Understanding RNNs

Prof. Forrest Davis {fdavis@colgate.edu}

October 25, 2023

Contents

1 RNN by Hand

1.1 Central Equation

$$\text{input} = \mathbf{E}(\text{tokenID})$$

$$h_0^t = \text{ReLU}(h_0^{t-1} \cdot \mathbf{W} + \text{input} \cdot \mathbf{U})$$

$$\text{logits} = h_0^t \cdot \mathbf{V}$$

1.2 Central Assumptions

The vocab has 6 words with the following token ids:

word	id
BOS	0
the	1
cat	2
is	3
outside	4
EOS	5

Our **E** (embedding matrix) is as follows:

$$\mathbf{E} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 2 & -1 & 3 & 5 \\ -1 & 2 & 1 & 4 \\ 5 & 2 & -1 & 3 \\ 4 & 0 & 1 & 1 \\ 1 & 3 & 4 & -1 \end{pmatrix}$$

Our **U** (input transformation) is as follows:

$$\mathbf{U} = \begin{pmatrix} 3 & -1 & 0 \\ 1 & 4 & -3 \\ 2 & 5 & -1 \\ 4 & 2 & 3 \end{pmatrix}$$

Our **W** (hidden transformation) is as follows:

$$\mathbf{W} = \begin{pmatrix} 2 & -3 & 1 \\ 4 & 1 & 4 \\ -5 & 3 & 2 \end{pmatrix}$$

Our **V** (output transformation) is as follows:

$$\mathbf{V} = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 4 & 2 & 1 & 5 & 2 & 3 \\ 0 & -3 & -1 & 4 & 3 & -1 \end{pmatrix}$$

Finally, our h_0^0 (initial hidden representation) is as follows:

$$h_0^0 = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$$

The entire model (ignoring the weight values) can be depicted as in Figure .

1.3 Questions

0. Give a diagram for the model (ignoring the weight values) above. In other words, plot the nodes with their connections, and label the figure with the weight matrix names. Note: Recurrence can be tricky to add, begin first pretending there is no recurrence.

See Figure 1.

Assume for the following questions, we are only concerned with the sentence "BOS the cat is outside EOS".

1. What is the new hidden state and output after reading in BOS?

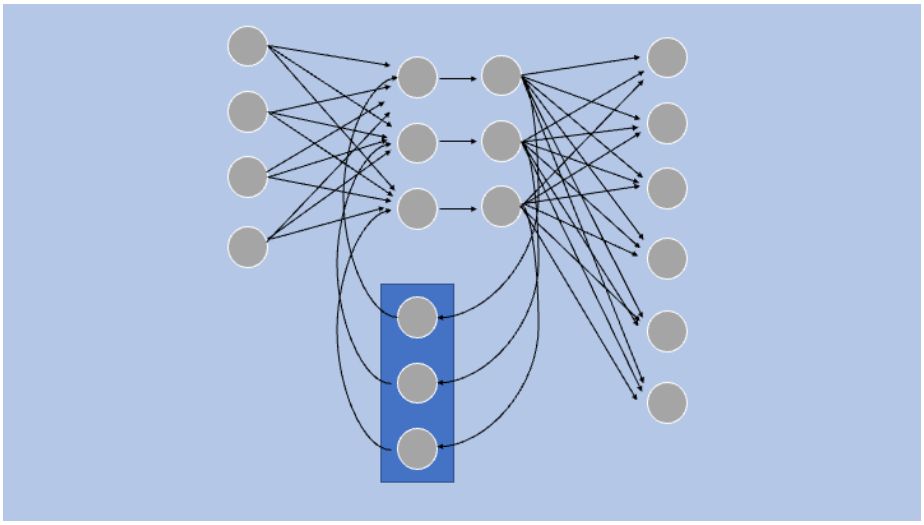

Figure 1: A schematic of the RNN detailed above

$$h_0^1 = \begin{bmatrix} 5 & 4 & 0 \end{bmatrix}$$

$$\text{logit} = \begin{bmatrix} 21 & 8 & 9 & 20 & 13 & 12 \end{bmatrix}$$

2. What is the model predicting as the next token?

The largest value is associated with index 0, so the model would predict BOS.

3. What is the new hidden state and output after reading in the? And what is the model's favored next token?

$$h_0^2 = \begin{bmatrix} 57 & 8 & 36 \end{bmatrix}$$

$$\text{logit} = \begin{bmatrix} 89 & -92 & 29 & 184 & 181 & -12 \end{bmatrix}$$

The largest value is associated with index 3, so the model favors 'is'.