1. a. i. Momentum makes oscillation at irrelevant directions (with highly various gradients) less influential. As a result moving towards the optimal point will be less distracted and faster (smoothing gradients using exponentially weighted average). Mathematically, it is equivalent to making average over N = $\frac{1}{1-\beta_1}$ values. (More on: ruder's website)

1. a. ii. By using adaptive learning rates technique, smaller gradients will get bigger updates and bigger gradients will get smaller updates (gradient normalization using moving average). Therefore, steps towards the optimal point will be smaller and smaller which makes sense because it prevents oscillation around the optimal point. In other words, by the given formulation when gradients get smaller, the learning rate increases and vica versa, hence the name “adaptive”. (It also may be helpful to think of bigger gradients as a sign of being far from the optimal point.) (More on: mlfromscratch)

1. b. i.

$$E_{p_{drop}}[h_{drop}]_i = h_i$$

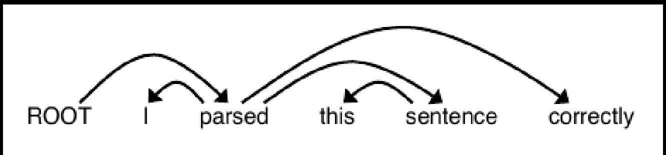
$$\gamma * h_i * (1 - p_{drop}) + \gamma * 0 * p_{drop} = h_i$$

$$\gamma * h_i * (1 - p_{drop}) = h_i$$

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

1. b. ii. In the training phase, dropout as a regularization technique helps neurons with wrong or very small weights participate in learning. In other words, it makes the NN not to rely on certain features. Hence, it will help the network prevent overfitting. However, in the testing phase, not only there is no value in making some weights zero, but also it will make predictions less accurate.

2. a.

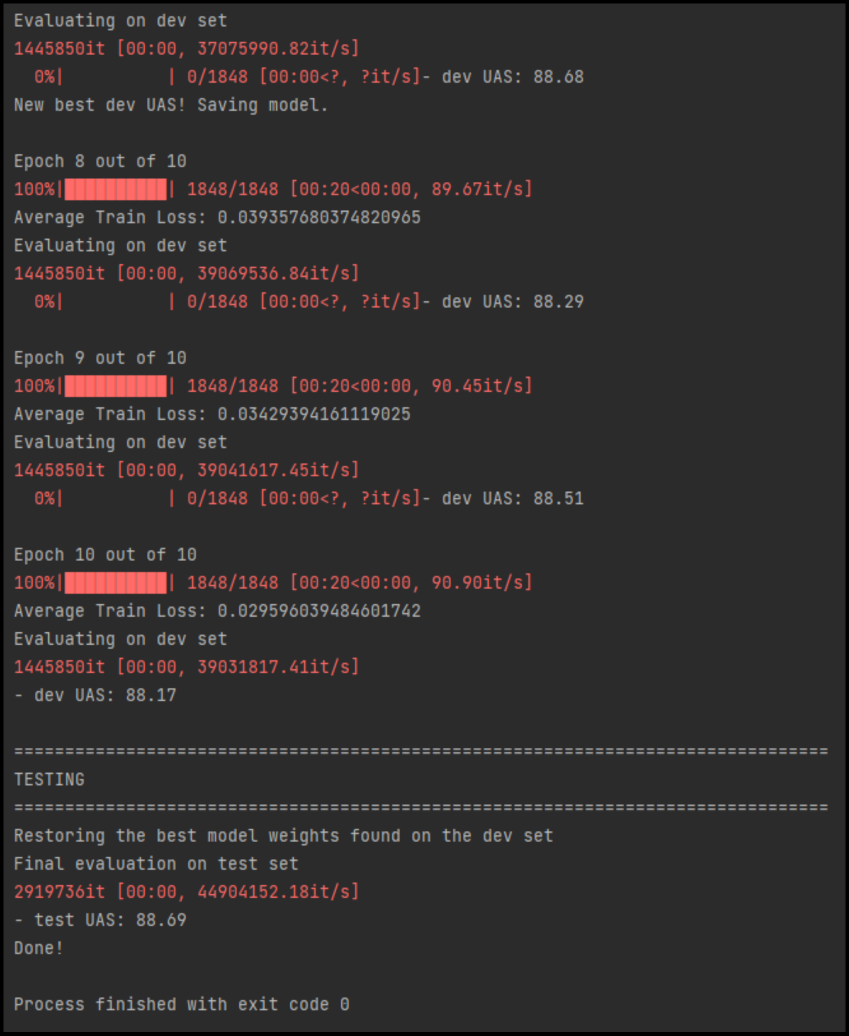


Stack	**Buffer**	**New dependency**	**Transition**
[ROOT]	[I, parsed, this, sentence, correctly]		Initial Configuration
[ROOT, I]	[parsed, this, sentence, correctly]		SHIFT
[ROOT, I, parsed]	[this, sentence, correctly]		SHIFT
[ROOT, parsed]	[this, sentence, correctly]	parsed → I	LEFT-ARC
[ROOT, parsed,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

2. b. It takes 2n steps. Because each word has to go to the stack and be removed from it one and only one time.

2. e. The training times are about 20s per epoch due to training on GPU. (See the image below)

Dev UAS	Test UAS
88.68	88.69



```
Evaluating on dev set
1445850it [00:00, 37075990.82it/s]
  0%|          | 0/1848 [00:00<?, ?it/s]- dev UAS: 88.68
New best dev UAS! Saving model.

Epoch 8 out of 10
100%|██████████| 1848/1848 [00:20<00:00, 89.67it/s]
Average Train Loss: 0.039357680374820965
Evaluating on dev set
1445850it [00:00, 39069536.84it/s]
  0%|          | 0/1848 [00:00<?, ?it/s]- dev UAS: 88.29

Epoch 9 out of 10
100%|██████████| 1848/1848 [00:20<00:00, 90.45it/s]
Average Train Loss: 0.03429394161119025
Evaluating on dev set
1445850it [00:00, 39041617.45it/s]
  0%|          | 0/1848 [00:00<?, ?it/s]- dev UAS: 88.51

Epoch 10 out of 10
100%|██████████| 1848/1848 [00:20<00:00, 90.90it/s]
Average Train Loss: 0.029596039484601742
Evaluating on dev set
1445850it [00:00, 39031817.41it/s]
- dev UAS: 88.17

================================================================================
TESTING
================================================================================
Restoring the best model weights found on the dev set
Final evaluation on test set
2919736it [00:00, 44904152.18it/s]
- test UAS: 88.69
Done!

Process finished with exit code 0
```

2. f.

PPAE: Prepositional Phrase Attachment Error
VPAE: Verb Phrase Attachment Error
MAE: Modifier Attachment Error
CAE: Coordination Attachment Error

#	Error Type	Incorrect dependency	Correct dependency
i	VPAE	wedding → fear	heading → fear
ii	CAE	makes → rescue	rush→ rescue
iii	PPAE	named → midland	guy → midland
iv	MAE	elements → most	crucial → most