The Exploration of CNN and ViT in Chinese Mnist Dataset

**Abstract**

1. **Introduction**

ML task:我们想要探究CNN和ViT在汉字识别任务上，各自存在什么优势。

Deep learning has revolutionized image classification, with CNNs historically being the dominant architecture due to their ability to extract hierarchical features efficiently. However, the emergence of Transformers, particularly ViT, has introduced a new paradigm that leverages self-attention mechanisms to model long-range dependencies. This paper examines CNN and ViT performance on a Chinese handwritten digit dataset, which presents a greater structural complexity compared to standard numerical digit datasets. We analyze the effects of image resolution on model performance and resource consumption.

1. **Problem Setting**

数据集为中文手写数字数据，选择其的缘由是相比简单的数字识别，中文数字的特征结构似乎更为复杂。并且中文数字包含简单到较复杂的汉字，像是一，二，三与零，四，五相比就简单很多。

The dataset used in this study consists of Chinese handwritten digits, which possess more complex stroke structures compared to standard Arabic numeral datasets such as MNIST. The choice of this dataset allows us to better evaluate how CNNs and ViTs capture intricate details in images. By experimenting with different image resolutions (64×64 and 128×128), we aim to investigate how each model adapts to changes in input size.

The Chinese MNIST dataset consists of handwritten digits from 0 to 9, each with distinct stroke-based structures. Unlike standard digit recognition, which mainly involves simple shapes, Chinese digits often require models to capture fine-grained patterns and long-range dependencies. The objective of this study is to evaluate how CNN and ViT architectures perform in extracting these patterns while adjusting for different image resolutions (64×64, 128×128, and 224×224).

1. **Methodology**

**CNNs** process images through convolutional layers that detect spatial features, pooling layers that reduce dimensionality, and fully connected layers that perform classification. A standard CNN architecture consists of:

1. **Convolutional layers**: Extract local patterns such as edges and textures.
2. **Pooling layers**: Downsample feature maps to reduce computation.
3. **Fully connected layers**: Convert spatial features into classification outputs.

Unlike CNNs, **ViTs** divide input images into fixed-size patches and treat them as sequences, similar to words in NLP. The main components of ViT include:

1. **Patch Embedding**: Converts image patches into tokenized representations.
2. **Multi-Head Self-Attention (MSA)**: Captures long-range dependencies across the entire image.
3. **Feedforward Layers**: Process extracted global information for classification.

**4.1 CNN**

卷积神经网络（CNN）是一种专门用于图像处理的神经网络结构，它通过层级化的方式学习局部特征，并逐步整合成更高层次的抽象信息。CNN 结构通常包括以下模块：

卷积层（Convolutional Layers）：用于提取局部模式，如边缘、纹理等关键信息。

池化层（Pooling Layers）：对特征图进行下采样，以减少计算量并提高模型的平移不变性。

全连接层（Fully Connected Layers）：将提取到的空间特征转换为分类结果。

**CNN 适用于本任务的原因**

在手写汉字分类任务中，CNN 由于其局部感受野和权重共享机制，能够高效地提取字符的局部结构特征，如笔画形状、边缘轮廓和基本组成部分。这些特征对于区分不同的手写数字和汉字至关重要。此外，CNN 通过池化层减少了计算复杂度，并增强了对字符形变的鲁棒性，使其在此任务中具有良好的适应性。

CNN 凭借着其独特的卷积操作，从图像中提取特征图

本研究中采用的自定义 CNN 由两个卷积块组成，每个卷积块包含两个卷积层，并在末尾接一个最大池化层（Max-Pooling）。之后，经过全局平均池化（Global Average Pooling, GAP）层，使特征映射缩减至固定尺寸，并最终通过两层全连接网络进行分类。模型的最终输出层包含 15 个类别，适用于本研究任务。

**4.2 ViT**

视觉变换器（ViT）采用 Transformer 结构进行图像处理，不同于 CNN 依赖局部特征提取的方法，ViT 直接将输入图像划分为固定大小的图像块（patches），并将其作为序列输入 Transformer 结构，类似于自然语言处理（NLP）中的词嵌入表示。ViT 的主要组件包括：

Patch Embedding：将图像划分为固定大小的 patch，并通过线性投影映射到高维特征空间。

多头自注意力（Multi-Head Self-Attention, MSA）：通过多个注意力头学习图像块之间的长程依赖关系，实现全局特征提取。

前馈神经网络（Feedforward Layers）：用于进一步处理提取到的全局信息，提高模型的表达能力。

**ViT 适用于本任务的原因**

ViT 的自注意力机制可以有效捕捉字符的全局依赖关系，例如汉字笔画之间的相对位置和整体结构，这对于手写汉字分类具有重要意义。此外，相较于 CNN 主要依赖局部感受野，ViT 可以直接从整体图像中学习全局信息，减少因局部特征相似而导致的混淆，提高分类的准确性。在手写字符任务中，汉字之间的区分不仅依赖于局部笔画形状，也与整体结构相关，因此 ViT 在该任务中具有较大优势。

本研究中使用的 ViT 模型基于 transformers 库中的 ViTForImageClassification 类，并采用以下参数进行配置：

Patch size = 8：将输入图像划分为 8×8 大小的 patch。

Hidden size = 256：Transformer 编码器中的隐藏层维度设为 256。

Num hidden layers = 10：模型包含 10 层 Transformer 编码器。

Num attention heads = 8：自注意力机制使用 8 个注意力头，以捕获不同 patch 之间的全局依赖。

Intermediate size = 512：每个 Transformer 块中的前馈网络（FFN）具有 512 维的中间层。

**方法选择的合理性**

本研究选择 CNN 和 ViT 进行对比实验，主要是因为两者在特征提取方式上存在显著差异。CNN 依赖局部感受野，适合提取笔画级别的细节特征，而 ViT 依赖自注意力机制，可以捕获汉字的整体结构信息。通过对比两种方法的分类表现，可以更全面地理解不同网络在手写汉字分类任务上的优势与局限性。

**4. Experiment**

每次以**64x64, 128x128, 224x224**的图像大小进行训练，采用5折交叉验证，记录训练时间，模型性能(准确率，损失值)，内存占用，得到的五个结果取平均为最终结果。

分析CNN和ViT之间的模型性能差异，以及图像大小对它们各项指标的影响程度

实验细节：

**Optimizer: Adam** with an initial learning rate of 0.0001

**Learning Rate: 0.0001**

**Batch Size: 64**

**Loss Function:** Cross-entropy loss

**Epochs: 10**

**5. Result**

**5.1 Model Performance**

**折线图：展示 两个模型分别绘制，在最好情况下的 Epoch-Acc/ Loss 图像**

**表格图：展示，在各个图像大小下的表现**

**5.2 Computational Efficiency**

**三线图**：展示**~~参数量(待定)，~~训练时间，内存占用，以及增长率**

**5.3 Key Observations(预期)**

1. **CNNs perform better on lower resolutions: CNNs excel in small-scale images due to efficient local feature extraction but struggle as image resolution increases.**
2. **ViTs benefit from larger images: ViTs leverage attention mechanisms, allowing them to extract more meaningful features when image size increases.**
3. **ViTs are more memory-efficient at high resolutions: Unlike CNNs, which require large fully connected layers, ViTs maintain a stable parameter size regardless of image resolution.**

分析结果：

解释：

**Conclusion**

总的来说，本研究简单探讨了CNN和ViT两种底层技术相差较大的机器/ 深度学习技术。我们观察到…

**~~不足：~~**

~~虽然本研究能清晰得出这些结论，但是中文手写数字识别任务并不能完全得出CNN和ViT的优劣。在实验中，我们也发现由于数据集的分割问题，测试集在训练过程中几乎全部优于训练集.~~