Visual concept learning with deep belief and feed forward networks

Pavel Ianko
pavel.ianko@studenti.unipd.it
13.01.2022
Student ID: 2041301

1. Introduction

In a rapidly developing sphere of machine learning and artificial intelligence, a significant part of research corresponds to unsupervised learning models. Learning without answers allows models to perform feature extraction, making them more resembling of 'natural' learning process.

In this work, we take advantage of a deep belief network (DBN), used for capturing hidden data representations, based on letter images from EMNIST dataset. Using DBN model allowed for analysing hidden data representations through visualising neuron receptive fields at several depths. Moreover, linear read-outs are performed, to compare DBN classification performance with a supervised feed forward neural network (FFNN).

To assess a DBN's ability to grasp similar visual concepts, an analysis with hierarchical clustering methods and confusion matrices is performed. Finally, both DBN and FFNN robustness are tested, providing gaussian noise, salt & pepper noise, and adversarial attacks.

2. Data

Model training is based on letters, provided by EMNIST dataset. Each data instance represents a greyscale-value image of 28 × 28 pixels.

For this study, we limited the whole dataset with 10 classes and 4800 images for each class. Thus, both models are trained on a set of 48000 images, while the accuracy is reported based on 8000 unseen data instances. An example of different labels is presented at Fig.1

Before training models, it was made sure that the class distribution is uniform for both train and test sets. This proves accuracy to be a valid metrics for models comparison (Fig.2).

Preprocessing procedure includes max-scaling the data, by dividing the images greyscale value by the maximum of 255.

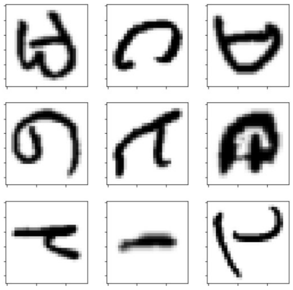

Figure 1: First nine labels of EMNIST letters dataset

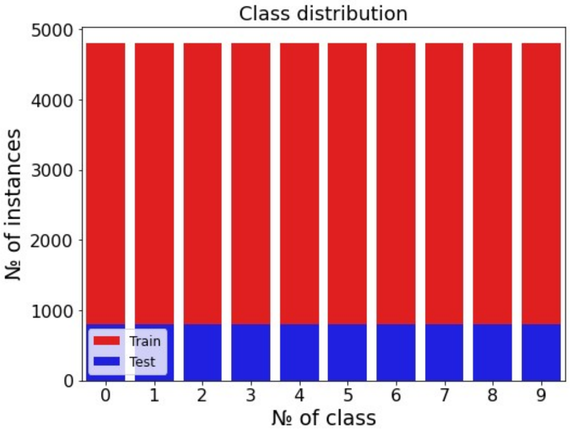

Figure 2: Uniform class distribution for train and test set

3. Model

3.1. Architecture

In this work, two types of architectures are presented - a shallow DBN model with 500 hidden units, versus a deep belief network with two hidden layers of 500 units each. One needs to choose an FFNN architecture accordingly with the chosen DBN structure. Hence, the presented FFNN model has one hidden layer of 500 neurons, and 784 input neurons, representing flattened pixels of an input image.

In addition, several perceptrons were trained for multi-classification task, with DBN hidden layers as input. In latter parts, FFNN performance will be compared with perceptrons, trained on DBN hidden representations, instead of raw images.

3.2. Training and testing method

For comparison purposes, FFNN model was trained during approximately the same time as the shallow DBN model. This is why number of training epochs varies between two types of architectures - 1500 and 120 epochs perceptron and FFNN model respectively.

Since the datasets are balanced (Fig.2), accuracy was chosen to compare models performance.

4. Results and discussion

4.1. Layer receptive fields

In Fig.3, neurons receptive fields of a deep belief network are presented. Each receptive field represents a layer connections matrix W, projected at two dimensional space of 28×28 pixels. Thus, a greyscale value highlights the neurons connection strength, capturing certain visual concepts.

Fig.3 clearly visualises how an unsupervised model captures more specific details with deeper hidden layers. While the first layer receptive fields are virtually 'greyish' (not specified), deeper layers focus on particular types of letters features, as it is demonstrated by bright dots and lines on virtually dark receptive fields.

Thus, deeper level receptive fields indicate a distribution of responsibility between neurons. Certain groups of hidden units are responsible for particular features detection.

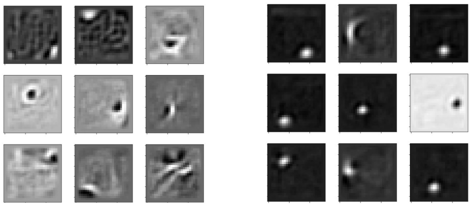
(a) First hidden layer (b) Second hidden layer

Figure 3: Hidden neurons receptive fields for DBN model

4.2. Clustering internal representations

In order to understand, if an unsupervised model was able to capture similar visual features, one can calculate mean hidden representations, corresponding to each class. Afterwards, an hierarchical clustering algorithm is to be applied to mean representations, to evaluate their similarity (Fig.4).

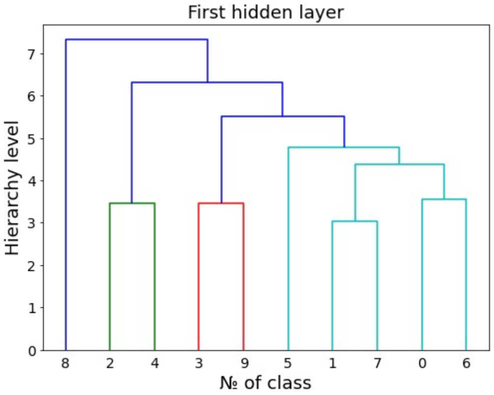

Figure 4: Class similarity, based on centroids of hidden representation

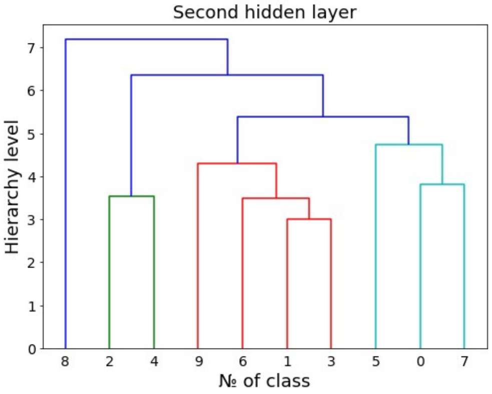

Figure 5: Second layer representation captured another similarity concepts, compared to first hidden layer

Below, Fig.6 demonstrates image instances, paired accordingly to the dendrogram (Fig.4), calculated over hidden layer representations. Indeed, from visual analysis one can conclude, that unsupervised model successfully encodes similarity of visual concepts.

4.3. Linear read-out performance

In this chapter, we report models performance on unseen test data. Internal representations of hidden DBN layers serve as input data for perceptrons, trained for multiclassification problem. However, while perceptrons learn from hidden encoded data, FFNN model is trained on raw images. Thus, hidden layer of FFNN network is both responsible for feature extraction, as well as for tuning for the multiclassification task.

Thus, Fig.7 reports accuracies on the test dataset.

Figure 7 clearly demonstrates benefits of learning from data with pre-extracted features, generated by unsupervised models. Linear read-outs outperform FFNN model by around 10%, with an accuracy above 80% on unseen data.

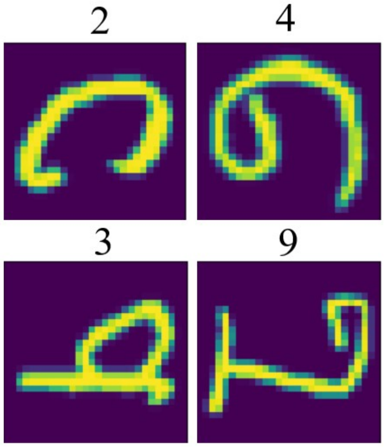

Figure 6: Similar class instances, according to the mean internal representations (Fig. 4)

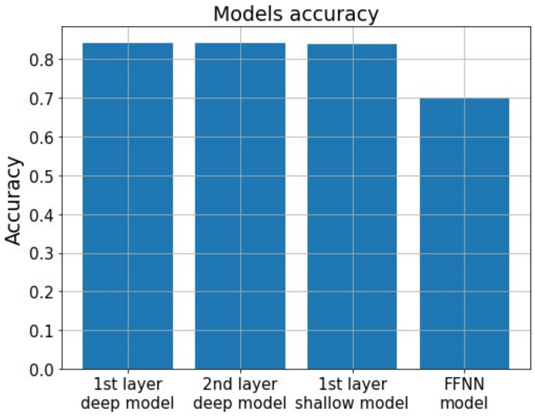

Figure 7: Accuracies of feed forward neural network and linear read-outs on the test data

Moreover, internal representations are favored over raw images, since they serve as 'shared' data for multiple tasks (e.g. letter and digit recognition) [1]. Thus, using shared representation is beneficial, because one can flexibly readjust model for another task, instead of retraining model from scratch.

4.4. Noise robustness

Here we compare models' performance with unseen data, distorted by random noise insertions. Two types of noise are introduced. Figure 8 illustrates distortions, imposed by Salt & Pepper and Gaussian random noise.

For noise coverage, ranged from 0% (clear image) to 50%, we perform accuracy computation on the unseen test data. Thus, figures 9 and 10 report models' test accuracies. In summary, perceptrons, learned on hidden representations outperform FFNN model for noise intensities up to 30%. In addition, models behaviour was consistent for both Gaussian and Salt and Pepper noise.

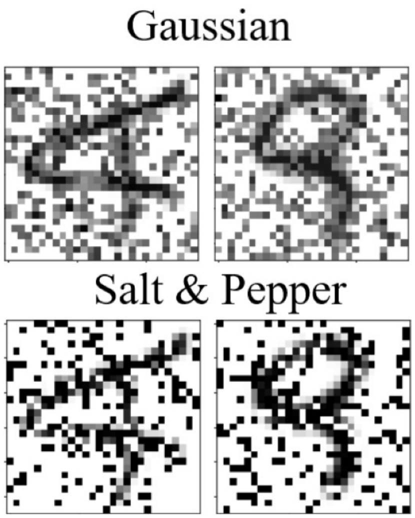

Figure 8: Two types of noise, applied correspondingly to two test instances (noise intensity is equal to 40%)

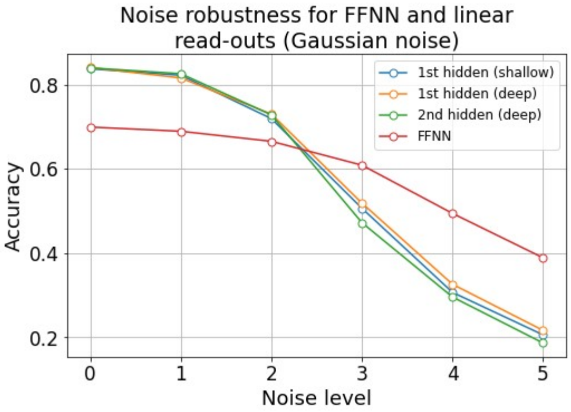

Figure 9: Accuracy of FFNN and perceptrons, corresponding to Gaussian noise intensities from 0% to 50%

4.5. Adversarial attacks

One beneficial trait of DBN model is the ability for data reconstruction, which is essential for adversarial attacks. For instance, Fig.11 demonstrates that, by adding specific noise, accounting for loss function gradient, it is feasible to negatively affect model's predictions.

However, DBN capacity for reconstruction allows the this model to stay robust against adversarial examples. For instance, after two steps, performed by shallow model, resulting image is almost free of adversarial distortions (Fig. 12).

To evaluate DBN robustness against adversarial

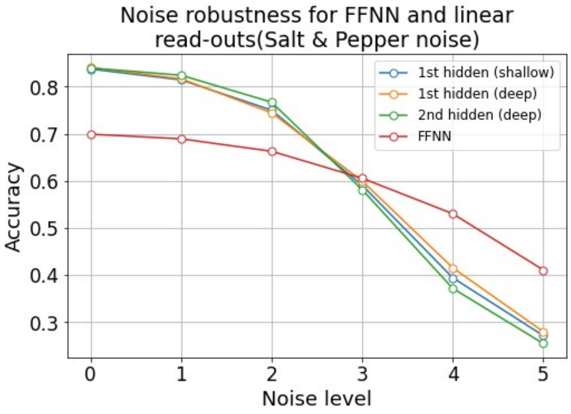

Figure 10: Accuracy of FFNN and perceptrons, corresponding to Salt and Pepper noise intensities from 0% to 50%

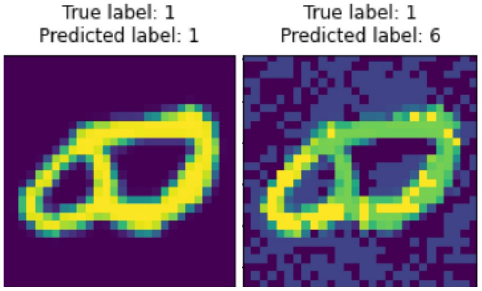

Figure 11: Example of image misclassification, performed by FFNN model on the image, exposed to adversarial attack

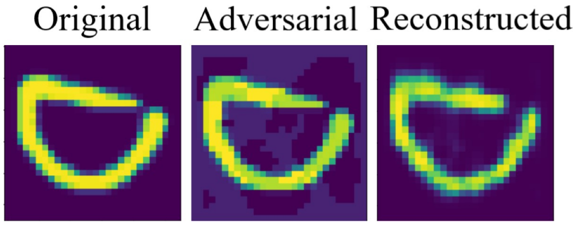

Figure 12: Example of DBN image two-step reconstruction. Image after reconstruction is more resembling of the original

samples, its accuracy was assessed on the test set, exposed to attack intensities ϵ ranged from 0 to 2.25, where ϵ stands for gradient multiplication coefficient. Thus, a plot on Fig. 13 emphasizes importance of reconstruction. Compared to FFNN model with the steepest accuracy decline, two reconstruction steps allow perceptrons to maintain accuracy of around 70% for weak attacks with ϵ less than 0.05.

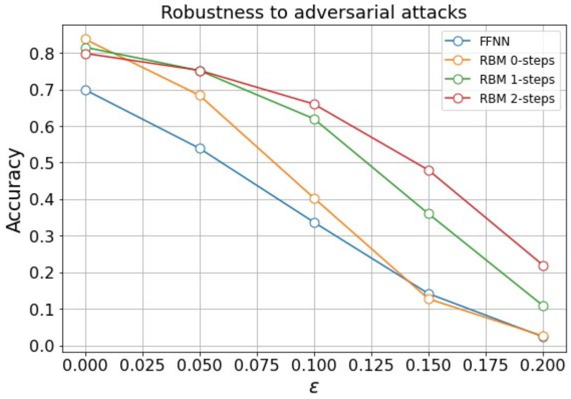

Figure 13: Accuracies of FFNN and perceptrons, for several reconstruction steps and attack intensities

4.6. Confusion matrix analysis

Confusion matrices allow to understand nature of misclassifications, made by studied models, and identify similar alphabet symbols (Fig. 14).

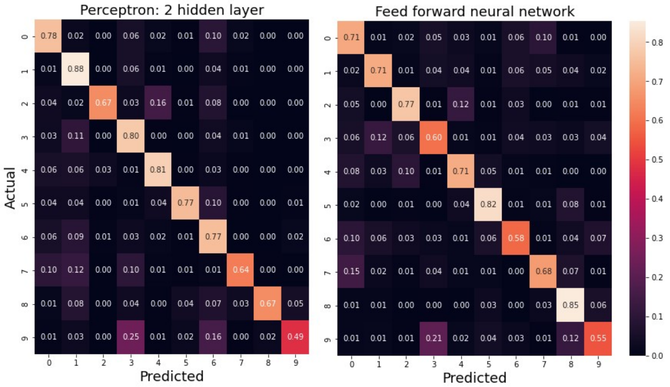

Figure 14: Confusion matrices for FFNN and the perceptron, learned from second layer hidden representation

Along the main axes of confusion matrices, the underperformance for certain classes is clear. For instance, 18% of FFNN misclassifications account for classes 9 and 3, which are actually visually similar (Fig. 15).

All confusion matrices exhibit the same pattern. For instance, also the model, learned from internal representations of the first hidden layer, in 17% of cases labels instances of class 9 as letters, corresponding to class 3 (Fig. 16). One of possible explanations for underperforming of models is a greater variety of written letters, as opposed to digits in MNIST dataset. EMNIST letters encapsulate both upper-case and lower-case letters as instances, corresponding to the same class, which ensures a greater statistical variety.

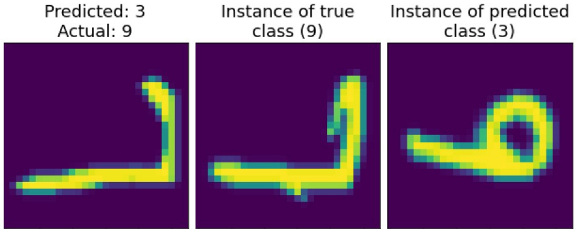

Figure 15: Instances of mistaken class pair (9 and 3) for FFNN model

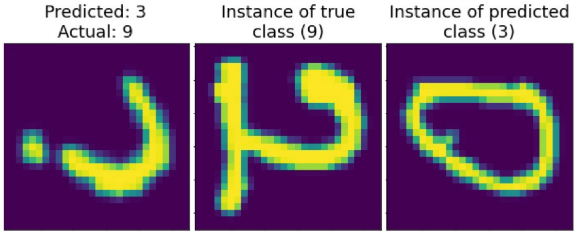

Figure 16: Instances of mistaken class pair (9 and 3) for perceptron, learned from first hidden layer representation

5. Conclusion

In this study, we compared visual concept learning, implemented by deep belief network and feed forward neural network. Neuron receptive fields visualizations emphasized how DBN units infer visual features. Mean hidden representations clustering proves DBN ability for capturing similarities between different classes. Finally, perceptrons, learned on internal representations, demonstrate higher accuracy (80%), as opposed to feed forward neural network with 70% accuracy on the test set. Moreover, DBN model exhibits highter robustness to two types of noise and adversarial attacks. Confusion matrix analysis revealed that mostly, all studied models misclassify visually similar instances.

References

[1] Marco Zorzi, Alberto Testolin, and Ivilin Peev Stoianov. Modeling language and cognition with deep unsupervised learning: a tutorial overview. *Frontiers in psychology*, 4:515, 2013.