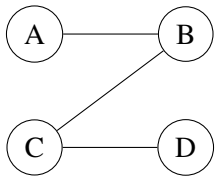
EECS 182 Deep Neural Networks
Fall 2025 Anant Sahai and Gireeja Ranade

Discussion 7

1. Graph Neural Network Forward Pass

Consider the following undirected graph G:



In this problem, we are going to work with an undirected graph without edge weights. We are imposing these limitations to make the problem simpler and to let us focus on the core ideas behind the forward pass. In practice, edges can be directed and have weights.

When a GNN layer is applied to a graph, it produces a new graph with the same topology as the original graph but with (potentially) different values in the nodes and the edges. After passing a graph through a number of these GNN layers, we can use the graph embedding for a variety of downstream tasks. In this problem, we will walk through a simplified forward pass of graph neural networks to help build concrete intuition for how GNNs operate (and so we will not be thinking about the downstream tasks). We will gradually add layers of complexity to the forward pass to allow GNNs to become more expressive.

To begin with, let us assign values v_A, v_B, v_C, v_D to nodes A, B, C, D respectively. Let:

$$v_A = 1, v_B = -2, v_C = -1, v_D = 2$$

We will define our update function for our nodes as follows:

$$f_v(v_i) = \begin{bmatrix} 1 & 2 \end{bmatrix} \text{ReLU}\left(\begin{bmatrix} -1 \\ 1 \end{bmatrix} v_i \right)$$

In practice, we could have different update functions at different layers of our network (and more generally, these update functions are learnable). For the sake of this problem, we will reuse the same update function at every layer of the network.

For example, to produce the node value at timestep $t+1$, we must apply the update rule to the node from timestep t, and so we have:

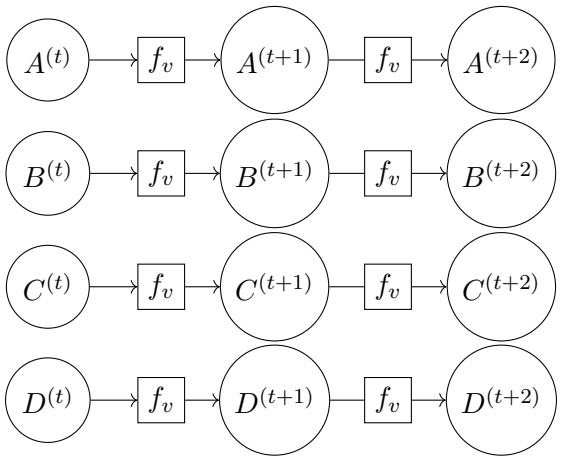
$$v_i^{(t+1)} = f_v(v_i^{(t)})$$

Thus, to compute $v_A^{(1)}$, we have:

$$v_A^{(1)} = f_v(v_A^{(0)}) = \begin{bmatrix} 1 & 2 \end{bmatrix} \text{ReLU}\left(\begin{bmatrix} -1 \\ 1 \end{bmatrix} v_A^{(0)} \right) = \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 2$$

In general, to produce the graph at timestep $t+1$, we will apply the update rules to each node in our graph from timestep t. Suppose that at timestep 0, the graph G is as above. Let us denote the state of G at timestep t by $G^{(t)}$.

(a) We can visualize two iterations of our current update rule with a diagram such as the following:



For what value of k would the update function defined for each node above be a counterpart to a $k \times k$ convolutional net?

Solution: $k = 1$, so a 1x1 convolution since no information from neighboring nodes is being used.

(b) From this diagram, it is clear to see that our GNN is not leveraging the topology of our graph in its forward pass. The way we overcome this is through message passing. In practice, we can apply message passing to both nodes and edges. In this problem, we will only consider applying message passing to nodes to simplify things. Let us consider the new update rule for nodes:

$$v_i^{(t+1)} = f_v(v_i^{(t)}) + \sum_{v_j \in N(v_i)} f_v(v_j^{(t)})$$

where $N(v_i)$ is the set of neighbors of node v_i.

Compute $\mathbf{G^{(1)}}$ under this new update rule.

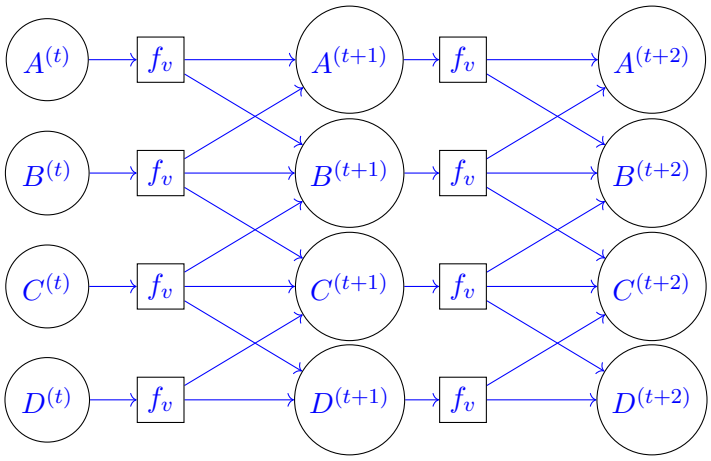
Solution:

$$v_A^{(1)} = f_v(v_A^{(0)}) + f_v(v_B^{(0)}) = 4$$
$$v_B^{(1)} = f_v(v_A^{(0)}) + f_v(v_B^{(0)}) + f_v(v_C^{(0)}) = 5$$

$$v_C^{(1)} = f_v(v_B^{(0)}) + f_v(v_C^{(0)}) + f_v(v_D^{(0)}) = 7$$
$$v_D^{(1)} = f_v(v_C^{(0)}) + f_v(v_D^{(0)}) = 5$$

(c) **Draw a diagram like the one in part (a) reflecting two iterations of our new update rule.**

Solution:



(d) Suppose the shortest path between nodes u and v in a graph traverses K edges. **How many iterations of updates must we do for information from node u to reach node v**? *(Hint: Consider the network diagram you made in part (c))*

Solution: From looking at the network diagram, we can see that it will take K iterations for these nodes to communicate with each other.

(e) Given a graph where the diameter (the longest shortest path between any two nodes) is L edges, **how many layers at least should our GNN have to ensure that the receptive field of each node is the entire graph?**

Solution: For information to propagate between any two nodes in the graph, the number of layers in the GNN should be at least equal to the distance (in terms of edge hops) between those two nodes.

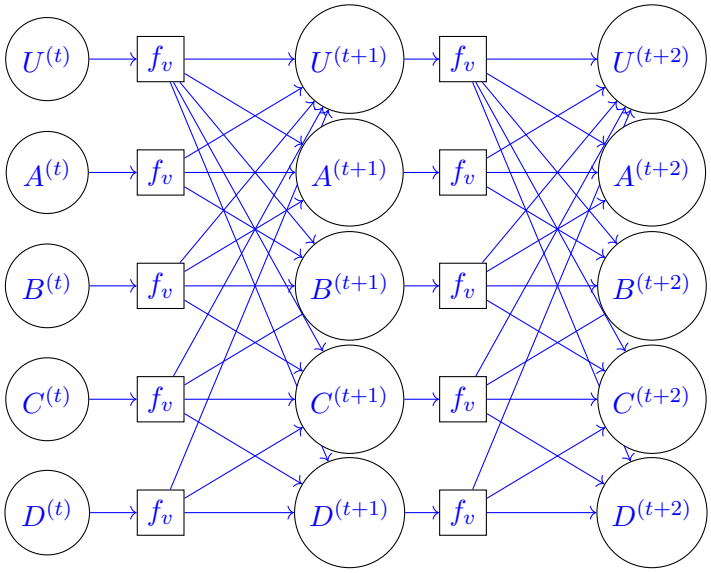
The maximum of such distances in a graph is termed the diameter, which is given as L in this case. Therefore, to ensure that any two nodes in the graph, even those maximally distant from each other, can exchange information, the GNN should have at least L layers.

(f) To speed up information flow between nodes, once can add a dummy node to our graph that is maximally connected to all of the original nodes in the graph. **How many layers are needed now to ensure that the receptive field of each node is the entire graph?**

Solution: With the addition of a dummy node that is connected to all other nodes in the graph, any two nodes can communicate with each other in at most 2 layers. This is because a node can send its information to the dummy node in one layer, and then the dummy node can relay that information to any other node in the next layer. Therefore, with this setup, only 2 layers are needed to ensure that the receptive field of each node encompasses the entire graph.

(g) **Draw a diagram like the one from part (c) reflecting the addition of this new dummy node.**

Solution:



We see that an arbitrary pair of nodes can communicate with each other in at most 2 layers.

2. Understand GNNs Through Graph Theory

You've seen in lecture how Graph Neural Nets (GNN) naturally generalize the mechanisms of ConvNets. While it is indeed helpful to understand from a ConvNet perspective, the point of GNNs is actually to solve

problems with graphs. This question is designed to teach you how the concepts in GNNs could emerge from just solving traditional graph problems.

Consider finding a *shortest path*. This kind of problem, to predict a function of a graph, is common in multiple applications such as in social network studies and molecular biology. Each instance of this problem:

- Has a given graph $G(V, E, W)$ where V is the set of vertices and E is the set of edges. The weight matrix $\mathbf{W}$ defines real-valued labels on the edges and is defined as

$$\mathbf{W}_{ij} = \begin{cases} \text{non-negative weight of edge from vertex } i \text{ to } j, & \text{if edge exists;} \\ \infty, & \text{otherwise.} \end{cases}$$

- Has an input $\mathbf{X}$ that represents the special *source vertex*.
- Has as desired output a vector $\mathbf{Y} \in [0, +\infty]^{|V|}$, where the i-th entry is is the shortest distance from the i-th vertex to specified source vertex.

Suppose we decided to frame this problem as a graph neural net. The network needs to learn a *mapping function* $f_\theta(\mathbf{X}) = \mathbf{Y}$, where the underlying function f is the shortest path.

The first question is whether or not this function is even representable by a GNN. Consider the pseudocode below for Dijkstra's algorithm, a well-known way of computing shortest paths in an iterative manner (notice the while loop). You can also look at the animation on Wikipedia.

```
def dijkstras(source):
    PQ.add(source, 0)
    For all other vertices, v, PQ.add(v, infinity)
    while PQ is not empty:
        p = PQ.removeSmallest()
        relax(all edges from p)
```

```
def relax(edge p,q):
   if q is visited (i.e., q is not in PQ):
       return

   if distTo[p] + weight(edge) < distTo[q]:
       distTo[q] = distTo[p] + w
       edgeTo[q] = p
       PQ.changePriority(q, distTo[q])
```

Figure 1: Pesudocode from https://joshhug.gitbooks.io/hug61b

(a) **Is this a node-level, edge-level, or graph-level prediction?**

Solution: This is a node-level prediction problem since we are predicting the distance for each node in the graph. We can imagine the GNN as emiting a number for each node — namely how far it is from the specified source.

If we are only interested in the shortest distance between any two nodes, we can change the problem to be graph-level prediction.

If we want to predict the shortest path between two nodes, we can change the problem to be edge-level prediction, where the output vector is a 0-1 vector indicating whether a given edge is on the shortest path between the two nodes.

(b) In GNN, a key property is message passing along edges of the graph. **Which part of the pseudocode above resembles the idea of message passing?**

How many iterations (forward pass) does it at most take to propagate the required information through the graph?

Solution:

The *relax* method acts like message passing in GNN since it updates the distances according to the rule of `distTo[q] = distTo[p] + w` where p and q are neighbors. Therefore it will take at most L iterations to propagate the information through the graph where L is the longest shortest path from source node s.

(c) In the *relax* function we need a $\min$ function, which is not differentiable. **Can you think of a way to approximate this with non-linearities in a neural net?**

Solution: There could be multiple answers. One way to create $\min$ function using differentiable non-linearities is with softmin, which is essentially softmax but negative, namely:

$$x_v^{t+1} = -\frac{1}{\lambda} \log \sum_{v' \in \mathcal{N}(v)} \exp\Big(-\lambda f(x_{v'}^t)\Big),$$

where $\lambda > 0$.

3. Backprop through a Simple RNN

Consider the following 1D RNN with no nonlinearities, a 1D hidden state, and 1D inputs u_t at each timestep. (Note: There is only a single parameter w, no bias). This RNN expresses unrolling the following recurrence relation, with hidden state h_t at unrolling step t given by:

$$h_t = w \cdot (u_t + h_{t-1}) \tag{1}$$

The computational graph of unrolling the RNN for three timesteps is shown below:

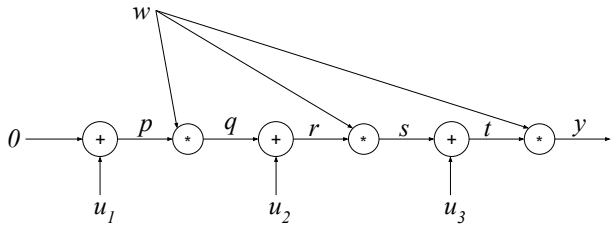


Figure 2: Illustrating the weight-sharing and intermediate results in the RNN.

where w is the learnable weight, u_1, u_2, and u_3 are sequential inputs, and p, q, r, s, and t are intermediate values.

(a) **Fill in the blanks for the intermediate values during the forward pass, in terms of w and the u_i's:**

$$t = \_\_\_\_\_$$

$$y = \_\_\_\_\_$$

Solution: $t = u_3 + s = u_3 + w \cdot u_2 + w^2 \cdot u_1$ **Solution:** $y = w \cdot t = w \cdot u_3 + w^2 \cdot u_2 + w^3 \cdot u_1$

(b) **Using the expression for y from the previous subpart, compute $\frac{dy}{dw}$.**

Solution: $\frac{dy}{dw} = u_3 + 2wu_2 + 3w^2u_1$

(c) **Fill in the blank for the missing partial derivative of y with respect to the nodes on the backward pass.** You may use values for p, q, r, s, t, y computed in the forward pass and downstream derivatives already computed.

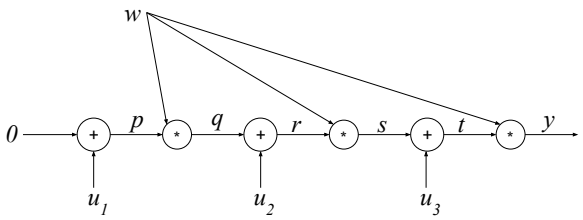
$$\frac{\partial y}{\partial p} = \_\_\_\_\_$$

Solution: $\frac{\partial y}{\partial p} = \frac{\partial y}{\partial q} \cdot w = w \cdot w \cdot w = w^3$

(d) **Calculate the partial derivatives along each of the three outgoing edges from the learnable w in Figure 2, replicated below.** (e.g., the right-most edge has a relevant partial derivative of t in terms of how much the output y is effected by a small change in w as it influences y through this edge. You need to compute the partial derivatives for the other two edges yourself.)

You can write your answers in terms of the p, q, r, s, t and the partial derivatives of y with respect to them.

Use these three terms to find the total derivative $\frac{dy}{dw}$.



(HINT: You can use your answer to part (b) to check your work.)

Solution: Along the right edge, we have $t = u_3 + wu_2 + w^2u_1$ (This is provided for you).
Along the middle edge, we have $r \cdot \frac{\partial y}{\partial s} = r \cdot w = (u_2 + wu_1)w = wu_2 + w^2u_1$
Along the left edge, we have $p \cdot \frac{\partial y}{\partial q} = p \cdot w^2 = u_1w^2$
Adding all of these up, we have

$$\frac{dy}{dw} = t + r \cdot \frac{\partial y}{\partial s} + p \cdot \frac{\partial y}{\partial q} \tag{2}$$

$$= (u_3 + wu_2 + w^2u_1) + (u_2 + wu_1)w + w^2u_1 \tag{3}$$

$$= 3w^2u_1 + 2wu_2 + u_3 \tag{4}$$

Which is same as answer to (b) so everything checks out.

(e) **What is the number of computations performed during one forward step of the RNN? Assume that you already have access to the current hidden state.**

Solution: Using the update equation, we can see that the number of computations is 1 multiplication and 1 addition per timestep.

(f) **What is the number of computations performed during T forward steps of the RNN? Assume that you already have access to the current hidden state.**

Solution: Using the answer to the previous part, we can see that the number of computations is T multiplications and T additions.

(g) **Now, given a T timestep sequence of inputs, what is the number of computations performed during the forward pass of the RNN? How many of these computations can be done in parallel? Do know worry about exact implementation details, just give a high-level overview. Solution:**

Naively, each timestep of the recurrence

$$h_t = w\big(u_t + h_{t-1}\big)$$

requires one addition and one multiplication once (h_{t-1}) is known. Hence, over (T) timesteps,

$$T \text{ additions} + T \text{ multiplications} = 2T \text{ operations.}$$

Because each (h_t) depends on (h_{t-1}), this direct implementation is not parallelizable across timesteps. However, by noticing that

$$h_t = w^t u_1 + w^{t-1} u_2 + \cdots + w u_t$$

is *linear*, one can exploit "prefix" or "divide-and-conquer" methods to compute all (h_t) in parallel.

So the total work remains $2T$ operations, but lot of this work can be done in parallel. Particularly, the multiplications can be done in parallel, and the additions can be performed in a divide-and-conquer way.

Contributors:

- Olivia Watkins.
- Anrui Gu.
- Kevin Li.
- Suhong Moon.
- Anant Sahai.
- Saagar Sanghavi.
- Naman Jain.

