American Sign Language Recognition using 3D Residual Networks

Shaelynn Brown Supervisor: Kwang Moo Yi
University of Victoria
{shaeb, kyi}@uvic.ca

Abstract

Action and gesture recognition is a challenging task which becomes increasingly difficult as the size of the vocabulary increases. Sign language, as a type of gesture recognition problem, is particularly interesting due to it being a means of communication for millions of deaf and hard-of-hearing people across the world. We address this problem using techniques from the computer vision field such as 3D Residual Networks and Spatial Transformer Networks. We evaluate the model by constructing a data-set using the National Center for Sign Language and Gesture Resources (NCSLG) Corpus. The work has resulted in an accuracy of 52%, between 51 different glosses.

1. Introduction

In Canada there is approximately 3.57 million people who are deaf or hard-of-hearing [1], many of which use sign language as their main form of communicating. Sign languages used in Canada include American Sign Language (ASL) and Quebec Sign Language (LSQ).

Due to ”hearing patronization, inappropriate educational methodology, and systemic discrimination” [2], people who are deaf experience a high unemployment rate. In 1998, a study concluded that 80% of people who are deaf in Canada did not have full-employment (under-employed or unemployed), and this number has been increasing since [2].

The use of an interpreter is a very effective form of communication between people who use spoken and signed languages. However, in Canada there is a dire lack of interpreters available and the cost of hiring one starts at approximately $100/hr. A computer program which could translate between spoken and signed languages in both directions, could allow interpretation to be more accessible for deaf people who are low income, under-employed, or unemployed.

In order to translate full signed sentences, we would need a program to recognize individual signs with high accuracy. This work focuses on classifying signs by their gloss with a limited vocabulary set, using modern computer vision methodologies.

2. Dataset

Initially, the American Sign Language Lexicon Video Dataset (ASLLVD) [3, 4] was used to construct a data set. However, due to the lack of data per gloss, sufficient results were not being achieved. In order to get more training data, the National Center for Sign Language and Gesture Resources (NCSLGR) Corpus [4] was used instead. Both data sets were published by the University of Boston. Although there is more data in the NCSLGR Corpus, the videos were less clean because they were clipped from an utterance.

In the repository, there is a script to organize the structure of the download zip into folders for each gloss/label. Afterwards, another script is used to create an H5 file using the directory of videos. The number of frames to sample from each videos, and processing options can be specified for each image.

The final data set used has 51 classes, with 31-263 videos per gloss, and a total of 3232 videos. Sixteen frames are taken from each video; each image is rescaled by 0.42, cropped to 128x128, and converted to greyscale. The metadata on the bottom is removed to ensure the its information is not used by the model.

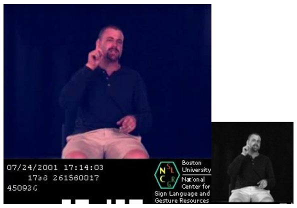

Figure 1. Left: original data, Right: data after processing

In order to create a class-balanced testing and validation set, two examples per gloss are chosen for the testing set at random, and five examples per gloss for the validation set.

3. Residual Building Block

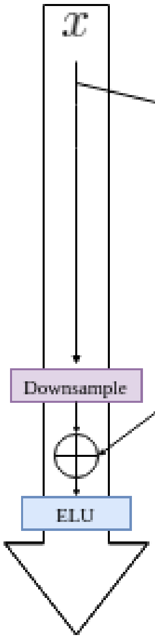

Figure 2. Residual Network Block

Figure 2 refers to the residual block used in the network. Each convolutional layer has a padding of 1, and a kernel size of (3,3,3). The stride of the first convolutional layer, and the output features of the block are optional parameters. If the size of the input data does not match the size of the output before the summation, it is downsampled.

4. Network Architecture

4.1. Architecture A

The first architecture was heavily influenced by the paper authored by L. Pigou et al. [5], where they used deep learning to classify Dutch and Flemish Sign Languages. The work takes the temporal difference of each image/frame, however better results were found by just taking the greyscale image. Another difference is the depth of the network; their final Residual Block outputs 128 features, where ours outputs 512 features. Additionally, they used a kernel size of (1,3,3) and (3,1,1) for each convolutional layer in the ResNet block respectively. Instead, we use a kernel size of (3x3x3) in each layer.

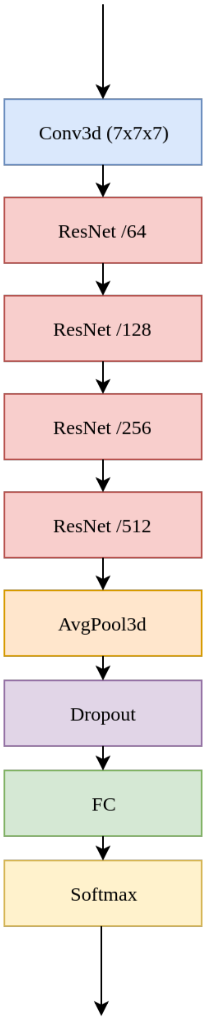

Figure 3. Network architecture

Figure 3 shows the layers of the network, where each ’ResNet’ block refers to layers in Figure 2.

4.2. Architecture B

The repository by Kensho Hara et al. [6] was referenced while implementing the architecture in Figure 3. Their work evaluates the results of a 3D ResNet on various gesture recognition data sets. The architecture they used varied a bit from Figure 3, and I was curious to see if it would out-preform it. Firstly, the network had a Batch Norm, ReLu and a Max Pool layer before the residual blocks. Secondly, it lacked a dropout layer. Thirdly, it allowed to optionally add various amounts of ResNet blocks per layer, where each layer increases the number of output features. Fourthly, Relu activation layers were used instead of Elu. Architecture B can be tested by setting the `--pooling_before_resnet`, `--layers`, and `--activation` flags.

4.3. Training

The model was evaluated using the cross entropy loss function. It was optimized by using Adam with a learning rate of 0.001. Different batch sizes were used depending on the complexity of the network being trained.

5. Results

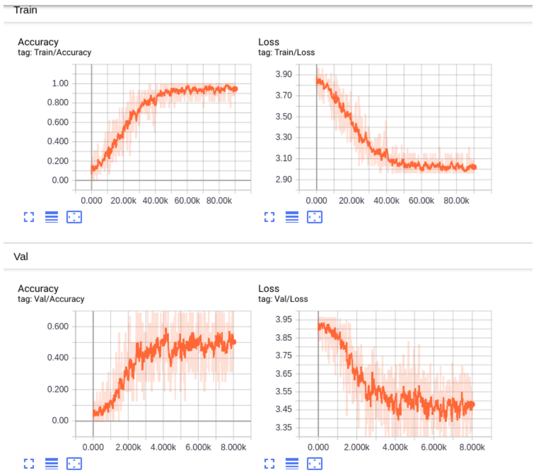

Figure 4. Loss and accuracy curves. Top: train, Bottom: validation

Architecture A achieved a 52% accuracy on the test set.

Architecture B was also tested. With the ReLu activation function, it seemed to suffer from the "dying" ReLu problem" due to the loss increasing instead of decreasing. The model is currently being trained with Elu activation function instead, but is taking too long to include in the report.

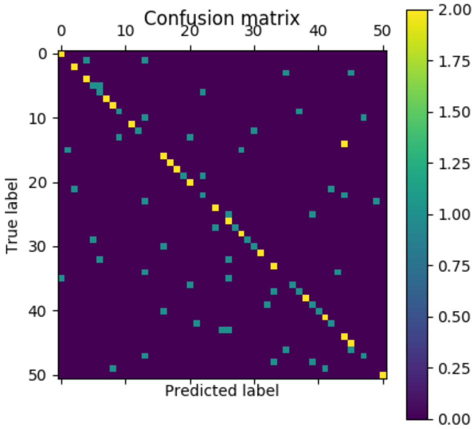

Figure 5. Confusion Matrix

The yellow pixel in Figure 5 that is outside diagonal represents the confusion between 'for' and 'think' glosses. The following demonstrates that the model has difficulty classifying between signs that have very similar handshapes.

5.1. Spatial Transformer Network

A spatial transformer network (STN) [7] learns to select regions of an image that are most relevant but also to transform those regions to an expected pose. They can be injected into a network and be trained end-to-end with the rest of the model. Due to the amount of background information in the data set, it was thought that prepending a STN to the model would increase performance.

Appending an STN to the network in architecture A in Figure 3 resulted in negligible increase in performance. Although, it does take a 324x250 frame input, and learns to crop and transform to a 128x128 output.

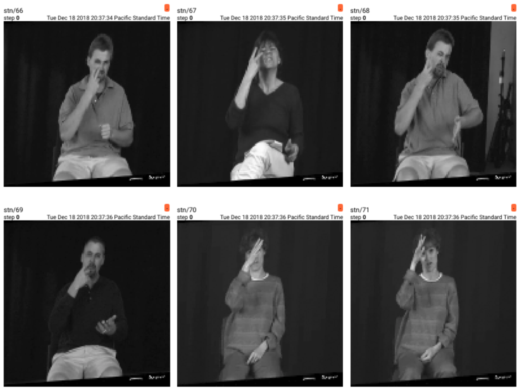
stn/66
step 0
Tue Dec 18 2018 20:37:34 Pacific Standard Time
stn/67
step 0
Tue Dec 18 2018 20:37:35 Pacific Standard Time
stn/68
step 0
Tue Dec 18 2018 20:37:35 Pacific Standard Time
stn/69
step 0
Tue Dec 18 2018 20:37:36 Pacific Standard Time
stn/70
step 0
Tue Dec 18 2018 20:37:36 Pacific Standard Time
stn/71
step 0
Tue Dec 18 2018 20:37:36 Pacific Standard Time

Figure 6. Output of the STN module

The STN crops and centers the actor, and slightly rotates the image. The rotation can be seen from the tilted black bar on the bottom of the image. Using the STN along with Architecture B may result in better performance.

6. Future Work

Using data augmentation would prove beneficial. Specifically, horizontal flipping could be used to emulate more hand-dominance variance in the data set. In addition, more architecture experiments and hyper parameter searching could find a better performing model. Improvements on the data set could include more training data, and a larger vocabulary. If a future model could detect individual signs with a large vocabulary set and high accuracy, further work in translating full sentences could be done. Each sign could be better predicted by also using probabilities based on the context of the sentence.

References

[1] "Canadian Association of the Deaf". (2015, June) Statistics on deaf canadians. [Online]. Available: http://cad.ca/issues-positions/statistics-on-deaf-canadians/

[2] Canadian Association of the Deaf. (2015, June) Employment and employability. [Online]. Available: http://cad.ca/issues-positions/employment-and-employability/

[3] C. Neidle, A. Thangali, and S. Sclaroff, "Challenges in development of the american sign language lexicon video dataset (asllvd) corpus," 2012. [Online]. Available: https://open.bu.edu/handle/2144/31899

[4] C. Vogler and C. Neidle, "A new web interface to facilitate access to corpora: development of the asllrp data access interface," 2012, http://www.bu.edu/asllrp/ and http://secrets.rutgers.edu/dai/queryPages/. [Online]. Available: https://open.bu.edu/handle/2144/31886

[5] L. Pigou, M. Van Herreweghe, and J. Dambre, "Gesture and sign language recognition with temporal residual networks," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW 2017)*, 2017, pp. 3086–3093. [Online]. Available: http://dx.doi.org/10.1109/ICCVW.2017.365

[6] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6546–6555.

[7] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," 2015.