The "why" of deep learning

1. Usually: We tried it and it helped.
2. Sometimes: Intuition.
3. Rarely: Theory. (post-hoc)

Residual connection (or "skip"):

Assume we have a "block" $f(x)$

Residual connection: $x + f(x)$

Note: $f(x)$ could change the shape.

So, we can apply a "simple" transformation along the residual path.

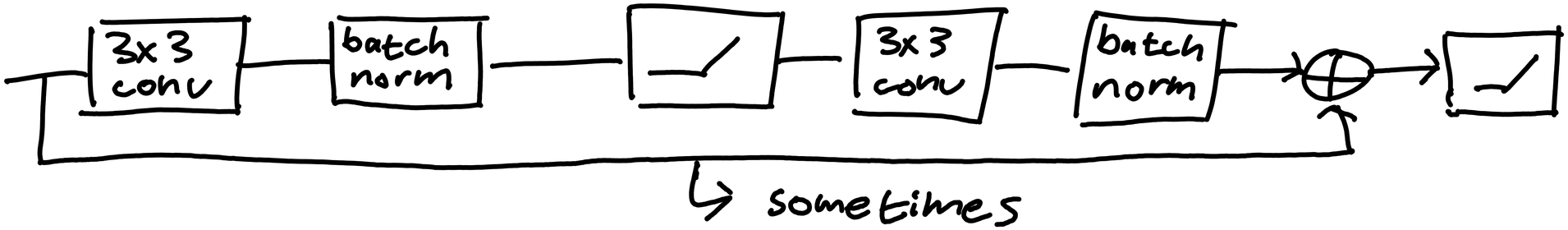
ex 1×1 convolution.

(filter size 1×1)

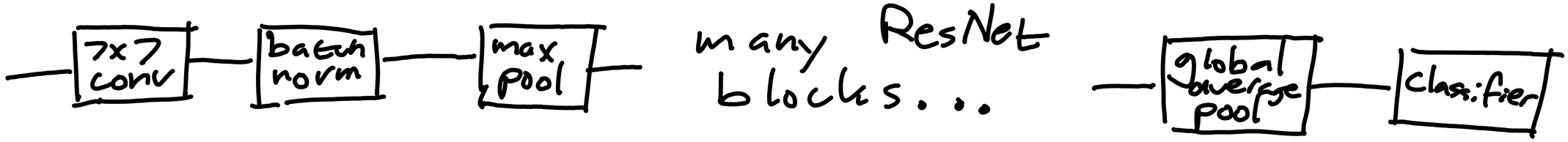
Can change # channels and stride.

ResNet

block:



ResNet:



global average pool averages over length and width.

Sequences and RNNs

So far: $p(y \mid x)$

- y: output (class)
- x: input (features)

Now, consider a <u>sequence</u>, a series of random variables that have a natural order:

$x_1, x_2, x_3, \ldots x_T$ ↳ length-T sequence

$p(x_1, x_2, x_3, \ldots, x_T) = p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_2, x_1) \ldots$

"order" implies natural decomposition

(any other order feels unnatural)

Note:

If we model $p(x_1)$, $p(x_2 \mid x_1)$, $p(x_3 \mid x_2, x_1)$...

Then we can sample autoregressively:

$$x_1 \sim p(x_1)$$

$$x_2 \sim p(x_2 \mid x_1)$$

$$x_3 \sim p(x_3 \mid x_2, x_1)$$

$$\vdots$$

Main goal:

$$p(x_t \mid \underbrace{x_{t-1}, x_{t-2}, \ldots, x_1}_{\text{variable-length prefix}})$$

single model

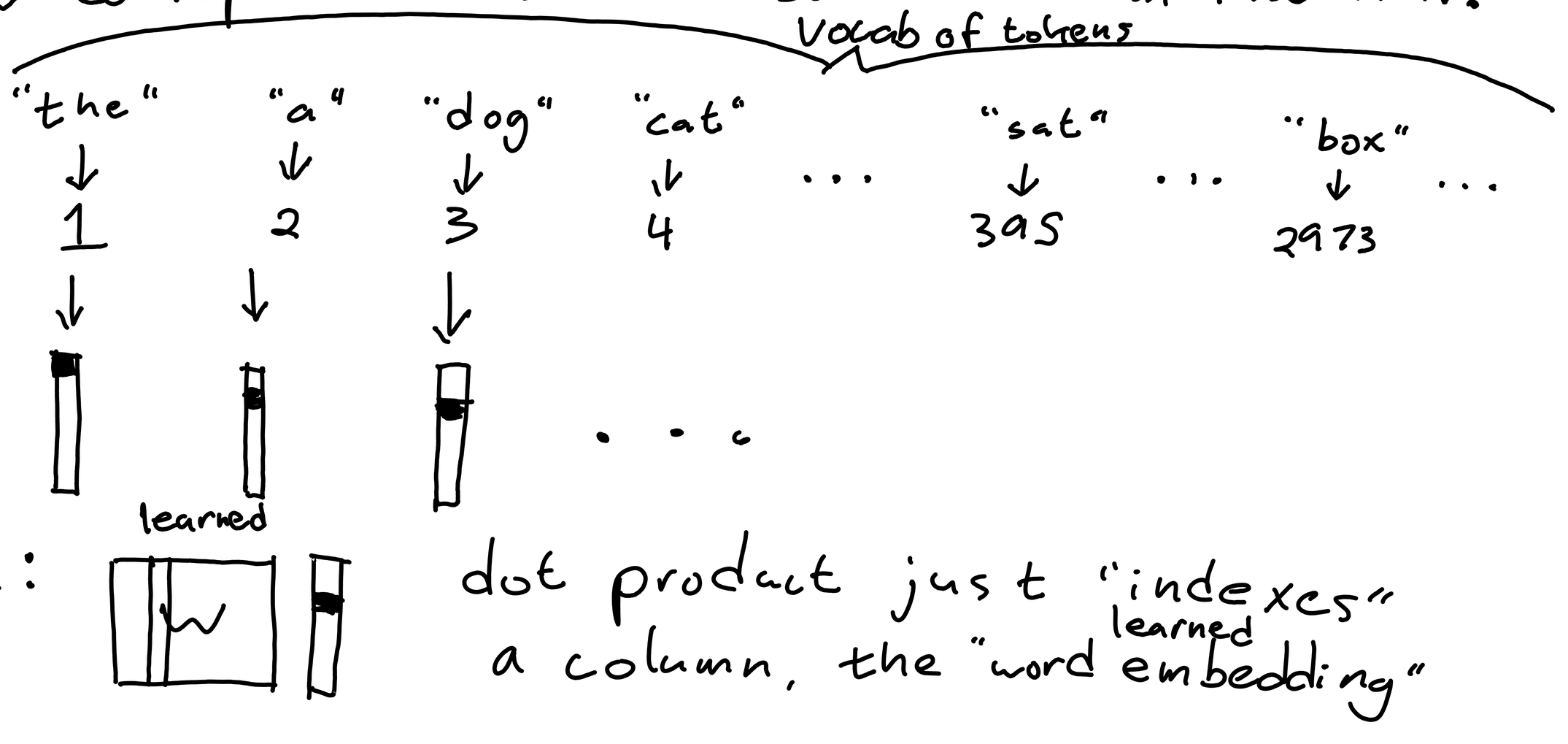
ex/ Text data (sequence of words/characters/...)

the cat sat

x_1 x_2 x_3

$p(\text{the cat sat}) = p(\text{the})\, p(\text{cat}|\text{the})\, p(\text{sat}|\text{the cat})$

How to represent words to feed them into a NN:



Recall that we want to model

$$p(x_t \mid x_{t-1}, \dots, x_1)$$

This is $|V|$-way classification.

Vanilla recurrent neural network

Recall: MLP with 1 hidden layer

$$h = \phi(W_h x + b_h)$$
$$o = W_o h + b_o$$

Simplest possible change to process a sequence:

$$h_t = \phi(W_{hx} x_t + \underbrace{W_{hh} h_{t-1}}_{\text{recurrence!}} + b_h)$$
$$o_t = W_o h_t + b_o$$

h_t can contain information about every past input! h_t "summarizes" the sequence up to step t.

Note: Can process any length input.
# of parameters doesn't change

Backpropagation through time:

$$h_t = f(x_t, h_{t-1}, w_h)$$

↳ same as w_{hh} from before

$$o_t = g(h_t, w_o)$$

$$L = \frac{1}{T}\sum_{t=1}^{T} \ell(y_t, o_t)$$

$$\frac{dL}{dw_h} = \frac{1}{T}\sum_{t=1}^{T} \frac{d\ell(y_t, o_t)}{dw_h}$$

$$= \frac{1}{T}\sum_{t=1}^{T} \frac{d\ell(y_t, o_t)}{do_t} \frac{do_t}{dh_t} \underbrace{\frac{dh_t}{dw_h}}_{\text{hard one}}$$

$$\frac{dh_t}{dw_n} = \frac{df(x_t, h_{t-1}, w_n)}{dw_n} + \frac{df(x_t, h_{t-1}, w_n)}{dh_{t-1}} \frac{dh_{t-1}}{dw_n}$$

$$a_t = \frac{dh_t}{dw_n} \quad b_t = \frac{df(x_t, h_{t-1}, w_n)}{dw_n} \quad c_t = \frac{df(x_t, h_{t-1}, w_n)}{dh_{t-1}}$$

$$a_t = b_t + c_t a_{t-1}$$

$$a_0 = 0$$

$$a_1 = b_1 + c_1 a_0 = b_1$$

$$a_2 = b_2 + c_2 a_1 = b_2 + c_2 b_1$$

$$a_3 = b_3 + c_3 a_2 = b_3 + c_3 b_2 + c_3 c_2 b_1$$

$$a_4 = b_4 + c_4 a_3 = b_4 + c_4 b_3 + c_4 c_3 b_2 + c_4 c_3 c_2 b_1$$

$$a_t = b_t + \sum_{i=1}^{t-1} \left(\prod_{j=i+1}^{t} c_j \right) b_i$$

$$\frac{dh_t}{dw_h} = \frac{df(x_t, h_{t-1}, w_n)}{dw_n} + \sum_{i=1}^{t-1} \left(\prod_{j=i+1}^{t} \frac{df(x_j, h_{j-1}, w_n)}{dh_{j-1}} \right) \frac{df(x_i, h_{i-1}, w_h)}{dw_n}$$

$$\frac{dh_t}{dw_h} = \frac{df(x_t, h_{t-1}, w_h)}{dw_h} + \sum_{i=1}^{t-1} \left(\prod_{j=i+1}^{t} \frac{df(x_j, h_{j-1}, w_h)}{dh_{j-1}} \right) \frac{df(x_i, h_{i-1}, w_h)}{dw_h}$$

ex/ Linear RNN:

$$f(x_t, h_{t-1}, w_h) = w_x x_t + w_h h_{t-1} + b_h$$

$$\frac{dh_t}{dw_h} = h_{t-1} + \sum_{i=1}^{t-1} \underbrace{\left(\prod_{j=i+1}^{t} w_h \right)}_{\text{gnarly!}} h_{i-1}$$

w_h can explode or vanish

Long short-term memory: (LSTM)

We want explicit mechanisms for:

- Remember/ignore the previous state
- Update/don't update state based on a new input
- Forget the past completely

LSTM adds a "cell" which has explicit read/write/update

$$I_t = \sigma(W_{xi} x_t + W_{hi} h_{t-1} + b_i) \quad \text{input gate}$$

(σ: sigmoid nonlinearity)

$$F_t = \sigma(W_{xf} x_t + W_{hf} h_{t-1} + b_f) \quad \text{forget gate}$$

$$O_t = \sigma(W_{xo} x_t + W_{ho} h_{t-1} + b_o) \quad \text{output gate}$$

$$\tilde{C}_t = \tanh(W_{xc} x_t + W_{hc} h_{t-1} + b_c) \quad \text{candidate cell}$$

$$C_t = F_t \odot C_{t-1} + I_t \odot \tilde{C}_t$$

$$h_t = O_t \odot \tanh(C_t)$$