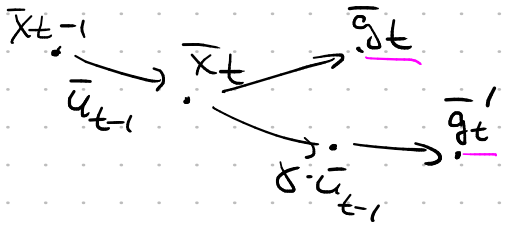
SGD / momentum / NAG

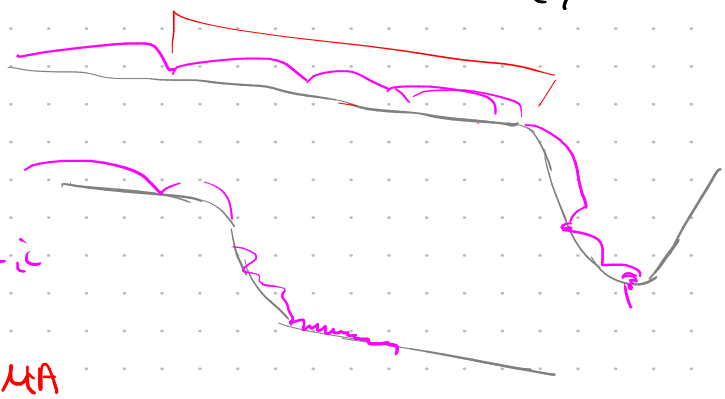


① Adaptive SGD $F(\bar{x})$ $\bar{x}_t$

- **Adagrad** $G_0 = 0$

$$G_{t,i} = G_{t-1,i} + g_{t,i}^2$$

$$x_{t+1,i} = x_{t,i} - \frac{\alpha}{\sqrt{G_{t,i} + \varepsilon}} \cdot g_{t,i}$$



- Exponential moving average EMA

$$G_{t,i} = G_{t-1,i} \cdot \gamma + (1-\gamma) \cdot g_{t,i}^2 =$$

$$= (1-\gamma) g_{t,i}^2 + (1-\gamma)\gamma g_{t-1,i}^2 + (1-\gamma)\gamma^2 g_{t-2,i}^2 + \dots$$

RMSprop

- **AdaDelta** $\bar{x}$ - coordinates $F(\bar{x})$ - loss $\bar{g}$ - м/с

$$\bar{x} := \bar{x} - \alpha \cdot \bar{g}_{м/с}$$

$$\bar{x} := \bar{x} - \frac{\alpha}{\sqrt{\bar{g}^2}_1} \bar{g}$$

✓ $$\bar{x} := \bar{x} - \frac{1}{2} H^{-1} \bar{g}$$

$$\left(\frac{м}{с^2}\right)^{-1} \quad м/с$$

$$x_{t+1,i} = x_{t,i} - \alpha \cdot \frac{\sqrt{R_{t,i} + \varepsilon}}{\sqrt{G_{t,i} + \varepsilon}} g_{t,i}$$

$$R_{t,i} = \rho \cdot R_{t-1,i} + (1-\rho) u_{t,i}^2$$

- **Adam / Nadam**

$$G_{t,i} = \beta_2 G_{t-1,i} + (1-\beta_2) g_{t,i}^2$$

$$x_{t+1,i} = x_{t,i} - \frac{\alpha}{\sqrt{G_{t,i} + \varepsilon}} \cdot m_{t,i}$$

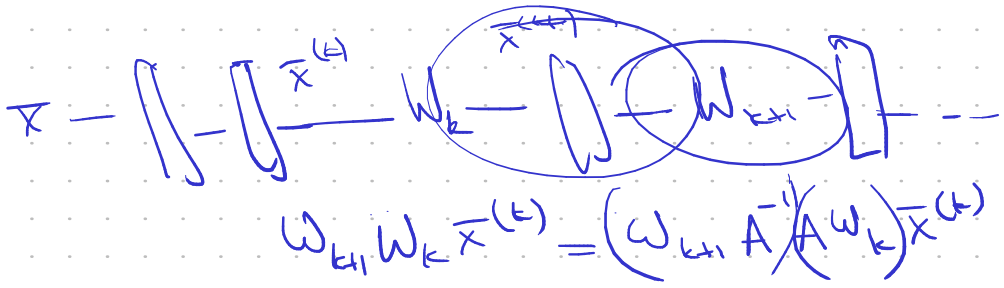
$$m_{t,i} = \beta_1 m_{t-1,i} + (1-\beta_1) g_{t,i}$$

$$\beta_1 = 0.9, \quad \beta_2 = 0.999, \quad \varepsilon = 10^{-6,7}$$

- Weight decay $L(\bar{w}) + \alpha \cdot \|\bar{w}\|^2$

AdamW $\hookrightarrow \bar{w} := (1-\eta)\bar{w} - \alpha \nabla_w L(\bar{w}) = \bar{w} - \alpha \cdot \nabla_w \left(L(\bar{w}) + \frac{\eta}{2\alpha} \|w\|^2 \right)$

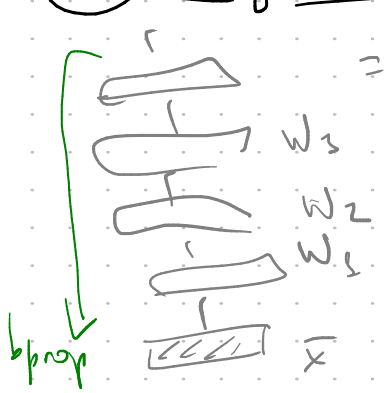
(2) Dropout



$$W_{k+1} W_k \bar{x}^{(k)} = \left(W_{k+1} A^{-1}\right)\left(A W_k\right)\bar{x}^{(k)}$$

(3) Weight initialization

2006–2014 unsupervised pretraining



$$= W_k W_{k-1} W_{k-2} W_{k-3} \cdots W_1 \bar{x}$$

- exploding gradients
- vanishing gradients

$$\bar{x} \rightarrow \Sigma \rightarrow \bar{w}^T \bar{x} = y \rightarrow h \rightarrow h(\bar{w}^T \bar{x})$$

$\mathrm{Var}[y] \approx \mathrm{Var}[x_i]$

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

$\mathrm{Var}[y] \approx n \cdot \mathrm{Var}[y_i]$

$$\mathrm{Var}[y_i] = \mathrm{Var}[w_i x_i] = \mathbb{E}[w_i^2 x_i^2] - \left(\mathbb{E}[w_i x_i]\right)^2 =$$

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

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

$\mathrm{Var}[w_i] \approx O(1/n)$ $\quad \approx 1$

$\mathrm{Var}(\mathrm{Unif}([a,b])) = \frac{(b-a)^2}{12}$

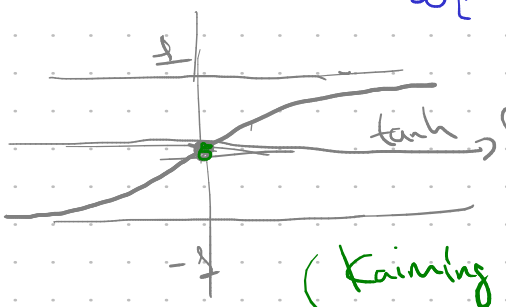
$w_i \sim \mathrm{Unif}\left(\left[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right]\right)$

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

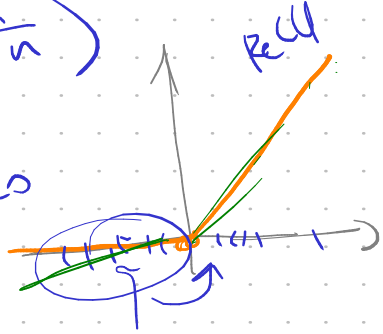
(Glorot) Xavier init

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

$w_i \sim N\left(0, \frac{1}{n}\right)$



$\mathbb{E}[x_i] = 0$
$\mathbb{E}[w_i] = 0$



$\mathbb{E}[\mathrm{ReLU}(\_)] > 0$

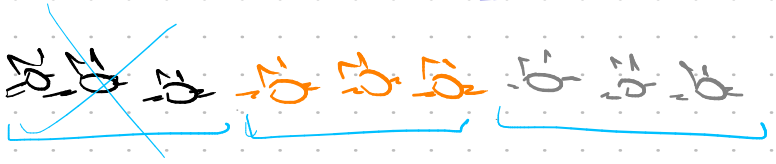
(Kaiming He)

He init: $\mathrm{Var}[w_i] = 2/n$

(4) Batch normalization

internal covariate shift

covariate shift: $\bar{x}^{(k+1)} \rightarrow \square \rightarrow (1)$



Norm $\cdot \bar{x}^{(k+1)} = \text{Norm}(\bar{y}^{(k)}) = \frac{\bar{y}^{(k)} - \mathbb{E}_D[\bar{y}^{(k)}]}{\sqrt{\text{Var}_D[\bar{y}^{(k)}]}}$

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

F — $\bar{w}^{(k)}$

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

По выходам слоя — layer normalization

$\mathbb{E}$, Var по D

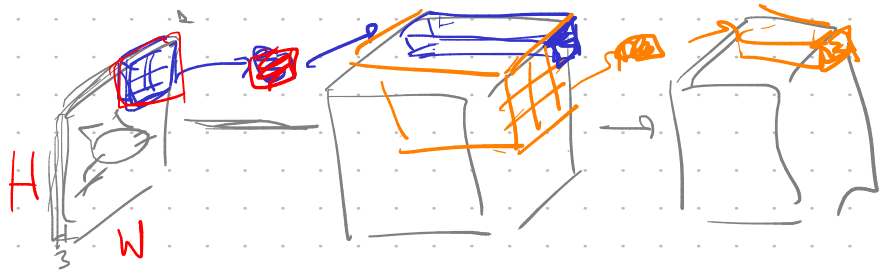
B.Norm. — $\bar{\gamma}, \bar{\beta}$

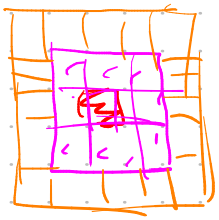
$BN(\bar{y}^{(k)}) = \frac{\bar{y}^{(k)} - \mathbb{E}_m[\bar{y}^{(k)}]}{\sqrt{\text{Var}_m[\bar{y}^{(k)}]}} \odot \bar{\gamma} + \bar{\beta}$

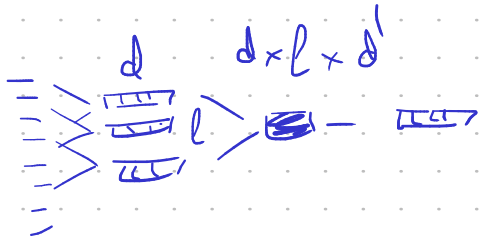
$\bar{y}_1^{(k)}$ $\bar{y}_2^{(k)}$ ... $\bar{y}_m^{(k)}$

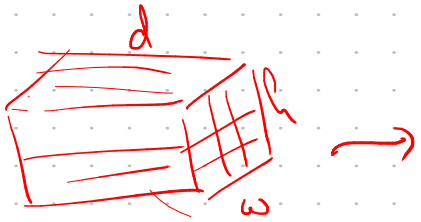
$\bar{x}_1$ $\bar{x}_2$ — — $\bar{x}_m$

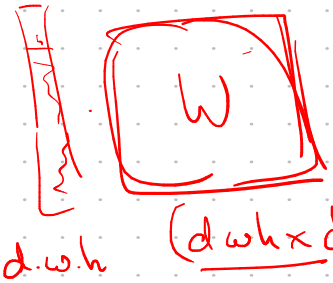
(5) CNN

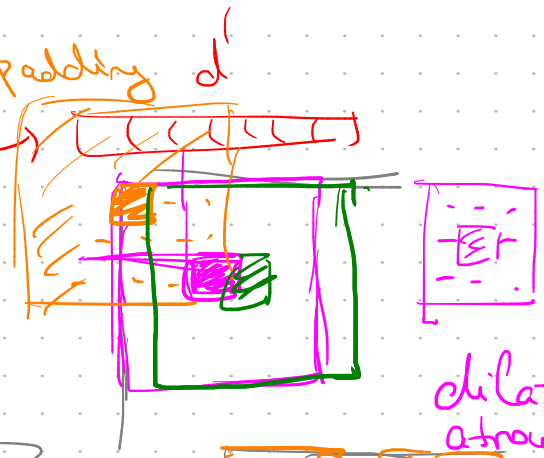






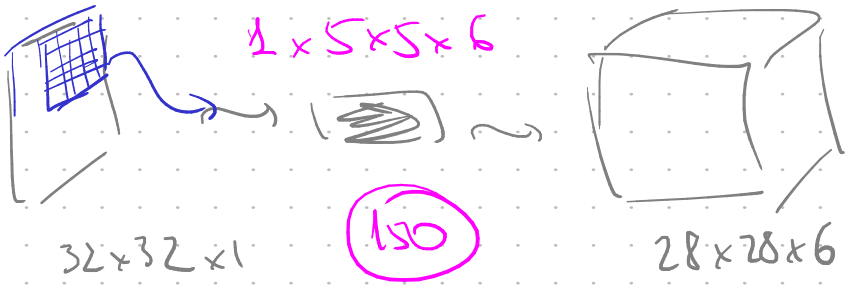


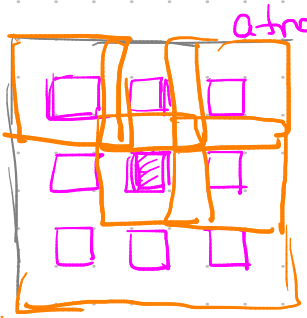




(d × w × h) × (d')

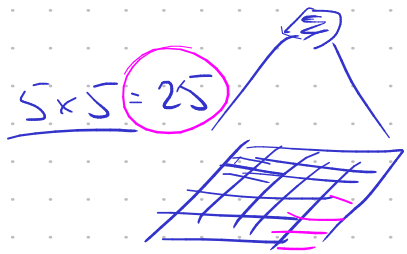
dilated atrous



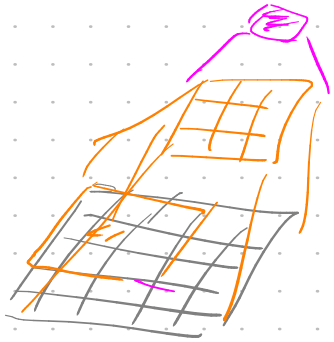


$(32 \times 32) \times (28 \times 28 \times 6)$

(6) VGG



$5 \times 5 = 25$



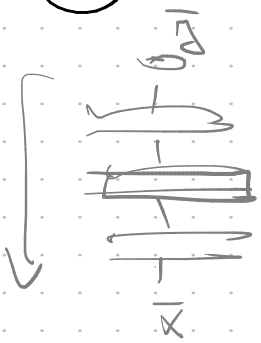
$(3 \times 3) \cdot 2 = 18$

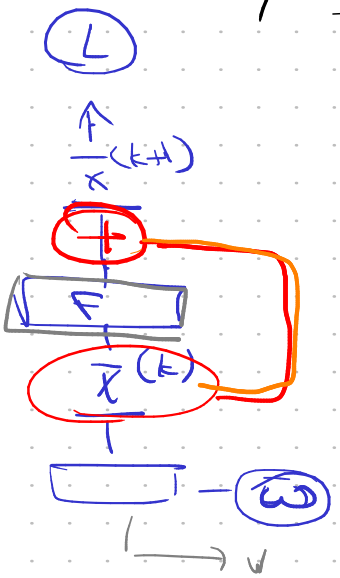
$11 \times 11 = 121$

$(3 \times 5) \cdot 5 = 45$

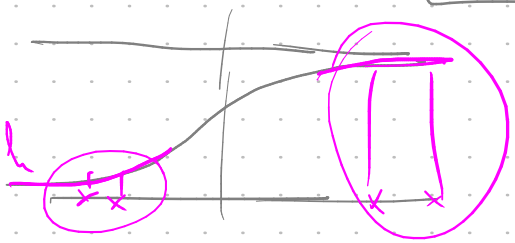
(7) Network in Network / Inception

(8) Residual connections





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



$\nabla_x F \to 0$

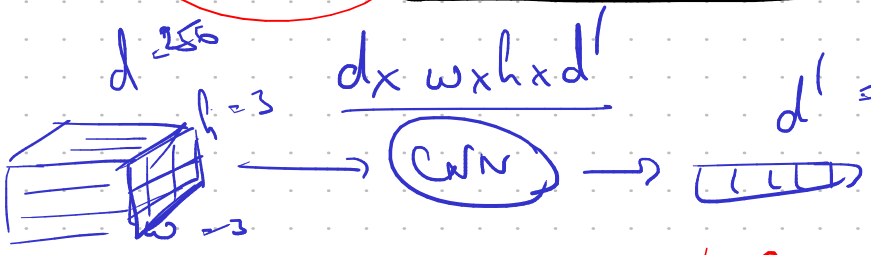
$$x^{(k+1)} = f(x^{(k)}) + x^{(k)}$$

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

residue

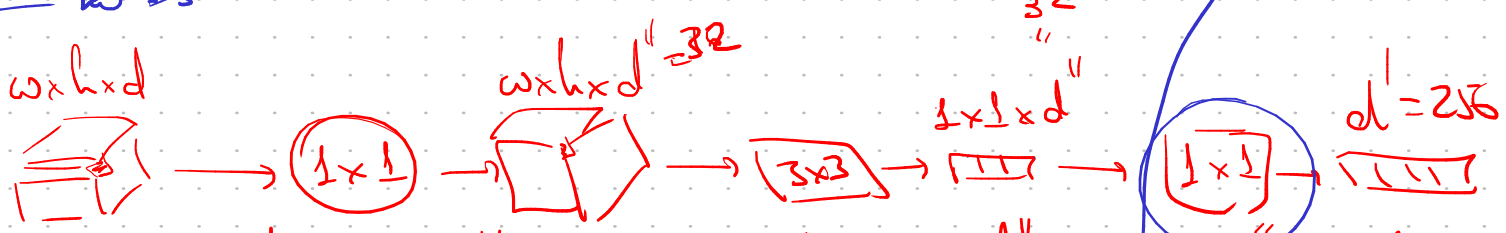
$$\frac{\partial x^{(k+1)}}{\partial x^{(k)}} = \nabla_x f + I$$

(9) Bottleneck / split-transform-merge



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

$256 \cdot (256 \cdot 9)$



$d \times 1 \times 1 \times d''$

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

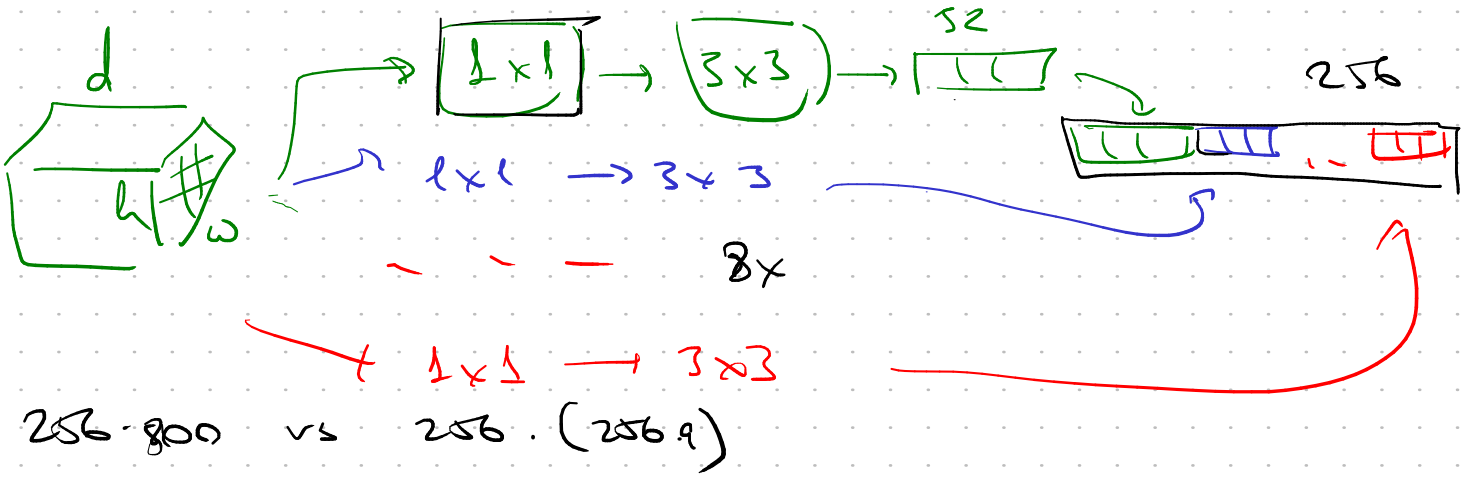
$d'' \times 1 \times 1 \times d'$

$256 \cdot 32$

$32 \cdot 3 \cdot 3 \cdot 32$

$256 \cdot 32$

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



$256 \cdot 800$ vs $256 \cdot (256 \cdot 9)$