③ RBF kernel (Gaussian kernel)

For any $\alpha > 0$ $\quad \phi: \mathbb{R}^d \rightarrow \mathbb{R}^\infty$

$$K(x, x') \quad \phi(x)^T \phi(x') = \exp\left(-\frac{\|x - x'\|_2^2}{2\sigma^2}\right)$$

$\boxed{\text{Deep Network 1}}$

$\boxed{\text{Basic structure}}$:

iterated linear predictor.

1 layer: $\quad x \rightarrow W_1 x + b_1$

2 layers: $\quad x \rightarrow W_2 (W_1 x + b_1) + b_2$

3 layers: $\quad x \rightarrow W_3 (W_2 (W_1 x + b_1) + b_2) + b_3$

L layers $\quad x \rightarrow W_L (\cdots (W_1 x + b_1) \cdots) + b_L$

$$W_2(W_1 x + b_1) + b_2$$
$$= W_2 W_1 x + [W_2 b_1 + b_2]$$

$$W_L(\cdots(W_1 x + b_1)\cdots) + b_L$$
$$= \boxed{(W_L \cdots W_1)x + (b_L + W_L b_{L-1} + \cdots + W_L \cdots W_2 b_1)}$$
$$= w^T \begin{bmatrix} x \\ 1 \end{bmatrix}$$

$w \in \mathbb{R}^{d+1}$

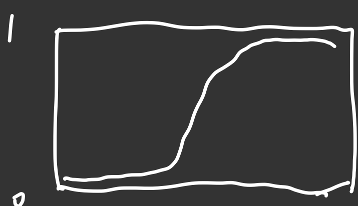
$w^T_{1:d} = W_L \cdots W_1$ $\quad$ $w_{d+1} = b_L + W_L b_{L-1} + \cdots + W_L \cdots W_2 b_1$

Just a linear predictor

$\boxed{\text{Activations / Nonlinearities}}$

$$\Pr[Y=1 \mid X=x] = \frac{1}{1+\exp(-w^T x)} =: \sigma_s(w^T x)$$

σ_s logistic or sigmoid function



$$W_2 \sigma(W_1 x + b) + b_2$$

$\boxed{\text{Classical deep network}}$

$$x \to \sigma_L(W_L \sigma_{L-1}(\cdots(W_2 \sigma_1(W_1 x + b_1) + b_2)\cdots) + b_L)$$

R^{d_2}

R^{d}

R^{d_1}

$$\equiv x \to (f_L \circ \cdots \circ f_1)(x) \qquad f_i(z) = \sigma_i(W_i z + b_i)$$

weights $\quad (W_i)_{i=1}^{L} \in R^{d_i \times d_{i-1}}$

biases $\quad (b_i)_{i=1}^{L}$

activations $\quad (\sigma_i)_{i=1}^{L} \qquad \sigma_i : R^{d_i} \to R^{d_i}$

$\boxed{(W_i, b_i)_{i=1}^{L} \ \text{params}}$

choices

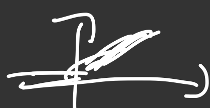
① binarization: $z \to 1[z \geq 0] \in \{0, 1\}$

② sigmoid: $\sigma(z) := \frac{1}{1+\exp(-z)}$

③ Hyperbolic tangent: $z \to \tanh(z)$

④ Rectified Linear Unit (ReLU)

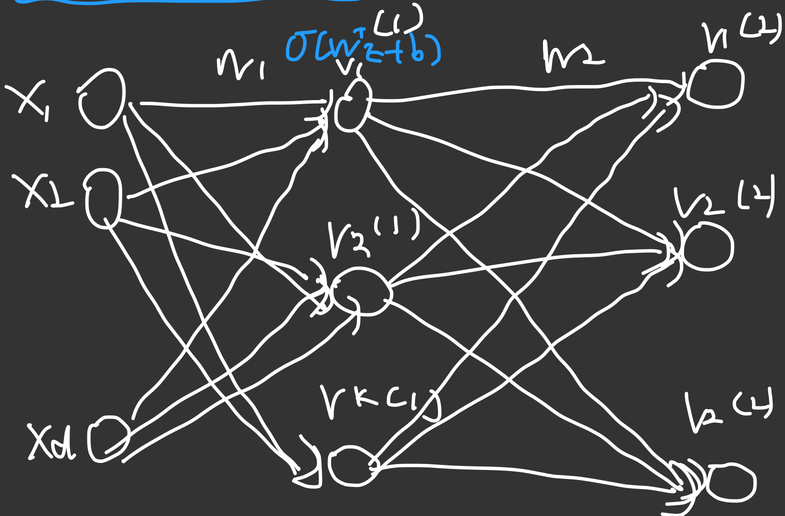
$\sigma_r(z) = \max\{0, z\}$ E.g. (imagenet)



⑤ identity: $z \to z$ last layer. when cross entropy (or.

Multilayer Neural network

$[W_1][x] \to \mathbb{R}$



① Columns of $W_1 \in \mathbb{R}^{d \times k}$: params of original logistic regression models.

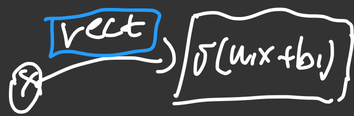
② columns of $W_2 \in \mathbb{R}^{k \times k}$: params of new logistic regression models to combine prediction of original models.

③ $\boxed{\text{noninput}}$ nodes compute $z \to \underbrace{\sigma(w^T z + b)}$ for some (w, b)

④ non-input or non-output units are $\boxed{\text{hidden}}$

Current "computation graph" perspective.

① Edges can pass full tensors



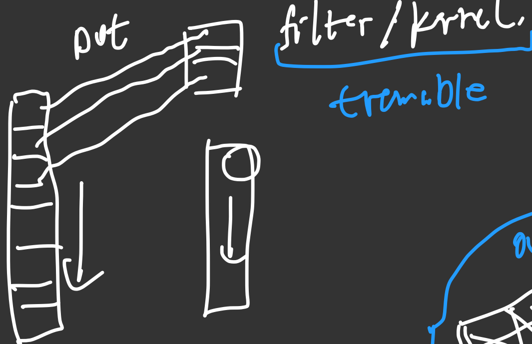
② Nodes are more general primitives

③ Edges skip layers

Convolutional layers

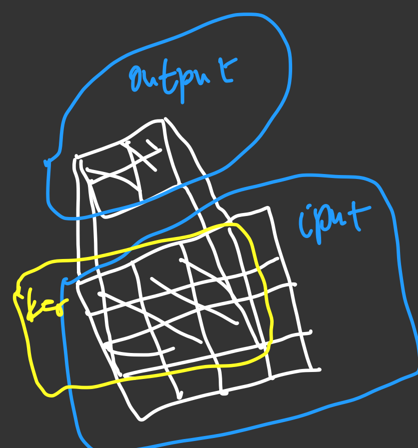
(linear)

① 1D convolution



② 2D convolution

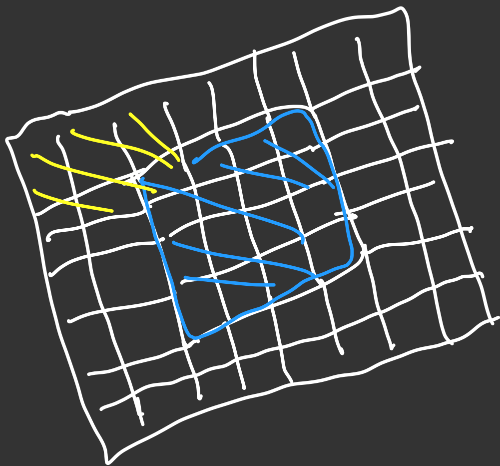
not trainable →



③ padding

kernal

org



④ Strides : step size

$\boxed{\text{channels}}$

extra below