American Sign Language Recognition System

Vagdevi K
S20150010029, IIITS
vagdevi.k15@iiits.in

D. Swathi Reddy
S201501015, IIITS
swathireddy.d15@iiits.in

P. Sonia
S201501038, IIITS
sonia.p15@iiits.in

Tapan Krish
S20150010044, IIITS
tapankrish.s15@iiits.in

I. Abstract

An American Sign Language (ASL) Translator opens the door for communication between the deaf or dumb individual and a normal human being. An Automated American Sign Language Recognition System has been implemented to aid physically challenged people. The system has been implemented using convolutional neural networks. A CNN architecture called VGG-16 has been trained on a data set. We have been able to build an ASL Translator, which correctly classifies the four letters a-d.

II. Introduction

Sign language has been a major boon for the people who are hearing and speech impaired. But this could serve its purpose only when the other party could understand the sign language. Since such people exist in limited number, we aim to build an automated American Sign Language System, which could convert the sign symbols into the English letters.

A technique called transfer learning has become so much handy for the computer vision geeks. This technique essentially helps us in two ways: overcoming the issue of training millions of parameters in the network from the scratch and critical demand for the gigantic computational resources. The technique of transfer learning is an approach not only to overcome these shortcomings but also to transfer the knowledge obtained from other datasets to the one we want the current task be attained. Instead of training the entire neural network from scratch, we can use the pre-trained network. Convolutional neural network is both a feature extractor and a classifier. The technique of transfer learning helps to use the pre-trained networks. Transfer Learning is nothing but training a model on one problem and re-use the obtained weights on a second related, but a different problem. There are three major transfer learning methods. They are:

- ConvNet as fixed feature extractor.
- Fine-tuning the ConvNet.
- Pretrained models.

In our project, we have used AlexNet and VGG-16 architecture. Two of the neural network architectures that worked very well in the famous Image Net Database Challenge (ILSVRC) are AlexNet and VGG-16.

III. Related work

There has been a significant advancement in this area of research. Pratibha et al [1] have proposed an pure image processing based technique, in which they have separated the background and the hand region which is considered as an object in the frames of the input video. Once the hand region was obtained, a vigorous feature extraction is carried out. These features are then fed to a feed forward neural network to achieve classification. Nevertheless, this method requires a very good feature engineering, for which huge amount of data is required. Also, higher the number of features leads us to face the curse of dimensionality, thus requires us to use techniques like PCA to reduce the dimensionality.

Ching-Hua Chan et al.[2] suggested the use of palm sized 3D leap motion sensor for accomplishing the same task. The motion sensor is used for hand and finger movements in 3D space. The sensor plays the role of feature extractor and reports data such as position, spread of palm and fingers based on the sensor's coordinate system. The obtained features were used by KNN and SVM for the classification task. However, some of the reported data such as the position of the palm is irrelevant, since the sign can be placed anywhere as long as it is in the detectable range

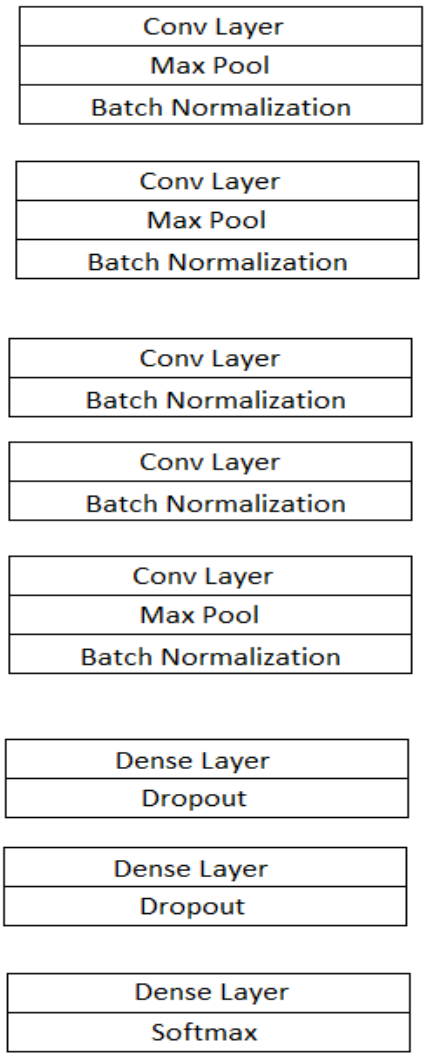
Sharma et al. used piece-wise classifiers (Support Vector Machines and k-Nearest Neighbors) to characterize each color channel after background subtraction and noise removal. They attained an accuracy of 62.3% using an SVM on the segmented color channel model.
Bayesian networks like Hidden Markov Models have also achieved high accuracies.Starner and Pentland used a Hidden Markov Model (HMM) and a 3-D glove that tracks hand movement [5]. Since the glove is able to obtain 3-D information from the hand regardless of spatial orientation, they were able to achieve an impressive accuracy of 99.2% on the test set.

Some neural networks have been used to tackle ASL translation. Arguably, the most significant advantage of neural networks is that they learn the most important classification features. However, they require considerably more time and data to train.It is in the recent times that different architectures of the Convolutional Neural Networks(CNNs) have been standing as a major breakthrough in solving image or video related problems. They account for the spatiality of features in the images by considering the neigbourhood pixels within a window size of filter used. The parameter sharing phenomenon of the CNNs stand as another important factor to help reducing the number of parameters. CNNs need huge amounts of data and resources like GPUs during their training phase. However, since they are able to outperform the traditional neural networks and any other image processing and machine learning algorithms because of their capability

to capture very powerful features from the input images, they stand as the state of art solutions to many computer vision problems.

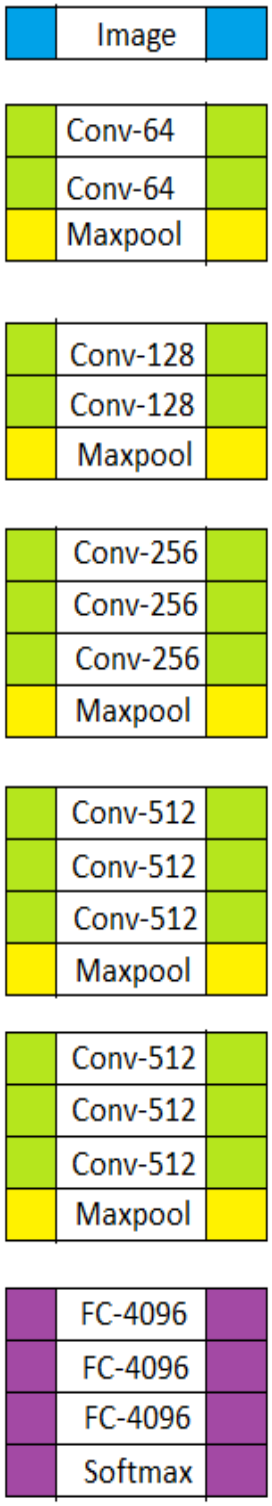
IV. Methodology

Fig. 1. AlexNet architecture



The heart of this capability is the convolutional neural network, often used for modern image processing. Technically, the work done by convolutional neural network off-screen is really astonishing. Convolutional neural networks outperform any other neural network architecture for performing object recognition within images. The ImageNet Challenge poses an open challenge where teams work to come up with various methods to classifying objects contained within millions of images into one of 1,000 different categories. The network that has been trained on such vast datasets essentially learn and extract such important features which are independent and crucial for object classification amongst the various objects available. Each layer has several filters, which may be considered as the feature detectors of the input image, right from the low-level features (in shallow layers) and high-level

Fig. 2. VGG-16 architecture



features (in deeper layers) and using non-linear combinations of these feature detections to recognize objects. The feature

maps in the convolution layers of the network can be seen as the networks internal representation of the image content. The shallow layers represent simpler features whereas the deeper layers represent more complex features. Thus the CNNs are able to capture all the features, right from the simplistic ones like colors, edges, texture etc to the higher order or complex ones as progressed through deeper filters and layers in the network. Thus we have adopted this technique of the Convolutional Neural Networks to solve out problem of recognizing the American Sign Language via hand gestures in order for it to be understood by the common people in the real world English letters.

With the millions of data samples available in the database, the model has attained a high power to represent the features of images. The technique of transfer learning can be applied in three different ways.

Pre trained neural network: It is very difficult to run the entire dataset since it takes a very long time, so the technique of transfer learning can be used. Transfer learning is nothing but the transfer of knowledge. That is, weights obtained by training the model for a dataset (ImageNet database) can be utilized for the similar problem.

Fixed Feature Extractor: In this method of transfer learning, we use the weights obtained by the transfer learning technique for the initial and the middle layers. That is, we take a convolutional neural network which was already pretrained on the famous ImageNet challenge, we remove the last fully-connected layer from it. For the classification part, we use machine learning classification algorithms like Support Vector Machines (SVM) or KNN. So, finally we obtain the classification label for the given input image.

Fine Tuning: In this strategy of transfer learning, we retrain and also replace the classifier on top of the convolutional neural network. We also fine-tune the weights of the pretrained network by continuing the backpropagation. We feed the features obtained by the middle layer to the final layer for the further feature extraction and classification. We basically feed them to the dense layer which is considered as the final layer in the network, and we train the dense layer and get the classification label. It is also possible to apply fine-tuning strategy to all the layers of the convolutional neural network, or we can just keep input and middle layers fixed and only apply fine tuning method on some higher level portion of the network.

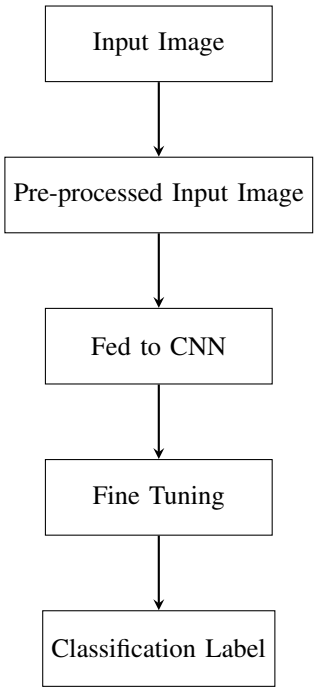
In our project, we have used Fine Tuning method of Transfer Learning, but the challenge we have faced here is to apply what sort of transfer learning strategy. After rigorous searching, we came to a conclusion that the fine tuning method would best suit for our problem statement.

Taking these ideas from the concepts of the convolutional neural networks and transfer learning, we have implemented two of the famous and leading CNN architectures, called the AlexNet and VGG 16. AlexNet is the pioneer of the winner-list of the CNN architectures which established impressive accuracies with lower error rates on the famous ImageNet Database Challenge.ImageNet is a gigantic database of pictures for scholarly analysts. Consistently the researchers who run ImageNet have a image recognition competition.

Fig. 3. An image from the dataset

The objective is to compose a bit of programming nowadays typically a neural system that can effectively recognize the class for the test images. AlexNet came into lime-light in the year 2012 when Alex Krizhevsky et al have mentioned this in their work ImageNet Classification with Deep ConvolutionalNeural Networks. It has achieved a very impressive top 5 error rate of 16.4%. It has 5 convolution layers, 3 max pools along with 2 norm layers, and 3 dense layers at the end. On the other hand, the VGG 16 architecture is trained by Oxford's renowned Visual Geometry Group. The architecture is vividly presented in the figure. As evident, it is more complex than the AlexNet since VGG 16 has more layers thus making it analyze deeper. This is an important property that made it achieve higher accuracies than the AlexNet on the same ImageNet database. It has successfully achieved a top error rate of 7.3% which is quite boisterous to think that this helps achieve any computer vision task.

Fig. 4. WorkFlow



The major work we have done is to implement the AlexNet and VGG16 to achieve the task of recognizing American Sign Languages through hand gestures. The main objective is to classify the input image of the American Sign Language hand gesture into the real world English letter. As suggested in many papers we have refered, we initially opted to go with the VGG 16 architecture. We have applied the transfer learning technique by freezing some layers and fine-tuned the final dense layers in order to prepare the network to learn the patterns from our dataset. This has given as good as 100% accuracy on our dataset. We have later discovred that the dataset we have choosen is very strutured. Elaborately, the images are all already preprocessed and really structured, against the reality where the images are expected to have many other factors in them like light conditions, zoom in and zoom out versions of the sign, etc. We tried by imparting noise to the images in the dataset we already had. But it had shaped out the images making it more of noise rather than the image of the sign and hard to even preproces it. We then found a new and good enough data set, containing as many as 3000 images for each sign with varied factors like those mentioned above. We have also wanted to reduce the network complexity and train, validate and test it using the dataset. We got 0.99 accuracy on a scale of 0 to 1 for both train and test accuracies.

V. Results

Thanks to the modern technology of the Deep Convolution Neural Networks like the above mentioned ones, we are able to train the according to our datasets and use them for achieving the classification task. We have achieved 99% accuracy on both training and validation sets.

VI. conclusion

Transfer learning in convolutional neural networks are really a boon to the computer vision society. We are now able efficiently use the knowledge acquired from the other datasets through the pre-trained weights, and able to use all the essential features extracted by the neural network and use them in the final classification tasks. The combination of transfer learning and deep convolutional neural networks has been a major breakthrough in solving such important tasks.

References

[1] An Efficient Algorithm for Sign Language Recognition by Pratibha Pandey et al.

[2] American Sign Language Recognition usinf Leap Motion Sensor by Ching-Hua Chan et al.

[3] Real-time American Sign Language Recognition with Convolutional Neural Networks by Brandon Garcia et al.