ECE 498 ICC: IoT and Cognitive Computing

Spring 2020, Homework 1

Due in class on March **25, 2020**

Question 1: Cloud vs. Edge (10 pts)

1. TensorFlow uses a dataflow graph (`tf.Graph`) to represent the computation. Explain the benefit of using dataflow graph to represent the computation in detail. [**2 pts**]

2. List 2 advantages/disadvantages of using Cloud Computing vs Edge computing for an IoT system. [**4 pts**]

3. For the following applications which processing (Cloud Computing or Edge computing) is more suitable? Briefly explain why. [**4 pts**]

   - Health data collected by an Apple watch/Fitbit to track your daily activities
   - Temperature sensors placed inside a refrigerated storage container to regulate the temperature during the shipping process
   - License plate readers at toll plazas
   - Medical wearable that detects when you fall.

Question 2: Tensorflow (10 pts)

1. $step(x)$ is defined as $step(x) = 0$ for $x < 0$ and $step(x) = 1$ for $x \geq 1$. Imagine that you are doing gradient descent with Tensorflow using a loss function $step(1-x)x^2 + step(x-1)[-(x-1)^2+1]$. x is a `tf.Variable` of type `tf.float32`. What happens:

   - when x is initialized to -2? [**3 pts**]
   - when x is initialized to $+3$? [**3 pts**]

   Assume that learning rate is very small. Justify your answer.

2. Assume that we are trying to solve an equation $x^2 + x = -1$ for scalar x, using gradient descent and TensorFlow. We setup x as `tf.Variable` of type `tf.float32`, and proceed to find the solution following the methodology in lab 1, part 3. Will this method be able to find the correct solution for this equation (Yes/No)? Prove or justify your answer. [**4 pts**]

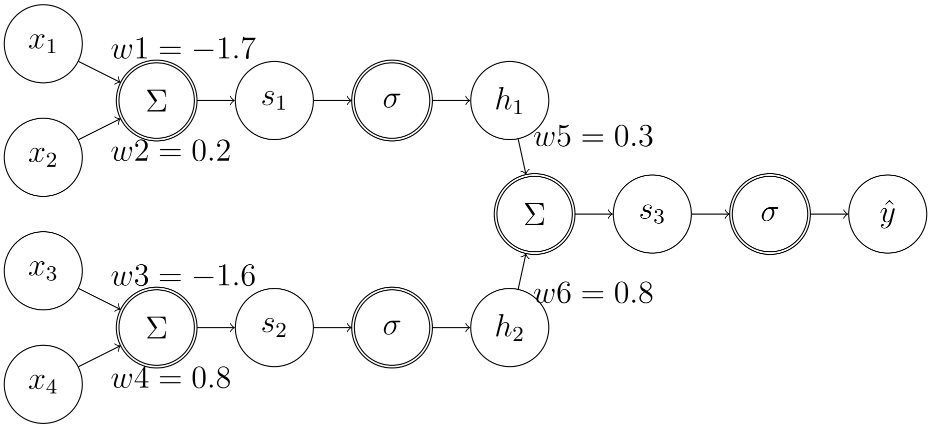

Figure 1: A neural network pipeline

Question 3: Backpropagation (10 pts)

Consider the neural network in Fig 1. Single-circled nodes denote variables (e.g. x_1 is an input variable, h_1 is an intermediate variable, $\hat{y}$ is an output variable), and double-circled nodes denote functions (e.g. Σ takes the sum of its inputs, and σ denotes the logistic function $\sigma(x) = \frac{1}{1+e^{-x}}$.
Suppose we have an MSE loss $L(y, \hat{y}) = \|y - \hat{y}\|_2^2$, We are given a data point $(x1, x2, x3, x4) = (0.3, 1.4, 0.9, -0.6)$ with true label 0.37. (Hint: : the gradient of an MSE loss function is $2\|y - \hat{y}\|$.)

1. Use the backpropagation algorithm to compute the partial derivative$\frac{\partial L}{\partial w_1}$, $\frac{\partial L}{\partial w_2}, \frac{\partial L}{\partial w_3}, \frac{\partial L}{\partial w_4}, \frac{\partial L}{\partial w_5}, \frac{\partial L}{\partial w_6}$. **[6 pts]**

2. Explain what are vanishing gradients and exploding gradients. You can watch this video before answering the question. (link) [**4 pts**]

Question 4: Neural Network (10 pts)

1. Consider two classification problems:
   Problem A: Items belonging to label '0' : {(0,0), (0,1)}, Items belonging to label '1': {(1,0), (1,1)}.
   Problem B: Items belonging to label '0' : {(0,0), (1,1)}, Items belonging to label '1': {(1,0), (0,1)}.
   Would linear classifiers be able to learn the patterns in these two classification problems? Justify your answer. [**2 pts**]

2. For a classification problem with 16 feature inputs and 16 classes, compare the computational complexity of:

(1) a deep network with eight hidden layers with 32 nodes each, and (2) a shallower network that has two hidden layers with 128 nodes each (all layers are fully-connected). What are the memory requirements of these two configurations? Assume ReLU activation is used in all layers except the output layer. Which one would you prefer to deploy on an edge device? [**2 pts**]

3. You have trained a DNN model for an image classification application for static images, and you are deploying the model in the field. However, you find that the camera in the field was shifted in space compared to the camera used for capturing training images due to some logistic issues. What happens to the classification accuracy if you (1) used a Multilayer Perceptron as your model or (2) a Convolutional Neural Network as your model. Justify your answer. [**2 pts**]
4. You are building a classifier and are comparing different activation functions. What are the trade-offs among the following three? (1) ReLU; (2) tanh; (3) unit step function? [**2 pts**]
5. Are there any disadvantages in treating an N-class classification problem as a regression in N dimensions? Please explain. [**2 pts**]

Question 5: Stassen's Algorithm (15 pts)

In machine learning applications, one of the most basic operations is matrix multiplication. Many so-called "AI chips" have large, dedicated on-chip regions for accelerating matrix multiplications. While Professor Hwu will teach us about parallel algorithms of matrix multiplications on GPU, there are fast algorithms that mathematicians had developed decades ago to out-perform naive $O(n^3)$ complexity. One of the earliest and most famous is Strassen's algorithm.
For matrices A, B, and $C \in R^{2^n \times 2^n}$, such that $C = AB$, we partition A, B, and C as follows:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

Strassen's algorithm defines new matrices M_k as follows:

$$\begin{aligned}
M_1 &:= (A_{11} + A_{22})(B_{11} + B_{22}) \\
M_2 &:= (A_{21} + A_{22})B_{11} \\
M_3 &:= A_{11}(B_{12} - B_{22}) \\
M_4 &:= A_{22}(B_{21} - B_{11}) \\
M_5 &:= (A_{11} + A_{12})B_{22} \\
M_6 &:= (A_{21} - A_{11})(B_{11} + B_{12}) \\
M_7 &:= (A_{12} - A_{22})(B_{21} + B_{22})
\end{aligned}$$

and expresses C_{ij} in terms of M_k:

$$\begin{aligned}
C_{11} &:= M_1 + M_4 - M_5 + M_7 \\
C_{12} &:= M_3 + M_5 \\
C_{21} &:= M_2 + M_4 \\
C_{22} &:= M_1 - M_2 + M_3 + M_6
\end{aligned}$$

1. Express the outcome matrix C_{ij} in terms of A_{ij} and B_{ij} as in the naive algorithm. [**2 pts**]

2. We now have A, B, and $C \in R^{4\times4}$; in other words, the matrices all have a dimension of 4×4. How many multiplications and additions are needed in the Strassen's algorithm? What about the naive algorithm? [**4 pts**]

3. Based on what you have observed so far, compare Strassen's algorithm and the naive algorithm. [**3 pts**]

4. Now we come back to the more basic case where A, B, and $C \in R^{2\times2}$. Let's try an alternative way to express the Strassen's algorithm. First, we reshape the input matrices A, B into vectors:

$$A = \begin{bmatrix} A_{11} \\ A_{12} \\ A_{21} \\ A_{22} \end{bmatrix}, B = \begin{bmatrix} B_{11} \\ B_{12} \\ B_{21} \\ B_{22} \end{bmatrix}$$

Then, we are able to express the outcome C as:

$$C = F \otimes ((G^T \otimes A) * (H^T \otimes B))$$

where $\otimes$ stands for matrix product, $*$ stands for element-wise product, and F, G, and $H \in R^{4\times7}$. We call F decoding matrix and G and H encoding matrices. Find F, G, H, and describe any observations on those matrices. [**6 pts**]
Hint: Look closely at the the original equations. $M_1 = (A_{11} + A_{22})(B_{11} + B_{22})$ tells you that it is the *element-wise product* of $(A_{11} + A_{22})$ and $(B_{11} + B_{22})$. Also, you notice that only A_{11} and A_{22} appear in the left side of the product. Therefore, the first column of G is [1 0 0 1].

Question 6: Curse of dimensionality and PCA (20 pts)

1. **Curse of dimensionality**[**2 pts**]
Briefly explain the curse of dimensionality.

2. **Statistical interpretation of PCA[2 pts]**
   Briefly explain the statistical interpretation of PCA.

3. **Linear algebra interpretation of PCA[4 pts]**
   PCA can also be interpreted as: finding a subspace, represented with a set of orthonormal basis $v_1, ..., v_d$, such that the Frobenius norm of the difference between the original input and the projection onto the subspace[1] is minimized.

   Given the description above, derive the objective function in terms of X, V, and Frobenius norm, where $V \in \mathbb{R}^{d \times k}$ whose column vectors are the orthonormal basis $v_1, ..., v_k$, and $X \in \mathbb{R}^{n \times d}$ whose rows are the input vectors $x_1, ..., x_n$. Notice that $V^T V = I$.

   Hint: What does matrix V do? How to derive the projection matrix onto a subspace given a set of basis? Does orthonormal basis make this projection it simpler?

4. **Linear algebra derivation of PCA**

   - **[6 pts]** Now that we have derived the loss function of PCA. Show that minimizing the objective function can be written as maximizing $||XV||_F^2$ You might need the following properties of Frobenius norm and trace [2]:
     - $||A||_F^2 = tr(A^T A)$
     - $Ctr(A) = tr(CA)$ (C is a constant)
     - $tr(A + B) = tr(A) + tr(B)$
     - $tr(A^T A) = tr(AA^T)$
   - **[6 pts]** Now, we have $||XV||_F^2 = \text{tr}(V^T X^T X V)$. Assuming that $X^T X$ has an eigendecomposition, that is, $X^T X = QDQ^T$. We can then construct the vectors of V with the orthonormal eigenvectors of the matrix $X^T X$. That is, we can have $V = QZ$, where $Z \in \mathbb{R}^{d \times k}$ and $Z^T Z = I$.
     - First, state the reason why the eigendecomposition of $X^T X$ exists and the eigenvectors can form an orthonormal basis of the d-dimensional space.
     - Then, show that the objective function is can be transformed to $\text{tr}(Z^T D Z)$.

   Notice that, the maximum value of $\text{tr}(Z^T D Z)$ is the sum of the largest K eigenvalues of $X^T X$. (you don't need to prove this step)

[1] https://ocw.mit.edu/courses/mathematics/18-06sc-linear-algebra-fall-2011/least-squares-determinants-and-eigenvalues/projections-onto-subspaces/MIT18_06SCF11_Ses2.2sum.pdf

[2] https://en.wikipedia.org/wiki/Trace_(linear_algebra)

Question 7: IoT system design (25 pts)

You are designing a security system. Your security system uses deep learning to recognize faces and detect unauthorized accesses, and a network of cameras, spread across 3 floors, each of size 3000 ft by 2000 ft. You are given that:

- Each camera sensor costs \$10, has viewing radius of 100 ft, and generates 10 frames per second. Each frame is ∼3KB.
- You have a bandwidth of 10MB/s to the internet.
- You may choose from the following four devices. Their cost and the performance of the deep learning algorithm is given:

  Raspberry Pi: \$55, Can run the deep network at 5FPS (frames per second)

  Xilinx PYNQ Z1: \$200, 20FPS

  Nvidia Jetson TX1: \$350, 35FPS

  Xilinx Cloud FPGA: \$2000, 350FPS
- To detect an intrusion, you must look at 2 seconds of footage.

You must detect intrusions as soon as possible. How will you design such a system with minimal cost?

1. Provide a brief description of your system. You may provide a simple figure or block diagram. [**5 pts**]
2. How many cameras will you use? [**5 pts**]
3. What processing engine will you choose and how many cameras will be coupled to each? [**5 pts**]
4. What is your final cost? [**5 pts**]
5. What is the average worst-case latency (in milliseconds) for detecting an intrusion? [**5 pts**]

NOTE: You are allowed to make any reasonable assumptions and approximation about your solution. The points are not necessarily to get the exact solution, but to see how you would approach a problem at that scale using some back-of-the-envelope computation.