COMP 551 Assignment 2

Abstract

In this report, we implemented multi-layer perceptron (MLP) models and a convolutional neural network to compare testing accuracy on image classification using the Fashion-MNIST dataset. To optimize the MLP and CNN, we performed Bayesian hyperparameter optimization. For the MLP performance, we looked at the effect of the number of hidden units, hidden layers, activation functions, regularization, learning rate, and normalized data. Similarly, we optimized the kernel size, dilation factor, and the number of filters for the CNN model and compared its performance to the MLP. We found that the choice of learning rate and regularization coefficient had the most profound impact on testing accuracy, representing $> 85\%$ of the variance in results. Overall, the CNN model achieved higher accuracy than the MLP with 90% vs 87%, respectively.

1 Introduction

The multi-layer perceptron (MLP) algorithm overcomes some of the limitations of the perceptron algorithm first described in 1958 by Rosenblatt [Rosenblatt, 1958]. In particular, the perceptron algorithm is a linear classifier that does not converge if the data is not linearly separable. A famous example of this is the XOR problem. To solve this problem, an MLP stacks layers of perceptrons together, each passing through a non-linear differentiable activation function, creating an "composite function" or directed acyclic graph which can map inputs to outputs effectively for non-linearly-separable data [Murphy, 2022]. To train an MLP, the gradient of the output with respect to the parameters for each layer can be calculated through backpropagation and trained through gradient descent.

Despite the flexibility of the MLP, problems such as image classification should have a model invariant to irrelevant details such as translations or illumination differences in an image [LeCun et al., 2015]. The Convolutional Neural Network (CNN) uses ideas including shared weights, pooling, and stacking of many layers to go "deep" to solve these problems. CNN specializes in analyzing images and videos since it can learn filter weights which correspond to features like edges, textures, and patterns. It is another image classification method that applies convolution operations to input and hidden layers. In general, CNN performs better than MLPs on complicated images due to its ability to understand the spatial relations of pixels better and recognize features such as edges despite translations.

In this report, we implement three MLP models: no hidden layer, one hidden layer, and two hidden layers. We compare the accuracy of these three models using ReLU, tanh, and LeakyReLU activations. Lastly, we perform Bayesian hyperparameter optimization for both CNN and MLP architectures and compare their accuracies.

2 Datasets

The Fashion-MNIST dataset is analogous to the MNIST dataset; however, it contains images of fashion items from 10 categories rather than digits [Xiao et al., 2017]. The dataset contains 70,000 images split into training and test sets with 60,000 and 10,000 images, respectively. Each image is 28x28 pixels and grayscale with values between 0 and 255 representing intensity. To normalize the pixel values between 0 and 1, division by the max value of 255 was used, which improved the results.

3 Results

3.1 MLP

The 2-layer MLP with ReLU activations and L2 regularization on the testing dataset achieved the highest accuracy of 78.2%. Table 1 shows the accuracies of all models specified in the assignment requirements. The MLP with no hidden layers achieved the lowest accuracy at 67.3%, and a positive correlation between the number of hidden layers and accuracy was observed.

Table 1: Test Accuracy Results for Required Models

Hidden Layers	Activation(s)	Test Accuracy	Regularization
0	Softmax	67.3	No
1	ReLU	71.2	No
2	ReLU	73.2	No
2	Tanh	77.9	No
2	LeakyReLU	73.1	No
2	ReLU	78.2	Yes

The choice of a Tanh activation function between hidden layers resulted in an $\approx 4.7\%$ increase in accuracy compared to ReLU and LeakyReLU. All results were obtained after training for 100 epochs with a learning rate chosen for each model using the learning-rate range test as described by [Smith, 2017]. Figure 1 shows an example of the learning rate range test for a two-layer MLP with ReLU activations. Based on the graph, a learning rate is chosen approximately one order of magnitude before the loss reaches a minimum.

An example of the loss profile during training of an MLP with L2 regularization (Figure 2) decreased over time. Since no divergence was observed, this indicated that the choice of learning rate from the learning rate range test was suitable.

Un-normalized images were used during training and testing using the MLP with two layers and ReLU activations. The testing accuracy achieved was 10%. Thus the model performed no better than a random classifier.

To optimize the MLP model, hyperparameter optimization was performed with the Optuna library. Table 2

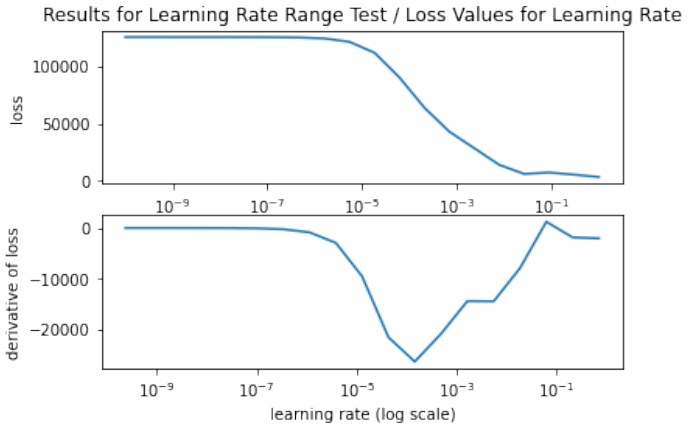

Figure 1: LR Range Test for 2 layer MLP with ReLU Activations

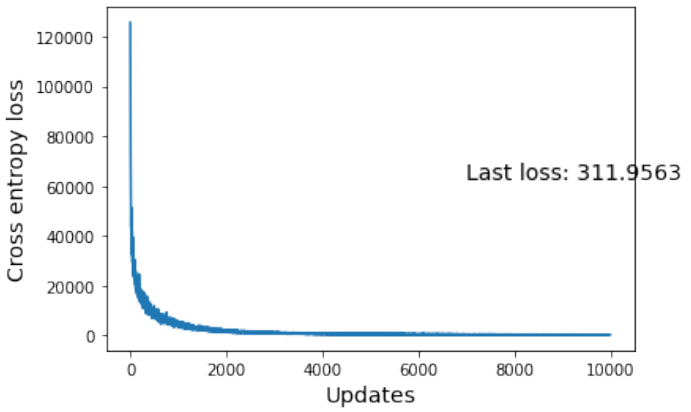

Figure 2: Training cross entropy loss vs number of updates for the 2 layer MLP with regularization. The number of updates is the number of optimizer steps = # batches * # epochs.

shows the best choice of hyperparameters selected using a Bayesian sampler, whereas Figure 3 indicates the relative importance of those parameters in determining the testing accuracy. The regularization coefficient was the most important feature, accounting for 85% of the variance among results. After training the optimized MLP model, a testing accuracy of 87% was achieved.

Table 2: MLP Optimized Hyperparameters

Hidden Layers	Layer Widths	LR	Regularization Coeff.
2	256	0.51	8.6e-4

Figure 4 shows a plot of the training loss and training and testing accuracy as a function of the number of epochs. The loss plot shows significant error increases due to the mini-batch gradient going only "on average" in the right direction rather than in the right direction for each step. No overfitting was detected as the testing accuracy mirrored the training accuracy over 100 epochs.

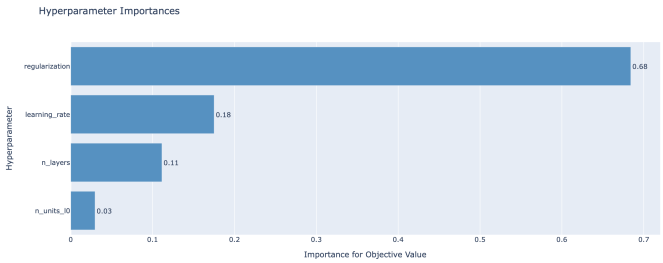

Figure 3: MLP hyperparameter relative importance for testing accuracy

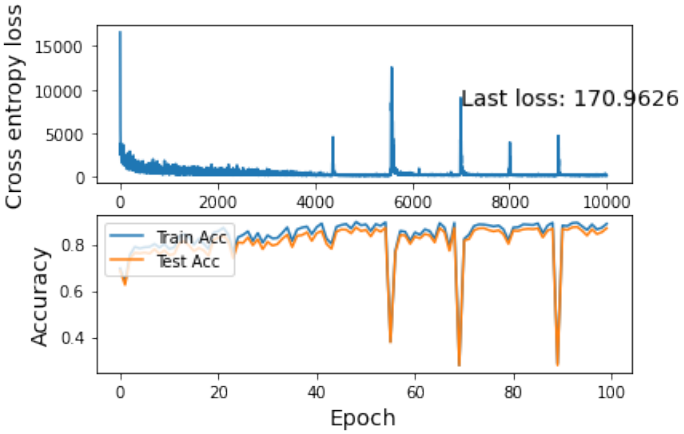

Figure 4: Loss and accuracy plot for the optimized MLP. Training loss decreased with occasional increases due to the use of mini-batch gradient descent. Training and testing accuracy was monitored for overfitting.

3.2 CNN

Similarly, Optuna was used for hyperparameter optimization of the Keras CNN architecture, where parameters such as the learning rate, dropout, dilation factor, kernel size, and the number of filters were varied. 93% of the variance in results was due to the choice of learning rate (Figure 5), where the optimal value was found to be 4.5e-3. Table 3 shows all the selected hyperparameters for the CNN model. Using the hyperparameter optimized model, 90% testing accuracy was achieved.

Table 3: CNN Optimized Hyperparameters

Filter size	Kernel size	Dilation	Learning rate
32	3	2	4.5e-3

4 Discussion

From the experiments conducted during this project we found that utilizing more hidden layers lead to a higher classification accuracy on the Fashion-MNIST data-set (Table 1). This is generally expected for complex image vision

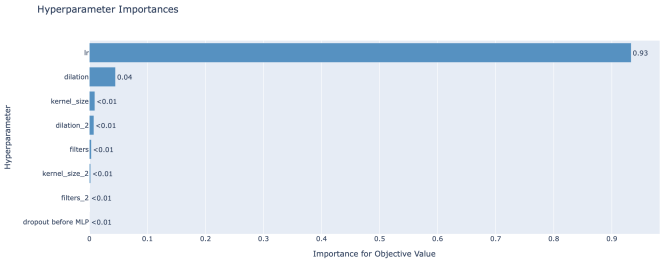

Figure 5: CNN hyperparameter relative importance for testing accuracy

tasks, in fact the very introduction of deep neural networks was the reason for multiple major breakthroughs in the field of image recognition [He et al., 2016]. The importance of a high number of layers in neural network models is indeed of such importance that residual connections were introduced in order to train even deeper models than those that had lead to the initial breakthroughs in the field of image recognition [He et al., 2016]. Depths of such magnitude are outside the scope of this current work, though even within the relatively shallow MLP experiments conducted here, the importance of multiple layers were appreciated.

MLPs with different activation functions were tested during this study, principly ReLU, LeakyReLU and Tanh were tested as activation functions. The advantage of using non-sigmoid and non-tanh activation functions was researched in an effort to develop deeper MLPs which had been found to be useful, while their activation made them difficult to train [Glorot and Bengio, 2010]. The ReLU function was found as a more robust alternative [Glorot and Bengio, 2010], while many modifications for various application have been made to the ReLU activation, leading to the invention of the LeakyReLU. However, in the comparatively shallow models that were used to compare the different cost functions in Table 1 we find that Tanh performs significantly better than both ReLU versions. This is likely due to the advantages of ReLU-type functions only appearing when relatively deep models are trained, as per their design.

Regularization was found to improve the accuracy of MLPs to allow them to reach within 3% of the CNN accuracy, which is usually used for image recognition tasks [He et al., 2016]. The addition of regularization has been found to cause simpler models to outperform some much more complex state of the art models in other fields, such as natural language processing [Howard and Gugger, 2020], therefore it is reasonable that MLP with regularization would outperform MLPs without and may even begin competing with CNNs as it has in this work.

When images without normalization were used for training MLPs in this study they performed with the accuracy of a random classifier, which is as good as having no model at all.

The CNN was found to be the highest performing model on the data with 90% accuracy, only the regularized MLP could come somewhat close to it with 87%. CNN based architectures are the basis for most modern image classifiers and are hence expected to perform better than a simple MLP, which our results are in agreeance with [He et al., 2016, Howard and Gugger, 2020]. Hyperparameter tuning using a Bayesian samples was conducted for both model types. The most important parameter during hyperparameter optimization was found to be the learning rate of the CNN classifier, for the MLP model it was the regularization coefficient that required the most attention to be chosen. As explained earlier the regularization is the primary reason for the ability of the MLP based architecture to compete with the CNN model, wherefore it is reasonable that the best use of regularization is of utmost importance to the performance of the MLP.

4.1 Conclusion

In conclusion it was found that despite their advantages for very deep MLPs the Tanh activation outperforms ReLU-type activation functinos for shallower models (2 hidden layers). Furthermore, though CNNs are the primary model architecture used for image recognition we found that for the Fashion-MNIST data-set an MLP with regularization was able to nearly compete with its accuracy.

5 Statement of contribution

Code: Joel Harms, Justin Charney Report: Luke Qian, Justin Charney, Joel Harms

References

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL https://proceedings.mlr.press/v9/glorot10a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

J. Howard and S. Gugger. *Deep Learning for Coders with Fastai and Pytorch: AI Applications Without a PhD*. O'Reilly Media, Incorporated, 2020. ISBN 9781492045526. URL https://books.google.no/books?id=xd6LxgEACAAJ.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Kevin P Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.