American Sign Language to Text Conversion Using CNN

Pujita Peri
Computer Science
SRH University of Applied Sciences
Berlin,Germany
3103755@stud.srh-campus-berlin.de

Akash Mane
Computer Science
SRH University of Applied Sciences
Berlin,Germany
310253@stud.srh-campus-berlin.de

Suraj Naik
Computer Science
SRH University of Applied Sciences
Berlin,Germany
3105252@stud.srh-campus-berlin.de

Kaushal Dabhi
Computer Science
SRH University of Applied Sciences
Berlin,Germany
3105065@stud.srh-campus-berlin.de

***Abstract*— American Sign Language (ASL) serves as the predominant sign language of Deaf communities in the United States and most of Anglophone Canada. ASL is a language completely separate and distinct from English and is closely related to French Sign Language i.e. Langue des Signes Française (LSF). In this project we will be using Convolutional Neural Network (CNN), a machine learning model as a base, which will translate the sign language into a readable text format thus helping normal individuals communicate and understand the language of a deaf-mute person.**

Keywords- ASL, Hand Gestures, Kaggle, Interface, Deaf Community, Computer Vision, CNN, ReLu, Pyttsx3, SQLite3

I. *Introduction*

American sign language (ASL) is a natural language which uses hand gestures and facial expressions to help people communicate. It is different from the English language in terms of grammar. ASL originated in the early 19th century in the American School for the Deaf (ASD) in West Hartford, Connecticut, from a situation of language contact. Since then, ASL use has propagated widely by schools for the deaf and Deaf community organizations. It is primarily used by North Americans who are deaf or hard on hearing. It contains all the fundamental features of language, with its own rules for pronunciation, word formation, and word order. While every language has ways of signaling different functions, such as asking a question rather than making a statement, languages differ in how this is done. For example, English speakers may ask a question by raising the pitch of their voices and by adjusting word order; ASL users ask a question by raising their eyebrows, widening their eyes, and tilting their bodies forward.

II. *Relevant Work*

Convolutional Neural Networks have been extremely successful in image recognition and classification problems and have been successfully implemented for human gesture recognition in recent years. There has been work done in the realm of sign language recognition using deep CNNs, with input-recognition that is sensitive to more than just pixels of the images. With the use of cameras that sense depth and contour, the process is made much easier via developing characteristic depth and motion profiles for each sign language gesture [1].The use of depth-sensing technology is quickly growing in popularity, and other tools have been incorporated into the process that have proven successful. Developments such as custom-designed color gloves have been used to

facilitate the recognition process and make the feature extraction step more efficient by making certain gestural units easier to identify and classify [2].

Until recently, however, methods of automatic sign language recognition were not able to make use of the depth-sensing technology that is as widely available today. Previous works made use of very basic camera technology to generate datasets of simple images, with no depth or contour information available, just the pixels present. Attempts at using CNNs to handle the task of classifying images of ASL letter gestures have had some success [3], but using a pre-trained GoogLeNet architecture.

III. *Design Framework and Theory*

The dataset consists of ASL gestures from Fig1 .The dataset consists of 2524 images with 70 images per category.Each category represents a different character of ASL.

A. *American Sign Language*

American Sign Language (ASL) is a finished, characteristic language that has indistinguishable etymological properties from being communicated in dialects, with sentence structure that varies from English. ASL is communicated by developments of the hands and face. It is the essential language of numerous North Americans who are hard of hearing and in need of a hearing aid and is utilized by many hearing individuals too. All in all, ASL is pretty much used everywhere in the world, thus, used in this project. The five ways that ASL can benefit anyone are [4]:

1. It improves spelling skills in the children. It provides the children with another tool in which way they can remember the spelling of any word.
2. It overall improves classroom behavior. Research suggests teachers find it easier to manage class with students armed with words like toilet and excuse me, so that it does not affect the flow of class.
3. It is said that ASL vastly improves Motor skills in children since it incorporates the use of hand gestures and mouth to convey meaning.
4. It overall increases the communication skills in everyone.
5. Helps in creating a better vocabulary in children as it increases the efforts taken by them to spell any word.

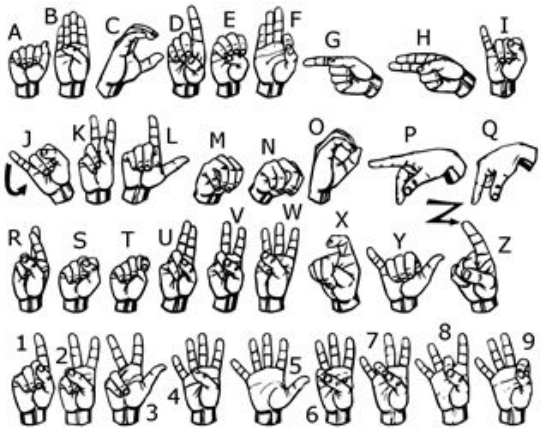

Fig 1: ASL Sign Language

B. *CNN (Convolutional Neural Networks)*

Convolution is the first layer to extract features from an input image. Convolution preserves the relationship between pixels by learning image features using small squares of input data. It is a mathematical operation that takes two inputs such as image matrix and a filter or kernel. The convolutional neural networks are very similar to the neural networks of the previous posts in

the series: they are formed by neurons that have parameters in the form of weights and biases that can be learned. But a differential feature of the CNN is that they make the explicit assumption that the entries are images, which allows us to encode certain properties in the network.

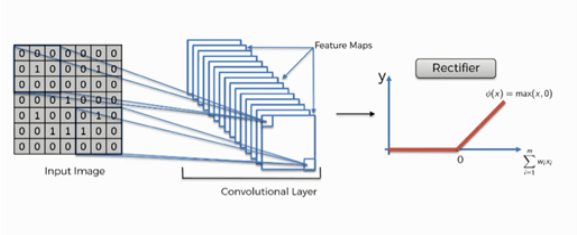

Fig 2: CNN Architecture

C. *ReLU*

ReLU stands for rectified linear unit and is a type of activation function. Mathematically, it is defined as $y = max(0, x)$. ReLU is the most used activation function in neural networks, especially in CNNs.Visually, it looks like the following:

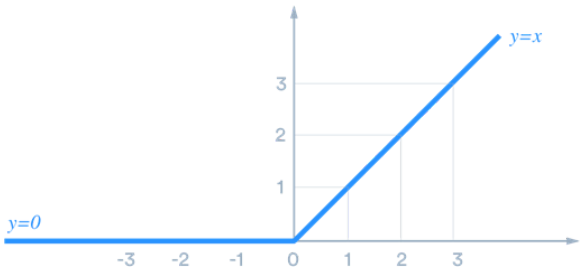

Fig 3: Relu

The Rectified Linear Unit, or ReLU, is not a separate component of the convolutional neural networks' process. The reason for applying the rectifier work is to build the non-linearity in our pictures. The explanation we need to do that will be that pictures are normally non-direct.

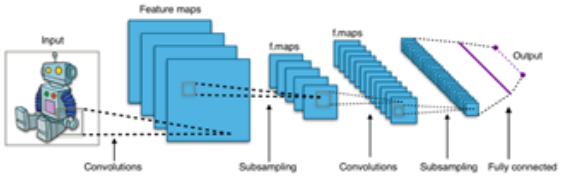

Fig 4: Relu Layer in CNN

At the point when you take a gander at any picture, you'll see it contains a great deal of non-direct highlights (for example the progress between pixels, the outskirts, the hues, and so forth.)The rectifier serves to separate the linearity significantly further so as to compensate for the linearity that we may force a picture when we put it through the convolution activity. To perceive how that really plays out, we can take a gander at the accompanying picture and witness the progressions that to it as it experiences the convolution activity followed by amendment.

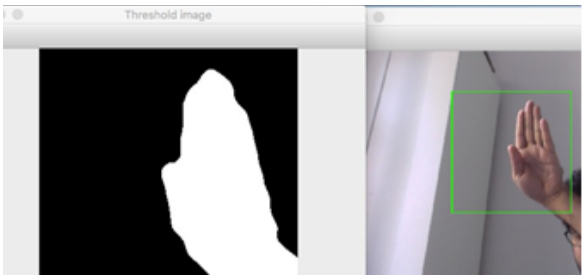

Fig 5: Threshold image

D. *Softmax*

Just like Relu, Softmax is also an activation function and is mainly used when there is only a single class. It gives probabilities of that single class over all the other classes and is generally used at the output layer.It is given by the following formula.

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$

It is mostly used for normalizing the Neural Network outputs between 0 and 1. It is a great way of knowing which input is getting the highest probability and accordingly make some changes for the desired output.

E. *Keras*

It is a vast used python library focused mainly on Machine and Deep learning algorithms. It is very flexible and offers high computational power since it already has most of the packages pre-installed in it. It also has support for Convolutional and Recurrent neural networks which makes it a great fit for most of the learning algorithms.

F. *Pyttsx*3

Python Library has its own text to speech conversion package. Though the important thing being it is written in Python 2, it's still being used widely around the programming circle due to the versatility of the package. It is easy to install and pip is already present in the system, one just has to use pip install Pyttsx3, and the package will be downloaded instantaneously.

G. *SQLite3*

SQLite finds its use almost everywhere and can be used in Android, IOS devices likewise. It is a simple database and can be easily understood by everyone. SQLite3 comes preinstalled in the Python library and one only must import it into the python script. It performs all the functions on a database for the Python script and thus makes functionality of any program easier.

IV. *Workflow*

This section sought to explain the practicality of the project and the basic design of how it was achieved. Here Python is used as a main programming language. Along with python, frameworks such as Keras are used to implement CNN (Convolutional Neural Network) Algorithm and SQLite3 database to store the image ID (0-25) and label (A-Z) to which the input is passed as ID and we extract the label from it thus displaying the output on the screen. The basic idea is if the sign is consistent for 9 frames we detect the letter. For example: A A A A A A A A A = predicated text is A . Also, if there are 25 frames without any input we consider it as a space or end of the word.

In the case of gesture recognition, the entire background region of the image is not of interest, so only the set of pixels with the presence of the human hand must be maintained. One method

For this, segmentation is the implementation of background removal, where image samples are collected from the environment and then objects are added to the scene. In this way, the pixels of the new images are compared with the images of the scenario. The regions that show a large amount of pixel differences, possibly containing the hand and gesture, are considered foreground. Although it is a good method, this type of segmentation is quite susceptible to variations in lighting.[5]

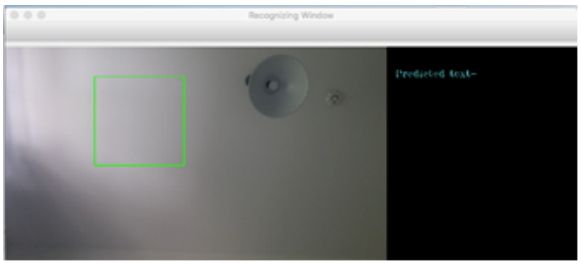

Fig 6: Predicted Text

In the above Fig 6 , the left green box is input and the right black section will show predicted text. We input our hand gesture sign in the green box accurately and according to the dataset we get the predicted result in the right side terminal which alphabet or number it is.

V. *Conclusions and future work*

Image preprocessing: We believe that the classification task could be made much simpler if there is very heavy preprocessing done on the images. This would include contrast adjustment, background subtraction and potentially cropping. A more robust approach would be to use another CNN to localize and crop the hand.

Language Model Enhancement: Building a bigram and trigram language model would allow us to handle sentences instead of individual words. Along with this comes a need for better letter segmentation and a more seamless process for retrieving images from the user at a higher rate.[6]

VI. *Acknowledgment*

This paper was done in accordance with another group headed by Ameya Mote who are working on an Android app for Text to ASL Sign. The project was developed under the guidance of Prof. Sreeganesh Thottempudi of SRH Berlin University of Applied Sciences.

VII. *References*

[1] Agarwal, Anant & Thakur, Manish. Sign Language Recognition using Microsoft Kinect. In IEEE International Conference on Contemporary Computing, 2013.

[2] Cao Dong, Ming C. Leu and Zhaozheng Yin. American Sign Language Alphabet Recognition Using Microsoft Kinect. In IEEE International Conference on Computer Vision and Pattern Recognition Workshops, 2015.

[3] Garcia, Brandon and Viesca, Sigberto. Real-time American Sign Language Recognition with Convolutional Neural Networks. In Convolutional Neural Networks for Visual Recognition at Stanford University, 2016.

[4] Dr. Johannes Fellinger, Daniel Hollzinger,Robert Pollard(2012).Mental Health of deaf people,The Lancelet, Volume 379, Issue 9820, 17–23 March 2012, Pages 1037-1044

[5] Carlos D. B. Borges, Antônio M. A. Almeida, and Iális C. Paula Static Hand Gesture Recognition Based on Convolutional Neural NetworksVolume 2019 |Article ID 4167890 | 12 pages

[6] Brandon Garcia, Sigberto Alarcon Viesca Real-time American Sign Language Recognition with Convolutional Neural Networks