CNN favours the bold – the untold story about seeing sign language in layers

Tristan Møller, Martin Jespersen & Marcus Gasberg

Aarhus University

Introduction

Ever felt the urge to understand sign language?

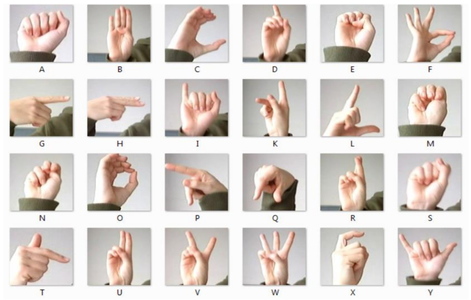
Stay a while and we'll tell you how a simple ML system can translate signs to words! Are CNN models best for image prediction or could logistic regression or SVM perform just as well or even better?

Data

Sign-language use visual manual modality to convey meaning. The system is fed pictures of signs made by a hand: 28x28px, 784 features.. The training data had a total of 27,455 instances, where the training set had a 7172 instances, so the data did not need to be split up

The dataset has no J or Z, because this requires movement, so the models should be able to classify 24 difference signs.

All images has labels attaches that shows the actual value of each image, so the type of machine learning is supervised.



Taking the models for a spin

One does not simply choose one model, when investigating a ML-problem. It is necessary to try several, optimize them and find the best! All models are measured with a f1_micro score since the balance between the precision and recall with the use of a f1 score was important, where the micro part of f1 is specific to multiclass classification and weights the 24 classes equally in the calculation.

Logistic Regression Classifier – the 'benchmarker'

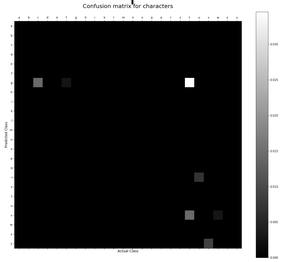
Scaling/normalization and GridSearch is used to find an optimal classifier. A solid score on 0.964. This model is unable to predict non-linear relations, which may cause the worse performance.

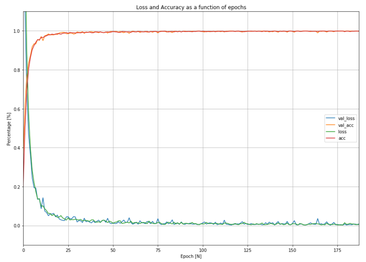
Support Vector Classifier – the smarter brother

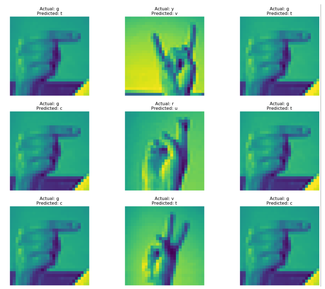
Like before we use GridSearch to find a 'mean' SVM classifier. SVC can predict non-linear relations compared to logistic regression. The best model found has a score of 0.986.

MMTNet (CNN Architecture) – The chosen one

A combination of data augmentation, convolutional/pooling layers & a 'mad' CNN architecture & voila, you have yourself a fully flexed ML system that can identify sign language with an impressive score of 0.996! Here is the performance of this bad boy:

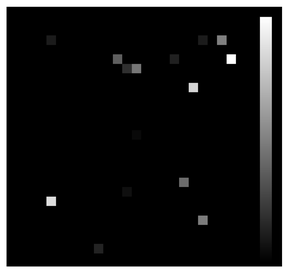


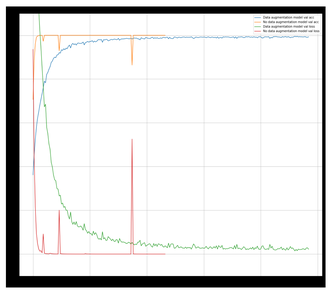


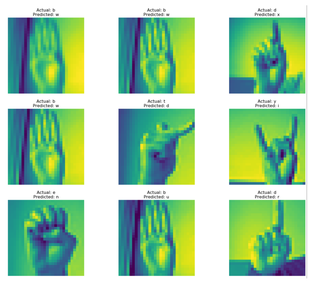


LeNet–5 (CNN Architecture)

This is the CNN architecture chosen for comparison of the MMTNet model. It was expected to perform well on images with the same size, but with a final f1_micro score on 0.993 it was impressive, but not as accurate as MMTNet.







Data Augmentation

It turns out CNN requires a lot of data to avoid overfitting. It was necessary to use real-time generation of new signs when training the model. This increased the performance by 12%!

Under- and overfitting

Besides data augmentation and the use of GridSearch to explore the hyperparameter space, the CNN architectures also made use of early stopping to avoid overtraining the models.

The logistic regression and SVC model both use scaling to reduce the sensitivity of input features as well as Lasso regularization to constrain the weights.

Conclusion

CNN was designed for image recognition. Realistic data augmentation made the model less prone to overfitting. Convolution and pooling layers made it possible to detect higher-level features. These building blocks enabled the model to differentiate between signs and predict the letters.

Our results show that LeNet-5 is possible to beat and it looks like a deeper neural network in combination with relu activation functions is able to give a slightly better result on the Sign Language dataset.

As expected the simpler models logistic regression and SVC did not perform as well, but still with impressive f1_micro scores.