705.641.81: Natural Language Processing Self-Supervised Models

Homework 3: Building Your Neural Network!

For homework deadline and collaboration policy, please see our Canvas page.  
Name: \_\_Madihah Shaik\_\_\_\_\_\_\_\_\_\_\_   
Collaborators, if any: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
Sources used for your homework, if any: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

This assignment is focusing on understanding the fundamental properties of neural networks and their training.

**Homework goals:** After completing this homework, you should be comfortable with:

* thinking about neural networks
* key implementation details of NNs, particularly in PyTorch,
* explaining and deriving Backpropagation,
* debugging your neural network in case it faces any failures.

# 

# Concepts, intuitions and big picture

1. Suppose you have built a neural network. You decide to initialize the weights and biases to be zero. Which of the following statements are True? (Check all that apply)  
    Each neuron in the first hidden layer will perform the same computation. So even after multiple iterations of gradient descent each neuron will be computing the same thing as other neurons in the same layer.  
    Each neuron in the first hidden layer will perform the same computation in the first iteration. But after one iteration of gradient descent they will learn to compute different things because we have “broken symmetry”.  
    Each neuron in the first hidden layer will compute the same thing, but neurons in different layers will compute different things, thus we have accomplished “symmetry breaking” as described in lecture.  
    The first hidden layer’s neurons will perform different computations from each other even in the first iteration; their parameters will thus keep evolving in their own way.
2. Vectorization allows you to compute forward propagation in an -layer neural network without an explicit for-loop (or any other explicit iterative loop) over the layers . True/False?  
    True  
    False
3. The tanh activation usually works better than sigmoid activation function for hidden units because the mean of its output is closer to zero, and so it centers the data better for the next layer. True/False?  
    True  
    False
4. Which of the following techniques does NOT prevent a model from overfitting?  
    Data augmentation Dropout Early stopping None of the above
5. Why should dropout be applied during training? Why should dropout NOT be applied during evaluation?

Dropout layers set the output of a few random neurons to 0. Dropout should be applied during training to drop the computation of a few neurons so that the model does not memorize the training data. This helps prevent overfitting. However, when evaluating, we want to use all of the neurons and what they have learned in the training process.

1. Explain why initializing the parameters of a neural net with a constant is a bad idea.  
   If the parameters are initialized with a constant, the computation at each neuron will be the same. Then, the neuron weights will not update via gradient descent, and the neural network will not learn.
2. You design a fully connected neural network architecture where all activations are sigmoids. You initialize the weights with large positive numbers. Is this a good idea? Explain your answer.  
   This is not a good idea. Exploding gradients will occur with weights of large positive numbers. During backpropagation, the gradients will be multiplied and continue to get larger. As the gradients grow exponentially, the weight updates will also be large. This causes the model’s learning to be unstable and overshoot the optimal solution.
3. Explain what is the importance of “residual connections”.  
   Residual connections are added to neural networks to apply simple functions when the network gets too complicated. There is a direct, linear connection from input to output. A network with residual connections is easier to optimize and is more stable.
4. What is cached (“memoized”) in the implementation of forward propagation and backward propagation?  
    Variables computed during forward propagation are cached and passed on to the corresponding backward propagation step to compute derivatives.  
    Caching is used to keep track of the hyperparameters that we are searching over, to speed up computation.  
    Caching is used to pass variables computed during backward propagation to the corresponding forward propagation step. It contains useful values for forward propagation to compute activations.
5. Which of the following statements is true?  
    The deeper layers of a neural network are typically computing more complex features of the input than the earlier layers.  
    The earlier layers of a neural network are typically computing more complex features of the input than the deeper layers.

# Revisiting Jacobians

Recall that Jacobians are generalizations of multi-variate derivatives and are extremely useful in denoting the gradient computations in computation graph and Backpropagation. A potentially confusing aspect of using Jacobains is their dimensions and so, here we’re going focus on understanding Jacobian dimensions.

#### Recap:

Let’s first recap the formal definition of Jacobian. Suppose is a function takes a point as input and produces the vector as output. Then the Jacobian matrix of is defined to be an matrix, denoted by , whose th entry is , or:

#### Examples:

The shape of a Jacobian is an important notion to note. A Jacobian can be a vector, a matrix, or a tensor of arbitrary ranks. Consider the following special cases:

* If is a scalar and is a column vector, the Jacobian of with respect to **w** is a row vector with dimensions.
* If is a column vector and is a column vector, the Jacobian of with respect to , or is a matrix.
* Suppose and . Then the Jacobian is a tensor of shape . More broadly, the shape of the Jacobian is determined as (shape of the output)(shape of the input).

#### Problem setup:

Suppose we have:

* , an matrix, correspond to the rows of
* , a matrix
* , a matrix and , a vector

For the following items, compute (1) the shape of each Jacobian, and (2) an expression for each Jacobian:

1. (constant)  
   All entries in the matrix would be 0 (derivative of a constant). The size would be **1xd.**
2. (squared L2-norm)  
   The shape of the Jacobian matrix would be 1xd.
3. (vector dot product)  
   , and the shape 1xd
4. (matrix-vector product)  
   , and the shape is nxd
5. (vector identity function)  
   , and the shape is dx1.
6. (element-wise power)  
   , and the shape is dx1
7. **Extra Credit:** (matrix multiplication)

# Activations Per Layer, Keeps Linearity Away!

Based on the content we saw at the class lectures, answer the following:

1. Why are activation functions used in neural networks?  
   The output of a neuron is fed to the activation function, typically to ensure the output is in some range, and to introduce non-linearity to the network.
2. Write down the formula for three common action functions (sigmoid, ReLU, Tanh) and their derivatives (assume scalar input/output). Plot these activation functions and their derivatives on .  
     
   A close-up of math equations

   AI-generated content may be incorrect. A graph on a piece of paper

   AI-generated content may be incorrect.
3. What is the “vanishing gradient” problem? (respond in no more than 3 sentences) Which activation functions are subject to this issue and why? (respond in no more than 3 sentences).  
   When computing gradient with respect to a loss function, the gradients may be very small. As they are multiplied, the gradients continue to get smaller. These are vanishing gradients. Some of the activation functions that are prone to this issue are the sigmoid and the tanh functions, because the derivatives become close to 0.
4. Why zero-centered activation functions impact the results of Backprop?  
   Zero-centered activation functions allow the network to be more balanced. The neurons start with similar activations, with a mean around 0. This helps prevent the gradients from getting too large or too small during backpropagation.
5. Remember the Softmax function and how it extends sigmoid to multiple dimensions? Let’s compute the derivative of Softmax for each dimension. Prove that:

* where is the Kronecker delta function.[[1]](#footnote-1)  
  A math equations on a piece of paper

  AI-generated content may be incorrect.

1. Use the above point to prove that the Jacobian of the Softmax function is the following:

where turns a vector into a diagonal matrix. Also, note that .  
When i=j, the derivative is , which is the diagonal entries of a diagonal matrix= diag *)* . When I does not equal j, the derivative is

# Simulating XOR

1. Can a single-layer network simulate (represent) an XOR function on ?

* Explain your reasoning using the following single-layer network definition:
* For x=(1,0), y should be 1. Then ReLU(w1+b) should be greater than 0, so w1+b>0. For x=(0,1), y should be 1. Then ReLU(w2+b) should be greater than 0, so w2+b>0. This means w1+w2+b must be greater than 0. However, for x=(1,1), y should be 0, ReLU(w1+w2+b)=0, w1+w2+b<= 0. This contradicts w1+w2+b>0, and no such variables exist.

1. Repeat (1) with a two-layer network:

* Note that this model has an additional layer compared to the earlier question: an input layer , a hidden layer with ReLU activation functions that are applied component-wise, and a linear output layer, resulting in scalar prediction . Provide a set of weights and and biases and such that this model can accurately model the XOR problem.  
  If biases are 0, and , then the hidden layer is 1 when exactly one x is 1. Then y\_hat is 1 for weights of 1 and bias of 0 when only one x is active, which is the XOR function.

1. Consider the same network as above (with ReLU activations for the hidden layer), with an arbitrary differentiable loss function which takes as input and , our prediction and ground truth labels, respectively. Suppose all weights and biases are initialized to zero. Show that a model trained using standard gradient descent will not learn the XOR function given this initialization.  
   When all weights and biases are set to 0, then for every input x, parameters are 0. The gradient at this point may be nonzero, but the gradient with respect to the output weights will be 0. At each step, the weights will stay 0, and the hidden layer’s activation will remain 0 for every input. The output will be a constant function, which cannot represent XOR.
2. **Extra Credit:** Now let’s consider a more general case than the previous question: we have the same network with an arbitrary hidden layer activation function:

* Show that if the initial weights are any uniform constant, then gradient descent will not learn the XOR function from this initialization.

*A computation graph, so elegantly designed  
With nodes and edges, so easily combined  
It starts with inputs, a simple array  
And ends with outputs, in a computationally fair way  
  
Each node performs, an operation with care  
And passes its results, to those waiting to share  
The edges connect, each node with its peers  
And flow of information, they smoothly steer  
  
It’s used to calculate, complex models so grand  
And trains neural networks, with ease at hand  
Backpropagation, it enables with grace  
Making deep learning, a beautiful race*

*–ChatGPT Feb 3 2023*

# Neural Nets and Backpropagation

Draw the computation graph for . Each node in the graph should correspond to only one simple operation (addition, multiplication, exponentiation, etc.). Then we will follow the forward and backward propagation described in class to estimate the value of and partial derivatives at . For each step, show your work.

1. Draw the computation graph for . The graph should have three input nodes for and one output node . Label each intermediate node .  
   A diagram of a diagram

   AI-generated content may be incorrect.
2. Run the forward propagation and evaluate and () at .

A screenshot of a notebook

AI-generated content may be incorrect.

1. Run the backward propagation and give partial derivatives for each intermediate operation, i.e., , , and . Evaluate the partial derivatives at .  
   A math equations and formulas on a piece of paper

   AI-generated content may be incorrect.
2. Aggregate the results in (c) and evaluate the partial derivatives with chain rule. Show your work.

A math equations on a piece of paper

AI-generated content may be incorrect.

# Programming

In this programming homework, we will

* implement MLP-based classifiers for the sentiment classification task of homework 1.

#### Skeleton Code and Structure:

The code base for this homework can be found at [this GitHub repo](https://github.com/JHU-CLSP/CS-601-471-671-Sp24-Public/) under the hw3 directory. Your task is to fill in the missing parts in the skeleton code, following the requirements, guidance, and tips provided in this pdf and the comments in the corresponding .py files. The code base has the following structure:

* mlp.py reuse the sentiment classifier on movie reviews you implemented in homework 1, with additional requirements to implement MLP-based classifier architectures and forward pass .
* main.py provides the entry point to run your implementations mlp.py
* hw3.md provides instructions on how to setup the environment and run each part of the homework in main.py

**TODOs** — Your tasks include 1) generate plots and/or write short answers based on the results of running the code; 2) fill in the blanks in the skeleton to complete the code. We will explicitly mark these plotting, written answer, and filling-in-the-blank tasks as **TODOs** in the following descriptions, as well as a **# TODO** at the corresponding blank in the code.  
**TODOs** (Copy from your HW1). We are reusing most of the model.py from homework 1 as the starting point for the mlp.py - you will see in the skeleton that they look very similar. Moreover, in order to make the skeleton complete, for all the **# TODO (Copy from your HW1)**, please fill in the blank below them by copying and pasting the corresponding implementations you wrote for homework 1 (i.e. the corresponding **# TODO** in homework 1.)

#### Submission:

Your submission should contain two parts: 1) plots and short answers under the corresponding questions below; and 2) your completion of the skeleton code base, in a .zip file

## MLP-based Sentiment Classifier

In both homework 1 & 2, our implementation of the SentimentClassifer is essentially a single-layer feedforward neural network that maps input features directly to 2-dimensional output logits. In this part of the programming homework, we will expand the architecture of our classifier to multi-layer perceptron (MLP).

### Reuse Your HW1 Implementation

**TODOs** (Copy from your HW1): for all the **# TODO (Copy from your HW1)** in mlp.py, please fill in the blank below them by copying and pasting the corresponding implementations you wrote for homework 1 (i.e. the corresponding **# TODO** in the model.py in homework 1).

### Build MLPs

Remember from the lecture that MLP is a multi-layer feedforward network with perceptrons as its nodes. A perceptron consists of non-linear activation of the affine (linear) transformation of inputs.  
  
**TODOs**: Complete the \_\_init\_\_ and forward function of the SentimentClassifier class in mlp.py to build MLP classifiers that supports custom specification of architecture (i.e. number and dimension of hidden layers)  
**Hint**: check the comments in the code for specific requirements about input, output, and implementation. Also, check out the document of [nn.ModuleList](https://pytorch.org/docs/stable/generated/torch.nn.ModuleList.html) about how to define and implement forward pass of MLPs as a stack of layers.

### Train and Evaluate MLPs

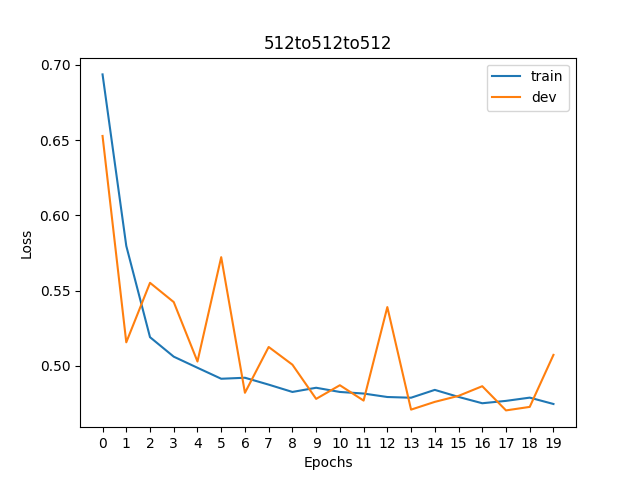
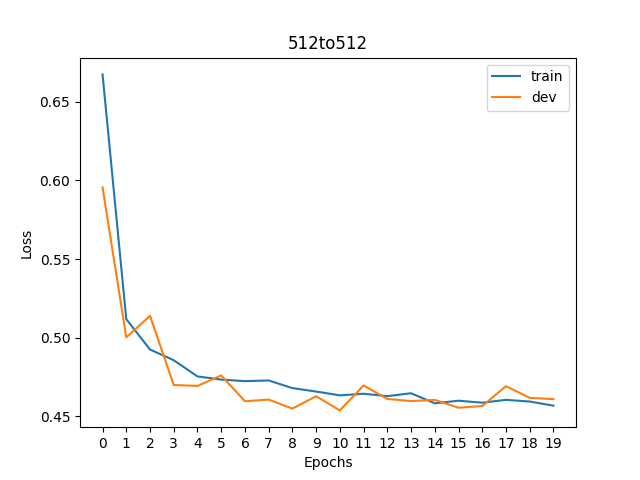
We provide in main.py several MLP configurations and corresponding recipes for training them.  
  
**TODOs** Once you finished [6.1.2](#subsubsec: build mlps), you can run load\_data\_mlp and explore\_mlp\_structures to train and evaluate these MLPs and paste two sets of plots here:

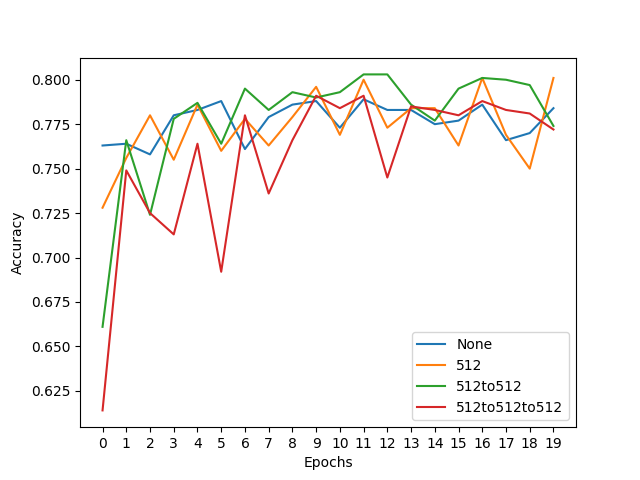
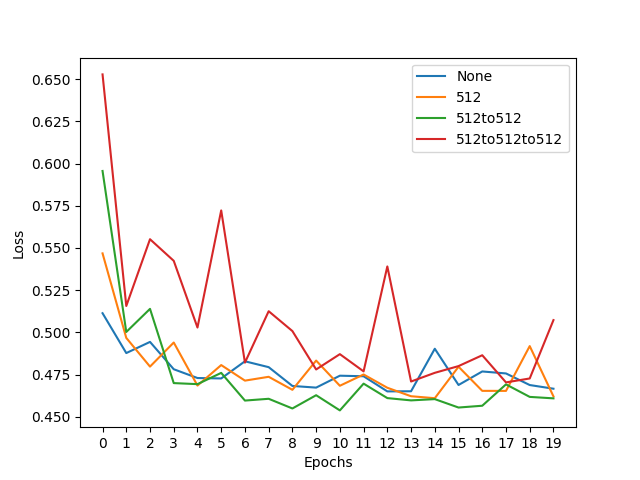
* 4 plots of train & dev loss for each MLP configuration

A graph with numbers and lines

AI-generated content may be incorrect.

A graph with numbers and lines

AI-generated content may be incorrect.

* 2 plots of dev losses and accuracies across MLP configurations

and describe in 2-3 sentences your findings.  
**Hint**: what are the trends of train & dev loss and are they consistent across different configurations? Are deeper models always better? Why?  
The trends differ across different configurations. As more hidden layers are introduced, the network’s accuracy fails to stabilize on the validation data. This suggests there is overfitting to the training data in these denser models. Based on these graphs, I would use the model with 2 hidden layers of 512 nodes (green in the chart).

### Embrace Non-linearity: The Activation Functions

Remember we have learned why adding non-linearity is useful in neural nets and gotten familiar with several non-linear activation functions both in the class and [3](#sec: activation). Now it is time to try them out in our MLPs!  
**Note: for the following TODO and the TODO in** [**6.1.5**](#subsubsec: lr)**, we fix the MLP structure to be with a single 512-dimension hidden layer, as specified in the code. You only need to run experiments on this architecture**.  
  
**TODOs**: Read and complete the missing lines of the two following functions:

* \_\_init\_\_ function of the SentimentClassifier class: define different activation functions given the input activation type.  
  **Hint**: we have provided you with a demonstration of defining the Sigmoid activation, you can search for the other nn.<activation> in PyTorch documentation.
* explore\_mlp\_activations in main.py: iterate over the activation options, define the corresponding training configurations, train and evaluate the model, and visualize the results. Note: you only need to generate the plots of dev loss and dev acc across different configurations, by calling visualize\_configs, you **do not** need to plot the train-dev loss curves for each configuration (i.e. no need to call visualize\_epochs). We provide you with a few choices of common activation functions, but feel free to try out the others.  
  **Hint**: You can refer to explore\_mlp\_structure as a demonstration of how to define training configurations with fixed hyper-parameters & iterate over hyper-parameters/design choices of interests (e.g. hidden dimensions, choice of activation), and plot the evaluation results across configurations.

Once you complete the above functions, run explore\_mlp\_activations and paste the two generated plots here. Describe in 2-3 sentences your findings.  
A graph with lines and numbers

AI-generated content may be incorrect.A graph with numbers and lines

AI-generated content may be incorrect.

These charts suggest that the sigmoid activation function was best for this problem. The loss and accuracy are the most stable. I would be interested to see how this varies over the learning rates or model structures.

### Hyper-parameter Tuning: Learning Rate

The training process mostly involves learning model parameters, which are automatically performed by gradient-based methods. However, certain parameters are “unlearnable" through gradient optimization while playing a crucial role in affecting model performance, for example, learning rate and batch size. We typically refer to these parameters as *Hyper-parameters*.

We will now take the first step to tune these hyper-parameters by exploring the choices of one of the most important one - learning rate, on our MLP. (There are lots of tutorials on how to tune the learning rate manually or automatically in practice, for example [this note](https://www.kaggle.com/code/residentmario/tuning-your-learning-rate) can serve as a starting point.)  
  
**TODOs**: Read and complete the missing lines in explore\_mlp\_learning\_rates in main.py to iterate over different learning rate values, define the training configurations, train and evaluate the model, and visualize the results. Note: same as above, you only need to generate the plots of dev loss and dev acc across different configurations, by calling visualize\_configs, you **do not** need to plot the train-dev loss curves for each configuration (i.e. no need to call visualize\_epochs). We provide you with the default learning rate we set to start with, and we encourage you to add more learning rate values to explore and include in your final plots curves of **at least 4 different representative learning rates.**  
**Hint**: again, you can checkout explore\_mlp\_structure as a demonstration for how to perform hyper-parameter search.  
Once you complete the above functions, run explore\_mlp\_learning\_rates and paste the two generated plots here. Describe in 2-3 sentences your findings.A graph of different colored lines

AI-generated content may be incorrect.A graph with numbers and lines

AI-generated content may be incorrect.

It is interesting to see the variation across different learning rates. The charts suggest that the loss and accuracy are most stable and consistently low when using a learning rate of 0.025. To expand on this, I would probably look at the performance of models with different configurations of learning rate combined with activation functions.

1. <https://en.wikipedia.org/wiki/Kronecker_delta> [↑](#footnote-ref-1)