Continuous Sign Language Recognition from Wearable IMUs using Deep Capsule Networks and Game Theory

Karush Suri, Rinki Gupta
Department of ECE, Amity School of Engineering & Technology,
Amity University, Noida, India

Abstract

The work presented here proposes a novel one-dimensional deep capsule network (CapsNet) architecture for continuous Indian Sign Language recognition by means of signals obtained from a custom designed wearable IMU system. The performance of the proposed CapsNet architecture is assessed by altering dynamic routing between capsule layers. Performance of the model is compared to that of foundational convolutional neural networks (CNNs) in terms of accuracy, loss and learnt activations. The proposed CapsNet yields improved accuracy values of 94% (for 3 routing) and 92.50% (for 5 routings) in comparison to CNNs which yield 87.99%. Improved learning of the architecture is also validated by spatial activations depicting excited units at the predictive layer. For the purpose of evaluating relative performance and competitive nature of models, a novel non-cooperative pick game is constructed. Both models compete with each other in order to reach their best responses. Higher value of Nash equilibrium for CapsNet as compared to CNN indicates the suitability of the proposed approach.

Proposed Algorithm

A. Data Corpus:

The dataset of IMU signals was constructed by making use of the GY-80 multiboard and Arduino UNO board, which contains the ATmega328P microcontroller. The GY-80 board consists of tri-axial accelerometer ADXL345 and gyroscope L3G400D, which measure acceleration (in m/s^2) and turn rate (in deg/s), respectively at 100 Hz. Both accelerometer and gyroscope have 3 degree-of-freedom each, giving a total of 6 channels of data. Fig. 1a depicts the GY-80 board consisting of IMU instruments. Signals from IMU sensors were recorded from a total of 10 different subjects out of which 5 are female and 5 are male. All the subjects fall in the age range of 21-49 years and 2 subjects were left-handed. Signs were recorded from the Indian Sign Language (ISL) in a sequence, resulting in the formation of a complete sentence. A total of 20 sentences were gathered with each sentence consisting of 2-4 signs spaced approximately 1s apart. Each subject performed 10 repetitions of a sentence.

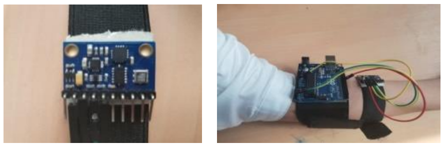

Fig 1. (a) GY-80 sensor board, (b) Placement of experimental apparatus

B. CapsNet Architecture:

The CapsNet architecture proposed in this work for the recognition of IMU signals consists of a 12x12 convolutional filter for both the capsule layers with a stride of 1. A total of 256 filters have been used for convolution. The Primary Caps layer has 20 channels with 5 dimensional capsules indicating that each capsule consists of 5 convolutional filters of size 12x12 with a stride of 2. Each primary capsule receives the input of all 256x144 1st convolutional layer (Conv1) units. Both the layers are activated with rectified linear unit (ReLU) activation. The Digit Caps layer receives 20x1 capsule outputs from the Primary Caps layer as 5 dimensional vectors. Each capsule in the Digit Caps grid shares its weight with other capsules. A total of 10 dimensions are accumulated per class in the Digit Caps layer. In the case of routing, all the logits are initialized to zero, thus indicating equal probability for the capsule output to be routed to the next unit. For prediction purpose a fully connected Sigmoid layer consisting of 9324 units is inserted. One-dimensional scalar outputs from the Digit Caps layer are received by the fully connected units which lead to the prediction of the correct class. Optimization of the model is conducted by making use of the Adam optimizer with Amsgrad gradient optimization.

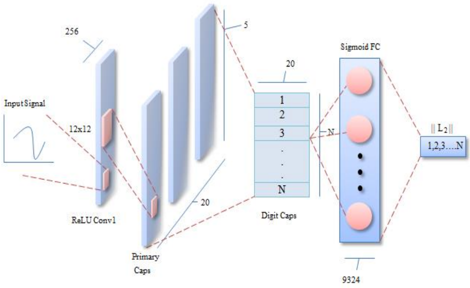

Fig.2: One-Dimensional CapsNet Architecture for Signed Sentence Recognition.

C. Non-Cooperative CapsNet Games:

Optimized models may be compared using competitive or non-competitive games [28]. Here, a non-cooperative pick game is constructed for comparing the performance of different CapsNet architectures and the conventional CNN. If there are '*x*' players competing in a game then the condition '*P*' for a player to win the game is mathematically expressed as-

$$P(1): (c_j)_1 = c_{true} ;\ P(2): (c_j)_2 = c_{true};\ .\ .\ .\ .\ .\ P(x): (c_j)_x = c_{true}$$

Now, for a player 'z' to win the game-

$$\exists! z: W(z)\ ;\ \ z \epsilon \{1,2,.\ .\ .\ .\ x\}$$

Thus, there is exactly one player '*z*' in the set of possible winning criteria '*W*' which satisfies the condition. However, the condition for the game to end up as a draw is expressed as-

$$\exists z \mid \vee z \mid \vee !z : W(z)\ ;\ \ z \epsilon \{1,2,.\ .\ .\ .\ x\}$$

Algo 1.

```
for $i$=1:$n$
    $(c_i)_{\text{CNN}} \leftarrow \text{pred}((D_i)^{train})$
    $(c_i)_{\text{CapsNet}} \leftarrow \text{pred}((D_i)^{train})$
for $j$=1:$m$
    $(c_j)_{\text{CNN}} \leftarrow \text{pred}((D_i)^{test})$
    $(c_j)_{\text{CapsNet}} \leftarrow \text{pred}((D_i)^{test})$
    if $((c_j)_{\text{CapsNet}} == (c_j)_{\text{true}}$ & $(c_j)_{\text{CNN}} \neq (c_j)_{\text{true}})$
        Winner $\leftarrow$ CapsNet
    else if $((c_j)_{\text{CNN}} == (c_j)_{\text{true}}$ & $(c_j)_{\text{CaspNet}} \neq (c_j)_{\text{true}})$
        Winner $\leftarrow$ CNN
```

Experimental Results

A. CapsNet Recognition:

The performance of the proposed model is compared to the conventional state-of-the-art CNN having similar hyper-parameters. Fig 3a and Fig 3b depict the training and validation accuracies of the models over 50 iterations respectively. Peak accuracy values for the CapsNet architecture with 3 routings are observed to be 99.72% during training and 94.00% during validation. Figs. 4a and 4b highlight the same variation by means of optimization loss during training and validation, respectively. CapsNet with 3 routings depicts better performance with approximately 0.01 units of loss during training and validation. Another informative measure to assess the predictive behavior of these models is the number of false predictions.

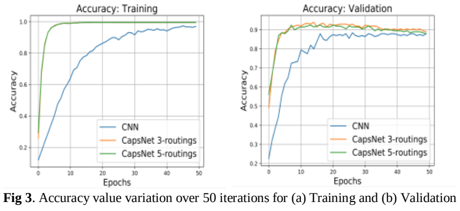

Fig 3. Accuracy value variation over 50 iterations for (a) Training and (b) Validation

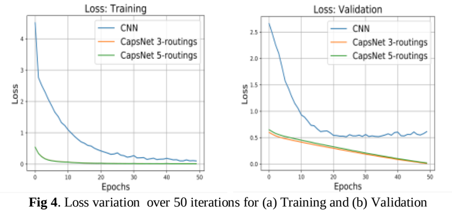

Fig 4. Loss variation over 50 iterations for (a) Training and (b) Validation

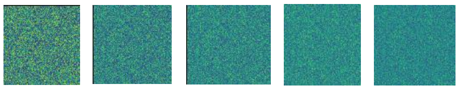

Fig 5. Learnt Activations at the final layer of CNN over (a) 1 iteration, (b) 10 iterations, (c) 20 iterations, (d) 30 iterations and (e) 40 iterations

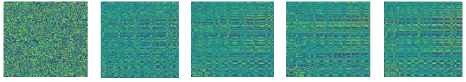

Fig 6. Learnt Activations at the final layer of CapsNet over (a) 1 iteration, (b) 10 iterations, (c) 20 iterations, (d) 30 iterations and (e) 40 iterations

Analyzing learnt activations for the model plays a significant role in the evaluation of learning as these weights tend to change over each iteration. Fig. 5 and 6 indicate the improvement in weights after every 10th iteration for CNN and CapsNet architecture. Activation of the last layer tends to improve by virtue of the learnt weights. As seen from Fig. 6a to 6d, the multiplicative values modify over passing iterations producing a regular pattern in Fig. 6d, which depicts the weights in the spatial domain and a definite structure in the background for CapsNet architecture. This definite structure indicates the activated units in the layer. However, no such regular pattern is observed to emerge in the case of CNN (Fig. 5a to 5d).

B. Non-Cooperative Games:

A total of 3 games are played, each between two sets of players at a given time. Each game of the three games consist of the two players. These games based on their players may be mathematically summarized as,

$$Game\text{-}1\text{-}\ P(CNN):(c_j)_{CNN}=c_{true}\ ;\ P(CapsNet\text{-}3):(c_j)_{CapsNet\text{-}3}=c_{true}$$
$$Game\text{-}2\text{-}\ P(CNN):(c_j)_{CNN}=c_{true};\ P(CapsNet\text{-}5):(c_j)_{CapsNet\text{-}5}= c_{true}$$
$$Game\text{-}3\text{-}\ P(CapsNet\text{-}3):(c_j)_{CapsNet\text{-}3}=c_{true};P(CapsNet\text{-}5):(c_j)_{CapsNet\text{-}5}=c_{true}$$

Table 1. Performance Comparison of CapsNet architecture with CNN

Architecture	Optimization Loss		Classification Accuracy		Nash Equilibrium	
	T	V	T	V	T	V
CNN	0.07	0.11	93%	87%	0.93	0.88
CapsNet 3 routings	**0.01**	**0.01**	**99%**	**94%**	**0.99**	**0.94**
CapsNet 5 routings	0.01	0.02	99%	92%	0.99	0.93

Table 1 summarizes the performance of CapsNet models in comparison to CNN. Here, 'T' and 'V' indicate the results for training and validation phases respectively. Improved results of the CapsNet architecture in comparison to the conventional CNN highlight the appropriateness of dynamic routing with nested layers in the capsule theory.

Conclusions

In this work, recognition of sentences signed according to the Indian sign language is performed using signals recorded from a wearable IMU device. The sentence recognition is carried out using a novel one-dimensional CapsNet architecture. Improved accuracy value of 94% is observed for CapsNet with 3-routings in comparison to CapsNet with 5-routings and CNN, which yield 92.5% and 87.99% accuracy values, respectively. Furthermore, a non-cooperative pick game is constructed for assessing the relative performance of the models. CapsNet architecture presents a better value of the best response at Nash Equilibrium asserting the suitability of the proposed approach.

Acknowledgement

The authors would like to thank the volunteers who helped in recording the data. The authors also recognize the funding support provided by the Science & Engineering Research Board, a statutory body of the Department of Science & Technology (DST), Government of India (ECR/2016/000637).