Task 2

a

Consider an image of 224x224 and a patch size of 16x16.
We can easily calculate the number of patches along the width and height of the image doing $\frac{224}{16} = 14$. So, the total number of patches is 14x14 = 196

b

Let's consider these token vectors: $\mathbf{p_0} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \cdot \mathbf{p_1} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} \cdot \mathbf{p_2} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \cdot \mathbf{p_3} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.
And the projection matrices $\mathbf{W_q} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\mathbf{W_k} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ $\mathbf{W_v} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$ First of all, we need to calculate query, key and values. We can do this doing the following: $(q_i, k_i, v_i) = (W_q \cdot p_i, W_k \cdot p_i, W_v \cdot p_i)$. Doing this, we get that:

$$(q_1, k_1, v_1) = (\begin{bmatrix} 1 \\ 2 \end{bmatrix}), (\begin{bmatrix} 1 \\ 0 \end{bmatrix}), (\begin{bmatrix} 0 \\ 2 \end{bmatrix})$$

$$(q_2, k_2, v_2) = (\begin{bmatrix} 2 \\ 0 \end{bmatrix}), (\begin{bmatrix} 2 \\ 0 \end{bmatrix}), (\begin{bmatrix} 0 \\ 0 \end{bmatrix})$$

$$(q_3, k_3, v_3) = (\begin{bmatrix} 0 \\ 1 \end{bmatrix}), (\begin{bmatrix} 0 \\ 0 \end{bmatrix}), (\begin{bmatrix} 0 \\ 1 \end{bmatrix})$$

$$(q_4, k_4, v_4) = (\begin{bmatrix} 1 \\ 1 \end{bmatrix}), (\begin{bmatrix} 1 \\ 0 \end{bmatrix}), (\begin{bmatrix} 0 \\ 1 \end{bmatrix})$$

Now we need to calculate the attention score. This is calculated as $q_{i,0} \cdot k_{j,0} + q_{i,1} \cdot k_{j,1}$. In this case, all k has 0 as second component so every score would be just $q_{i,0} \cdot k_{j,0}$.

$$i = 0 \text{ scores to j} \rightarrow [1 \cdot 1, 1 \cdot 2, 1 \cdot 0, 1 \cdot 1]$$

$$i = 1 \text{ scores to j} \rightarrow [2 \cdot 1, 2 \cdot 2, 2 \cdot 0, 2 \cdot 1]$$

$$i = 2 \text{ scores to j} \rightarrow [0 \cdot 1, 0 \cdot 2, 0 \cdot 0, 0 \cdot 1]$$

$$i = 3 \text{ scores to j} \rightarrow [1 \cdot 1, 1 \cdot 2, 1 \cdot 0, 1 \cdot 1]$$

So, the confusion matrix is:

$$\begin{bmatrix} 1 & 2 & 0 & 1 \\ 2 & 4 & 0 & 2 \\ 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 1 \end{bmatrix}$$

Now we need to apply softmax. The formula for softmax is: $\frac{e^{s_j}}{\sum_l e^{s_l}}$. For row 1 and 4, we have:
Exponential: $[e, e^2, 1, e]$. If we sum them we get circa 13.8256.

$$w_{1,1} = e/13.8256 = 0.1967$$

$$w_{1,2} = e^2/13.8256 = 0.5343$$

$$w_{1,3} = 1/13.8256 = 0.0723$$

$$w_{1,4} = e/13.8256 = 0.1967$$

For row 2:
Exponential: $[e^2, e^4, 1, e^2]$. If we sum them we get circa 70.3763.

$$w_{2,1} = e^2/70.3763 = 0.1050$$

$$w_{2,2} = e^4/70.3763 = 0.7757$$

$$w_{2,3} = 1/70.3763 = 0,0142$$

$$w_{2,4} = e^2/70.3763 = 0,1050$$

For row 3:

$$w_{3,j} = 0.25$$

.

Finally, we can compute the output for each v vector. Since they have 0 as first component, the first component of the output will be 0. We can compute the second as $out_i = \sum_j w_{i,j} \cdot v_j$. I will show the steps only for the first output.

$$out_1 = 0.1967 \cdot 2 + 0.5343 \cdot 0 + 0.0723 \cdot 1 + 0.1967 \cdot 1 = \begin{bmatrix} 0 \\ 0.6624 \end{bmatrix}$$

Similarly:

$$out_2 = \begin{bmatrix} 0 \\ 0.3292 \end{bmatrix}$$

$$out_3 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$out_4 = \begin{bmatrix} 0 \\ 0.6624 \end{bmatrix}$$

c

Property	CNNs	RNNs	Transformers
Parallelization	High (across spatial dims)	Low (sequential processing)	High (full parallel attention)
Long-range Dependencies	Limited (by receptive field)	Difficult (vanishing gradients)	Excellent (direct connections)
Computational Complexity	$O(k^2 \cdot d \cdot n)$ for convolution	$O(n \cdot d^2)$ for sequence	$O(n^2 \cdot d)$ for self-attention
Inductive Bias	Strong (locality, translation equivariance)	Moderate (temporal order)	Weak (requires more data)
Best Use Cases	Image processing, spatial data	Time series, sequential tasks with short context	NLP, long sequences, multimodal tasks

Table 1: Comparison of CNNs, RNNs, and Transformers. Here n is sequence/spatial length, d is feature dimension, and k is kernel size.