(a) A Long Short-Term Memory (LSTM) network has the following settings.

Initial hidden state, $\mathbf{h}_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, Initial cell state, $\mathbf{c}_0 = \begin{bmatrix} 0.1 \\ 0.2 \end{bmatrix}$,

Forget gate weight matrix, $\mathbf{W}_f = \begin{bmatrix} \mathbf{W}_{hf} & \mathbf{W}_{xf} \end{bmatrix} = \begin{bmatrix} 0.1 & 0.2 & 0.5 & 0.6 \\ 0.3 & 0.4 & 0.7 & 0.8 \end{bmatrix}$,

Input gate at timestep $t = 1$, $\mathbf{i}_1 = \begin{bmatrix} 0.3 \\ 0.4 \end{bmatrix}$,

Gate gate at timestep $t = 1$, $\mathbf{g}_1 = \begin{bmatrix} 0.5 \\ 0.6 \end{bmatrix}$,

Output gate at timestep $t = 1$, $\mathbf{o}_1 = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix}$,

Input at timestep $t = 1$, $\mathbf{x}_1 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$.

Assume no bias is used in the computation of the LSTM. The sigmoid and tanh functions are given as follows.

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

(i) Find the forget gate $\mathbf{f}_1$ at timestep $t = 1$. Comment on your obtained result.

(ii) Find the cell state $\mathbf{c}_1$ at timestep $t = 1$.

(iii) Find the hidden state $\mathbf{h}_1$ at timestep $t = 1$.

(13 Marks)

(b) Briefly describe the function(s) of (i) transformer encoder and (ii) position embedding in a Vision Transformer (ViT).

简要描述(i)变压器编码器和（ii）视觉变压器（VIT）中的位置嵌入的功能。 (6 Marks)

(c) Briefly discuss the key advantage(s) of using Swin Transformer when compared with Vision Transformer (ViT) to perform image classification.

(6 Marks)

(c)简要讨论Swin Transformer与Vision Transformer （VIT）进行图像分类时的主要优势。

Solution: LTSM

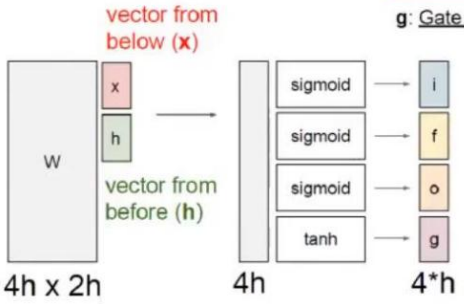
(i)

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}$$

$$c_t = f \odot c_{t-1} + i \odot g$$

$$h_t = o \odot \tanh(c_t)$$

i: Input gate, whether to write to cell
f: Forget gate, Whether to erase cell
g: Gate gate (?), How much to write to cell



$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}$$

$$c_t = f \odot c_{t-1} + i \odot g$$

$$h_t = o \odot \tanh(c_t)$$

$$\mathbf{f}_t = \sigma\left(\mathbf{W}_f \begin{pmatrix} \mathbf{h}_{t-1} \\ \mathbf{x}_t \end{pmatrix}\right) = \sigma\left(\mathbf{W}_{hf}\mathbf{h}_{t-1} + \mathbf{W}_{xf}\mathbf{x}_t\right)$$

$$\mathbf{f}_1 = \sigma\left(\mathbf{W}_f \begin{pmatrix} \mathbf{h}_0 \\ \mathbf{x}_1 \end{pmatrix}\right) = \sigma\left(\mathbf{W}_{hf}\mathbf{h}_0 + \mathbf{W}_{xf}\mathbf{x}_1\right)$$

$$= \sigma\left(\begin{bmatrix} 0.1 & 0.2 \\ 0.3 & 0.4 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0.5 & 0.6 \\ 0.7 & 0.8 \end{bmatrix}\begin{bmatrix} 2 \\ 1 \end{bmatrix}\right) = \sigma\left(\begin{bmatrix} 1.6 \\ 2.2 \end{bmatrix}\right) = \begin{bmatrix} 0.8320 \\ 0.9002 \end{bmatrix}$$

$$\mathbf{f}_1 = \begin{bmatrix} 0.832 \\ 0.900 \end{bmatrix} \text{ (round to 3 decimal places)}$$

The cell state at t=1 retains most of the memory from the previous cell state t=0.

t=1时的单元格状态保留了前一个单元格状态t-0的大部分内存。

(ii)

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t$$

$$\mathbf{c}_1 = \mathbf{f}_1 \odot \mathbf{c}_0 + \mathbf{i}_1 \odot \mathbf{g}_1$$

$$= \begin{bmatrix} 0.832 \\ 0.900 \end{bmatrix} \odot \begin{bmatrix} 0.1 \\ 0.2 \end{bmatrix} + \begin{bmatrix} 0.3 \\ 0.4 \end{bmatrix} \odot \begin{bmatrix} 0.5 \\ 0.6 \end{bmatrix}$$

$$= \begin{bmatrix} 0.233 \\ 0.420 \end{bmatrix}$$

(iii)

$$\mathbf{h}_1 = \mathbf{o}_1 \odot \tanh(\mathbf{c}_1) = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix} \odot \tanh\begin{pmatrix} 0.233 \\ 0.420 \end{pmatrix} = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix} \odot \begin{bmatrix} 0.229 \\ 0.397 \end{bmatrix} = \begin{bmatrix} 0.0916 \\ 0.2382 \end{bmatrix}$$

(i) Transformer Encoder in Vision Transformer (ViT):

(i) Vision Transformer (ViT) 中的 Transformer 编码器：

The transformer encoder in ViT serves as the core component that models relationships between image patches. After an image is divided into fixed-size patches and each patch is embedded into a vector, these embeddings are passed to the transformer encoder. The encoder utilizes **multi-head self-attention mechanisms** and **feed-forward neural networks** to capture global dependencies among all patches. This allows the model to weigh the importance of each patch relative to others, effectively learning contextual information and enabling the recognition of complex patterns across the entire image.

ViT 中的 Transformer 编码器是模拟图像块之间关系的核心组件。将图像划分为固定大小的块并将每个块嵌入到向量中后，这些嵌入将传递到转换器编码器。编码器利用**多头自注意力机制**和**前馈神经网络**来捕获所有补丁之间的全局依赖性。这使得模型能够权衡每个补丁相对于其他补丁的重要性，有效地学习上下文信息并能够识别整个图像中的复杂模式。

(ii) Position Embedding in Vision Transformer (ViT):

(ii) Vision Transformer (ViT) 中的位置嵌入：

Position embeddings in ViT are crucial for incorporating spatial information about the image patches. Since the transformer architecture lacks inherent positional awareness (it treats input tokens as a set without order), position embeddings are added to the patch embeddings to encode the positional relationships. By injecting this positional information, the model understands the arrangement of patches, which is essential for tasks like image classification where the spatial structure of features contributes significantly to the meaning of the image.

ViT 中的位置嵌入对于合并有关图像块的空间信息至关重要。由于 Transformer 架构缺乏固有的位置感知（它将输入标记视为无序的集合），因此将位置嵌入添加到补丁嵌入中以对位置关系进行编码。通过注入位置信息，模型可以理解斑块的排列，这对于图像分类等任务至关重要，其中特征的空间结构对图像的含义有显着贡献。

Key Advantages of Swin Transformer over Vision Transformer (ViT):

Swin Transformer 相对于 Vision Transformer (ViT) 的主要优势：

1. **Hierarchical Feature Representation:** Swin Transformer constructs hierarchical feature maps similar to convolutional neural networks (CNNs). This hierarchy enables the model to capture representations at multiple scales, making it more effective for recognizing objects at different sizes and improving performance on downstream tasks beyond classification, such as detection and segmentation.

   分层特征表示： Swin Transformer 构造类似于卷积神经网络 (CNN) 的分层特征图。这种层次结构使模型能够捕获多个尺度的表示，从而更有效地识别不同尺寸的对象，并提高分类之外的下游任务（例如检测和分割）的性能。

2. **Local Self-Attention with Shifted Windows:** By computing self-attention within local non-overlapping windows and shifting these windows between transformer layers, Swin Transformer reduces computational complexity from quadratic to linear with respect to image size. The shifted window mechanism allows for cross-window connections, enhancing the model's ability to capture both local and global context efficiently.

   具有移动窗口的局部自注意力：通过在局部非重叠窗口内计算自注意力并在 Transformer 层之间移动这些窗口，Swin Transformer 将计算复杂度从相对于图像大小的二次方降低为线性。移动窗口机制允许跨窗口连接，增强模型有效捕获本地和全局上下文的能力。

3. **Scalability to High-Resolution Images:** Swin Transformer is more scalable to high-resolution images compared to ViT. Its linear computational complexity makes it feasible to process larger images without a significant increase in resource demands, which is particularly advantageous for tasks requiring high-resolution inputs.

   高分辨率图像的可扩展性：与 ViT 相比，Swin Transformer 对于高分辨率图像的可扩展性更高。其线性计算复杂性使得可以在不显着增加资源需求的情况下处理更大的图像，这对于需要高分辨率输入的任务特别有利。

4. **Improved Performance:** Due to its efficient modeling of local and global interactions and hierarchical structure, Swin Transformer often achieves better accuracy on image classification benchmarks than ViT. It effectively balances the trade-off between model capacity and computational efficiency.

   改进的性能：由于其对局部和全局交互以及层次结构的有效建模，Swin Transformer 在图像分类基准上通常比 ViT 取得更好的准确性。它有效地平衡了模型容量和计算效率之间的权衡。

By addressing some of the limitations of ViT, such as high computational cost and lack of hierarchical features, Swin Transformer provides a more efficient and versatile architecture for image classification and other vision tasks.

通过解决 ViT 的一些局限性，例如高计算成本和缺乏分层特征，Swin Transformer 为图像分类和其他视觉任务提供了更高效、更通用的架构。