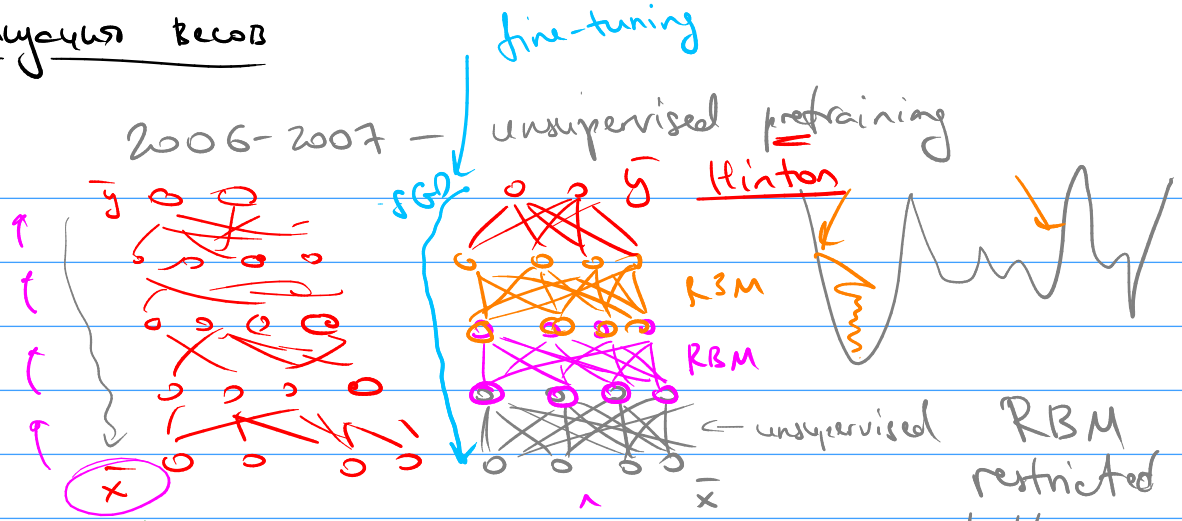
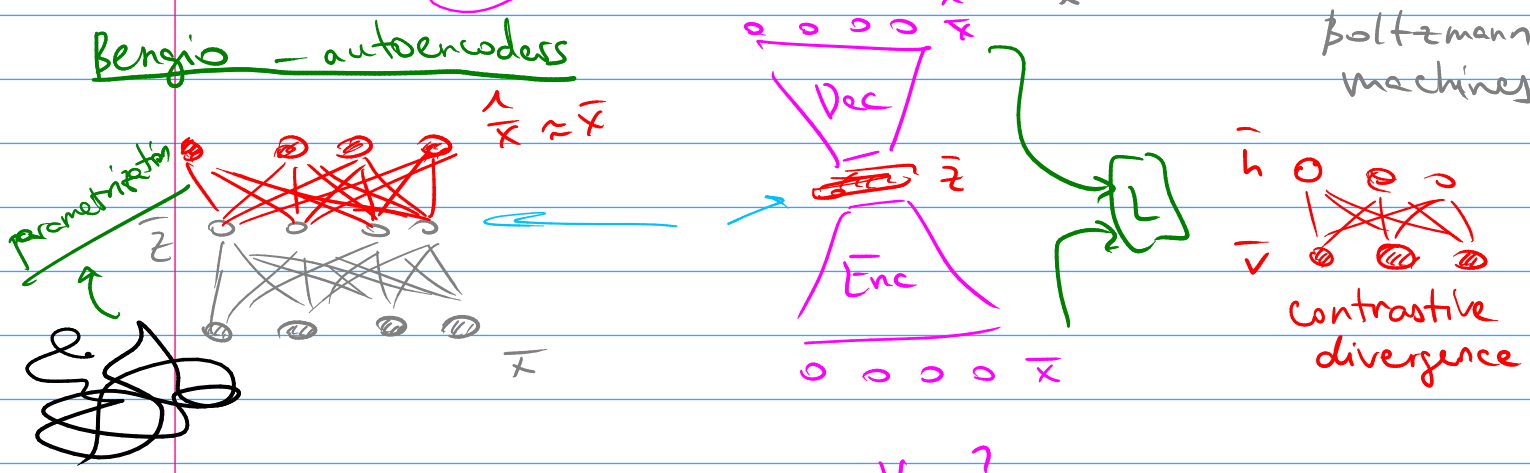
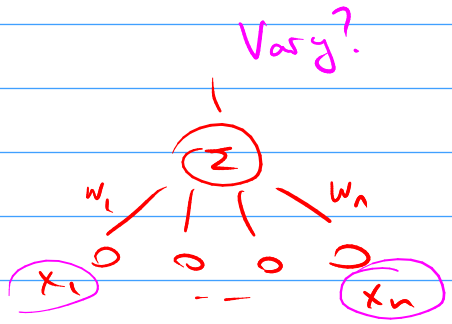
0 Инициализация весов





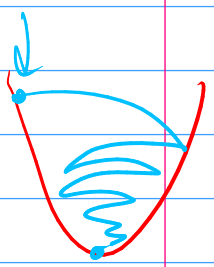
(Glorot, Bengio, 2010)



$$y = \bar{w}^T \bar{x} = \sum_{i=1}^{n} w_i x_i$$

$$Var[w_i x_i] = E[w_i^2 x_i^2] - (E[w_i x_i])^2 \neq$$

$$\neq E[x_i]^2 Var[w_i] + E[w_i]^2 Var[x_i] + Var[w_i] Var[x_i] =$$



$$= Var[w_i] \cdot E[x_i^2] + E[w_i]^2 \cdot Var\, x_i$$

$$(E(w_i^2) - (E w_i)^2) E[x_i^2] + E[w_i]^2 (E x_i^2 - (E x_i)^2) =$$

$$= E[w_i^2 x_i^2] - (E[w_i x_i])^2$$

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

(0 and 0)

$$\mathrm{Var}[w_i x_i] = (\mathrm{Var}\, w_i)(\mathrm{Var}\, x_i)$$

$$\mathrm{Var}\, y = \mathrm{Var}\left[\sum w_i x_i\right] = \sum \mathrm{Var}[w_i x_i] = n \cdot \mathrm{Var}\, w_i \cdot \mathrm{Var}\, x_i$$

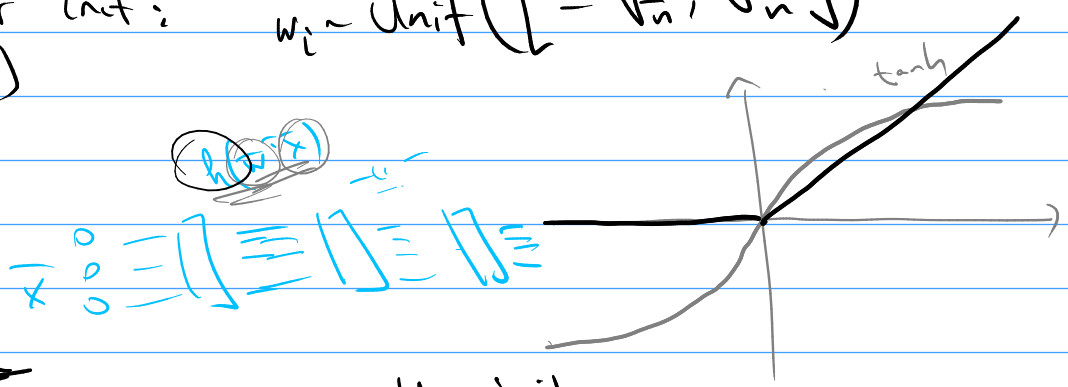
$$\mathrm{Var}\, y = (n \cdot \mathrm{Var}\, w_i) \cdot \mathrm{Var}\, x_i$$

$$w_i \sim \mathrm{Unif}\left(\left[-\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}\right]\right) \qquad \mathrm{Var}[\mathrm{Unif}[a,b]] = \frac{(b-a)^2}{12}$$

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

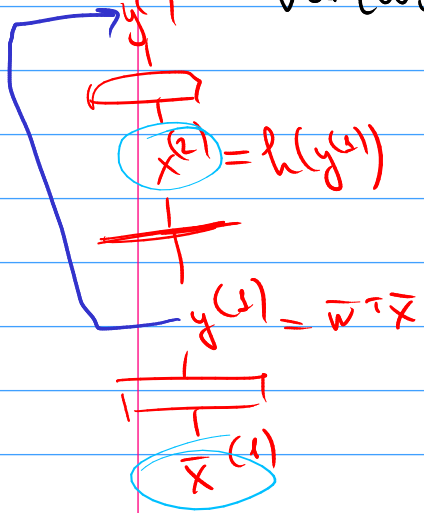
Xavier init: (Glorot)

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



Kaiming He (He et al., 2015): He init

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



$$x^{(l)} = h(y^{(l-1)}) \qquad \mathbb{E}[y^{(l)}] = 0 \quad (= \sum x_i w_i^{(l)})$$

$$x^{(l)} = \max(0, y^{(l-1)})$$

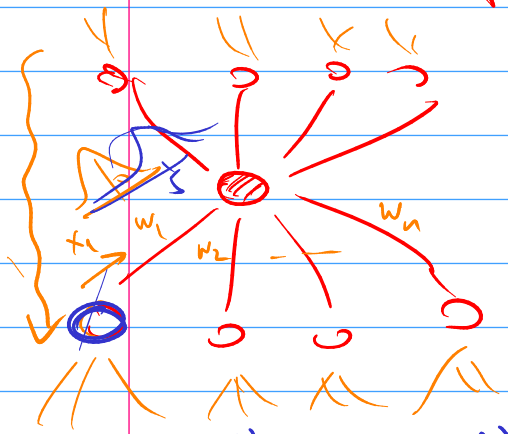
$$\mathbb{E}[x^{(l)2}] = \frac{1}{2}\mathrm{Var}[y^{(l-1)}]$$

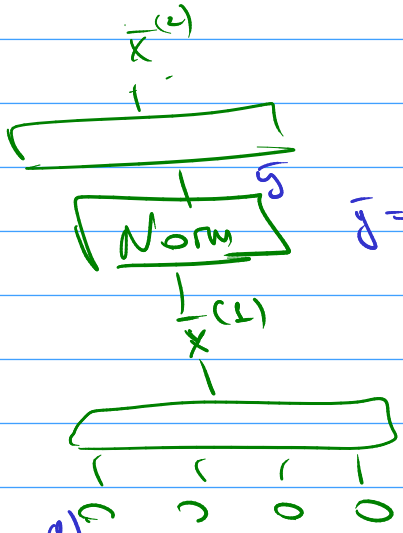
$$\mathrm{Var}[y^{(l)}] = \frac{n}{2}\mathrm{Var}[w^{(l)}]\mathrm{Var}[y^{(l-1)}]$$

$$w_i^{(l)} \sim N\left(0, \sqrt{\frac{2}{n}}\right) \qquad \mathrm{Var}[w^{(l)}] = \frac{2}{n}$$

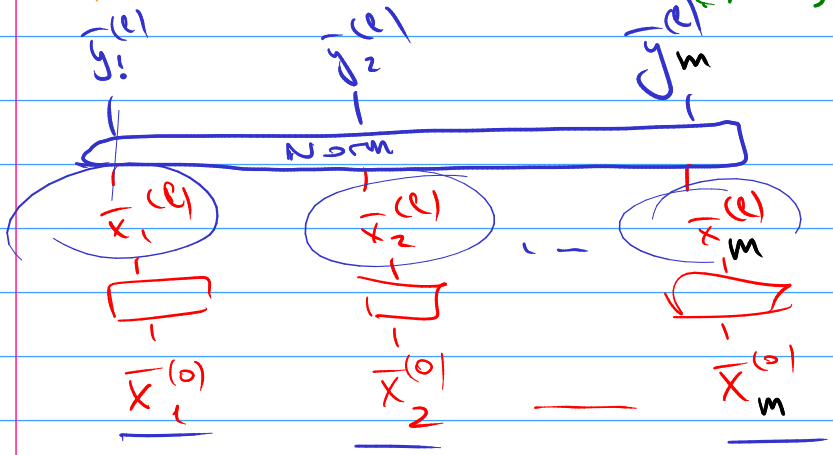
Batch Normalization

Internal covariate shift

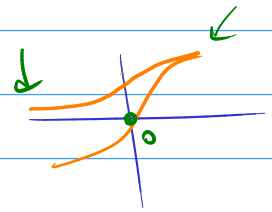


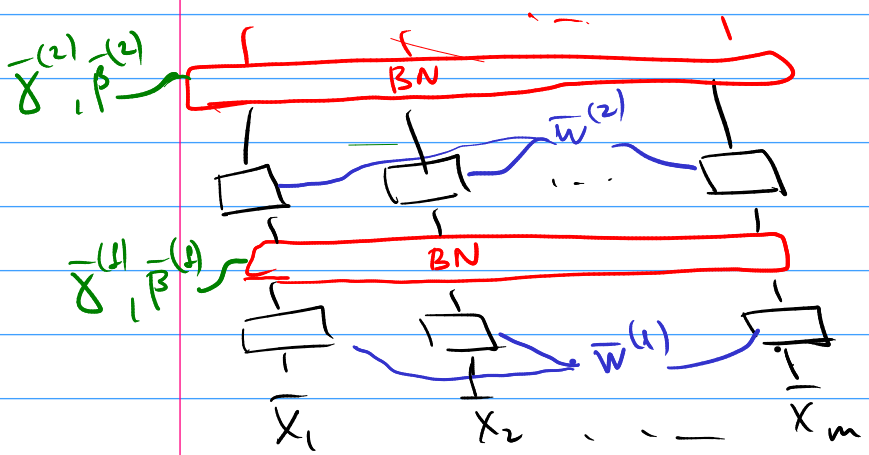


$$\bar{y} = \frac{\bar{x} - E\bar{x}}{\sqrt{Var\,\bar{x}}}$$

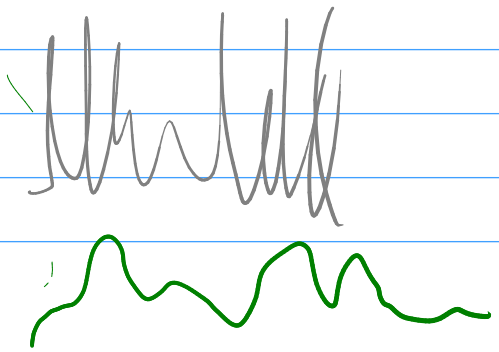


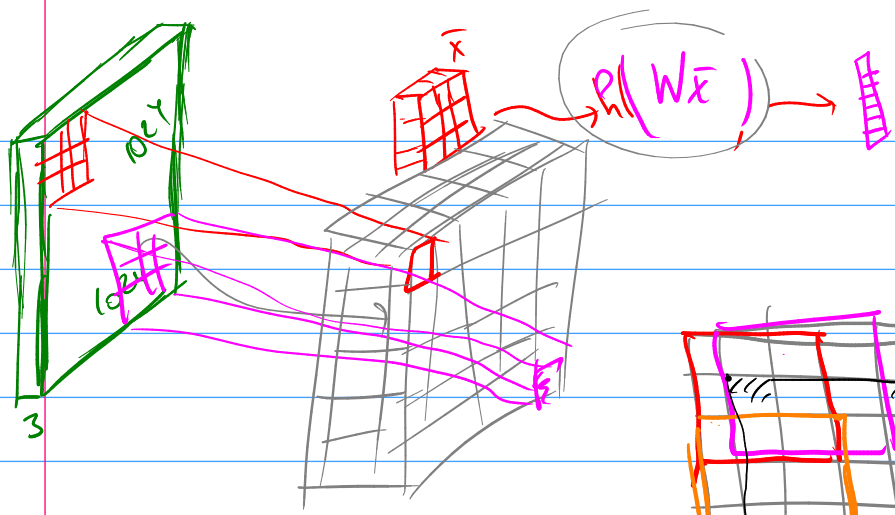
$$y_i = \frac{x_i - E_0 x_i}{\sqrt{Var_0 x_i}}$$

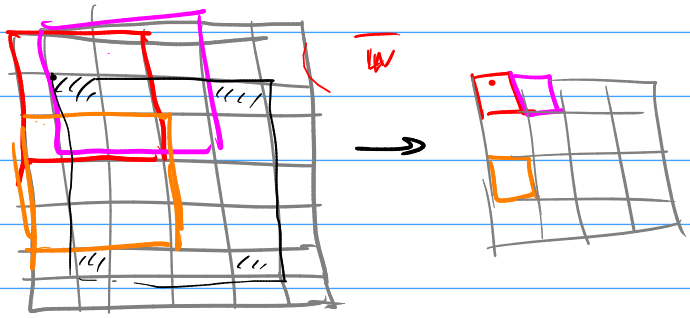


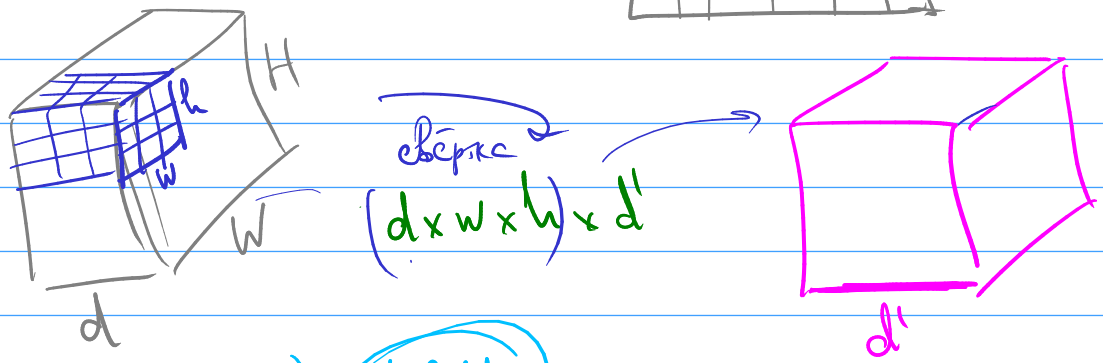


$$y_i = \frac{x_i - Avg(x_i)}{\sqrt{\widehat{Var}(x_i)}}\gamma_i + \beta_i$$



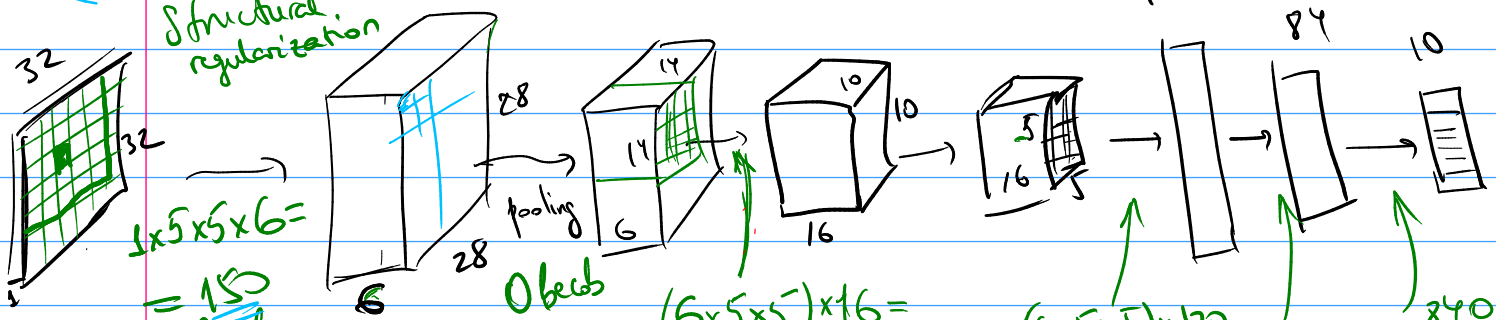


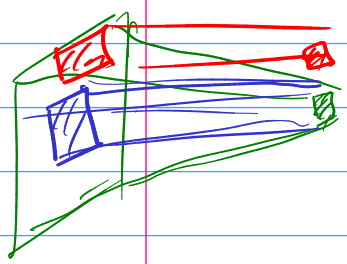


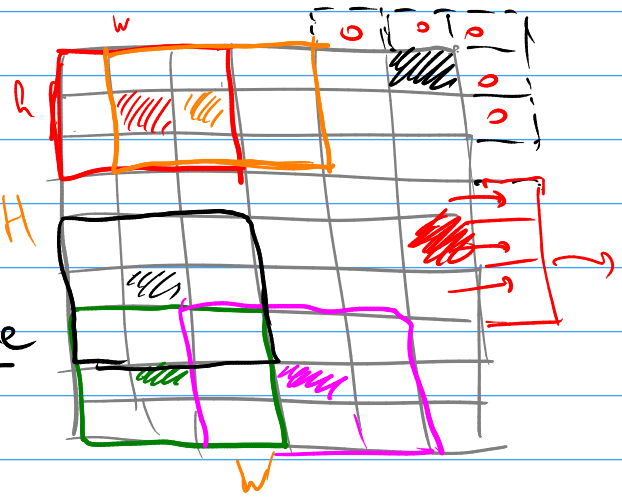


$(32 \times 32) \times (6 \times 28 \times 28) \approx 4.8M$

Structural regularization







padding