Layer Normalization is just the “transpose of batch normalization”.

1. transpose the input, and set the factor of moving average of mean and std to 0. Then the calculations in this layer is the same as batch norm.
2. layer normalization dose not need to save the mean and std, and the training and testing process are the same.
3. It seems that many codes can be reused between batch normalization and layer normalization.

Forward Pass

- x_i is the vector representation of the summed inputs to the neurons in a layer

$$\mathbf{x} = \mathbf{W}^T \mathbf{h}$$

- H is number of neurons in the layer
- m is number of training samples in one mini-batch
- in mini-batch training, $\mathbf{x}$ is a matrix whose size is: $m \times H$
- In the forward pass, first compute the layer normalization statistics over the hidden units in the same layer:

$$\mu = \frac{1}{H}\sum_{i=1}^{H} x_i$$

$$\sigma^2 = \frac{1}{H}\sum_{i=1}^{H}(x_i - \mu)^2$$

- normalize the output:

$$\hat{\mathbf{x}} = \frac{\mathbf{x} - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

$$\mathbf{y} = \gamma \odot \hat{\mathbf{x}} + \beta$$

Note that both μ and σ^2 are matrix and they have the same size as $\hat{\mathbf{x}}_{m\times H}$

Some Notes:

- learnable parameters: γ and β
- all the hidden units in a layer share the same normalization terms
- different training sample have different normalization terms

Backward Pass

- the backward pass computes two things:
  1. partial derivatives of loss function with regard to learnable parameters: $\frac{\partial \mathcal{L}}{\partial \gamma}$ and $\frac{\partial \mathcal{L}}{\partial \beta}$
  2. partial derivatives of the loss function with regard to input: $\frac{\partial \mathcal{L}}{\partial \mathbf{x}}$

1. partial derivatives of $\mathcal{L}$ with respect to learnable parameters:

$$\frac{\partial \mathcal{L}}{\partial \gamma} = \sum_{i=1}^{m} \frac{\partial \mathcal{L}}{\partial y_i} \odot \hat{\mathbf{x}}_{\mathbf{i}} \tag{1}$$

$$\frac{\partial \mathcal{L}}{\partial \beta} = \sum_{i=1}^{m} \frac{\partial \mathcal{L}}{\partial y_i} \tag{2}$$

2. partiasl derivative of $\mathcal{L}$ with respect to input $\mathbf{x}$:

- $\mathcal{L}$ can be regarded as a function of $\mathcal{L}(\hat{\mathbf{x}}, \sigma^2, \mu)$
- $\hat{\mathbf{x}}, \sigma^2$, and μ are all functions of $\mathbf{x}$
- according to the chain rule, we can obtain the following formular:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}} = \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}} \frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{x}} + \frac{\partial \mathcal{L}}{\partial \mu} \frac{\partial \mu}{\partial \mathbf{x}} + \frac{\partial \mathcal{L}}{\partial \sigma^2} \frac{\partial \sigma^2}{\partial \mathbf{x}} \tag{3}$$

let's calculate $\frac{\partial \mathcal{L}}{\partial \mathbf{x}}$ step by step.

1. the first part:

$$\frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}} = \frac{\partial \mathcal{L}}{\partial y} \odot \gamma$$

$$\frac{\partial \hat{\mathbf{x}}}{\partial \mathbf{x}} = (\sigma^2 + \epsilon)^{-\frac{1}{2}}$$

2. the second part:

$$\frac{\partial \mathcal{L}}{\partial \mu} = \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}}\frac{\partial \hat{\mathbf{x}}}{\partial \mu} + \frac{\partial \mathcal{L}}{\partial \sigma^2}\frac{\partial \sigma^2}{\partial \mu} \tag{4}$$

$$\frac{\partial \hat{\mathbf{x}}}{\partial \mu} = -(\sigma^2 + \epsilon)^{-\frac{1}{2}} \tag{5}$$

$$\begin{aligned} \frac{\partial \sigma^2}{\partial \mu} &= \frac{-2}{H}\sum_{i=1}^{H}(x_i - \mu) \\ &= \frac{-2}{H}(\sum_{i=1}^{H} x_i - \sum_{i=1}^{H} \mu) \\ &= 0 \end{aligned} \tag{6}$$

- substitude (5), (6) into (4):

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mu_i} &= \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}}\frac{\partial \hat{\mathbf{x}}}{\partial \mu_i} \\ &= -(\sigma_i^2 + \epsilon)^{-\frac{1}{2}} \sum_{j=1}^{H} \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x_j}}} \end{aligned} \tag{7}$$

- $\frac{\partial \mathcal{L}}{\partial \mu}$ is a matrix shown as below, and equation (7) is one row of this matrix.

$$\begin{pmatrix} \frac{\partial \mathcal{L}}{\partial \mu_1} & \dots & \frac{\partial \mathcal{L}}{\partial \mu_1} \\ & \dots & \\ \frac{\partial \mathcal{L}}{\partial \mu_m} & \dots & \frac{\partial \mathcal{L}}{\partial \mu_m} \end{pmatrix}$$

- it is easy to get:

$$\frac{\partial \mu}{\partial \mathbf{x}} = \frac{\mathbf{1}}{\mathbf{H}} \tag{8}$$

3. the third part

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \sigma_i^2} &= \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{x}}} \frac{\partial \hat{\mathbf{x}}}{\partial \sigma_i^2} \\
&= -\frac{1}{2}(\sigma_i^2 + \epsilon)^{\frac{3}{2}} \sum_{j=1}^{H} \frac{\partial \mathcal{L}}{\partial x_j}(x_j - \mu)
\end{aligned} \tag{9}
$$

- $\frac{\partial \mathcal{L}}{\partial \sigma^2}$ is a matrix as shown below, and equation (9) is a row of this matrix.

$$
\begin{pmatrix}
\frac{\partial \mathcal{L}}{\partial \sigma_1^2} & \dots & \frac{\partial \mathcal{L}}{\partial \sigma_1^2} \\
 & \dots & \\
\frac{\partial \mathcal{L}}{\partial \sigma_m^2} & \dots & \frac{\partial \mathcal{L}}{\partial \sigma_m^2}
\end{pmatrix}
$$

- for $\frac{\partial \sigma^2}{\partial x}$:

$$
\frac{\partial \sigma^2}{\partial \mathbf{x}} = \frac{2}{H}(x_j - \mu), j = i \dots H \tag{10}
$$

- $\frac{\partial \sigma^2}{\partial x}$ is a matrix shown as below, and equation (10) is one row of this matrix:

$$
\begin{pmatrix}
\frac{\partial \sigma_1^2}{\partial x_{1,1}} & \dots & \frac{\partial \sigma_1^2}{\partial x_{1,H}} \\
 & \dots & \\
\frac{\partial \sigma_H^2}{\partial x_{m,1}} & \dots & \frac{\partial \sigma_H^2}{\partial x_{m,H}}
\end{pmatrix}
$$