Updated on: 2nd March, 2023

# 1 Machine Learning & Neural Networks

## Question (a)

### Part (i)

In each step $t$, parameters $\theta_t$ are updated along the direction of momemtum $\boldsymbol{m_{t+1}}$ in the next step, which is a weight mean between momentum $\boldsymbol{m_t}$ and gradient $\nabla_{\theta_t} J(\theta_t)$ of the current step. Here, the momentum takes a much higher weight (usually 90%) and so a large proportion of previous direction is preserved. The change in each update direction (fluctuation) is reduced and hence achieving a higher rate of convergence towards the minima.

### Part (ii)

$\boldsymbol{v_t}$ keeps track of the rolling average of the magnitude of gradients in each step. For parameters having small gradients, the corresponding momentums are divided by smaller $\boldsymbol{v_t}$ elements and hence the parameters get boosted updates. On the other hands, parameters with frequent updates are divided by a larger $\boldsymbol{v_t}$ and therefore update with smallers steps. The bias towards parameters in the learning process is reduced.

## Question (b)

### Part (i)

$$\begin{aligned}
\mathrm{E}_{P_{drop}}[\boldsymbol{h}_{drop}]_i &= h_i \\
\mathrm{E}_{P_{drop}}[\lambda \boldsymbol{d} \odot \boldsymbol{h}]_i &= h_i \\
\mathrm{E}_{P_{drop}}[d_i] \times \lambda h_i &= h_i \\
(1 - p_{drop}) \times \lambda h_i &= h_i \\
\lambda &= \frac{1}{1 - p_{drop}}
\end{aligned}$$

### Part (ii)

In the training process, dropout should be adopted to prevent overfitting the model on the training data. When overfitting occurs, some neurons dominate the model output. When dominated neurons are turned off randomly, other neurons are forced to learn other hidden patterns from the training data. Thus, all neurons are trained to contribute to the model output based on different aspects of the input. Dropout should therefore not be used during evaluation such that all neurons are involved in affecting the model output.

# 2 Neural Transition-Based Dependency Parsing

## Question (a)

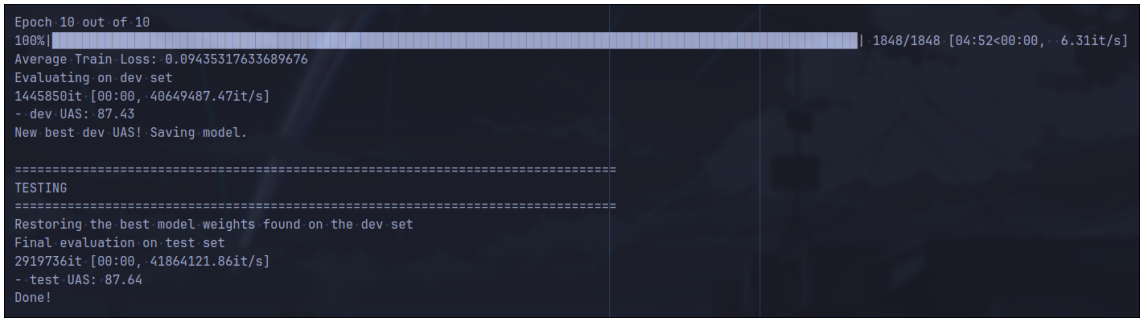
| Stack | Buffer | New dependency | Transition |
|---|---|---|---|
| [ROOT] | [I, attended, lectures, in, the, NLP, class] | | Initial Configuration |
| [ROOT, I] | [attended, lectures, in, the, NLP, class] | | SHIFT |
| [ROOT, I, attended] | [lectures, in, the, NLP, class] | | SHIFT |
| [ROOT, attended] | [lectures, in, the, NLP, class] | attended→I | LEFT-ARC |
| [ROOT, attended, lectures] | [in, the, NLP, class] | | SHIFT |
| [ROOT, attended] | [in, the, NLP, class] | attended→lectures | RIGHT-ARC |
| [ROOT, attended, in] | [the, NLP, class] | | SHIFT |
| [ROOT, attended, in, the] | [NLP, class] | | SHIFT |
| [ROOT, attended, in, the, NLP] | [class] | | SHIFT |
| [ROOT, attended, in, the, NLP, class] | [ ] | | SHIFT |
| [ROOT, attended, in, the, class] | [ ] | class→NLP | LEFT-ARC |
| [ROOT, attended, in, class] | [ ] | class→the | LEFT-ARC |
| [ROOT, attended, class] | [ ] | class→in | LEFT-ARC |
| [ROOT, attended] | [ ] | attended→class | RIGHT-ARC |
| [ROOT] | [ ] | ROOT→attended | RIGHT-ARC |

## Question (b)

One step is required to shift each word to the stack from the buffer. An extra step is then required to move each word out of the stack to create a dependency. Therefore, $2n$ steps are required.

## Question (e)

dev UAS: 87.43
test UAS: 87.64



# Question (f)

**Part (i)**

**Error type:** Verb Phrase Attachment Error

**Incorrect dependency:** acquisition → citing

**Correct dependency:** blocked → citing

**Part (ii)**

**Error type:** Modifier Attachment Error

**Incorrect dependency:** left → early

**Correct dependency:** afternoon → early

**Part (iii)**

**Error type:** Prepositional Phrase Attachment Error

**Incorrect dependency:** declined → decision

**Correct dependency:** reasons → decision

**Part (iv)**

**Error type:** Coordination Attachment Error

**Incorrect dependency:** affects → one

**Correct dependency:** plants → one