Based on the above data, answer the given subquestions.

Consider an RNN employing the following formulae to compute the state vector and output at time step t:

$$s_t = \sigma(Ux_t + Ws_{t-1} + b)$$
$$\hat{y}_t = O(Vs_t + c)$$

Here, σ and O denote the sigmoid and softmax functions, respectively. Assume that $s_t \in \mathbb{R}^2$, $\hat{y}_t \in \mathbb{R}^5$ and $x_t \in \mathbb{R}^6$

With a total of 7 time steps (T = 7), what is the total count of parameters (including bias) within the network?

Answer (Numeric):

Answer

Accepted Answer : 33

If all the parameters (including the bias) in the network are initialized to zero, what will be the predicted $\hat{y}_2$ at time step 2 (assuming indices start at 1) for the input $[0, 0, 1, 0, 0, 0]^T$?

OPTIONS :

- ○ $[0.2, 0.2, 0.2, 0.2, 0.2]^T$
- ○ $[0, 0.5, 0.5, 0, 0]^T$
- ○ $[0.1, 0.1, 0.6, 0.1, 0.1]^T$
- ○ $[0, 0, 1, 0, 0]^T$

Your score : 0

If all the parameters (including bias) in the network are initialized to zero, what will be the total loss after 10 time steps (assume that indices start with 1) for the input $[0, 0, 1, 0, 0, 0]^T$? The ground truth ($y \in \mathbb{R}^5$) for each time step is given by the following sequence $[y_0, y_4, y_1, y_2, y_3, y_4, y_4, y_4, y_2, y_3]$. Assume the loss to be cross-entropy at each time step. (Use natural log and enter the answer correct up to two decimal places.) Each y is one-hot vector (i.e y_0 means $y = [1, 0, 0, 0, 0]^T$, y_4 means $y = [0, 0, 0, 0, 1]^T$ and so on).

Answer (Numeric):

Answer

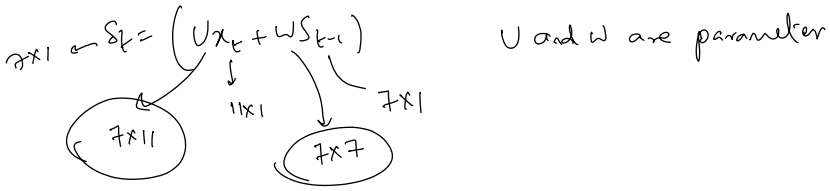
Accepted Answer : 16 to 16.2

In the context of RNNs, what structural feature of LSTMs helps reduce the impact of vanishing gradients?

OPTIONS :

- ◯ Skip connections
- ◯ Gated mechanisms like forget and input gates
- ◯ Weight sharing across layers
- ◯ Use of dropout

Your score : 0



An encoder RNN in a sequence-to-sequence model has the following specifications:

77 + 49 = 126

Embedding dimension = 11
Hidden state dimension = 7
Vocabulary size = 15
How many parameters does the encoder RNN have? (Assume no biases and assume we already have embeddings of all the input words).

Answer (Numeric):

Answer

Accepted Answer : 126

Based on the above data, answer the given subquestions.

Consider an RNN employing the following formulas to compute the state vector and output at time step t:

$$s_t = \sigma(Ux_i + Ws_{t-1} + b)$$
$$\hat{y}_t = O(Vs_t + c)$$

Here, σ and O denote the sigmoid and softmax functions, respectively. Assume that $s_t \in \mathbb{R}^2$, $\hat{y}_t \in \mathbb{R}^5$ and $x_i \in \mathbb{R}^6$

With a total of 7 time steps (T = 7), implying prediction for a word of length 6), what is the total count of parameters (including bias) within the network?

Answer (Numeric):

Answer

Accepted Answer : 33

Your score : 0

If all the parameters (including the bias) in the network are initialized to zero, what will be the predicted $\hat{y}_2$ at time step 2 (assuming indices start at 1) for the input $[0, 0, 1, 0, 0, 0]^T$?

OPTIONS :

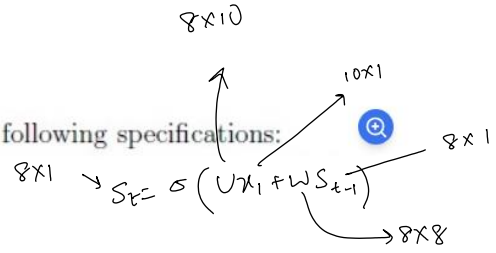
- ○ $[0, 0.5, 0.5, 0, 0]^T$
- ○ $[0.1, 0.1, 0.6, 0.1, 0.1]^T$
- ○ $[0, 0, 1, 0, 0]^T$
- ○ $[0.2, 0.2, 0.2, 0.2, 0.2]^T$

If all the parameters (including bias) in the network are initialized to zero, what will be the total loss after 10 time steps (assume that indices start with 1) for the input $[0, 0, 1, 0, 0, 0]^T$? The ground truth ($y \in \mathbb{R}^5$) for each time step is given by the following sequence $[y_0, y_4, y_1, y_2, y_3, y_4, y_4, y_4, y_2, y_3]$. Assume the loss to be cross-entropy at each time step. (Use natural log and write the answer correct up to two decimal places.) each y is one-hot encoder (i.e $y_0$means $= [1, 0, 0, 0, 0]^T$)

Answer (Numeric):

Answer

Accepted Answer : 16 to 16.2



An encoder RNN in a sequence-to-sequence model has the following specifications:
Embedding dimension = 10
Hidden state dimension = 8
Vocabulary size = 12

How many parameters does the encoder RNN have? (Assume no biases and assume we already have embeddings of all the input words).

Answer (Numeric):

Answer

Accepted Answer : 176

Since we **already have embeddings** and there are **no biases**, we only count the RNN weights:

- ✓ **Hidden size = 8**
- ✓ **Input (embedding) size = 10**

A basic RNN cell has two parameter matrices:

1. **Input → hidden weights**
   $W_{xh} \in \mathbb{R}^{8\times10} = 80$
2. **Hidden → hidden weights**
   $W_{hh} \in \mathbb{R}^{8\times8} = 64$

☑ **Total parameters**

$$80 + 64 = 144$$

✓ **Answer: The encoder RNN has 144 parameters.**

Based on the above data, answer the given subquestions.

In a time series prediction task using a GRU (Gated Recurrent Unit) network, the GRU processes input sequences where each input is represented by a 2-dimensional vector ($x_t \in \mathbb{R}^2$). The GRU uses the following formulas for the hidden state and output at time step t:

$$i_t = \sigma(W_i s_{t-1} + U_i x_t + b_i)$$
$$o_t = \sigma(W_o s_{t-1} + U_o x_t + b_o)$$
$$\tilde{s}_t = \tanh(U x_t + W(o_t \odot s_{t-1}) + b)$$
$$s_t = (1 - i_t) \odot s_{t-1} + i_t \odot \tilde{s}_t$$
$$\hat{y}_t = V h_t + c$$

where $\odot$ denotes element-wise multiplication. Assume that $h_t \in \mathbb{R}^3$ and $\hat{y}_t \in \mathbb{R}^2$.

Given that the GRU processes sequences of length 6 (T = 6), what is the total number of parameters (including biases) in the network?

Total = **62 parameters.**

Breakdown:

- For each gate i_t and o_t (two gates):
  $W \in \mathbb{R}^{3\times3} \to 9$, $U \in \mathbb{R}^{3\times2} \to 6$, bias $b \in \mathbb{R}^3 \to 3$
  Per gate = $9 + 6 + 3 = 18$. Two gates → $18 \times 2 = 36$.
- For $\tilde{s}_t$: $W \in \mathbb{R}^{3\times3} \to 9$, $U \in \mathbb{R}^{3\times2} \to 6$, bias → 3 →
  $9 + 6 + 3 = 18$.
- Output layer: $V \in \mathbb{R}^{2\times3} \to 6$, bias $c \in \mathbb{R}^2 \to 2$.

Sum: $36 + 18 + 6 + 2 = 62$.

Answer (Numeric):

Answer

Accepted Answer : 113

Your score : 0

If all parameters (including biases) are initialized to zero, what will be the predicted $\hat{y}_4$ at time step 4 (assuming indices start with 1) for the input vector $[1, 0]^T$?

OPTIONS :

- ○ $[0.5, 0.5]^T$
- ○ $[0.3, 0.7]^T$
- ○ $[0.4, 0.6]^T$
- ○ $[0.25, 0.75]^T$

Your score : 0

If the GRU is modified such that $\hat{y}_t \in \mathbb{R}^4$, and all parameters are initialized randomly, what will be the dimensionality of the weight matrix V?

OPTIONS :

- ◯ 3 × 2
- ◯ 3 × 4
- ◯ 4 × 3
- ◯ 2 × 4

Your score : 0

How many weight matrices does a LSTM unit and GRU unit learns respectively during backpropagation?

OPTIONS :

- ◯ 2,3
- ◯ 3,4
- ◯ 4,3
- ◯ 8,6

Your score : 0

Which statements about LSTM and GRU units in handling the problem of exploding/vanishing gradients are correct?

OPTIONS :

- ◯ LSTM units are more susceptible to exploding gradients compared to GRUunits due to their complex gating mechanisms.
- ◯ The gating mechanisms in LSTM and GRU units implicitly handle explodinggradients without the need for additional techniques.
- ◯ GRU units mitigate both vanishing and exploding gradient problems moreeffectively than LSTM units due to their simplified architecture.
- ◯ Both LSTM and GRU units utilize gating mechanisms to address vanishinggradient problems, but GRU units typically converge faster due to their simpler design.

Your score : 0

Based on the above data, answer the given subquestions.

Given a scenario where there are 5 possible letters represented as one-hot encoded vectors of length 5, the RNN employs the following formulas for the state vector and output at time step t:

$$s_t = \sigma(Ux_i + Ws_{t-1} + b)$$
$$\hat{y}_t = O(Vs_t + c)$$

Here, σ and O denote the sigmoid and softmax functions, respectively. Assume that $s_t \in \mathbb{R}^2$ and $y_t \in \mathbb{R}^5$

With a total of 10 time steps (T = 10, implying prediction for a word of length 9), what is the total count of parameters (including bias) within the network?

Answer (Numeric):

Answer

Accepted Answer : 31

Your score : 0

If all the parameters (including bias) in the network are initialized to zero, what will be the predicted $\hat{y}_2$ at time step 2 (assume that indices start with 1) for the input $[0, 0, 1, 0, 0]^T$?

OPTIONS :

- $[0, 0.5, 0.5, 0, 0]^T$
- $[0.1, 0.1, 0.6, 0.1, 0.1]^T$
- $[0, 0, 1, 0, 0]^T$
- $[0.2, 0.2, 0.2, 0.2, 0.2]^T$

Your score : 0

If all the parameters (including bias) in the network are initialized to zero, what will be the total loss after 10 time steps (assume that indices start with 1) for the input [0, 0, 1, 0, 0]T ? assume the loss to be cross-entropy at each time step. Write your answer correct to two decimal places.

Answer (Numeric):

Answer

Accepted Answer : 16 to 16.2

Select the correct statement regarding the vanishing and exploding gradient problem in RNN.

OPTIONS :

- [ ] The vanishing gradient problem in RNNs occurs when the gradient approacheszero during backpropagation, hindering the training of long sequences.
- [ ] The exploding gradient problem in RNNs occurs when the gradient growsuncontrollably during backpropagation, leading to numerical instability and difficulty in training.
- [ ] The vanishing gradient problem in RNNs can be mitigated by using theRectified Linear Unit (ReLU) activation function
- [ ] The vanishing gradient problem in RNNs can be mitigated by using gradientclipping, where gradients are capped to a maximum value during training.
- [ ] Both vanishing and exploding gradient problems in RNNs can occur due to thenature of the recurrent connections and the repeated multiplication of weight matrices.

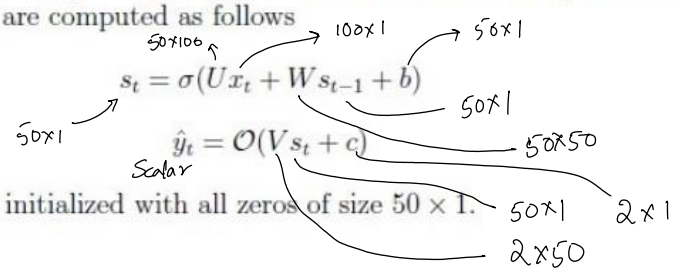
Based on the above data, answer the given subquestions.

Suppose that we need to develop an RNN model for sentiment classification tasks. The input to the model is a **sentence composed of 15 words** and the output is the sentiment (positive or negative). Assume that each word is represented as a vector of length 100×1 and the output labels are one-hot encoded. Further, the state vector s_t and the prediction $\hat{y}_t$ are computed as follows

$$s_t = \sigma(Ux_t + Ws_{t-1} + b)$$

$$\hat{y}_t = \mathcal{O}(Vs_t + c)$$

The state vector s_t is initialized with all zeros of size 50×1.



U → 5000
W — 2500
b — 50
V — 100
c — 2
7652

Your score : 0

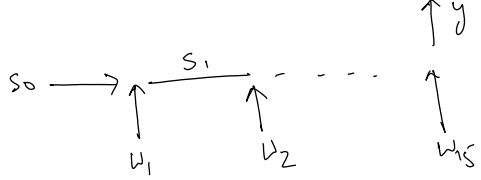
How many parameters (including bias) are there in the network?

Answer (Numeric):

Answer

Accepted Answer : 7652

Your score : 0



For the given input sentence containing 15 words, how many sequential time steps does RNN take to make a final prediction?

Answer (Numeric):

Answer

Accepted Answer : 15

Your score : 0

Back propagation through time

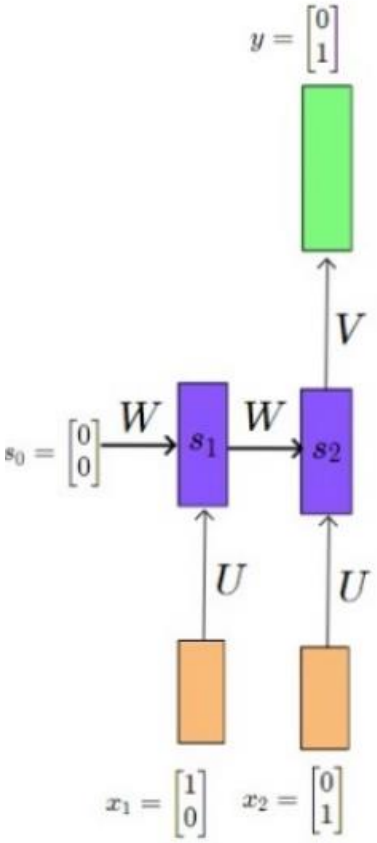
Suppose we train the model using the BPTT algorithm for 100 iterations. In each iteration, we feed the input sentence and make a prediction, compute the loss, back-propagate through time and update the parameter. How many times does the parameter matrix W get updated over 100 iterations?

Answer (Numeric):

Answer

Accepted Answer : 100

Your score : 0



$$s_1 = \tanh\left(\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \end{pmatrix}\right)$$

$$= \tanh\begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} -0.76 \\ 0.76 \end{pmatrix}$$

$$s_2 = \tanh\left(\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} -0.76 \\ 0.76 \end{pmatrix}\right)$$

$$= \begin{pmatrix} 0.99 \\ -0.99 \end{pmatrix}$$

$$\hat{y} = O\left(\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 0.99 \\ -0.99 \end{pmatrix}\right) = O\begin{pmatrix} -1.98 \\ 1.98 \end{pmatrix}$$

$$= \begin{pmatrix} 0.02 \\ 0.98 \end{pmatrix} \quad \hat{y} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Suppose the weight matrices U, V, W are initialized as follows

$$W = U = V = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$L = -0\ln(0.02) - 1\ln(0.98)$$

$$= 0.02$$

The state vector s_t is computed as follows

$$s_t = tanh(Ux_t + Ws_{t-1})$$

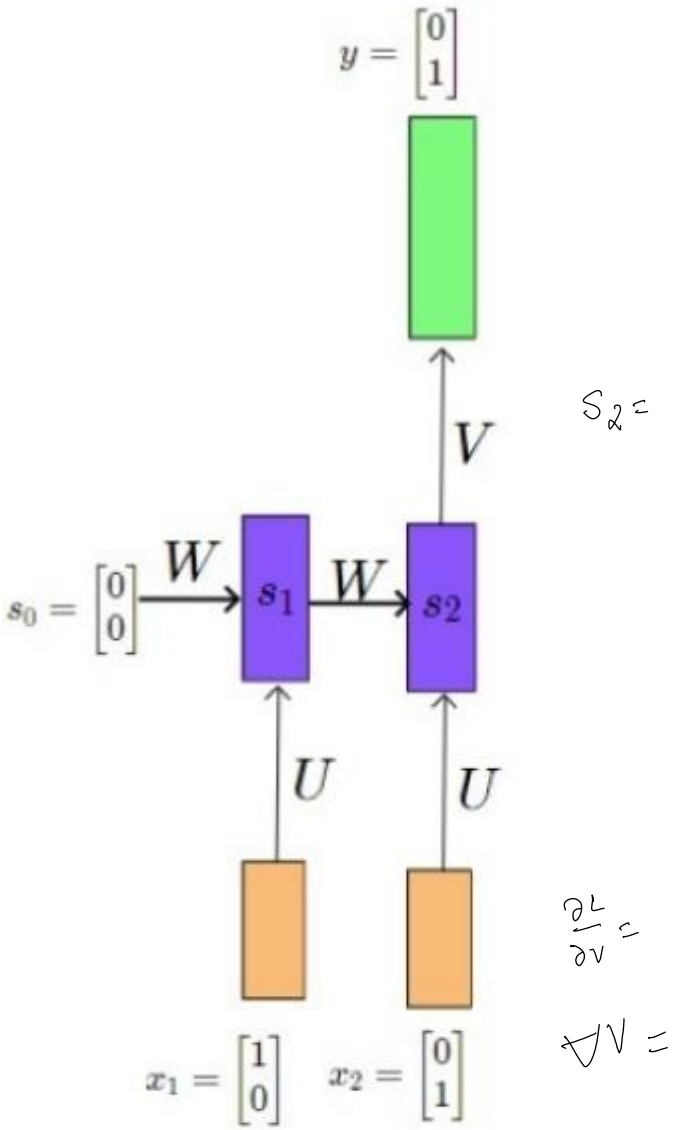
What is the loss value? Use cross entropy loss with natural logarithm.

Note: In all your calculations, consider only the first two decimal places of any number (such as inputs, intermediate results..). That is, if the number is -1.0234, take it as -1.02.

Answer (Numeric):

Answer

Accepted Answer : 0.12 to 0.26



Suppose the weight matrices U, V, W are initialized as follows

$$W = U = V \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

The state vector s_t is computed as follows

$$s_t = tanh(Ux_t + Ws_{t-1})$$

Note: In all your calculations, consider only the first two decimal places of any number (input, intermediate results..). That is, if the number is -1.0234, take it as -1.02.

What is the loss value? Use cross entropy loss with natural logarithm.

Answer (Numeric):

Answer

Accepted Answer : 2 to 2.1

Your score : 0

Compute the gradients for the weight matrix, that is ∇V and enter the sum of its (main) diagonal elements?

Answer (Numeric):

Answer

Accepted Answer : 0.80 to 0.92

Your score : 0