Data C182 Designing, Visualizing & Understanding DNN
Fall 2024 Eric Kim, Naveen Ashish

Discussion 8

This discussion covers select questions from the midterm exam.

1. Initialization

In class, we discussed how, when initializing neural network weights, we tend to choose them randomly from e.g., a Gaussian distribution of a certain variance and mean. Why is this the case? Let's walk through some alternatives.

For the sake of simplicity, assume that your neural network consists only of consecutive affine layers and ReLU non-linearities, and that there is at least one such non-linearity. All hidden layers can have an arbitrary number of elements ≥ 1. You can also assume batch sizes of 1 for training (though your answers should hold for arbitrary batch size). Finally, assume that there is some loss function $L(y)$ that takes in the output of your neural network y, and that loss is used to train your neural network with standard gradient descent (i.e., no momentum, gradient clipping, RMSProp, etc).

For this problem, use this small two affine layer neural network, where x is a two-element column vector:

$$\text{out} = W_2\left[\text{ReLU}\left(W_1 x + b_1\right)\right] + b_2$$

Where $W_1 \in \mathbb{R}^{2\times 2}$, $b_1 \in \mathbb{R}^2$, $W_2 \in \mathbb{R}^{2\times 2}$, $b_2 \in \mathbb{R}^2$.

0.1 Part A.i

Suppose that all weights W_1, W_2 and biases b_1, b_2 for all layers are initialized to zero. The input $x = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

What is the output of the neural network?

0.2 Part A.ii

Suppose that the final gradient $\frac{dL}{d_{\text{out}}} = \begin{bmatrix} 0.1 \\ 0.2 \end{bmatrix}$. What are the gradients over final Linear layer's weights and biases $\frac{dL}{dW_2}$ and $\frac{dL}{db_2}$?

0.3 Part A.iii

What are the gradients over first Linear layer's weights and biases $\frac{dL}{dW_1}$ and $\frac{dL}{db_1}$?

0.4 Part A.iv

Describe why would this be a problem? (Hint: think about the next forward and backward pass. Your answer should contain little-to-no complicated math and should be at most a few sentences.)

0.5 B

(a) Now, suppose that for each weight matrix and bias vector, all elements are set to the same constant

$$W_1 = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}, \quad b_1 = \begin{bmatrix} -2 \\ -2 \end{bmatrix}, \quad W_2 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \text{and} \quad b_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

The input $x = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. What is the output of the neural network?

(b) Suppose that the final gradient $\frac{\partial L}{\partial d_{\text{out}}} = \begin{bmatrix} 0.1 \\ 0.2 \end{bmatrix}$. What are the gradients over the final Linear layer's weights and biases $\frac{\partial L}{\partial W_2}$ and $\frac{\partial L}{\partial b_2}$?

(c) What are the gradients over the first Linear layer's weights and biases $\frac{\partial L}{\partial W_1}$ and $\frac{\partial L}{\partial b_1}$?

(d) Describe why this would be a problem? Hint: think about the next forward and backward pass. How do the results of the previous question impact the expressiveness of the model? Your answer should contain little-to-no complicated math and should be at most a few sentences.

2. Multiple Choice

0.6 Q1

A model for classifying different objects is getting a high training set error. Which of the following is the most likely way to improve the classifier?

- ◯ **A**: Use more training data.
- ◯ **B**: Increase the regularization being used.
- ◯ **C**: Use a bigger network.
- ◯ **D**: Use a smaller network.

0.7 Q2

How many model parameters are in a Convolution2D layer that uses a 4x4 filter with 5 output channels and a bias, and takes as input a three-channel color RGB image with height=32 pixels, width=32 pixels?

- ◯ **A**: 16
- ◯ **B**: 245
- ◯ **C**: 80
- ◯ **D**: 240
- ◯ **E**: 21
- ◯ **F**: 85

0.8 Q3

Which of the following can lead to vanishing gradients?

- ◯ **A**: Sigmoid activations.
- ◯ **B**: Very deep neural network with skip connections.
- ◯ **C**: Batch normalization layers.
- ◯ **D**: Leaky ReLU activations.

0.9 Q4

What is the primary motivation for adding masks in "masked self-attention" in the Transformer decoder?

- ◯ **A**: To better-condition the intermediate activation values to avoid the vanishing/exploding gradient problem.
- ◯ **B**: To avoid the decoder from "cheating" and using information from future token positions.
- ◯ **C**: To avoid the decoder from "cheating" and using information from other batch samples.
- ◯ **D**: To improve representation power by adding more model parameters.

0.10 Q5

In the Transformer self-attention block, when computing the attention weights $A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$, what is the primary motivation for dividing by $\sqrt{d_k}$, where d_k is the embedding dimensionality?

- ◯ **A**: To better-condition the intermediate activation values to avoid the vanishing/exploding gradient problem.
- ◯ **B**: To avoid dividing by 0.
- ◯ **C**: To add additional regularization to avoid overfitting.
- ◯ **D**: To improve representation power by adding more model parameters.

