Natural Language Processing
Assignment №3

Pooya Kabiri

Department of Computer Science

Iran University of Science and Technology

April 2021

1 Machine Learning & Neural Networks

a.

i. Instead of using a single learning rate for all the weights, a different learning rate is maintained for each parameter, By using exponential moving averages of both first (mean) and second (uncentered variance) moments.

When using a moving average with decaying, the value of all the gradients in the past until now can be used for optimization, with past values contributing less and less over time in the optimization process, rather than GD which only relies on the current gradient.

When using Adam, the step size in each iteration is invariant to the magnitude of the gradient at the moment, so this algorithm can perform better (escape faster) when going through areas with very high or very small gradients, hence the convergence can happen a lot better and faster.

ii. As I mentioned earlier, Adam uses the uncentered variance of the gradients. When dividing m by $\sqrt{v}$, it means that when the variance in gradient values is small (algorithm is walking a flat surface) the update value is in order of m, and the walking is stable and steady, with a relatively large step size.

But, when the variance is big (algorithm is walking down a hill) the update value is much smaller, so in this way the algorithm can adapt step sizes based on the situation and it takes more careful steps, so the overall training procedure is more robust and stable.

b.

i. When using the expected value formula, we have:

$$\mathbb{E}_{p_{drop}}[\mathbf{h}_{drop}]_i = h_i$$
$$\Rightarrow \mathbb{E}_{p_{drop}}[\gamma d \odot h] = h_i$$

We know the expected value for d:

$$\mathbb{E}_{p_{drop}}[d] = 0 * p_{drop} + 1 * (1 - p_{drop}) = 1 - p_{drop} = p_{keep}$$

Using the above equation we have:

$$\Rightarrow \gamma = \frac{h_i}{h_i * \mathbb{E}_{p_{drop}}[d]} = \frac{1}{p_{keep}} = \frac{1}{1 - p_{drop}}$$

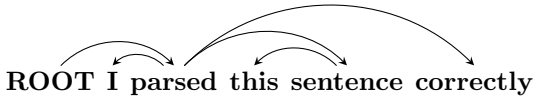
ii. When the training can result in a overfit, we apply dropout to the network so the network can't use all of its computation complexity power, so learning becomes harder for the network, and this reduction in computation power is equal for all the neurons in the network (dropout is random). In this way the network is less likely to overfit the data.

When using the model for inference, we want the model to use all of it's power so that it can generalize the data at it's best result. If we apply dropout to the model, because every time random neurons are turned off and deactivated, our model can have different inference result each time for the same test example, this can result in a inconsistent and lower performance in inference.

2 Neural Transition-Based Dependency Parsing

a.

Sentence:



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.

In procedure of dependency parsing, each word first needs to be shifted, and then one arc be decided for the word, so each word is parsed in 2 steps, hence for a sentence with n words it takes $2n$ steps until the sentence is parsed completely.

e.

- **Minimum Average Loss:** 0.043
- **Best UAS on Dev:** 88.78
- **UAS on Test:** 88.94

f.

i. **Error type:** Verb Phrase Attachment Error
Incorrect Dependency: wedding → fearing
Correct Dependency: heading → fearing

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: Midland → guy

iv. **Error type:** Modifier Attachment Error
Incorrect Dependency: elements → most
Correct Dependency: crucial → most