Recognizing Handshapes using Small Datasets

Ulises Jeremias Cornejo Fandos[1⋆], Gaston Gustavo Rios[1,3⋆], Franco Ronchetti[1], Facundo Quiroga[1], Waldo Hasperué[1,2], and Laura Lanzarini[1]

[1] Instituto de Investigación en Informática III-LIDI, Facultad de Informática, Universidad Nacional de La Plata
[2] Investigador Asociado - Comisión de Investigaciones Científicas (CIC)
[3] Becario de entrenamiento - Comisión de Investigaciones Científicas (CIC)
{ucornejo,grios}@lidi.info.unlp.edu.ar

Abstract. Advances in convolutional neural networks have made possible significant improvements in the state-of-the-art in image classification. However, their success on a particular field rests on the possibility of obtaining labeled data to train networks. Handshape recognition from images, an important subtask of both gesture and sign language recognition, suffers from such a lack of data. Furthermore, hands are highly deformable objects and therefore handshape classification models require larger datasets.
We analyze both state of the art models for image classification, as well as data augmentation schemes and specific models to tackle problems with small datasets. In particular, we perform experiments with Wide-DenseNet, a state of the art convolutional architecture and Prototypical Networks, a state of the art few-shot learning meta model. In both cases, we also quantify the impact of data augmentation on accuracy.
Our results show that on small and simple data sets such as CIARP, all models and variations of achieve perfect accuracy, and therefore the utility of the data is highly doubtful, despite its having 6000 samples. On the other hand, in small but complex datasets such as LSA16 (800 samples), specialized methods such as Prototypical Networks do have an advantage over other methods. On RWTH, another complex and small dataset with close to 4000 samples, a traditional and state-of-the-art method such as Wide-DenseNet surpasses all other models. Also, data augmentation consistently increases accuracy for Wide-DenseNet, but not fo Prototypical Networks.

Keywords: sign language, hand shape recognition,convolutional neural networks,densenet, prototypical networks, small datasets

1 Introduction

Sign Language Recognition is a field in the intersection of computer vision and language translation that seeks to create systems capable of translating videos of people speaking in sign language into text.

⋆ equal contribution

In recent years, new advances in machine learning using models such as convolutional and recurrent neural networks have improved our ability to tackle complex recognition problems such as speech recognition, image classification or object detection[5]. These advances are fueled by a combination of improvements in three areas; better datasets, better models, and more compute power. While the last two are mostly independent of a particular field, the availability of quality datasets for a given field limits the application of these new advances. For example, common image classification datasets such as MNIST, CIFAR10, CIFAR100 and ImageNet contain thousands of examples per class [4].

The process of recognizing a sign language consists of several steps, ranging from image preprocessing, body part detection, facial expression recognition, handshape recognition, language modeling and language translation. Of these steps, handshape recognition plays the most crucial role in the interpretation of signs[10,15]. However, sign language recognition cannot currently take full advantage of state-of-the-art models, since the availability of labeled, quality data for training models is very limited [10]. Lack of data also impairs the development of accurate handshape recognition models [10].

In particular, Convolutional Neural Networks (CNN), a type of neural network that takes advantage of convolutional layers to learn arbitrary convolutional filters, have proven very effective at image classification [5], including the classification of handshapes in images [14]. However, in most applications, convolutional neural networks are trained using thousands of images per class. In handshape recognition tasks, the datasets are considerably smaller and of lower quality, and therefore the performance of the models suffers accordingly [10,18,14].

In this work we propose to evaluate and compare new methods devoted to deal with small datasets in order to improve the current state-of-the-art in hand shape recognition for sign language.

Our approach consists of comparing different techniques for improving model performance in these conditions: data augmentation and prototypical networks for few shot learning and semi supervised learning. Our data augmentation scheme consists of basic augmentation operators such as rotations, translations and crops. We compare this with new technique approach.

In the following subsection we summarize previous efforts on training CNN on handshape datasets. Section 2 describes the datasets and models we employed in our experiments, which are detailed along with results in Section 3, and Section 4 contains the conclusion of our work.

1.1 Related Work

Recent years have seen the rise in the use of deep learning models for sign language recognition, specifically the use of convolutional neural networks to extract image features or directly classified hand images. [10] trained a CNN to recognize handshapes from the RWTH handshape dataset, which contains 3200 labeled samples and 50 different classes. The model was based on a pre-trained network with a VGG architecture, and employed a semi-supervised scheme to take advantage of approximately one million weakly labeled images, achieving

an accuracy of 85.50%. This constitutes the first attempt at adapting a model to overcome the low availability of labeled images for training. [16] employed a radon transform as a feature for an ad hoc classifier that employed clustering as a quantization step and K nearest neighbors for the final classification. They tested the model on the LSA16 dataset, which contains only 800 examples, obtaining an accuracy of 92.3%. [14] evaluated several CNNs on the LSA16 and RWTH datasets, including both vanilla and pre-trained models. The use of pre-trained models helps to alleviate the lack of labeled data, since pretraining the convolutional filters establishes a prior that a further classifier can exploit for handshape recognition. This work is the second and last instance we found where a specific strategy was employed to alleviate the lack of data. Their best models of an accuracy of 95.92% for LSA16 and 82.88% for RWTH. [12] trained a simple neural network to classify a new dataset they created, which contains 6000 examples and 10 classes, reaching an accuracy of 99.20%. [18] train a CNN on a custom dataset with 36 classes, 8 subjects and 57000 sample images. However, the samples correspond to video sequences and therefore are highly correlated; while there are approximately 2000 images per class, there are only eight image sequences, one for each subject. Each of this image sequences contains approximately 250 images which are highly correlated, and therefore it is best to consider the dataset as having eight image sequences per class. They obtained an accuracy of 94.17%, [2] trained a simple CNN with only 6 layers using the ASL Finger spelling dataset, obtaining an accuracy of 80.34%. The dataset consists of 60000 images of 25 different classes, but they were captured as videos so they are also highly correlated as in the previous case. [3] employed the Jochen Triesch Database (JTD), which contains only 10 classes and 72 samples per class, as well as the NAO Camera Hand Posture Database, which contains 4 classes and 400 examples per class. They trained a simple CNN with a multichannel image containing the results of the Sobel operator as input, obtaining an F-score of 94% and 98% in each dataset perspective. [1] trained a deep CNN on the Hand Gesture Dataset LPD, which contains 3250 images of only 6 classes, obtaining an accuracy of 99.73%.

This brief review confirms our previous statement that while CNN are being consistently applied to handshape recognition tasks, most of these datasets are small and ad hoc, that is, recorded specifically for the purpose of testing a single model and not developed with the intent of providing a benchmark and complete training set for handshape recognition models. It is also worth noticing that some datasets are so small that it is very easy to obtain near-perfect performance with simple models. Also, many datasets are not readily available, given that the authors have not publish the data and do not provide any means of obtaining it. We note that the RWTH and LSA16 are both publicly available and current models have been shown to achieve less than perfect accuracy for them. While the dataset in [12] has been easily solved, it is interesting because it targets general handshapes instead of those specific to sign language. We will call this dataset CIARP.

2 Datasets and Models

We selected three datasets, LSA16 [16], RWTH-PHOENIX-Weather (RWTH) [11] and CIARP [12], because they contain images whose setting varies greatly, have been evaluated already, and posses different quantities of examples or distributions of samples per class.

We employed two different classification models to analyze their ability to learn from these small handshape datasets; Prototypical Networks [17] and DenseNet [8]. Prototypical Networks is a model that was designed explicitly to deal with small sample sizes. On the other hand, DenseNet is currently the state of the art in image classification with convolutional neural networks, and while it has not been explicitly designed for small datasets, it has shown exceptional performance in many different tasks.

We also experimented with data augmentation to analyze its capacity to compensate for the lack of data.

In the following subsections we describe in more detail the selected datasets and models.

2.1 Datasets

LSA16 [16] contains images of 16 handshapes of the Argentinian Sign Language (LSA), each performed 5 times by 10 different subjects, for a total of 800 images of size 32x32. The subjects wore colored hand gloves and dark clothes on a white background. The dataset is balanced, with 50 images per class. There is only one hand in each image which are centered and isolated from the background.

RWTH [11] is composed of a selection of hand images of size 132x92 cropped from videos of the sign language interpreters at the German public tv-station PHOENIX. There are a total of 45 different hand signs. The interpreters wore dark clothes in front of an artificial grey background. Many images posses significant movement blur, others contain both hands of the interpreter and hands are not always perfectly centered.

The dataset is highly imbalanced with some classes having just 1 sample while others have as many as 529 samples. We removed classes that had less than 20 samples following [14], to guarantee a minimum amount of images per class for the networks to learn.

CIARP [12] contains 6000 images of size 38x38 acquired by a single color camera. The images were manually labeled and correspond to 10 classes of hand gestures. The hands are centered and were segmented from the background, which was replaced by black pixels. The small size of the images and low amount of classes give this dataset lower complexity compared to LSA16 and RWTH. The classes in the data set correspond to handshapes not based on sign language, but are similar enough that the comparison remains valid.

Fig. 1: Sample images from the LSA16 (first row), RWTH-PHOENIX-Weather (second row) and CIARP (third row) datasets.

2.2 Models

Prototypical Networks for Small Datasets Prototypical Networks [17] is a meta-learning model for the problem of few-shot classification, where a classifier must generalize to new classes not seen in the training set, given only a small number of examples of each new class. The ability of a algorithm to perform few-shot learning is typically measured by its performance on n-shot, k-way classification tasks. First a model is given a query sample belonging to a new, previously unseen class. Then, it's also given a support set, S, consisting of n examples, each from k different unseen classes. Finally, the algorithm then has to determine which of the support set classes the query samples belong to. Schemes for few shot classification tasks like Prototypical Networks can also be of use for training small datasets where all classes are known.

Prototypical Networks applies a compelling inductive bias in the form of class prototypes to achieve impressive few-shot performance. The key assumption is made is that there exists an embedding in which samples from each class cluster around a single prototypical representation which is simply the mean of the individual samples. This idea streamlines n-shot classification in the case of $n > 1$ as classification is simply performed by taking the label of the closest class prototype.

DenseNet We selected DenseNet as it is the current state of the art model in many domains and can handle small datasets with low error rate[13].

DenseNet [8] works by concatenating the feature-maps of a convolutional block to the feature-maps of all the previous convolutional blocks and using this value as input for the next convolutional block. This way each convolutional block receives all the collective knowledge of the previous layers maintaining the global state of the network which can be accessed.

Convolutional networks construct informative features by fusion both spatial and chanel-wise information within local receptive fields at each layer. Squeeze and excitation blocks (SE block) [7] focus on the chanel-wise information used in the convolutional layers. SE blocks improve the quality of representations produced by the network by modeling the interdependency between channels to

perform feature recalibration. SE blocks can be included in any model that uses convolutional layers to improve its performance at low computational cost. We added SE blocks to our DenseNet model to improve its performance.

Data Augmentation Image data augmentation is a set of techniques that aim at artificially augmenting the amount of data that can be obtained from the images in the dataset. These techniques modify the images in the dataset with a set of predefined operations to create new images that can be used to train a model. In this manner, we can compensate for the lack of variability in a small dataset[4].

3 Experiments

We performed classification experiments on LSA16, RWTH and CIARP handshape datasets. For each experiment, we split the dataset in training and test sets, with the latter taking 25% of the samples. The split was stratified, maintaining the proportion of samples of each class in both sets.

We applied normalization feature-wise substracting the mean and dividing by the standard deviation of each feature. For data augmentation we used horizontal flipping, a 10 and 30 degree rotation and a resampling of the images creating new versions of them with a different size reducing each image by 10% and 20% in width and height. We found that a 10 degree rotation gave better results because a rotation of 30 degrees showed to be too high for the nature of the datasets.

We made multiple experiments with Prototypical Networks and DenseNet to find out which hyperparameter configuration was the best for each dataset: with and without data augmentation. We describe the hyper parameters for each model/dataset combination.

3.1 Prototypical Network

As mentioned in section 2.2, we can use Prototypical Networks' ability to work with small datasets even if all samples are labeled.

Therefor we experimented with Prototypical Networks using an embedding architecture composed of four convolutional blocks. Each block comprises a {64, 128}-filter 3×3 convolution, batch normalization layer, a ReLU nonlinearity and a $2x2$ max-pooling layer.

We used the same encoder for embedding both support and query points. All of our models were trained with the ADAM[9] optimizer. We used an initial learning rate of 10^{-3} and cut the learning rate in half every 2000 episodes.

We trained prototypical networks using Euclidean distance in the 1-shot and 5-shot scenarios with training episodes containing 16, 20 and 10 classes (for LSA16, RWTH and CIARP respectively) and 5 query points per class. We found it advantageous to match the same value of n for train and test scenarios, and to use a higher value of k (more classes) per training episode. We computed

classification accuracy for our models by averaging over 1000 randomly generated episodes from the test set.

In the experiments performed with RWTH we used the same four-block embedding architecture by adding an eight-block architecture with the same layer composition with the idea of analyzing the need to increase the size of the network given the difficulty of the dataset. The difference in the results obtained in 1-shot and 5-shot scenarios for this dataset was very large. We found that 5-shot scenarios gave better results. Using this discovery we only used 5-shot learning in the remaining experiments.

The best configurations for all datasets is the 5-shot scenario with equals n for train and test scenarios by using more than or equal to 5 classes per training episode. Better results were obtained when the number of classes approaches the total amount of classes in the dataset except on CIARP where the best results were obtained when the number of classes per training episode is 5. In addition, the best configurations of the embedding architecture is a 64-filter for all datasets.

3.2 Wide-DenseNet

We employed a variation on DenseNet called Wide-DenseNet which follows the strategy used by wide residual networks.[6].

We employed a Wide-DenseNet including SE blocks after each dense and transition block. We performed a grid search of hyperparameters to find the model with the best accuracy, averaged over all datasets. We tried growth rate values of 32, 64 and 128 and depth of dense layers of no more than [6,12,24,16], where each number represents the number of dense blocks.

We trained the models using a batch size of 16, an initial learning rate of 10^{-3} with categorical cross entropy optimizer and 400 epochs with a maximum patience of 25. The resulting model used a growth rate of 64 and two dense blocks with 6 and 12 layers respectively, for all datasets.

3.3 Results

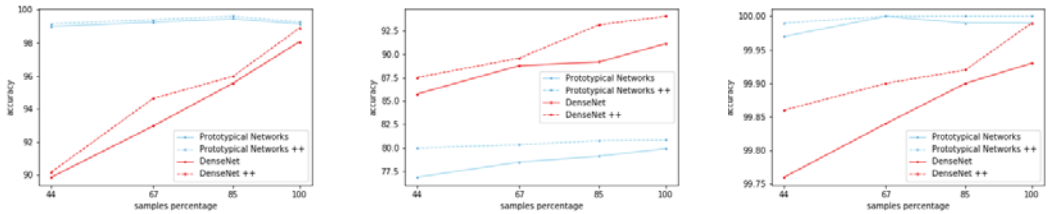
In table 1, we can observe that all models have a lower accuracy on the RWTH dataset, which is expected since it has more classes, unsegmented hand images and class imbalances. Prototypical Networks have similar accuracy for LSA16 and CIARP datasets beating the rest of the models, also expected since both datasets have very few examples. For LSA16 they achieve better accuracy than VGG16 and DenseNet; and for CIARP they achieve similar or better accuracy than LeNet CNN and DenseNet. The accuracy of DenseNet on the RWTH is slightly bigger than for other models. Our hypothesis was that Prototypical Networks obtained low accuracy because the images of the hands were unsegmented. It should be noted that the use of data augmentation did not bring significant improvements in the accuracy obtained in LSA16 and CIARP.

Another fact to consider is that better results were obtained with those parameters that reduced the size of the architectures.

Method	*LSA16*	*RWTH*	*CIARP*
LeNet [12]	-	-	99.20
Inception (fine-tuning) [10]	-	85.50	
VGG16 [14]	95.92	82.88	
Inception+SVM (pre-trained) [14]	93.67	78.12	-
DenseNet	98.07	91.10	99.93
DenseNet ++	98.90	**94.00**	99.99
Prototypical Networks	99.15	79.93	99.98
Prototypical Networks ++	**99.26**	80.85	**100.00**

Table 1: Accuracy of various convolutional neural network based models on three datasets: LSA16, RWTH and CIARP. Models with "++" used data augmentation as described in this section.

In figure 2, we can observe the accuracy of Prototypical Networks and DenseNet models trained by varying sample sizes. We performed experiments using the same embedding architectures and configurations described in this section varying the training sample sizes with percentages of 44%, 67% and 85% and a fixed test size of 25%. From the obtained results, we can see that the performance of the DenseNet models increases as more training examples are provided. From figure 2(b) we can see that the DenseNet model trained using data augmentation obtains better results than the one trained without. On the other hand, Prototypical Networks models do not show a significant increase in performance as the percentage of samples increases. In figures 2(b) and 2(c) we can observe how the use of data augmentation, on RWTH and CIARP datasets respectively, results in Prototypical Networks models with great accuracy improvement compared to the results obtained on LSA16, figure 2(a)), where the increase of performance from the use of data augmentation is minimal.



(a) LSA16 (b) RWTH (c) CIARP

Fig. 2: Accuracy of Prototypical Networks and DenseNet models trained by varying sample sizes on three datasets: LSA16, RWTH and CIARP. Each plot represents a different dataset where the x-axis is the percentage of samples used and the y-axis is the accuracy obtained. Models with "++" used data augmentation as described in this section.

4 Conclusion

We have performed experiments to evaluate the mean accuracy of Prototypical Networks and Wide-DenseNet on three handshape recognition datasets with and without data augmentation techniques. For all datasets we found models that showed a performance on par with or better than the state of the art. All models achieve near-perfect accuracy on CIARP. This shows that the dataset is too simple as a benchmark for handshape recognition. While it has more samples on the other datasets (6000), the samples are too homogeneous and do not have enough variation to generalize results to real-world application. Prototypical Networks provide a new state-of-the-art accuracy on the LSA16 dataset, surpassing all other known methods. Wide-DenseNets also improve upon the state of the art, and come close to prototypical networks. However, we can observe that the performance gap between the two datasets decreases sharply when the sample size increases. We have also obtained a new state-of-the-art on the RWTH dataset with Wide-DenseNet, while Prototypical Networks also improved upon all previous results; this shows that newer convolutional architectures can work better with less data, but there's still room for improvements using specialized models.

In future work, we will focus on comparing with other datasets to better understand the relationship between models and dataset complexities for handshape recognition. We also see the need to compare with pre-trained models, which are another way to alleviate the lack of data in a certain domain, as well as methods that can take advantage of unlabeled data. Finally, we will investigate the possibility of merging data sets from different sign languages to augment the sample size, as well as identify the types of data augmentation that lead to an improvement in state-of-the-art models.

References

1. Alani, A.A., Cosma, G., Taherkhani, A., McGinnity, T.M.: Hand gesture recognition using an adapted convolutional neural network with data augmentation. 2018 4th International Conference on Information Management (ICIM) pp. 5–12 (2018)
2. Ameen, S., Vadera, S.: A convolutional neural network to classify american sign language fingerspelling from depth and colour images. Expert Systems 34(3), e12197 (February 2017), http://usir.salford.ac.uk/id/eprint/41255/
3. Barros, P., Magg, S., Weber, C., Wermter, S.: A multichannel convolutional neural network for hand posture recognition. pp. 403–410 (09 2014)
4. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 113–123 (2019)
5. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. The MIT Press (2016)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2015)
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 7132–7141 (2017)

8. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014), http://arxiv.org/abs/1412.6980, cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015
10. Koller, O., Ney, H., Bowden, R.: Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3793–3802. Las Vegas, NV, USA (Jun 2016)
11. Koller, O., Ney, H., Bowden, R.: Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3793–3802. Las Vegas, NV, USA (Jun 2016)
12. Núñez Fernández, D., Kwolek, B.: Hand posture recognition using convolutional neural network. In: Mendoza, M., Velastín, S. (eds.) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. pp. 441–449. Springer International Publishing, Cham (2018)
13. Pham, H., Guan, M., Zoph, B., Le, Q., Dean, J.: Efficient neural architecture search via parameters sharing. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 4095–4104. PMLR, Stockholmsmässan, Stockholm Sweden (10–15 Jul 2018), http://proceedings.mlr.press/v80/pham18a.html
14. Quiroga, F., Antonio, R., Ronchetti, F., Lanzarini, L.C., Rosete, A.: A study of convolutional architectures for handshape recognition applied to sign language. In: XXIII Congreso Argentino de Ciencias de la Computación (La Plata, 2017). (2017)
15. Ronchetti, F., Quiroga, F., Estrebou, C., Lanzarini, L., Rosete, A.: Sign languague recognition without frame-sequencing constraints: A proof of concept on the argentinian sign language. In: Ibero-American Conference on Artificial Intelligence. pp. 338–349. Springer (2016)
16. Ronchetti, F., Quiroga, F., Lanzarini, L., Estrebou, C.: Handshape recognition for argentinian sign language using probsom. Journal of Computer Science and Technology 16(1), 1–5 (2016)
17. Snell, J., Swersky, K., Zemel, R.S.: Prototypical networks for few-shot learning. CoRR abs/1703.05175 (2017), http://arxiv.org/abs/1703.05175
18. Tang, A., Lu, K., Wang, Y., Huang, J., Li, H.: A real-time hand posture recognition system using deep neural networks. ACM Transactions on Intelligent Systems and Technology (TIST) 6(2), 21 (2015)