EECS 182 Deep Neural Networks
Spring 2023 Anant Sahai

Review: ConvNets & GNNs

1. CNN Design decisions

Consider a CNN for classification consisting of the layers shown below. "..." indicates appropriate arguments that have been passed in.

```
nn.Conv2d(...),
nn.ReLU(),
nn. MaxPool2d(...),
nn. Conv2d(...),
nn. ReLU(),
nn. MaxPool2d(...),
nn. Flatten(),
nn. Linear(...),
nn. Softmax(...)
```

In the questions below, explain how you would modify this design to achieve particular objectives.

(a) You would like to use fewer parameters in the fully connected layer at the end of the network. **Solution:** The parameters in the final fully connected layer are equal to (num classes) * (flattened output from the last conv block). We can reduce this flattened output by either reducing the number of channels in the last conv layer or by reducing the spatial dimension of the feature map. You can reduce the spatial dimension by increasing the stride in the conv/pooling layers, or by adding additional conv/pooling layers with stride > 1. Removing padding from the conv/pooling layers also slightly decreases the spatial dimension.

(b) You would like the model to be able to handle images of different sizes.

Solution: Replace the Flatten and fully connected layers with a global pooling layer.

(c) You would like to increase the receptive field of each pixel in the feature map output by the second conv layer. **Solution:** Increase the stride of the first conv or max pool layer. Increasing the kernel size of either conv layer or the first max pool layer has a small effect too.

(d) You would like every conv layer to leave the height and width dimensions of its input unchanged.

Solution: Use a stride of 1 and add padding to each conv layer. In Pytorch, this is 'SAME' padding. Manually, you would pad by $(K-1)/2$ on each side, where K is the kernel height/width.

(e) You would like to add 50 more layers to this network without suffering from vanishing gradients.

Solution: Add Resblocks (blocks with residual connections) to the network rather than simple conv layers.

2. 3D Convolutions

We've seen CNNs in 1D and 2D settings. Now, let's consider a 3D setting: you are building a classifier to detect the activity shown in short video clips.

(a) One approach is to build a CNN which uses 3D convolutions. What is the size of the parameters in each convolutional layer? Write your answer in terms of F (number of filters), S (stride), C (number of input channels), H (input height), W (input width), T (input time length), K (kernel size).

Solution: The weight will be (F, C, K, K, K) (Note: in some implementations, the C dimension comes last). The bias is F.

(b) Another approach is to process frames individually, then use an RNN to aggregate information over time. How would you combine a convolutional encoder with an RNN?

Solution: At each timestep, encode the current frame with the same convolutional encoder. This encoder should include a flatten or global pooling layer which collapses the input into a single vector. This vector is then passed as the input at that timestep to the RNN. Make the final classification using a linear layer on top of the finall hidden state of the RNN.

3. Feature Dimensions of Convolutional Neural Network

In this problem, we compute output feature shape of convolutional layers and pooling layers, which are building blocks of CNN. Let's assume that input feature shape is $W \times H \times C$, where W is the width, H is the height and C is the number of channels of input feature.

(a) A convolutional layer has 4 architectural hyperparameters: the filter size(K), the padding size (P), the stride step size (S) and the number of filters (F). **How many weights and biases are in this convolutional layer? And what is the shape of output feature that this convolutional layer produces?**

Solution:

The number of weights $= K^2CF$
The number of biases $= F$
$W' = \lfloor (W - K + 2P)/S \rfloor + 1$
$H' = \lfloor (H - K + 2P)/S \rfloor + 1$
$C' = F$

(b) A max pooling layer has 2 architectural hyperparameters: the stride step size(S) and the "filter size" (K). **What is the output feature shape that this pooling layer produces?**

Solution:
$W' = (W - K)/S + 1$
$H' = (H - K)/S + 1$
$C' = C$

(c) Let's take a real example. We are going to describe a convolutional neural net using the following pieces:

- CONV3-F denotes a convolutional layer with F different filters, each of size $3 \times 3 \times C$, where C is the depth (i.e. number of channels) of the activations from the previous layer. Padding is 1, and stride is 1.
- POOL2 denotes a 2×2 max-pooling layer with stride 2 (pad 0)
- FLATTEN just turns whatever shape input tensor into a one-dimensional array with the same values in it.
- FC-K denotes a fully-connected layer with K output neurons.

Note: All CONV3-F and FC-K layers have biases as well as weights. **Do not forget the biases when counting parameters.**

Now, we are going to use this network to do inference on a single input. **Fill in the missing entries in this table of the size of the activations at each layer, and the number of parameters at each layer. You can/should write your answer as a computation (e.g.** $128 \times 128 \times 3$**) in the style of the already filled-in entries of the table.**

Layer	Number of Parameters	Dimension of Activations
Input	0	$28 \times 28 \times 1$
CONV3-10	**Solution:** $3 \times 3 \times 1 \times 10 + 10$	$28 \times 28 \times 10$
POOL2	0	$14 \times 14 \times 10$
CONV3-10	$3 \times 3 \times 10 \times 10 + 10$	**Solution:** $14 \times 14 \times 10$
POOL2	**Solution:** 0	**Solution:** $7 \times 7 \times 10$
FLATTEN	0	490
FC-3	**Solution:** $490 \times 3 + 3$	3

4. Graph Neural Networks

For an undirected graph with no labels on edges, the function that we compute at each layer of a Graph Neural Network must respect certain properties so that the same function (with weight-sharing) can be used at different nodes in the graph. Let's focus on a single particular "layer" ℓ. For a given node i in the graph, let $\mathbf{s}_i^{\ell-1}$ be the self-message (i.e. the state computed at the previous layer for this node) for this node from the preceeding layer, while the preceeding layer messages from the n_i neighbors of node i are denoted by $\mathbf{m}_{i,j}^{\ell-1}$ where j ranges from 1 to n_i. We will use w with subscripts and superscripts to denote learnable scalar weights. If there's no superscript, the weights are shared across layers. Assume that all dimensions work out.

(a) **Tell which of these are valid functions for this node's computation of the next self-message $\mathbf{s}_i^{\ell}$. For any choices that are not valid, briefly point out why.**
Note: we are *not* asking you to judge whether these are useful or will have well behaved gradients. Validity means that they respect the invariances and equivariances that we need to be able to deploy as a GNN on an undirected graph.

(i) $\mathbf{s}_i^{\ell} = w_1 \mathbf{s}_i^{\ell-1} + w_2 \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{m}_{i,j}^{\ell-1}$
Solution: This is valid because it is permutation invariant to the ordering of neighbors. This is the classic averaging form. Notice that a dependence on the number of neighbors is fine.

(ii) $\mathbf{s}_i^{\ell} = \max(w_1^{\ell} \mathbf{s}_i^{\ell-1}, w_2 \mathbf{m}_{i,1}^{\ell-1}, w_3 \mathbf{m}_{i,2}^{\ell-1}, \ldots, w_{n_i-1} \mathbf{m}_{i,n_i}^{\ell-1})$ where the $\max$ acts component-wise on the vectors.
Solution: This is invalid. Since different scalar weights are applied to different $\mathbf{m}$, it is not permutation invariant to the ordering of neighbors.

(iii) $\mathbf{s}_i^{\ell} = \max(w_1^{\ell} \mathbf{s}_i^{\ell-1}, w_2 \mathbf{m}_{i,1}^{\ell-1}, w_2 \mathbf{m}_{i,2}^{\ell-1}, \ldots, w_2 \mathbf{m}_{i,n_i}^{\ell-1})$ where the $\max$ acts component-wise on the vectors.
Solution: This is valid. Since the same weight w_2 is applied to all $\mathbf{m}$, it is permutation invariant to the ordering of neighbors. The max is another classic permutation-invariant operation.

(b) Suppose we decide to use the following update rule for the internal state of the nodes at layer ℓ.

$$\mathbf{s}_i^{\ell} = \mathbf{s}_i^{\ell-1} + W_1 \frac{\sum_{j=1}^{n_i} \tanh\left(W_2 \mathbf{m}_{i,j}^{\ell-1}\right)}{n_i} \tag{1}$$

where the $\tanh$ nonlinearity acts element-wise.
For a given node i in the graph, let $\mathbf{s}_i^{\ell-1}$ be the self-message for this node from the preceeding layer, while the preceeding layer messages from the n_i neighbors of node i are denoted by $\mathbf{m}_{i,j}^{\ell-1}$ where j ranges from 1 to n_i. We will use W with subscripts and superscripts to denote learnable weights in matrix form. If there's no superscript, the weights are shared across layers.

(i) **Which of the following design patterns does this update rule have?**

☐ Residual connection

☐ Batch normalization

Solution: This rule shows the residual connection pattern since the previous layer's state gets added in. This kind of residual connection enables gradients to flow back to earlier layers more easily. None of the other patterns are present.

(ii) **If the dimension of the state $\mathbf{s}$ is d-dimensional and W_2 has k rows, what are the dimensions of the matrix W_1? Solution:** W_1 needs to return something that can add to the state. This means that it needs d rows. It also has to be able to act on vectors that are k dimensional since W_2 has k rows and we say that the $\tanh$ operation acts element-wise. This means that W_1 needs k columns. Putting this together, the dimensions of W_1 are $d \times k$.

(iii) If we choose to use the state $\mathbf{s}_i^{\ell-1}$ itself as the message $\mathbf{m}^{\ell-1}$ going to all of node i's neighbors, **please write out the update rules corresponding to (1) giving $\mathbf{s}_i^{\ell}$ for the graph in Figure 1 for nodes $i = 2$ and $i = 3$ in terms of information from earlier layers.** Expand out all sums.

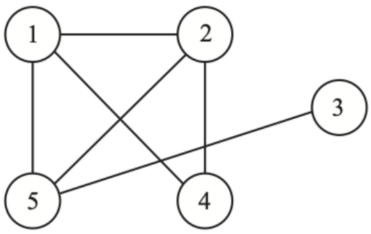

Figure 1: Simple Undirected Graph

Solution: We can write the mean-pooling/ average pooling as in the the neighborhood N_i is of node i:
Therefore:
i = 2:

$$s_2^l = s_2^{l-1} + \frac{W_1}{3}(\tanh\left(W_2 s_1^{l-1}\right) + \tanh\left(W_2 s_4^{l-1}\right) + \tanh\left(W_2 s_5^{l-1}\right))$$

i = 3:

$$s_3^l = s_3^{l-1} + \frac{W_1}{1}(\tanh\left(W_2 s_5^{l-1}\right))$$

The above equations can be found by reducing the sum in equation 1 and applying the accurate neighbors from Figure 1.

Contributors:

- Suhong Moon.
- Fei-Fei Li.
- Jerome Quenum.
- Anant Sahai.

