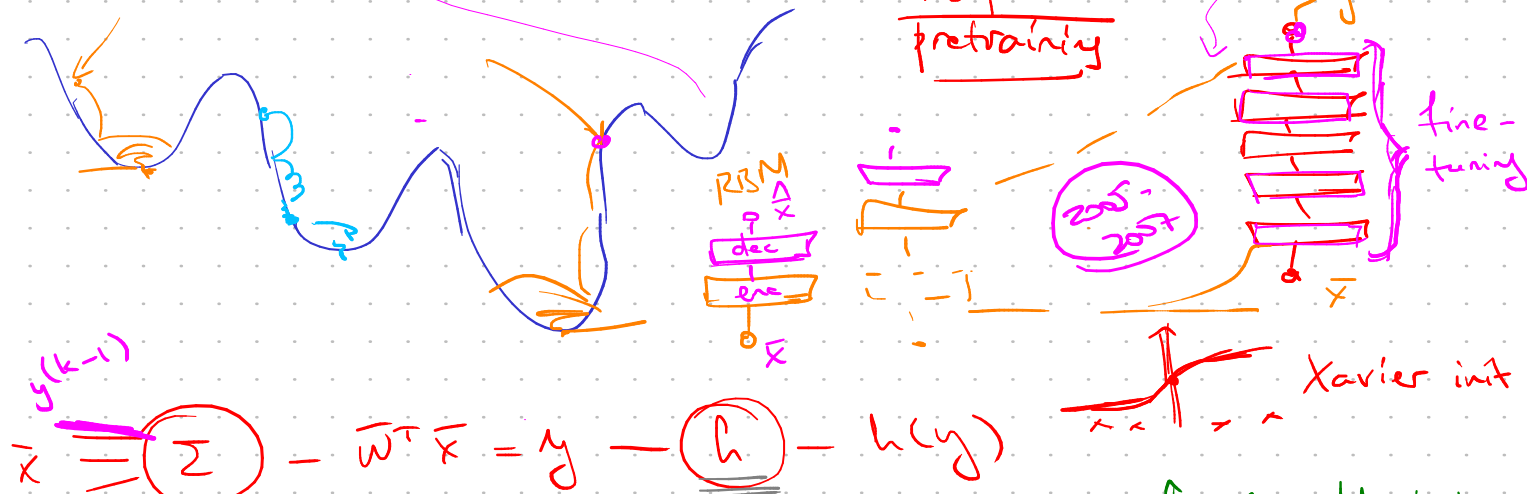


① Weight initialization



$$\bar{x} = \sum_{i=1}^n x_i - \bar{w}^T \bar{x} = y - \textcircled{h} - h(y)$$

$$y = \sum w_i x_i = \sum y_i$$

$$\text{Var } y = \sum \text{Var}[y_i] = \sum_{i=1}^n \text{Var}[w_i x_i]$$

$$\text{Var}[w_i x_i] = \underbrace{\mathbb{E}[x_i]^2}_{=0} \text{Var}[w_i] + \underbrace{\mathbb{E}[w_i]^2}_{=0} \text{Var}[x_i] + \text{Var}[w_i] \text{Var}[x_i]$$

$$\text{Var}[w_i x_i] = \text{Var}[w_i] \text{Var}[x_i]$$

$$\text{Var}[y] = \textcircled{n} \text{Var}[w_i] \cdot \text{Var}[x_i]$$

$\text{Var}[y^{(k+1)}]$
 $\text{Var}[y^{(k)}]$

$$\text{Var}[w_i] = O(1/n)$$

$$w_i \sim \text{Unif}\left(\left[-\frac{\sqrt{3}}{\sqrt{n}}, \frac{\sqrt{3}}{\sqrt{n}}\right]\right)$$

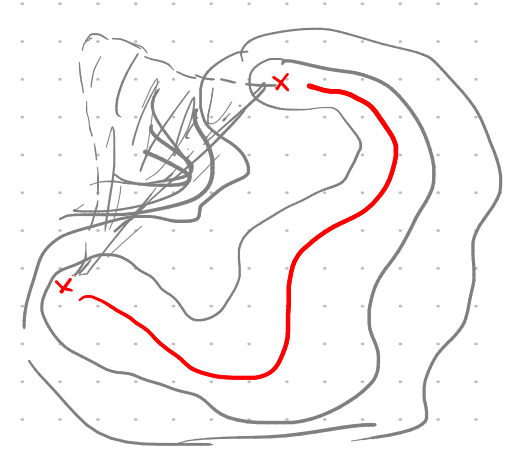
2010 Xavier Glorot
Y. Bengio

Xavier init

$$\text{Var}[w_i] = \frac{(b-a)^2}{12} = \frac{4/n}{12} = \frac{1}{3n}$$

Dmitry Vetrov

$\bar{w}$



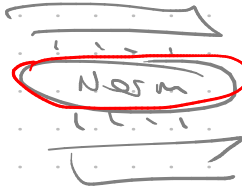
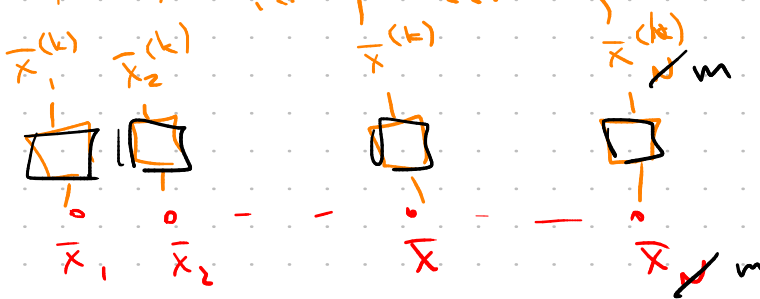
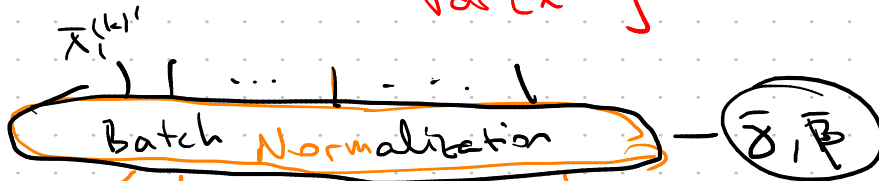
② Batch normalization

Covariate shift

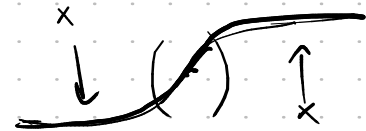
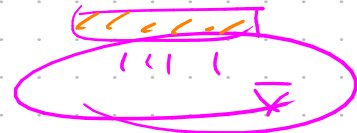


Internal covariate shift

$$\bar{x}^{(k)'} := \frac{\bar{x}^{(k)} - E[\bar{x}^{(k)}]}{\text{Var}[\bar{x}^{(k)}]}$$



$A(\bar{w})$



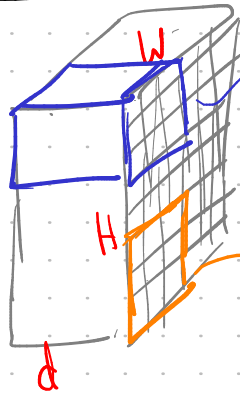
$$\bar{x}^{(k)'} := \frac{\bar{x}^{(k)} - E_m[\bar{x}^{(k)}]}{\text{Var}_m[\bar{x}^{(k)}]} \odot \bar{\gamma} + \bar{\beta}$$

variance

mean

③ Convolutional neural networks

$W \times H \times 3$



Hello-world



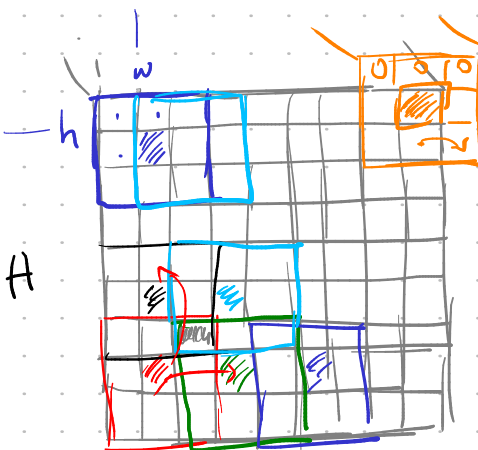
$d \times l \times d'$



$W = (dwh) \times d'$

d'

$d \times w \times h \times d'$

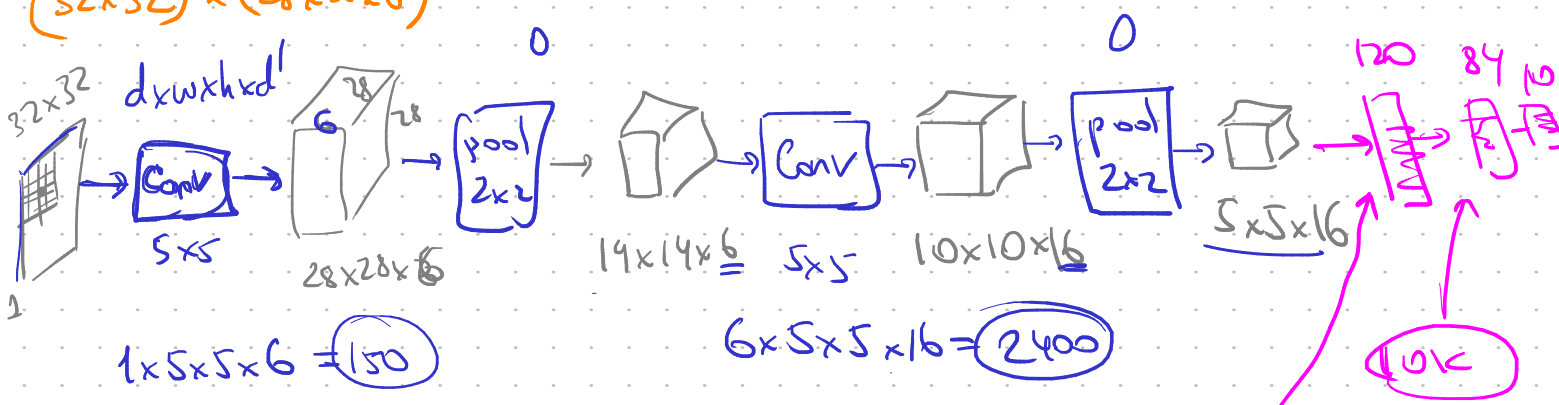
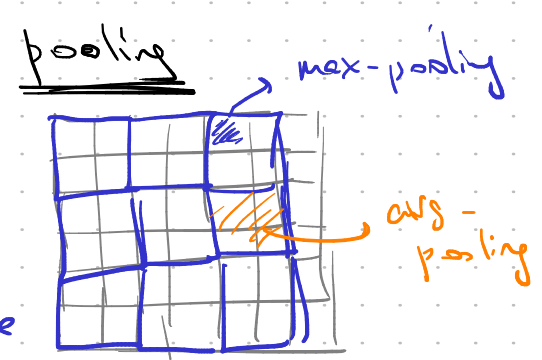
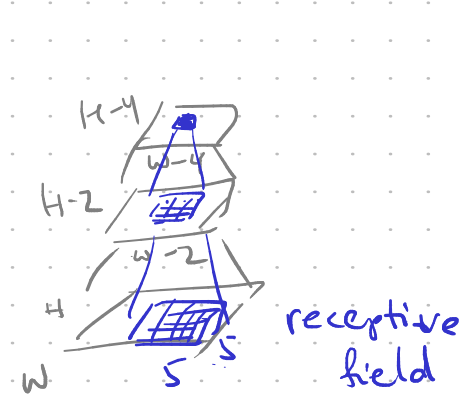


$$3 \times 3 \rightarrow (W-2) \times (H-2)$$

$$(2k+1) \times (2k+1) \rightarrow (W-2k) \times (H-2k)$$

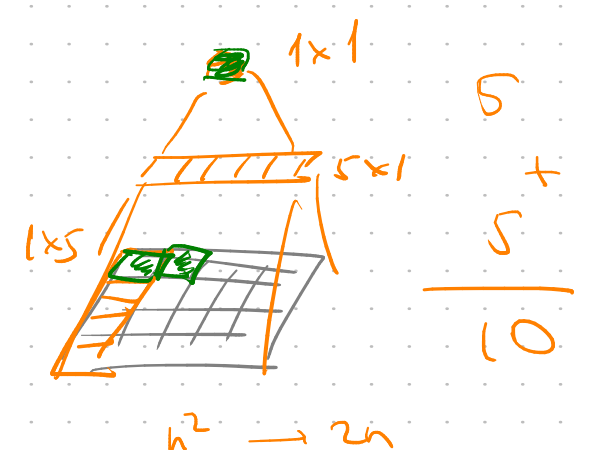
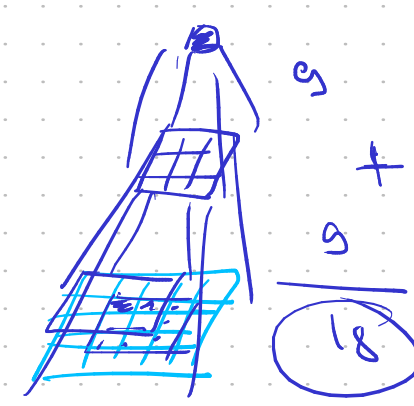
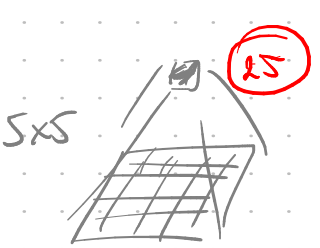
Stride (1, 1)

$$(32 \times 32) \times (28 \times 28 \times 6)$$

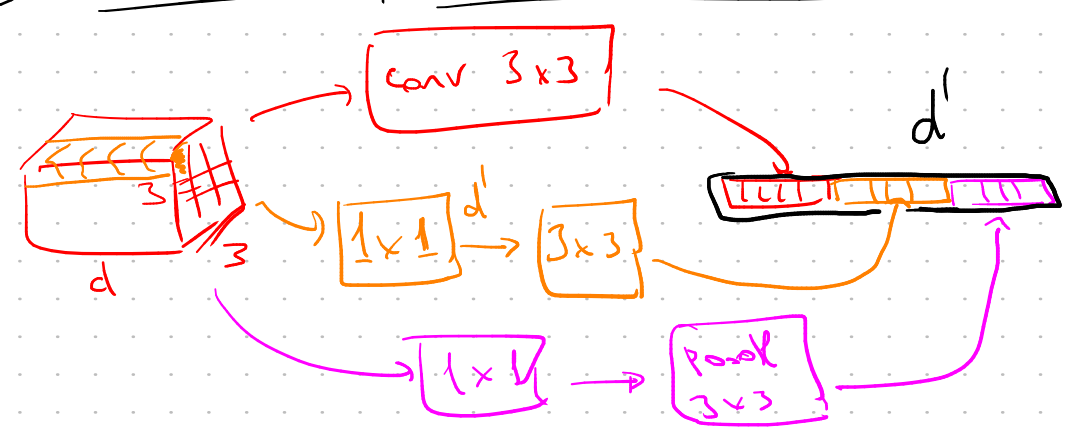


$$(16 \times 5 \times 5) \times 120 = 48000$$

(4) VGG

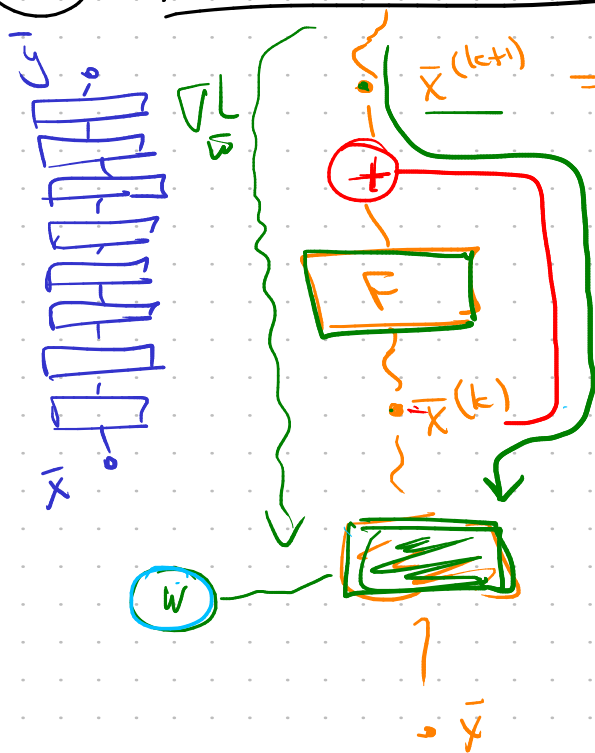


(5) Inception / network in network



⑥ Residual connections / ResNet

minimizing the



$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial \bar{x}^{(k+1)}} \frac{\partial \bar{x}^{(k+1)}}{\partial w} = \frac{\partial L}{\partial \bar{x}^{(k+1)}} \frac{d \times d}{\partial \bar{x}^{(k)}} \frac{\partial \bar{x}^{(k)}}{\partial w}$$

$$F' \left(- \frac{\partial F'}{\partial x_i} - \right)$$

$\|F'\| < 1$

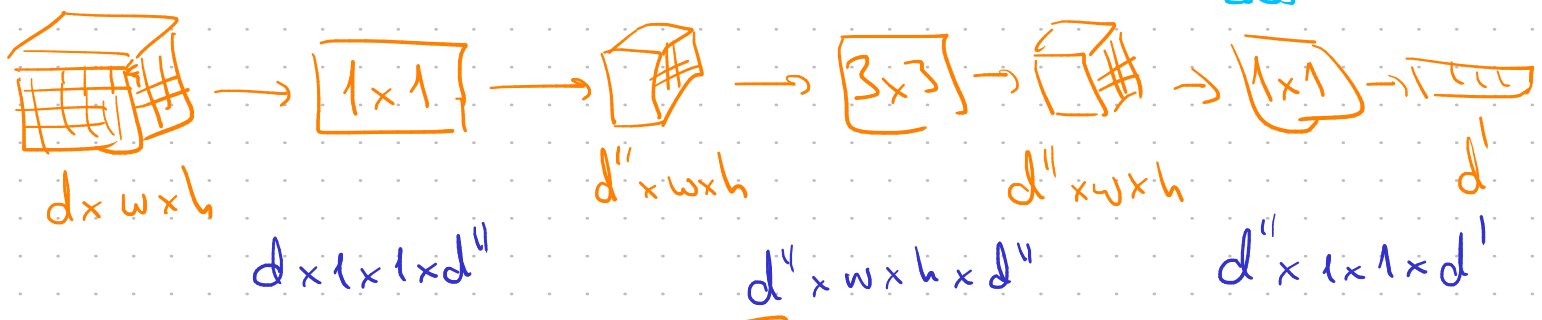
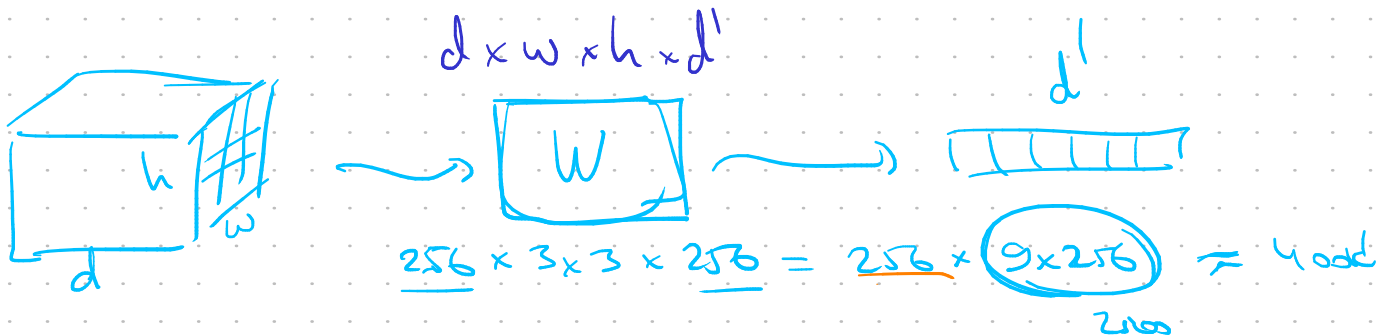
$$\bar{x}^{(k+1)} = F(\bar{x}^{(k)}) + \bar{x}^{(k)}$$

$$F \sim \bar{x}^{(k)} \mapsto (\bar{x}^{(k+1)} - \bar{x}^{(k)})$$

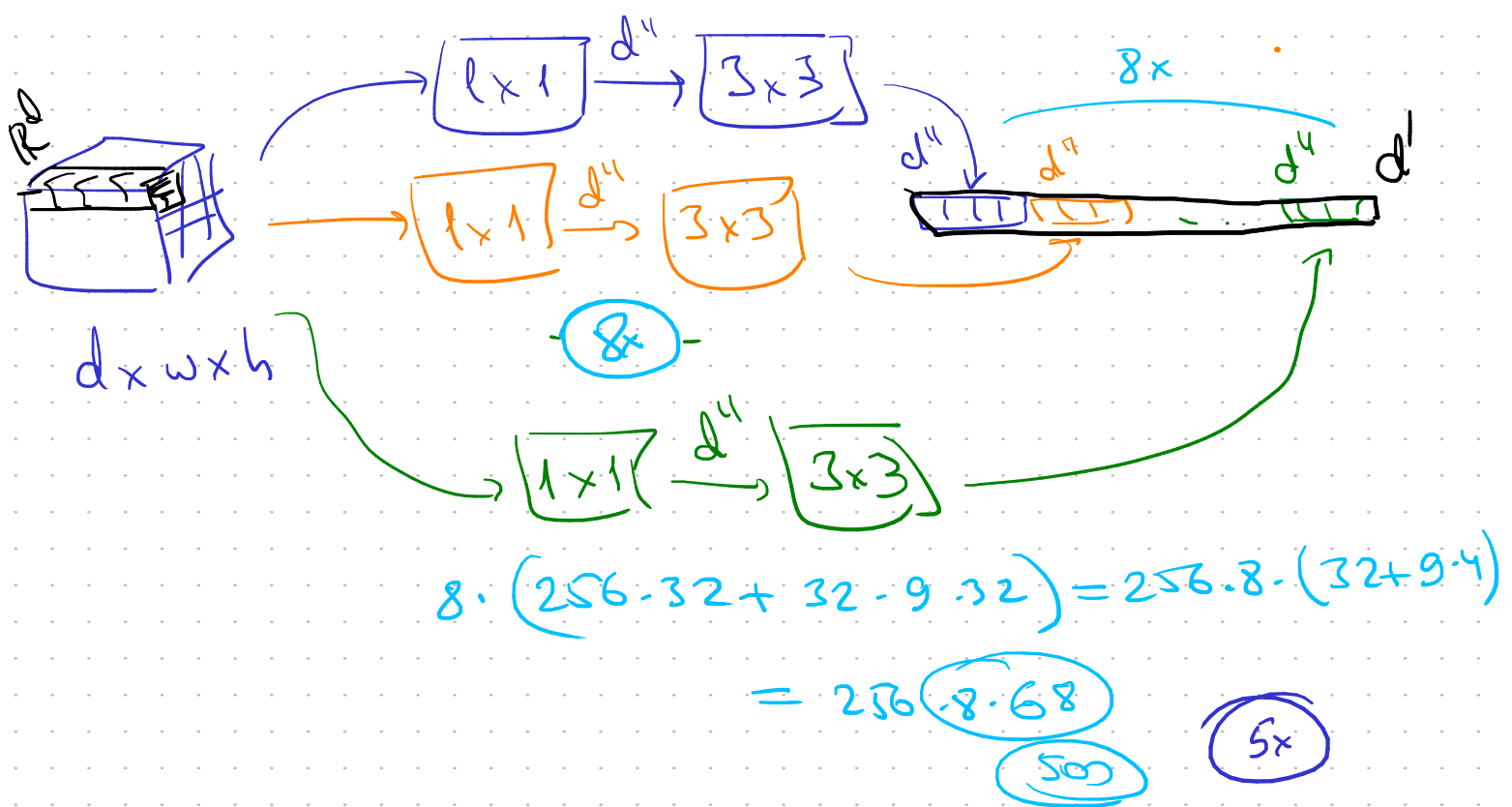
residual connection

$$\frac{\partial \bar{x}^{(k+1)}}{\partial \bar{x}^{(k)}} = F' + I$$

⑦ Bottleneck / split-transform-merge



$$256 \cdot 32 + 32 \cdot 9 \cdot 32 + 32 \cdot 256 = 256(32 + 9 \cdot 4 + 32) = 256 \cdot 100 = 25600$$



② EfficientNet

- depth d / #layers
- width w / channels
- resolution r / input size

$$= \frac{\alpha}{\beta} \cdot \frac{\gamma}{\delta} \cdot \frac{\epsilon}{\zeta}$$

Compound scaling

$$\text{FLOPs} \propto d \cdot w^2 \cdot r^2 = (\alpha \cdot \beta^2 \cdot \gamma^2) \cdot \frac{\epsilon}{\zeta}$$

NAS: $\left\{ \begin{array}{l} \alpha, \beta, \gamma \geq 1 \\ \alpha \cdot \beta^2 \cdot \gamma^2 = C \end{array} \right\} \rightarrow \boxed{\text{BO}}$