

Author: Benjamin Smidt  
Created: October 17th, 2022  
Last Updated: October 28th, 2022

## CS 224N A3: Dependency Parsing

Note to the reader. This is my work for assignment two of Stanford's course CS 224N: Natural Language Processing with Deep Learning. You can find the lecture Winter 2021 lectures series on YouTube here. This document is meant to be used as a reference, explanation, and resource for the assignment, not necessarily a comprehensive overview of Word Vectors. If there's a typo or a correction needs to be made, feel free to email me at benjamin.smidt@utexas.edu so I can fix it. Thank you! I hope you find this document helpful :).

### Contents

<b>1</b>	<b>Machine Learning and Neural Networks</b>	<b>2</b>
1.1	Adam Optimizer . . . . .	2
1.2	Dropout . . . . .	3
<b>2</b>	<b>Neural Transition-Based Dependency Parsing</b>	<b>4</b>
2.1	Problem A . . . . .	4
2.2	Problem B . . . . .	4
2.3	Problem C: Init and Parse Step . . . . .	5
2.4	Problem D: Minibatch . . . . .	5
2.5	Problem E: Training and Test . . . . .	5
2.6	. . . . .	5

# 1 Machine Learning and Neural Networks

## 1.1 Adam Optimizer

In our traditional Stochastic Gradient Descent, the update rule is

$$\theta \leftarrow \theta - \alpha \nabla_{\theta} J_{\text{minibatch}}(\theta)$$

The Adam optimizer modifies SGD such in an effort to improve convergence. The first addition is the use of *momentum*. Adam keeps a rolling average of the gradients instead of using only the current gradient.

$$m \leftarrow \beta_1 m + (1 - \beta_1) \nabla_{\theta} J_{\text{minibatch}}(\theta) \tag{1}$$

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

(i) *Briefly explain in 2-4 sentences how using  $m$  stops the updates from varying as much and why this low variance may be helpful to learning, overall.*

$m$  is a weighted average between all the previous updates, embedded in  $m$ , and the current update  $\nabla_{\theta} J_{\text{minibatch}}(\theta)$  (our  $\beta_1$  parameter specifies the weight to give each term,  $\beta_1 = 0.9$  is common). By keeping this weighted average, the update naturally gives higher weight to updating in directions that have been consistent while updates along dimensions that keep switching between positive and negative are given close to no weight. This improves optimization since our updates will minimize steps in dimensions that aren't getting us anywhere meaningful (flipping between positive and negative, can't decide which direction to go in) and maximize steps in dimensions that are getting us somewhere meaningful (nearly all updates have had this direction).

A second addition to Adam is *adaptive learning rates*, which keeps track of  $v$ , a rolling average of the magnitude of the gradients.

$$m \leftarrow \beta_1 m + (1 - \beta_1) \nabla_{\theta} J_{\text{minibatch}}(\theta)$$

$$v \leftarrow \beta_2 v + (1 - \beta_2) (\nabla_{\theta} J_{\text{minibatch}}(\theta) \odot \nabla_{\theta} J_{\text{minibatch}}(\theta))$$

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

$\odot$  is elementwise multiplication and  $/$  is elementwise division.  $\beta_2$  is our second hyperparameter (often set to 0.99).

(ii) *Since Adam divides the update by  $\sqrt{v}$ , which of the model parameter will get larger updates? Why might this help with learning?*

If  $v$  is quite small, then dividing by  $\sqrt{v}$  will make the term  $\alpha m/\sqrt{v}$  large. This improve learning because often times we need a large learning rate if our gradient update is naturally small.

The vice versa is also true. When the gradient is very large ( $v$  is very large), then we don't need a very large learning rate and often a small learning rate will be better. In this case, by dividing by  $\sqrt{v}$ , we actually reduce the magnitude of the update to counterbalance the already large gradient.

## 1.2 Dropout

Dropout is a form of regularization wherein we “drop” random connections within the hidden layers of our network during each update (different connections are dropped for each update). We do this mathematically with the following

$$h_{\text{drop}} = \gamma d \odot h$$

where  $h$  is a hidden layer,  $d \in \{0, 1\}^{D_h}$  ( $D_h$  is the size of  $h$ ) is a mask vector with each entry being 0 (with probability  $p_{\text{drop}}$ ) or 1 (with probability  $1 - p_{\text{drop}}$ ), and  $\gamma$  is a constant chosen such that the expected value of  $h_{\text{drop}}$  is  $h$

$$\mathbb{E}_{p_{\text{drop}}}[h_{\text{drop}}]_i = h_i \quad \forall i \in \{1, \dots, D_h\}$$

(i) *What must  $\gamma$  equal in terms of  $p_{\text{drop}}$ ? Briefly justify your answer or show your math derivation using the equations given above*

$$\mathbb{E}_{p_{\text{drop}}}[h_{\text{drop}}] = h$$

$$\mathbb{E}_{p_{\text{drop}}}[\gamma d \odot h] = h$$

$$\gamma \mathbb{E}_{p_{\text{drop}}}[d \odot h] = h$$

$$\gamma [hp_{\text{drop}} + (1 - p_{\text{drop}})0] = h$$

$$\gamma hp_{\text{drop}} = h$$

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

(ii) *Why should dropout be applied during training? Why should dropout NOT be applied during evaluation?*

Dropout should be applied during training so the network learns different pathways that lead to the same prediction. By closing different connections randomly, the network is forced to produce multiple paths in which data can flow to achieve the correct prediction, theoretically making it more robust.

We wouldn't want to apply dropout during evaluation however because our results would be non-deterministic. Due to the randomness of the dropout connections, it's possible (and may even be likely depending on the network) that evaluating the same input twice yields different predictions. Obviously this is an undesirable trait to have in a machine learning model so we don't apply dropout during evaluation.

## 2 Neural Transition-Based Dependency Parsing

### 2.1 Problem A

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

### 2.2 Problem B

There are only two options at a given step: push a word onto the stack or pop a word (and create a dependency) off of the stack. Since every word must be pushed onto the stack a single time and popped off the stack a single time, this leaves us with  $2n$  steps. This is concurrent with our table which has 11 rows:  $2(5) + 1$  where  $n = 5$  and we have an additional row for our initial state of the stack with only ROOT.

## 2.3 Problem C: Init and Parse Step

I'm not going to go over the code in any details because using implementing what we just talked about is pretty trivial (you can see the code yourself if you want to see how its done). The only important thing to remember is that we do NOT want to modify the sentence we're parsing so we must make a COPY of the sentence when initializing the buffer. Otherwise, the buffer will point to the same memory location as the sentence and we'll be modifying our original sentence, which we don't want.

## 2.4 Problem D: Minibatch

Kind of the same deal as part C, the pseudo code is given so we just have to work out the actual implementation which isn't too difficult, particularly since we're using Python. Just see the code for how this is done, the instructions explain high level what's happening already.

## 2.5 Problem E: Training and Test

I know this seems a bit lazy maybe, but I'm not going to explain this section either. I spent a tremendous amount of time learning and explaining all of this in the EECS 498: Deep Learning for Computer Vision course. The details aren't the same (obviously), but building neural networks, using PyTorch, optimizing, etc. are all covered in DEPTH in the assignments (for which I have written very detailed explanations to accompany them).

Thus, I will suffice to say that you need only look at the code for this assignment to understand what I'm doing (provided you understand DL). If you don't, I'd recommend the EECS 498: Deep Learning for Computer Vision as it is the best introduction to Deep Learning I've found so far (even if it is a little "older" (2019)).

## 2.6