cs224n 2019 Assignment 3

Zhou Xiaorui - jarvan000@gmail.com

March 22, 2019

Notice: The answer is done by myself, and I'm not a Stanford student.

(a)

i. with β_1 be set to 0.9, $\mathbf{m}$ will keep most of the previous step value, which will also keep most of its previous step value, this mechanism keeps $\mathbf{m}$ from varying too much.

ii. RMSProp divide every gradient by its magnitude, that means larger gradient will divide a big number, and smaller gradient will divide a small number. So we get near the same magnitude of gradient no matter how big or small that particular gradient is. It helps to handle the saddle points problem.

(b)

i. since $\mathbf{h}_{drop} = \gamma \mathbf{d} \circ \mathbf{h}$, we can get following equation (notice that p_{drop} is the drop probability):

$$\mathbb{E}_{p_{\text{drop}}}[\mathbf{h}_{drop}]_i = \mathbb{E}_{p_{\text{drop}}}[\gamma \mathbf{d} \circ \mathbf{h}]_i = \gamma(1 - p_{\text{drop}}) \mathbb{E}_{p_{\text{drop}}}[\mathbf{h}]_i = h_i$$

so the value of γ is

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

ii. During training process, we need to apply dropout to train a robust model, but when the training is finished, we need the full network to get a good prediction. So during evaluation, we do not apply dropout.

2. Neural Transition-Based Dependency Parsing

(a) see table below

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) for a sentence with n words, every word need to go through **SHIFT** process and **ARC** process, so there will be **2n** steps needed.

(c)