization to enhance its ability to carry out cross-data recognition. Data from different sources are collected and used to train and test the model. This study also explores the effects of data augmentation in the optimization process. Consequently, the improved CNN model can recognize the hand gestures for different alphabetic under ideal conditions like solid background colors. In daily contexts, when the backgrounds are much more cluttered, the model loses some accuracy when tested but can still correctly recognize most of the pictures given, which contain sign language representations.

2. Method

2.1. Dataset preparation

In this section, the composition of datasets and the pre-processing methods will be introduced.

The training dataset and the major testing dataset are obtained from Kaggle [5], including 21,455 cases used for training and 7,172 cases used for testing, representing 24 classes of alphabetic language letter (J and Z are excluded because of dynamic expressions). All the data is stored in CSV files, marked with labels. Each case is a 28*28-pixel grayscale image. Figure 1 presents the first 5 training cases.

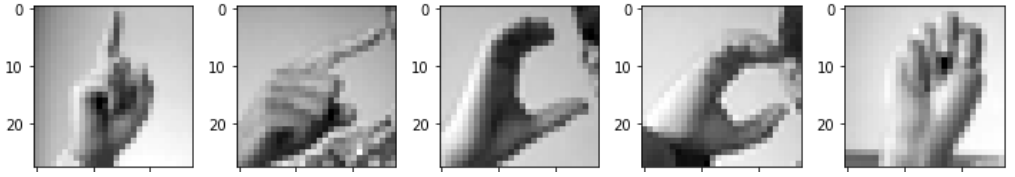


Figure 1. Sample training data.

It is worth mentioning that, in this study, both the original testing dataset corresponding to this training set and an extra testing dataset are utilized to evaluate the ability of this CNN model. The extra

DOI: 10.54254/2755-2721/15/20230847

testing data set is composed of 96 images, 4 for each alphabetic letter, which are collected in real-life situations. This is because the implementation of sign language recognition tool in daily life is rather likely to be accompanied with cluttered backgrounds instead of ideal solid colors. By doing this, this study further examines the proposed CNN model's performance in practice.

To prepare for training and testing, all data, especially the real-life images, is firstly normalized: resized and converted to 28*28-pixel grayscale images, which are consistent with the format of the provided training dataset on Kaggle. Then the training data is augmented by rotating, zooming, and shifting a little bit, to increase the amount of data and avoid overfitting, hence better training the model and improving its capability to do recognition.

2.2. The proposed CNN model

A Convolutional Neural Network (CNN) is a type of neural network that takes advantage of convolutional structures to do feature extraction [6]. CNNs were considered suitable for constructing image recognition models due to their robust feature extraction capabilities. Many computer-vision-based recognition tasks, like face recognition, have been achieved with CNNs [6]. Typically, besides input and output, a CNN model is composed of three major types of layers: convolutional layers, pooling layers, and fully connected layers. In a convolutional layer, a filter, which is spatially smaller than the original image, shifts across the image and carries out operations, creating feature maps. A pooling layer helps reduce spatial information, hence rendering the model less complex and reducing the risk of overfitting. Finally, a fully connected layer builds links between previous layers and the output.

In this work, a CNN model with three convolutional layers is built after accepting grayscale images with a resolution of 28*28 pixels, and the filter size for each layer is set to be 3*3. A max-pooling layer of size 2*2 follows each convolutional layer. After accepting the data from the previous convolutional layer, it reduces the dimensions of feature maps. To avoid overfitting, a dropout layer with a rate of 0.2 is placed next to the pooling layer. The multidimensional output from the preceding layers is then flattened and entered into the fully linked layer after these. In the end, the CNN model can recognize and output the corresponding alphabetic letter for each sign language gesture in the input data.

2.3. Implementation details

This part describes some relevant details in this study.

TensorFlow and Keras are used to build this recognition model since their convenience has been proved by many studies [7-9]. The model is trained across 50 epochs, with each batch with a size of 128. The patience parameter for learning rate reduction is set to 2, which implies that if the validation accuracy does not increase after two epochs, the learning rate will be automatically reduced. The Adam optimizer, which is a computationally efficient method for gradient-based optimization, is used in this model [10]. And the categorical cross-entropy is selected as the loss function. Furthermore, to optimize the model and enhance its performance on the test datasets, several changes are made to the model's parameters. All loss and accuracy data are recorded in the model's history during the training phase. Finally, the provided dataset and additionally collected photos are used to test the model.

3. Results and discussion

Data augmentation is done on the original training dataset to avoid overfitting and increase the amount of data. The Figure 2 and Figure 3 below show the training and validation loss and accuracy during model training. The loss diagram illustrates that there is a rapid decrease in training and validation loss at the beginning, then these figures both approaches to zero. And in the accuracy graph, both training and validation accuracy rise to approximately one at the end of training. The trends shown in these two graphs imply the parameter settings are appropriate.

DOI: 10.54254/2755-2721/15/20230847

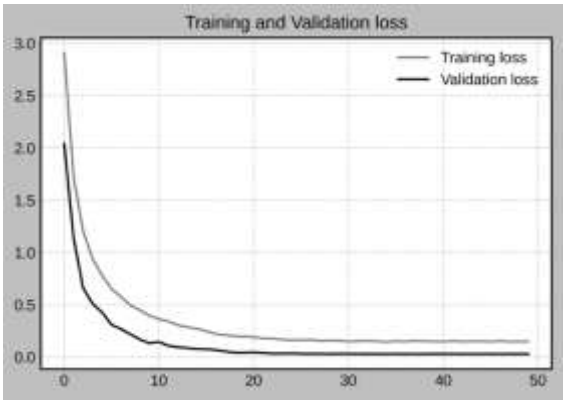


Figure 2. Loss diagram.

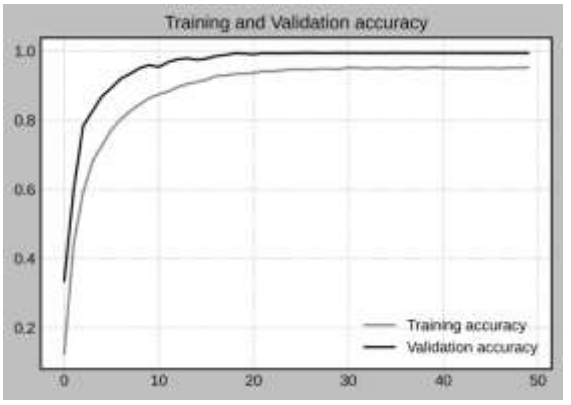


Figure 3. Accuracy diagram.

After testing the model using the provided datasets on Kaggle, the results reveal that the accurate rate of this CNN model in recognizing fingerspelling letters from the corresponding test cases on Kaggle reaches approximately 99.83%. From this, it can be speculated that the model's ability in recognizing fingerspelling gestures under ideal conditions is strong.

Then the model is tested with 96 real-life photographs, it recognizes 81 of them correctly, which means the precision rate of it drops to around 84.38%. This reveals that the model can still accurately recognize the majority of testing photographs, but the cluttered backgrounds negatively affects its performance, to some extent.

The Table 1 below compares its performance on different datasets.

Table 1. Results comparison.

	Provided testing dataset	Collected real-life photos
Recognition accuracy	99.83%	84.38%

The testing result suggests that the model is highly acute when performing recognition tasks under ideal conditions—tiny backgrounds and clearly presented gestures. Nevertheless, when faced with real-life photos with more complex backgrounds, the accuracy rate experiences an acceptable drop. This also points out that future studies may need to concentrate on improving the recognition methods' practical value.

Comparatively, this study discusses both the model's performance on the original dataset, which is rather high, and in real-life scenarios, which is relatively lower. It tackles the constraints in previous studies which did not consider the practical utility of the recognition model.

4. Conclusion

In this work, CNN is selected as the algorithm to implement American sign language recognition based on computer vision. Datasets are collected from online resources and real life, to overcome the limitations that many previous researchers did not take the practical utility of the model into account. The results show that the recognition capability of this model in solid color backgrounds is high, while it loses some accuracy when recognizing real-life photographs taken in more cluttered backgrounds. This suggests that the model needs further adjustments and optimizations to truly put into daily-life use. To improve this model, future researchers may need to consider designing better models by combining the baseline CNN algorithm with other methods like LSTM, or adopting other algorithms like 3D-CNN, and YOLO. Furthermore, as the situations of daily communication are likely to be much more complex than training data contained in existing sign language datasets, collecting a wider range of data in real-life conditions may hone the trained model's recognition capability in practice.

DOI: 10.54254/2755-2721/15/20230847

References

[1] World Health Organization 2023 Deafness and Hearing Loss
[2] Suharjito and Gunawan H and Thiracitta N and Nugroho A 2018 Indonesian Association for Pattern Recognition International Conference pp 1–5
[3] Voulodimos A and Doulamis N and Doulamis A and Protopapadakis E 2018 Computational Intelligence and Neuroscience pp 1-13
[4] Li T and Yan Y and Du W 2022 IEEE International Conference on Artificial Intelligence and Computer Applications pp 927-31
[5] Tecperson 2017 Sign Language MNIST (Kaggle)
[6] Li Z Liu F Yang W et al. 2021 A survey of convolutional neural networks: analysis, applications, and prospects IEEE transactions on neural networks and learning systems
[7] Yu Q Wang J Jin Z et al. 2022 Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training Biomedical Signal Processing and Control 72: 103323
[8] Grattarola D Alippi C 2021 Graph neural networks in tensorflow and keras with spektral [application notes IEEE Computational Intelligence Magazine 16(1): 99-106
[9] Ramasubramanian K Singh A Ramasubramanian K et al. 2019 Deep learning using keras and tensorflow Machine Learning Using R: With Time Series and Industry-Based Use Cases in R, 667-688
[10] Kingma D P and Ba J 2015 the 3rd International Conference for Learning Representations (San Diego)