1) Input: $x^t = [0\,1\,0\,1\,1\,0]$ expected output $y^t = [0\,1\,1\,0\,1\,1]$

y^{t-1}	x^t	y^t
0	0	0
0	1	1
1	0	1
1	1	0
0	1	1
1	0	1

Notice this is an XOR function. Not linearly seperable so we can use more than 1 hidden states to build our RNN

$\therefore$ XOR can be formed using AND, OR, and NOT

3 hidden states for each operation

h_1^t: h_3^{t-1} AND x^t

h_2^t: h_3^{t-1} OR x^t

h_3^t: NOT(h_1^t) AND h_2^t,

y^t : h^t

So our equations, weights, and biases are as follows:

$$h_1^t = f(h_3^{t-1}, x^t) = h_3^{t-1} + x^t - 1.5$$

$$W_1 = [1\ 1] \quad b_1 = -1.5$$

$$h_2^t = f(h_3^{t-1}, x^t) = h_3^{t-1} + x^t - 0.5$$

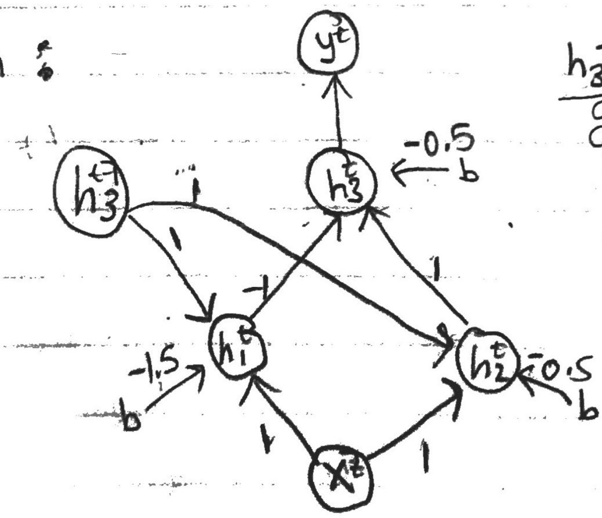
$$W_2 = [1\ 1] \quad b_2 = -0.5$$

$$h_3^t = f(h_1^t, h_2^t) = -h_1^t + h_2^t - 0.5$$

$$W_3 = [-1, 1] \quad b_3 = -0.5$$

$$y_3^t = h_3^t$$

Diagram:



h_3^{t-1}	x^t	h_1^t	h_2^t	h_3^t	y^t
0	0	0	0	0	0
0	1	0	1	1	1
1	0	0	1	1	1
1	1	1	1	0	0
0	1	0	1	1	1
1	0	0	1	1	1

2. Notice $C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$

closley resembles the XOR function $\bar{x} \wedge y \vee x \wedge \bar{y}$

lets assume $C_{t-1} = h_{t-1}$

lets assume $\tilde{C}_t$ = NOT (h_{t-1}) (this is saying $\tilde{C}_t$ is the opposite of our previous cells parity, following our $\bar{y}$ notation.

We can say $C_{t-1} = y$

This leaves us with making $f_t = \bar{x}$ = NOT (x^{t-1})

and $i_t = x = (x^{t-1})$.

Since our cell state is already representing the bit parity, $h_t = C_t$ and

O_t is going to affect h_t so we just have to ensure that it produces a value that causes h_t to be C_t.

Now, lets use these assumptions to solve for weights & biases:

① W_f and b_f:

f_t = NOT (x^{t-1}) so $W_f = [0 \ -1]$ $b_f = 1$

② W_i and b_i:

$i_t = x^{t-1}$ so $W_i = [0 \ 1]$ $b_i = 0$

③ W_c and b_c:

$\tilde{C}_t$ = NOT(h_{t-1}) so $W_c = [-1 \ 0]$ $b_c = 1$

note since h_{t-1} is equal to C_{t-1}, $\tilde{C}_t$ is just NOT(C_{t-1})

④ W_o and b_o:

since we want O_t to not affect the outcome of h_t and $h_t = O_t * \tanh(C_t)$, we can enforce O_t to always be equal to 1.

so $W_o = [0 \ 0]$ $b_o = 1$

3) If at time t, the current highest scoring beam scores worse than or equal to $best_{\le t}$

$score(B_{i,1}) \le best_{\le t}$, that means that for all $j = 1 \dots t$, $score(B_{i,j}) \le best_{\le t}$ because the score is defined as the sum of log probabiliteres. probability is always ≤ 1 so $\log(x)$ when $x \le 1$ is always zero or negative so anything after $B_{i,1}$ has equal or lower score than $B_{i,1}$.

Therefore, all future steps will result in a lower score.

4) $h_0 \quad h_1 = W^T h_0 \quad h_2 = W^T h_1 = W^T(W^T h_0) = (W^T)^2 h_0$

so on.--- $\quad h_t = (W^T)^t h_0$

(W^T) eigen decomposition $\rightarrow \quad W^T = (Q \Lambda Q^{-1})^T = Q^T \Lambda^T (Q^{-1})^T$

$$(W^T)^2 = Q^T \Lambda \underbrace{\cancel{(Q^{-1})^T Q^T}}_{=1} \Lambda (Q^{-1})^T = Q^T \Lambda^2 (Q^{-1})^T$$

$$(W^T)^t = Q^T \Lambda^t (Q^{-1})^T$$

So $\quad h_t = (Q \Lambda^t Q^{-1})^T h_0 \quad$ where $\Lambda = \begin{bmatrix} \lambda_0 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$

Note that Λ is a diagonal matrix so $\Lambda^t = \begin{bmatrix} \lambda_0^t & & 0 \\ & \ddots & \\ 0 & & \lambda_n^t \end{bmatrix}$

for all i, if any eigenvalue λ_i is <1, $\lim_{t\to\infty} \lambda_i^t = 0$

if any eigenvalue λ_i is >1, $\lim_{t\to\infty} \lambda_i^t = \infty$

1. Therefore, if $\rho(W) < 1$, then all eigenvalue λ_i will tend to 0 as $t \to \infty$, therefore $(W^T)^t \to 0$ so $h_t \to 0$ and we will see vanishing gradients
2. if $\rho(W) = 1$, must check the other eigenvalues to gain insight.
3. On the other hand, if $\rho(W) > 1$, then some eigenvalue λ_i will tend to ∞, so $(W^T)^t \to \infty$, so $h_t \to \infty$ as $t \to \infty$.

6.

The key contributions of this paper was to introduce multimodal explanations for the decisions a deep neural network was making in the task of Visual Question Answering and Activity Recognition Tasks.They realized that traditional methods were unimodal and therefore for questions such as 'what color is the banana?', just descriptions were hard to justify the answer, and the same idea applies for just visualizations about where the network was 'looking' when it made its decision. Furthermore, current datasets had ground truth as descriptions of what was happening within an image, and there was no ground truth curated specifically for justifications for VQA and ACT. Therefore, in this paper a large contribution was that 2 new datasets were first collected VQA-X and ACT-X where tuckers had to describe the reasoning behind the ground truth answers, and then had to annotate the areas in the image corresponding to those reasonings. After that, the Pointing and Justification Explanation (PJ-X) was trained on these two datasets for the two different tasks to produce a multimodal justification model. The results were very promising. For textual justifications, the researchers performed ablation and compared with related approaches. The result was that the full version of the PJ-X with attention and descriptions outperformed the related approaches and all ablations. Similarly, the visual point justifications were compared and it was found that PJ-X performed the best.

7.

One key takeaway that I have from this paper is the power of combining different modes of understanding to produce more informative justification systems. It is very similar to the concept of how people learn in very different ways, and therefore algorithms need to be designed to support that. Furthermore, this paper clarified distinction between justification and introspective systems for me. I wonder what work has been in introspective systems and how we can use these same datasets for introspective systems. Finally, one more question I have is whether new datasets need to be developed for each type of task beyond VQA and ACT or if we can generalize this method to all types of language/vision tasks.