Batch Normalization

Related to input transformation. So that loss landscape looks smooth

Question: should we use moving avgs instead of current avgs for normalize

answered below

① $\text{mean} = \frac{1}{m}\sum A_i$ — all activations

$\text{Std} = \sqrt{\frac{1}{m}\sum (A_i - \mu)^2}$

Innovation: each batch norm layer allows for different mean and variance

② normalize

$\hat{A}_i = \frac{A_i - \mu_i}{\sigma_i}$

③ $BN_i = \gamma \odot \hat{A}_i + \beta$

learnable parameters

④ Exponential Moving Average

$\mu_{movi} = \alpha \mu_{movi} + (1-\alpha)\mu_i$

(EMA) $\sigma_{movi} = \alpha \sigma_{movi} + (1-\alpha)\sigma_i$

hyperparameter "momentum"

these parameters are not used during training they are only used for inference

Why EMA?

Ideally If I want averages of all activations I can keep track of entire data for every minibatch. But that's expensive. So a exponential moving average is a good approximation

Batch Normalization.

Related to input transformation. So that loss landscape looks smooth

Question
should we use moving avgs instead of current avgs for normalize

answered below

① $\text{mean} = \mu_i = \frac{1}{m}\sum A_i$ — all activations

$\text{std} = \sigma_i = \sqrt{\frac{1}{m}\sum (A_i - \mu)^2}$

② normalize

$\hat{A}_i = \frac{A_i - \mu_i}{\sigma_i}$

Innovation

each batch norm layer allows for different mean and variance

③ $BN_i = \gamma \odot \hat{A}_i + \beta$

learnable parameters

④ Exponential Moving Average (EMA)

$\mu_{movi} = \alpha \mu_{movi} + (1-\alpha)\mu_i$

$\sigma_{movi} = \alpha \sigma_{movi} + (1-\alpha)\sigma_i$

hyperparameter "momentum"

these parameters are not used during they are only used for inference

Why EMA?

Ideally If I want averages of all activations I can keep track of entire data for every minibatch. But that's expensive. So a exponential moving average is a good approximation