Aryan Agarwal   19CS30005

2. (a)

Rank	Correct	Precision	Recall
1	True +ve	1·00	0·25
2	True +ve	1·00	0·50
3	False +ve	0·67	0·50
4	False +ve	0·50	0·50
5	False +ve	0·40	0·50
6	True +ve	0·50	0·75
7	False +ve	0·43	0·75
8	True +ve	0·50	1·00

(b)

$W \rightarrow 10 \times 20$ Total $\rightarrow 200 + 2 \times 20 = 240$

Batch Norm $\rightarrow 2 \times 20$

$\hookrightarrow$ (2 params per dimension)

(1)(a) We have two classes $j \in \{0, 1\}$

$$P(y = 0 | x) = \frac{e^{w_0^T x}}{e^{w_0^T x} + e^{w_1^T x}}$$

$$P(y = 1 | x) = \frac{e^{w_1^T x}}{e^{w_0^T x} + e^{w_1^T x}}$$

$$= 1 - P(y = 0 | x)$$

[Since only two classes]

$$P(y=0|x) = \frac{e^{w_0^T x}}{e^{w_0^T x} + e^{w_1^T x}}$$

$$\left(\text{divide by } e^{w_0^T x}\right) = \frac{1}{1 + e^{(w_1 - w_0)^T x}}$$

Let $w_0 - w_1 \triangleq w$

$$\therefore P(y=0|x) = \frac{1}{1 + e^{-w^T x}}$$

This is a logistic regression model with parameter w.
$(w = w_0 - w_1)$

$$P(y=1|x) = 1 - \frac{1}{1 + e^{-w^T x}}$$

(b)

Architecture	Region Proposal	Classifier
RCNN	External	External
Fast RCNN	External	CNN
Faster RCNN	RPN	CNN

(c)   Bias — High
       Variance — Low

(d)   Preactivation $= a(x) = w_1 x_1 + w_2 x_2 + b$
       Output $= h(x) = h(w_1 x_1 + w_2 x_2 + b)$

(e)   (iv) $y_i(w^T x_i + b) \geq 0$ for all i

③ (a)
Key → $n \times 100 \times 36 = 3600n$
Query → $n \times 100 \times 36 = 3600n$
Value → $n \times 100 \times 25 = 2500n$

~~$\forall i j \to \text{lij} \to n \times n \times \text{dimq} = 36n^2$~~

$\forall i j \to \text{lij} \to n \times n \times \text{dimq} = 36n^2$

weighted values → $n \times n \times \text{dim}_v = 25n^2$

Total $= 61n^2 + (7200 + 2500n)$
$= 61n^2 + 9700n$
$= (61n + 9700)n$
$= 1070625 \ \text{ops}$

(b)
$W_K \to 36 \times 100$ {$\text{dim}_K = 36$}
$W_q \to 36 \times 100$ {$\text{dim}_q = 36$}
$W_v \to 25 \times 100$ {$\text{dim}_v = 25$}

Total ⇒ $(36 + 36 + 25) \times 100 = 9700 \ \text{parameters}$

(c). ~~There are 2 scalar values for~~ word
2 scalar values are present for each ~~word~~ ∧
in one layer.

∴ Total $2K = 2 \times 4 = 8$
∴ Total $2n = 150$

(d) i) Total multiplication operations $= 100 \times 200 + 200 \times 100$
$= 40000 \ \text{ops}$

ii) Total trainable parameters $= 100 \times 200 + 200 \times 100$
$+ 100 + 200$
$= 40300$ ✓

iii) Total $= 4 \times 40300 . + 9700 + 4n$ $= 50300$

$\underbrace{\qquad}_{\text{FFN}}$ $\underbrace{1 \cdot}_{\substack{\text{Self} \\ \text{attention}}}$ $\underbrace{\qquad}_{\substack{\text{Layer} \\ \text{Norm}}}$

(4) (a) $J(\theta) = \dfrac{1}{N} \sum\limits_{i=1}^{N} \left(\theta^T x^{(i)} + b y^{(i)} \right) + \dfrac{1}{2} \theta^T A \theta$

$J'(\theta) = \dfrac{\partial}{\partial \theta} \left[\dfrac{1}{N} \sum\limits_{i=1}^{N} \left(\theta^T x^{(i)} + b y^{(i)} \right) + \dfrac{1}{2} \theta^T A \theta \right]$

$= \dfrac{1}{N} \sum\limits_{i=1}^{N} x^{(i)} + A \theta$

(b). $J'(\theta) = 0$ {closed form}

$0 = - A^{-1} \left(\dfrac{1}{N} \sum\limits_{i=1}^{N} x^{(i)} \right) = \dfrac{-1}{N} \left(A^{-1} \sum\limits_{i=1}^{N} x^{(i)} \right)$

(c)   GPT
Reason: Its ability to generate text, whereas BERT cannot.

(d)   Single representation can be obtained in 2 ways:-

(i)

Word2Vec

(d) Word to vec simply averages context and center representation to get a single representation.