Project 2: A Comparative Study Between Traditional and Information Theoretic Neural Network Models for Fashion-MNIST Classification

Connor McCurley
Deep Learning, Fall 2019
University of Florida
Gainesville, FL, USA 32611
Email: cmccurley@ufl.edu

***Abstract*—**

***Index Terms*—Fashion MNIST, Autoencoder, Information Theoretic Learning, Support Vector Machine**

I. Introduction

II. Methodology

This section describes the methodology implemented in this work. Analysis of the data is performed, dimensionality reduction and classification procedures are described, various network architectures under analysis are elaborated on and the experimental procedure is outlined.

A. Data Analysis

The following data analysis was originally described in [1], but is pertinent to this work and is thus re-analyzed here. The data was plotted as shown in Figure **??** to gain an understanding of the format. Each sample in the Fashion-MNIST dataset is a 28x28, gray-scale image of a clothing item belonging to one of ten classes [2]. This translates to 784 length feature vectors with values ranging between 0-255. There were exactly 60000 training images included in the training dataset and 10000 which were held-out for test. The 60000 samples were later sub-divided in the experimentation for cross-validation. Histograms of one-versus-all Euclidean distances to each of the classes are shown in Figure **??** to gain a sense of class separability using the raw images, solely. Given that the classifier would be a multi-layer perceptron artificial neural network, it was determined that dimensionality reduction should be utilized to combat the Curse of Dimensionality, while potentially improving class discriminability.

B. Network Architecture

C. Experiments

Experiments

1) *Baseline:* Model trained on the original data (described in section II).
2) *PCA 100:* Principal vectors estimated from the training set were used to project the data down to 100 dimensions, thus retaining 91.1% of the original variance. The model was the same as the baseline excluding the input layer.
3) *PCA 400:* Principal vectors estimated from the training set were used to project the data down to 400 dimensions, thus retaining 98.5% of the original variance. The model was the same as the baseline excluding the input layer.
4) *UMAP 100:* UMAP was used to project the data down to 100 dimensions. 15 neighbors were used in the local neighborhood and the minimum distance of points in the latent space was constrained to 0.1, which was measured by Euclidean distance. The model was the same as the baseline excluding the input layer.
5) *UMAP 400:* UMAP was used to project the data down to 400 dimensions. 15 neighbors were used in the local neighborhood and the minimum distance of points in the latent space was constrained to 0.1, which was measured by Euclidean distance. The model was the same as the baseline excluding the input layer.
6) *AE 100:* An additional multi-layer perceptron was first trained in the form of an autoencoder. This MLP demonstrated layers as (784-500-400-150-100) in the encoder with the reverse in the decoder. A hyperbolic tangent activation was applied at the output layer. The network was trained with mean-square error loss and updated with Adamax. The original dataset was passed through the encoder to reduce dimensionality. The model was the same as the baseline excluding the input layer.
7) *AE 400:* An additional multi-layer perceptron was first trained in the form of an autoencoder. This MLP demonstrated layers as (784-500-400) in the encoder with the reverse in the decoder. A hyperbolic tangent activation was applied at the output layer. The network was trained with mean-square error loss and updated with Adamax. The original dataset was passed through the encoder to reduce dimensionality. The model was the same as the baseline excluding the input layer.

III. Results

In this section, classification results for each experimental method tested are presented in the form of confusion matrices. A confusion matrix demonstrates the discrepancies between predicted and true class values for groups of samples. Essentially, it is a way to measure how accurate a classifier is, while providing insight into how the network confuses samples. A diagonal matrix signifies zero mis-classifications among all categories.

A. Confusion Matrices

Figures

TABLE I: Classification Accuracies for Different Neural Model/ Classification Systems

Classification Model	**Accuracy**
Baseline CNN	0.90
SVM 100D	0.86
SVM 75D	0.85
SVM 50D	0.84
SVM 25D	0.81
SVM 10D	0.76

IV. Discussion

In this sections, observations are made on results and insight is given to potential influences.

A. Results

B. Potential Improvements

In

V. Conclusions

Three

Future research endeavors toward this topic include,

Honor Statement

* I confirm that this assignment is my own work, it is not copied from any other person's work (published or unpublished), and has not been previously submitted for assessment either at University of Florida or elsewhere.

References

[1] C. H. McCurley, "Project 1: Manifold learning for fashion-mnist classification with multi-layer perceptrons," 2019.

[2] H. Xiao, K. Rasul, and R. Vollgraf. (2017) Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms.