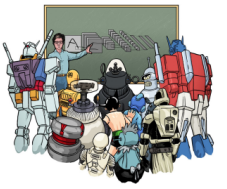
METU EE 449
Computational Intelligence



Saving the Princess with a Reinforcement Learning Agent

Ozgur Gulsuna

[a]*Middle East Technical University, Electrical and Electronics Engineering, Ankara, Turkey*

Introduction

This report explores the basics of ANNs and implement a convolutional layer using the NumPy, and PyTorch libraries. We experiment with different architectures, such as Multi-Layer Perceptron (MLP) and Convolutional Neural Network (CNN), and analyze the weights of the first layers. We also investigate the impact of different activation functions and learning rates on the performance. Additionally, we explore scheduled learning and how it can improve the performance of our models.

Keywords: PyTorch; NumPy; Multi Layer Perceptron; Convolutional Neural Network; Activation Functions; Learning Rate; Scheduled Learning;

1. Basic Concepts

1.1. Which Function ?

An ANNs classifier that is trained with cross-entropy loss approximates the conditional probability distribution function. More specifically, for an input data, the output of the classifier is a probability distribution for the classes. The cross-entorpy loss function is a measure between the predicted probability distribution and the true distribution. The form of the loss function is decreasing, smooth and differentiable, which makes it easier to optimize using gradient-based methods. This form is also known as the negative log-like function.

1.2. Gradient Computation

High number of iterations, when the difference between the weights are small, the gradient calculation can be made with basic slope calculation.

$$\gamma \nabla \mathcal{L}_{\omega_k} = \omega_k - \omega_{k+1} \qquad \textit{hence,}$$
$$\nabla_\omega \mathcal{L}\big|_{\omega=\omega_k} = \frac{\omega_k - \omega_{k+1}}{\gamma}$$

E-mail address: ozgur.gulsuna@metu.edu.tr

1.3. *Some Training Parameters and Basic Parameter Calculations*

1. The batch refers to a subset of the training data that is used to compute the weights for one iteration. More specifically, the batch size is the number of training samples in a batch. The epoch on the other hand refers to the number of times the entire training data is used to update the weights. In training, there are generally multiple epoch iterations where the weights are updated with different batches/subsets of the training data.
2. For the N number of training samples, the number of batches per epoch is N/B, where B is the batch size. A little side note that the solution is rounded up to the higher integer if N/B is not an integer.
3. For the optimization iterations, such as SGD, for E number of epochs, the total number of iterations is $E \times N/B$. Again, a practical side note states that the N/B is rounded up to the higher integer.

1.4. *Computing Number of Parameters of ANN Classifiers*

1. Starting from the initial layer of the MLP, we have D_{in} number of input neurons and H_1 number of neurons in first hidden layer. Also there are biases associated with each neuron. Therefore, the number of parameters of the each layer is,

$$\begin{aligned} \text{Input Layer} &= D_{in} \times H_1 \ + \ H_1 \\ \text{Hidden Layers} &= H_1 \times H_2 + H_2 \\ &\cdots \\ \text{More Hidden Layers} &= H_{k-1} \times H_k + H_k \\ \text{Output Layer} &= H_k \times D_{out} \ + \ D_{out} \end{aligned}$$

   The total sum can be written as, where K is the number of hidden layers.

$$\text{Total Number of Parameters} = D_{in} \times H_1 + \sum_{k=2}^{K} (H_{k-1} \times H_k + H_k) + H_k \times D_{out} + D_{out}$$

2. CNN structure is more complicated. The number of parameters of a CNN layer is calculated as follows:
   For the input layer, the number of parameters is,

$$\text{Input Layer} = [(H_{in} - H_1 + 1) \times (W_{in} - W_1 + 1) \times C_{in} \times C_1] + C_1$$

   where H_{in} and W_{in} are the height and width of the input image, and C_{in} is the number of channels of the input image. Each input of layer is the output of the previous layer. There exist also biases associated with each neuron added. For the convolutional layers, the number of parameters is calculated as,

$$\text{Convolutional Layer} = [(H_k - H_{k+1} + 1) \times (W_k - W_{k+1} + 1) \times C_k \times C_{k+1}] + C_{k+1}$$

   Combination of all layers is,

$$\text{Convolutional Layers} = \sum_{k=1}^{K} [(H_k - H_{k+1} + 1) \times (W_k - W_{k+1} + 1) \times C_k \times C_{k+1}] + C_{k+1}$$

   Here all the parameters are summed up. The output is assumed to be the last index of the array. The final equation for the total number of parameters is,

$$\text{Total Parameters} = [(H_k - H_{k+1} + 1) \times (W_k - W_{k+1} + 1) \times C_k \times C_{k+1}] + C_{k+1} + \sum_{k=1}^{K} [(H_k - H_{k+1} + 1) \times (W_k - W_{k+1} + 1) \times C_k \times C_{k+1}] + C_{k+1}$$

2. Implementing a Convolutional Layer with NumPy

The section involves implementing conv2d function using NumPy for forward propagation and testing it on a small batch of MNIST dataset. We downloaded and loaded input and kernel files, and created an output image using the part2Plots function. The implementation code can be found in the appendix named my_conv2d.py. We confirmed the correctness of our implementation by the output image.

2.1. Experimental Work

The generated output for t 1.

Fig. 1. Convolution over the number 8 of the MNIST dataset

2.2. Results and Discussion

1. The Convolutional Neural Networks are important for couple of reasons. First of all, when the input shape of the CNN is selected as 2D, it is well suited image processing. Second, the CNN is able to learn the spatial features of the input image. This also means that the CNN is able to learn and extract the features of the input image without any manual work. CNN's are also able to recognize the features of the input image even if the input image is rotated or scaled. Since the features are extracted from image, partial occlusion of the image affect the performance of the CNN less.
2. The kernel of a Convolutional Layer is essentially a matrix of weights that is convolved (inversely correlated) over with the input data to extract features. The size of the kernel refers to the number of rows and columns in the matrix. It corresponds to the reception of the filter, meaning that higher sizes can extract more complex features.
3. The output image shows that the convolution of pre-presented kernels for the number 8 of the MNIST dataset. Basically each filter is convoled over the images to grasp different meanings. Each column is another kernel with each row is different input image.
4. The numbers in the same column look like each other since they both have a representation of the same number and the same kernel is able to extract the features related to the number 8 other than the specific image.
5. The numbers in the same row look different althougn the input image the same. This is because the kernels are different and they are able to extract different features from the same image.
6. For more specific examples, the third column kernel represents that an 8 has two "islands" of white patches in the middle but the size, shape and location of these pathes differ for each 8 altough all of them represent the same thing in a different manner. Another column such as 6, implies the white track like feature of the number 8. In this sense some features are more dinstinctive than other however when different of these combined make the action work even though they do not seem to represent a clear feature. This is similar to human behaviour as we associate the similar patterns to the general inputs and this is the importance of the convolutaional layers, the features can be learned in a sense.

3. Experimenting ANN Architectures

3.1. Experimental Work

This experimental work focuses on testing various Artificial Neural Network (ANN) architectures for a classification task. The models will use adaptive moment estimation (Adam) with default parameters for the optimizer. The datasets will be preprocessed and split into three sets: training, validation, and testing. The ANN architectures to be tested are ‘mlp 1’, ‘mlp 2’, ‘cnn 3’, ‘cnn 4’, and ‘cnn 5’, each with their specific layers. For each architecture, the ANN will be trained for 15 epochs, and training loss, accuracy, validation accuracy, and test accuracy will be recorded. The best test accuracy and weights of the first layer will be recorded, and a dictionary object will be created and saved for each architecture. Performance comparison plots will be created, and the weights of the first layer of all architectures will be visualized.

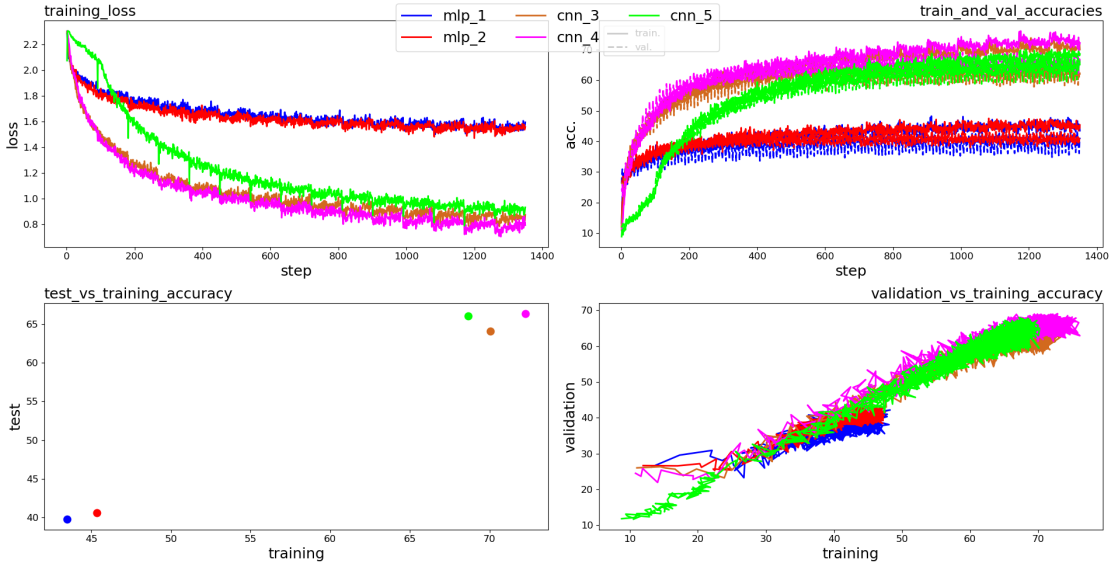
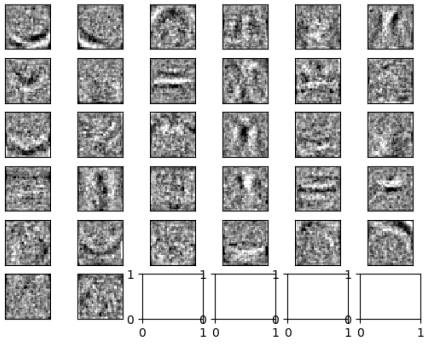


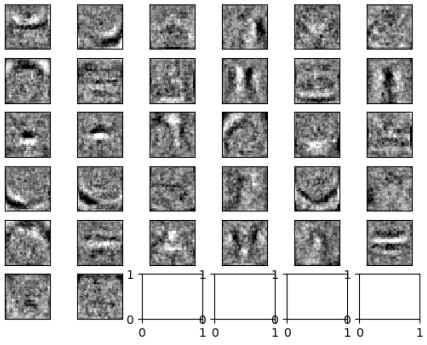
Fig. 2. Performance Comparison Plots for the ANN Architectures

3.2. Results and Discussion

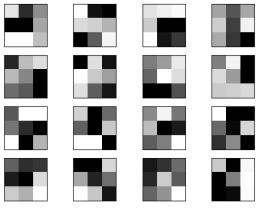
1. Generalization performance refers to the ability of the model to to classify the unseen data. It is used as the ability to recognize patterns and apply this knowledge to the new data.
2. The plots showing the results with the ”test” data show the generalization performance since the test data is not used in the training process and the model is not familiar with the test data. Validation data is also seemed to be a good indicator of the generalization performance. However, altough the validation data is not directly used in training process, it is used to tune the parameters of the model. Therefore, the model is familiar with the validation data in a sense. The first two curves and the last x-y plot give hint about the comparative generalization performance since these show the results with the validation data. The third scatter plot on the other hand is used with the test data, hence have the correct generalization performance. However it only shows the best run hence the variety of the results are a topic of discussion and the plot does not show that information.
3. Copmarative results show that the convolutional architectures perform better than the multi-layer perceptron architectures. This is because the convolutional layers are able to extract spatial features from the data with more grasping ability. The ”mlp_1” and ”mlp_2” are very similar in terms of performance although they have different size of parameters and number of layers. The ”cnn_3” and ”cnn_4” are also very similar again the latter has more layers. The ”cnn_5” is more of a slow learner and could not get to the same level of performance as the ”cnn_4” but with more epochs it is seem to be able to surpass the ”cnn_4” since the gradient of the accuracy is increasing.



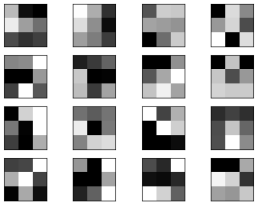
a) mlp_1



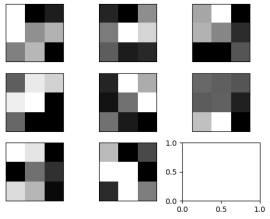
b) mlp_2



c) cnn_3



d) cnn_4



e) cnn_5

4. Higher the number of parameters, it is generally easier for model to learn complex features. However it also means that the model is more prone to overfitting. This is because the model is able to learn the training data in more depth, like its noise characteristics not the required features. Hence, it is not able to generalize well. This is called overfitting, the models with higher parameters have more ”memory” that they can memorize the unwanted characteristics that is specific to the training data. Another aspect is the distribution of the parameters, the convolutaional layers use the parameters more eficiently and make more of the increased number of parameters without easily falling into overfitting trap.
5. The depth of the architecture is also relevelant with the distribution of the parameters, how they organized in an architecture. The models with more depth are able to learn more complex features as well, however they are harder to train in terms of computation. The extremum of depth parameter results in not overfitting but underfitting. This is because the model is not able to learn the features of the data in depth and generalize well.
6. The first layer weights of the MLP structure has high resolution, and can be interpreted as some curves and horizontal lines are perceived. for the CNN structures the weights are more abstract but these architectures have more depth hence a meaning can be found when these are all evaluated together.
7. It is early to say that the units are related with the classes with initial layer weights but horizontal lines can be the plane wing identifier with the tail and body sections are interpreted in the latter layers.
8. At this stage, with the more information on MLP’s first layer, it is more interpretable.
9. For each arhchiecture, the depth is increased as explained in previous sections. CNN have an advantage in classification of an image input over the MLP. The higher depth suggests that the CNN is able to learn more complex features and generalize better.
10. I would go with the CNN architecture for image classification task. More specifically ”cnn_4” which performs around the same as others with faster training and less parameters as ”cnn_5”.

4. Experimenting Activation Functions

4.1. Experimental Work

This section compares the performance of artificial neural networks (ANNs) using the rectified linear unit (ReLU) and logistic sigmoid activation functions, trained with stochastic gradient descent (SGD) on a constant learning rate of 0.01, 0.0 momentum, and a batch size of 50 samples. For each architecture in section 3.1, two torch.nn.Module objects are created with ReLU and logistic sigmoid activations respectively. The ANNs are trained for 15 epochs, recording the training loss and magnitude of the loss gradient at every 10 steps. The results are saved as dictionary objects with filenames prefixed with 'part4'. Finally, performance comparison plots are generated.

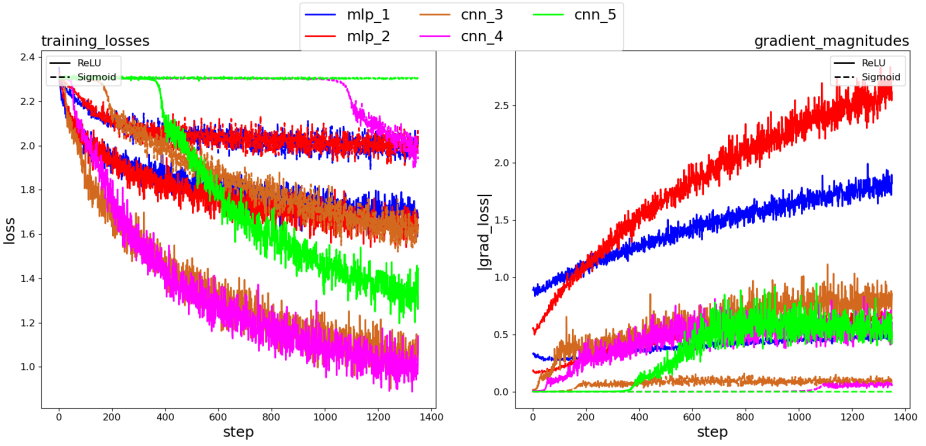


Fig. 3. Performance Comparison Plots for the Different Activation Functions over the Different Architectures.

4.2. Results and Discussion

1. As the depth increases, the normalized gradient of the loss decreases. Also for the further steps it does not change much, meaning that the learning gets slower. The most complex model "cnn_5" has the lower gradient of the lost for starting epochs. At a point, also loss gradient reduces after a peak value. This also suggests that the training will get slower. The shallow models "mlp_1" and "mlp_2" have relatively high gradients of loss. However, from the minimum loss plot, it is seen that the loss is not decreasing much after a point. This is due model is not deep enough to learn more of the features.
2. Increasing with the depth, gradients start to become harder to calculate. This is because the gradients are calculated by backpropagation and the gradients of the previous layers are needed to calculate the gradients of the current layer. Increasing the complexity at each iteration.
3. Bonus: Since the gradients are calculated from the weights, the unnormalized input can cause gradients to become very small and the learning could take longer. Also for different scales of inputs, the model can give bias to one or more features, which can also reduce the performance of both the training process and the overall system.

5. Experimenting Learning Rate

5.1. Experimental Work

This section explores the effect of varying learning rates in the stochastic gradient descent (SGD) optimization method with a ReLU activation function, 0.0 momentum, and batch size of 50 samples. Three torch.nn.Module objects are created with initial learning rates of 0.1, 0.01, and 0.001, respectively, and trained for 20 epochs. The training loss and validation accuracy are recorded every 10 steps to form loss and accuracy curves. Scheduled learning rate is then explored by training a classifier with a 0.1 learning rate until the epoch step where the validation accuracy stops

increasing. The learning rate is then reduced and training continues for another 30 epochs. Finally, the test accuracy of the trained model is compared to the same model trained with Adam in section 3.1.

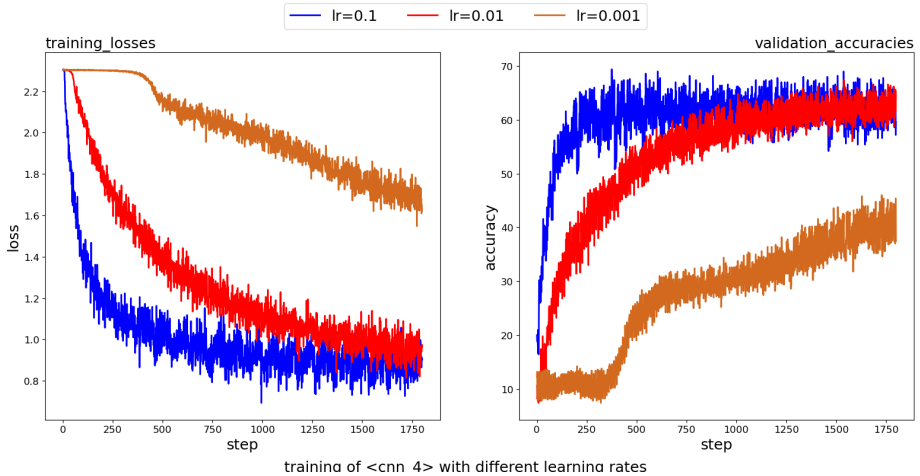


Fig. 4. Performance Comparison Plots for the Different Learning Rates of "cnn_4".

5.2. *Results and Discussion*

1. Higher learning rates lead to faster convergence, however the model can skip the global minimum. Lower learning rates can lead to a slower convergernce.
2. High learning rate might not get to the minimum point but oscillate around it. Lower has better chance to get to the minimum point.
3. The scheduled learning rate iteratively reduces the learning rate, promising a better convergence with a better step count.
4. With the scheduled learning rate, there is a jump at the epoch 7, which is the point where the learning rate is reduced. The next jump is at epoch 17, however the performance does not increase much after that, suggesting a minimal point is reached. One thing to notice that the training accuracy is higher than the respected model in part 3.1. But the validation accuracy is around the same at 66%.