Visual concept learning with deep belief and feed forward networks

Mehran Farajinegarestan
mehran.farajinegarestan@studenti.unipd.it
student ID: 2071980

June 27, 2023

1 Introduction

This study explores the efficacy of learning models that do not rely on explicit labels to train, thus enabling them to perform feature extraction and closely mimic the natural learning process seen in the human brain. Specifically, we employ a deep belief network (DBN) to capture hidden data representations from the EMNIST dataset, which comprises letter images. By utilizing the DBN model, we are able to analyze these hidden data representations by visualizing the receptive fields of neurons at various depths. Furthermore, we employ linear read-outs to compare the classification performance of the DBN with a supervised feed-forward neural network (FFNN) which has a similar number of parameters to the DBN model.

In order to evaluate the DBN's capacity to comprehend visually similar concepts, we employ hierarchical clustering methods and confusion matrices for analysis. Additionally, we subject both the DBN and FFNN models to robustness testing, which includes the introduction of Gaussian noise, salt and pepper noise, as well as adversarial attacks.

2 Data

In this project, we utilized the letter subset of the EMNIST dataset, an extended version of the well-known MNIST dataset. Similar to MNIST, each instance in the dataset represents a grayscale image with dimensions of 28×28 pixels.

To maintain simplicity and ensure comparability with the MNIST dataset, we focused on the first 10 classes of the EMNIST dataset. Consequently, the models were trained on a total of 48,000 images (with 4,800 instances per class), while the accuracy evaluation was conducted on a separate set of 8,000 unseen instances. Figure 1 provides a visual representation of the different labels included in the dataset.

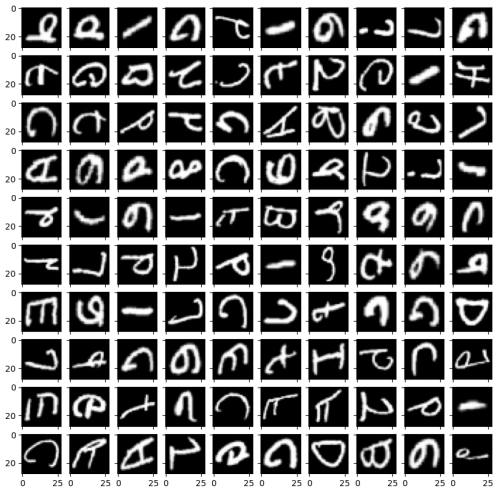


Figure 1: Sample data instances of EMNIST letter dataset

Prior to training the models, we carefully ensured that the class distribution was uniform across both the training and test sets. This approach guarantees

that accuracy serves as a reliable metric for evaluating the models' performance (Figure 2).

The pre-processing pipeline involved performing max-scaling on the data, whereby the grayscale values of the images were divided by the maximum value of 255. Additionally, we selected and retained only the first 10 labels from the dataset.

Figure 2: distribution of train and test sets

3 Model Architecture

In this study, we employed the same Deep Belief Network (DBN) architecture that was utilized during the laboratory sessions. The DBN consisted of three hidden layers, with dimensions of 400, 500, and 800 neurons, respectively. This decision was motivated by the similarity between the dataset used in the laboratory sessions and the dataset employed in this study. Furthermore, the author's keen interest lay in investigating the possibility of reproducing the results observed in the lab sessions using a dataset that closely resembled it. Similarly, the Feed-Forward Neural Network (FFNN) model utilized in this study maintained the same architecture as the one observed during the lab sessions. It consisted of hidden layers with 400, 500, and 800 neurons, along with 784 input neurons representing the flattened pixels of an input image. Moreover, three perceptrons were trained specifically for the multi-classification task, with the three hidden layers of the DBN serving as their input. In subsequent sections, we will compare the performance of the FFNN model with that of the perceptrons trained on the DBN's hidden representations, rather than directly using the raw images.

4 Simulations and Discussion

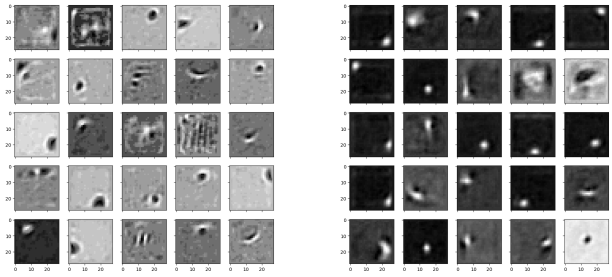
4.1 Layer receptive fields

Figure 3 displays the receptive fields of neurons in a deep belief network, offering insights into the network's connectivity matrix, represented in a two-dimensional space of 28 $\times$ 28 pixels. The grayscale values within each receptive field depict the strength of connections between neurons, thereby capturing distinct visual concepts.

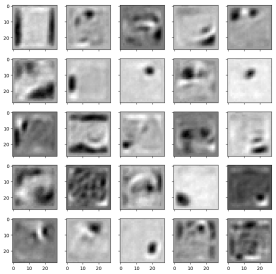
The visualization in Figure 3 effectively illustrates how an unsupervised model progressively captures finer details as it delves deeper into its hidden layers. In contrast to the receptive fields of the first layer, subsequent layers exhibit bright dots and lines against predominantly dark backgrounds, demonstrating a focus on specific features associated with letter recognition. Consequently, an examination of the receptive fields at deeper levels reveals that individual neurons within the network are assigned more specialized functions. In this regard, specific hidden units are allocated the responsibility of detecting distinct visual features.

4.2 Clustering internal representations

To assess to which extent an unsupervised model captures similar visual features, we a mean representation is calculated for each class, and these representations are subjected to a hierarchical clustering algorithm for evaluating their similarity, as illustrated in Figure 4. The resulting dendrogram guides the pairing of image instances depicted in Figure 5. Observing the image pairs in Figure 5 alongside the corresponding dendrogram, it is evident that label pairs (2,4), (1,7), (3,7), (3,1), and (6,9) exhibit a greater degree of similarity as they reside within the same cluster based on the representations from all three

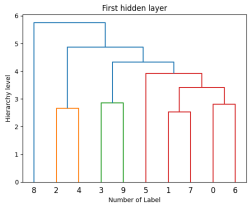


(a) First hidden layer (b) Second hidden layer

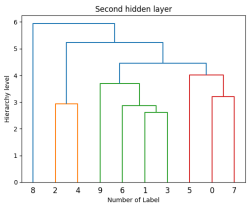


(c) Third hidden layer

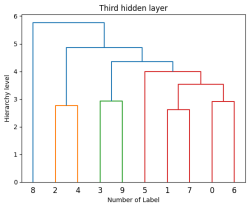
Figure 3: Hidden neurons receptive fields for DBN model



(a) Class similarity, based on clustering of hidden representations of the first hidden layer



(b) Class similarity, based on clustering of hidden representations of the second hidden layer



(c) Class similarity, based on clustering of hidden representations of the third hidden layer

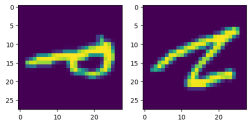
Figure 4: Comparison of three hidden layers of the DBN model using centroid (mean) of the representations of each class

hidden layers. Through this visual analysis, it can be concluded that the unsupervised model effectively encodes the similarity of visual concepts.

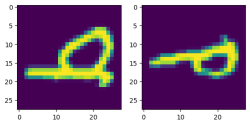
4.3 Linear read-out performance

In this section, we report models performance on a multi-class classification on test data. Internal representations of hidden DBN layers serve as input data for perceptrons, trained for classification problem. However, while perceptrons learn from hidden encoded data, FFNN model is trained end-to-end on train images. End-to-end means that hidden layer of FFNN learns to do feature extraction and multi-class classification in one step. We choose the number of epochs to match the training time of the DBN and the readout layer combined.

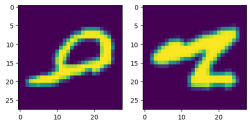
Figure 6 clearly demonstrates benefits of learning from data with pre-extracted features, generated by unsupervised models. Linear read-outs outperform FFNN model by around 10%, with an accuracy above



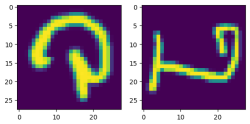
(a) comparison of two random instances of classes 1 and 7



(b) comparison of two random instances of classes 1 and 3



(c) comparison of two random instances of classes 3 and 7



(d) comparison of two random instances of classes 6 and 9

Figure 5: Similar class instances, according to the mean internal representations (Figure 4)

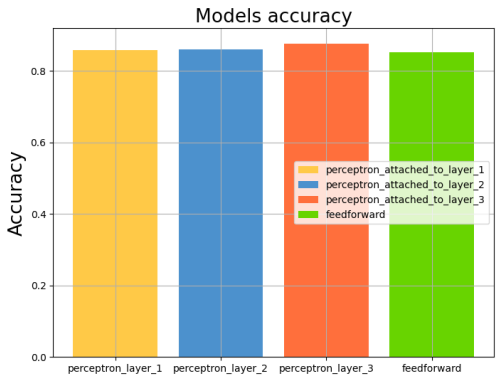


Figure 6: Accuracy of feed forward neural network and linear read-outs on the test data

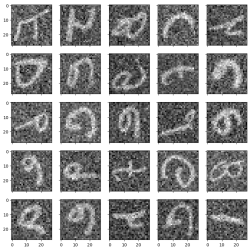
80% on unseen data.

Moreover, internal representations is beneficial, because one can use them for another task, instead of retraining model from scratch.

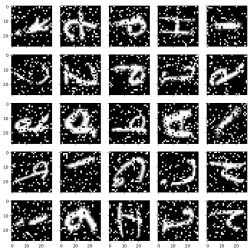
4.4 Noise robustness

This section focuses on evaluating the robustness of the models under the presence of random noise applied to the raw images. Two types of noise, namely Gaussian random noise and Salt & Pepper noise, are utilized for this purpose. Figure 7 visually demonstrates an example image corrupted by both types of noise.

To compare the performance of the models, various levels of noise ranging from 0 (no noise) to 1.3 are injected into the raw images of the test dataset. The accuracy of different models is then calculated and reported in Figures 8 and 9. It is observed that perceptrons, trained on hidden representations, outperform the Feedforward Neural Network (FFNN) model when subjected to Gaussian noise more than 10%. Even the perceptron trained on internal representaions of the first hidden layer out perform the FFNN model, which show how much representaions learned by DBN models are helpful to learn a generalization of simple shapes seen in nature. Surprisingly, the perceptrons using first and hidden layer of DBN as their input out performed the perceptron using third hidden layer of DBN and the FFNN model when they were subjected to Salt & Pepper noise.



(a) Gaussian random noise injected to test image



(b) Salt & Pepper noise injected to test image

Figure 7: Hidden neurons receptive fields for DBN model

4.5 Adversarial attacks

One advantageous characteristic of the Deep Belief Network (DBN) model is its capability for data reconstruction, which proves to be crucial in scenarios involving adversarial attacks. For instance, Figures 10 and 11 exemplify how the model's predictions can be negatively influenced by introducing specific noise that takes into account the gradient of the loss function.

Nevertheless, the DBN's ability to reconstruct data enables it to maintain robustness against adversarial attacks. By subjecting the resulting image to two steps performed by the DBN model, the adversarial distortions are effectively mitigated, resulting in an image that is nearly free from such distortions.

To assess the DBN model's robustness against ad-

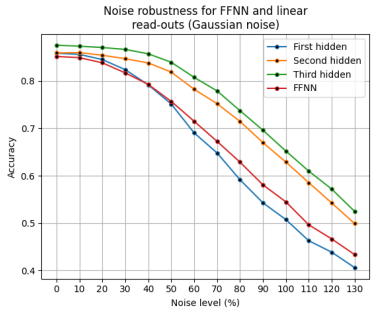


Figure 8: Accuracy of FFNN and perceptrons, corresponding to Gaussian noise intensities from 0% to 130%

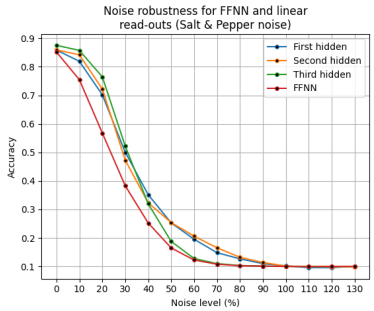


Figure 9: Accuracy of FFNN and perceptrons, corresponding to Salt & Pepper noise intensities from 0% to 130%

versarial attacks, its accuracy on the test set was evaluated under varying attack intensities denoted as epsilon, ranging from 0 to 0.25. Here, epsilon represents the coefficient used to multiply the gradient. The plot in Figure 12 highlights the significance of reconstruction. In comparison to the Feedforward Neural Network (FFNN) model, which experiences the sharpest decline in accuracy, the incorporation of two reconstruction steps enables the perceptron-based DBN model to maintain a higher level of accuracy at each epsilon level.

4.6 Confusion matrix analysis

Confusion matrices serve as valuable tools for analyzing the misclassifications made by the studied models and identifying similarities between alphabet symbols, as depicted in Figure 13.

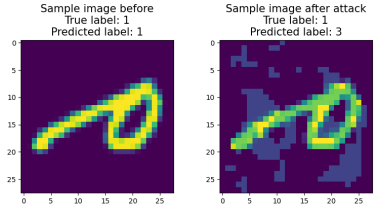


Figure 10: Example of image misclassification, performed by DBN model on the image, exposed to adversarial attack

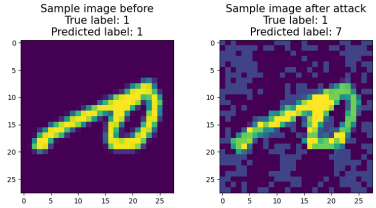


Figure 11: Example of image misclassification, performed by FFNN model on the image, exposed to adversarial attack

By examining the primary axes of the confusion matrices, it becomes apparent that certain classes exhibit underperformance. For instance, 18% of the misclassifications made by the FFNN model are attributed to classes 9 and 3, which are visually similar, as it was shown in 4.2

The observed pattern of the confusion matrices is consistent across all models. For instance, even the model trained on internal representations from the first hidden layer shows a tendency to label instances of class 9 as letters corresponding to class 3 in 17% of cases. One plausible explanation for the underperformance of the models lies in the greater diversity of written letters compared to digits in the MNIST dataset. The EMNIST dataset encompasses both upper-case and lower-case letters as instances belonging to the same class, thereby introducing a higher degree of statistical variety.

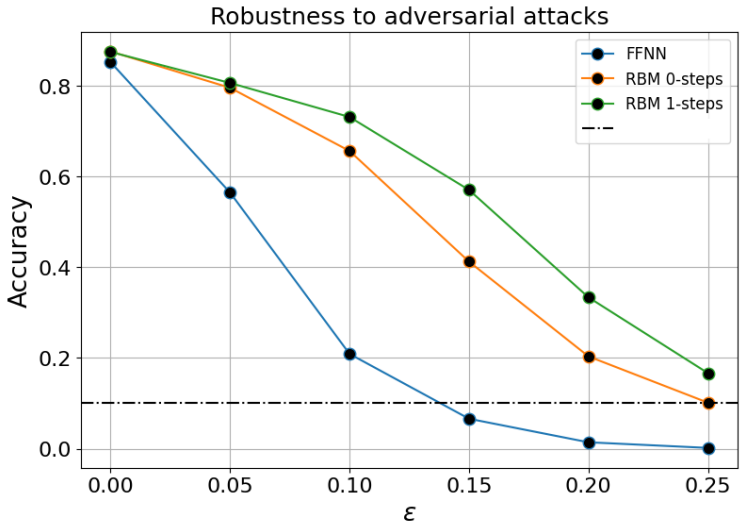
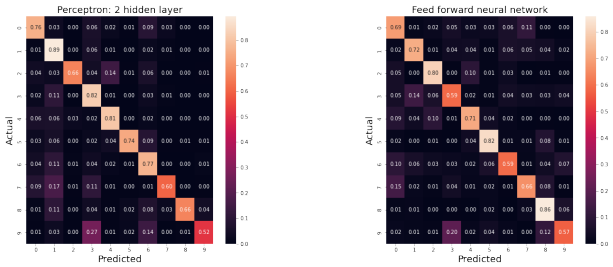


Figure 12: Accuracy of FFNN and perceptrons, for several reconstruction steps and attack intensities



(a) (b)

Figure 13: Confusion matrices for FFNN and the perceptron, learned from second layer hidden representation

timately, perceptrons trained on internal representations exhibited a superior accuracy rate of 80% on the test set, surpassing the FFNN model with an accuracy of 70%. Furthermore, the DBN model showcased higher robustness in the face of two types of noise and adversarial attacks. The analysis of confusion matrices revealed that all the studied models predominantly misclassified visually similar instances.

5 Conclusion

This study undertook a comparison between deep belief networks (DBN) and feedforward neural networks (FFNN) in terms of their capacity for visual concept learning. The visualization of neuron receptive fields in DBNs provided insights into how these units infer visual features. Additionally, the clustering of mean hidden representations demonstrated the DBN's ability to capture similarities across different classes. Ul-