UNIVERSITY OF PENNSYLVANIA

ESE 546: PRINCIPLES OF DEEP LEARNING

Changelog

Read the following instructions carefully before beginning to work on the homework.

- You will submit solutions typeset in LaTeX on Gradescope (strongly encouraged). You can use hw_template.tex on Canvas in the "Homeworks" folder to do so. If your handwriting is unambiguously legible, you can submit PDF scans/tablet-created PDFs.
- Clearly indicate the name and Penn email ID of all your collaborators on your submitted solutions.
- Start a new problem on a fresh page and mark all the pages corresponding to each problem. Failure to do so may result in your work not graded completely.
- For each problem in the homework, you should mention the total amount of time you spent on it. This helps us keep track of which problems most students are finding difficult.
- You can be informal while typesetting the solutions, e.g., if you want to draw a picture feel free to draw it on paper clearly, click a picture and include it in your solution. Do not spend undue time on typesetting solutions.
- You will see an entry of the form "HW 2 PDF" where you will upload the PDF of your solutions. You will also see entries like "HW 2 Problem 3 Code" where you will upload your solution for the respective problems.
- **For each programming problem/sub-problem, you should create a fresh .py file**. This file should contain **all** the code to reproduce the results of the problem/sub-problem, e.g., it should save the plot that is required (correctly with all the axes, title and legend) as a PDF in the same directory. You will upload the .py file as your solution for "HW 2 Problem 3 Code". Name your file as pennkey_hw2_problem3.py, e.g., I will name my code as pratikac_hw2_problem3.py. Note, we will not accept .ipynb files (i.e., Jupyter notebooks), you should only upload .py files. If you are using Google Colab to do your homework (and I suggest that you don't...), you can export the notebook to a .py file.
- **This is very important**. Note that the instructors will download your code and execute it themselves, so your code should be such that it can be executed independently without any errors to create all output/plots required in the problem.

Credit The points for the problems add up to 110. You only need to solve for 100 points to get full credit, i.e., your final score will be $\min(\text{your total points}, 100)$.

Problem 1 (25 points). The torchvision library at https://pytorch.org/vision/stable/models.html implements a number of popular architectures that you can use quickly in your code. In this problem, you will take a deeper look at residual networks at https://github.com/pytorch/vision/blob/master/torchvision/models/resnet.py. Understand how this architecture is coded up.

(a) (2 points) Note down all the peculiarities that you notice in the code. E.g., which one is better: using a batch-normalization layer before ReLU or after ReLU; what does this code do?

(b) (2 points) What do the calls "model.train()" and "model.eval()" do and where do you use them in a typical training and validation code? Why did we not have them in HW 1 when we wrote our own library for training deep networks?

(c) (3 points) Draw a rough picture of the Resnet-18 architecture and note down the number of parameters in each layer; you will find it easier to to write a function that computes the number of parameters in each layer of the network.

(d) (3 points) Weight decay should not be applied to the biases of the different layers in the network, argue why this is the case.

(e) (5 points) Write the code to iterate over all the network layers in Resnet-18 and separate out the parameters in three groups: (i) batch-norm affine transform parameters, (ii) biases of convolutional and fully-connected layers, and (iii) all the rest. There is no need to submit the code separately in this case, since these are a few lines of code just copy them out into your PDF solutions.

(f) (10 points) Augmentations are increasingly becoming much more important for deep learning that what we initially believed. And therefore there are a number of sophisticated ways to perform random augmentation. Go through the paper https://arxiv.org/abs/1909.13719 (you don't have to read the details in order to do this question) and its implementation in torchvision at https://pytorch.org/vision/main/generated/torchvision.transforms.RandAugment.html. The code of this implementation is at https://pytorch.org/vision/main/_modules/torchvision/transforms/autoaugment.html#RandAugment.

Take any one image of your choice (you can use the same image of the astronaut as that of the examples) and show the result of augmenting this image using the following 10 augmentations: (a) ShearX, (b) ShearY, (c) TranslateX, (d) TranslateY, (e) Rotate, (d) Brightness, (e) Color, (f) Contrast, (g) Sharpness, (h) Posterize, (i) Solarize and (j) Equalize. You will use existing functions from torchvision to perform these augmentations (see the code of RandAugment linked here to understand how to call each of them). Each of these augmentations has some parameters that you will have to choose in order to run these corresponding augmentation functions. Any reasonable values are okay, feel free to experiment. The goal of this problem is to understand how these augmentations work. You don't have to submit code in this case, pictures in the PDF correctly annotated with the parameters of the augmentation that was used to create them are fine.

Problem 2 (20 points). Non-convex optimization problems are harder than convex optimization problems. There are however a few special non-convex problems that are easy. We will look at one of them here, namely unconstrained matrix factorization. Given a matrix $X \in \mathbb{R}^{m \times n}$ we would like to decompose it into two matrices of rank at most r

$$X = AB$$

where $A \in \mathbb{R}^{m\times r}$ and $B \in \mathbb{R}^{r\times n}$. Think of arranging all your data as columns of X. Columns of the matrix A are like elements of a dictionary, they correspond to different patterns in the data and are called atoms. The matrix B chooses which patterns to collect together in order to create a particular datum, i.e., column of X. Solving for factors A, B is usually done with constraints, e.g., B is typically forced to be a sparse matrix which enables regenerating data X using as few atoms as possible. We will solve a simpler problem:

$$A^*, B^* = \operatorname*{argmin}_{A\in\mathbb{R}^{m\times r},\ B\in\mathbb{R}^{r\times n}} \|X - AB\|_F^2.$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

(a) (4 points) Why is the above problem not convex?

(b) (8 points) The global optimum for this loss function can be obtained easily in spite of it being non-convex; find it. You may find it useful to write down the SVD of X.

(c) (8 points) Is the solution to the above optimization problem unique? Given one solution A^*, B^* name one way using which you can obtain another solution.

Problem 3 (45 points). Use Google Colab to train the neural network, but after debugging everything on your laptop first. If you have a new laptop then you should also be able to do everything on it without Colab. Neural networks are high-dimensional classifiers. While we may be able to train and regularize SVMs to have a large margin between two classes, it is often difficult to do the same for neural networks. A consequence of this is that, roughly speaking, all samples in typical training datasets lie close to the decision boundaries after training. This makes it quite easy to make minor perturbations to the input image—perturbations that are imperceptible to the human eye—and send the sample across the decision boundary so as to cause the network to mis-predict. We will synthesize such adversarial perturbations in this problem. You can read more about it at https://arxiv.org/abs/1412.6572; however be wary of the heuristic generalizations in this paper that are incorrect.

We will find the best adversarial perturbation to a given image x and its target y, this amounts to solving the optimization problem

$$\max_{\|x'-x\|_\infty \leq \epsilon} \ell(x', y) \tag{1}$$

where x' is the *variable of optimization* and it is the adversarially perturbed image corresponding to x, the quantity $\ell(x', y)$ is the loss computed on the image x' for the label y. We have chosen to make the parameters of the classifier w implicit for sake of clarity. This optimization problem searches for all images within an ϵ-ball of the original image x. We will use the ℓ_∞-norm

$$\|x\|_\infty = \max_k |x_k|$$

in this problem. Let us perform the Taylor expansion of the objective Eq. (1)

$$\ell(x', y) = \ell(x, y) + \epsilon\, d^\top \nabla \ell(x, y) + \mathcal{O}(\epsilon^2);$$

here $d = (x' - x)/\epsilon$. We can now write an approximate problem for finding adversarial perturbations as

$$\max_{\|d\|_\infty \leq 1} d^\top \nabla \ell(x, y).$$

Notice that the constraint implies that any element of the vector d can be perturbed by at most 1; there is no limit on the number of elements perturbed. The value of d that maximizes this objective is therefore the "signed gradient"

$$d_k = \frac{\nabla \ell(x, y)_k}{|\nabla \ell(x, y)_k|};$$

where $|\cdot|$ denotes the element-wise absolute value. If $\nabla \ell(x, y)_k < 0$, the corresponding $d_k < 0$ and vice versa. The maximal objective is $\|\nabla \ell(x, y)\|_1$. The perturbation d is what we want to compute.

(a) (25 points) We will first train a convolutional neural network for this problem with all the bells and whistles. You can use the code for the model and training provided on Canvas. This is a small model with about 1.6M parameters. Train this model on the CIFAR-10 dataset for 100 epochs, you should try to get a validation error below 12%. You should use a GPU on Colab for training; else your code will be very slow. Roughly speaking, running for 100 epochs will take 1-2 hours, so be patient. You can use data augmentation such as mirror flips and brightness and contrast changes to improve your validation accuracy. Plot the training and validation losses and errors as a function of the number of epochs. Some hints for choosing hyper-parameters:

- Learning rate of 0.1 for the first 40 epochs, then 0.01 for the next 40 epochs and then 0.001 for the final 20 epochs.
- Weight decay of 10^{-3}.
- No need to perform data augmentation, although you can do so if you wish.
- Use SGD with Nesterov's momentum of 0.9 to train the network.

Make sure you save the parameters of the network because we will need them for the next part.

(b) (10 points) We will next compute the backprop gradient of the loss with respect to the input. We know that code of the form

```
...
yh = net.forward(x)
loss = loss.forward(yh, y)

loss.backward()
```

computes the gradient of the loss with respect to the weights, i.e., it computes $\overline{w}$ in our notation. You can get the gradient $\overline{x}$ very easily by adding the following line after loss.backward().

```
dx = x.grad.data.clone()
```

Plot this gradient $\mathrm{d}x$ for a few input images which the network classifies correctly and also for a few images which the network misclassifies. Comment on the similarities or the differences.

Note that each pixel of the RGB image x lies in $[0, 255]$, we will pick $\epsilon = 8$. Pick a particular mini-batch $\{x_1, \ldots, x_b\}$ with $b = 100$. For every image in this mini-batch perform the "5-step signed gradient attack", i.e., perturb that image 5 times using the signed gradient, at each step you feed in the perturbed image from the previous step and perturb it a bit more. Your pseudo-code will look as follows.

```
xs, ys = mini-batch of inputs and targets
for x,y in zip(xs, ys):
```

```python
    for k in range(5):
        # forward propagate x through the network
        # backprop the loss
        dx = ...
        x += eps*sign(dx)
        # record loss on the perturbed image
        ell = loss(x, y)
```

Plot the loss on the perturbed images as a function of the number of steps in the attack averaged across your mini-batch.

(c) (10 points) Compute the accuracy of the network on 1-step perturbed images, i.e., for every image in the validation set, perturb the image using a 1-step attack and check the prediction of the network. How does this accuracy on adversarially perturbed images compare with the accuracy on the clean validation set?

(d) (0 points) The human brain also has a lot of neurons and is likely a high-dimensional classifier. Are humans susceptible to adversarial perturbations? Can you give examples of images that fool the human visual system? Are these "small", i.e., is $\|\epsilon/x\|_\infty$ small for these examples?

Problem 4 (20 points). Training of recurrent neural networks (RNNs) is often difficult because of the vanishing or exploding gradient problem. We will study this in a simple setting without any nonlinearities.

(a) (5 points) If the input to an RNN at the t^{th} timestep is $x^t \in \mathbb{R}^d$ and the hidden vector is $z^t \in \mathbb{R}^p$, the hidden vector z^{t+1} is given by

$$z^{t+1} = \sigma\left(w_x x^t + w_z z^t\right)$$

where $w_x \in \mathbb{R}^{p \times d}$ and $w_z \in \mathbb{R}^{p \times p}$ are weights and σ is a nonlinearity. If $\sigma(z) = z$, i.e., there is no nonlinearity, and $w_x = 0$ then the update boils down to

$$z^{t+1} = w_z z^t.$$

Write down the back-propagation gradient for w_z if the loss function is only a function of the hidden vector at time T , i.e., the loss function is $\ell(z^T)$. Compute the conditions on the weight matrix w_z under which the gradient explodes and vanishes.

(b) (2 points) Argue how the nonlinearity $\sigma(\cdot)$ affects the exploding/vanishing gradients. Which nonlinearities are well-suited for training RNNs?

(c) (3 points) Updates to the weights are computed using SGD as

$$w_z^{\text{new}} = w_z^{\text{old}} - \eta \, \frac{\mathrm{d}\ell}{\mathrm{d}w_z}.$$

We would like to protect the weights w_z^{new} from blowing up even if the gradient $\overline{w_z}$ explodes. Can you think of a way to do this? Similarly, can you modify these updates to handle the vanishing gradient problem?

(d) (5 points) Explain how an LSTM solves the problem of vanishing gradients.

(e) (5 points) Explain how a self-attention layer works for solving the problem of vanishing gradients.