Article

Data Augmentation for Human Keypoint Estimation Deep Learning based Sign Language Translation

Chan-Il Park * and Chae-Bong Sohn *

Department of Electronics and Communications Engineering, Kwangwoon University, Seoul 01897, Korea

* cksdlf78@kw.ac.kr (C.-I.P.); cbsohn@kw.ac.kr (C.-B.S.)

Received: 8 July 2020; Accepted: 4 August 2020; Published: 5 August 2020

Abstract: Deep learning technology has developed constantly and is applied in many fields. In order to correctly apply deep learning techniques, sufficient learning must be preceded. Various conditions are necessary for sufficient learning. One of the most important conditions is training data. Collecting sufficient training data is fundamental, because if the training data are insufficient, deep learning will not be done properly. Many types of training data are collected, but not all of them. So, we may have to collect them directly. Collecting takes a lot of time and hard work. To reduce this effort, the data augmentation method is used to increase the training data. Data augmentation has some common methods, but often requires different methods for specific data. For example, in order to recognize sign language, video data processed with openpose are used. In this paper, we propose a new data augmentation method for sign language data used for learning translation, and we expect to improve the learning performance, according to the proposed method.

Keywords: deep learning; data augmentation; sign language; openpose

1. Introduction

Recently, the field of artificial intelligence has been actively applied in the sign language of expressing language by hand, as shown in Figure 1.

Figure 1. Sign language example of the letter i.

In the sign language area, the method of translating sign language recognized by the camera into the language of text or voice is performed through artificial intelligence (AI). Sufficient learning must be done in order for deep learning [1] to properly translate the word into text voice. One of the most important parts of this learning is the amount of sign language data. Sign language data are difficult to obtain. Since sign language is a motion, not a stationary image, the data production process is difficult compared to an image, because data must be produced in the form of a video. In addition, since sign language is not a language that everyone basically understands, such as English or Korean, it is necessary to fully understand the language to produce data. In addition, as sign language has many words, collecting every word is difficult. For these reasons, manufacturing and securing a sufficient amount of sign language data is a great difficulty. Therefore, this paper proposes a method to increase the sign language data set. We expect to improve translation accuracy by increasing the dataset through three methods to increase the sign language data and learning artificial intelligence.

The composition of this paper begins with an introduction, and Chapter 2 describes the research related to the paper. The next chapter describes in detail the data augmentation [2,3] methods for augmenting sign language data and explains how they were applied. In Chapter 4, the performance of the proposed method is verified based on the simulation results by comparing and simulating with augmented data based on the proposed method, and Chapter 5 summarizes the proposed method and experiment. Finally, Chapter 6 dictates the evaluation and opinion of the paper.

2. Related Work

Frame skip sampling was used as a data augmentation method in an existing paper on sign language translation [4]. Frame skip sampling is a random frame skip sampling, which is mainly used when handling video data, and when all frames are divided into n frames, one is extracted and integrated for each section. In the paper described above, frame skip sampling increases the data by 10 to 100 times. The learning process of this paper is shown in Figure 2.

Figure 2. Architecture of sign language translation.

The video increased through data augmentation is converted into a video composed of only keypoints using openpose [5]. After that, the transformed video is trained on the GRU-based encoder decoder model. As a result of the experiment, the accuracy was best when learning by increasing the video by 50 times using data augmentation. Since the random frame skip sampling method is used when dealing with general video data, this paper proposes a data augmentation method that generates data according to the characteristics of the sign language.

3. Materials and Methods

The data augmentation proposed in this paper is divided into the following three types.

1. Finger length conversion;
2. Random keypoint removal;

3. Camera angle conversion.

As a dataset that performs the proposed data augmentation, learning in the process shown in Figure 3 was performed.

Figure 3. Learning process of sign language translation.

Data augmentation was performed by combining one or two of the videos to be trained, camera angle conversion, finger length conversion, and random keypoint removal. Then, we input the increased data into openpose [6] and transform generated keypoint video to frame image. By passing the frame image data through convolutional neural network (CNN) [7,8] for each frame, the feature was extracted [9,10] and the extracted feature was input to the long short-term memory models (LSTM) [11] to learn what the image means [12]. Learning was progressed through the input features and translated into what words the video means. In addition, in order to recognize only sign language motion, a pre-processing process of changing a video to a video composed of keypoints, as shown in Figure 4, through openpose was performed in all data augmentation steps.

Figure 4. Video created with openpose.

The generated video consisted of the upper body, arms and fingers of both hands needed to recognize the sign language. The reason for learning with the above video is that if openpose was successfully performed, the result was not greatly affected by the surrounding environment.

3.1. Finger Length Conversion

Data augmentation of an image of the type shown in Figure 4 is very limited, and the first proposed method was to randomly increase or decrease the finger length. Finger length varies by person. Therefore, by randomly reducing or increasing the length of a finger, a single video can be increased as if taken by various people.

3.2. Random Keypoint Removal

The random keypoint removal method is a randomly removing keypoints from a video created using openpose, as shown in Figure 4. As shown in Figure 5b, the keypoint was randomly removed to appear as if there was no specific node of the finger.

Figure 5. Example of random keypoint removal. (a): original (b): remove keypoint.

The reason for increasing data using a random keypoint removal as above is as follows. Because openpose may not be recognized successfully, we randomly removed keypoints, thereby lowering our learning dependency on openpose. In other words, the recognition rate of openpose was lowered, so the effect of when the correct answering rate of the translation decreased due to the missing keypoint. In addition, as the keypoint randomly disappeared, various videos can be generated, so the generalization of learning can be enhanced.

3.3. Camera Angle Conversion

The data used to learn for sign language translation were in video format. This video was taken using a camera. When shooting with a camera, various angles were possible. Since the size or perspective of the object looks different depending on the shooting angle, various data can be produced by using it for data augmentation. If the shooting angle is changed, the location of the keypoint or the length of the node is also changed in the video shown in Figure 4, so data augmentation can be successfully performed. However, in order to change the camera angle, it is difficult to shoot again. Therefore, instead of changing the angle of the actual camera, the method of processing the photo itself as shown in Figure 6 was used so that the camera's shooting angle was changed.

Figure 6. Example of camera angle conversion. (a): original (b): convert to an overlooking angle.

The camera angle change method was as shown in the following for Equation (1). The equation expresses the changed x and y values by multiplying the determined 3D square matrix by the x and y values corresponding to each pixel. To illustrate the equation using Figure 6 as an example, x , y are the x and y coordinates of each pixel in the original Figure 6a, and x' , y' are the x and y coordinates of each pixel in the converted Figure 6b. t is the remaining value which is not actually used, and h is the total height value in the video, while w is the total width value in the video. Additionally, the value a is the weight value for how much the camera angle is to be changed. As the value of a increases, the effect of the camera looking down from a higher angle can be obtained.

$$\begin{pmatrix} x' \\ y' \\ t \end{pmatrix} = \begin{pmatrix} 1 & \frac{-a * w}{2 * a * h - h * w} & 0 \\ 0 & \frac{w}{w - 2 * a} & 0 \\ 0 & \frac{-2 * a}{2 * a * h - h * w} & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \quad (1)$$

In the same way as above, videos could be generated at various angles through camera angle conversion, and accordingly various learning data were generated. In addition, the generalization of the learned network was better because the videos were diversified.

4. Experimental Results

To verify the proposed method, a sign language dataset from KETI (Korea Electronics Technology Institute) was used. Among the KETI data sets, 20 classes were used: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, bear, ear, gas, leak, leg, next year, police, eyes, stairs, and stun did. There were a total of 20 images in each class, and a total of 400 data were used. The network to be used in the experiment is shown in Figure 3.

The experiment was conducted in the following four ways, and the results were compared. The rest of the learning environment was the same, except for the number of data sets.

- Learning is done with only the original data. That is, using 400 data in total, the learning process is as shown in Figure 3.
- Data augmentation is performed using camera angle conversion and random keypoint removal at the same time before learning. Using camera angle conversion, the original data are augmented about five times, and this is doubled through the random keypoint removal method to learn using 4000 data, which is 10 times the total of the original data set.
- Finger length conversions and random keypoint removal are used simultaneously to perform data augmentation before learning. Using the finger length conversion, the original data are augmented by about five times, and this is doubled through the random keypoint removal method to learn using 4000 data, which is 10 times the total of the original data set.
- All three methods are used simultaneously: camera angle conversion, finger length conversion, and random keypoint removal, and data augmentation is performed before learning. Using the finger length conversion, the original data are augmented about five times, and this is doubled through the random keypoint removal method, and the data are again augmented five times using the camera angle conversion. As a result, learning is performed using 20,000 data, which is 50 times the original data set.

In addition, three methods of camera angle conversion, finger length conversion, and random keypoint removal describe how to increase data. First, in the camera angle conversion method, data of a camera angle weight a is combined with the original using four values of 50, 100, 150, and 200, respectively, and the data are increased five times. The finger length conversion method uses four types of random finger length increase and decrease in left hand, and random finger length increase and decrease in right hand, which is combined with the original to increase data five times. Lastly, the random keypoint removal method doubles the data by combining the original and one method of randomly removing keypoint from the whole.

The results of the experiment using the above method are shown in Table 1 below. The total number of trainings was 1000, but an early stop was set so that if the loss converged in the middle, the training was terminated, the batch size was set to 32, and the learning rate was set to e^{-6} .

Table 1. Compare training result.

Method	Number of Data	Learning Availability	Accuracy
a	400	X	-
b	3200	O	41.9%
c	4000	O	87.2%
d	16,000	O	96.2%

In addition to the results, when a was trained with 400 original data sets, the learning itself was not performed properly, so the results could not be confirmed. The reason for not learning was that the learning process was not confirmed because the learning loss did not change due to the small data set and the learning result was not confirmed. Then, b and c both increased the data by 10 times—c increased the data successfully, while b increased the data only by eight times, with 3200 data. The reason that the data were not increased by 10 times was that the openpose was not recognized because the camera's angle changed excessively, so the dataset was not properly created. In addition, it can be seen that the accuracy is also low as the data are reduced. The learning process of b is shown in Figure 7. The total number of learnings was 240, and loss converged at 2.48.

Figure 7. Training process of b.

c shows that the data are successfully generated and the learning is successful, showing a considerable accuracy of 80%, and the learning process is shown in Figure 8. The total number of learning sessions was 80, and the loss converged at 0.365.

Figure 8. Training process of c.

d shows a high accuracy of 96.2% by using enough datasets by combining all the previous methods, and the learning process is shown in Figure 9. The total number of learning sessions was 500, and the loss converged at 0.272.

Figure 9. Training process of d.

As can be seen in Table 1, increasing the number of data through data augmentation has a great influence on the learning result, and shows that good learning results can be obtained by learning enough data by properly combining the proposed methods.

In addition, the comparison with the frame skip sampling data augmentation method of sign language translation [4] mentioned in Related Work is shown in Table 2 below.

Table 2. Compare method d and frame skip sampling.

Method	Augmentation Factor	Number of Data Per Class	Accuracy
d	40	800	96.2%
Frame skip sampling	50	1000	93.2%

The frame skip sampling method has an accuracy of up to 93.2% when the data are increased 50 times, and the d method has an accuracy of up to 96.2% when the data are increased 40 times. d has a higher accuracy with fewer data.

Figure 10 shows the result of the program execution and estimates the sign language estimated through the model in Figure 3 and outputs it as text.

Figure 10. Program execution example.

5. Conclusions

In this paper, we proposed an augmentation method for sign language data to improve learning performance. Three methods were proposed: finger length conversion, random keypoint removal, and camera angle conversion, and these methods were combined with b, c, and d to augment the data by as little as eight times and as much as 40 times. Through this, it showed various performance improvements from the lowest accuracy of 41.9% to the highest of 96.2%, and using all the combinations of d, 54.3% higher accuracy improvement than b. Through these results, it was verified that the method of this paper was very helpful when there were insufficient sign language data.

6. Discussion

In this paper, we verify new three data augmentation methods used in sign language translation. As you can see from the above experiment results, the combination of two of the three proposed methods, such as b and c, can be used to collect the minimum data for learning. When all three methods are used in combination, a high accuracy of 96.2% is obtained. In addition, when looking at the results in Table 2, the learning results may vary depending on the data augmentation method. If the data augmentation method is more suitable for the data, the maximum results can be obtained with less data. Using small data to achieve maximum results means better learning efficiency per data.

Therefore, if the data augmentation method proposed is used when using openpose processed sign language video as training data, you can save time and effort to collect data with high performance.

Author Contributions: Conceptualization, C.-I.P.; funding acquisition, C.-B.S.; investigation, C.-I.P.; methodology, C.-I.P.; project administration, C.-I.P.; software, C.-I.P.; supervision, C.-I.P.; writing—original draft, C.-I.P.; writing—review and editing, C.-I.P. and C.-B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2020-2016-0-00288) supervised by the IITP (Institute for Information and Communications Technology Planning and Evaluation).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [[CrossRef](#)] [[PubMed](#)]
2. Perez, L.; Wang, J. The effectiveness of data augmentation in image classification using deep learning. *arXiv* **2017**, arXiv:1712.04621.
3. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [[CrossRef](#)]
4. Ko, S.K.; Kim, C.J.; Jung, H.; Cho, C. Neural sign language translation based on human keypoint estimation. *Appl. Sci.* **2019**, *9*, 2683. [[CrossRef](#)]
5. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. OpenPose: Realtime multi-person 2D pose estimation using Part Affinity Fields. *arXiv* **2018**, arXiv:1812.08008. [[CrossRef](#)] [[PubMed](#)]
6. Liu, Y.H. Feature extraction and image recognition with convolutional neural networks. *J. Phys. Conf. Ser.* **2018**, *1087*, 062032. [[CrossRef](#)]
7. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
8. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Comlulus, OH, USA, 23–28 June 2014; pp. 1725–1732.
9. Jogin, M.; Madhulika, M.S.; Divya, G.D.; Meghana, R.K.; Apoorva, S. Feature extraction using Convolution Neural Networks (CNN) and Deep Learning. In Proceedings of the 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bengaluru, India, 18–19 May 2018; pp. 2319–2323.

10. Chen, Y.; Jiang, H.; Li, C.; Jia, X.; Ghamisi, P. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks. *IEEE Trans. Geosci. Remote. Sens.* **2016**, *54*, 6232–6251. [[CrossRef](#)]
11. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)] [[PubMed](#)]
12. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).