Stanford CS 224n Assignment 3

Shivanshu Shekhar

December 2021

1 Machine Learning & Neural Networks (8 points)

(a) (i) As we can see that typical value of β_1 around 0.9 so what this means is that the value of m varies very less with current gradients and maintains a sort of ”memory” of previously computed gradients and this makes the update more stable and reduces variance of m, also low variance is better as it leads to reduced vibrations of updates and also helps to reduce overfitting.

(ii) As we can see that for smaller value of **v** the effective learning rate will be higher, so we can deduce form this that for less frequent parameters the gradient accumulation will be low and the learning rate will be high which handles the sparse gradient problem and also on the opposite side for frequent parameters the learning rate decays and leads to convergence and prevents overfittiing of those parameters.

(b) (i)

$$\begin{aligned} \mathbb{E}_{pdrop}[h_drop]_i &= \mathbb{E}_{pdrop}[\gamma \mathbf{d} \circ \mathbf{h}]_i \\ &= \gamma \mathbb{E}_{pdrop}[d_i h_i] \end{aligned}$$

We want $RHS = h_i$

$$\begin{aligned} h_i &= \gamma((1-p)h_i) \\ \gamma &= \frac{1}{1-p_drop} \end{aligned}$$

(ii) We don’t use dropout during evaluation as we can end up getting random results for every run, this randomness helps to reduce overfitting during training but has no use when we are evaluating the model.

2 Neural Transition-Based Dependency Parsing

(a) The complete table:

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]	this → sentence	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) Initially every word will be in the buffer and stack will contain only ”ROOT” so for moving n words from buffer to stack we need n ”SHIFT” operations and es every word will we removed from stack by ”LA” or ”RA” so we need 1 move per pair of words so we have total n + 1 words(including ”ROOT”) so we will need n operations to get the stack back to 1 so in total we need 2n operations for parsing a n word sentence.

(c) CODE

(d) CODE

(e) CODE

(f) (i)

- **Error Type:** Verb Phrase Attachment Error
- **Incorrect dependency:** wedding → fearing
- **Correct dependency:** heading → fearing

(ii)

- **Error Type:** Coordination Attachment Error
- **Incorrect dependency:** rescue → and
- **Correct dependency:** rescue → rush

(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