Intelligent Recognition of American Sign Language System Based on Deep Learning

GitHub repository: https://github.com/dinglunz/CV-for-American-Sign-Language

Zhanqin Ai
University of Southern California
Zhanqina@usc.edu

Dinglun Zhang
University of Southern California
dinglunz@usc.edu

Abstract

American Sign Language (ASL) is a comprehensive natural language that is the primary sign language of the deaf community in the United States.

This project focuses on building a deep learning model to recognize and interpret American Sign Lan-guage from images. The goal is to leverage the data set available on Kaggle, which is made up of images representing 29 categories that correspond to 26 le-tters of the alphabet, a space, deleted gestures, noth-ing. The dataset contains 3,000 images for each category, all captured as 200x200 RGB images. The stru-cture of the dataset is divided into training subsets and test subsets, with special emphasis on model trai-ning without test sets to encourage the use of real-world images.

We also describe the steps of the method: the necessary pre-processing of the dataset images, a literature survey of existing work in the field to un-derstand the network architectures used, experi-ments on different architectures and hyperparameters, and the results of testing the model on an external image set due to the small size of the test set provided.

We also discussed the possibility of extending the project to develop an application capable of cap-turing video and providing real-time transcription, which could be further used to generate speech, com-bined with features such as Google text-to-speech.

Keywords: EE 541, ASL Recognition, Deep Learning, Computer Vision, Accessibility

1. Introduction

Sign language is designed to allow normal communication between deaf and non-deaf people. However, with the development of society, the popularity of sign language is decreasing. Because most people do not understand sign language, there are no special services about sign language in many public places, and many deaf Americans feel inconvenience in life, and even cause obstacles in life, such as education, medical care and workplace environment. This is largely because the development of intelligent recognition systems in ASL has lagged behind [1].

At the same time, the lack of effective learning tools is also a problem for those who wish to learn ASL. While there are a few apps and online resources available for learning ASL, most lack the real-time feedback and personalized learning experiences that are essential for language learning.

Given these challenges and needs, our goal is to develop an intelligent ASL recognition system based on deep learning that will provide real-

time, effective communication tools for the deaf community and those who wish to learn ASL.

We expect that the implementation of this system will significantly improve the quality of life of the deaf and deaf-mute community, make it easier for them to communicate in daily life, and provide an efficient learning platform for those who are interested in American hand language.

2. Problem statement

Given a sign language image (captured hand movements), we hope to build a system that can:

- Detect hand movements in the picture
- The adjustment of image input size and color can be converted into processable data
- Different sign language image frames may vary in size, but it is assumed that the image frame of a particular sign language image is stable. In this article we use 200x200 pixel RGB image frames.

3. Mini Literature Review

The problem of American Sign Language (ASL) recognition using deep learning has been tackled by various researchers. Kshitij Bantupalli and Ying Xie at the College of Computing and Software Engineering approached ASL recognition by employing machine learning and computer vision techniques, contributing to the body of knowledge in robotics and intelligent systems [2].

In the realm of ASL alphabet recognition, significant progress has been reported. Researchers have proposed models that recognize the ASL alphabet from RGB images, focusing on the pre- processing of images and employing a squeeze net architecture suitable for mobile devices [3]. This indicates a trend towards developing more accessible and portable ASL recognition systems.

Further research conducted a comprehensive review of automated sign language recognition based on machine and deep learning methods. This study spanned publications from 2014 to 2021, suggesting that despite advancements, there is still a need for improved conceptual classification within current methods [4].

Another study focused on the development of a sign language translation system based on deep learning, underscoring the importance of sign language as the primary communication tool for people with hearing and language impairments [5].

A comprehensive overview of sign language recognition, including various types and modalities, is presented in a recent review paper. This paper covers two decades of progress in isolated and continuous sign language recognition (SLR), discussing various sensing approaches, SLR datasets, and current trends in SLR. This review also analyzes current issues in SLR and advises on future research directions [6].

[7] focuses on improving sign recognition with insights from phonology. It discusses the importance of structural units in sign language, like handshape, location, and movement, and how they are similar to phonological elements in spoken languages. This research is significant as it attempts to bridge the gap between natural language processing and theories of sign language phonology.

4. Dealing with data

We found that the size of test set in this dataset was too small (only 1 image per class), while the size of training set was sufficient. So we decided to split a portion of the training set to the validation set and test set, which had a minimal impact on the training dataset.

4.1 Labeling

The ASL dataset utilized in this project is sourced from Kaggle, containing 87,000 images representing 29 labels, including the 26 letters of the English alphabet, along with three additional signs: SPACE, DELETE, and NOTHING. Each

image is in RGB format with a resolution of 200 by 200 pixels.

The labeling process involved associating each image with its corresponding ASL category. The dataset was divided into training, validation, and test subsets, ensuring a representative distribution of categories in each subset. This meticulous labeling allows the model to learn the intricate gestures associated with each sign language category during the training phase.

4.2 Augmentation

Data augmentation is a critical step to enhance the model's ability to generalize to diverse real-world scenarios. The augmentation techniques applied to the images during training include:

- **Resize**: All images are resized to a consistent 200x200 pixel resolution.
- **Color Jittering**: Random adjustments to brightness, contrast, and saturation are applied, introducing variability to the color distribution in the dataset.
- **ToTensor Transformation**: Conversion of images to PyTorch tensors for compatibility with the neural network.
- **Normalization**: Pixel values are normalized to have a mean of [0.485, 0.456, 0.406] and a standard deviation of [0.229, 0.224, 0.225], aligning with standard pre-processing practices for image data.

These augmentations ensure that the model is exposed to a diverse range of visual conditions, contributing to its robustness in handling real-world images.

4.3 Handling

The dataset, comprising 87,000 images, was divided into subsets for training, validation, and testing. A split of 70%, 15%, and 15% was employed, respectively, ensuring a balanced distribution of categories in each subset. The emphasis on training without a dedicated test set encourages the model to adapt to real-world scenarios, enhancing its performance on unseen data.

Additionally, steps were taken to handle potential imbalances or noise in the dataset, ensuring that the model's training is both effective and robust. The careful management of the dataset, coupled with thoughtful data augmentation, lays the foundation for the successful training and evaluation of the ASL recognition model.

5. Model Architecture

The SignLanguageModel is a convolutional neural network (CNN) designed to recognize and interpret American Sign Language (ASL) gestures from images. The architecture encompasses several key components tailored for image classification tasks.

5.1 Convolutional Layers

The model begins with the first convolutional layer (conv1), which takes RGB input images of size 200x200 pixels. This layer consists of 16 filters with a kernel size of 3x3 and padding of 1. The choice of padding ensures that the spatial dimensions of the input are preserved after convolution. The rectified linear unit (ReLU) activation function is applied after each convolution to introduce non-linearity.

The output of the first convolutional layer undergoes max-pooling (max_pool2d), reducing the spatial dimensions by a factor of 2. This process helps the model capture essential features while downsizing the computational complexity.

The second convolutional layer (conv2) follows a similar structure as the first, with 32 filters, a 3x3 kernel, and padding of 1. ReLU activation and max-pooling are again applied to further extract hierarchical features.

5.2 Fully-Connected Layer

The output from the last convolutional layer is flattened (view) to a one-dimensional tensor.

This flattened representation is fed into a fully-connected layer (fc1) with 29 output nodes, corresponding to the 29 categories in the ASL dataset. The choice of 29 nodes aligns with the unique labels present in the dataset.

5.3 Dropout Layer

To prevent overfitting and enhance generalization, a dropout layer (dropout) with a dropout rate of 0.3 is incorporated after flattening the tensor. Dropout randomly sets a fraction of input units to zero during training, forcing the model to learn robust features and preventing reliance on specific neurons.

5.4 Activation and Output

Throughout the architecture, the ReLU activation function is applied to introduce non-linearity, allowing the model to capture complex relationships in the data. The final layer outputs logits for each category, and the model employs the cross-entropy loss function during training.

5.5 Model Training

The model is trained using the Adam optimizer with a learning rate of 0.001. The choice of Adam optimizer facilitates adaptive learning rates, enabling faster convergence during training.

The training loop involves forward and backward passes, updating model parameters to minimize the cross-entropy loss. Training is conducted over multiple epochs, with the model learning to recognize ASL gestures through the iterative adjustment of weights.

6. Results and Discussion

The model exhibits remarkable performance, achieving an accuracy of 98% after 10 epochs. Learning curves demonstrate consistent improvement in both training and validation phases, indicating effective training and generalization. The confusion matrix provides insights into the model's performance across various ASL categories, highlighting its proficiency in distinguishing between gestures.

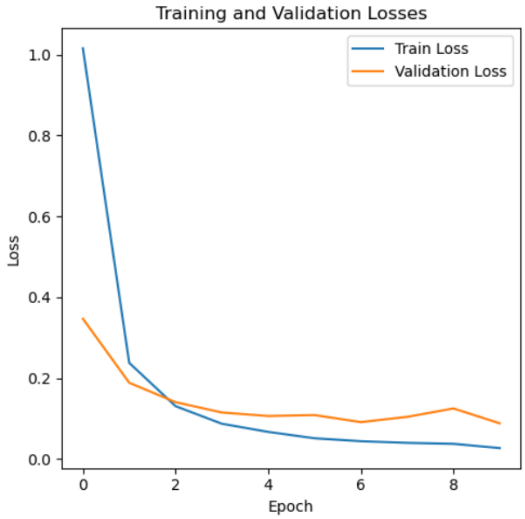

Figure 1. Learning curves – Losses

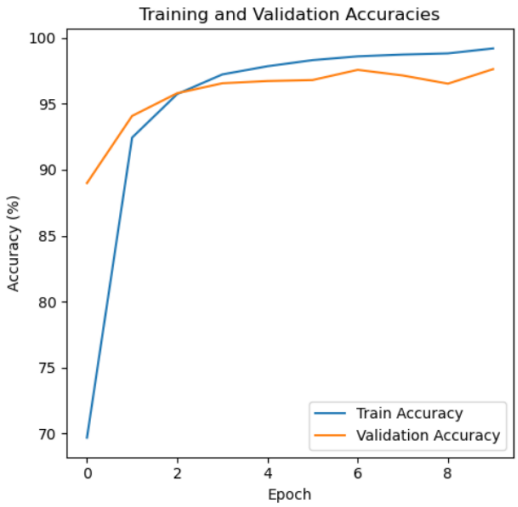

Figure 2. Learning curves – Accuracies

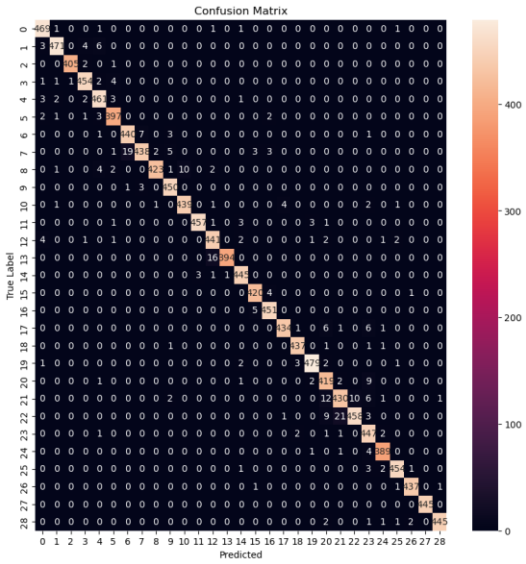

Figure 3. Confusion matrix

The success of the model showcases the efficacy of the chosen architecture, data augmentation strategies, and training parameters in building an intelligent ASL recognition system. The robustness of the model positions it as a valuable tool for real-world applications, promising improved communication and learning experiences for the deaf community and ASL enthusiasts alike.

7. Future work

The successful implementation of this ASL recognition model paves the way for impactful real-world applications. Future work could involve:

- Integration into communication platforms for real-time interpretation.
- Expansion to support additional sign languages.
- Enhancements in the model for more nuanced gesture recognition.
- Collaboration with educators to create interactive ASL learning applications.

In conclusion, the intelligent recognition of American Sign Language through deep learning holds great promise in transforming communication for the deaf community and fostering a more inclusive society.

References

[1] S. He, "Research of a Sign Language Translation System Based on Deep Learning," 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), Dublin, Ireland, 2019,

[2] Bantupalli, K., & Xie, Y. (2019). American Sign Language Recognition Using Machine Learning and Computer Vision. Kennesaw State University.

[3] Kasukurthi, N., Rokad, B., Bidani, S., & Dennisan, A. (2019). American Sign Language Alphabet Recognition using Deep Learning. arXiv.

[4] Al-Qurishi, M., Khalid, T., & Souissi, R. (2021). Research of a Sign Language Translation System Based on Deep Learning. IEEE Access, 9, 126917-126951.

[5] He, S. (2019). Deep Learning for Sign Language Recognition: Current Techniques. In 2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM) (pp. [Page numbers not provided]). Dublin, Ireland: IEEE.

[6] Madhiarasan, M., & Roy, P. P. (2022). A Comprehensive Review of Sign Language Recognition: Different Types, Modalities, and Datasets. Indian Institute of Technology Roorkee. Manuscript received April , 2022.

[7] Kezar L, Thomason J, Zed Sevcikova Sehyr. Improving Sign Recognition with Phonology. arXiv.org. Published online 2023. doi:10.48550/arxiv.2302.05759