Exercise: CNN

(a) In a simple Convolutional Neural Network (CNN), an input image **A** passes through a convolutional layer, followed by an activation layer and a max pooling layer. The output from the max pooling layer is then used for further processing. The grayscale image **A** is given by:

$$\mathbf{A} = \begin{bmatrix} 4 & 0 & 1 \\ 4 & 0 & 2 \\ 0 & 2 & 2 \end{bmatrix}.$$

The convolutional layer has the following settings: the current filter is given by **F** below, the amount of zero padding at each side of the image is 1, and the stride both horizontally and vertically is 2.

$$\mathbf{F} = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$

The activation function used in the activation layer is given by a sigmoid function:

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

The max pooling layer uses 2×2 max pooling with a stride of 2.

Exercise: CNN

(i) Find the output after the convolution layer.

(ii) Briefly discuss the effect of filter **F** when applied to the input image.

(iii) Find the output after the activation layer.

(iv) Find the output after the max pooling layer.

(v) A student would like to make the following changes to the input image and the convolutional layer:

- Change the grayscale image **A** to an RGB image **B** with a spatial dimension of 100×100.
- Change the channel number of the output feature maps to 6 for the new convolutional layer. Assume the spatial dimension of the filter remains the same.

Find the number of trainable parameters of the new convolutional layer after the changes. Assume no bias is used in the calculation.

(18 Marks)

Solution

(a)

$$\mathbf{A}_{\text{padded}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 1 & 0 \\ 0 & 4 & 0 & 2 & 0 \\ 0 & 0 & 2 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \mathbf{F} = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix}$$

(i)

$$\text{Output after convolution layer} = \begin{bmatrix} 0 & 0 \\ -4 & 4 \end{bmatrix}$$

Solution

(ii)

The effect of filter F is that it computes the horizontal gradient information which reflects the edge information.

(iii)

Output after activation layer = $\begin{bmatrix} 0.5 & 0.5 \\ 0.018 & 0.982 \end{bmatrix}$

(iv)

Output after max pooling layer = $[0.982]$

(v)

Number of trainable parameters of the new convolutional layer
= 3 x 3 x 3 x 6
= 162

(b) A student would like to develop an Artificial Intelligence (AI) model to perform short video clip genre classification with high accuracy performance. The type of genre may include comedy, action, romance, etc. Assume the visual feature from each video frame has been embedded into a feature vector. State clearly which of the following models is most likely to satisfy the student's need: (i) CNN, (ii) Vanilla Recurrent Neural Network (RNN), or (iii) Transformer. Briefly justify your answer.

(7 Marks)

Solution 解决方案

Answer: (ii) Vision Transformer (ViT).

Vision Transformer (ViT) is based on Transformer architecture that uses attention mechanism and can achieve very good accuracy.

VGG is a CNN that does not leverage on global attention. It uses convolutional layers to progressively extract higher-level abstraction features.

LSTM is a model that uses memory to analyse sequential data. Hence it is not suitable for image classification application.

答案：（ii）视觉变压器（ViT）。

视觉变压器（Vision Transformer，ViT）

是基于Transformer架构，利用注意机制，可以达到很好的精度。

VGG是一个不利用全局关注的CNN。

它使用卷积层逐步提取更高级别的抽象特征。

LSTM是一种使用内存分析顺序数据的模型。

因此不适合用于图像分类。

The **Vision Transformer (ViT)** is the most suitable model for the user's needs.

Vision Transformer (ViT)是最适合用户需求的模型。

Justification:理由:

- **Use of Attention Mechanism:** ViT leverages the Transformer architecture, which is built upon self-attention mechanisms. This allows the model to capture long-range dependencies and relationships within image data effectively.

  使用注意力机制： ViT 利用基于自注意力机制的 Transformer 架构。这使得模型能够有效地捕获图像数据中的远程依赖性和关系。

- **Image Classification Capability:** ViT is specifically designed for image classification tasks and has demonstrated strong performance on benchmarks, making it well-suited for developing image classification applications.

  图像分类能力： ViT专为图像分类任务而设计，在基准测试中表现出强大的性能，非常适合开发图像分类应用。

- **Comparison with Other Models:与其他型号比较:**

  - **VGG:** While VGG is a powerful convolutional neural network for image classification, it does not incorporate attention mechanisms.

    VGG： 虽然 VGG 是一个强大的用于图像分类的卷积神经网络，但它没有包含注意力机制。

  - **LSTM:** LSTMs are tailored for sequential data and are not typically used for image classification. Although they can utilize attention mechanisms, they are less effective for processing image data compared to ViT.

    LSTM： LSTM 是为序列数据量身定制的，通常不用于图像分类。尽管它们可以利用注意力机制，但与 ViT 相比，它们处理图像数据的效率较低。

LSTMs can utilize attention mechanisms to better handle complex sequential data by allowing the model to focus selectively on different parts of the input sequence. This combination has been highly effective in various natural language processing tasks.

LSTM 可以利用注意力机制，让模型选择性地关注输入序列的不同部分，从而更好地处理复杂的序列数据。这种组合在各种自然语言处理任务中非常有效。