Name Date 11/08/2025 07:32

$emb = C[X_b]$

$X_b = n \times k$ : $n$ : no of samples

$k$ : context size

$m$ : no of features

$V$ : Vocabulary size

$C : V \times m$

$emb : n \times k \times m$

$embcat : n \times (k \times m)$ convert 3D to 2D

$hprebn = embcat\ W_1 + b_1 : n \times n_1$

$W_1 : (k \times m, n_1)$ $b_1 : (n, n_1)$

$bmean_i = \frac{1}{n} \sum_{i=1}^{n} hprebn_{ij}$

$\begin{pmatrix} x_{11} & \cdots & x_{1n_1} \\ x_{n1} & \cdots & x_{nn_1} \end{pmatrix}$ ↓ mean across n samples

$bndiff = hpreb n - bnmean_i : n \times n_1$

$bndiff_2 = bndiff^2 : n \times n_1$

$bnvar = \frac{1}{n-1} \sum_{i=1}^{n} bndiff_{ij}^2 : 1 \times n_1$

$bnvar\_inv = \frac{1}{\sqrt{bnvar + \varepsilon}} : 1 \times n_1$

$\varepsilon$ is set at $1e^{-5}$

$bnraw = bndiff \times bnvar\ inv. \quad n \times n_1$

$hpreact = bngain \times bnraw + bnbias$

$n \times n_1 \qquad (\gamma) \quad n \times n_1 \qquad (\beta)$

$1 \times n_1 \qquad 1 \times n_1$

$h = Tanh(hpreact) : n \times n_1$

$logits = h W_2 + b_2 : n \times V$

$W_2 : n_1 \times V \qquad b_2 : n \times V$

Cross entropy losses : Note this is all element w'se operations.

$$\text{logit maxes} = \max_{j=1}^{V}(\text{logits}_{ij}) : n \times 1$$

$$\text{norm\_logits} = \text{logits} - \text{logit maxes} : n \times V$$

$$\text{counts} = \exp(\text{norm\_logits}) : n \times V$$

$$\text{count\_sum} = \sum_{j=1}^{V} \text{counts}_{ij} : n \times 1$$

$$\text{count\_sum\_inv} = \frac{1}{\text{count sum}} : n \times 1$$

→ implicit broadcasting, duplicate across all columns

$$\text{probs} = \text{counts} \times \text{counts sum inv} : n \times V$$

$$\text{logprobs} = \log(\text{probs}) : n \times V$$

$$L = -\frac{1}{n} \sum_{i=1}^{n} \text{logprobs}_{ij}$$

$j$ = index of 1 in each row of $Yb$

$$Yb = \begin{pmatrix} 0 & 1 & \cdots \\ 0 & 1 & 0 \\ & \vdots & \end{pmatrix} = n \times V$$

1 appears on the character output index

# backproping

$$dlogprobs = \frac{dL}{dlogprobs}$$

$$= \begin{pmatrix} 0 & \cdots & \frac{-1}{n} & \cdots & 0 \\ 0 & & \frac{-1}{n} & & 0 \\ 0 & & & & \frac{-1}{n} \end{pmatrix}$$

$n \times v$

$$dprobs = dlogprobs \times \frac{1}{probs} \quad : n \times v$$

(elementwise multiplication)

$$dcountssum_{inv} = \sum_{i=1}^{v} dprobs \times counts$$

(sum across colums)

$n \times 1$

reflect the broadasting operation

$d\text{counts} = d\text{probs} \times \text{counts\_sum\_inv}$

$n \times V \quad \times \quad n \times 1$

→ broadcasting

(counts is used multiple times
→ still missing some gradient)

$d\text{counts\_sum} = d\text{counts\_sum\_inv} \times \left(\frac{-1}{\text{countsum}^2}\right)$

$n \times 1 \qquad n \times 1$

$d\text{counts} = d\text{countsum} \times$

$n \times V \qquad n \times 1 \qquad (n \times V)$

$$\begin{pmatrix} 1 & \cdots & 1 \\ 1 & \cdots & 1 \\ 1 & \cdots & 1 \end{pmatrix}$$

reflect the summing up across colums

→ rebroadcasting back to $n \times V$

$$dnorm\_logits = dcounts \times \exp(norm\_logits)$$

$$dlogits\_maxes = -\sum_{j=1}^{V} dlogits_{ij}$$

$n \times 1$ $\quad$ $n \times V \rightarrow$ need to sum across columns.

$$dlogits = dnorm\_logits + dlogitmaxes \times max\_indices$$

$logits^{max} = \max($

$$\frac{dlogits\_maxes}{dlogits} = \begin{pmatrix} 0 & \cdots & 1 & \cdots & 0 \end{pmatrix}$$

↓

indices of the max

$$dh = dlogits \cdot W_2^T$$

$$n \times n_1 \quad \cdot \quad \frac{d\,logits}{dh}$$

$$n \times V \quad \times \quad V \times n_1$$

$$dh_{preact} = dh \times (1 - h^2)$$

$$n \times n_1 \quad \text{(element wise)}$$

$$db_{ngain} = \sum_{i=1}^{n} dh_{preact_{ij}} \; bn_{raw_{ij}}$$

$$1 \times n_1$$

$$db_{nraw} = dh_{preat} \times b_{ngain}$$

$$n \times n_1 \qquad n \times n_1 \quad 1 \times n_1$$

(broad casting)

$$bnvar = \frac{1}{n-1} \sum_{i=1}^{n} bndiff2_{ij}$$

sum across row

$$dbndiff = dbnvar \times \frac{1}{n-1} bn$$

$$(x_t - \bar{x})^2 = (x_t - \mu)^2 - 2(\bar{x} - \mu)(x_t - \mu) + (\bar{x} - \mu)^2$$

$$\sum_{t=1}^{T} (x_t - \bar{x})^2$$

$$= \sum_{t=1}^{T} (x_t - \mu)^2 - 2(\bar{x} - \mu) \sum_{t=1}^{T} (x_t - \mu) + \sum_{t=1}^{T} (\bar{x} - \mu)^2$$

$$= \sum_{i=1}^{n} (x_i - \mu)^2$$

$$- 2n(\bar{x} - \mu)^2 + n(\bar{x} - \mu)^2$$

$$= \sum_{i=1}^{n} (x_i - \mu)^2 - n(\bar{x} - \mu)^2$$

$$E\left[\sum (\bar{x} - \mu)^2\right]$$

$$= n\left(\text{Var}(\bar{x} - \mu)\right)$$

$$= n\left(\text{Var}(\bar{x})\right)$$

$$= n \,\text{Var}\left(\frac{x_1 + y_2 + \cdots x_n}{n}\right)$$

$$= \frac{\sigma^2 n}{n^2} \cdot n$$

$$bnvar = \frac{1}{n-1} \sum_{i=1}^{n} bndiff2_{ij}$$

$j: 1 \to n_1$

$$= \frac{1}{n-1} \begin{pmatrix} bndiff2_{11} + bndiff2_{21} \\ bndiff2_{12} + bndiff2_{22} \end{pmatrix}^T$$

$$\rightarrow \frac{dbnvar}{dbndiff2} = \frac{1}{n-1} \times \begin{pmatrix} 1 & \cdots & 1 \\ 1 & \cdots & 1 \end{pmatrix}$$

$32 \times 64$

( elementwise derivative)

$dhpreb = dbndiff.clone()$

important otherwise it will change the variables.

demb = dembcat.view(emb.shape)

→ just need to undo the operation.

emb = C[X]

n x m x h  V x m  n x h

C is a look up table for characters in X

→ demb[i,j]

↓

multiple occurences of the characters

dC_ij +=

Cross entropy backward loss

$$loss = \frac{-1}{n} \sum_{i=1}^{n} \log \frac{\exp(z_{i y_i})}{\sum_{j=1}^{V} \exp(z_{ij})}$$

$$logit_{iy} = z_{i y_i} \quad i : 1 \to N$$

$$y_i \in Y_b$$

$$\begin{pmatrix} z_{11} & \cdots & z_{1 y_1} & \cdots & z_{1 27} \\ \vdots & & & & \\ z_{n1} & \cdots & z_{n y_n} & \cdots & z_{n 27} \end{pmatrix} \xrightarrow{softmax} \begin{pmatrix} p_{11} & \cdots & p_{1 y_1} & \cdots & p_{1 27} \\ \vdots & & & & \\ p_{n1} & \cdots & p_{n y_n} & \cdots & p_{n 27} \end{pmatrix}$$

$$\xrightarrow{-\log} \begin{bmatrix} -\log p_{11} & \cdots & -\log p_{1 y_1} & \cdots & -\log p_{1 27} \\ \vdots & & & & \\ -\log p_{n1} & \cdots & -\log p_{n y_n} & \cdots & -\log p_{n 27} \end{bmatrix}$$

$$\frac{\partial loss}{\partial z_{ij}} = \frac{-1}{n} \frac{\partial}{\partial z_{ij}} -\log p_{i y_i}$$

$$p_{i y_i} = \frac{\exp(z_{i y_i})}{\sum_{j=1}^{V} \exp(z_{i y_i})}$$

Case1:

$$j = y_i = \frac{\partial}{\partial z_{ij}} \left[ -\log P_{ij} \right]$$

$$= -\frac{1}{P_{ij}} \frac{\exp(z_{ij}) \left( \sum_{j=1}^{V} \exp(z_{ij}) - \exp(z_{ij}) \right)}{\left( \sum_{j=1}^{V} \exp(z_{ij}) \right)^2}$$

$$= P_{ij} - 1 = P_{iy_i} - 1$$

Case 2:

$$j \neq y_i \quad \frac{\partial}{\partial z_{ij}} \left[ -\log P_{iy_i} \right]$$

$$= \frac{-1}{P_{iy_i}} \frac{-\exp(z_{iy_i})^2}{\left[ \sum_{j=1} \exp(z_{ij})^2 \right]}$$

$$= P_{iy_i}$$

$y_i$ index

$$\rightarrow \frac{\partial \text{loss}}{\partial z_{ij}} = \frac{1}{n} \left( P_{iy_i} \cdots P_{iy_i - 1} \cdots P_{iy_i} \right)$$

$\Rightarrow$ dlogits

$$= \frac{1}{n} \begin{pmatrix} p_{1y_1} & \cdots & p_{1y_1}-1 & \cdots & p_{1z_{y_1}} \\ \vdots & & & & \\ p_{ny_n} & & p_{n\cdot y_n}-1 & \cdots & p_{ni_{y_n}} \end{pmatrix}$$