

③ RBF kernel (Gaussian kernel)

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

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

Deep Network 1

Basic structure:

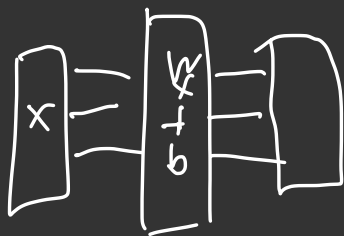
iterated linear predictor.

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

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

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

L layers: $x \rightarrow W_L (\dots (W_1 x + b_1) \dots) + b_L$



$$W_2(W_1x + b_1) + b_2$$

$$= W_2W_1x + [W_2b_1 + b_2]$$

$$W_L (\dots (W_1x + b_1) \dots) + b_L$$

$$= [W_L \dots W_1]x + [b_L + W_L b_{L-1} + \dots + W_L \dots W_2 b_1]$$

$$= W^T \begin{bmatrix} x \\ 1 \end{bmatrix}$$

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

$$\downarrow$$

$$W_{1:d}^T = W_L \dots W_1$$

$$w_{d+1} = b_1 + W_L b_{L-1}$$

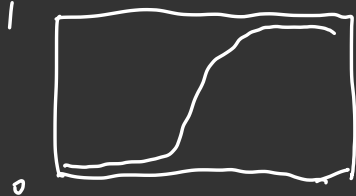
$$+ \dots + W_L \dots W_2 b_1$$

Just a linear predictor

Activations / Nonlinearities

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

σ_s logistic or sigmoid function



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

Classical deep network

$$x \mapsto \sigma_L \left(W_L \sigma_{L-1} \left(\dots \left(W_2 \sigma_1 \left(\boxed{W_1 x + b_1} \right) + b_2 \right) \dots \right) + b_L \right)$$

Diagram illustrating the nested structure of a classical deep network. The input x is passed through a series of layers. The innermost operation is $W_1 x + b_1$, which is then followed by σ_1 , W_2 , σ_2 , and so on, up to W_L and σ_L . The diagram uses nested boxes to show the sequence of operations, with dimensions d_1 and d_2 indicated for the intermediate layers.

$$\equiv x \mapsto (f_L \circ \dots \circ f_1)(x) \quad f_i(z) = \sigma_i(W_i z + b_i)$$

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

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

activations $(\sigma_i)_{i=1}^L$

$$\sigma_i: \mathbb{R}^{d_i} \rightarrow \mathbb{R}^{d_i}$$

$(W_i, b_i)_{i=1}^L$ parameters

choices (b)indization: $z \mapsto \mathbb{1}[z \geq 0] \in \{0, 1\}$

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

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

④ Rectified Linear Unit (RELU)

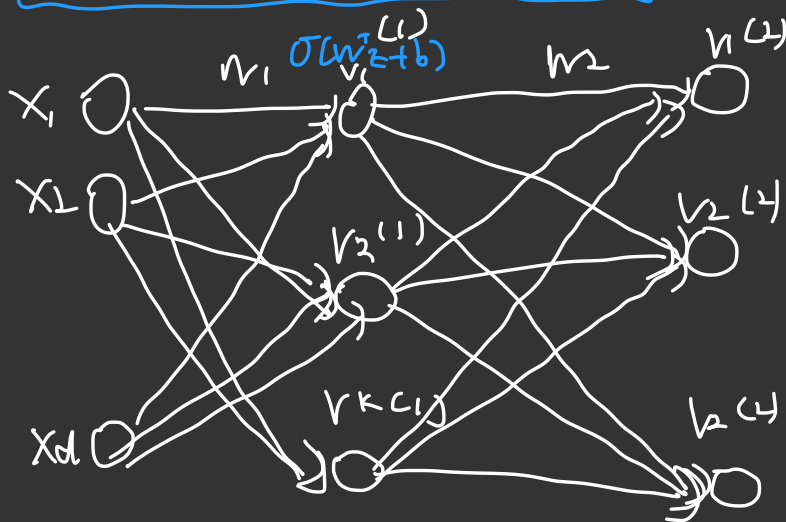
$$\sigma_r(z) = \max\{0, z\} \quad \text{E.g. Cimgenet}$$



⑤ identity: $z \mapsto z$ last layer when cross entropy (or)

Multi-layer Neural network

$$[w_1] [x] = z$$



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 manipulate nodes compute $z \rightarrow \sigma(w^T z + b)$ for some (w, b)

④ non-input or non-output units are hidden

Current "computational graph" perspective.

① Edges can pass full tensors

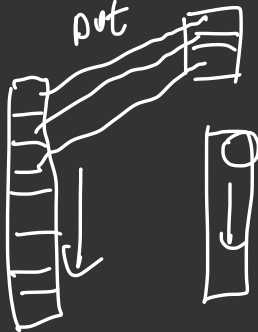
⊗ $\xrightarrow{\text{vect}}$ $\sqrt{5}(ux+bi)$

② Nodes are more general primitives

③ Edges skip layers

Convolutional layers (Linear)

① 1D convolution

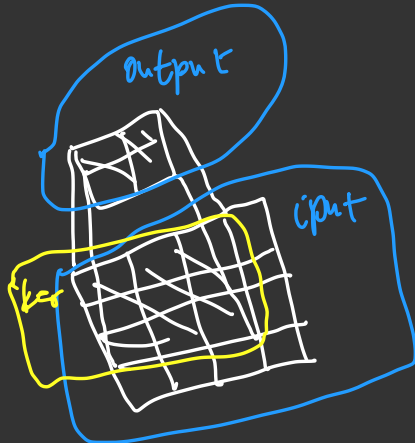


filter/kernel,
stridable

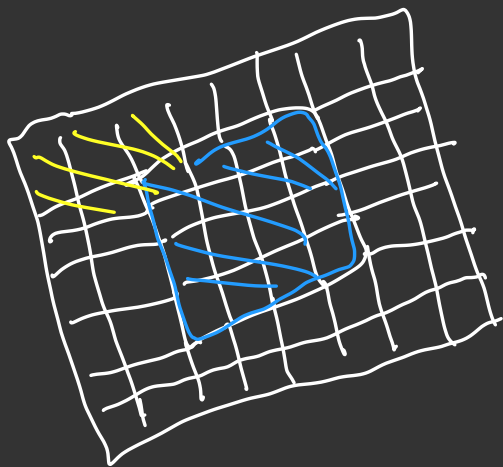
② 2D

convolution

not stridable →



③ padding



~~kernel~~ kernel

0 or 1

④ Strides: step size

channels

extra below

