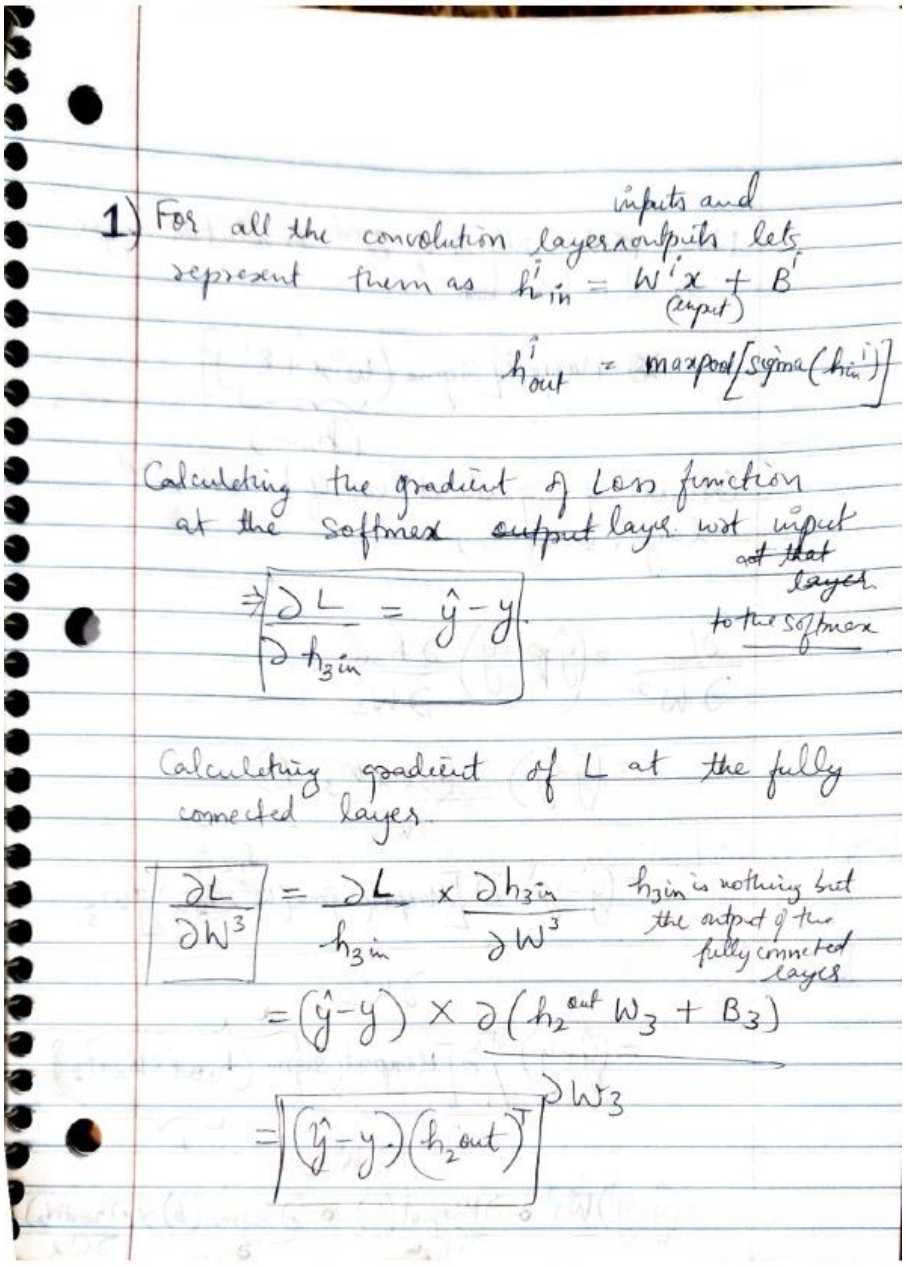
ECE 792 HW 3
Kriti Singh(ksingh23)

1.



1) For all the convolution layers inputs and outputs lets represent them as $h^i_{in} = W^i x + B^i$ (input)

$h^i_{out} = \text{maxpool}[\text{sigma}(h^i_{in})]$

Calculating the gradient of Loss function at the softmax ~~output~~ layer wrt input ~~at that layer~~ to the softmax

$$\Rightarrow \boxed{\frac{\partial L}{\partial h_{3in}} = \hat{y} - y}$$

Calculating gradient of L at the fully connected layers.

$$\boxed{\frac{\partial L}{\partial W^3}} = \frac{\partial L}{h_{3in}} \times \frac{\partial h_{3in}}{\partial W^3}$$

h_{3in} is nothing but the output of the fully connected layers

$$= (\hat{y} - y) \times \partial (h_2^{out} W_3 + B_3)$$

$$= \boxed{(\hat{y} - y).(h_2 out)^T} \partial W_3$$

Now output of 2nd convolution layer h_2 out is

$$\text{Maxpool}\left[\text{Sigma}\left(\underbrace{W^2 x + B^2}_{(h_2 in)}\right)\right]$$

Calculating gradient of L wrt 2nd layer weights,

$$\frac{\partial L}{\partial W^2} = (\hat{y} - y)\frac{\partial h_{3in}}{\partial W_2}$$

$$= (\hat{y} - y)\frac{\partial (h_{2out} W_3 + B_3)}{\partial W_2}$$

$$= (\hat{y} - y)\frac{\partial [\text{Maxpool}(\text{sigma}(\overset{h_{2in}}{\cancel{W \times h_{in}}})]W_3}{\partial W_2}$$

$$= (\hat{y} - y)\left[\frac{\partial [\text{Maxpool}(\text{sigma}(\overbrace{h_{1out} \times W_2 + B_2}^{a}))]W_3}{\partial W_2}\right]$$

(b = sigma argument)

$$= (\hat{y} - y) W_3^T \circ \frac{\partial \text{Maxpool}(a)}{\partial W_2} \circ \frac{\partial \text{sigma}(b)}{\partial W_2} \times \frac{\partial (h_{out} W_2)}{\partial W_2}$$

Here '$\odot$' is element wise product

$$\frac{\partial L}{\partial W_2} = (h_1 out)^T \cdot \underbrace{(\hat{y}-y) \cdot W_3^T \cdot \frac{\partial maxpool(a)}{\partial W_2} \cdot \frac{\partial sigma(b)}{\partial W_2}}_{\text{let us call this } z_2}$$

Gradient of Loss wrt layer 1.

$$\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial h_{3in}} \times \frac{\partial h_{3in}}{\partial W_1}$$

$$= \underbrace{(\hat{y}-y) \cdot W_3^T \odot \frac{\partial [max(sigma(h_1^{out} W_2 + B_2))]}{\partial W^2}}_{z_2} \times \frac{\partial (h_1 out W_2 + B_2)}{\partial W_1}$$

$$= z_2 \times \frac{\partial [max(sigma(\overbrace{h_0 out\, W_1 + B_1}^{b}))^{a} W_2]}{\partial W_1}$$

$$= z_2 \times h_0 out \times \frac{\partial max(a)}{\partial W_1} \times \frac{\partial sigma(b)}{\partial W_1}$$

$a = sigma(h_0 out\, W_1 + B_1)$ $\quad$ $b = h_0 out\, W_1 + B_1$

Here $h_0 out$ is nothing but the my input given at the zeroth layer that is the 28×28 image which we consider

x

gradient wrt Biases →

$$\frac{\partial L}{\partial B_3} = (\hat{y} - y) \times \frac{\partial h_{3in}}{\partial B_3}$$

$$= (\hat{y} - y) \times \frac{\partial h_2 out \cdot W_3 + B_3}{\partial B_3}$$

$$= (\hat{y} - y)$$

$$\frac{\partial L}{B^2} = (\hat{y} - y) \frac{\partial h_{3in}}{\partial B_2}$$

$$= (\hat{y} - y) \frac{\partial (h_2 out \cdot W_3 + B_3)}{\partial B_2}$$

$$= (\hat{y} - y) \times \frac{\partial [(\max(\text{sigma}(h_{in}))) W_3]}{\partial B_2}$$

$$= (\hat{y} - y) \times \frac{\partial \text{maxpool}(\text{sigma}(h_1 \text{out} \times W_2 + B_2))}{\partial B_2} \times \frac{\partial \text{sigma}(h_{\text{out}} W_2 + B_2)}{\partial W_2}$$

$$\times \frac{\partial (h_{1\text{out}} \cdot W_2 + B_2)}{\partial W_2} \quad [\text{chain rule}] \times W_3^T$$

$$\boxed{\frac{\partial L}{\partial B_2}} = (\hat{y} - y) \times \frac{\partial \text{maxpool}(a)}{\partial B_2} \times \frac{\partial \text{sigma}(b)}{\partial B_2} \times W_3^T$$

$$a = \text{sigma}[h_{1\text{out}} \times W_2 + B_2]$$

$$b = [h_{1\text{out}} \times W_2 + B_2]$$

$$\frac{\partial L}{\partial B_1} = (\hat{y} - y) \frac{\partial h_{3in}}{\partial B_1}$$

$$= [illegible]$$

$$= \frac{\partial L}{\partial B_2} \times W_2^T \times \frac{\partial [\text{maxpool}(\text{sigma}(h_{0\text{out}} W_1 + B_1))]}{\partial B_1}$$

2. We have, at the encoder side,

$$h(x) = a(wx+b)$$

$$\hat{x} = \hat{a}(w^* h(x) + c)$$

Loss function $\rightarrow L(x, \hat{x}) = ||x - \hat{x}||_2^2$

Now calculating the gradient of cost w.r.t weight parameters w and b,

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial \hat{x}} \times \frac{\partial \hat{x}}{\partial h} \times \frac{\partial h}{\partial w}$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \hat{x}} \times \frac{\partial \hat{x}}{\partial h} \times \frac{\partial h}{\partial b}$$

$$\frac{\partial L}{\partial \hat{x}} = 2(\hat{x} - x)$$

$$\frac{\partial \hat{x}}{\partial h} = \frac{\partial \hat{a}(w^* h(x) + c)}{\partial (w^* h + c)} + \frac{\partial (w^* h + c)}{\partial h}$$

$$= \hat{a}'(w^* h + c) \times w^*$$

$$\frac{\partial h}{\partial W} = \frac{\partial a(Wx+b)}{\partial (Wx+b)} \times \frac{\partial (Wx+b)}{\partial W}$$

$$= a'(Wx+b) \times * \; x^T$$

We know $W^* = W^T$

$$\frac{\partial L}{\partial W} = x^T \times a'(Wx+b) \times \hat{a}'(W^* \hat{x} + c) \times 2(\hat{x}-x)$$

$$\frac{\partial L}{\partial b} = a'(Wx+b) \times \hat{a}'(W^* \hat{x} + c) \times 2(\hat{x}-x)$$

The gradient of L wrt W and b are products of gradients of encoder and decoder.

So we can say gradient ~~with~~ of L wrt W and W^* is sum of ~~encoder~~ gradients of encoder and decoder.

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial W^*} = x^T \cdot a'(Wx+b) + \hat{a}'(W^* \hat{x} + c) \times 2(\hat{x}-x)$$

3). KL Divergence between two gaussian dist^r,

$$D_{KL}[N(\mu_0, \Sigma_0) \| N(\mu_1, \Sigma_1)] = \frac{1}{2}\left(tr(\Sigma_1^{-1} \times \Sigma_0) + (\mu_1 - \mu_0)^T (\Sigma_1^{-1})(\mu_1 - \mu_0) - k + \log\left(\frac{\det(\Sigma_1)}{\det(\Sigma_0)}\right)\right)$$

where k is dimension of Gaussian dist^r.

To show this, we see the definition of KL divergence,

$$D_{KL}[N(\mu_0, \Sigma_0) \| N(\mu_1, \Sigma_1)] = \int p(x) \log\left[\frac{p(x)}{q(x)}\right] dx$$

where $p(x)$ is true distr of $N(\mu_0, \Sigma_0)$ and $q(x)$ is approximate dist^r $N(\mu_1, \Sigma_1)$

$$D_{KL}[N(\mu_0, \Sigma_0) \| N(\mu_1, \Sigma_1)] = -\int p(x) \log\left[\frac{q(x)}{p(x)}\right] dx$$

we know, $q(x) = N(x | \mu_1, \Sigma_1) = \frac{1}{(2\pi)^{k/2} (\Sigma_1)^{1/2}} e^{-\frac{1}{2}\left[\frac{(x-\mu_1)^T (x-\mu_1)}{\Sigma_1}\right]}$

substituting this expression for $q(x)$,

$$D_{KL}[N(\mu_0, \Sigma_0) \| N(\mu_1, \Sigma_1)] = -\int p(x) \log \left[\frac{\frac{1}{(2\pi)^{k/2}(\Sigma_1)^{1/2}} e^{-1/2 \frac{(x-\mu_1)^T}{\Sigma_1}(x-\mu_1)}}{\frac{1}{(2\pi)^{k/2}(\Sigma_0)^{1/2}} e^{-1/2 \frac{(x-\mu_0)^T}{\Sigma_0}(x-\mu_0)}} \right] dx$$

Simplifying the above expressions,

$$D_{KL}[N(\mu_0, \Sigma_0) \| N(\mu_1, \Sigma_1)] = \frac{1}{2}[tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T \Sigma^{-1}(\mu_1 - \mu_0) - k + \log\left(\frac{\det \Sigma_1}{\det \Sigma_0}\right)$$

$$\rightarrow D_{KL}[N(\mu_0, \Sigma_0) \| N(\mu_1, \Sigma_1)] = \frac{1}{2}[tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0) - k + \log\left(\frac{\det \Sigma_1}{\det \Sigma_0}\right)$$

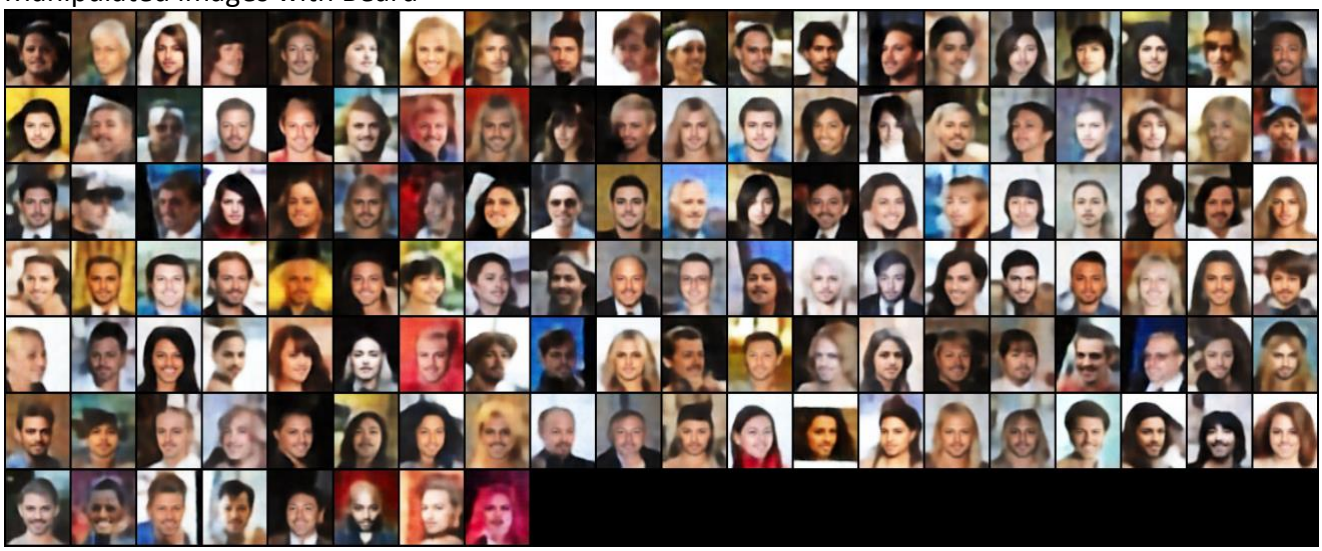
$$\rightarrow D_{KL}[N(\mu_1, \Sigma_1) \| N(\mu_0, \Sigma_0)] = \frac{1}{2}[tr(\Sigma_0^{-1}\Sigma_1) + (\mu_0 - \mu_1)^T \Sigma_0^{-1}(\mu_0 - \mu_1) - k + \log\left(\frac{\det \Sigma_0}{\det \Sigma_1}\right)]$$

We see that the two k dimensional distr is symmetric with only the subscripts of μ and Σ flipped.

4.Wrote the code for **TASK 1** and ran the model for **8 epochs** due to processing constraints. The batch size was **128** and learning rate was **0.001.**

TASK 2:

Manipulated images with Beard



Original Images

Original Images

Manipulated images for smiling faces

Manipulated Images for eyeglasses

Original Images

TASK 3:

Took 2 images and got the following interpolation for 10 samples.