**Thesis Proposal**

**(論文選題報告)**

**Research and Develop of Offline Handwritten Signature Verification Algorithm based on Transformer**

**(基於Transformer的離線手寫簽名驗證算法研究與部署)**

**Abstract（摘要）**

**1. Introduction（引言）**

**1.1. Research Background（研究背景）**

生物特征技术（Biometrics technology）是一种根据特征、面部、虹膜等个体生理特征或声纹、手写签名等行为特征进行识别或验证的技术。该技术被广泛应用于企业的安全认证、金融交易、门禁系统等安全领域。

生物特征技术主要应用于识别和验证场景。在第一种场景下，用户仅需提供个体或生理特征样本，生物特征识别系统根据样本在注册用户中识别提供特征的用户。而第二种场景则是在第一种场景的基础上，向系统声明自己的身份，随后系统将根据以上信息检查注册用户中是否存在提供特征的用户。

手写签名是日常生活中较为重要的个体行为特征，因为其作为在法律、金融、行政等领域验证个人身份的主要特征。同时也因为它在收集过程中是无法侵入的，诸多验证个人身份的技术中也将手写签名视为主要特征之一。

基于上述，手写签名验证属于生物特征技术验证场景的一种具体应用，它将根据收集途径不同分为离线和在线。离线手写签名验证（Offline Handwritten Verification）的个人签名收集过程是通过用户在纸张上进行书写过程后得到，而在线手写签名验证（Online Handwritten Verification）采用数字化台方式收集用户的手写签名特征，因此收集的手写签名图像可能会受到设备的影响，如笔的位置、倾斜度、压力等。由此可以看出离线手写签名的优势，可以随时随地不受设备限制地收集用户个人签名图像。

**1.2. Motivation（动机）**

离线手写签名验证在收集数据的过程，虽然不会受到数字化台等设备因素影响，但是用户使用纸张书写个人签名的时候不能保证多次书写得到的签名完全一模一样，可能也会受到实体笔、纸张厚度等影响。因此收集数据后进行的预处理过程一定程度上会影响系统验证效果，一个良好的图像预处理技术可能改善系统验证性能。

在过去的离线手写签名验证研究中，人为设计图像预处理后的特征输入支持向量机（SVM）分类器中以鉴别签名是真实的还是伪造的。由于深度学习和卷积神经网络（CNN）的发展，学者们放弃了以往人工设计特征，采取CNN作为特征提取器提取图像多通道特征的方式以进行手写签名验证，并且取得了不错的效果。但是基于CNN的特征提取方式，共享参数卷积核的方法导致在全局特征上学习效果较差。

离线手写签名验证还需要验证提供样本的用户，因此验证任务中分为两种：作者依赖（Writer Dependent）和作者独立（Writer Independent）。WD任务是针对每一个作者的手写签名训练一个分类器，由此将增加模型训练成本和推理时间。WI任务是不需要针对某个作者，仅仅分析参考签名和提供签名之间的特征，从而验证其是否为伪造的。

**1.3. Research Objectives（研究目标）**

离线手写签名验证的模型训练过程中输入的是一对特征图像，即参考签名和查询签名。验证、测试过程中可能仅仅需要输入单个签名图像（查询签名）以鉴别是否为伪造的。因此特征提取器在模型架构上针对上述参考签名和查询签名图像对的推理过程是否需要共享参数以更好地学习个人手写签名特征，后续将验证模型共享参数后验证准确率是否有提高。

传统机器学习方法的离线手写签名验证很大程度依赖于人工设计的特征进行训练，最终鉴别的分类准确性与这些特征直接相关，一旦效果欠佳排查问题将耗费巨量的时间。由此人工设计的特征提取方式逐渐更替为深度学习的卷积神经网络方式，输入的单通道或者三通道图像会经过许多卷积层，从而得到一个多通道的特征图，以此作为后续鉴别签名是否赝品的特征。这种方式由于卷积算法的卷积核共享参数，将导致图像一定程度上会丢失全局特征。而个人手写签名每一次都不一定相同，因此局部共享参数的卷积神经网络在学习个人手写签名图像特征时会丢失整体签名的特征，如书写粗度、笔迹、字体风格等等问题。由此本文将提出基于Transformer中关注全局特征的多头注意力机制特征提取器，从而更有效地学习手写签名图像特征。

针对WI和WD任务，以往基于传统机器学习方法的离线手写签名验证模型是采取SVM作为分类器以完成task。更替为深度学习的深度神经网络进行特征提取后需要验证SVM能否作为分类器以完成对个人手写签名多通道特征图的鉴别是否赝品任务。同时将基于图像分类的卷积神经网络最后分类部分以改良分类器架构，验证改良是否更好契合多通道特征图来完成手写签名离线验证。

**2. Literature Review（文献综述）**

**3. Methods（论文主题算法等）**

在本章节中，首先总结使用的方法并详细介绍所提出的模型架构。深入解析所提出的包含编码器、卷积模块、解码器架构的特征提取器具体细节。随后将介绍模型参数训练过程的损失函数、优化器、作者独立分类器等细节。最后描述是否共享特征提取器模型参数的实验、分类器种类的实验，以达到效果良好的手写签名离线验证任务。

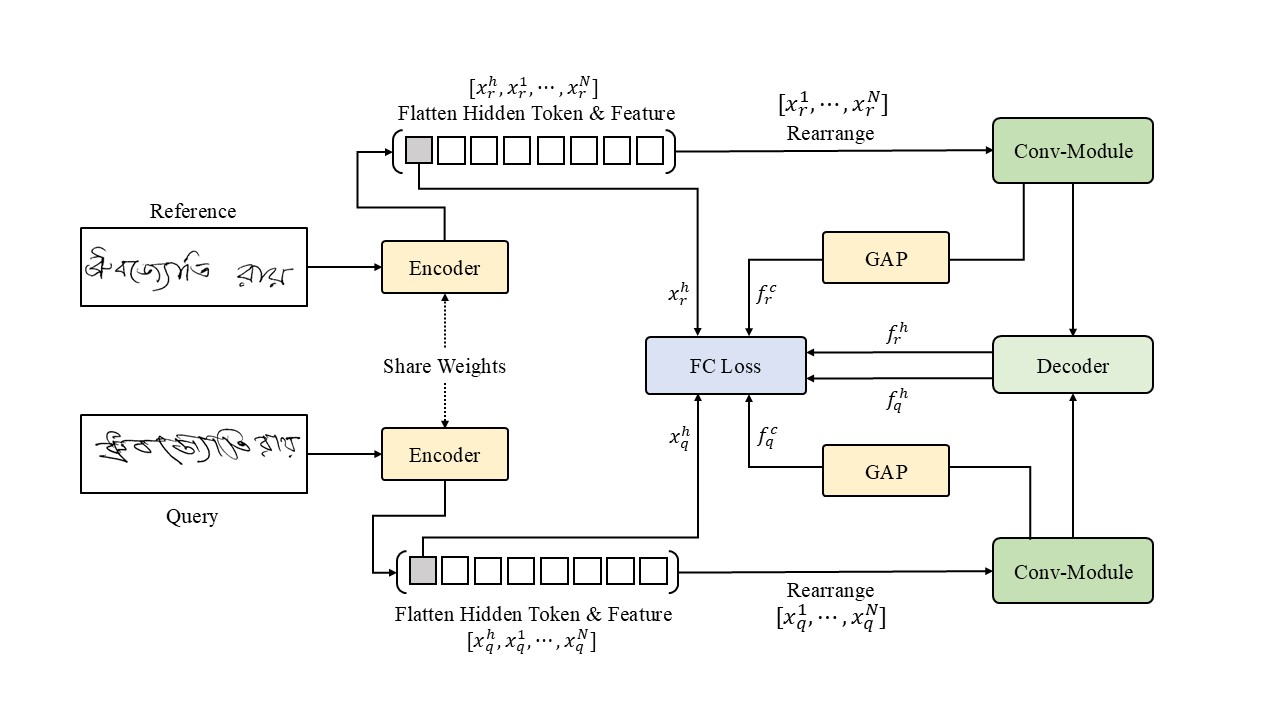
**3.1. Overview**

图1: Offline Signature Verification TransFormer (OSVTF)

提出的OSVTF Extractor是基于TransOSV改良的完成离线手写签名验证任务的特征提取器如图1所示。与TransOSV输入部分相同，输入一对尺寸相同的手写签名图像 Reference和Query。起始会经过权重共享的Encoder以学习图像特征，该部分推理将提取得一对平坦特征token 和一对平坦特征图。随后平坦特征图会经过reshape到二维图像向量再进行卷积模块推理。其中卷积模块的输出将分为两个方向：第一个方向将通过Global Average Pooling层得到平坦卷积特征；另一个方向将成对进入Decoder得到平坦注意力解码特征。训练过程将收集推理过程中平坦特征token 、平坦卷积特征和平坦注意力解码特征以进行Focal Contrast Loss (FC)损失计算，从而进行模型权重的训练。接下来将深入解析Encoder、Decoder和Conv-Module结构。

**3.2. Encoder**

OSVTF中的Encoder是沿用了ViT(Vision Transformer)的模型架构如图2所示。

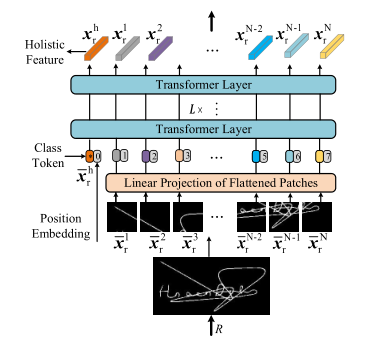


图2: Vision Transformer

以Reference（形状为）举例，其中表示图像通道数、高度和宽度。ViT首先将输入图片分成块和平坦化，其中patch number如式（1）所示。

其中表示floor function。和分别表示patch size和卷积层步数，因为ViT原文架构是采取卷积层以完成分块操作。其次，拼接一个可学习权重Class Token和patch图像累加Position Embedding，形状是。由于Transformer Encoder要求输入的是Token向量，图像patch后会丢失相对位置信息，因此累加了Position Embedding重新赋值了相对位置特征。经过上述处理后进入层Transformer Encoder Layer推理得到Encoder的平坦特征token与特征图，整个Encoder推理如式（2）。

其中表示Embeddings。表示Position Embeddings。每个Encoder结构相同，由个堆叠形成，如图3所示。

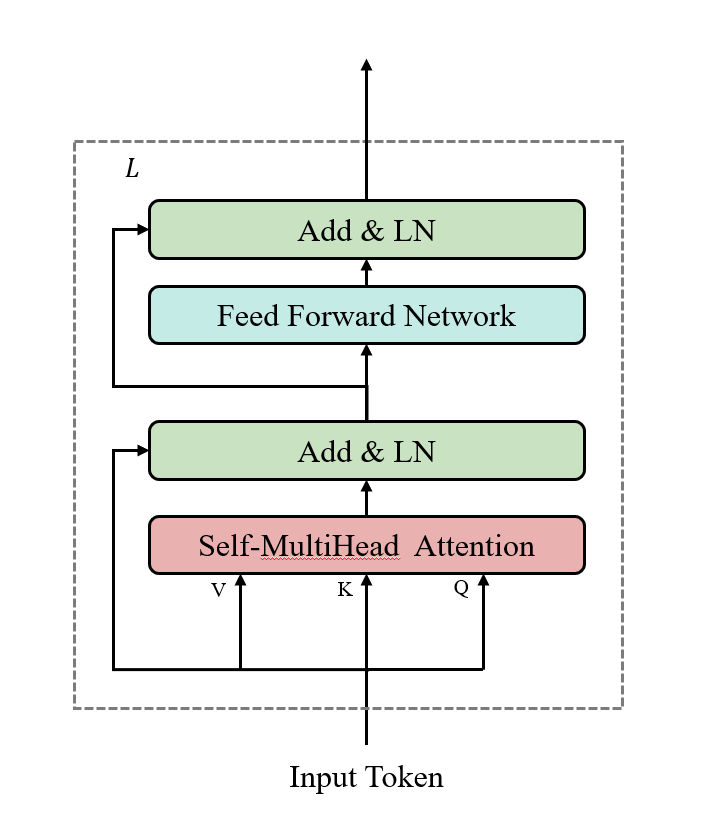


图3: Transformer Encoder

每个Encoder Layer均由一个MultiHead-Self Attention (MHSA)、Feed Forward Network (FFN) 和两个残差累加&Layer Normalization (LN) 组成。首先，输入Token进入Encoder之前，会经过三个Linear层将输入向量进行映射得到，将映射后的三个向量输入MHSA中进行注意力特征计算。Encoder中MHSA采取的和[Attention is all you need]的MHSA相同，均是使用了Scaled Dot Product Attention，具体计算公式如式（3）。

其中表示MHSA的维度数。由于Scaled Dot Production Attention在计算过程涉及两次规模较大的矩阵乘法，因此在输入过程会根据定义头的数量以生成头数量乘3的维度线性映射权重，以此将输入向量映射成较低维度数的三个矩阵。在每个头上都进行运算，最终通过一个大的线性层将注意力计算后得到的矩阵以映射到输入向量维度数。由此堆叠Transformer Encoder能够多次进行MHSA计算从而更好学习整体向量特征。

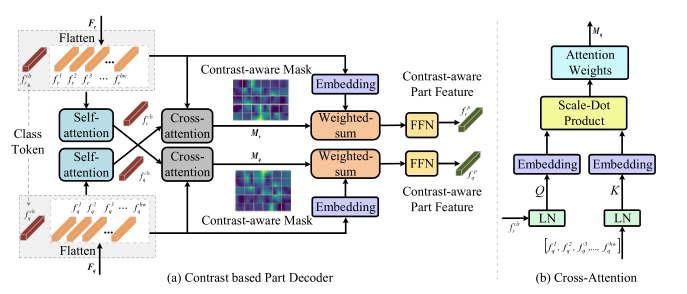
**3.3. Conv-Module**

输入图像经过Encoder后得到的均是平坦化特征向量，因此需要对其进行reshape才能够进行卷积运算。同时平坦化Token后会丢失部分二维信息，因此在reshape后经过卷积模块能够弥补部分二维信息。卷积模块的组成和卷积神经网络相似，由四个Conv2D+ReLU和两个Max Pooling 2D组成，顺序如表1。

|  |  |  |
| --- | --- | --- |
| 层 | Kernel Size | 输出图像形状 |
| Conv 2D+ReLU |  |  |
| Conv 2D+ReLU |  |  |
| Max Pooling 2D |  |  |
| Conv 2D+ReLU |  |  |
| Conv 2D+ReLU |  |  |
| Max Pooling 2D |  |  |

经过Conv-Module后得到的输出会进行两个方向的传播：第一个方向是进行Global Average Pooling (GAP) 计算和平坦化，从而得到平坦卷积特征；另一个方向是成对进入Decoder进行解码注意力计算。

**3.4. Decoder**



Decoder

**4. Schedule for the thesis（论文时间安排）**

**5. Publication（出版物）**

**References (参考文献)**