AI501 Homework 4

20200130 Yujun Kim

May 2023

1 Exercise #8.2

The loss is cross entropy between two distribution. The sum disappears as we progress to the second line in the equation below because y_i are 0 except the index i corresponding to the expected word. The probability of getting correct word given continuous back of words is calculated as in the third line of the equation as the result pass softmax function at the last. The last line hold by basic property of log function.

$$\begin{aligned}
\mathcal{L}_\theta &= -\sum y_i \log p(x_i | x_k, x \in \mathcal{I}_C(i)) \\
&= -\log p(x_i | h_i) \\
&= -\log\left(\frac{exp(\tilde{w_{t_i}}^T h_i)}{\sum_k exp(\tilde{w_{t_k}}^T h_k)}\right) \\
&= -\tilde{w_{t_i}}^T h_i + \log\left(\sum_k exp(\tilde{w_{t_k}}^T h_k)\right)
\end{aligned}$$

2 Exercise #8.3

Let G be a simple connected graph with 5 vertices and 5 edges. A connected graph with v vertices and e edges is a tree if and only if $v = e + 1$. A connected graph contains a single cycle if and only if $v = e$. This is because for a connected graph with $v = e$, the graph is not a tree and thus, we can remove an edge while keeping it connected. Then the new graph is tree as the number of vertices becomes the number of edges plus 1. The original graph contains a cycle that contains an edge that we have removed.

As $e = v = 5$ for our case, the connected graph should contain a single cycle. The size of cycle is either 3, 4, or 5. Figure 1 shows the possible five simple connected graph with 5 edges and 5 vertices up to isomorphism. It is clear that within fixed cycle size(3, 4, 5), those are the only possible graphs. Now, we show that those five graphs are not isomorphic. Graph with different cycle size are not isomorphic. Thus, we only need to show the first three graphs are not isomorphic. The first graph has two vertices with degree 3, while the second graph has no vertices with degree 3 and the last graph has one vertices with degree 3. Hence, those three graphs are not isomorphic.

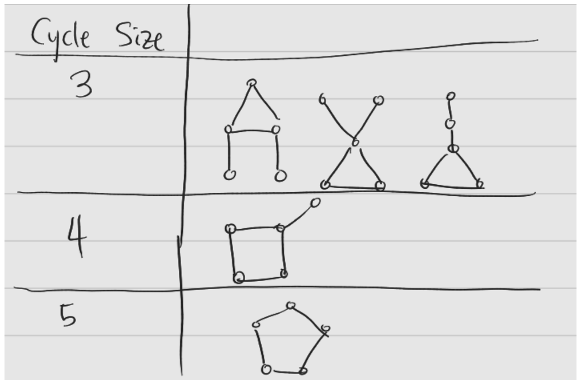

Figure 1: Possible Five Simple Connected Graph up to Isomorphism

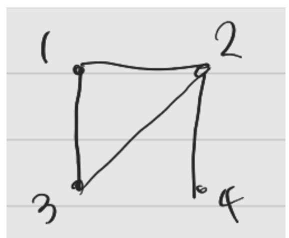

Figure 2: Graph Corresponding to A

3 Exercise #8.4

For fixed number of vertices, tree are graphs that use minimum number of edges while keeping the graph connected. Hence, for fixed number of edges, trees are the graphs that use maximum vertices while keeping the graph connected. If a connected graph has e edges and v vertices, then $v \leq e+1$, as trees satisfy $v = e+1$.

Let G be a graph with 20 edges and 4 component. Let e_i be the number of edges in $i-th$ component. Then, for number of vertices in $i-th$ component v_i, we have $v_i \leq e_i + 1$. Hence, $\sum v_i \leq \sum(e_i+1) = (\sum e_i) + 4 = 24$. The equality is achieved when each component of G is a tree. Hence, maximum possible number of vertices in G is 24.

4 Exercise #8.5

(a) The graph looks like figure 2.

(b) $\epsilon = 0.1$. $x_v^0 = e_v$ for $v = 1,2,3,4$. Let $z_v^k = (1+\epsilon)x_v^k + \sum_{u \in \mathcal{N}(v)} x_u^k$. Then let $Z^k = [z_1^k z_2^k z_3^k z_4^k]$. Note that $x_v^k = MLP^k(z_v^{k-1})$ and $MLP^k(x) = \sigma(W^2\sigma(W^1x))$, where σ is a ReLU. Through MATLAB computation, we obtain

X^1, X^2 as below.

$$Z^0 = \begin{bmatrix} 1.1 & 1 & 1 & 0 \\ 1 & 1.1 & 1 & 1 \\ 1 & 1 & 1.1 & 0 \\ 0 & 1 & 0 & 1.1 \end{bmatrix}, X^1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.123 \\ 0.182 & 0.229 & 0.172 & 0.307 \\ 0.182 & 0.229 & 0.172 & 0.115 \end{bmatrix}$$

$$Z^1 = \begin{bmatrix} 0.007 & 0.02 & 0.007 & 0.020 \\ 0.005 & 0.061 & 0.048 & 0.036 \\ 0.127 & 0.137 & 0.127 & 0.040 \\ 0.027 & 0.041 & 0.027 & 0.040 \end{bmatrix}, X^2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0.123 & 0 & 0.135 \\ 0.601 & 0.913 & 0.6 & 0.567 \\ 0.601 & 0.721 & 0.6 & 0.356 \end{bmatrix}$$

5 Exercise #9.2

(a) The mean μ is 0.083 and the standard deviation σ is 0.860. Then, the layer normalization is

$$Y = \frac{\gamma}{\sigma}(X - \mu\mathbb{1}) + \beta\mathbb{1}$$

Simply taking $\gamma = 1, \beta = 0$ gives

$$Y = \begin{bmatrix} 1.07 & 2.23 & 3.39 \\ -1.26 & -3.58 & -0.10 \\ 5.71 & -2.42 & 1.07 \\ -0.10 & -0.10 & -5.91 \end{bmatrix}$$

(b) Now we perform instance normalization. MATLAB calculates the means and standard deviation column-wise.

$$\mu = \begin{bmatrix} 1.25 \\ -0.75 \\ -0.25 \end{bmatrix}, \sigma = \begin{bmatrix} 2.63 \\ 2.22 \\ 3.40 \end{bmatrix}$$

Then,

$$y_c = \frac{\gamma_c}{\sigma_c}(x_c - \mu_c\mathbb{1}) + \beta_c\mathbb{1}$$

Taking simply $\gamma = 1, \beta = 0$ gives

$$Y = \begin{bmatrix} -0.095 & 1.24 & 0.955 \\ -0.856 & -1.015 & 0.074 \\ 1.426 & -0.564 & 0.367 \\ -0.475 & 0.338 & -1.40 \end{bmatrix}$$

6 Exercise #9.4

(a) Use $Q = X * W_Q, K = X * W_K$ to calculate key and query. After calculating QK^T, divide each row by the sum of each row. This results

$$A = \begin{bmatrix} -3.4 & 0.025 & -1.25 & 5.625 \\ -3.512 & -0.030 & -1.33 & 5.87 \\ -3.09 & 0.176 & -1.03 & 4.95 \\ NaN & NaN & NaN & NaN \end{bmatrix}$$

Note that NaN appears as the last row of Q is zero.

(b)

$$Y = AX = \begin{bmatrix} -9.68 & -4.38 & -39.6 \\ -10.1 & -4.27 & -41.2 \\ -9.41 & -4.66 & -35.0 \\ NaN & NaN & NaN \end{bmatrix}$$

(c) Using masked attention, we have

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.992 & 0.998 & 0 & 0 \\ 0.784 & -0.045 & 0.260 & 0 \\ NaN & NaN & NaN & NaN \end{bmatrix}$$

7 Exercise #9.5

By simple MATLAB calculation, we have positional encoding vector

$$p_n = \begin{bmatrix} 0.9999 \\ -0.0141 \\ 0.2486 \\ 0.9686 \\ 0.0398 \\ 0.9992 \\ 0.0063 \\ 1.0000 \\ 0.0010 \\ 1.0000 \end{bmatrix}$$

8 Exercise #9.7

The matrix formulation of AdaIN and noise for styleGAN is

$$Y = XT_xT_s + B_{x,s} + \epsilon$$

where

$$T_x = \begin{bmatrix} 1/\sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_C \end{bmatrix} \in \mathbb{R}^{C\times C}, T_s = \begin{bmatrix} \gamma_1^s & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \gamma_C^s \end{bmatrix} \in \mathbb{R}^{C\times C}$$

and

$$B_{x,s} = [\mathbb{1} \cdots \mathbb{1}] \begin{bmatrix} \beta_1^s - \frac{\gamma_1^s}{\sigma_1}\mu_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \beta_C^s - \frac{\gamma_C^s}{\sigma_C}\mu_C \end{bmatrix}$$

and ϵ is a zero-mean error(e.g. Gaussian)