Batch Normalization

1. BN的动机：让 Error Surface 更好训练

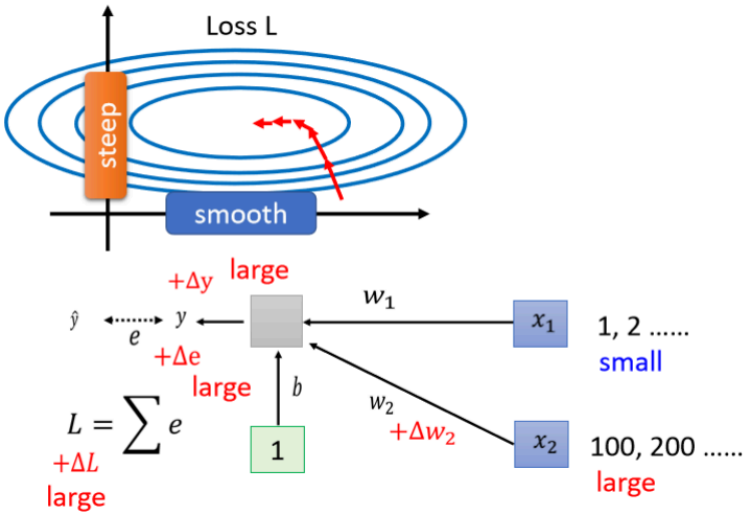
训练神经网络时，参数空间的不同方向斜率差异巨大，会造成：

- 某些方向梯度很小（走不动）
- 另一些方向梯度很大（震荡不收敛）

即使 error surface 是 convex，「不同方向的尺度不一致」也会造成训练困难。

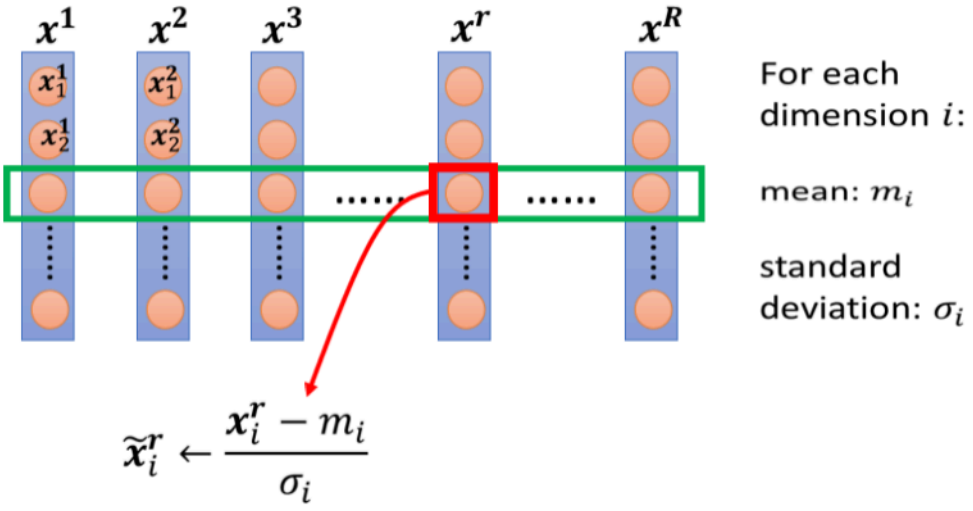
根本原因：

- **输入各维度的数值 scale 差异太大**



> 通过归一化 feature，使各维度尺度一致，让 error surface 更平滑，更适合优化

2. Feature Normalization（BN 的概念基础）



对训练数据每个维度 i，计算：

$$\mu_i = \mathbb{E}[x_i], \quad \sigma_i = \sqrt{\mathbb{E}[(x_i - \mu_i)^2]}$$

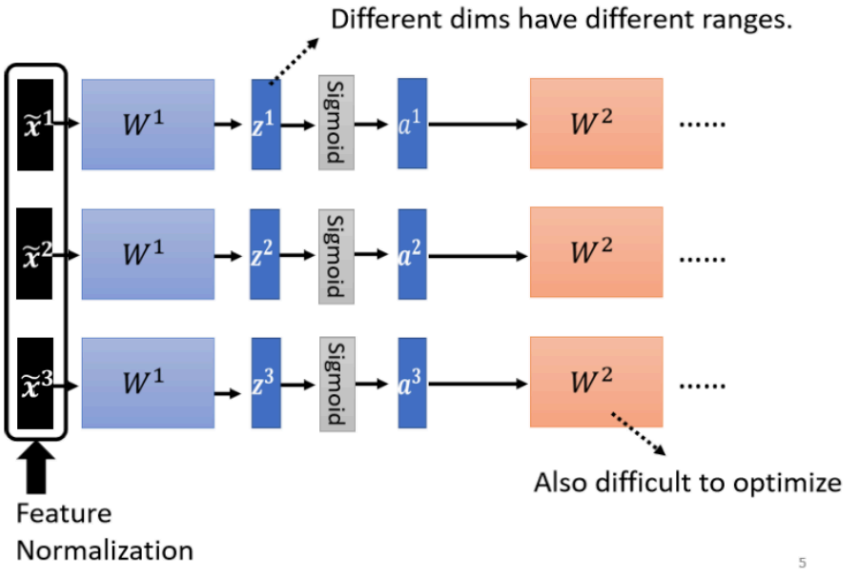
然后对每个样本做 standardization：

$$\tilde{x}_i = \frac{x_i - \mu_i}{\sigma_i}$$

这样：

- 每一维的 mean → 0、variance → 1
- 不同维度的scale被统一
- 训练更容易、收敛更快

3. 深度网络中的问题：中间层的z也需要 normalization

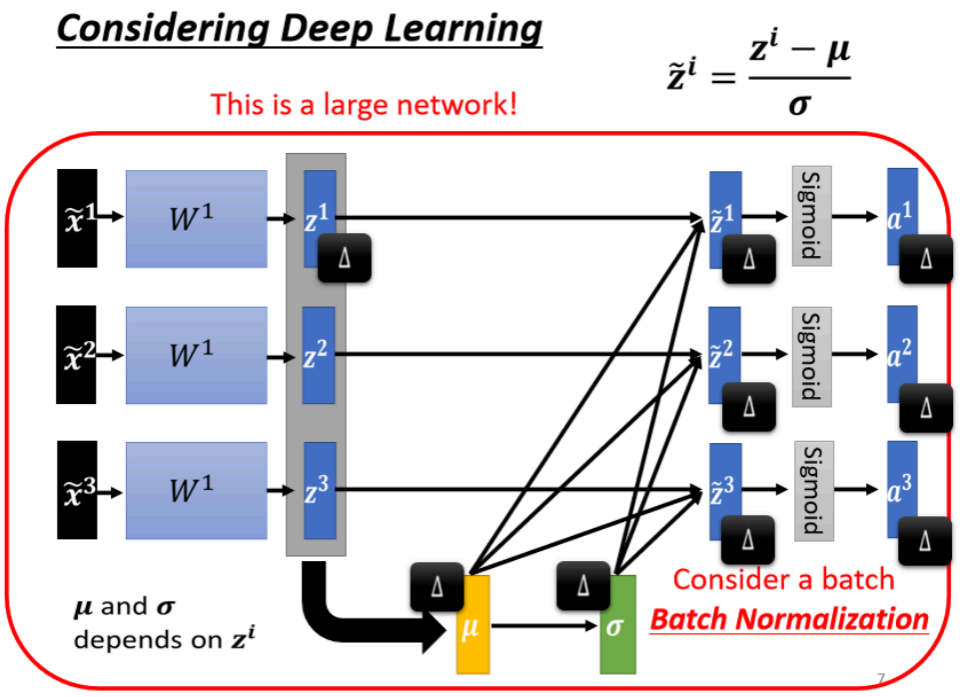


深度网络中：$z = Wx + b \quad \Rightarrow \quad a = \text{Activation}(z)$，即便输入 x 被 normalize，经过 W 后的 z 仍可能：

- 不同维度分布差异巨大
- 导致下一层训练困难

所以 Feature Normalization 应该放在：z（activation 前） 或 a（activation 后），一般选择在 **activation 前（对 z 做 BN）**

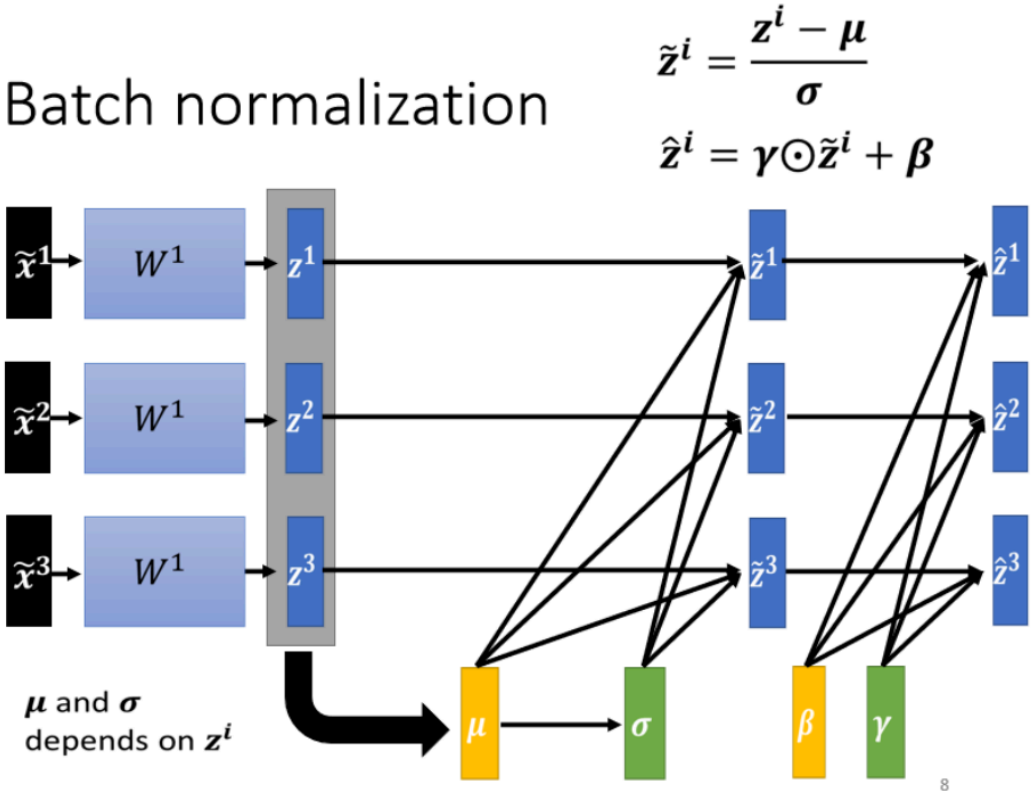
4. BN = Normalization + Learnable Shift & Scale



对一个 batch（大小=64，例如）中所有样本的同一维度，计算：$\mu = \frac{1}{m}\sum_{i=1}^{m} z^{(i)}$，$\sigma = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(z^{(i)} - \mu)^2}$，对每个样本：$\tilde{z}^{(i)} = \frac{z^{(i)} - \mu}{\sigma}$

这就是 **Batch Normalization** 名字的来源：

> 只利用当前 batch 的数据估计 mean/variance



- **Normalization 会强制 mean=0、std=1，但模型可能需要不同分布**

  所以引入 learnable 参数：$y = \gamma\tilde{z} + \beta$

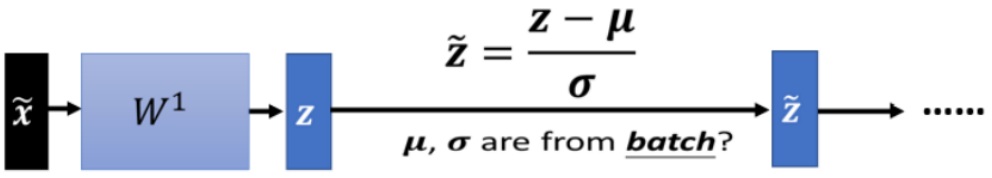
  其中：

  - γ 初值 = 1
  - β 初值 = 0

  确保初期模型行为与未使用 BN 相近。

  训练过程中 γ 与 β 会被学习，用于恢复或调整分布。

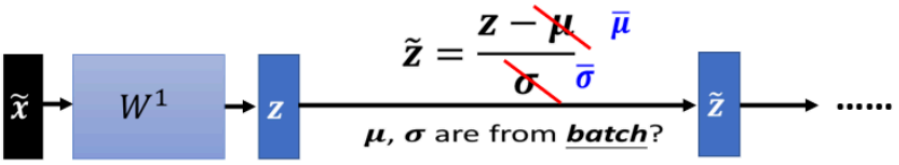
5. Testing（Inference）阶段如何处理 BN?



- 测试时：

- 不会等待“凑满一个 batch”再计算 mean & variance
- 必须独立处理每一笔样本

- 解决方式（PyTorch、TensorFlow 都是这样做的）：



We do not always have ***batch*** at testing stage.

Computing the moving average of μ and σ of the batches during training.

$\mu^1 \quad \mu^2 \quad \mu^3 \quad ..... \quad \mu^t$

$$\bar{\mu} \leftarrow p\bar{\mu} + (1-p)\mu^t$$

在训练过程中维持 moving average：

$\mu_{\text{running}} = p \cdot \mu_{\text{running}} + (1-p) \cdot \mu_{\text{batch}}$

$\sigma_{\text{running}} = p \cdot \sigma_{\text{running}} + (1-p) \cdot \sigma_{\text{batch}}$

测试阶段使用：

$\tilde{z} = \frac{z - \mu_{\text{running}}}{\sigma_{\text{running}}}$

> **测试时 BN 不依赖 batch，而是依赖训练时累积的 running mean / variance。**