Examples: Named Entity Recognition

1) Benedict lives in Easthampton, Massachusetts.

Classify each word as a $\underset{0}{\text{person}}$, $\underset{1}{\text{location}}$, or $\underset{2}{\text{other}}$

$x^{<1>}$ = "Benedict", $x^{<2>}$ = "lives", $x^{<3>}$ = "in", $x^{<4>}$ = "Easthampton", $x^{<5>}$ = "Massachusetts"

$y^{<1>} = 0 \quad y^{<2>} = 2 \quad y^{<3>} = 2 \quad y^{<4>} = 1 \quad y^{<5>} = 1$

$T_x = 5, \ T_y = 5$

---

2) This movie was the best!

Classify whole sentence as $\underset{1}{\text{positive}}$ or $\underset{0}{\text{negative}}$

$T_x = 5, \quad T_y = 1$

$x^{<1>}$ = "This", ..., $x^{<5>}$ = "best" $\qquad y^{<1>} = 1$

Order matters!

"The cat chased the mouse."

vs.

"The mouse chased the cat."

---

3) Translation:

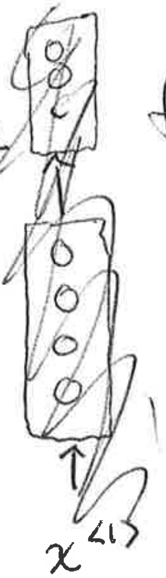
Why is the cat so cute? → Porqué es el gato tan lindo?

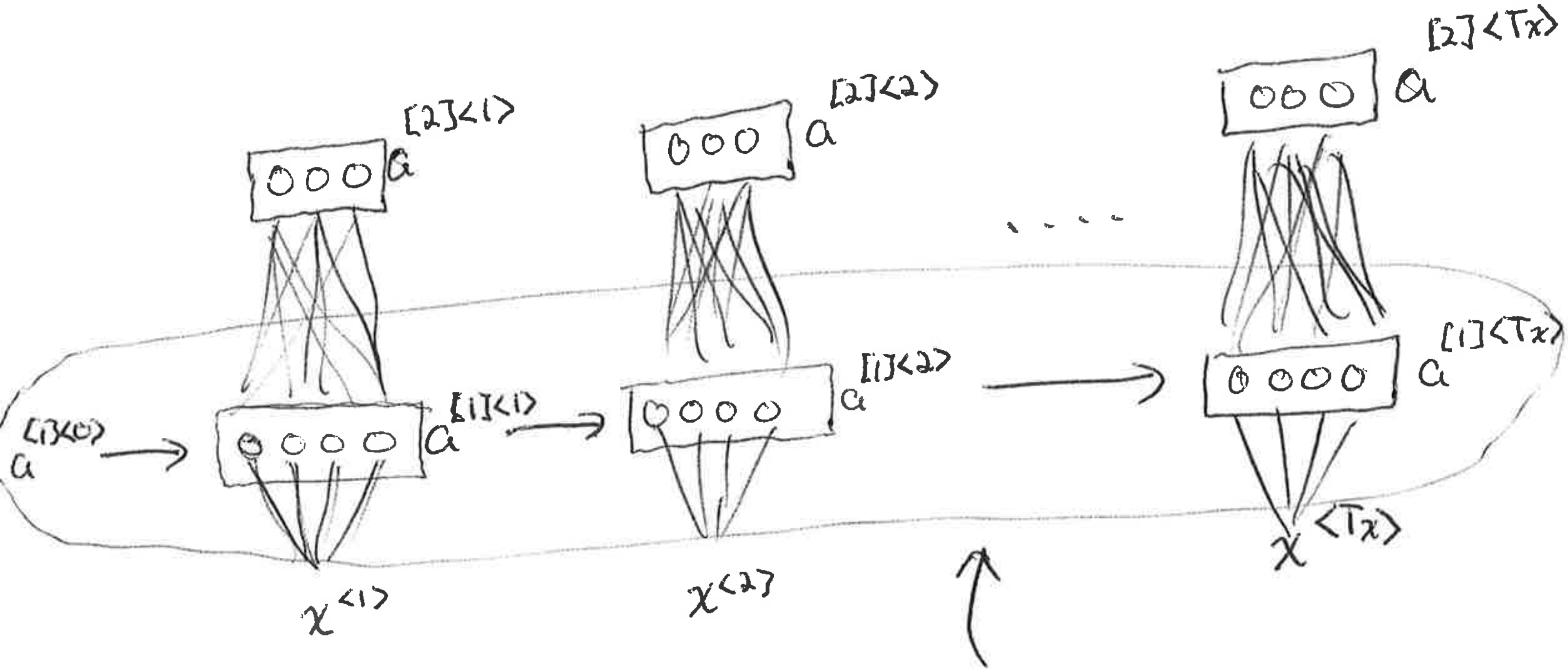
$T_x = 6 \qquad T_y = 7$

Notation:

$a^{(i)[l]<t>}$: activation in layer l at time t for observation i.

Picture: for named entity recognition example





a NN that takes in first word, predicts type of word (person, location, other)

this (circled) is a <u>recurrent layer</u>

Forward Prop.:

$$a^{[1]\langle t\rangle} = g^{[1]}\left((W_{aa}^{[1]})^T \cdot a^{[1]\langle t-1\rangle} + (W_{ax}^{[1]})^T x^{\langle t\rangle} + b^{[1]}\right)$$

$$= g^{[1]}\left(\left[(W_{aa}^{[1]})^T \; (W_{ax}^{[1]})^T\right]\begin{bmatrix} a^{[1]\langle t-1\rangle} \\ x^{\langle t\rangle} \end{bmatrix} + b^{[1]}\right)$$

For example suppose we have 4 units in each box of the recurrent layer, and 100 input features.

W_{aa} is 4×4

W_{ax} is 100×4

b_a is 4×1

For recurrent layers, g is almost always tanh activation (ReLU also occasionally used)
- ↳ prevents exploding gradients
- ↳ vanishing gradients is a serious problem we will address through other strategies.

In our example,

$$a^{[2]\langle t\rangle} = g^{[2]}\left((W^{[2]})^T a^{[1]\langle t\rangle} + b^{[2]}\right)$$

$g^{[2]}$ is whatever appropriate activation for your task (eg. softmax for named entity recognition with 3 classes)

Loss Function:

$$J(b,\omega) = \frac{1}{T_y} \underbrace{\sum_{i=1}^{m} \sum_{t=1}^{T_y^{(i)}} J^{(i)<t>}(b,\omega)}$$

basically, add up losses across all subjects and time points.

Formally, assumes all examples are independent. $(i = 1, \dots, m)$

Within one example, allows for dependence; sort of.

Distribution of $Y^{(i)<t>}$ makes use of $x^{(i)<1>}, \dots, x^{(i)<t>}$

Assumes conditional independence:

$$P(Y^{<1>} = y^{<1>}, Y^{<2>} = y^{<2>}, \dots, Y^{<t>} = y^{<t>} \mid x^{<1>}, \dots, x^{<t>})$$

$$= P(Y^{<1>} = y^{<1>} \mid x^{<1>}) \cdot P(Y^{<2>} = y^{<2>} \mid x^{<1>}, x^{<2>})$$

$$\cdot \dots \cdot P(Y^{<t>} = y^{<t>} \mid x^{<1>}, x^{<2>}, \dots, x^{<t>})$$

"If I know all of the inputs up to the current time, $x^{<1>}, \dots, x^{<t>}$, knowing $Y^{<t-1>}$ would not give me any additional information about $Y^{<t>}$."

↑ not realistic, but better than ignoring time.

Backward Propagation Through Time.

Start with the cost function <u>at the last time point.</u>

$$\frac{\partial J}{\partial w} = \sum_{i=1}^{m}\sum_{t=1}^{T_y^{(i)}} \cancel{J^{(i)<t>}} \frac{\partial}{\partial w} J^{(i)<t>}$$

$$\frac{\partial J^{(i)<t>}}{\partial w} = \frac{\partial J^{(i)<t>}}{\partial a^{[2]<t>}} \cdot \frac{\partial a^{[2]<t>}}{\partial a^{[1]<t>}} \cdot \frac{\partial a^{[1]<t>}}{\partial w}$$

$$a^{[1]<t>} = g^{[1]}\left((w^{[1]})^T \begin{bmatrix} a^{[1]<t-1>} \\ x^{<t>} \end{bmatrix} + b^{[1]}\right)$$

So

$$\frac{\partial}{\partial w} a^{[1]<t>} = \frac{\partial a^{[1]<t>}}{\partial z^{[1]<t>}} \circ \frac{\partial z^{[1]<t>}}{\partial w}$$

$$= \frac{\partial a^{[1]<t>}}{\partial z^{[1]<t>}} \left(\frac{\partial}{\partial w} w_a^{[1]T} a^{[1]<t-1>} + \frac{\partial}{\partial w} (w_x^{[1]})^T x^{<t>} \right)$$

W also need to calculate $a^{[1]<t-1>}$.