Flipped Classroom Questions on Recurrent Networks

Michael Collins

Question 1: Consider the equations for a simple recurrent model mapping an input $x_1 \ldots x_n$ to a label y:

> **Inputs:** A sequence $x_1 \ldots x_n$ where each $x_j \in \mathbb{R}^d$. A label $y \in \{1 \ldots K\}$. An integer m defining size of hidden dimension. Parameters $W^{hh} \in \mathbb{R}^{m \times m}$, $W^{hx} \in \mathbb{R}^{m \times d}$, $b^h \in \mathbb{R}^m$, $h^0 \in \mathbb{R}^m$, $V \in \mathbb{R}^{K \times m}$, $\gamma \in \mathbb{R}^K$. Transfer function $g : \mathbb{R}^m \rightarrow \mathbb{R}^m$.
>
> **Computational Graph:**
>
> - For $t = 1 \ldots n$
>   - $h^{(t)} = g(W^{hx} x^{(t)} + W^{hh} h^{(t-1)} + b^h)$
> - $l = V h^{(n)} + \gamma, \;\; q = \text{LS}(l), o = -q_y$

Question 1a: Draw the computational graph for the above equations with $n = 3$.

There are three directed paths in the computational graph from W^{hx} to o. One of them is

$$W^{hx} \rightarrow h^{(3)} \rightarrow l \rightarrow q \rightarrow o$$

What are the other two directed paths?

Question 1b: For each pair of variables (a, b) with a directed arc from a to b in the computational graph, define $D(a \rightarrow b)$ to be the Jacobian associated with the edge from a to b, as calculated in the backpropagation algorithm. For example, $D(q \rightarrow o)$ is the Jacobian on the edge from q to o.

For convenience, define the matrix A to be

$$A = D(q \rightarrow o) \times D(l \rightarrow q) \times D(h^{(3)} \rightarrow l)$$

and in addition define matrices

$$B^2 = D(h^{(2)} \rightarrow h^{(3)})$$
$$B^1 = D(h^{(1)} \rightarrow h^{(2)})$$

Now write down the expression for

$$\frac{\partial o}{\partial W^{hx}}$$

using the fact that we can sum over all paths from W^{hx} to o, taking a product of Jacobians along each path.

Question 1c: For which edge or edges in the graph does the Jacobian vary as the value for x_1 varies? That is, which Jacobian or Jacobians are sensitive to the input x_1?

Question 2: Consider the following set of equations for a Bidirectional recurrent network:

> **Inputs:** A sequence $x_1 \dots x_n$ where each $x_j \in \mathbb{R}^d$. A label $y \in \{1 \dots K\}$ for position i.
>
> **Computational Graph:**
>
> - For $t = 1 \dots n$, $h^{(t)} = g(W^{hx}x^{(t)} + W^{hh}h^{(t-1)} + b^h)$
> - For $t = n \dots 1$, $\eta^{(t)} = g(W^{bhx}x^{(t)} + W^{bhh}\eta^{(t+1)} + b^{bh})$
> - $l = V \times \text{CONCAT}(h^{(i)}, \eta^{(i)}) + \gamma, \quad q = \text{LS}(l)$, $o = -q_y$

Question 2a: Complete the pseudo-code below to give a recurrent model with *two levels* of recurrent units, where the second level depends on the sequences $h^{(1)} \dots h^{(n)}$ and $\eta^{(1)} \dots \eta^{(n)}$.

> **Inputs:** A sequence $x_1 \dots x_n$ where each $x_j \in \mathbb{R}^d$. A label $y \in \{1 \dots K\}$ for position i.
>
> **Computational Graph:**
>
> - For $t = 1 \dots n$, $h^{(t)} = g(W^{hx}x^{(t)} + W^{hh}h^{(t-1)} + b^h)$
> - For $t = n \dots 1$, $\eta^{(t)} = g(W^{bhx}x^{(t)} + W^{bhh}\eta^{(t+1)} + b^{bh})$
> - For $t = 1 \dots n$, $h^{(2,t)} = \underbrace{\hspace{10em}}_{\text{Complete code here}}$
> - For $t = n \dots 1$, $\eta^{(2,t)} = \underbrace{\hspace{10em}}_{\text{Complete code here}}$
> - $l = V \times \text{CONCAT}(h^{(2,i)}, \eta^{(2,i)}) + \gamma, \quad q = \text{LS}(l)$, $o = -q_y$

Question 2b: Complete the pseudo-code below to give a recurrent model which takes as input a sequence $x_1 \dots x_n$, a position i, and in addition **a sequence of tags** $y_1 \dots y_{i-1}$, and computes the probability of a label y_i.

Inputs: A sequence $x_1 \dots x_n$ where each $x_j \in \mathbb{R}^d$. A label $y \in \{1 \dots K\}$ for position i. A sequence of tags $y_1 \dots y_{i-1}$.

Computational Graph:

- For $t = 1 \dots n$, $h^{(t)} = g(W^{hx} x^{(t)} + W^{hh} h^{(t-1)} + b^h)$
- For $t = n \dots 1$, $\eta^{(t)} = g(W^{bhx} x^{(t)} + W^{bhh} \eta^{(t+1)} + b^h)$
- For $j = 1 \dots (i-1)$, $\beta^{(j)} = \underbrace{\hspace{4cm}}_{\text{Complete code here}}$
- $l = V \times \text{CONCAT}(h^{(i)}, \eta^{(i)}, \beta^{(i-1)}) + \gamma, \quad q = \text{LS}(l), o = -q_y$

Question 3: Consider the following equations that define a *gated recurrent unit*, which takes an input $x^{(t)}$ together with the previous hidden state $h^{(t-1)}$, and returns a new hidden state $h^{(t)}$:

$$\begin{aligned}
z^{(t)} \in \mathbb{R}^m &= \sigma^m(W^z x^{(t)} + U^z h^{(t-1)} + b^z) \\
r^{(t)} \in \mathbb{R}^m &= \sigma^m(W^r x^{(t)} + U^r h^{(t-1)} + b^r) \\
h^{(t)} \in \mathbb{R}^m &= (1 - z^{(t)}) \odot h^{(t-1)} + z^{(t)} \odot g(W^h x^{(t)} + U^h (r^{(t)} \odot h^{(t-1)}) + b^h)
\end{aligned}$$

Here we have followed the conventions in the slides. $a \odot b$ is the element-wise product of vectors a and b: that is, if $c = a \odot b$ then $c_i = a_i \times b_i$ for all i.

$\sigma^m : \mathbb{R}^m \rightarrow \mathbb{R}^m$ maps a vector v to a vector $\sigma^m(v)$ with components

$$\sigma_i^m(v) = \frac{e^{v_i}}{1 + e^{v_i}}$$

Question 3a: Explain the role of the $z^{(t)}$ and $r^{(t)}$ vectors in these updates.