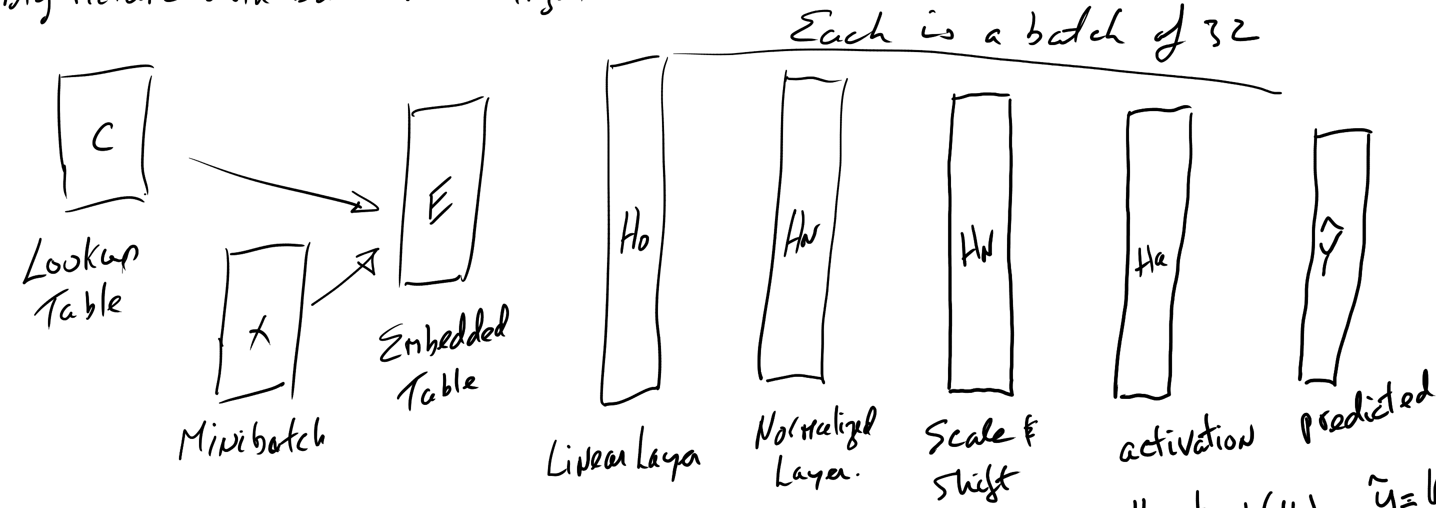
Big Picture with Batch Normalization



$Ha = \tanh(Ha)$ $\qquad$ $\tilde{y} = W_2 Ha + b_2$

$H_0 = W_1 \cdot E + b_1$ $\qquad$ $H_N = \dfrac{H_0 - \mu}{\sigma}$ $\qquad$ $H_N = \gamma H_N + \beta$

Start with the lookup table - C - Batch size × Length of Alphabet
- Initially Random. 32 27

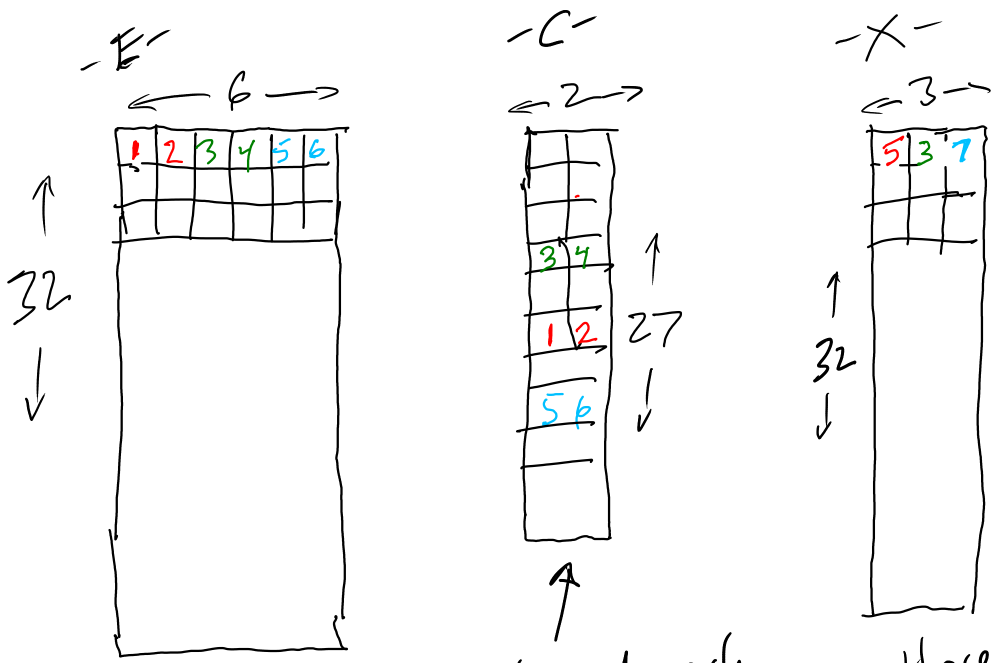
Get a Random batch of bigrams - X - Batch size × Blocksize
32 × 3

Create Embedded table -

E - Batch Size × Dimensions • Blocksize

$\underbrace{32 \times 2 \cdot 3}_{32 \times 6}$

E will now be our inputs to the Network.



Filled with the vectors from -C- in order given from -X-

think of each row as a 2D vector (x,y)

these bigrams determine the order of -C- vectors placed into the embedded table -E-

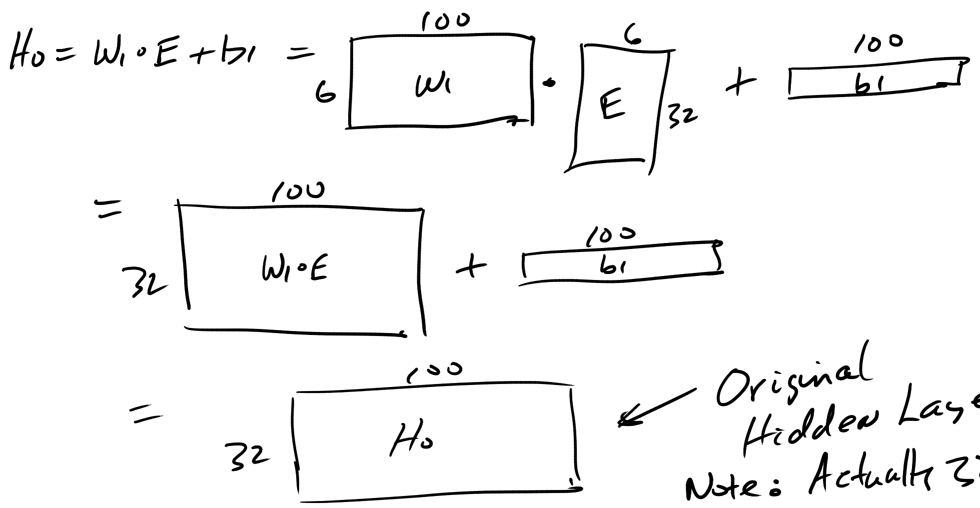
Start forward Pass -

100 hidden Neurons - Ho -

6×100 Weight 1 - W1 -

100 bias 1 - b1 -

Note: Batch Size = 32×6

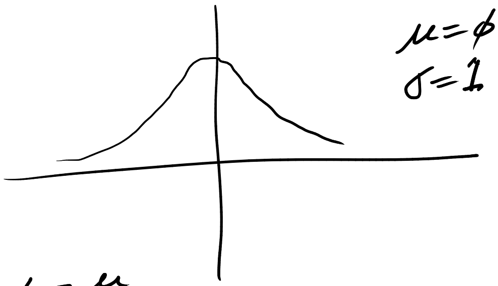


Original Hidden Layer.

Note: Actually 32 Hidden Layers.

We now want to normalize the entire batch ⇒ 32×100 Nodes

After normalization, the values in the hidden Layer will have a mean (μ) of zero & a standard deviation (σ) of one.



$\mu = \phi$
$\sigma = 1$

$$H_{Normal} = \frac{H_{original} - \mu}{\sigma}$$

$$\mu = \frac{1}{n}\sum_{1}^{n} H_o^i$$

$$\sigma^2 = \frac{1}{n}\sum_{1}^{n-1}\left(H_o^i - \mu\right)^2$$

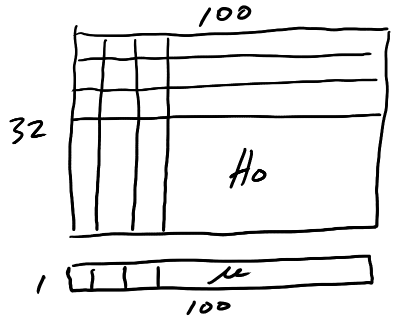
n-1 Bessel's Correction

$\sigma^2 \rightarrow$ Variance

$$H_N = \frac{H_o - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

10^{-5} (avoid ÷ zero)

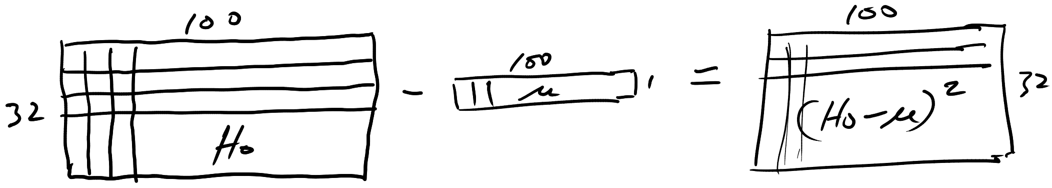
First calculate the mean (μ) for the batch



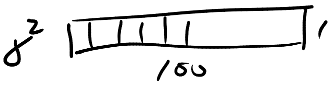
Sum the columns & divide by batch size (32)

$$\mu = \frac{1}{n}\sum_{1}^{n} H_o^i$$

Next calculate the variance of the batch



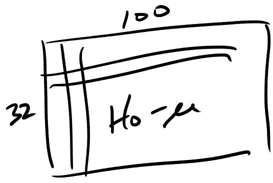
Subtract the mean from each row of the batch & square the result then add the columns & divide by batchsize.



$$\sigma^2 = \frac{1}{32}\sum_{1}^{32-1}\left(H_o^i - \mu\right)^2$$

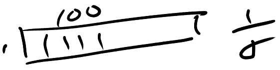
Finally Complete the Normalization

$$H_N = \frac{H_o - \mu}{\sqrt{\sigma^2 + \epsilon}} =$$



$\times \frac{1}{\sigma}$ (if $\epsilon = \phi$)

multiply each row by $\frac{1}{\sigma}$



Scale γ & Shift β

Not actually needed for Batch Normalization but used in most cases.

Allows the normalized values: $\mu = 0$ & $\sigma^2 = 1$ to be altered by these two parameters.

mean, variance

i.e $H_z = \gamma H_N + \beta \implies$ $\gamma \to$ scale (or gain)
$\beta \to$ shift (or bias)

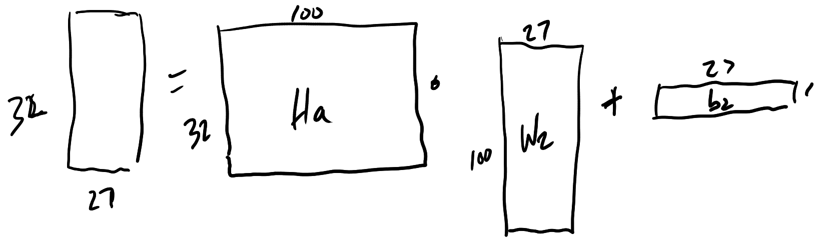
In our case both γ & β are vectors of 100 length.
And we set γ to all one's & β to all zeros

We can now implement the activation function — in our case it's $\tanh(x)$

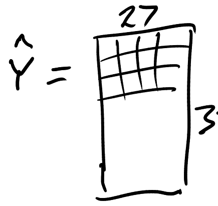
So $H_{activated} = H_a = \tanh(H_z)$

Finally we can compute the last linear layer of our network.

$\text{Logits} = W_2 \cdot H_a + b_2$ where W_2 is 100×27 & b_2 is 27



Use softmax to convert logcounts (Logits) to probabilities



$\implies$ probabilities.
- 32 Rows
- each row should sum to $\underline{\underline{1.00}}$

Now start back propogation;

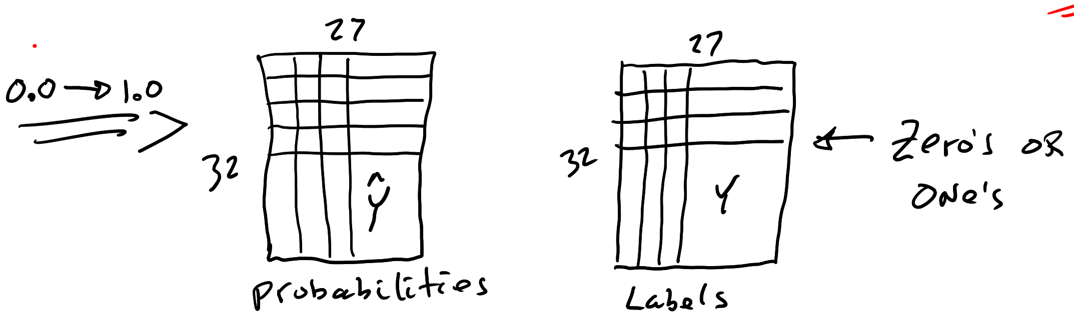
First we need to determine $\frac{\partial L}{\partial \hat{Y}}$ derivative of Loss w.r.t output - probabilities

$$L = \frac{1}{2}(\hat{Y} - Y)^2 = \frac{1}{2}(\hat{Y}^2 - 2\hat{Y}Y + Y^2)$$

$$\frac{\partial L}{\partial \hat{Y}} = \frac{1}{2}(2\hat{Y} - 2Y) = \underline{\underline{\hat{Y} - Y}}$$

probability (0-1)

actual (0-1) Labels



So now we have $\frac{\partial L}{\partial \hat{Y}}$ but we need the following:

$$\frac{\partial L}{\partial b_2}, \frac{\partial L}{\partial W_2}, \frac{\partial L}{\partial \beta}, \frac{\partial L}{\partial \gamma}, \frac{\partial L}{\partial b_1}, \frac{\partial L}{\partial W_1}, \frac{\partial L}{\partial C}$$

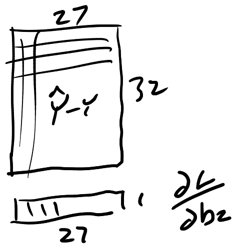
& these deltas will be used to update

$$b_2, W_2, \beta, \gamma, b_1, W_1, C$$

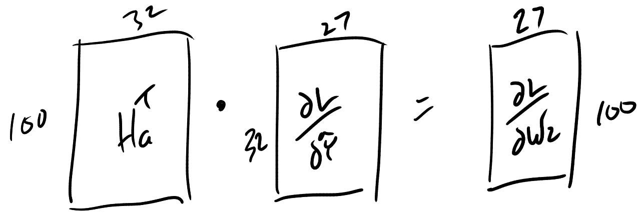
but to do this we'll also need several intermediate derivatives such as $\frac{\partial L}{\partial H_N}$, $\frac{\partial L}{\partial \sigma^2}$, $\frac{\partial L}{\partial \mu}$ & $\frac{\partial L}{\partial H_o}$

Start with $\frac{\partial L}{\partial b_2}$: $\hat{Y} = W_2 \cdot H_a + b_2 \Rightarrow \frac{\partial \hat{Y}}{\partial b_2} = 1$

so by chain rule: $\frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial b_2} = \frac{1}{n}\sum_{1}^{n}(\hat{Y}-Y) \cdot 1$



And $\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial W_2} = \underline{\underline{(\hat{Y}-Y) \cdot H_a}}$



Note: need to transpose $\underline{\underline{H_a}}$

Now need: $\frac{\partial L}{\partial \beta}$: $H_a = \tanh(H_z)$ $\quad \hat{Y} = W_2 H_a + b_2$

$H_z = \gamma H_N + \beta$

by chain Rule: $\frac{\partial L}{\partial \beta} = \underbrace{\frac{\partial L}{\partial \hat{Y}}}_{\hat{Y}-Y} \cdot \underbrace{\frac{\partial \hat{Y}}{\partial H_a}}_{W_2} \cdot \underbrace{\frac{\partial H_a}{\partial H_z}}_{1-\tanh^2(H_z)} \cdot \underbrace{\frac{\partial H_z}{\partial \beta}}_{1}$

and $\frac{\partial L}{\partial \gamma} = \underbrace{\frac{\partial L}{\partial \hat{Y}}}_{\hat{Y}-Y} \cdot \underbrace{\frac{\partial \hat{Y}}{\partial H_a}}_{W_2} \cdot \underbrace{\frac{\partial H_a}{\partial H_z}}_{1-\tanh^2(H_z)} \cdot \underbrace{\frac{\partial H_z}{\partial \gamma}}_{H_N}$

$H_z = \gamma H_N + \beta$

$H_N = \frac{H_o - \mu}{\sqrt{\sigma^2}}$

$H_o = E \cdot W_1 + b_1$

Next we need

$$\frac{\partial L}{\partial H_N} = \frac{\partial L}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial H_a} \cdot \frac{\partial H_a}{\partial H_z} \cdot \underbrace{\frac{\partial H_z}{\partial H_N}}_{\gamma}$$

Now $\frac{\partial L}{\partial \sigma^2} = \frac{\partial L}{\partial H_N} \cdot \frac{\partial H_N}{\partial \sigma^2} = -\frac{1}{2} \cdot (H_o - \mu)(\sigma^2 + \epsilon)^{-3/2}$

We also need $\frac{\partial L}{\partial \mu} = \frac{\partial L}{\partial H_N} \cdot \frac{\partial H_N}{\partial \mu} + \frac{\partial L}{\partial \sigma^2} \cdot \frac{\partial \sigma^2}{\partial \mu}$

$\frac{\partial H_N}{\partial \mu} = \frac{-1}{\sqrt{\sigma^2+\epsilon}}$

$\frac{\partial \sigma^2}{\partial \mu} = -2\frac{(H_0-\mu)}{n}$ from $\sigma^2 = \frac{1}{N}\sum(H_0-\mu)^2$

And we need $\frac{\partial L}{\partial H_0} = \frac{\partial L}{\partial H_N} \cdot \frac{\partial H_N}{\partial H_0} + \frac{\partial L}{\partial \sigma^2} \cdot \frac{\partial \sigma^2}{\partial H_0} + \frac{\partial L}{\partial \mu} \cdot \frac{\partial \mu}{\partial H_0}$

$\frac{\partial H_N}{\partial H_0} = \frac{1}{\sqrt{\sigma^2+\epsilon}}$, $\frac{\partial \sigma^2}{\partial H_0} = -2\frac{(H_0-\mu)}{n}$, $\frac{\partial \mu}{\partial H_0} = 1/n$ from $\mu = \frac{1}{N}\sum H_0$

So $\frac{\partial L}{\partial H_0} = \frac{\partial L}{\partial H_N} \cdot \frac{1}{\sqrt{\sigma^2+\epsilon}} + \frac{\partial L}{\partial \sigma^2} \cdot \frac{-2(H_0-\mu)}{n} + \frac{\partial L}{\partial \mu} \cdot \frac{1}{n}$

$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial H_0} \cdot \frac{\partial H_0}{\partial b_1}$, $\frac{\partial H_0}{\partial b_1} = 1$ from $H_0 = E \cdot W_1 + b_1$

& $\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial H_0} \cdot \frac{\partial H_0}{\partial W_1}$, $\frac{\partial H_0}{\partial W_1} = E$

then $\frac{\partial L}{\partial E} = \frac{\partial L}{\partial H_0} \cdot \frac{\partial H_0}{\partial E}$, $\frac{\partial H_0}{\partial E} = W_1$

Now use $\frac{\partial L}{\partial E}$ to unembed the gradients into C giving $\frac{\partial L}{\partial C}$

QED!

Proceed to update $C, W_1, b_1, W_2, b_2, \gamma, \beta$