ECE 491 Project Report
American Sign Language Gesture Classification

Amir Seidakhmetov, Joshua Wilbur

Abstract

American sign language (A.S.L.) is a common form of communication that isn't understood by many individuals outside of the deaf community. A VGG19 base mode was fine tuned to create a deep learning network that specializes in interpreting sign language gestures. This model performed well and was 96% accurate during training. When run with a novel testing dataset, the network was 77% accurate.

1. Introduction

Many individuals struggle to effectively communicate with those who are deaf. Furthermore, there are people who become deaf at some point in their lives and do not know sign language. This lack of understanding can impact quality of living and people who are close to the deaf individual. Currently there isn't a "best" solution for this problem, although companies, such as Google, are working on this issue as well.

American sign language is a gesture based language, with all letters of the alphabet being represented with hand gestures. The model employed for this task takes an image as an input and outputs what letter is being represented. This model is a fine tuned convolutional neural network based on the VGG19 model.

2. Related Work

In "Artificial Intelligence Technologies for Sign Language", the authors provide a comprehensive review of already existing methods of identifying the sign language, discussing their advantages, limitations and relations between them. The paper mainly summarizes the theoretical methods of sign language capturing, recognition, translation, and representation.

In "AI at the Edge for Sign Language Learning Support", the authors proposed creating a robust CNN model for correctly identifying ASL letters in "most cases". The authors also proposed using such a network for creating a software aimed at assisting people in ASL learning. Using transfer learning and AlexNet, the authors were able to achieve validation accuracy of 99.96%.

In "The FATE Landscape of Sign Language AI Datasets: AnInterdisciplinary Perspective ", the authors discuss the ethics of using sign language datasets in the context of personal nature of the data, as well as historical oppression of the deaf community. The papers highlight the need for more involvement of the deaf community in the AI research to ensure ethical adoption.

Like in the works mentioned before, in "Sign Language Recognition and Translation: A Multidisciplined Approach From the Field of Artificial Intelligence", the authors discuss existing technology used for ASL recognition and translation, as well its' impact on education of the deaf and hard-of-hearing community. The authors also propose future improvement on existing translation systems like "THEOS" and "Tessa".

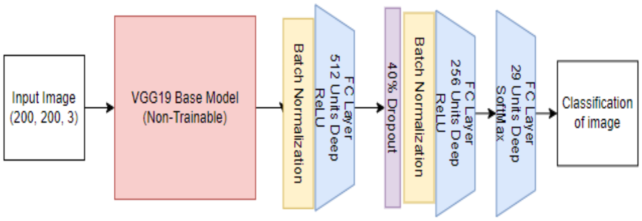
In "Human–computer interaction based on visual hand-gesture recognition using volumetric spatiograms of local binary patterns", the authors discuss a vision-based hand gesture recognition system for Human-Computer-Interaction. The paper introduced a new video descriptor VSLBP that is designed to recognise hand gestures in color imagery.

3. Dataset and Features

Two datasets were used from Kaggle. First dataset consisted of training and test data. Training data is 4.5GB, consisting of 29 classes of data. There are 26 classes for each letter of the alphabet, plus additional classes for "space", "delete", and "nothing" classes. Each class consists of about 7500 images each. Test data consisted of 29 images, one image per class. Due to the small size of the testing dataset, the training data was split into 80% training and 20% validation. The second dataset was used as a testing dataset, it consisted of 30 images for each class. For both datasets, images within a data class differ in hand orientation, position, background, and lighting, as well as by the shape and position of fingers. Before being used in the model, all images were resized to 200x200, and augmented with shear and zoom to improve model generalization.

4. Model Architecture

The deep learning model used for this task is a fine tuned VGG19 base model. Six layers are placed at the output of the VGG19 model. These layers consisted of three fully connected (F.C.) layers, one dropout layer and two batch normalization layers. A high level overview of the model is below.

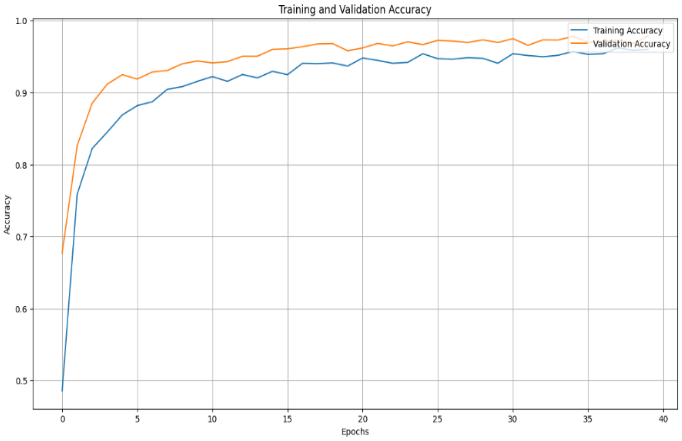


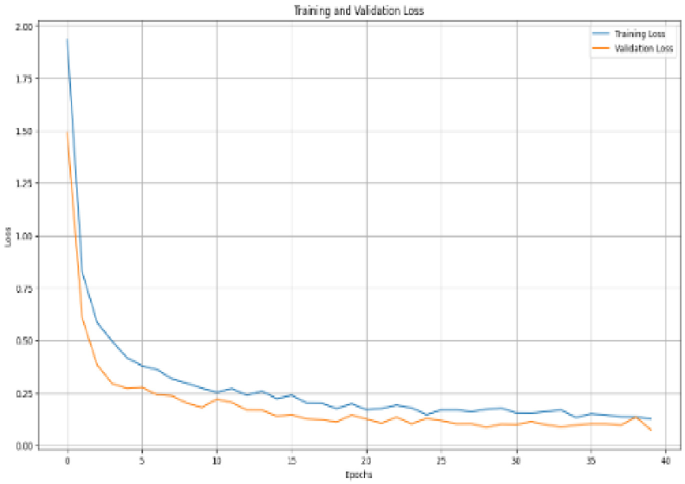
The F.C. layers have depths of 512, 256 and 29 units respectively. The depth choice for the first two F.C. layers were found through experimentation, while the final F.C. layer had 29 units to classify images into 29 classes. The first two F.C. layers have ReLU activation and L1 bias regularization. The ReLU activation function is used as it reliably converges and isn't computationally expensive. L1 bias regularization was used to ensure neuron biases didn't become extreme. The final F.C. layer uses the softmax activation function. Batch normalization layers were placed between the initial two F.C. layers. This helped accelerate training and generalize the model. A single dropout layer, using 40% dropout, was placed before the 256 unit F.C. layer to fight overfitting.

The Adam optimizer algorithm, provided by Keras, was used with a learning rate of 0.0014. This value for learning rate is slightly higher than the default of 0.001. It was found through experimentation that using a 0.0014 learning rate sped up training and allowed the model to fully train in about 40 epochs. The categorical crossentropy loss function was used to gain insight into the difference between predictions and ground truth.

5. Experiments/Results/Discussion

After construction of the model, it needed to be trained and validated to ensure good performance. The model was trained for 40 epochs with 128 steps per. Training occurs in the initial half of every epoch. This is where the model learns from the images. Validation occurs in the back half of every epoch. This consists of the model predicting gestures for unseen images using knowledge it gained during training. A batch size of 64 was used to improve training speed and convergence. Metrics measured during training were training accuracy, training loss, validation accuracy and validation loss. Accuracy measures how well the model predicts the correct hand gesture class compared to the ground truth. Loss measures the difference between the predicted gesture probabilities and the true gesture. After the 40 epochs were run, the following two graphs show the accuracy and loss of the model.





As seen in the graphs above, the model performed quite well. The peak values for accuracy were 96.19% for training and 97.85% for validation. These occurred on epochs 38 and 40 respectively. The lowest loss values were 0.125 for training and 0.074 for validation. These both occurred in epoch 40.

Prior to getting the above results, the network was tested on three different base models to see which performed best. ResNet101 and EffiecientNetV2S were compared against VGG19 over ten epochs. ResNet V2 models were tested as well, however these models performed poorly on this task. The results showed that VGG19 trained more efficiently than the others. Results for it are shown in a table below.

Model	Avg. Epoch Run Time	Training Loss	Training Accuracy
VGG19	70 sec.	0.2716	91.55 %
ResNet101	73 sec.	0.3183	89.43 %
EfficientNetV2S	68.5 sec.	0.3969	86.96 %

The VGG19 base model is pacing the others by at least 2%. Loss is also lowest for VGG19, which gave reason for using it in this case.

The results obtained thus far showed that the model had learned from the dataset. Concerns about overfitting led to use of a novel dataset to test model performance in unique situations. The validation dataset is split off from the training dataset, which had many images that were very similar. If the model is memorizing inputs, which is an issue with overfit models, it could perform poorly with new data. The testing dataset, which is described in section three, was run for five epochs to get a sense of the accuracy and loss. Over the five epochs, the peak accuracy was 77.01 % and lowest loss was 1.034. This showed that the model isn't perfect, however it is able to identify hand gestures in new images with decent accuracy. Section six provides some ways that could further generalize the model.

6. Conclusion and Future Work

In summary, our model has achieved a validation accuracy of 97.85%, demonstrating the model's robustness in interpreting American Sign Language gestures and underscoring the effectiveness of our fine-tuning approach on the VGG19 base model. Interestingly, the VGG19 model outperformed both ResNet101 and EfficientNetV2S in our experiments, with the lowest average epoch run time and highest training accuracy. This can be attributed to VGG19's architecture, which is well-suited for image recognition tasks, and its ability to train more efficiently on our dataset. Looking ahead, we could explore additional datasets, experiment with more complex architectures, and involve the deaf community to ensure ethical and effective communication tools. Our ambitious goal is training and optimizing the model for translation of videos or live video feeds, as well as integrating a language model for interpretation of complete sentences, including based on the context.

Contribution

Both members contributed greatly to the development of this model. Database uploading and preprocessing was done by Amir. Initial construction and training of the model was a joint effort. Further hyperparameter tuning and training was done by Josh. Once the model was performing well enough, the testing dataset was integrated and tested jointly.

References

Battistoni, Pietro, et al. "AI at the Edge for Sign Language Learning Support | IEEE Conference Publication | IEEE Xplore." Ieeexplore.ieee.org, 27 Dec. 2019, ieeexplore.ieee.org/abstract/document/8940852.

Bragg, Danielle, et al. "The FATE Landscape of Sign Language AI Datasets." ACM Transactions on Accessible Computing, vol. 14, no. 2, 30 June 2021, pp. 1–45, https://doi.org/10.1145/3436996.

Maqueda, Ana I., et al. "Human–Computer Interaction Based on Visual Hand-Gesture Recognition Using Volumetric Spatiograms of Local Binary Patterns." Computer Vision and Image Understanding, vol. 141, Dec. 2015, pp. 126–137, https://doi.org/10.1016/j.cviu.2015.07.009.

Papastratis, Ilias, et al. "Artificial Intelligence Technologies for Sign Language." Sensors (Basel, Switzerland), vol. 21, no. 17, 30 Aug. 2021, p. 5843, www.ncbi.nlm.nih.gov/pmc/articles/PMC8434597/, https://doi.org/10.3390/s21175843.

Parton, B. S. "Sign Language Recognition and Translation: A Multidisciplined Approach from the Field of Artificial Intelligence." Journal of Deaf Studies and Deaf Education, vol. 11, no. 1, 12 Oct. 2005, pp. 94–101, academic.oup.com/jdsde/article/11/1/94/410770, https://doi.org/10.1093/deafed/enj003.