CS-224n, Assignment-3.

1. Machine Learning and Neural Networks.

a) Adam optimizer

i) $m \leftarrow \beta_1 m + (1-\beta_1)\nabla_\theta J_m(\theta)$

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

'm' is the weighted average of past gradients. In each step of update we not only consider the present gradient but also the previous gradients. So if $\beta = 0.9$ past gradients weights 90% and present gradient direction weights 10%. This leads to the decrease of variance in gradients because when present gradient is in same direction of past gradients we are accelerated in the corresponding direction. But if present gradient is opposite to that of past gradients instead of suddenly varying the direction, we deaccelerate and slightly move in the present gradient direction, since we are averaging gradients over minibatch. a noisy gradient at a particular step will not lead to abrupt change in direction as past gradients are also considered.

ii) $m \leftarrow \beta_1 m + (1-\beta_1)\nabla_\theta J_m(\theta)$

$v \leftarrow \beta_2 v + (1-\beta_2)(\nabla_\theta J_m(\theta) \odot \nabla_\theta J_m(\theta))$ → This is the RMSprop method.

$\theta \leftarrow \theta - \alpha \odot m/\sqrt{v}$.

Here we are taking the history of gradient as well as squared gradients into account. The term 'm' is momentum which does the things mentioned above (i). The term 'v' is history of squared gradients. Regardless of the gradient direction this term acts as there is a square in it. It makes the learning rate a adaptive one by changing it in each step. Whenever there is a oscillation in direction of gradients for a particular dimension, 'v' makes the

Update in that direction smaller as the term 'm' is devided by $\sqrt{v}$. So, irregularity in magnitude and direction of gradient for a particular dimension will affect the 'v' which in turn reduces the amount of update in that direction, In turn dimensions whose gradient are constant will be updated more largely with time

b) Dropout

i) Y must be equal to $\frac{1}{1 - p_{drop}}$

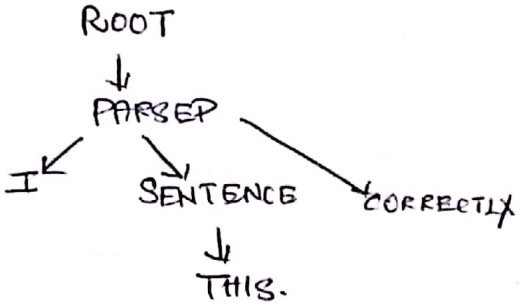
$1 - p_{drop}$ is the probability of retaining a neuron in a layer. Here $p_{drop} \propto Y$ So if p_{drop} is more it means more neurons are dropped and effective output value of that layer is less. To make the effective output value of that layer equal to the actual output distribution we are dividing it by $1 - p_{drop}$. More the p_{drop} more the value of Y, more will be the amount in which output is scaled. So we are scaling h_{drop} in a amount proportional to p_{drop} such that it's output distribution is similar to actual value.

ii). During training we apply dropout so that co-adaptive learning between neurons is decreased and each neuron becomes more powerful individually. During testing we shouldn't apply dropout, as it is a random process and may lead to masking of neurons, which are necessary to obtain the desired output for the input given. So we Consider all the neurons during evaluation time.

Assignment- 3.

Neural transition based dependency parsing

a) The given tree structure is



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 steps.

This is because each word has to be pushed into the stack and then removed from the stack based on dependency exactly once. So, 'n' words will have 'n' shift operations and 'n' arc operations (right or left) totalling '2n' operations.

c) Got an UAS of 88.29 in dev set and 88.96 in test set

d) i) Error type - Verb phrase attachment error

Incorrect dependency - wedding → feasting

Correct dependency - heading → feasting

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 - elements → most

Correct dependency - crucial → most