MOTIVATONS → REDUCE INTERNAL COVARIATE SHIFT

2 INPUTS TO A NETWORK VARY OVER TIME RESULTING IN CHANGES TO THE RANGE OF NETWORK ACTIVATION OUTPUTS

↓

THIS IS INTERNAL COVARIATE SHIFT

↓

ANSWE IS BATCH NORM

NORMALIZATION GENERALLY IS DONE TO INPUTS TAKE IT A STEP FURTHER AND DO IT TO EACH LAYERS INPUTS

# NORMALIZATION VIA MINI-BATCH STOCHASTIC GRADIENT DESCENT

- GET MEAN AND VARIANCE FOR ACTIVATIONS ONLY THROUGH THE ACTIVATIONS OF THAT MINIBATCH
  - AVOIDS GETTING COMPUTATIONALLY EXPENSIVE AND NON-CONVEX
- LEARNING PARAMETERS $\gamma$ & $\beta$ TO MAINTAIN A SPECIFIC REPRESENTATION FOR THE NETWORK

**Input:** Values of $x$ over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: $\gamma, \beta$
**Output:** $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

**Algorithm 1:** Batch Normalizing Transform, applied to activation $x$ over a mini-batch.

SHOULD HAVE DISTRIBUTION OF 0 AND VARIANCE OF 1

→ $Y = BN_{\gamma,\beta}(x)$

↓

DEPENDS ON EXAMPLE $x$ BUT ALSO ALL EXAMPLES IN THE MINI BATCH

$\sum_{i=1}^{m} \hat{x}_i = 0, \quad \frac{1}{m}\sum_{i=1}^{m} \hat{x}_i^2 = 1$

# TRAINING AND INFERENCE

NO BATCH NORM DURING INFERENCE

↳ CHANGES DISTRIBUTION OF INPUT WHICH IS BAD DURING INFERENCE

BN TRANSFORM GOES IMMEDIATELY BEFORE THE NONLINEARITY

IN CONVOLUTIONS $\gamma$ & $\beta$ ARE LEARNED PER FEATURE MAP RATHER THAN ACTIVATION

↳ ACTIVATION AS IN NODE