

CS 224n Assignment #3 Written Part

1.

(a)

(i)

When $\nabla_{\theta} J_{minibatch}(\theta)$ first reaches zero, the algorithm will not stop immediately, so that it is possible to escape from local optimum. Then when sign of $\nabla_{\theta} J_{minibatch}(\theta)$ becomes negative, \mathbf{m} will be diminished and the speed of the shrinking of θ will slow down. Therefore, the update will not vary as much. In a nutshell, \mathbf{m} helps to make sure that the update will not vary too much when the direction is opposite to the accumulated gradient, and that the learning will converge.

(ii)

The model parameter \mathbf{v} will get larger updates when $\nabla_{\theta} J_{minibatch}(\theta)$ is big, while when $\nabla_{\theta} J_{minibatch}(\theta)$ is small the model parameter \mathbf{m} will get larger ones. This help with learning in that it slows down the speed of learning with the increase of iterates, which is of benefit as it is expected to be nearer the optimum.

(b)

(i)

$$\gamma = \frac{1}{1 - p_{drop}}$$

$$\because \mathbb{E}_{p_{drop}}[h_{drop}]_i = p_{drop} \cdot 0 + (1 - p_{drop}) \cdot 1 \cdot \gamma \circ h_i = \frac{1 - p_{drop}}{1 - p_{drop}} \cdot h_i = h_i$$

(ii)

During training we apply dropout so as to train with loads, thus during evaluation when loads are taken off out model may have a better performance.

2.

(a)

Stack	Buffer	New Dependency	Transition
[ROOT]	[I, parsed, this, sentence, correctly]		Initial Configuration
[ROOT, I]	[parsed, this, sentence, correctly]		SHIFT
[ROOT, I, parsed]	[this, sentence, correctly]		SHIFT
[ROOT, parsed]	[this, sentence, correctly]	parsed→I	LEFT-ARC
[ROOT, parsed, this]	[sentence, correctly]		SHIFT
[ROOT, parsed, this, sentence]	[correctly]		SHIFT
[ROOT, parsed, sentence]	[correctly]	sentence→this	LEFT-ARC
[ROOT, parsed]	[correctly]	parsed→sentence	RIGHT-ARC
[ROOT, parsed, correctly]	[]		SHIFT
[ROOT, parsed]	[]	parsed→correctly	RIGHT-ARC
[ROOT]	[]	root→parsed	RIGHT-ARC

(b)

A sentence containing n words will be parsed in $2n$ times, for that every step we either do *SHIFT* or do *LEFT/RIGHT-ARC*. Every word need **one** *SHIFT* and **one** *LEFT/RIGHT-ARC* to be parsed. Therefore, n words need to be parsed in $2n$ times.

(f)

(i)

Error type: Verb Phrase Attachment Error

Incorrect dependency: wedding→fearing

Correct dependency: heading→fearing

(ii)

Error type: Coordination Attachment Error

Incorrect dependency: makes→rescue

Correct dependency: rush→rescue

(iii)

Error type: Prepositional Phrase Attachment Error

Incorrect dependency: named→Midland

Correct dependency: guy→Midland

(iv)

Error type: Modifier Attachment Error

Incorrect dependency: element→most

Correct dependency: crucial→most

My Model Report

```
MINGW64:/d/I love study/Self-taught/NLP-cs224n/PPT/Assignments/...
Average Train Loss: 0.06955559689599972
Evaluating on dev set
1445850it [00:00, 32949385.70it/s]
- dev UAS: 88.39

Epoch 10 out of 10
100%|████████████████████| 1848/1848 [05:51<00:00, 5.26it/s]
Average Train Loss: 0.06654114008465886
Evaluating on dev set
1445850it [00:00, 32949027.66it/s]
- dev UAS: 88.59

=====
TESTING
=====
Restoring the best model weights found on the dev set
Final evaluation on test set
2919736it [00:00, 52276807.55it/s]
- test UAS: 88.78
Done!
(pytorch)
hp@DESKTOP-N9FA2IC MINGW64 /d/I love study/Self-taught/NLP-cs224n/PPT/Assignment
s/Kortez's work/a3
$ |
```

dev UAS: 88.59

test UAS: 88.78