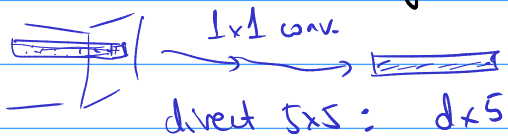
Ideas:

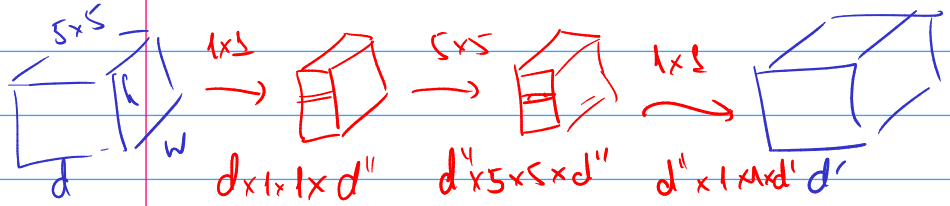
1) 5×5 ⟶ (3×3) ∘ (3×3) - VGG idea

2) Network in network

3) 1×1 convolution - fully conn. layer shared across pixels



direct 5×5 : $d \times 5 \times 5 \times d'$ weights



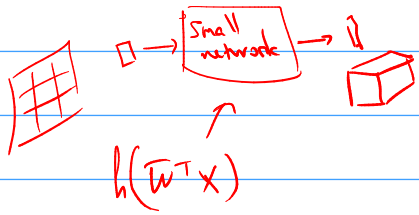
$256 \times 32 + 32^2 \times 25 + 256 \times 32$

$256 \times (64 + 4 \times 25) = 256 \times 164$

$256^2 \times 25 \leq 100 \cdot 256 \cdot 64$

256×6400

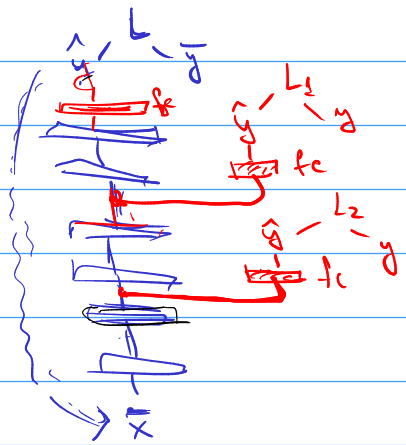
5×5	25	121
2×3×3	18	2× 45



Network in network

4) Auxiliary classifiers

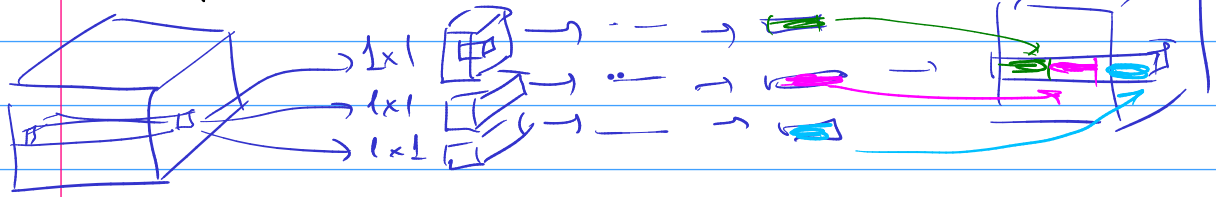
GoogleNet



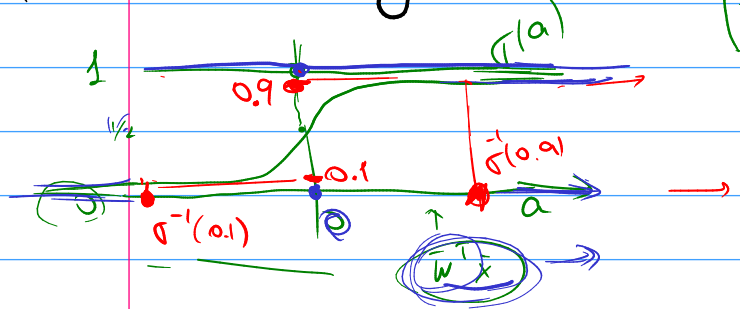
$L = L(\bar{w}) + \alpha_1 L_1(\bar{w}) + \alpha_2 L_2(\bar{w})$

α_1: 1 → 0, α_2: 1 → 0

5) Separate processing, concat results



6) Label smoothing



$\sigma(a) = \frac{1}{1+e^{-a}}$

7) Residual connections

Kaiming He

$$\frac{\partial L}{\partial W_i^{(k-1)}} = \sum_j \frac{\partial L}{\partial x_j^{(k)}} \frac{\partial x_j^{(k)}}{\partial W_i^{(k-1)}}$$

$$\frac{\partial L}{\partial x_j^{(k)}} = \dots \frac{\partial f}{\partial x_j}$$

Constant error carousel

LSTM, GRU 1995

∇f

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

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

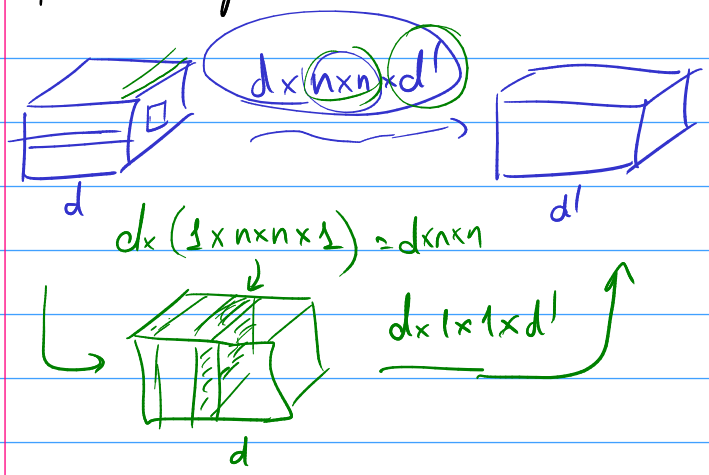
$$\left(\frac{\partial x_i^{(k+1)}}{\partial x_j^{(k)}}\right) = \nabla f + I$$

$$f : x^{(k)} \to x^{(k+1)}$$

$$\tilde{f} : x^{(k)} \to x^{(k+1)} - x^{(k)}$$

8) Neural architecture search

9) Depthwise separable convolutions



10) DenseNet