ECBM 4040: Fall 2016 Midterm Solutions

Problem 1

(a)

$$\begin{bmatrix} 4 & 3 \\ 8 & 6 \end{bmatrix} = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} \sqrt{125} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0.8 & 0.6 \\ 0.6 & -0.8 \end{bmatrix}$$

$$A^+ = V\Sigma U^T$$

$$= \begin{bmatrix} 0.8 & 0.6 \\ 0.6 & -0.8 \end{bmatrix} \begin{bmatrix} 1/\sqrt{125} & 0 \\ 0 & 0 \end{bmatrix} \left(\frac{1}{\sqrt{5}} \begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 0.8 & 0.6 \\ 0.6 & -0.8 \end{bmatrix} \begin{bmatrix} 1/25 & 0 \\ 0 & 0 \end{bmatrix} \left(\frac{1}{\sqrt{5}} \begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 0.8 & 0.6 \\ 0.6 & -0.8 \end{bmatrix} \begin{bmatrix} 1/25 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & -1 \end{bmatrix} = \frac{1}{125} \begin{bmatrix} 4 & 8 \\ 3 & 6 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \frac{1}{25} \begin{bmatrix} 4 \\ 3 \end{bmatrix}$$

(b)

$$A = U\Sigma V^T$$

$$A^T A = V\Sigma^2 V^T = V \begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \cdots & \\ & & & \sigma_n^2 \end{bmatrix} V^T$$

We have

$$V = I$$

and U is the matrix whose columns are w_i/σ_i.

(c)

$$Ae_i = \lambda_i e_i$$

$$e_i^T A e_i = \lambda_i$$

Problem 2

This problem directly follows the Regularization lecture slides.

Problem 3

1. Consider a convolutional layer that takes a $32 \times 32 \times 3$ image as input. All 128 filters in the convolutional layer have size 9×9 and there is a 3×3 max-pooling operation after the convolution. What is the size of the output of the max-pooling operation?

A. After convolution, one has a $24 \times 24 \times 128$ dimensional output. Max-pooling reduces the output to $8 \times 8 \times 128$.

2. State two things that are characteristic for stochastic gradient descent methods as opposed to classical (batch) gradient descent.

A. Stochastic gradient descent uses a single random example, or mini-batches. There are many secondary characteristics one could consider; for example, in class, we had specifically discussed convergence related differences.

3. Describe deep learning in two sentences.

A. Deep learning methods utilize multilayer, or architecturally complex, networks with a large number of parameters and/or layers to solve supervised or unsupervised machine learning problems. Usually, the networks utilize backpropagation for training. We have accepted answers as long as they made sense and showed knowledge that implied they understood the fundamental characteristics of deep learning approaches.

4. What is adversarial training?

A. Basically, adversarial training is an augmentation method that allows one to train the neural network on examples along the classification margin, which the network is less "certain" about. A good way to generate adversarial examples is to add an imperceptible amount of noise that shares the sign of the gradient at the input for an input example. Networks trained via this approach often show higher testing accuracy and are resistant to the problem wherein the addition of small, imperceptible perturbations could change the output label. More details can be given, but answers that lined up with this explanation were accepted.

5. List two examples of data augmentation.

A. Translation, rotation, horizontal/vertical flipping, zooming, blurring, spatial transformations, perhaps even some more interesting approaches like adversarial generative samples.

6. Compare the frequency of occurrence of functional minima in low-dimensional and high-dimensional spaces. Provide intuitive explanation.

A. Frequency of occurence of local minima, compared to the frequency of occurence of saddle points, is much smaller in higher dimensional spaces than in low-dimensional spaces.

7. Draw a vectorial figure which illustrates the method of Nesterov momentum.

A. You can check the course materials or the internet.

8. Sketch the behavior of the verification error as the model becomes overfit.

A. You can check the course materials or the internet.

9. Write down the expressions for the softmax activation function and its derivative.

A. You can check the course materials or the internet.

10. What are the advantages of ReLU as compared to the sigmoid activation function?

A. ReLU is computationally cheaper and has been shown to accelerate convergence.

Problem 4

1. While highway networks are capable of attaining a higher accuracy than Fitnets on CIFAR-10 and MNIST datasets, highway networks need slightly more parameters to do so. **(False)**

2. Highway networks successfully utilize a gating mechanism to pass information almost unchanged through many layers. **(True)**

3. While capable of giving impressive results, a significant drawback of generative adversarial networks is that they cannot be trained via backpropagation. **(False)**

4. Convolutional neural networks require max-pooling to give good results. **(False)**

5. A big advantage of CNN over MLP is that it allows for fewer model parameters. **(True)**

6. Weight normalization is inspired by batch normalization, thus it is a probabilistic method that shares batch normalizations property of adding noise to the gradients. **(False)**

7. In weight normalization, no additional memory is required and additional computation is quite negligible. **(True)**

8. The first couple of hidden layers of a CNN captures the high-level content of the image, while lower level content is captured by the latter subsequent layers. **(False)**

9. Increasing the number of layers while keeping the number of parameters the same always increases the maximum testing accuracy of the network. **(False)**

10. There cannot be a convolutional layer after a fully connected layer. **(False)**

* Backpropagation

Problem 5

1 First we calculate $\frac{\partial \mathcal{L}}{\partial h^2}$.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial h^2} &= \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial h_i^3} \frac{\partial h_i^3}{\partial h^2} \\ &= \sum_{i=1}^{N} \frac{\partial}{\partial h_i^3} \frac{1}{k} \sum_{j=1}^{k} \sum_{l=1}^{N} (t_l^j - h_l^3)^2 \frac{\partial}{\partial h^2} \sigma(h^2 W_i^3) \\ &= -\frac{2}{k} \sum_{i=1}^{N} \sum_{j=1}^{k} (t_i^j - h_i^{3^j}) \sigma(h^2 W_i^3)(1 - \sigma(h^2 W_i^3)) W_i^{3T}\end{aligned}$$

where W_i^3 is the i^{th} column vector of W^3.

Let $c_i = -\frac{2}{k} \sum_{j=1}^{k} (t_i^j - h_i^3) \sigma(h^2 W_i^3)(1 - \sigma(h^2 W_i^3))$ be a scalar, then we have

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial h^2} &= \sum_{i=1}^{N} c_i W_i^{3T} \\ &= C W^{3T}\end{aligned}$$

where $C = [c_1, c_2, ..., c_N]$.

Next we calculate $\frac{\partial \mathcal{L}}{\partial h^1}$.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial h^1} &= \sum_{i=1}^{M} \frac{\partial \mathcal{L}}{\partial h_i^2} \frac{\partial h_i^2}{\partial h^1} \\ &= \sum_{i=1}^{M} \frac{\partial \mathcal{L}}{\partial h_i^2} \frac{\partial}{\partial h^1} \sigma(h^1 W_i^2) \\ &= \sum_{i=1}^{M} [CW^{3T}]_i \sigma(h^1 W_i^2)(1 - \sigma(h^1 W_i^2)) W_i^{2T}\end{aligned}$$

where W_i^2 is the i^{th} column vector of W^2.

Let $d_i = [CW^{3T}]_i \sigma(h^1 W_i^2)(1 - \sigma(h^1 W_i^2))$ be a scalar, then we have

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial h^1} &= \sum_{i=1}^{M} d_i W_i^{2T} \\ &= DW^{2T}\end{aligned}$$

where $D = [d_1, d_2, ..., d_M]$.

2 Since now the number of hidden units in the first hidden layer is the same as the output layer, the residual connection between these two layers is an identity mapping (i.e. $g(x) = x$).

Let the new output of the network be $z = h1 + h3$ and the new objective be $\hat{\mathcal{L}}$. We have

$$\begin{aligned}\frac{\partial \hat{\mathcal{L}}}{\partial h^1} &= \frac{\partial \hat{\mathcal{L}}}{\partial z} \frac{\partial z}{\partial h^1} \\ &= \frac{\partial \hat{\mathcal{L}}}{\partial z} (I + \frac{\partial h^3}{\partial h^1}) \\ &= \frac{\partial \hat{\mathcal{L}}}{\partial z} (I + \frac{\partial h^3}{\partial h^2} \frac{\partial h^2}{\partial h^1})\end{aligned}$$

and we have

$$\begin{aligned}\frac{\partial \hat{\mathcal{L}}}{\partial z_i} &= \frac{\partial}{\partial z_i}\frac{1}{k}\sum_{i=1}^{N}\sum_{j=1}^{k}(t_i^j - z_i)^2 \\ &= -\frac{2}{k}\sum_{i=1}^{N}\sum_{j=1}^{k}(t_i^j - z_i)\end{aligned}$$

$$\frac{\partial \hat{\mathcal{L}}}{\partial z} = [\frac{\partial \hat{\mathcal{L}}}{\partial z_1}, \frac{\partial \hat{\mathcal{L}}}{\partial z_2}, ..., \frac{\partial \hat{\mathcal{L}}}{\partial z_N}]$$

The second term in the parentheses is the same as what we use for the network without residual, hence we have

$$\frac{\partial h_i^3}{\partial h^2} = \sigma(h^2 W_i^3)(1 - \sigma(h^2 W_i^3))W_i^{3T}$$

$$\frac{\partial h_i^2}{\partial h^1} = \sigma(h^1 W_i^2)(1 - \sigma(h^1 W_i^2))W_i^{2T}$$

$$\frac{\partial h^3}{\partial h^2} = \begin{bmatrix} \frac{\partial h_1^3}{\partial h^2} \\ \frac{\partial h_2^3}{\partial h^2} \\ ... \\ \frac{\partial h_N^3}{\partial h^2} \end{bmatrix}$$

$$\frac{\partial h^2}{\partial h^1} = \begin{bmatrix} \frac{\partial h_1^2}{\partial h^1} \\ \frac{\partial h_2^2}{\partial h^1} \\ ... \\ \frac{\partial h_M^2}{\partial h^1} \end{bmatrix}$$

3 There are two possible reasons for gradient vanishing. First, when sigmoid function saturates at either tails, the gradient will become almost zero, which may cause gradient vanishing within even a single layer. Second, since the derivative of sigmoid function satisfies $\sigma()(1-\sigma()) \leq \frac{1}{4}$, stacking multiple sigmoid layers (even they are not saturated) will cause the gradient to decrease rapidly.

From the results above, we can find that by adding the residual connection, the new gradient for h_1 has two parts: one part directly propagates information from the output (without passing through intermediate layers), and another part pass through intermediate layers (which is the original gradient). Even if there's gradient vanishing problem in the second part, since the first part does not involve any activation functions, it ensures that the total gradient will not easily vanish towards zero.