**Use CNN for image classification**

**Sign Language Recognition using Convolutional Neural Networks – (temporal segmentation 1 wadagath)**

The system that (8) has implemented use the Microsoft Kinect, convolutional neural networks (CNNs) and GPU acceleration.

able to recognize 20 Italian gestures cross-validation accuracy of 91.7%.

use the data set from the ChaLearn Looking at People 2014 [5]

preprocessing with filtering, resizing.

CNN,

use max-pooling.

using 2D convolutions resulted in a better validation accuracy than 3D convolutions.

During training, dropout [9] and data augmentation are used as main approaches

to reduce overfitting.

**Dataset**

**(8)** used the data set from the ChaLearn Looking at People 2014 [5]

this type of communication is impersonal and slow in face-to-face conversations. For example, when an accident occurs, it is often necessary to communicate quickly with the emergency physician where written communication is not always possible. {for introduction}

Preprocessing –

First they have crop the highest hand and the upper body using the given joint information. results in four video samples (hand and body with depth and gray-scale) of resolution 64x64x32 (32 frames of size 64x64). Furthermore, the noise in the depth maps is reduced with thresholding, background removal using the user index, and median filltering.

Proposed architecture -

For the pooling method, we use max-pooling: only the maximum value in a local neighborhood of the feature map remains. Use 2D convolutions resulted in a better validation accuracy than 3D convolutions. The architecture of the model consists of two CNNs with Each CNN is three layers deep. Finally have a classical ANN with one hidden layer. Also, local contrast normalization (LCN) as in [10] is applied in the first two layers and all artificial neurons are rectified linear units (ReLUs [14], [6])

Generalization and Training –

During training dropout and data augmentation are used as main approaches to reduce overfitting.

Result –

observed a validation accuracy of 91.70%. The accuracy on the test set is 95.68% and we observe a 4.13% false positive rate, caused by the noise movements.

American Sign Language Character Recognition Using Convolution Neural Network

There are basically two types of approaches for hand gesture recognition: vision-based approaches and data glove approaches.

The reason for choosing a system based on vision relates to the fact that it provides a simpler and more intuitive way of communication between a human and a computer.

The image dataset consists of 2524 ASL gestures. Out of this dataset 75% images were used for training and remaining 25% images were used for testing. Depth image dataset

Preprocessing -

read and resize each of the image to the similar size of 224x224 pixel. Only when all of the images in the dataset are of the same size can the images be fed into a neural network for training. The mean value of RGB over all pixels was subtracted from each pixel value. Normalization?

Augmentation –

augmented to produce several images from each image, thus increasing the size of the dataset and also tackling the problem of overfitting. Maximum ranges or degrees for shear, zoom, horizontal and vertical shifting were specified.

Architecture –

[The Model Used for training the dataset was inspired from VGG16 model]

VGG 16 model which is a deep convolutional neural network model used as the architecture. max pooling layer used as the pooling layer. only the maximum value in a local neighborhood of the feature map remains.

ReLU used.

The accuracy of the model obtained using Convolution Neural Network was 96%.

**Real-time American Sign Language Recognition with Convolutional Neural Networks**

many of them require a 3-D capture element with motion-tracking gloves or a Microsoft Kinect, and only one of them provides real-time classifications. The constraints imposed by the extra requirements reduce the scalability and feasibility of these solutions. {introduction}

We utilize a pre-trained GoogLeNet architecture trained on the ILSVRC2012 dataset, as well as the Surrey University and Massey University ASL datasets in order to apply transfer learning to this task. We produced a robust model that consistently classifies letters a-e correctly with first-time users and another that correctly classifies letters a-k in a majority of cases. Given the limitations of the datasets and the encouraging results achieved, we are confident that with further research and more data, we can produce a fully generalizable translator for all ASL letters.

uses Convolutional Neural Networks (CNN) in real time to translate a video of a user’s ASL signs into text. This is done by 3 steps

1. Obtaining video of the user signing (input)

2. Classifying each frame in the video to a letter

3. Reconstructing and displaying the most likely word from classification scores (output)

Challenges

Environmental concerns (e.g. lighting sensitivity, background, and camera position)

• Occlusion (e.g. some or all fingers, or an entire hand can be out of the field of view)

• Sign boundary detection (when a sign ends and the next begins)

• Co-articulation (when a sign is affected by the preceding or succeeding sign)

Our system features a pipeline that takes video of a user signing a word as input through a web application. We then extract individual frames of the video and generate letter probabilities for each using a CNN (letters a through y, excluding j and z since they require movement). With the use of a variety of heuristics, we group the frames based on the character index that each frame is suspected to correspond to. Finally, we use a language model in order to output a likely word to the user.

Approach and Methods

Our ASL letter classification is done using a convolutional neural network (CNN or ConvNet). CNNs are machine learning algorithms that have seen incredible success in handling a variety of tasks related to processing videos and images. Since 2012, the field has experienced an explosion of growth and applications in image classification, object localization, and object detection. A primary advantage of utilizing such techniques stems from CNNs abilities to learn features as well as the weights corresponding to each feature. Like other machine learning algorithms, CNNs seek to optimize some objective function, specifically the loss function. We utilized a softmax-based loss function

Use transfer learning, Transfer Learning is a machine learning technique where models are trained on (usually) larger data sets and refactored to fit more specific or niche data.

**Dataset Description**

Use color images. They are close-ups of hands that span the majority of the image surface.

How dataset separate?

Since there was little to no variation between the images for the same class of each signer, we separated the datasets into training and validation by volunteer. Four of the five volunteers from each dataset were used to train, and the remaining volunteer from each was used to validate. We opted not to separate a test set since that would require us to remove one of four hands from the training set and thus significantly affect generalizability. Instead, we tested the classifier on the web application by signing ourselves and observing the resulting classification probabilities outputted by the models.

**Pre-processing and data augmentation’**

Both datasets contain images with unequal heights and weights. Hence, we resize them to 256x256 and take random crops of 224x224 to match the expected input of the GoogLeNet.

**Experiments, Results and Analysis**

**5.1. Evaluation Metric**

We evaluate two metrics in order to compare our results with those of other papers. The most popular criterion in the literature is accuracy in the validation set, i.e. the percentage of correctly classified examples. One other popular metric is top-5 accuracy, which is the percentage of classifications where the correct label appears in the 5 classes with the highest scores. Additionally, we use a confusion matrix, which is a specific table layout that allows visualization of the performance of the classification model by class.

Conclusion

We implemented and trained an American Sign Language translator on a web application based on a CNN classifier. We are able to produce a robust model for letters a-e, and a modest one for letters a-k (excluding j). Because of the lack of variation in our datasets, the validation accuracies we observed during training were not directly reproducible upon testing on the web application. We hypothesize that with additional data taken in different environmental conditions, the models would be able to generalize with considerably higher efficacy and would produce a robust model for all letters.

**Real-time sign language recognition based on neural network architecture**

Mekala et al. classified video of ASL letters into text using advanced feature extraction and a 3-layer Neural Network [8]. They extracted features in two categories: hand position and movement. Prior to ASL classification, they identify the presence and location of 6 “points of interest” in the hand: each of the fingertips and the center of the palm. Mekala et al. also take Fourier

Transforms of the images and identify what section of the frame the hand is located in. While they claim to be able to correctly classify 100% of images with this framework, there is no mention of whether this result was achieved in

the training, validation or test set.

**In real-time, it is highly essential to have an autonomous translator that can process the images and recognize the signs very fast at the speed of streaming images. In this paper, architecture is being proposed using the neural networks identification and tracking to translate the sign language to a voice/text format. Introduction of Point of Interest (POI) and track point provides novelty and reduces the storage memory requirement.**

One important means of communication method for the hearing impaired community is the use of sign language, as in [1]

{introduction}

Signing takes place in a 3D space, called signing space close to the trunk and the head, as in [7]. {intro}

They will get a video sequence of the signer as the input from camera.

Preprocessing –

In order to satisfy the memory requirements and the environmental scene conditions, preprocessing of the raw video content is highly important [14].

A moving average or median filter used as filteration preprocess. Next did the Background subtraction.

Feature Extraction –

Because this is Neural network. They did feature extraction. **Introduction of Point of Interest (POI) where** The state of the hand gestures are given by the attributes called Point of Interest (POI) of the hands. The feature vector consists of 55 features.

Neural Network system –

They used a new scheme called combinational neural networks (CNN). A three layer network called back propagation is used to build the CNN.

In general, as the output classes are increased and the non-linearity of differentiation in the classes increases, more feature vectors are necessary for the object recognition technique. (fact)

Results

This sign language recognition approach requires a computer with at least 1GHz processor and at least 256 MB of free RAM. The training set consists of all alphabets A to Z (26 patterns).