1.1

h_{t-1}	x_t	h_t
0	0	0
0	1	1
1	0	1
1	1	0

$$h_t = (h_{t-1} \wedge \neg x_t) \vee (\neg h_{t-1} \wedge x_t)$$

1. 2

$$C_t = (\neg C_{t-1} \wedge x_t) \vee (C_{t-1} \wedge \neg x_t)$$

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

$$\Rightarrow \begin{cases} f_t = \neg C_{t-1} \\ x_t = i_t \\ \tilde{C}_t = \neg C_{t-1} \end{cases} \Rightarrow \begin{cases} W_f = [-1\ 0],\ b_f = 0.5 \\ W_i = [1\ 0],\ b_i = 0.5 \\ W_c = [0\ 1],\ b_c = 0 \\ W_o = [0\ 0],\ b_o = 1 \end{cases}$$

1.3 For step i and a given B_i,

$$p(y_i | x, y_{<i}) = S$$

As we know, $\sum_i x\, p(y_{i+2} | x, y_{<i+2}) = 1$, thus for a given x, $p(y_{i+1} | x, y_{<i+1}) = p(x) p(y_{i+1} | y_{<i+1})$, which means with $p(x) \in [0, 1]$, the current score S at B_i will lead to $S_{i+1} = p(x) \cdot S \leq S$ for B_{i+1}. In other word, B_{i+1} will have score that is at its best as high as S for B_i. Therefore, $best_{\leq i}$ is the overall highest-probability completed hypothesis and future steps will be no better than the best completed hypothesis.

1.4 $h_t = (W^T)^t h_{t-1}$

$\frac{\partial h_t}{\partial h_0} = (W^T)^t$, perform the eigen value decomposition

$(W^T)^t = V \Lambda^t V^T$

The equation above shows that if any one of the eigenvalues of W^T is larger than 1, Λ^t will make the gradient explode, and on the other hand, if none of them are greater than 1, the gradient will eventually vanish.

1.5 Paper Review.

The paper proposes a new model which can jointly generate visual and textual explanations using an attention mask to localise salient regions when generating textual rationales. A multimodal approach is proposed to explain, and the paper also argue that the two modalities provide complementary explanatory strength. The proposed model called Pointing and Justification model, incorporates answering model and multimodal explanation model, the first of which generates the answer to the question and the second produces visual pointing and text justification based on the answer given.

In the paper, it is shown that the model proposed above outperforms other existing (then) unimodality models.