A deep learning approach to sign language recognition using stacked autoencoders and neural networks

Pablo Rivas

Jan/24/17

1 Introduction

Language is an essential part of being human as it enables us to communicate with others. Some people have to develop additional means of communication for different reasons. A popular alternative language is through signs. This type of language is based on symbols or figures produced using the hands.

Sign language recognition is a task that has been studied in the last few years [**?**]. We believe this is an important problem to address because it directly affects the quality of life of people who need to communicate using sign language. The task is to develop algorithms that can recognize different signs in an alphabet and that can do so in an efficient manner. Ideally, the methodologies should be fast to the point of enabling the near real-time processing of signs [**?**].

In this research we deal with the recognition of signs over the american sign language (ASL). The proposed approach uses depth images of subjects making different signs, building upon the work of B. Kang, et.al. in [**?**].

Typical approaches involve the using hidden Markov models [**?**], and a combination of them with other discriminative functions for feature extraction in multi-stage architectures [**?**]. Other major alternatives included the exploration of neural strategies combined with fuzzy systems [**?**]. For a more detaile review of alternatives for generic hand gesture recognition we can turn to the work in [**?**].

However, with the dramatic attention that deep learning has gained recently in the machine learning and the immage processing for pattern recognition communities, we are motivated to similarly explore this new algoritmic alternative to see if it offers better solutions to open problems. Most recently, the authors in [**?**] have explored a deep learning approach based on convolutional neural networks (CNNs) achieving outstanding results. The research presented here also explores a novel deep learning approach but based in autoencoders and neural networks. We will show that this alternative approach achieves great performance and efficiency.

The rest of the paper is organized as follows.

2 Background and Related Work

Recently, J. Nagi et.al. [**?**] studied a deep neural network approach for recognizing six different hand gestures given to a robot. Using CNNs the authors achieve a 96% accuracy rate. Then, M. Van den Bergh et.al. [**?**] explored the idea of using depth-sensor imagery to establish an algorithm for hand segmentation. The authors ultimately used this to recognize six different hand gestures in 3D. They achieved a 99.54% recognition rate. Some of the key commonents of the overall architecture were the usage of wavelets for image filtering and neural networks.

Moreover, B. Kang et.al. [**?**] took advantage of the recent success of CNNs for pattern recognition over images. They propose a system that makes use of depth-sensor images to study further the problem of hand gesture recognition. When recognizing signs within the same subject they averaged a 99.99% accuracy rate over five subjects. However, when recognizing signs across different subjects, their accuracy rate droped to 83.58%.

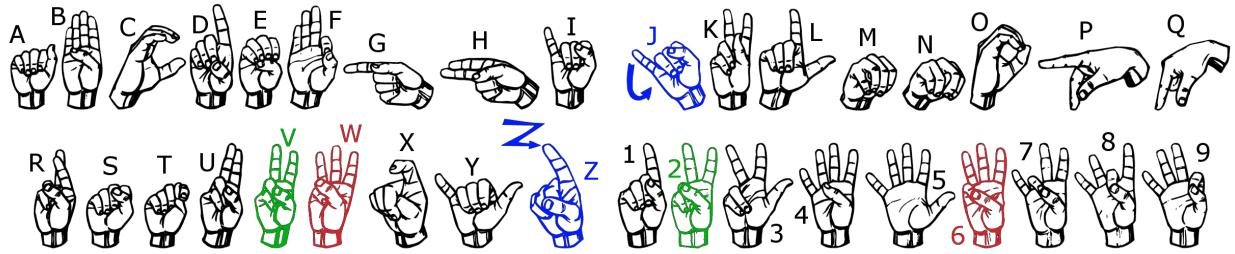

Figure 1: The American Sign Language (ASL) displaying 26 signs corresponding to the English alphabet (A to Z) and 9 signs for numbers (1 to 9).

Figure 2: Examples of ASL hand gestures corresponding to the number one and number two with variations in the horizontal axis and its inclination with respect to a depth-sensor camera.

Our research aims to further study the problem posed in [**?**] of increasing accuracy rates over different subjects. In the next section we discuss the dataset used and propose a deep learning approach using autoencoders and a neural network.

3 Methodology

3.1 Data

This research focuses in the ASL as a case study [**?**]. The ASL has 26 hand gestures corresponding to the letters in the English alphabet; it also contains 9 hand gestures for every single number. Figure 1 depicts the gestures pertaining to our study.

We use the data provided in [**?**] that was released in 2015. This dataset consists of images acquired with a three-dimensional depth-sensor camera. There is a total of $31,000$ images available with a resolution of 256×256. The images correspond to already segmented dept-images that contain the area of the hand. The authors use a rather simple algorithm to make the segmentation based on the fact that in all depth images the hand is the closest object with respect to the camera.

Not all the 35 signs (26 letters and 9 numbers) were considered in this study. Some signs were exluded and others considered jointly as follows.

The signs corresponding to J and Z clearly require a sequence of images rather than a single instance, as can be seen in Figure 1. This goes beyond the scope of this project; although the detection of signs from live-video sequences is part of a bigger project, we will not address it in this paper. For that reason these two signs were exluded.

The signs corresponding to V and 2, by observation, are nearly identical and their distinction requires contextual information. Similarly, the signs for letter W and number 6 require being interpreted in context. For this reason, these signs were considered as the same sign, leaving a total of 31 signs.

There is a total of five different subjects and each subject produced 200 images of each sign. This represents a total 1,000 images per sign and 31,000 images overall. The images collected capture each subject making a hand gesture moving across the horizontal axis and varying inclination of the hand gesture, as depicted in Figure 2.

3.2 Sparse Autoencoders

A sparse autoencoder is usually categorized as a neural network with unsupervised learning [**?**]. In a general sense an autoencoder is trained to output an approximation of the input provided a deep architecture of layered neurons that encode and decode based on the input stimuli. In our research we use a sparse autoencoder that specifically minimizes a modified loss function based on the mean squared error.

Let $\mathbf{x} \in \mathbb{R}^d$ be a d-dimensional input vector. Then the loss function of the common sparse autoencoder is defined as follows:

$$L = \frac{1}{N} \left\| \mathbf{x}_n - \hat{\mathbf{x}}_n \right\|_2^2 + \theta_w \frac{1}{2} \sum_{l=1}^{L} \left\| \mathbf{w}^l \right\|_2^2 + \theta_s \sum_{m=1}^{M} KL\left(\theta_\alpha \left\| \bar{\alpha}_m \right.\right) \tag{1}$$

where N is the total number of training samples; $\hat{\mathbf{x}}_n$ is the learned (or encoded) approximation of the n-th input vector $\mathbf{x}_n$; θ_w controls the sparsenes of the weights of the network, $\mathbf{w} \in \mathbb{R}^d$; L denotes the number of layers in the deep network; θ_s regulates the sparsity of the activation functions' output, α, in every neuron in the network; M is the total number of neurons in the deep network; and $KL(\cdot)$ is the Kullback-Leibler divergence function [**?**] used to measure how much the observed average activation of the m-th neuron, $\bar{\alpha}_m$, actually deviates from the desired average output, θ_α.

The Kullback-Leibler divergence function can be defined as follows [**?**]:

$$\sum_{m=1}^{M} KL\left(\theta_\alpha \left\| \bar{\alpha}_m \right.\right) = \sum_{m=1}^{M} \theta_\alpha \log\left(\frac{\theta_\alpha}{\bar{\alpha}_m}\right) + (1-\theta_\alpha) \log\left(\frac{1-\theta_\alpha}{1-\bar{\alpha}_m}\right) \tag{2}$$

where the average output of the m-th neuron, $\bar{\alpha}_m$, at the l-th layer is given by

$$\bar{\alpha}_m = \frac{1}{N} \sum_{n=1}^{N} \psi\left(\mathbf{w}_m^{(l)T} \mathbf{x}_n + b_m^{(l)}\right) \tag{3}$$

where $\psi(\cdot)$ is the neuron's activation function, and $b_m^{(l)}$ is the bias term for the m-th neuron at the l-th layer. In this research we specifically use logistic (sigmoid) activation functions, $\psi(z) = 1/(1-e^{-z})$, for any given value of z.

By observing the mathematical form of (1), it is feasible to apply an accelerated form of guided learning known as scaled conjugate gradient (SCG) descent [**?**]. SCG has proven to be a reliable and efficient form of minimizing loss functions such as the one for the sparse autoencoder, overcoming the shorcomings of a traditional conjugate gradient and a back-propagation with gradient descents [**?**].