Recurrent Neural Networks

Consider the following problem:

input: beginning of a sentence

output: next word

e.g. The answer to your (question)

A feedforward NN is not good at solving this kind of problem.

Reasons – Cannot handle variable length inputs
– No parameter sharing

Solution: Recurrent Neural Networks

States Systems: Consider a dynamical system of the form

$$h_t = f(h_{t-1}; \theta) \qquad t \geq 1 \quad (h_0 \text{ given})$$

- $f: \mathbb{R}^k \times \mathbb{R}^m \to \mathbb{R}^k$ (Borel Measurable)
- θ index of t

E.g. $f(x; \theta) = \sigma(\theta x)$ $\quad \theta \in \mathbb{R}$ $\quad |\theta| < 1$

(What is .5708?)

consider $h_0 = 2$ $\quad \theta = .5$

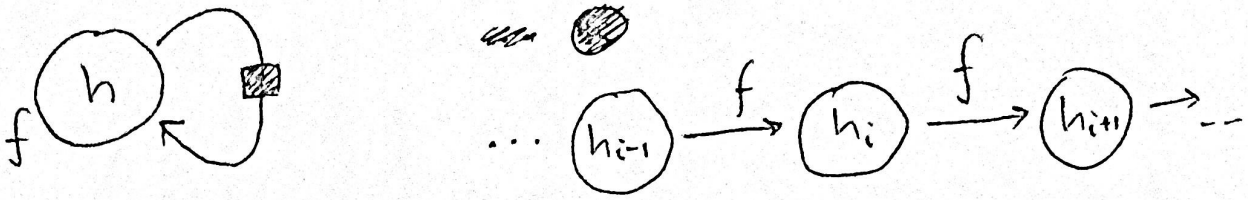
$h_1 = \sigma(1) \approx .73$

$h_2 = \sigma\left(\frac{.73}{2}\right) \approx .59$

$\ldots h_{100} = .5708 \ldots$

converges to .5708 (true for any value of h_0)

Two ways to "graph" a dynamical system like this:

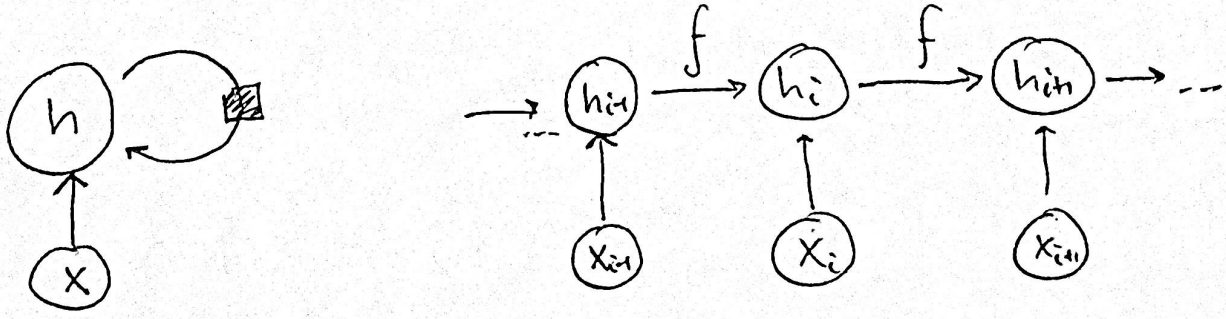


▨ : time step

Circuit diagram unfolded

Can also consider a dynamical system "driven by external signal" X_t

$$h_t = f(h_{t-1}, X_t; \theta)$$



circuit diagram unfolded

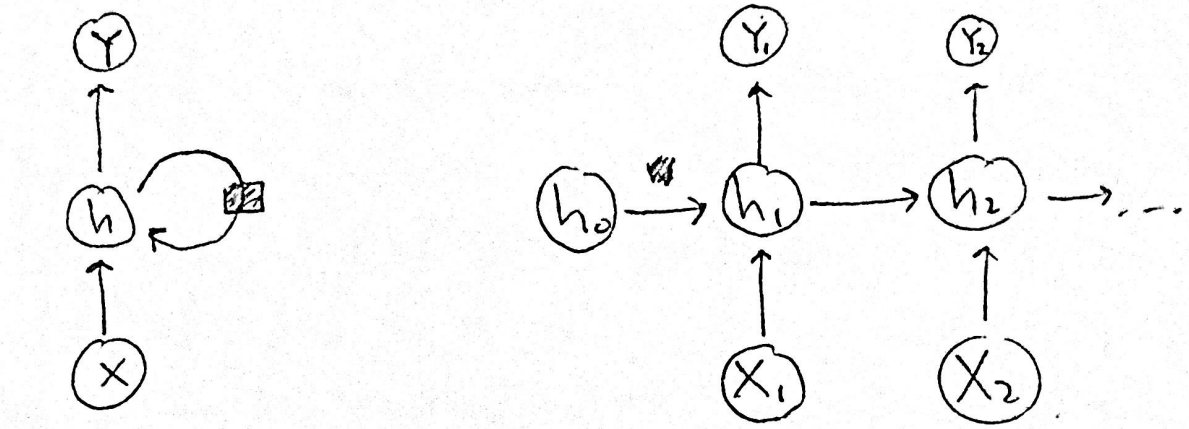
Recurrent Neural Networks (RNN)

An RNN is a state system $h_t = f(h_{t-1}, X_t; \theta)$ together with a prescribed way to associate an ~~output~~ output Y_t at each time step $t > 0$

h_t : hidden state at time t

terminology: X_t : t^{th} input

Y_t : t^{th} output



Circuit

unfolded

Standard Equations:

$$h_t = \tanh(W h_{t-1} + U X_t + b)$$

$$Y_t = V h_t + C$$

$$\theta = (\underbrace{W, U}_{\text{matrices}}, \underbrace{b, c}_{\text{vectors}})$$

W: hidden-to-hidden params

U: input-to-hidden "

V: hidden-to-output * (fixed)

b, c: bias vectors

Another way to think of an RNN:

a sequence of feedforward neural networks

(each with one hidden layer)

first: $\begin{pmatrix} h_0 \\ X_1 \end{pmatrix} \longrightarrow h_1 \xrightarrow{V} Y_1$

$\begin{pmatrix} h_1 \\ X_2 \end{pmatrix} \longrightarrow h_2 \xrightarrow{V} Y_2$

$\vdots$

Loss functions Consider an RNN with inputs $X_1, \dots, X_T$ and outputs $Y_1, \dots, Y_T$ and consider target $Z_1, \dots, Z_T$.

Common choice for loss function is

$$L((Y_1, \dots, Y_T), (Z_1, \dots, Z_T)) = \sum_{t=1}^{T} \frac{1}{2} \| Y_t - Z_t \|_2^2$$

but can also use CE BCE (depends on context)

E.g. Classifying reviews of a restaurant

"The food was just terrible"

"It was great"

"I had a good time but the waiter was annoying"

E.g. Classifying reviews of a restaurant

"The food was just terrible"

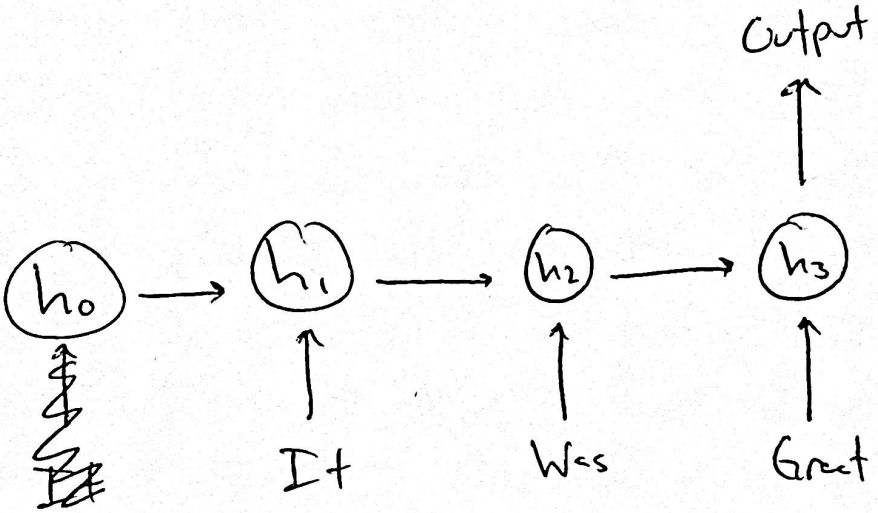
"It was great"

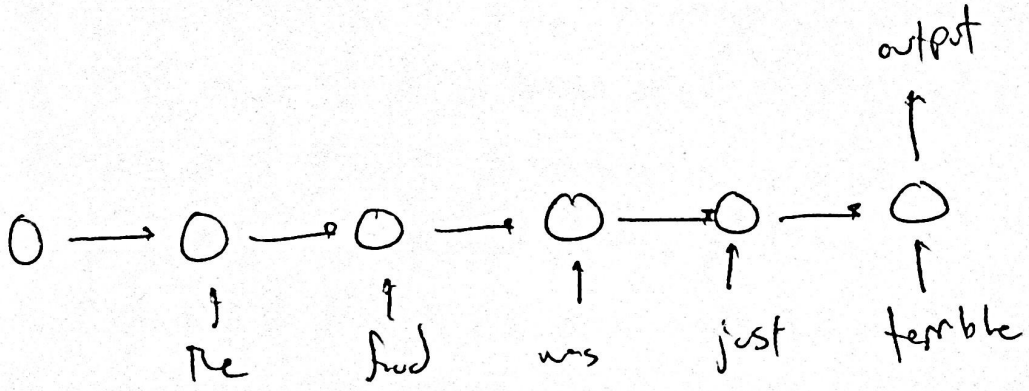
"I had a good time but the waiter was annoying"

X_1 = "~~It~~"

X_2 = "was"

X_3 = "great"





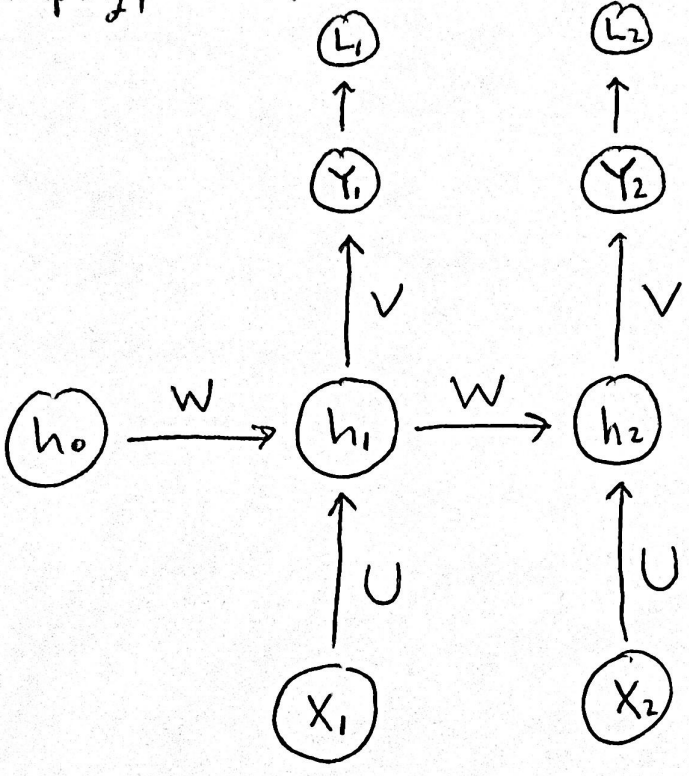
(So input length can vary!)

RNN Backpropagation

Mistake last time: V is not fixed, it is a parameter

so $\theta = (W, V, U, b, c)$

To simplify, consider an RNN with 2 steps



Write $a_1 = Wh_0 + UX_1 + b$ $\quad h_1 = \tanh(a_1)$ $\quad Y_1 = Vh_1 + c$

$a_2 = Wh_1 + UX_2 + b$ $\quad h_2 = \tanh(a_2)$ $\quad Y_2 = Vh_2 + c$

Using square error loss $L = \frac{1}{2}(Y_1 - Z_1)^2 + \frac{1}{2}(Y_2 - Z_2)^2$

$= L_1 + L_2$

We want to compute $\nabla_\theta L = \left(\frac{\partial L}{\partial W}, \frac{\partial L}{\partial V}, \frac{\partial L}{\partial U}, \frac{\partial L}{\partial b}, \frac{\partial L}{\partial c}\right)$

L_1 depends on W only through h_1

L_2 " " " through h_1 and h_2, so

$$\frac{\partial L}{\partial W} = \frac{\partial L_1}{\partial W} + \frac{\partial L_2}{\partial W} = \frac{\partial L_1}{\partial h_1}\frac{\partial h_1}{\partial W} + \frac{\partial L_2}{\partial h_1}\frac{\partial h_1}{\partial W} + \frac{\partial L_2}{\partial h_2}\frac{\partial h_2}{\partial W}$$

$$\frac{\partial L}{\partial W} = \frac{\partial L_1}{\partial h_1}\frac{\partial h_1}{\partial W} + \frac{\partial L_2}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W} + \frac{\partial L_2}{\partial h_2}\frac{\partial h_2}{\partial W}$$

$$\frac{\partial L}{\partial V} = \frac{\partial L_1}{\partial V} + \frac{\partial L_2}{\partial V} = \frac{\partial L_1}{\partial Y_1}\frac{\partial Y_1}{\partial V} + \frac{\partial L_2}{\partial Y_2}\frac{\partial Y_2}{\partial V}$$

$$\frac{\partial L}{\partial V} = (Y_1 - Z_1)h_1 + (Y_2 - Z_2)h_2 = \sum_t (Y_t - Z_t)h_2$$

$$\frac{\partial L}{\partial U} = \frac{\partial L_1}{\partial U} + \frac{\partial L_2}{\partial U} = \frac{\partial L_1}{\partial h_1}\frac{\partial h_1}{\partial U} + \frac{\partial L_2}{\partial h_1}\frac{\partial h_1}{\partial U} + \frac{\partial L_2}{\partial h_2}\frac{\partial h_2}{\partial U}$$

$$\frac{\partial L}{\partial U} = \frac{\partial L_1}{\partial h_1}\frac{\partial h_1}{\partial U} + \frac{\partial L_2}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial U} + \frac{\partial L_2}{\partial h_2}\frac{\partial h_2}{\partial U}$$

$$\frac{\partial L}{\partial b} = \frac{\partial L_1}{\partial b} + \frac{\partial L_2}{\partial b} = \frac{\partial L_1}{\partial h_1}\frac{\partial h_1}{\partial b} + \frac{\partial L_2}{\partial h_1}\frac{\partial h_1}{\partial b} + \frac{\partial L_2}{\partial h_2}\frac{\partial h_2}{\partial b}$$

$$\frac{\partial L}{\partial b} = \frac{\partial L_1}{\partial h_1}\frac{\partial h_1}{\partial b} + \frac{\partial L_2}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial b} + \frac{\partial L_2}{\partial h_2}\frac{\partial h_2}{\partial b}$$

$$\frac{\partial L}{\partial c} = \frac{\partial L_1}{\partial c} + \frac{\partial L_2}{\partial c} = \frac{\partial L_1}{\partial Y_1}\frac{\partial Y_1}{\partial c} + \frac{\partial L_2}{\partial Y_2}\frac{\partial Y_2}{\partial c}$$

$$\frac{\partial L}{\partial c} = (Y_1 - Z_1) + (Y_2 - Z_2) = \sum_t (Y_t - Z_t)$$

Suffices to compute:

$$\frac{\partial L_1}{\partial h_1}, \frac{\partial h_1}{\partial W}, \frac{\partial L_2}{\partial h_2}, \frac{\partial h_2}{\partial h_1}, \cancel{\frac{\partial L_2}{\partial h_2}}, \frac{\partial h_2}{\partial W}, \frac{\partial h_1}{\partial U}, \frac{\partial h_2}{\partial U}, \frac{\partial h_1}{\partial b}, \frac{\partial h_2}{\partial b}$$

$$\frac{\partial L_1}{\partial h_1} = \frac{\partial L_1}{\partial Y_1}\frac{\partial Y_1}{\partial h_1} = (Y_1 - Z_1)V$$

$$\frac{\partial L_2}{\partial h_2} = \frac{\partial L_2}{\partial Y_2}\frac{\partial Y_2}{\partial h_2} = (Y_2 - Z_2)V$$

$$\frac{\partial h_1}{\partial W} = \frac{\partial}{\partial W}\tanh(a_1) = \operatorname{sech}^2(a_1)\frac{\partial a_1}{\partial W}$$

$$= (1 - \tanh^2(a_1))h_0$$

$$= (1 - h_1^2)h_0$$

$$\frac{\partial h_2}{\partial W} = \frac{\partial}{\partial W}\tanh(a_2) = \operatorname{sech}^2(a_2)\frac{\partial a_2}{\partial W}$$

$$= \left(1 - \tanh^2(a_2)\right)\frac{\partial a_2}{\partial W}$$

$$= \left(1 - h_2^2\right)h_1$$

$$\frac{\partial h_2}{\partial h_1} = \frac{\partial}{\partial h_1}\tanh(a_2) = \operatorname{sech}^2(a_2)\frac{\partial a_2}{\partial h_1}$$

$$= (1 - h_2^2)W$$

$$\frac{\partial h_1}{\partial U} = \frac{\partial}{\partial U}\tanh(a_1) = (1-h_1^2)\frac{\partial a_1}{\partial U} = (1-h_1^2)X_1$$

similarly $\frac{\partial h_2}{\partial U} = (1-h_2^2)X_2$

$$\frac{\partial h_1}{\partial b} = \frac{\partial}{\partial b}\tanh(a_1) = \operatorname{sech}^2(a_1)\frac{\partial a_1}{\partial b} = (1-h_1^2)$$

similarly $\frac{\partial h_2}{\partial b} = (1-h_2^2)$