CS 224N Assignmen # 3

1. (a)

(i) $m \leftarrow \beta_1 m + (1-\beta) \nabla_\theta J_{mini\text{-}batch}(\theta)$

$\theta \leftarrow \theta - \alpha m$

Using β_1 tracks the history of the gradients by the rolling average. Controlling β_1 on the previous state m_{i-1}, we can let the recent changes have a bit more importance than the current gradient, thus reduce the variance

(ii) Small v values will get larger updates. This can be helpful for learning because it can get recent parameters moving more efficiently along the axes and thus expediate the convergence.

(b) (i) $\gamma = \dfrac{1}{1-P_{drop}} = \dfrac{1}{P_{keep}}$

$E_{drop}[h_{drop}]_i = P_{drop} \cdot 0 + (1-P_{drop})\, \gamma \cdot h_i = h_i$

$\gamma = \dfrac{1}{1-P_{drop}}$

(ii) dropout is used for preventing overfitting.

2. (a)

Stack	Buffer	Arc	Transition
[ROOT]	[I, parsed, this, sentence, correctly]		Initial
[ROOT, I]	[parsed, this, sentence, correctly]		Shift
[ROOT, I, parsed]	[this, sentence, correctly]		Shift
[ROOT, parsed]	[this, sentence, correctly]	parsed → I	Left
[ROOT, parsed, this]	[sentence, correctly]		Shift
[ROOT, parsed, this, sentence]	[correctly]		Shift
[ROOT, parsed, sentence]	[correctly]	sentence → this	Left
[ROOT, parsed]	[correctly]	parsed → sentence	Right
[ROOT, parsed, correctly]	[ ]		Shift
[ROOT, parsed]	[ ]	parsed → correctly	Right
[ROOT]	[ ]	ROOT → parsed	Right

(b) $O(n)$

The worst case is $2 \cdot n$, which is linear.

(e) Final model performance

dev UAS	test UAS
88 32	88.47