Assignment 3 Comp 551

Angad Verma, Angela Moskal, Kai Turanski

Abstract

We investigate Multilayer Perceptrons (MLPs) and Convolutional Neural Networks (CNNs) for image classification on the OrganAMNIST dataset. The impacts of network architecture, regularization, data preprocessing, and image resolution on classification performance were explored. The optimal MLP architecture achieved 75.9% test accuracy. The CNNs outperformed MLPs, achieving 88.4% test accuracy. A ResNet model pretrained on ImageNet yielded 86.0% test accuracy. Surpassing all other models is a pretrained Medical Open Network for AI (MONAI) DenseNet, achieving a 96% test accuracy.

1 Introduction

This project explores the performance of MLPs and CNNs on classification of the OrganAMNIST image dataset [1, 2]. An optimal MLP architecture was determined to include 2 hidden layers, each with 256 units and ReLU activations, achieving a maximum test accuracy of 75.3%. To avoid the vanishing gradient problem, we implemented alternative weight initialization schemes such as kaiming. Adding L1 or L2 regularization improved performance to 75.9% by mitigating overfitting. Normalizing input data significantly enhanced accuracy, while increasing image resolution (128×128) decreased model efficiency due to computational overhead. The optimal MLP was outperformed by a CNN with appropriately tuned filters, stride and padding, achieving a test accuracy of 88.4%. A pre-trained ResNet CNN with appropriately tuned width and depth yielded a test accuracy of 86%. A MONAI DenseNet, pretrained on medical images, achieved an astonishing accuracy of 96%.

2 Datasets

The dataset used is OrganAMNIST, which contains grayscale medical images of various sizes, categorized into 11 classes representing different organs [1, 2]. Figure 1 reveals an imbalanced class distribution, with certain categories underrepresented, which could impact model performance. Likewise, the proportional distribution of classes is nearly identical in the training and test datasets. This suggests that our models' test scores may over-represent their ability to generalize.

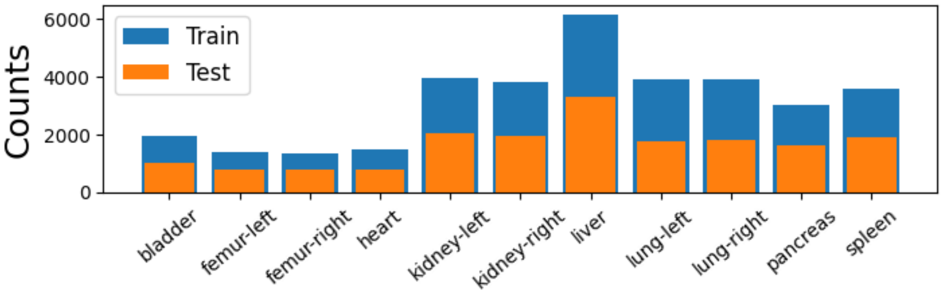

Figure 1: Distribution of data in different categories. There is an imbalance in category representation, with more images of livers than there are of hearts. The train and test distributions are similar.

3 Results

3.1 Hyperparameter tuning

We tune the learning rate, weight initialization and minibatch size for a MLP (2 hidden layers, ReLU activation) and find maximum test accuracy with a learning rate of 0.003 and minibatch size of 32, shown in figure 2(b). Initializing the weights to zero may lead to a vanishing gradient problem so we use Kaiming weight initialization, shown in figure 2(a) [3].

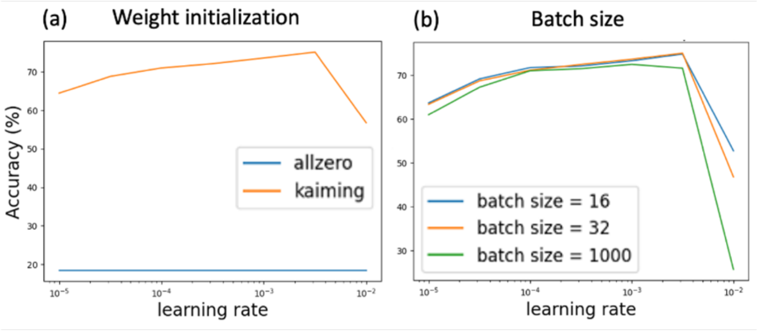

Figure 2: **a)** Test accuracy for kaiming and zero weight initialization across learning rates. **b)** Accuracy with different batch sizes across learning rates. All MLPs are trained for 10 epochs.

3.2 Network depth, non-linearity and network width

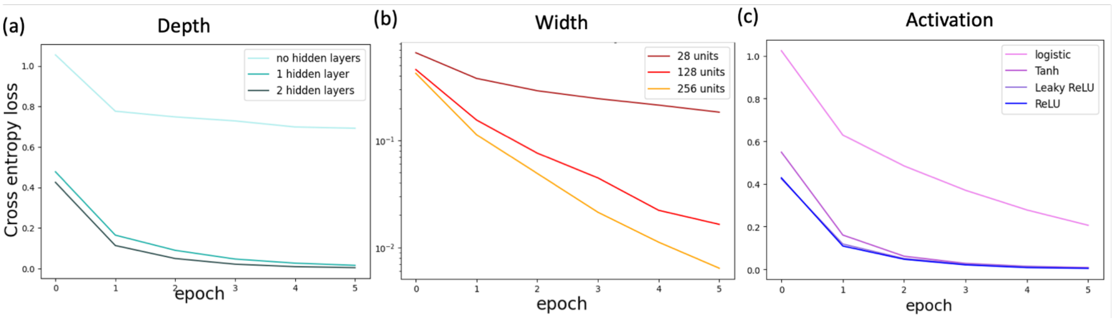

Figure 3: **a)** Cross entropy loss (CEE) of MLPs with various depth. Performance improves with depth. **b)** CEE of MLPs with various widths. Performance improves with width. **c)** CEE of MLPs with various activation functions. Performance is maximized with ReLU activation.

Depth (# of layers)	Width (# of units)	Activation functions	Train accuracy (%)	Test accuracy (%)
0	256	ReLU	80.6	55.7
1			99.9	73.8
2			100.0	75.3
2	28	ReLU	94.8	67.8
	128		99.6	73.3
	256		99.8	74.7
2	256	ReLU	100.0	75.3
		Leaky ReLU	99.9	75.1
		Tanh	99.8	70.6
		Logistic	95.5	70.6

Table 1: Accuracy for various network depths, widths and activation functions in an MLP trained for 6 epochs (learning rate = 0.003, minibatch size = 32, momentum=0.99). The highest test accuracy is 75.3%, achieved with 2 hidden layers that are 256 units wide with ReLU activation.

3.3 L1 and L2 regularization

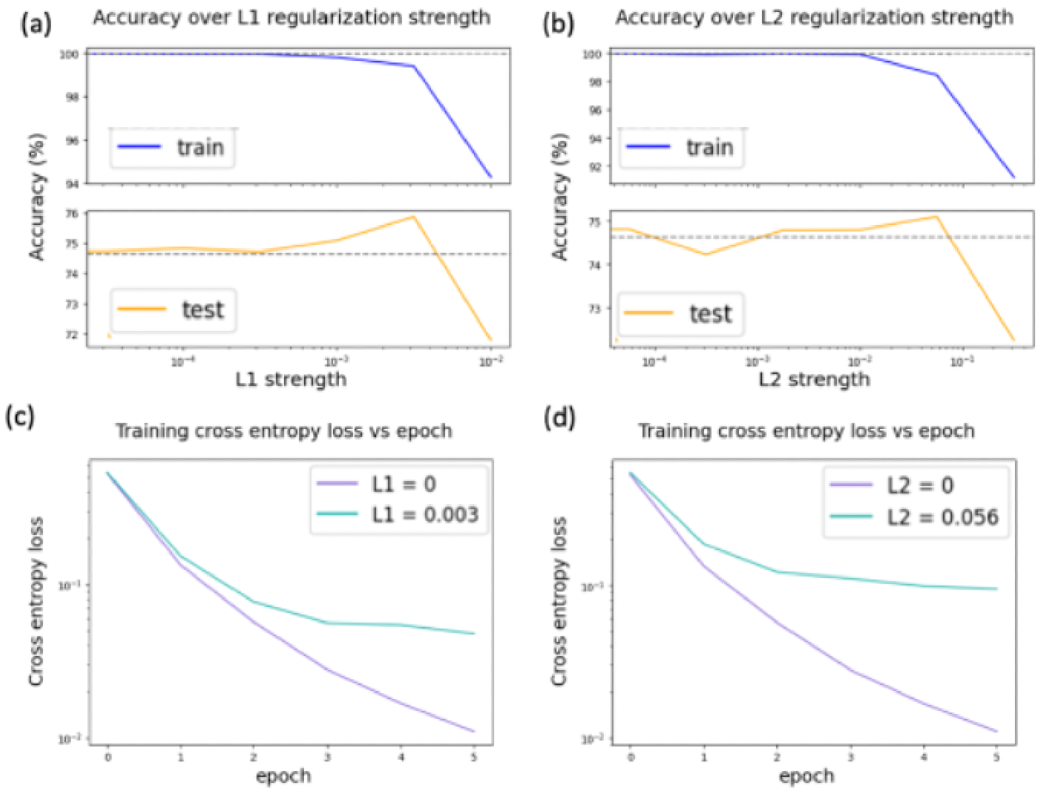

Figure 4: **a)b)** Training and test accuracy across regularization strength. Maximum test accuracy is 75.9%, achieved at L1 strength = 0.003 and L2 strength = 0.06. The dotted grey line shows the accuracy without regularization. **c)d)** Evolution of CEE across training epoch with and without optimal regularization.

The general trends for L1 and L2 regularization are similar, with weak regularization leading to overfitting, intermediate regularization leading to optimal test accuracy and overly strong regularization leading to a drop off in both train and test accuracy (figure 4(a)(b)). As the regularization strength increases, the model is less prone to overfit and so the training performance suffers, as seen in figure 4(c),(d).

3.4 Data normalization and image size

Before having trained any MLP on the data, we normalized the images. Figure 5 demonstrates the consequences of not normalizing data in pre-processing.

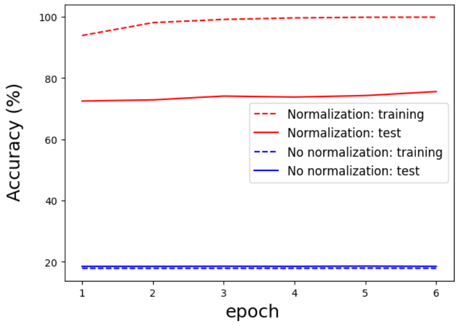

Figure 5: Evolution of train and test accuracy for a MLP (2 hidden ReLU layers with 256 units each, learning rate=0.003,momentum=0.99) with and without normalization in pre-processing. Normalizing the data before model training has a profound impact on the model performance as we are unable to surpass 20% accuracy without it.

Furthermore, training our MLP on a dataset that contains larger 128×128 pixel images results in decreased accuracy and slower training, as shown in figure 6.

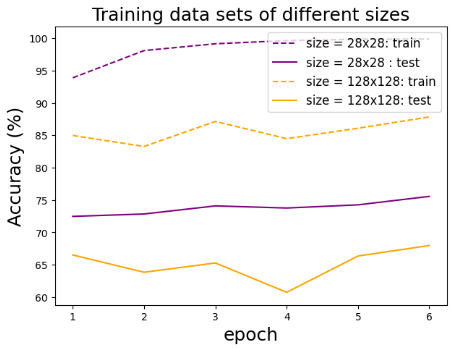

Figure 6: Evolution of train and test accuracy for a MLP (2 hidden ReLU layers with 256 units each, learning rate=0.003,momentum=0.99) trained on 28x28 and 128x128 pixel images, achieving maximum test accuracies of 75.3% and 69%, respectively. Training with the small images took 4.2 seconds per epoch. In contrast, training with the larger images took 98.3 seconds per epoch.

3.5 CNN

We now train a CNN, tuning the number of filters, filter size, stride and padding according to validation accuracy (Table 2). We implement batch norm, dropout and max pooling in all models. Training a CNN on the 28x28 OrganAMNIST dataset significantly improved performance, achieving a test accuracy of 88.4%, at the expense of slower training times than the MLP.

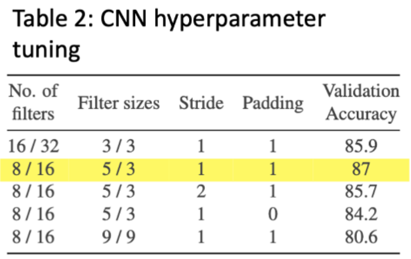

Table 2: CNN hyperparameter tuning

No. of filters	Filter sizes	Stride	Padding	Validation Accuracy
16 / 32	3 / 3	1	1	85.9
8 / 16	5 / 3	1	1	87
8 / 16	5 / 3	2	1	85.7
8 / 16	5 / 3	1	0	84.2
8 / 16	9 / 9	1	1	80.6

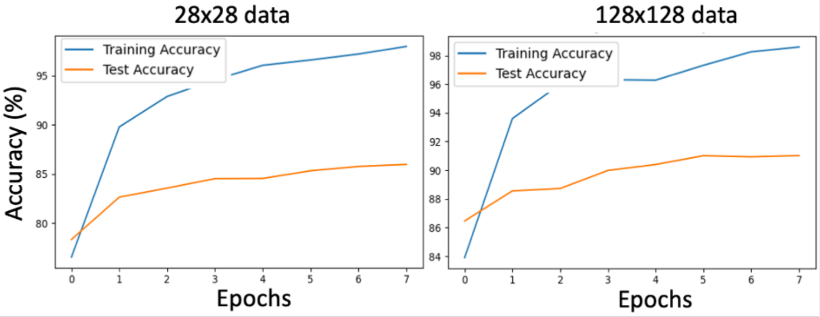

Figure 7: Evolution of training and test accuracy for 28x28 and 128x128 image data of the CNN selected in Table 2. The maximum test accuracy for the 28x28 model and the 128x128 model is 87% and 88.4%, respectively. The training time for the 28x28 and 128x128 data was 12 and 40 seconds per epoch, respectively.

3.6 Pre-trained models

We now test ResNet50, pretrained on ImageNet, with all layers frozen to leverage pre-trained features. The modified fully connected layers consist of two layers: the first has 2048 neurons, followed by Batch Normalization, ReLU activation, and Dropout (0.5), while the second layer has 1024 neurons. Table 3 shows how the number of neurons was selected by validation accuracy. Mixed precision training is implemented using PyTorch's GradScaler and autocast, which speeds up training and reduces memory usage. Each epoch takes 125 seconds and achieves a test accuracy of 86%. While this is an improvement over the MLP, the pre-trained model does not outperform the CNN. This is because ResNet50 is pretrained on ImageNet, which contains images such as animals, objects, and landscapes, which are feature-wise distinct from organ data. If we instead use a pretrained model trained on medical images such as MONAI DenseNet, we achieve a test accuracy of 96%.

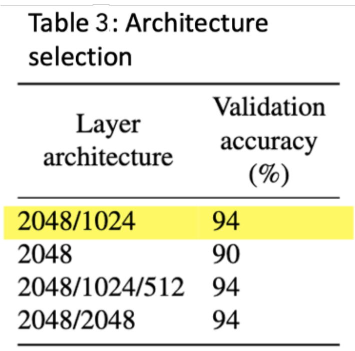

Table 3: Architecture selection

Layer architecture	Validation accuracy (%)
2048/1024	94
2048	90
2048/1024/512	94
2048/2048	94

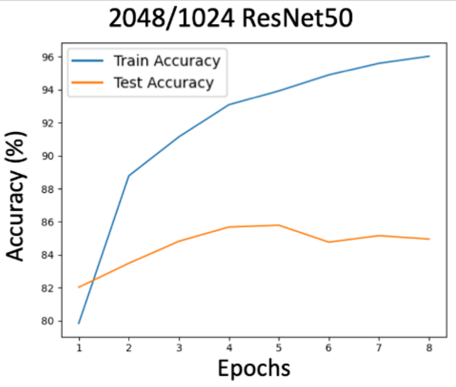

Figure 8: Evolution of accuracy achieved by the two-layer ResNet selected in Table 3 . The maximum test accuracy is 86%.

4 Discussion and Further Improvements

This project explored image classification using MLPs and CNNs on the OrganAMNIST dataset. Key findings include the importance of proper weight initialization (e.g., Kaiming[3]), data normalization, and regularization (L1 and L2) to achieve robust model performance. The optimal MLP configuration, with two hidden layers of 256 units and ReLU activation, indicates the necessity of non-linearity and deep-network abstraction in predicting organ image data. Likewise, the CNNs' improved performance compared to MLPs highlights how convolutional layers are critical to analyzing image data, given the transposability of identifiable regions. The marginal increase in accuracy gained by training on 128×128 images is largely outweighed by the dramatic increase in training time. Additionally, the improved performance of pre-trained convolutional layers highlights the efficacy of transfer learning over similar computer vision tasks. Future work could investigate regularization techniques such as dropout, data augmentation, and fine-tuning pre-trained models for enhanced accuracy and scalability.

Statement of contributions: All team members independently completed the entire assignment, including data preprocessing, model implementation, and experiments.

References

[1] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni, "Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification," *Scientific Data*, vol. 10, no. 1, p. 41, 2023.

[2] J. Yang, R. Shi, and B. Ni, "Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis," in *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 191–195, 2021.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.