Benjamin Ye

CS/CNS/EE 156a: Learning Systems (Fall 2023)

November 17, 2023

**Bonus Exercise**

**1 Classifying digits with NNs**

**1.3 Number of Parameters**

What happens to the learning curve when you vary the number of hidden units? Specify the number of hidden units you use in each layer, and report learning curves for each different architecture you try. What trends do you notice?

As the number of hidden units in the first layer decreases, the accuracy decreases and the loss increases on both the training and validation sets, but much more significantly on the training set. For example, the figure on the next page shows a downwards shift of both accuracy curves and an upwards shift of both loss curves as decreases from to . This is likely caused by the order of magnitude difference in the number of parameters (see tabulated results in the following pages). In the other direction, we get diminishing returns; when the number of hidden units doubles in the first layer, the training accuracy and loss improves nearly imperceptibly, and the validation accuracy and loss worsen slightly. This is indicative of a sufficient number of parameters in the original model ( and ) and suggests that the default dimensionalities are the optimal values since they maximize the accuracies while minimizing the losses and the number of parameters.

The performance of the model is extremely insensitive to the number of hidden units in the second layer for in the range of explored, with minimal fluctuations in the accuracies and losses as increases or decreases, because the changes in the number of parameters are inconsequential compared to when is varied. It is expected that when is extremely small, i.e., , the model will perform poorly because the number of parameters will have decreased an order of magnitude.

In the following figures, the dimensionality of the first layer is adjusted from the default .

A graph of a number of numbers and a line

Description automatically generated with medium confidenceA graph of a number of numbers and a line

Description automatically generated with medium confidence

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generatedA graph of a number of data

Description automatically generated with medium confidence

In the following figures, the dimensionality of the second layer is adjusted from the default .

A graph of a number of numbers and a line

Description automatically generated with medium confidenceA graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generatedA graph of a number of data

Description automatically generated with medium confidenceA graph of a number of data

Description automatically generated with medium confidence

The table below summarizes the number of parameters and the performance of the models as and are varied. The values original/default model are bolded for comparison.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 25 | 25 | 20535 | 0.9692 | 0.1026 | 0.9612 | 0.1264 |
| 25 | 50 | 21435 | 0.9739 | 0.0865 | 0.9661 | 0.1139 |
| 25 | 100 | 23235 | 0.9758 | 0.0788 | 0.9672 | 0.1101 |
| 25 | 200 | 26835 | 0.9786 | 0.0692 | 0.9694 | 0.1025 |
| 50 | 25 | 40785 | 0.9835 | 0.0541 | 0.9718 | 0.0941 |
| 50 | 50 | 42310 | 0.9849 | 0.0500 | 0.9745 | 0.0880 |
| 50 | 100 | 45360 | 0.9845 | 0.0500 | 0.9701 | 0.1001 |
| 50 | 200 | 51460 | 0.9888 | 0.0360 | 0.9738 | 0.0902 |
| 100 | 25 | 81285 | 0.9911 | 0.0299 | 0.9643 | 0.1240 |
| 100 | 50 | 84060 | 0.9911 | 0.0288 | 0.9758 | 0.0843 |
| 100 | 100 | 89610 | 0.9928 | 0.0244 | 0.9747 | 0.0858 |
| 100 | 200 | 100710 | 0.9930 | 0.0224 | 0.9760 | 0.0860 |
| 200 | 25 | 162285 | 0.9947 | 0.0181 | 0.9761 | 0.0823 |
| 200 | 50 | 167560 | 0.9954 | 0.0150 | 0.9741 | 0.1060 |
| **200** | **100** | **178110** | **0.9950** | **0.0154** | **0.9801** | **0.0722** |
| 200 | 20 | 199210 | 0.9951 | 0.0150 | 0.9789 | 0.0791 |
| 400 | 25 | 324285 | 0.9960 | 0.0123 | 0.9802 | 0.0773 |
| 400 | 50 | 334560 | 0.9958 | 0.0131 | 0.9786 | 0.0839 |
| 400 | 100 | 355110 | 0.9955 | 0.0141 | 0.9802 | 0.0756 |
| 400 | 200 | 396210 | 0.9960 | 0.0121 | 0.9750 | 0.1024 |
| 47 | 25 | 38355 | 0.9808 | 0.0628 | 0.9700 | 0.1037 |

What is the smallest number of model parameters (not hidden units; view this via the summary function) for which you can achieve over 97% validation accuracy? An estimate is fine.

Based on the tabulated results above and some cursory testing, it appears that (rounded up to the nearest ten thousand) will give in ten epochs.

**1.4 Regularization**

Explore the usefulness of weight decay regularization (minimizing instead of just ) on the large neural net implementation given in the support code. You can do so changing the REGULARIZATION parameter in the IPython notebook, or by specifying the -r command line option in train.py in the command shell. Report learning curves for different choices of (just a few).

The value was varied from to . Below, the accuracy and loss curves are shown for a subset of the runs, and the final accuracies and losses are tabulated.

A graph of a graph of a number of individuals

Description automatically generated with medium confidenceA graph of a number of numbers

Description automatically generated with medium confidence

A graph of a number of numbers

Description automatically generated with medium confidenceA graph of a number of numbers

Description automatically generated with medium confidence

(Discussion continued on the following page.)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **178110** | **0.9950** | **0.0154** | **0.9801** | **0.0722** |
|  | 178110 | 0.9952 | 0.0143 | 0.9786 | 0.0754 |
|  | 178110 | 0.9954 | 0.0145 | 0.9791 | 0.0720 |
|  | 178110 | 0.9950 | 0.0170 | 0.9790 | 0.0755 |
|  | 178110 | 0.9920 | 0.0723 | 0.9809 | 0.1126 |
|  | 178110 | 0.9154 | 0.7611 | 0.9163 | 0.7514 |
|  | 178110 | 0.1124 | 2.3013 | 0.1135 | 2.3011 |

Compared to the model with no regularization, the models with and performed slightly better in the training accuracy and loss but regressed slightly in the validation accuracy. All other models with performed worse in all metrics, suggesting significant underfitting as .

Can you use regularization to get a similar result to one of the smaller neural nets you implemented in the previous part?

No. Regardless of the value used, the validation accuracy using the smaller model with and hovers around 0.97 and is always lower than the 0.995 attainable using the larger model.

(My understanding of this question is whether the smaller neural net from 1.3 that can achieve a validation accuracy of 0.97 is able to compete with the larger neural net if regularization with the optimal value is used. If the question is be interpreted literally, then the answer is yes, the large neural net can perform worse and achieve a validation accuracy of 0.97 like the smaller neural net if a value between and is used.)

**1.5 Activations**

In class, we have primarily discussed using the and sigmoid activations. In practice nowadays, people often use the ReLU activation, which is defined as . We use this activation function in the support code, and in practice, people have observed that it results in much faster convergence. Verify this for yourself by replacing the ReLU activation with a activation. Why do you think this happens? (Hint: How does behave as ? Compare this to .)

The ReLU and activations performed similarly in terms of convergence. The accuracy and loss curves are shown below on the right, with those for the default neural net on the left for comparison.

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generatedA graph of a number of data

Description automatically generated with medium confidence

Generally, the ReLU activation converges faster than the activation because at large values of , plateaus to a value of 1 while continues to grow linearly. As such, the gradient of vanishes, which impedes the learning process, but not the gradient of ReLU.

**1.6 Different Architectures**

Now, let’s say that I give you a fixed budget of 200 hidden units. What is the best validation accuracy you can achieve? Feel free to vary the number of layers, the kind of regularization (e.g., which is based on the 1-norm of as opposed to which is the usual weight decay based on the 2-norm) and its strength (), the activation, and the optimizer you use.

**1.7 Convolutional Neural Nets**

By learning a few smaller convolutional filters instead of a series of huge matrices, a convolutional neural net performs much better on this image classification task with a fraction of the parameters. Verify this for yourself. Run the train script and the evaluate script using the convolutional neural net and no regularization. Compare the number of parameters between any of the convolutional and dense neural nets, and report learning curves.

Indeed, the convolutional neural net can achieve comparable accuracies and losses as the default neural net, but with two orders of magnitude fewer parameters (7,240 vs. 178,110). Notably, the validation accuracy is higher and now comparable to the training accuracy. The learning curves are shown below on the right, with those for the default neural net on the left for comparison.

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generatedA graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

Description automatically generated

**2 Generative Adversarial Network**

**2.2 How does a GAN work?**

1. Since this solution occurs at a saddle point, let’s take the derivative of the loss with respect to . Since is a function, take a functional derivative (like a vector derivative, don’t worry about being rigorous). Use the following loss,

Hint: We can interchange an integral and a functional derivative:

Don’t forget the chain rule!

1. Now that we have the derivative, we solve for in terms of the probability distributions and when . Hint: Move one integral to the other side of the equality. Notice that if the values under both integrals are the same pointwise, we also have equality with integrals. Solve with respect to .
2. You should get

Do you think this is reasonable?

Seems reasonable. Consider the case when , or when the distribution of images as classified by the discriminator matches that of real and generated images created by generator . As expected at this quasi-Nash-equilibrium state, assigns the (expected) probability of to both real and generated images because it is unable to distinguish between them, indicating that has created images realistic enough to not be detectable as fakes.

**2.4 Reflecting on GAN Behavior**

How would you describe the training process in terms of the images that the model generated?

At the beginning (epoch 0), the images are grayscale artifacts that do not resemble anything.

As training commences, the images start to resemble numbers (by around epoch 10) but are far from being able to be recognized as any specific number. Some simpler numbers, like 1, becomes legible due to their single stroke.

More complex numbers become distinguishable at around epoch 20. However, most look unnatural because of the gray uncertainty between the white digits and the black background and can easily be identified as generated images.

Further refinement occurs all the way up to the final epoch (150). In the final generated image, many of the digits look like they could have been written by a human (with bad handwriting), largely because the edges between the numbers and the backgrounds are much cleaner and pronounced.

What about the learning dynamics in the loss values for the generator and the discriminator?

A large concern in training GANs is a phenomenon known as “mode collapse”. Look up this term and explain it in your own words (it doesn’t have to be rigorous).

Mode collapse occurs when the generator continuously creates data samples are that are too similar or even identical because it has learned that certain features are more likely to fool the discriminator. As a result, the distribution of samples does not align with that of the real-world data. Worse yet, the discriminator learns to classify all samples coming from the generator as real images, so the discriminator does not provide the right feedback to the generator so that it can generate more realistic samples.

Skimming over the code, did the current generator’s implementation manage to avoid mode collapse?

The BatchNormalization layers in the generator can help stabilize the training process by preventing a vanishing gradient (which causes the generator to continue producing similar samples).

**3 Feedback**

Overall, this was a fantastic way to get hands-on experience with perhaps the most widely used machine learning package (Keras and TensorFlow). Furthermore, the exercise reinforced most of the concepts covered in class and provided a practical application to showcase why certain algorithms or techniques lead to better trained models.

However, there are a few technical issues with the provided scripts, such as outdated Keras and TensorFlow packages, that caused some classes and object to not load correctly without some debugging. Moreover, the default neural net for Part 1 performed perhaps a bit too brilliantly, so much so that I was unable to get better results with weight decay regularization, show that the inferior activation should have worse convergence, or highlight the improvements of a convolutional neural net. (It is also entirely possible that I implemented the approaches incorrectly.)