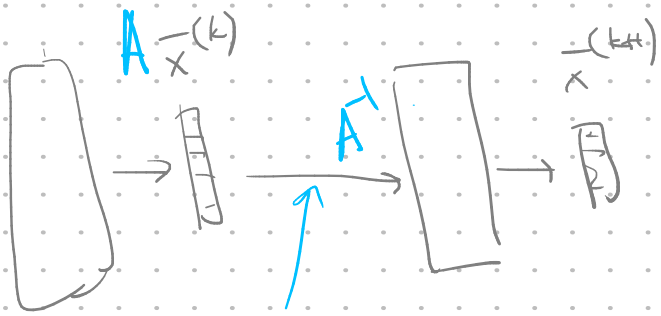
$p(D|\theta) \to \max$

$p(\theta|D) \to \max$

$\log p(\theta|D) = \log p(D|\theta) + \log p(\theta) + \text{const}$

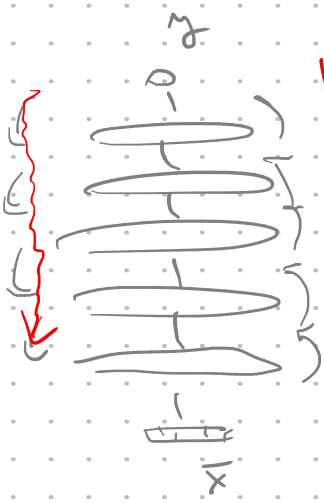
$L + L_{reg} \to \min$

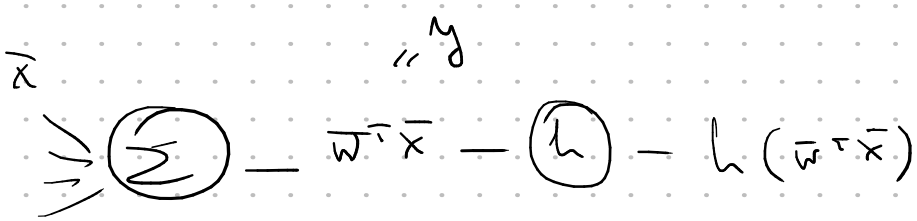


Unsupervised pretraining

exploding gradients

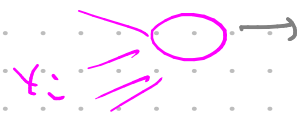
vanishing gradients





$y = \sum_{i}^{n} y_i \qquad y_i = w_i x_i$

$\mathrm{Var}[y_i] = \mathrm{Var}[w_i x_i] = \mathbb{E}[x_i]^2 \mathrm{Var}[w_i] + \mathbb{E}[w_i]^2 \mathrm{Var}[x_i] + \mathrm{Var}[x_i]\mathrm{Var}[w_i]$



1) $E[x_i] = 0$ $\quad Var[y_i] = Var[x_i] Var[w_i]$

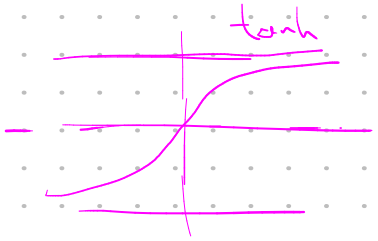
$Var[y] = n \cdot Var[w_i] Var[x_i]$

$Var[y] = Var[x]$ $\quad \Rightarrow \quad n \cdot Var[w_i] \approx 1$

$w_i \sim Unif\left(\left[-\frac{\sqrt{3}}{\sqrt{n}}, \frac{\sqrt{3}}{\sqrt{n}}\right]\right)$ $\quad \times \frac{1}{3}$

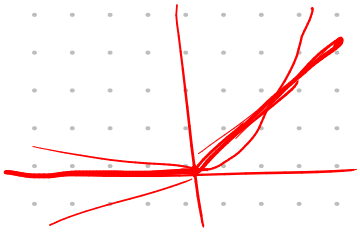
$Var[w_i] = \frac{(2/\sqrt{n})^2}{12} = \frac{1}{3n}$

→ Xavier init (Glorot) 2010



$E[y] = 0 \Rightarrow E[h(y)] = 0$

2) ReLU



$E[ReLU(\cdot)] \neq 0$

ReLU → $Var[w_i] = \frac{2}{n}$

He init $\quad$ Kaiming He

Batch normalization

Covariate shift

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

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

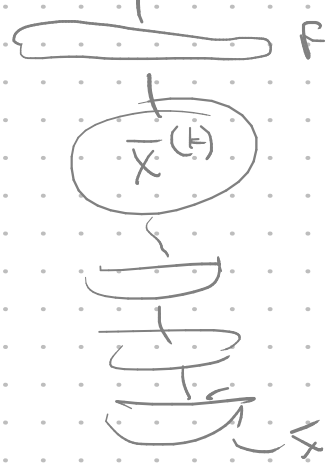
Norm

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

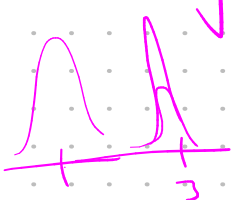
f

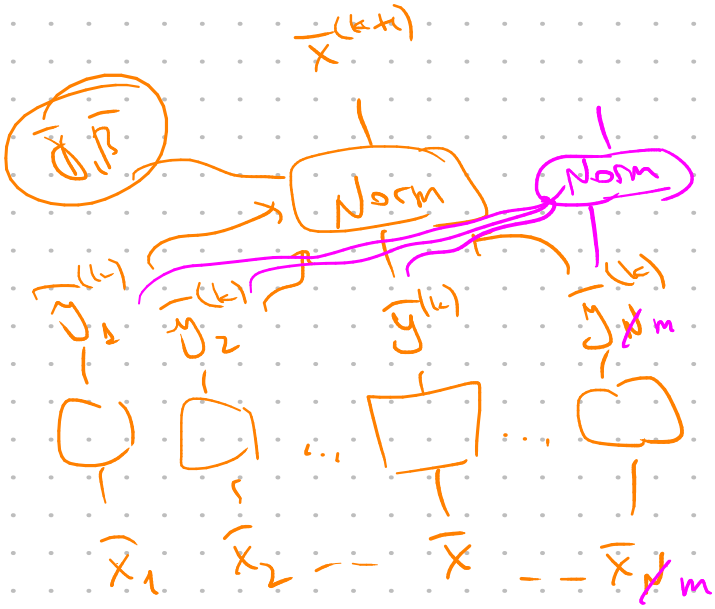
$\bar{x}^{(k)}$

Internal covariate shift



$\hat{x}^{(k+1)} = \frac{\bar{y}^{(k)} - E_D \bar{y}^{(k)}}{\sqrt{Var_D \bar{y}^{(k)}}}$



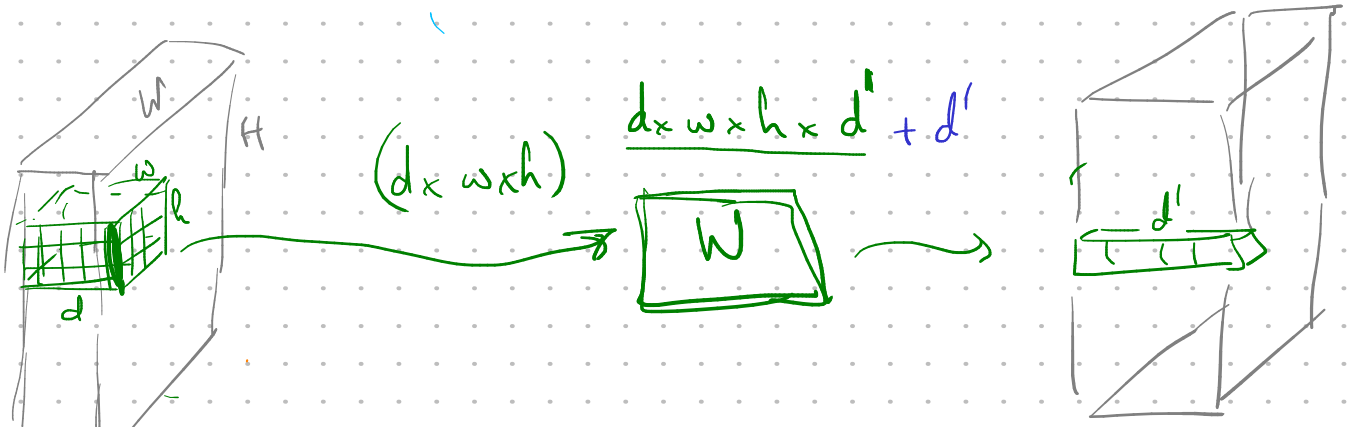


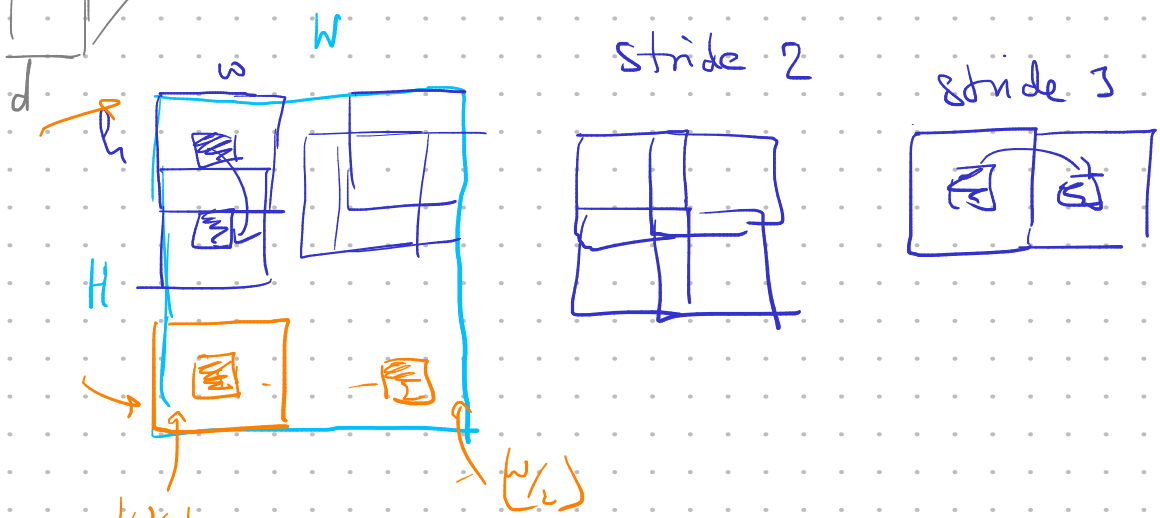
variance

mean

$$BN(\bar{x}) = \frac{\bar{x} - \mathbb{E}_m \bar{x}}{\sqrt{Var_m \bar{x}}} \odot \bar{\gamma} + \bar{\beta}$$

CNN



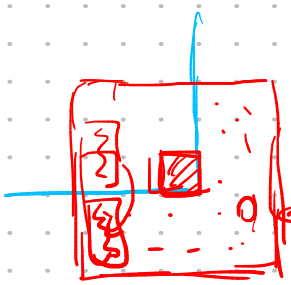


stride 1

W x H

w x h

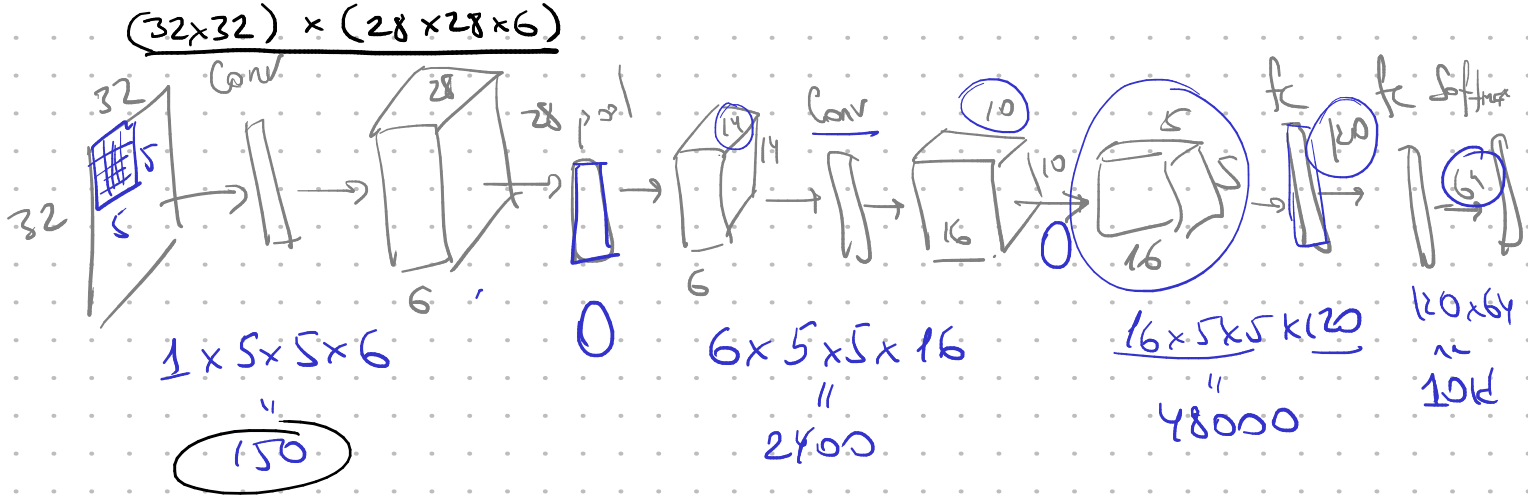
$\rightarrow \left(W - (w-1)\right) \times \left(H - (h-1)\right)$

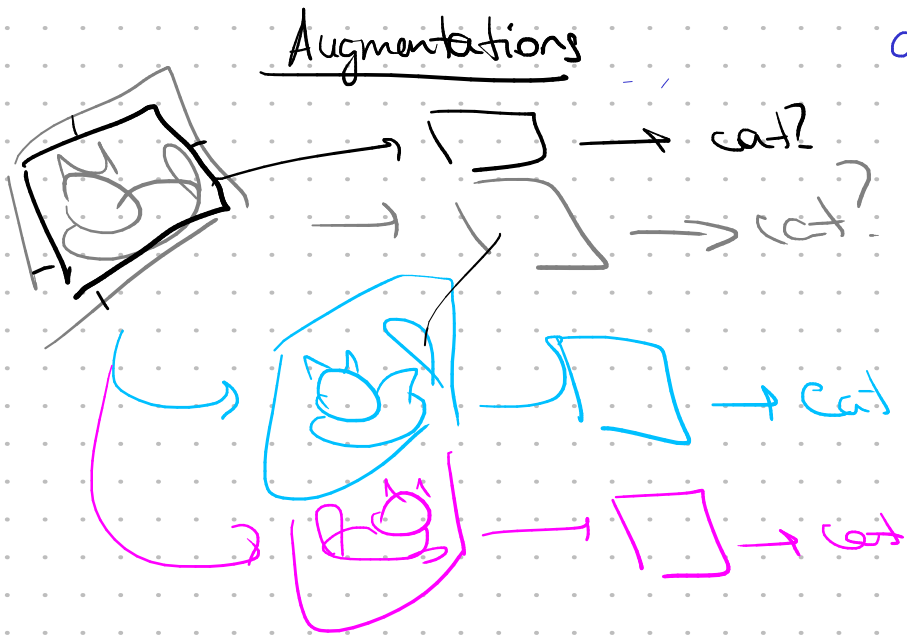


padding

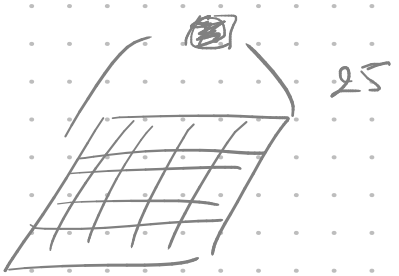
zero

W x H $\rightarrow$ W x H



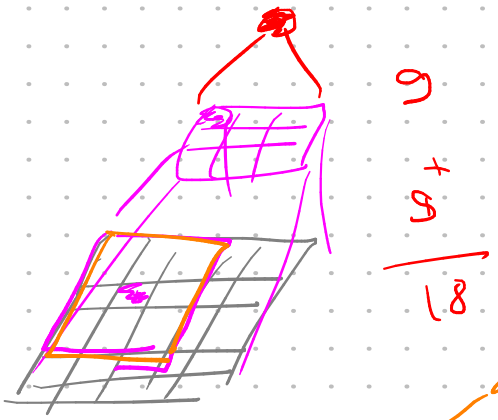


① VGG - Oxford

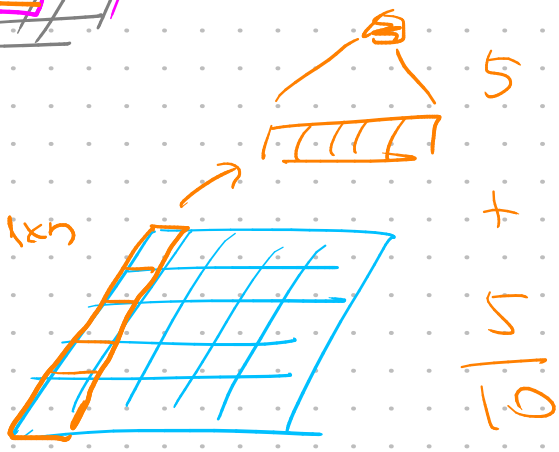


$(11 \times 11) \approx (3 \times 3) \cdot 5$

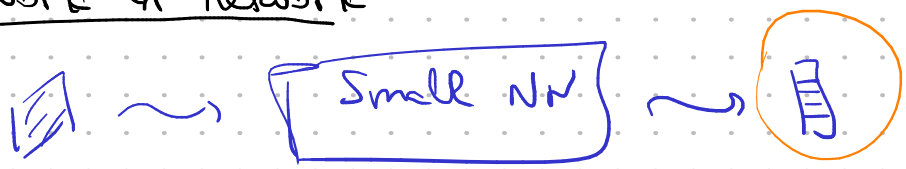
$121 \simeq 45$



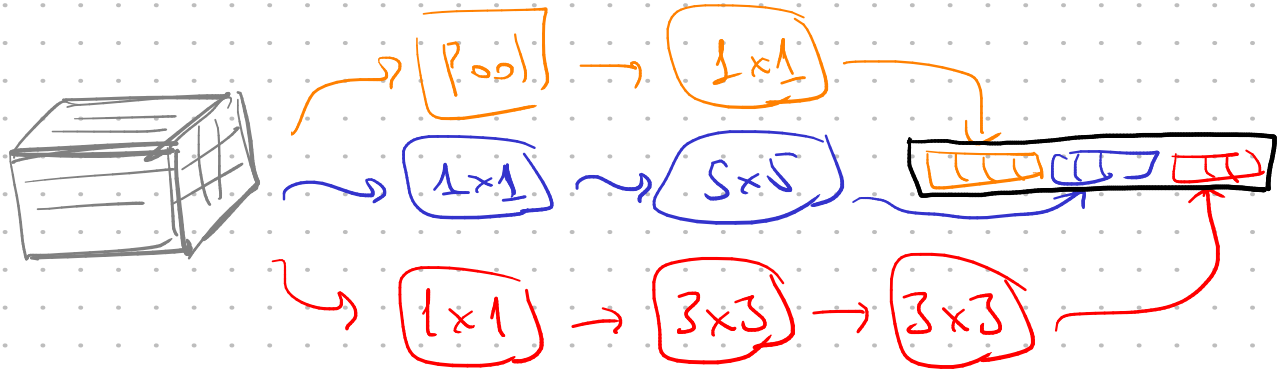
$n \times n \sim n \times 1 + 1 \times n$



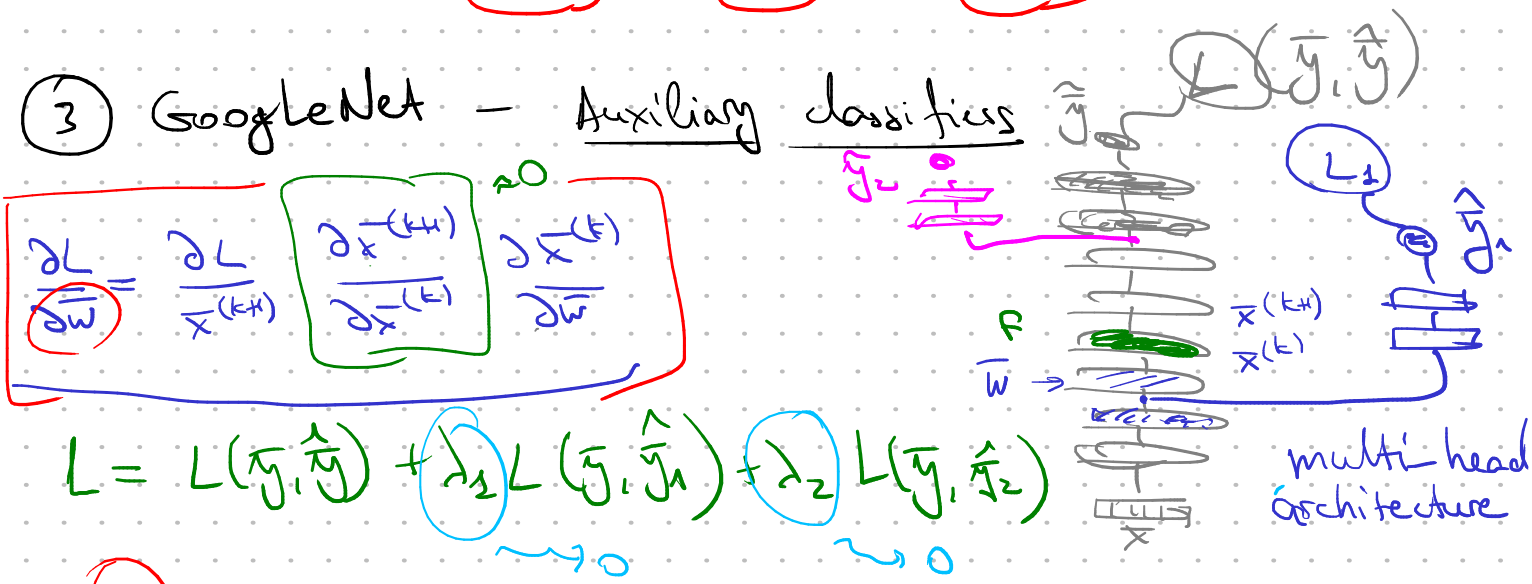
(2) Network in network



Inception module



(3) GoogLeNet – Auxiliary classifiers



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

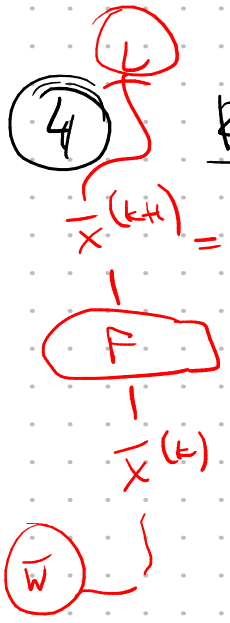
$$L = L(y, \hat{y}) + \lambda_1 L(y, \hat{y}_1) + \lambda_2 L(y, \hat{y}_2)$$

$\lambda_1 \to 0$, $\lambda_2 \to 0$

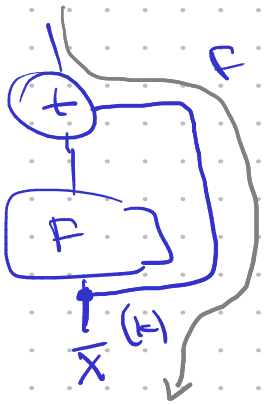
multi-head architecture

(4) Residual connections — training ResNet

$x^{(k+1)} = f(x^{(k)})$



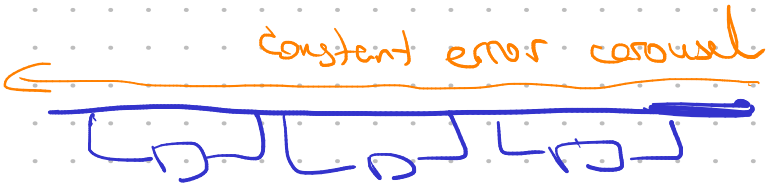
$x^{(k+1)} = x^{(k)} + f(x^{(k)})$



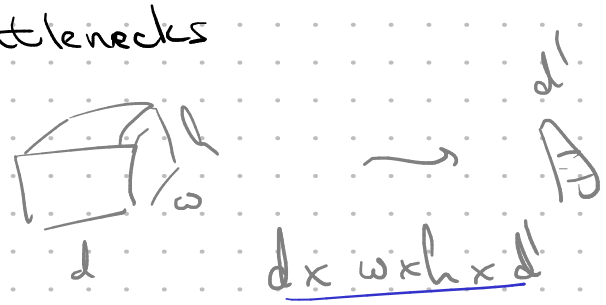
$f \approx x^{(k)} \mapsto (x^{(k+1)} - x^{(k)})$ residue

$$\frac{\partial x^{(k+1)}}{\partial x^{(k)}} = \frac{\partial f}{\partial x^{(k)}} + I$$

constant error carousel

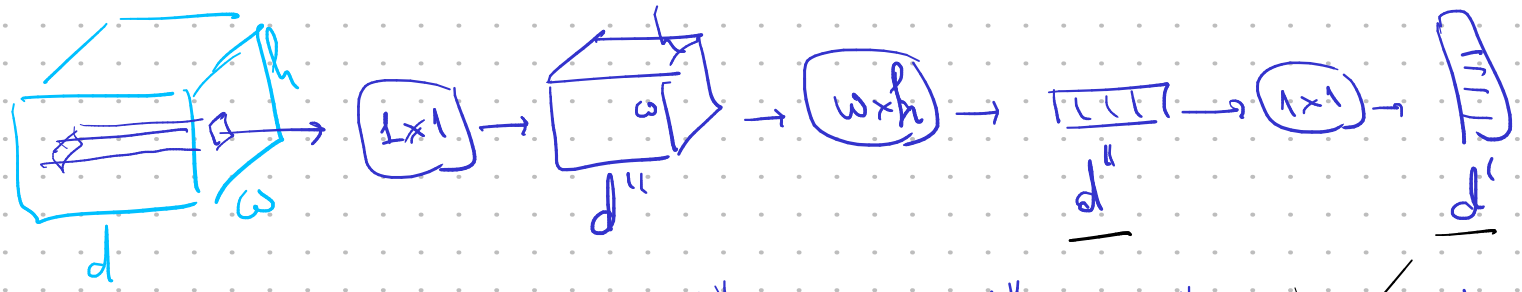


(5) Bottlenecks



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

$256 \times 3 \times 3 \times 256 = 256(256 \cdot 9)$

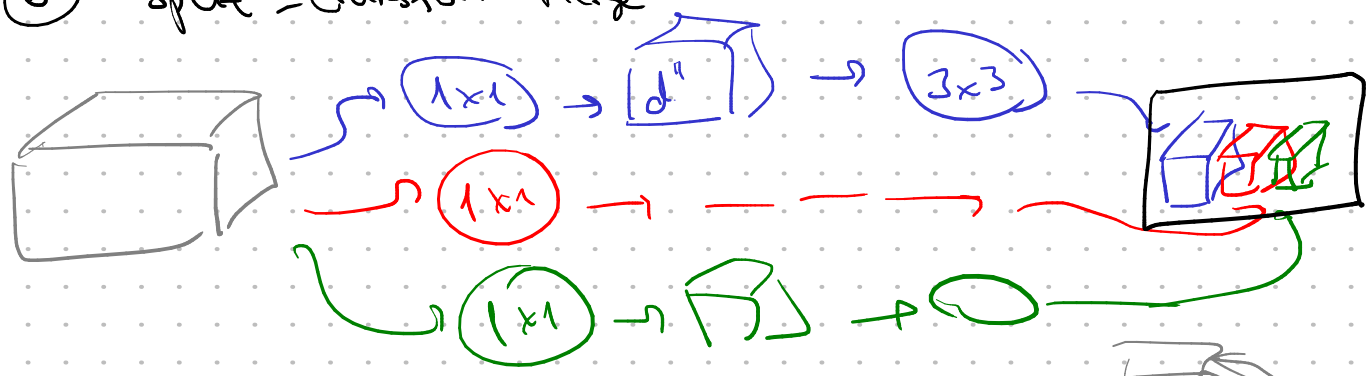


$d \times 1 \times 1 \times d'' \;+\; d'' \times w \times h \times d'' \;+\; d'' \times 1 \times 1 \times d'$

256 32 32 32 32 256

$256(32 + 4 \cdot 9 + 32) \cdot 8$

(6) Split-transform-merge



(7) DenseNet - skip connections