1 DS GA 1008 HW 2

2 Joby George

3 Due 10/13 4 PM

4 Section 1: Theory

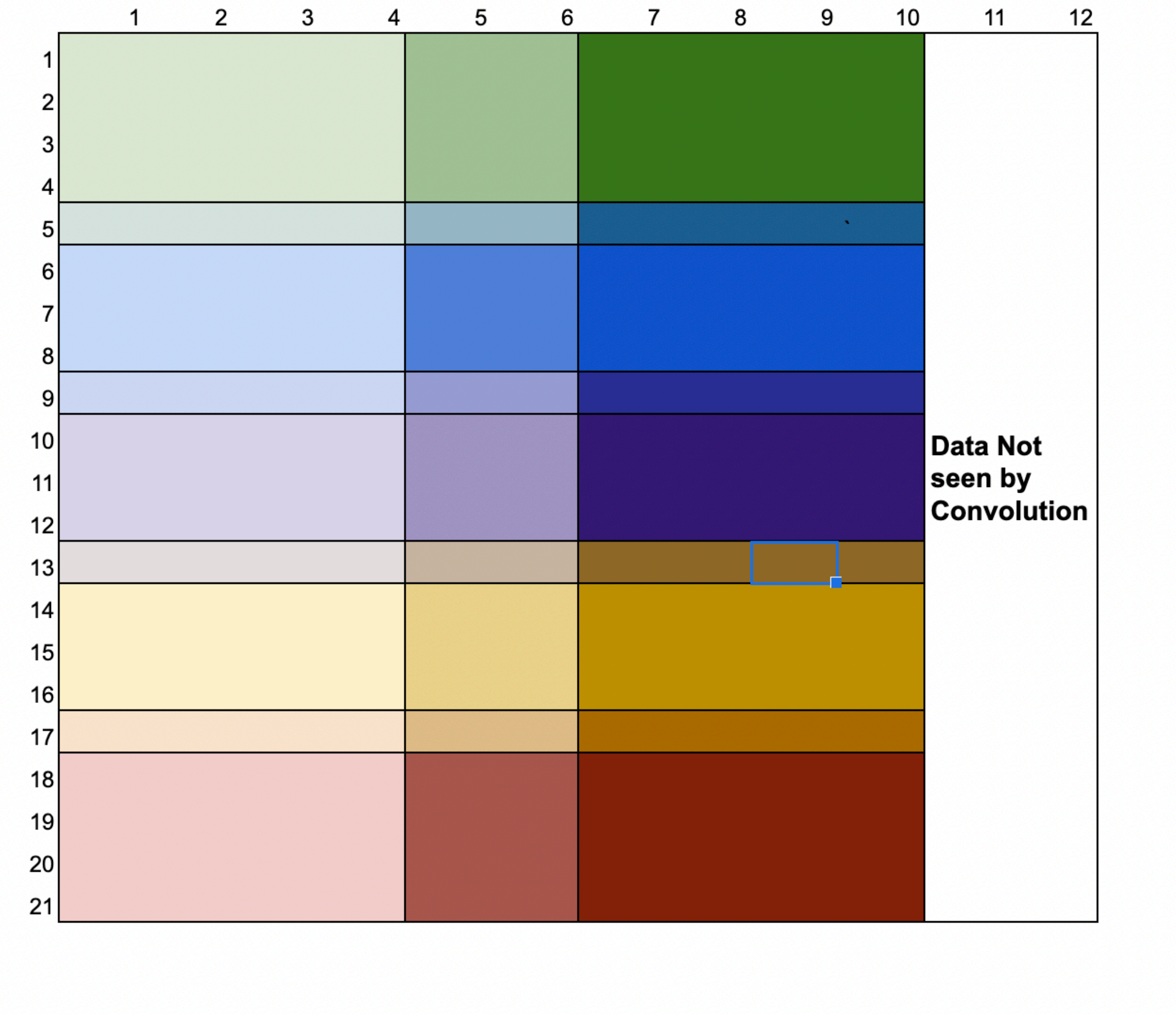
5 Section 1.1 Convultional Neural Networks

5.1 1.1.A

Given an input image of dimmension 21 x 12 what will the output dimmension be after applying a convolution with a 4x5 kernel, stride of 4 and no padding?

5.2 1.1.A Answer

To help in my understanding of this problem, I've attached a visual diagram that shows how this convolution would impact the input image.



We can see that there are two differentiations of colors, light vs dark, indicating the column output dimmension of this convolution would be 2.

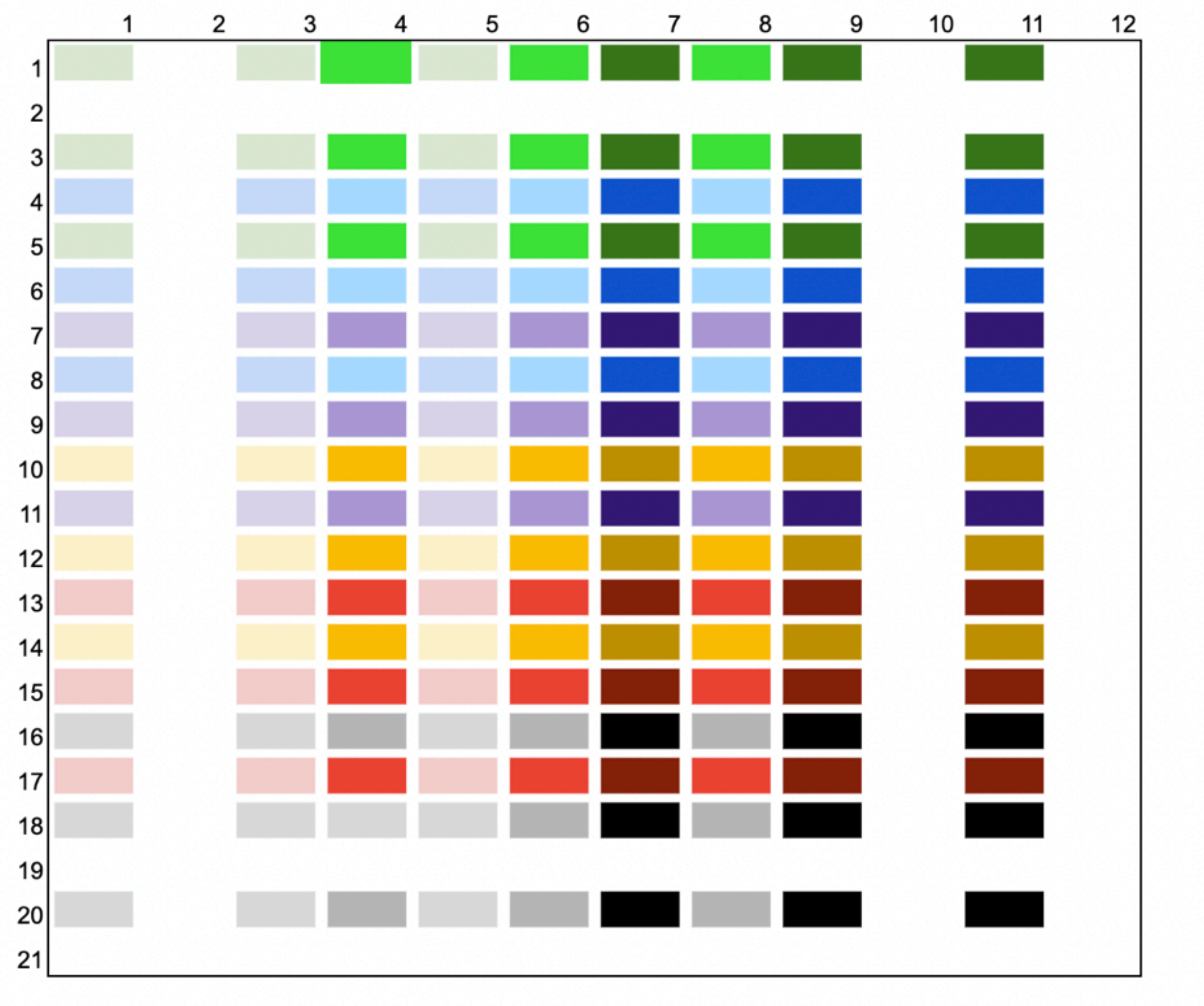
There are 5 primary types of colors, indicating the row dimmension would be 5. Thus our output would be a **cx5x2** where c is the number of channels for this input image.

5.3 1.1 B

Given an input of dimmension C X H X W what will be the dimension of the output of a convolutional layer with kernel of size K x K, padding P, stride S, dilation D, and F filters. Assume that H $\geq K, W \geq K$

5.4 1.1.B Answer

To help me understand visually how stride and dilation worked together, i created a visualization.



It became apparent after applying a convolution of size K X K, with padding 0, stride S, Dilation D, and filters (or num convolutions) F we get:

$$ouput \in R^{(F\text{x}C\text{x}\lfloor (H-DK+0+S)/S\rfloor \text{x}\lfloor (W-DK+0+S)/S\rfloor)} \tag{1}$$

where $\lfloor x \rfloor$denotes the floor function on x

Therefore, after applying a **convolution of size K X K, with padding P, stride S, Dilation D, and filters (or num convolutions) F** we get:

$$ouput \in R^{(F\text{x}C\text{x}\lfloor (H-DK+P+S)/S\rfloor \text{x}\lfloor (W-DK+P+S)/S\rfloor)} \tag{2}$$

5.4.1 QED

5.5 1.1.C

In this section we are going to work with 1-dimensional convolutions. Discrete convolution of 1-dimensional input x[n] and kernel k[n] is defined as follows:

$$s[n] = (x * k)[n] = \sum_m x[n-m]k[m] \tag{3}$$

However, in machine learning convolution is usually implemented as cross-correlation which is defined as follows:

$$s[n] = (x * k)[n] = \sum_m x[n+m]k[m] \tag{4}$$

Note the difference in signs, which will get the network to learn a "flipped" kernel. In general it doesn't change much, but it's important to keep it in mind. In convolution neural networks, the kernel k[n] is usually 0 everywhere, except a few values near 0: $\forall_{|n|>M} k[n] = 0$. Then the formula becomes

$$s[n] = (x * k)[n] = \sum_{m=-M}^{M} x[n+m]k[m] \tag{5}$$

Let's consider an input x[n] ∈ R^5, with $1 \leq n \leq 7$ e.g. it is a length 7 sequence with 5 channels. We consider the convolutional layer f_{W} with one filter, with kernel size 3, stride of 2, no dilation and no padding. The only parameters of the convolutional layer is the weight W, w ∈ R^{1x5x3}, there's no bias and no non-linearity

5.5.1 1.1.C.I

What is the dimension of the output $f_W(x)$? Provide an expression for the value of elements of the convolutional layer output $f_W(x)$ Example answer format here and in the following problems

$$f_W(x) \in R^{42X42X42}, f_W(x)[i,j,k] = 42 \tag{6}$$

5.5.2 1.1.C.I Answer

$$f_W(x) \in R^{1x5X3} \tag{7}$$

Going forward, let $\hat{y} = f_W(x)$

Thinking about the first convolution, we take a 3 unit interval of the first channel's input to get a vector ∈ R^{1x3}. We take the dot product with W_1 a vector ∈ R^{3x1} and perform the dot product giving us a scalar value.

We can represent all 5 channels by taking the first 3 units of all 5 channels, giving us a data matrix ∈ R^{5X3} and perform the dot product of each channel with the corresponding weight, matrix each with dimmension ∈ R^{3x1} giving us a resulting vector ∈ R^{5x1}

Because there is a stride of 2, we then take inputs x[3],x[4],x[5] and repeat this process giving us another R^{5x1}, and lastly we repeat this operation one last time using inputx x[5],x[6],x[7] to get our third vecotr ∈ R^{5x1}.

$$\hat{y}_{1,j,k} = \sum_{m=1}^{3} W_{j+m} \cdot X_{j,2k-1+m} \tag{8}$$

While the notation is dense, essentially we are taking the dot product between the R^{1x3} weight matrix for channel j and element wise multiplying it by the associated inputs, for a given i,j in our output matrix.

i.e.

$$\hat{y}_{1,2,3} = W_{21} * x_{25} + W_{22} * x_{26} + W_{23} * x_{27} \tag{9}$$

5.5.3 1.1.C.II

What is the dimension of $\frac{\partial \hat{y}}{\partial W}$? Provide an expression for $\frac{\partial \hat{y}}{\partial W}$

5.5.4 1.1.C.II Answer

$\frac{\partial \hat{y}}{\partial W} \in R^{5x3x3}$

Thinking about our convolution operator on one channel,

$$\hat{y}_{(1,1)} = W_{11} * x_{11} + W_{12}x_{12} + W_{13}x_{13} \tag{10}$$

$$\hat{y}_{(1,2)} = W_{11} * x_{13} + W_{12}x_{14} + W_{13}x_{15} \tag{11}$$

$$\hat{y} + (1,3) = W_{11} * x_{15} + W_{12}x_{16} + W_{13}x_{17} \tag{12}$$

If we were to take the $\frac{\partial \hat{y}_1}{\partial W_1}$we get:

$$\frac{\partial \hat{y}_1}{\partial W_1} = \begin{bmatrix} \frac{\partial \hat{y}_{11}}{\partial W_{11}} & \frac{\partial \hat{y}_{11}}{\partial W_{12}} & \frac{\partial \hat{y}_{11}}{\partial W_{13}} \\ \\ \frac{\partial \hat{y}_{12}}{\partial W_{11}} & \frac{\partial \hat{y}_{12}}{\partial W_{12}} & \frac{\partial \hat{y}_{12}}{\partial W_{13}} \\ \\ \frac{\partial \hat{y}_{13}}{\partial W_{11}} & \frac{\partial \hat{y}_{13}}{\partial W_{13}} & \frac{\partial \hat{y}_{13}}{\partial W_{13}} \end{bmatrix} \tag{13}$$

Re-expressing this:

$$\frac{\partial f \hat{y}_1}{\partial W_1} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{13} & x_{14} & x_{15} \\ x_{15} & x_{16} & x_{17} \end{bmatrix} \tag{14}$$

Expanding this to the general form:

$$\frac{\partial \hat{y}}{\partial W_{ijk}} = X_{1,i,2j+k-2} \tag{15}$$

5.5.5 1.1.C.III

What is the dimension of $\frac{\partial f_W(x)}{\partial x}$? Provide an expression for $\frac{\partial f_W(x)}{\partial x}$

5.5.6 1.1.C.III Answer

$$\frac{\partial \hat{y}}{\partial x} \in R^{5x3x7} \tag{16}$$

Recalling the work from the previous problem

$$\begin{align}
\hat{y}_{(1,1)} &= W_{11} * x_{11} + W_{12}x_{12} + W_{13}x_{13} \tag{17}\\
\hat{y}_{(1,2)} &= W_{11} * x_{13} + W_{12}x_{14} + W_{13}x_{15} \tag{18}\\
\hat{y}_{(1,3)} &= W_{11} * x_{15} + W_{12}x_{16} + W_{13}x_{17} \tag{19}
\end{align}$$

If we were to take the $\frac{\partial f_W(x)_1}{\partial W_1}$ we get:

$$\frac{\partial \hat{y}_1}{\partial x_1} = \begin{bmatrix} \frac{\partial \hat{y}_{11}}{\partial x_{11}} & \frac{\partial \hat{y}_{11}}{\partial x_{12}} & \cdots & \frac{\partial \hat{y}_{11}}{\partial x_{17}} \\ \\ \frac{\partial \hat{y}_{12}}{\partial x_{11}} & \frac{\partial \hat{y}_{12}}{\partial x_{12}} & \cdots & \frac{\partial \hat{y}_{12}}{\partial x_{17}} \\ \\ \frac{\partial \hat{y}_{13}}{\partial x_{11}} & \frac{\partial \hat{y}_{12}}{\partial x_{12}} & \cdots & \frac{\partial \hat{y}_{13}}{\partial x_{17}} \end{bmatrix} \tag{20}$$

Re-expressing this:

$$\frac{\partial \hat{y}_1}{\partial x_1} = \begin{bmatrix} W_{11} & W_{12} & W_{13} & 0 & 0 & 0 & 0 \\ \\ 0 & 0 & W_{11} & W_{12} & W_{13} & 0 & 0 \\ \\ 0 & 0 & 0 & 0 & W_{11} & W_{12} & W_{13} \end{bmatrix} \tag{21}$$

Expanding this to the general form:

$$\frac{\partial \hat{y}}{\partial x_{ijk}} = \begin{cases} W_{1,\lfloor k/j \rfloor} \text{ if } \frac{k}{j} < 2 \\ W_{1,\lceil k/j \rceil} \text{ if } 2 \leq \frac{k}{j} \leq 3 \\ 0 \text{ else} \end{cases} \tag{22}$$

5.5.7 1.1.C.IV

Now suppose you are given the gradient of the loss l w.r.t the output of the convolutional layer $f_W(x)$,i.e. $\frac{\partial l}{\partial f_W(x)}$.Whta is the dimension of $\frac{\partial l}{\partial W}$? Provide an expression for $\frac{\partial l}{\partial W}$. Explain similarities and differences of this expression and expression in (i).

5.5.8 1.1.C.IV Answer

6 Section 2 Recurrent Neural Networks

6.1 1.2.1

In this section we consider a simple recurrent neural network defined as follows:

$$c[t] = \sigma(W_c x[t] + W_h h[t-1]) \tag{23}$$
$$h[t] = c[t] \odot h[t-1] + (1 - c[t]) \odot W_x x[t] \tag{24}$$

Where σ is elmeent-wise sigmoid, $x[t] \in R^n, h[t] \in R^m, W_c \in R^{mxn}, W_h \in R^{mxm}, \odot$ is the Hadamard product (element wise multiplicaiton), h[0] = 0

6.2 1.2.1.A

Draw a diagaram for this recurrent neural network, similar to the diagarm of rNN we had in class. We suggest using diagrams.net

6.3 1.2.1.A Answer

6.4 1.2.1.B

What is the dimension of c[t]

6.5 1.2.1.B Answer

The dimmension of c[t] is R^m

6.6 1.2.1.C

Suppose that we run the RNN to get a sequence of h[t] for t from 1 to K. Assuming we know the derivative $\frac{\partial l}{\partial h[t]}$, provide the dimmensions of, and an expression for values of $\frac{\partial l}{\partial W_x}$. What are the similarities of backward pass and forward pass in this RNN?

6.7 1.2.1.C Answer

Writing out the gradient, using the chain rule, we see:

$$\frac{\partial l}{\partial W_x} = \frac{\partial l}{\partial h[t]} \frac{\partial h[t]}{\partial W} \tag{25}$$

Since we can assume the first partial is known, we focus in on $\frac{\partial h[t]}{\partial W_x}$. Writing out the expression for h[t] we see:

$$h[t] = c[t] \odot h[t-1] + (1 - c[t]) \odot W_x x[t] \tag{26}$$

Since we are taking the derivative of a h[t] ∈ $R^m$with a matrix $W_x \in R^{mxn}$, $\frac{\partial h[t]}{\partial W_x}$must be of dimmension R^{mxmxn}.

Looking at the first object of this R^{mxmxn}tensor, we see:

\begin{equation}\frac{\partial{\mathcal{h[t]}}}{\partial{W_x}}{1jk} = \begin{bmatrix} \frac{\partial{h[t]{1}}}{\partial{W_{x11}}} & \frac{\partial{h[t]{1}}}{\partial{W{x12}}} & ... & \frac{\partial{h[t]{1}}}{\partial{W{x1n}}} \ \ \frac{\partial{h[t]{1}}}{\partial{W{x21}}} & \frac{\partial{h[t]{1}}}{\partial{W{x22}}} & ... & \frac{\partial{h[t]{1}}}{\partial{W{x2n}}} \ \ ... & ... & ... & .... \ \ \frac{\partial{h[t]{1}}}{\partial{W{xm1}}} & \frac{\partial{h[t]{1}}}{\partial{W{xm2}}} & ... & \frac{\partial{h[t]{1}}}{\partial{W{xmn}}}

`\end{bmatrix}\end{equation}`

We know that the first index of h[t] is the Hadmarand product of 1-c[t] and the first element of W_xx.

Therefore, if we were to change the first row of $W_x$we would see resulting changes in the first index of h[t]. However, changes to values outside of the first row of $W_x$will not change the first index of h[t], giving us:

$$\frac{\partial h[t]}{\partial W_{x\ 1jk}} = \begin{bmatrix} (1-c[t])_1 * x[t]_1 & (1-c[t])_1 * x[t]_2 & \dots & (1-c[t])_1 * x[t]_n \\ \\ 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots. \\ \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Generalizing this, we see:

$$\frac{\partial h[t]}{\partial W_{x\ ijk}} = \begin{cases} (1 - c[t])_i * x[k] \text{ if } j = i \\ 0 \text{ else} \end{cases} \tag{28}$$

6.8 1.2.1.D

Can this network be subject to vanishing or exploding gradients? Why?

6.9 Answer

The network can be subject to vanishing gradients as a result of the multiplication by (1-c[t]). Since the sigmoid function has a range of (0,1) and pushes higher values closer to 0 and smaller values closer to 0, the result of this operation will likely result in a multiplication by a small number when c[t] takes a large value.

The problem compounds itself by the recurrent infrastructure. When we unravelling the derivative, the previous hidden state is used as an input, thus the number of times we perform this multiplication is a function of T.

6.10 Section 1.2.2

We define an AttentionRNN(2) as:

$$q_0[t], q_1[t], q_2[t] = Q_0 x[t], Q_1 h[t-1], Q_2 h[t-2] \tag{29}$$
$$k_0[t], k_1[t], k_2[t] = K_0 x[t], K_1 h[t-1], K_2 h[t-2] \tag{30}$$
$$v_0[t], v_1[t], v_2[t] = V_0 x[t], V_1 h[t-1], V_2 h[t-2] \tag{31}$$

$$w_i[t] = q_i[t]^T k_i[t] \tag{32}$$
$$a[t] = softargmax([w_0[t], w_1[t], w_2[t]]) \tag{33}$$
$$h[t] = \sum_{i=0}^{2} a_i[t] v_i[t] \tag{34}$$

Where $x_i[t], h[t] \in R^n$, and $Q_i, K_i, V_i \in R^{nxn}$. We define h[t] = 0 for t <1. You may safely ignore these base cases in the following questions.

6.11 Section 1.2.2.A

Draw a diagram for this recurrent neural network

6.12 1.2.2.A Answer

6.13 Section 1.2.2.B

What is the dimmension of a[t]?

6.13.1 1.2.2.B Answer

a[t] is $\in R^3$

6.14 Section 1.2.2.C

Extend this to, AttentionRNN(k), a network that uses the last k state vectors h. Write out the system of equations that defines it. You may use set notation or elipses in your definintion.

6.14.1 1.2.2.C Answer

k $\in 1....k, k \leq T$

$$q_0[t], q_1[t], q_2[t], \ldots q_k[k] = Q_0 x[t], Q_1 h[t-1], Q_2 h[t-2], \ldots Q_k h[t-k] \quad (35$$

$$k_0[t], k_1[t], k_2[t], \ldots k_k[t] = K_0 x[t], K_1 h[t-1], K_2 h[t-2], \ldots K_k h[t-k] \quad (36$$

$$v_0[t], v_1[t], v_2[t]. \ldots v_k[t] = V_0 x[t], V_1 h[t-1], V_2 h[t-2], \ldots V_k h[t-k] \tag{37}$$

$$w_i[t] = q_i[t]^T k_i[t] \tag{38}$$

$$a[t] = softargmax([w_0[t], w_1[t], w_2[t], \ldots w_k[t]]) \tag{39}$$

$$h[t] = \sum_{i=0}^{k} a_i[t] v_i[t] \tag{40}$$

6.15 Section 1.2.2.D

Modify the above network to produce AttentionRNN(∞), a network that uses every past state vector. Write out the system of equations that defines it. You may use set notation or elipses in your definintion. **HINT**: We can do this by tying together some set of parameters, e.g. weight sharing.

6.16 1.2.2.D Answer

k $\in 1....T$

$$q_0[t], q_1[t], q_2[t], \dots q_k[k] = Q_0 x[t], Q_1 h[t-1], Q_2 h[t-2], \dots Q_k h[t-k] \quad (41$$

$$k_0[t], k_1[t], k_2[t], \dots k_k[t] = K_0 x[t], K_1 h[t-1], K_2 h[t-2], \dots K_k h[t-k] \quad (42$$

$$v_0[t], v_1[t], v_2[t] \dots v_k[t] = V_0 x[t], V_1 h[t-1], V_2 h[t-2], \dots V_k h[t-k] \tag{43}$$

$$w_i[t] = q_i[t]^T k_i[t] \tag{44}$$

$$a[t] = softargmax([w_0[t], w_1[t], w_2[t], \dots w_k[t]]) \tag{45}$$

$$h[t] = \sum_{i=0}^{k} a_i[t] v_i[t] \tag{46}$$

6.17 Section 1.2.2.E

Suppose the loss l is computed, and we know the derivative $\frac{\partial l}{\partial h[t]}$ Please write down the expression for $\frac{\partial h[t]}{\partial h[t-1]}$for AttentionRNN(2)

6.18 1.2.2.E Answer

Since we are taking a derivative of a vector $\in R^n$w.r.t a vector $\in R^n$our Jacobian will be an R^{nxn}matrix, looking something like this:

\begin{equation}\frac{\partial{\mathcal{h[t]}}}{\partial{h[t-1]}_{jk}} = \begin{bmatrix}

```
\frac{\partial{h[t]_{1}}}{\partial{h[t-1]_{1}}} & \frac{\parti
al{h[t]_{1}}}{\partial{h[t-1]_{2}}} & ... &  \frac{\partial{h[
t]_{1}}}{\partial{h[t-1]_{n}}} \\
\\
\frac{\partial{h[t]_{2}}}{\partial{h[t-1]_{1}}} & \frac{\parti
al{h[t]_{2}}}{\partial{h[t-1]_{2}}} & ... & \frac{\partial{h[t
]_{2}}}{\partial{h[t-1]_{n}}}\\
\\
... & ... & ... & .... \\
\\
\frac{\partial{h[t]_{n}}}{\partial{h[t-1]_{1}}} & \frac{\parti
al{h[t]_{n}}}{\partial{h[t-1]_{2}}} & ... &  \frac{\partial{h[
t]_{n}}}{\partial{h[t-1]_{n}}}


\end{bmatrix}\end{equation}
```

To help understand this calculus, let's create a derived experiment. Let Q_1, K_1, V_1 equal the Id_3matrix. Let h[t-1], $v_0[t]$and $v_2[t]$be vectors $\overrightarrow{\backslash\mathrm{bold}1} \in R^3$and let $w_0, w_2 = 3$

$$q_1[t], k_1[t], v_1[t] = \overrightarrow{\backslash\mathrm{bold}1} \tag{47}$$

$$w_1[t] = 3 \tag{48}$$

$$\alpha_0, \alpha_1, \alpha_2 = .33 \tag{49}$$

$$h[t] = \alpha_0 * v_{0,1} + \alpha_1 * v_{1,1} + \alpha_2 * v_{2,1} \tag{50}$$

$$h[t] = .33 * \overrightarrow{\backslash\mathrm{bold}1} + .33 * \overrightarrow{\backslash\mathrm{bold}1} + .33 * \overrightarrow{\backslash\mathrm{bold}1} \tag{51}$$

$$h[t] = \overrightarrow{\backslash\mathrm{bold}1} \tag{52}$$

Now, imagine we change the first element of h[t-1] to a 2, i.e:

$$h[t-1] = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} \tag{53}$$

$$q_1[t], k_1[t], v_1[t] = h[t-1] \tag{54}$$

$w_1[t] = 6\begin{equation}

$v_{0,1}$denotes the $v_0$vector's first index

Looking at the above, the only terms that depends on h[t-1] are α_1 and $v_{1,1}$ \end{equation}\alpha_0, \alpha_2 = .045

(55)

\alpha_1 = .91

(56)

h[t] = .045 *\vec{\bold{1}} + .91 *v_1[t] + .045 *\vec{\bold{1}}

(57)

h[t] =

$$\begin{bmatrix}1.91\\1\\1\end{bmatrix}$$

\begin{equation}

Therefore, we can **clearly** observe that when we change h[t-1], we change every α term, impacting h[t].

Expressing this in more generalizable terms:

\end{equation}\frac{\partial{\mathcal{h[t]}}}{\partial{h[t-1]}} = \frac{partial{h[t]}}{\frac{partial{\alpha_0}}} + {\frac{partial{\alpha_1}}}

6.19 Section 1.2.2.F

Suppose we know the derivative $\frac{\partial h[t]}{\partial h[T]}$for all t > T. Please write down the expression for $\frac{\partial l}{\partial h[T]}$ for AttentionRNN(k)

6.20 1.2.2.F Answer

6.21 Section 1.3

6.22 1.3.1

What caused the spikes on the left?

6.23 1.3.1 Answer

6.24 1.3.2

How can they be higher than the initial value of the loss?

6.25 1.3.2 Answer

6.26 1.3.3

What are some ways to fix them?

6.27 1.3.3 Answer

6.28 1.3.4

Explain why the loss and accuracy are at these values before training starts. You mayn eed to check the task definition in the notebook.

6.29 1.3.4 Answer

Type *Markdown* and LaTeX: α^2

What is the dimension of the output $f_W(x)$? Provide an expression for the value of elements of the convolutional layer output $f_W(x)$. Example answer format here and in the following problems

$$f_W(x) \in R^{42X42X42}, f_W(x)[i, j, k] = 42 \tag{58}$$

6.29.1 1.1.C.I Answer

$$f_W(x) \in R^{5X3} \tag{59}$$

Thinking about the first convolution, we take a 3 unit interval of the first channel's input to get a vector ∈ R^{1x3}. We take the dot product with W_1 a vector ∈ R^{3x1} and perform the dot product giving us a scalar value.

We can represent all 5 channels by taking the first 3 units of all 5 channels, giving us a data matrix ∈ R^{5X3} and perform the dot product of each channel with the corresponding weight, matrix each with dimmension ∈ R^{3x1} giving us a resulting vector ∈ R^{5x1}

Because there is a stride of 2, we then take inputs x[3],x[4],x[5] and repeat this process giving us another R^{5x1}, and lastly we repeat this operation one last time using inputx x[5],x[6],x[7] to get our third vecotr ∈ R^{5x1}.

Visualizing this, using python indexing notation of a matrix X[i,j] with indicies ranging from 0 $\leq i \leq 7$ and 0 $\leq j \leq 5$

$$f_W(x) =< x[2j : 2j + 3][i], W[i] > \tag{60}$$

Re-express this to be a little nicer