# Assignment 2. Recurrent Neural Networks and Graph Neural Networks

**Xinyi Chen**
University of Amsterdam – Deep Learning Course
xinyi.chen@student.uva.nl

## 1 Recurrent Neural Networks

### 1.1 Vanilla RNNs

#### 1.1.1 a

$$\frac{\partial L^{(T)}}{[\partial W_{ph}]_{ij}} = \sum_i \frac{\partial L^{(T)}}{\partial p_i^{(T)}} \frac{\partial p_i^{(T)}}{[\partial W_{ph}]_{ij}}$$

$$\frac{\partial L^{(T)}}{\partial W_{ph}} = \frac{\partial L^{(T)}}{\partial p^{(T)}} [h^{(T)}]^T$$

#### 1.1.2 b

$$\frac{\partial L^{(T)}}{\partial W_{hh}} = \frac{\partial L^{(T)}}{\partial p^{(T)}} \frac{\partial p^{(T)}}{\partial h^{(T)}} \frac{\partial h^{(T)}}{\partial W_{hh}} = \frac{\partial L^{(T)}}{\partial p^{(T)}} \frac{\partial p^{(T)}}{\partial h^{(T)}} \frac{\partial h^{(T)}}{\partial W_{hh}} = \frac{\partial L^{(T)}}{\partial p^{(T)}} W_{ph} \sum_{t=0}^{T} ( \prod_{k=t+1}^{T} (1 - (h^{(k)})^2 h^{t-1})$$

#### 1.1.3 c

The $\frac{\partial L^{(T)}}{\partial W_{hh}}$ results depend on previous time steps. As the result is the product of previous time steps, the $\frac{\partial L^{(T)}}{\partial W_{hh}}$ will become smaller and smaller(if $h^{t-1} < 1$ ) or larger and larger ($h^{t-1} > 1$) with the timesteps increasing. So it might have gradient vanishing or gradient exploring problems.

### 1.2 Long Short-Term Memory (LSTM) network

#### 1.2.1 a

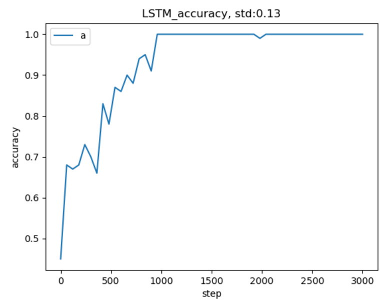
The input gate takes input information and current state to the network. The tanh function here is to centered the data around zero so that the model converge faster. The Forget gate uses the sigmoid activation function to determine whether the network should keep the data for current step. The output gate uses sigmoid function to determine wether to output the current status. Using these gates can prevent the vanishing gradient problem and make full use of its variables.

#### 1.2.2 b

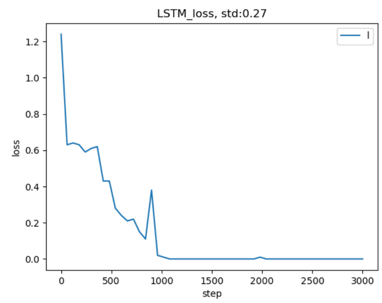
The total number of trainable parameters without the output layer is: $(n * d + n * n + n) * 4+$ where n is the nunmber of hiiden units, d is the input dimention of x.

### 1.3 LSTMs in PyTorch

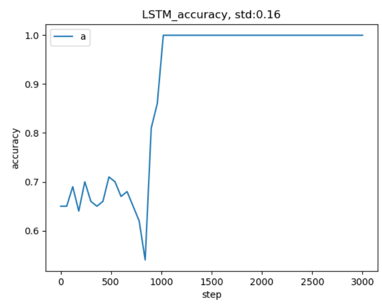
The LSTM model has been implemented in train.py. Using the required parameters and dataset. The accuracy and loss plotted in the figures are the averaged results with seed =[5, 10, 20], and their standard derivations are shown in the pictures.



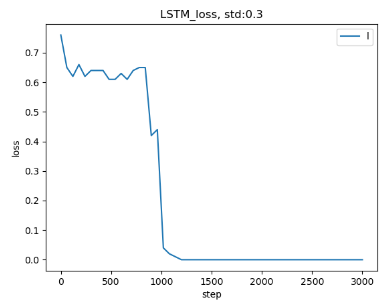

(a) LSTM accuracy curve of the averaged results with seed =[5, 10, 20], T = 10




(b) LSTM loss curve of the averaged results with seed =[5, 10, 20], T = 10



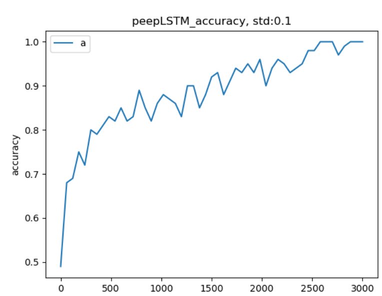

(c) LSTM accuracy curve of the averaged results with seed =[5, 10, 20], T = 20



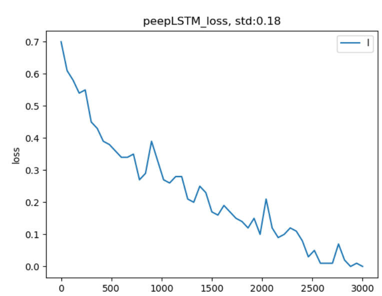

(d) LSTM loss curve of the averaged results with seed =[5, 10, 20], T = 20

### 1.4 LSTM with Peephole Connections

Overall, the LSTM model converge faster than the Peephole LSTM and achieve better performance with few steps.




(a) peepLSTM accuracy curve of the averaged results with seed =[5, 10, 20], T = 10




(b) peepLSTM loss curve of the averaged results with seed =[5, 10, 20], T = 10

# 2 Recurrent Nets as Generative Model

## 2.1

### 2.1.1 a

The hyperparameters for training of Fig.**??** and Fig.**??** are as follows: batch/$_{size} = 64 : usingbatchtospeedupthetrainingprocess.Settingittorelativelysmallvaluetopreventexceedingthememorysize.learni$ $2e{-}3 : Usingalearningratethatisnottoolargeorsmall.learning_{r}ate_{d}ecay, milestonesareusedforlearningscheduler.d$ $1 : keepsallthenodeinthenetwork, whichdoesn'tperformdropout.$

### 2.1.2 b

The generated sentences after 1/3, 2/3 and 3/3 of the training interations are shown in Fig.4. Overall, as there are more training steps, the model generated sentences with more meaningful words. Sentences become more similar to the ones in the corpus and even generated some sentences that make sense in human languages. With more training steps, the words in the sentences become more diverse as well. However, as observed in longer sentences, the repeated word problem still exists even after completing training. But they are not easily observed in short(less than 30 characters) sentences because such repeated words are not very close to the ones that first occur in the sequence. So the shorter sentences seem to have higher qualities than the longer ones.

```
Step = 2812
Sentence 0, len = 30, 8 The will soon and said to th
Sentence 0, len = 90, 8 The will soon and said to the world was the same to the wolf was the same to the wolf was
Sentence 1, len = 30, : ‘I will be a strange the sam
Sentence 1, len = 90, : ‘I will be a strange the same to the servants and said: ‘I will be a strange the same to
Sentence 2, len = 30, he said to the wolf was the sa
Sentence 2, len = 90, he said to the wolf was the same to the wolf was the same to the wolf was the same to the w
Sentence 3, len = 30, @NTHE THE THE THE THE THE THE
Sentence 3, len = 90, @NTHE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE THE T
Sentence 4, len = 30, , and the world was the same t
Sentence 4, len = 90, , and the world was the same to the wolf was the same to the wolf was the same to the wolf
```

(a) Generated sentences after 1/3 training interations

```
Step = 5625
Sentence 0, len = 30, he was a great beautiful fathe
Sentence 0, len = 90, he was a great beautiful father was already to be able to have a little tailor was already
Sentence 1, len = 30, Marleen was already to be may
Sentence 1, len = 90, Marleen was already to be may be able to have a little tailor was already to be able to hav
Sentence 2, len = 30, Now will not long that the bea
Sentence 2, len = 90, Now will not long that the bear to have a little tailor was already to be able to have a li
Sentence 3, len = 30, She had been and said: ‘I will
Sentence 3, len = 90, She had been and said: ‘I will not longer to be a golden before the world to be a golden be
Sentence 4, len = 30, And the bear to have a little
Sentence 4, len = 90, And the bear to have a little tailor was already to be able to have a little tailor was alr
```

(b) Generated sentences after 3/2 training interations

```
Step = 8438
Sentence 0, len = 30, the cook was to be able to the
Sentence 0, len = 90, the cook was to be able to the window, and the mother said: ‘I have nothing to see the wind
Sentence 1, len = 30, 4. Then the mother said: ‘I wi
Sentence 1, len = 90, 4. Then the mother said: ‘I will go to the countryman, and the morning the wood and said: ‘
Sentence 2, len = 30, So the mother said: ‘I will go
Sentence 2, len = 90, So the mother said: ‘I will go to the countryman, and the morning the wood and said: ‘I hav
Sentence 3, len = 30, I will not let the wild man sa
Sentence 3, len = 90, I will not let the wild man said: ‘I have nothing to see the window, and the morning the wo
Sentence 4, len = 30, When the morning the wood and
Sentence 4, len = 90, When the morning the wood and said: ‘I have nothing to see the window, and the morning the
```

(c) Generated sentences after 3/3 training interations

Figure 4: Generated sentences after different training steps

### 2.1.3 c

Using a lower temperature parameter will make the distribution of the softmax results more smooth. With random sampling, the lower the parameter is, it produces more surprising results. The higher the parameter is, it will produce more predictable results. The code is implemented in train.py. The generated sentences are shown in Fig.5. Using a low temperature(t = 0.5), the generated characters appear to be very random and the sentence and even the words don't make any sense. With a higher temperature(t = 2), the model generates more meaningful words and some examples are similar to human language. And surprisingly, it doesn't suffer from the word repetition problem as the greedy sample does, even when the sentence becomes longer. But the language it uses are less similar to the corpus compared with the greedy sampling method.

### 2.2 Bonus

The code has been implemented in train.py. Using –gen$_{i}nputtopassthebeginningtext.$

## 3 Graph Neural Networks

### 3.1 GCN Forward Layer

#### 3.1.1 a

Using the adjacency matrix of the graph is the key to exploit structural information, which contains nodes and edges information. By multiplying with the adjacency matrix(considering self-connections), the output of every node uses the sum of feature vectors of all its neighboring nodes and itself in one iteration. So in the next iteration, when node j uses the features of its neighboring node i, it actually takes in features of both node i and node i's neighboring nodes even if they are not neighbors of node j. In this way, the messages of each node can pass over the entire graph.

#### 3.1.2 b

The GCN treats the feature of every node connecting to it equally important. However, in real-world applications, the important of nodes are not equally important(for example, social networks might have more stronger connection neighbors). One way to improve it is to use attention networks.

### 3.2

#### 3.2.1 a

$$\tilde{A} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

#### 3.2.2 b

4 updates. The first update to C, D, the second update to C, D, B, F, the third update to A, C, D, B, F and the forth update to E, A, C, D, B, F.

### 3.3 Graph Attention Networks

The updated equation is $h_i^{(l+1)} = \sigma(\sum_{j \in N(i)} a_{ij} W^{(l)} h_j^{(l)})$. To add .where $a_{ij}$ is the final attention weight from node i to node j. The score function of attention weights is calculated using a one-layer MLP, which threats message from the node itself as a query, and the messages to average as both keys and values.

### 3.4 Applications of GNNs

1. Social recommendation: We can use nodes to represent users and items. Using edges to represent their connections. Such graph networks can be used to capture latent information between users and items, and such information can be used to make recommendations.

2. Chemical shfits prediction: The molecular structure can be represented by graph using nodes and edges. Using the GNN model can stimulate the chemistry shift of the molecular and make predictions based on it.

### 3.5 Comparing and Combining GNNs and RNNs

#### 3.5.1 a

Advanced AI requires greater interpretability. Standard neural networks (CNN, RNN) can generate composite images or documents, but not graphs, while GNN can generate unstructured data. For this reason, GNN usually performs better than RNN in visual reasoning tasks. Computer vision systems usually need to combine spatial information and semantic information to make reasoning. Therefore, it is common practice to generate diagrams for inference tasks. Visual Question answering (VQA) is a typical visual reasoning task, which requires the construction of image scene diagram and question syntax diagram respectively, and then the application of GNN to train the embedding of the prediction final answer. According to the spatial relation between objects, the relationship diagram based on problem condition is established. With the knowledge graph, more detailed relationship exploration and more explicable reasoning process can be carried out.

#### 3.5.2 b

## References

[1] T. Kipf and M. Welling, Semi-Supervised Classification with Graph Convolutional Networks (2017). arXiv preprint arXiv:1609.02907. ICLR

[2] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learn- ing to compose neural networks for question answering. In Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2016.

[3] Z. Wu, et. al., A Comprehensive Survey on Graph Neural Networks (2019).

```
Step = 8441
Sentence 0, len = 30, rking looks abovis( hapsiedla’
Sentence 0, len = 90, rking looks abovis( hapsiedla’s
afaurelthbaintial Hojob.

EtGliel; and
after your effourt h
Sentence 1, len = 30, KA–I Beorayed; howhwimmailled
Sentence 1, len = 90, KA–I Beorayed; howhwimmailled down like neot, andmother–saliggline5rary
voel; thite party–a
Sentence 2, len = 30, !
pecidedb.’ Jutterway virichi
Sentence 2, len = 90, !
pecidedb.’ Jutterway virichimine counse, so sing the, therplatspeme that intendze Mother,
Sentence 3, len = 30, Yerze,’ sil noisproluess by sp
Sentence 3, len = 90, Yerze,’ sil noisproluess by sparr!  UVEpY, gave some prejhy, O’RSSY–BIVE R5BBULLIN i Ato M
Sentence 4, len = 30, sburd_.”–turight:


‘Ta_n; 1ai
Sentence 4, len = 90, sburd_.”–turight:


‘Ta_n; 1ai“A Nrottil Droung
nar,’
At lasts
overwallens not’s, Kywly upi
```

(a) Generated sentences with temperature = 0.5

```
Step = 8441
Sentence 0, len = 30, that was the roast, I must bre
Sentence 0, len = 90, that was the roast, I must bread into the youngest, for no one looked not into
the parson.
Sentence 1, len = 30, 1968@(.E? Ball but may premed
Sentence 1, len = 90, 1968@(.E? Ball but may premed to seonather, ‘anted to the happy on his again.’ The bear pat
Sentence 2, len = 30, [913), man concerndid on, with
Sentence 2, len = 90, [913), man concerndid on, with men within ooce to dry upon, he could be a draught: ‘What ar
Sentence 3, len = 30, $X13I pockets of
bigged fell m
Sentence 3, len = 90, $X13I pockets of
bigged fell more in the states and more enough so much gave him a littly t
Sentence 4, len = 30, @M. ‘Take my told a few bride.
Sentence 4, len = 90, @M. ‘Take my told a few bride.’ At last he had her across agreed to light this, he doars.’
```

(b) Generated sentences with temperature = 1

```
Step = 8441
Sentence 0, len = 30, _.’ ‘All listen man took a res
Sentence 0, len = 90, _.’ ‘All listen man took a restrothed the water, and the miller said to her finger to the l
Sentence 1, len = 30, My son was a long the cat was
Sentence 1, len = 90, My son was a long the cat was so country, and then the cook, he said the princess was so th
Sentence 2, len = 30, Now the eldest silver have sha
Sentence 2, len = 90, Now the eldest silver have shall be found the tailor was asleep, and the
shoes were so that
Sentence 3, len = 30, ) that he was going to the sea
Sentence 3, len = 90, ) that he was going to the sea, and could not do you will not know where she was to be give
Sentence 4, len = 30, ZE
LIADDITE OR THUS
      CLEV
Sentence 4, len = 90, ZE
LIADDITE OR THUS
      CLEVER GRETEL
      THE FOUR CLEVER BROTHERS
     THE SILBIRD ICH
```

(c) Generated sentences with temperature = 2

Figure 5: Generated sentences with different temperature