Normalisations

In statistics, μ, σ is defined as below. Assume the input space as $\mathcal{X}$, and input samples $x \in \mathcal{X}$.

$$\mu = \frac{1}{|\mathcal{X}|}\sum_{i=1}^{|\mathcal{X}|} x_i = \mathbb{E}_{x\in\mathcal{X}}[x]$$
$$\sigma = \frac{1}{|\mathcal{X}|}\sum_{i=1}^{|\mathcal{X}|} \{x_i - \mu\} = \mathbb{E}_{x\in\mathcal{X}}[x-\mu]^2$$

A simple normalised $\tilde{x}$ can be calculated as:

$$\tilde{x} = \frac{x-\mu}{\sigma}$$

Batch Normalisation $BN_{\gamma,\beta}$ $[N, C, H, W]$

Consider input as $X\colon B \times d$ where B is batch size, d is feature dimension. In batch norm, x is normalised by each dimension. Note x's j dimension as $x^{(j)}$. We can write BN's normalisation as calculate each dimension's mean and variance then apply it to each dimension.

$$\mu_B^{(j)} = \mathbb{E}_{x\in\mathcal{B}}[x^{(j)}]$$
$$\sigma_B^{(i)} = \mathbb{E}_{x\in\mathcal{B}}[x^{(j)} - \mu^{(j)}]^2$$
$$\tilde{x}_i^{(j)} = \frac{x_i^{(j)} - \mu_B^{(j)}}{\sqrt{\left(\sigma_B^{(j)}\right)^2 + \epsilon}}$$

Where ϵ is a small number for numerical stability.

After normalisation, we do an affine transformation (also known as a scale) like what we have done in linear layer $(XW + b)$:

$$y_k^{(j)} = \gamma^{(j)}\tilde{x}_i^{(j)} + \beta^{(j)}$$

To be aware, γ, β are learnable vectors (i.e. they are the same shape as input features). Their initial values are 1 and 0 respectively.

In practice, we update the mean and variance via momentum. Momentum can be considered as an interpolation between old value and new value with weight λ. Update rule can be defined as:

$$v_{t+1} \leftarrow (1-\lambda)v_t + \lambda\tilde{v}_{t+1}$$

At here, let the historical mean and variance as μ_H, σ_H. During training, we may calculate the new mean and value $\tilde{\mu}, \tilde{\sigma}$. Then the updated mean and variance is

$$\mu \leftarrow (1 - \lambda)\mu_H + \lambda\tilde{\mu}$$
$$\sigma \leftarrow (1 - \lambda)\sigma_H + \lambda\tilde{\sigma}$$

Once we get the updated mean and variance, we use the new value to do normalisation. During inferencing stage, we freeze the parameter by let $\lambda = 0$.

In imaging field, we can consider $X: [N, C, H, W]$ where N is batch size, C is channel number, H, W are image's height and width. BN can be considered as normalised through channel. Therefore, mean and variance is calculated through $[N, H, W]$.

Its parameter number can be considering as 2 parts: $\mu, \sigma \in \mathbb{R}^C$, $\gamma, \beta \in \mathbb{R}^C$. μ, σ can be considered as statistics measures. The "actual" learning parameter is $\gamma, \beta \in \mathbb{R}^C$. So its parameter number is $2C$.

Layer Normalisation $LN_{\gamma,\beta}$ $[N, C, H, W]$

Different from the BN, LN do normalisation through each sample. That means to each image, it will be normalised through $[C, H, W]$. Consider mean and variance is calculated per sample, that means it will have the same behaviour during training and inferencing. Similar to BN, LN also introduces γ, β.

Its parameter number is $2C$ (γ, β)

Instance Normalisation $IN_{\gamma,\beta}$ $[N, C, H, W]$

IN calculates mean and variance per sample across channels. So, it will have the same behaviour during training and inferencing. Similar to BN and LN, IN also introduces γ, β.

Its parameter number is $2C$ (γ, β)