**Research on the Structure of Offline Handwritten Signature Verification Models Based on Transformer**

**(基於Transformer的離線手寫簽名驗證模型結構研究)**

**Abstract（摘要）**

*離線手寫簽名驗證是生物特徵技術的一個應用场景, 其根據用戶提供的手寫簽名与資料庫中該用戶存儲的手寫簽名進行對比以驗證用戶身份, 在日常生活中被廣泛用於安全認證, 金融交易等安全領域. 學術研究中, 離線手寫簽名驗證包含兩種類型的任務: 作者依賴和作者獨立. 第一種任務是將手寫簽名與資料庫的對應用戶手寫簽名進行對比驗證. 第二種任務是僅通過提供的手寫簽名來判斷是否偽造的. 本文在實驗部分會根據上述任務進行對比實驗, 以驗證提出的深度學習模型結構準確率.*

*本文研究内容如下: 1. 復現相關的離線手寫簽名驗證深度學習模型結構, 分析各個階段的模型存在的缺陷. 2. 提出CNN+Transformer的端到端多尺度特徵的離綫手寫簽名驗證特徵提取模型結構OSVTF, 解決CNN或Transformer在圖像特徵學習的部分缺陷, 並且這兩者相加的組合方式能夠擁有更優秀的圖像特徵學習能力. 3. 在分類器部分, 使用支持向量機、全局平均池化+分類頭兩種方法, 根據實驗結果選擇更優秀的分類器。*

*基於上述研究, 提出一個同時擁有CNN和Transformer圖像特徵學習能力的特徵提取器, 為圖像分類任務的衍生分支提供一個CNN+Transformer風格的模型結構思路. 在日常生活中能夠幫助安全領域人員更好地驗證手寫簽名是否為本人的工作, 從而更好保護個人隱私和財產安全.*

***關鍵詞****: 離線手寫簽名驗證; 端到端; CNN; Transformer; OSVTF.*

**1. Introduction（引言）**

**1.1. Research Background（研究背景）**

Biometrics technology是一种根据指纹、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安全认证、金融交易、门禁系统等安全领域**[1]**的个人身份识别或验证。在识别场景下，系统将根据用户提供的生理或行为特征，识别系统数据库中已有的用户资料。这种场景适用于指纹、虹膜识别个人身份等。其次在验证场景下，用户需要向系统提供验证的身份和特征信息，系统将根据已经存储的特征信息和用户当前提供的信息以此判断当前用户是否为声明的用户，适用于智能手机解锁、国际关口等需要声明个人身份和提供个人特征信息的场景。

手写签名是日常生活中较为重要的个体行为特征，其作为在法律、金融、行政等领域验证个人身份的主要特征，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。手写签名会根据个人的书写习惯产生不同的文字风格，如楷书和草书。甚至随着时间推移，个人书写风格可能会发生改变，相隔一段时间后用户提供的签名（定义为Query）可能和之前输入系统的参考签名（定义为Reference）存在一定的差异。由此在用户手写签名验证的对比工作上会出现一定困难，比如个人签名上每个字的拐弯笔画角度、直线长度和尾部勾起都会存在差异。上述也提到个人书写习惯不同最终得到的手写签名可能略有不同，因此会有一部分人为了能够得到某个用户的身份权限，会恶意伪造手写签名以通过安全隐私检查系统，甚至通过长时间练习以达到和某用户相似风格的手写签名。由此学术界展开了对手写签名验证的一系列研究，希望能够设计出相关模型以帮助相关工作人员更高效地完成流程重复较多的手写签名验证工作，从而让更多工作人员参与更核心的工作。

在学术界中，手写签名验证根据数据收集途径不同分为离线和在线两种类型。离线手写签名的收集过程是通过用户在纸张上进行书写过程后得到；在线手写签名是采用数字化台方式收集用户的手写签名，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**[3]**。除了数据收集途径不同分为不同类型的手写签名验证外，学者们定义了Writer-Independent (WI)和Writer-Dependent (WD)两种类型的分支工作以评价模型算法。WI任务指的是模型结构中输入一对手写签名图像，根据reference签名和query签名提取的特征从而判断query签名是否为伪造的。WD任务则是在WI任务输入的基础上增加作者id，根据作者id使用对应的作者分类器以判断query签名是否为伪造的。这两种分支任务虽然在最后输出的都是query签名的标签类别（genuine或forged），但是在不同程度上能够反映模型不同的性能，WI任务注重学习签名群体的全局特征，不需要针对每个用户单独学习一个作者分类器，具有新用户无需重新训练的优良拓展性，同时能够体现模型的泛化和判别能力强弱；WD任务则是额外为每个用户单独训练一个作者分类器，能够考虑到单独作者不同风格签名依然认为是真实签名的高精度性，但是需要每个用户收集足够多的样本，训练成本高，泛化能力较弱。由此WI和WD任务能够在不同程度上体现模型的整体性能，根据两种任务的评价指标来综合评价模型性能。

随着科技水平的发展，人工智能领域已经能够根据用户提供的样本图和关键词生成以假乱真的图片，正常人肉眼难以鉴别真假。因此会存在部分人为生成伪造签名地行为，从而达到以假乱真来通过安全系统的验证，造成个人隐私权、财产被侵犯、窃取等危险后果。单靠人工力量进行验证会存在一定的社交风险和判断错误，因此研究离线手写签名验证，一定程度上能够降低伪造签名通过验证的风险，通过相关算法模型的帮助，尽可能减少人工验证的错误，从而更好地保证个人隐私权和财产安全。

**1.2. Research Motivation and Importance（研究动机和重要性）**

本文主要研究的是离线手写签名验证任务，其本质上是一种图像分类任务的衍生分支，与传统图像分类任务不同的是，离线手写签名验证任务要求输入的是reference和query一对的签名图像样本，WD任务需要额外输入一个作者id以训练作者独立的分类器。离线手写签名验证模型将根据输入的reference和query签名图像提取相关的图像特征，根据特征进行对比从而输出genuine或forged的类别标签。由此该任务具有一定的挑战性和创新的需求，设计模型或算法架构时不仅需要考虑模型提取一对图像特征的关键性，同向双流的模型架构将直接影响reference签名和query签名图像提取的特征关键性，一旦提取的特征不具备那么重要的信息，那么将导致模型的判断能力急剧下降。同时由于WI和WD任务的定义以及现实生活中的运用场景，要求了其必须具备较高的准确率和泛化能力才能防止个人隐私财产侵犯，更是提升了对模型性能的需求，由此对于离线手写签名验证模型的设计与实验是一项极具挑战的图像任务。

在离线手写签名验证模型的设计上，分为特征提取器和WD任务的作者分类器两部分。在过去学者们的研究中，大多数根据样本分布等数据集群特征以人工设计图像特征提取方法，采取传统机器学习的方法以判断签名是否为伪造的**[1]**，如采取特征之间的距离或SVM以判断签名是否伪造的。但是这种人工设计特征的方式具有特定数据集群的缺陷，它必须要求数据集的作者签名风格统一，细微的差异将作为离群值导致传统机器学习方法的判断准确率下降，但是实际生产工作中是需要拥有这些一定差异包容度的，因此这种采取人工设计提取特征的方法逐渐淘汰。学者们希望能够拥有一种更关注某个部位并且不用多次人工干涉的图像特征提取器，随着近六年深度学习的快速发展，具有共享参数的局部视野运算的卷积神经网络 (CNN)在诸如MNIST、ImageNet这类传统图像分类任务上取得了不俗的成绩，与传统机器学习方法对比CNN的准确率和泛化能力更是得到了进一步验证。由此手写签名验证领域的学者们引入了CNN作为特征提取器，并且实验证明这种方式的准确率和泛化能力相较之前的方法有了很大的突破**[2]**。

即使CNN提取图像特征的能力出众，其核心思想的卷积核运算具有一定的局部视野强化学习能力，能够关注图像的局部特征；但是对于手写签名来说不仅需要关注到字体笔画的拐角程度，还需要关注签名图像整体的风格、字体等因素，可能会出现同一作者签名位置不同的情况，因此CNN作为特征提取器还具有一定的缺陷。随着自然语言处理领域的发展，具有全局特征学习的Transformer**[9]**的出现引起了学者们的注意，随即图像领域的学者们引入了Transformer的Encoder-Decoder架构对分类任务进行实验**[?]**，实验证明这种方式在模型参数收敛的情况下性能要优于CNN的特征提取器，但是训练达到模型参数收敛的条件较为苛刻，原因是Transformer的注意力机制需要对整体图片进行学习，CNN则是采取了共享参数的方式学习图像局部特征，从而能够更快达到模型收敛，但是在实际使用上泛化能力较差，需要不断微调数据集以达到更高的模型性能。由此衍生出CNN作为Backbone提取多通道特征图，拉平后作为特征向量进入Transformer中进全局行注意力特征的运算**[?]**，从而达到多通道特征图在缺乏整体信息的情况下，经过Transformer的注意力机制加强整体信息的特征，实验证明这种方式可以更有效地学习图像特征，模型从而达到更优秀准确率和泛化能力。综上所述，CNN+Transformer风格的模型架构成为了近几年图像领域的主流框架，离线手写签名验证任务中也开始引入这种框架并且实验证明这种风格的模型架构具有良好的模型性能**[?]**，因此本文提出的基于Transformer的离线手写签名验证模型同样是CNN+Transformer风格的模型架构，顺应近几年深度学习中该领域的发展趋势，并且在此基础上增加了多尺度融合特征，在满足日益增长的高清图片分辨率要求的同时，能够在多个尺度上学习签名图像字体笔画拐角等特征，弥补了经过CNN后多通道特征图在尺度方面上的不足。在最终输出分类阶段，与传统的图像分类任务不同的是，本文提出的模型在最后输出的特征中将收集之前各个模块的特征，进行整体拼接后进行类别标签预测或模型训练，这种方式已经被证实是有效的**[?]**。研究的挑战部分在于如何有效地运用多尺度融合特征进行类别预测和模型训练，这种运用整体模型特征的训练、预测方式能够在整体上更好地训练各个部份的模型权重，模型训练到收敛的阶段将会超越以往大部份深度学习模型架构，对于相关领域而言更是提供了一种新颖的思路，能够融合卷积神经网络和注意力机制的特征方式，为后续更简洁的融合模块提供了一定的设计思路；相反这种方式将增加的设备的运算推理压力和模型训练成本，需要经过多种实验以证明这种方式是有效的。

**1.3. Research Objectives（研究目标）**

本文中提出的OSVTF架构，整体上是基于孪生网络的思路设计的模型，如图1所示。图示

AI 生成的内容可能不正确。

图1：OSVTF模型架构流向图

孪生网络的核心思想是两个共享权重的子网络，分别接受两个输入，输出两个特征向量，随后使用距离函数计算相似度**[?]**。在离线手写签名验证的WI和WD任务中，输入reference和query签名的一对图像，因此整体架构方面参考了TransOSV **[?]**，在此基础上增加了backbone和FPN Fusion模块，对Encoder、Conv-Module和Decoder部分进行了一系列的优化和调整，更有利于各个部份特征的变化与学习，实验部分则是需要证实这些模块能够提升模型对于伪造签名的判断能力和不同场景的泛化能力。

在整体的研究计划和实验上，依旧采取WI和WD两种大方向的实验方案以整体评价模型性能和质量，在此基础上将针对CNN的多通道特征图与Transformer的注意力机制特征进行一定的可视化，需要关注这些模块在图像特征学习上是否能够更好地把握住图像的部份重要特征信息。此外，多尺度特征的融合方式有很多种，如Mask-RCNN**[?]**采取的特征金字塔网络风格（FPN-Style）**[?]**，这种融合方式包括但不限于均值累加、直接相加、拼接，因此后续会对这几种融合方式进行控制变量的模型训练以获得最佳性能的融合多尺度特征图方式。此外以往孪生网络的方式是根据距离函数计算相似度以完成相关任务，这种方式能否在增加了多尺度融合特征后依然有效需要进行一定的验证，本文在此基础上将参考传统图像分类任务的CNN处理分类方式，增加全局平均池化层的方式作为分类器对特征进行分类标签预测，以此判断是否会优于以往的距离相似度判断。

综上所述，本文将分为三部分研究阶段：1.初步对OSVTF模型架构进行模型周期训练，验证多尺度特征和模型的优化调整方案能否在原有架构上有所提升；2.针对FPN Fusion模块的多尺度融合方式，将在第一阶段的基础上采取控制变量法对模型进行小幅度的微调训练，筛选出最佳的多尺度特征融合方法；3.针对WI任务，在最后的分类器阶段增加全局平均池化的分类器，和以往离线手写签名验证的采取距离相似度预测的方法进行对比，从而判断两种分类器的优势和劣势，并且在WD任务中采取更优的分类器进行实验以验证是否奏效。

***2. Literature Reviewx（文献综述）***

*离线手写签名收集的图像数据，进行模型或算法验证时需要经过一定的预处理，而良好的数据预处理技术一定程度上会提高系统验证准确率。为了提高签名图像质量，学者们采取了RGB图像转换到GRAY单通道图像的处理，更有甚者为了去除噪点使用了平滑像素的图像处理****[4]****。这种处理方式能够很好的减少了白色纸张上收集的手写签名图像不必要的特征，如RGB三通道图像中某些通道像素点的值为255，而具备手写笔迹的像素点的值往往很小。但是这种处理方法存在具有笔迹特征的部分的像素点的值非常小，而其他空白部分的像素点的值非常大的缺陷，因此后续学者们对这些图像进行黑白反转的处理，经过验证这种处理方式能够得到更优秀的验证准确率****[4]。***

*离线手写签名早期均是采取机器学习的方式，如 Edson等人利用隐马尔可夫模型提出了离线签名验证系统****[5]****。这种传统机器学习方法存在缺陷：其依赖于人工设计的特征进行训练，分类的准确性与这些特征直接相关****[2]****。因此无论分类阶段模型如何更替，在数据预处理和特征提取阶段，仍需要花费大量时间提取签名图像特征。在2005年，Edson等人对SVM和HMM用于离线手写签名验证上进行了对比实验****[6]****，实验结果表明SVM的效果优于统计模型的HMM。SVM是一种广泛应用的机器学习技术，因其在二分类任务中的高效性而受到亲睐****[7]****，而离线签名验证本质上是属于鉴别签名是否为赝品的二分类任务，因此后续离线手写签名验证任务上大多数都采取SVM作为分类器，更专注于图像预处理和提取手写签名图像特征。*

*随着深度学习和卷积神经网络发展，2012年Krizhevsky A.等人提出AlexNet****[8]****在ImageNet比赛的图像分类任务上遥遥领先使用传统算法的第二名参赛者夺得冠军，使得卷积神经网络为代表的深度学习逐渐成为图像任务任务的主流方法。后续学术界推出更深层次的卷积神经网络VGG****[9]****，该架构是在AlexNet的基础上增加了更多的卷积层，提取更多通道的特征图。但是这种叠加卷积层的方式，将导致提取的特征图尺寸变得更小的同时，导致其他尺寸的特征丢失，为了解决这种问题，在2016年He Kaiming等人提出了残差连接的ResNet****[10]****。ResNet和VGG这些卷积神经网络类似，都是叠加了许多的卷积层，但是在每个尺寸特征图的卷积层之间，增加了一个残差连接。将原始输入卷积层的特征图累加到卷积层输出的特征图上，从而保留了该尺寸的一些原始特征。这种方式在监督学习的卷积神经网络中，经过权重共享的卷积核运算，能够一定程度上弥补原始图像特征信息，并且这种残差连接的方式成为后续神经网络模型的一种主流方式。卷积神经网络在图像分类任务上的优质表现，使得其他领域的任务也开始借鉴这种卷积神经网络的方式以完成任务，离线手写签名验证也是如此。在2017年，L. G. Hafemann等人使用卷积神经网络进行离线手写签名验证****[11]****，与图像分类任务的CNN相同，均是采取了卷积层+max pooling layer的方式提取手写签名图像特征图，最后输出部分略有不同；该模型是将最后全连接层的特征经过了两个输出部分：作者预测和是否为赝品的预测。并且要求输入是单张手写签名图像而不是成对手写签名图像，是属于WI任务，而WD任务需要Reference。该论文提供了一种监督学习的手写签名图像特征提取方式，并且实验结果证明这种方式显著提高了离线签名验证的准确性，为后续使用CNN和深度神经网络进行手写签名验证打下基础。*

*上述提到卷积神经网络是使用了卷积核对图像像素进行遍历以学习图像特征，但是卷积运算存在局部视野的缺陷。虽然它能够学习图像的局部特征，但是在学习全局特征的效果上较差，因此学者们相继提出ResNet的残差连接、用于目标检测的多尺度网络FPN-Style****[12]****来优化这些问题，但是并没有从根源上解决缺少全局特征能力，且图像一定会附带多余的信息和无用像素点。随着深度学习的自然语言处理领域发展，2017年Vaswani A.等人提出用于机器翻译的Transformer模型****[13]****。该模型采取了Encoder-Decoder的机器翻译主流架构，与以往机器翻译深度学习模型不同的是，Transformer引入了注意力机制。在注意力机制的作用下，能够对词向量进行全局特征感知，生成预测词语时会参考上一个词语的特征信息和输入词向量的特征信息，因此可以生成长短不一句子预测。并且实验证明了Transformer相较以往的Encoder-Decoder机器翻译模型拥有更优秀的全局泛化能力。*

*介于Transformer的注意力全局特征感知能力，Dosovitskiy A. 等人于2021年首次将Transformer Encoder用于图像分类任务中，命名为Vision Transformer (ViT)****[14]****。ViT的内部结构完全抛弃了卷积层，将输入图像进行patch和 Embeddings操作 后，每个patch的embedding当成词向量单个token，由此图像被拉平后当作词向量进入Transformer Encoder。经过若干个Encoder Layer，最后设置一个MLP以映射类别概率，从而完成图像分类任务。实验表明这种方式的模型在图像分类任务上得到了领先卷积神经网络的识别准确率，由此也衍生出CNN + Transformer的深度学习模型架构，如图像目标检测的DETR****[15]****、图像分割的MaskFormer****[16]****。这种方法采取的是摘除最后分类的Linear层卷积神经网络，输入图像以提取多通道特征，将该特征作为Transformer的输入向量从而进行全局特征感知。CNN + Transformer的开山之作DETR在MS COCO****[17]****的验证效果虽然没有以往采取NMS等后处理操作的R-CNN****[18]****，YOLO系列网络****[19]****识别准确率高，但是该方式的优势在于不需要任何的Archer Box等目标检测先验框预设置和NMS后处理，是一种端到端的模型架构。这种方式一定程度上简化了模型推理流程，但是也增加了模型训练时间成本，需要设计更精细的训练方案以保证模型参数收敛。*

*Transformer在图像任务上相较CNN具有优秀的全局特征感知能力，因此离线手写签名验证领域中学者们提出了一种基于Transformer的模型架构TransOSV****[20]****。该模型采取类似Encoder-Decoder架构，输入Reference和Query均经过RGB转GRAY图像预处理后，进入以ViT Encoder作为Holistic Encoder。其次将Holistic Encoder输出进行卷积运算，最后对Reference和Query的特征进行 Contrast based Part Decoder运算。在训练过程中，TransOSV是汇总了Transformer Encoder输出class特征、卷积模块输出特征和解码器输出特征以计算训练损失完成模型训练，最后特征分类器同样采取支持向量机进行验证签名是否伪造。该架构在Decoder的Cross-attention中将Reference特征和Query特征进行了注意力计算，对Reference和Query的特征进行关联度注意力学习，能够更好地联系Reference和Query的关系。但是这种架构由于输入图像是直接patch和Embeddings进入Transformer Encoder，一定程度上会缺少图像多通道特征信息。由于ViT其图像尺寸大小是全局约束的，当图像进行resize缩小的时候一定会丢失部分关键信息。*

*综上所述，本文将在TransOSV架构基础上，采取CNN + Transformer的方式，新增backbone以在图像进入Transformer Encoder之前先由CNN提取多通道特征图，随后提取backbone的多尺度特征图进行平坦化和拼接以补充手写签名图像特征，再进入Transformer Encoder。后续解码部分再引入多尺度特征图特征，从而形成适配多尺度手写签名图像特征的Transformer离线手写签名验证特征提取模型。特征提取后分类器将采取全局平均池化和分类头的方式以进行手写签名验证是否为赝品。*

**3. Methods（论文主题算法等）**

**3.1. Overview** 图示

描述已自动生成

图1: Offline Signature Verification Transformer (OSVTF)

提出的OSVTF如图1所示。输入一对尺寸相同的手写签名图像Reference 和Query ，首先经过Backbone提取多尺度特征图，将多尺度特征图进行平坦华和拼接后经过权重共享的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. Backbone**

图示

描述已自动生成

图2: Backbone structure

在Backbone上采取CNN的架构，将图像输入卷积神经网络，在卷积层运算后，将得到一个多通道特征图。如图2所示，以ResNet-50举例，输入参考签名图像，其中表示图像的高度、宽度和图像通道数。首先经过第一个卷积层，Conv1的输出特征图每一个像素点计算公式如下：

其中，表示backbone中第一个区块（Conv1）输出特征图在位置(I, j)的第k个通道上的值；表示Conv1的卷积核权重和偏置；表示卷积核的高和宽；表示卷积核位置索引。后续的Conv2、Conv3、Conv4、Conv5中每一个Conv2D运算均与上相同。

ResNet-50与传统CNN不同的是，在经过一定数量的卷积层运算后，如图2的Conv2部分，将进行一次残差连接，即将输入当前卷积层之前的特征图累加到当前卷积层运算后的特征图上，以此弥补部分卷积运算中缺少的部分特征。每个区块的参考签名图像特征计算如下：

其中，表示backbone中第个区块，如表示Conv1；表示特征图像经过的1个或多个卷积运算，如时，将经过3次由3个卷积运算堆叠的层。在经过一次堆叠的卷积运算后，为了解决训练难收敛问题，使用Rectified Linear Unit (ReLU) **[21]**激活函数和Batch Normalization (BN) **[22]**运算。ReLu激活函数的思想是当时，ReLU输出是线性的；相反，当时ReLU输出为0，计算公式如下：

BN的思想是在模型批次训练的过程中，对当前运算输出的结果进行当前批次的均值和方差进行标准化，而在推理阶段则使用训练中累计的全局均值和方差。计算公式如下：

在当前公式下，表示第个样本在第个特征的值。

在本文提出的OSVTF中，与DETR的方式不同的是，经过backbone运算后，将的所有输出特征图全部拼接起来，再输入至Encoder中进行后续运算，以此增加多尺度签名图像特征。

**3.3. Encoder**

Encoder是沿用了ViT(Vision Transformer) **[14]**的模型架构如图3所示。

图示

描述已自动生成

图3: Vision Transformer

ViT首先将输入特征图分成块和平坦化，其中patch number计算如下式。

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。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推理如下式。

其中表示Embeddings。表示Position Embeddings。每个Encoder结构相同，由个堆叠形成，如图4所示。

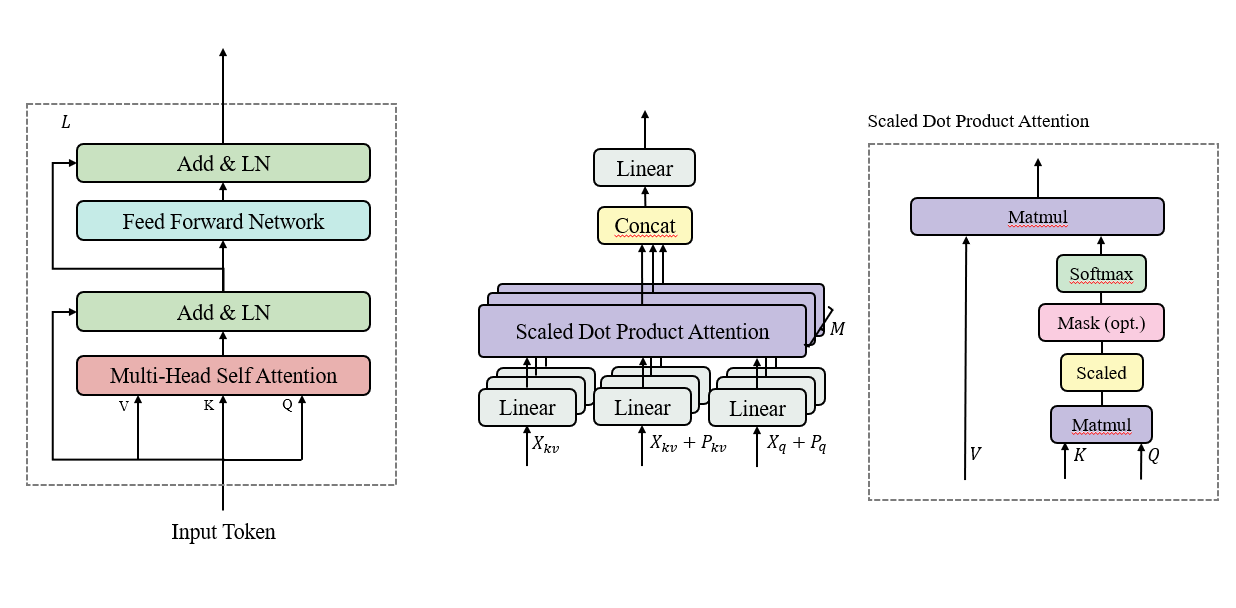


图4: Transformer Encoder

每个Encoder Layer均由一个Multi-Head 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Transformer**[13]**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。

**3.4. 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。

表1：Conv-Module结构信息

|  |  |  |
| --- | --- | --- |
| 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 |  |  |

其中每个Conv2D的运算与backbone的相同。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。

**3.5. Decoder**

**图示

描述已自动生成**

图5：Decoder & Cross-Attention

OSVTF中的Decoder与Transformer Decoder不同，它采取了两个MHSA和Cross-attention，如图5-(a)所示。与ViT输入阶段相似，在输入平坦化卷积特征之前，会初始化一对与平坦化卷积特征相同维度数的可学习参数，定义为Class Token。它的作用与ViT-Encoder的Class Token作用一致，均是定义了它在一个新的维度上学习图像是否为赝品的一个调整权重，从而使得模型能够更好地学习图像特征。

同上述MHSA一样，计算公式如式（3）。Decoder中先对插入了class token的平坦化卷积特征进行MHSA运算得到注意力机制的平坦化特征向量。随后为了能够学习reference和query之间的关系，引入了Cross-attention如图5-b，基于注意力机制定义的线性映射为Query矩阵，MHSA输出的平坦化注意力特征为Key矩阵。与MHSA相似，根据reference或query的可学习参数class token，和另外图像的平坦化特征进行注意力机制计算，一定程度上注意力机制内部的查询矩阵能够反应图像之间的关注度，由于没有输入值Value矩阵，因此Cross-attention得到注意力权重。该mask权重包含了reference和query之间各个Token（即各个像素点）的注意力权重，由此对输入Decoder的特征再经过Embeddings后与注意力权重进行广播后累加，从而得到的是一对注意力机制的平坦化注意力特征。经过FFN后得到最终的Decoder输出，这对特征后续将根据FC Loss来优化Decoder模型参数。

**3.6. Loss Function**

OSVTF的训练过程将采取了两种损失函数：Sparsity Loss, Focal Contrast Loss。

**Sparsity Loss**

由于OSV的图像中只有少数区域包含签名验证任务的判别信息，且Decoder中具有稀疏性，将采取计算交叉熵的方式以生成对比度感知掩码的多样化分布**[20]**，从而更好的训练Decoder的线性映射参数，计算公式如式（4）。

**Focal Contrast Loss**

在以往的非监督算法中，对比两个样本或特征之间的差异是采取计算distance的方式以衡量两者差异。定义为计算包括Encoder、Conv-Module和Decoder输出的一对平坦化特征向量之间差异。由此可以定义评价两个对象差异的Contrastive Loss**[23]**计算公式如式（5）。

其中，表示query的signature是forged，相反时表示赝品。为了抑制模型训练过程过拟合的问题，基于CaP**[24]**引入双倍边际损失如式(6)。

但是上述的双倍边际损失中存在缺点，当两对reference和query signature样品中query为正品，即的情况下，会得到两个此样本模型推理特征的距离。若，那么损失函数应该给予样本更大的损失/权重。但是在上述的损失函数中会平等地对待样本，因此基于Focal loss**[25]**的不平衡训练样本情况，改良得到最终的Focal Contrast Loss如式（7）。

其中表示Sigmoid function。为两个边距值。表示缩放因子。

**3.7. Dataset**

离线手写签名验证中使用的常见公开数据集如下BHSig-B、BHSig-H**[26]**和CEDAR**[27]**。本文在实验部分将使用这三个数据集以验证OSVTF的性能。

**BHSig-B & BHSig-H**

由印度理工学院古瓦哈提分校（IIT Guwahati）发布。其中BHSig-B是包含了100个用户的孟加拉语（Bengali）手写签名，BHSig-H是包含了160个用户的印地语（Hindi）。每个用户手写签名了24张真实签名和30张伪造签名。

**CEDAR**

由Center of Excellence for Document Analysis and Recognition开发并发布。该数据集包含了55个用户的英文手写签名。每个用户手写签名了24份真实签名和24份伪造签名。

表2: 现有复现模型结果对比

|  |  |  |  |
| --- | --- | --- | --- |
| Model | FRR | FAR | EER |
| BHSig-B dataset (50/50) | | | |
| SigNet | 13.89 | 13.89 | 13.89 |
| CaP | 3.96 | 3.96 | 3.96 |
| TransOSV[Base] | 11.26 | 11.26 | 11.26 |
| **My TransOSV [Reproduce]** | **14.73** | **14.73** | **14.73** |
| BHSig-H dataset (50/50) | | | |
| SigNet | 15.36 | 15.36 | 15.36 |
| CaP | 5.97 | 5.97 | 5.97 |
| TransOSV[Base] | 7.89 | 7.89 | 7.89 |
| **My TransOSV [Reproduce]** | **10.41** | **10.41** | **10.41** |

**References (参考文献)**

[1] L. G. Hafemann, R. Sabourin, and L. S. Oliveira, “Offline handwritten signature verification — Literature review,” in *2017 Seventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pp. 1–8, Nov. 2017.

[2] Y. Muhtar, W. Kang, A. Rexit, Mahpirat, and K. Ubul, “A Survey of Offline Handwritten Signature Verification Based on Deep Learning,” in *2022 3rd International Conference on Pattern Recognition and Machine Learning (PRML)* pp. 391–397, Jul. 2022.

[3] D. Banerjee, K. Dasgupta, D. Ganguly, and K. Chatterjee, “A Survey of Offline Handwriting Signature Recognition,” Mar. 2019.

[4] N. Y. Choudhary, R. Patil, U. Bhadade, and B. M. Chaudhari, "Signature Engineering and Applied Sciences Research (IJIEASR)," *International Journal of Innovative Engineering and Applied Sciences Research*, vol. 2, no. 1, Jan. 2013.

[5] J. Edson, R. Justino, E. Bortolozzi, and R. Sabourin, "An offline signature verification using HMM for random and skilled forgeries," in *Proc. 6th Int. Conf. Document Analysis and Recognition*, pp. 1031-1034, Sept. 2001.

[6] E. J. R. Justino, F. Bortolozzi, and R. Sabourin, “A comparison of SVM and HMM classifiers in the off-line signature verification,” *Pattern Recognition Letters*, vol. 26, no. 9, pp. 1377–1385, 2005.

[7] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, “Support vector machines,” *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, Jul. 1998.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS)*, Curran Associates Inc.,pp. 1097–1105, 2012.

[9] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[11] L. G. Hafemann, R. Sabourin, and L. S. Oliveira, “Learning features for offline handwritten signature verification using deep convolutional neural networks,” *Pattern Recognition*, vol. 70, pp. 163–176, 2017.

[12] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature Pyramid Networks for Object Detection,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, Jul. 2017.

[13] A. Vaswani *et al.*, “Attention is All You Need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.

[14] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[15] N. Carion *et al.*, “End-to-End Object Detection with Transformers,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 213–229, 2020.

[16] B. Cheng, E. Xie, H. Zhang, Y. Zhu, and Y. Qiao, “MaskFormer: Masked Image Modeling for Visual Tasks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[17] T.-Y. Lin *et al.*, “Microsoft COCO: Common Objects in Context,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 580–587, 2014.

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016.

[20] Y. Zhang *et al.*, “TransOSV: Offline Signature Verification with Transformers,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[21] V. Nair and G. E. Hinton, ‘Rectified linear units improve restricted boltzmann machines’, in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, in ICML’10. Madison, WI, USA: Omnipress, pp. 807–814, 2010.

[22] S. Ioffe and C. Szegedy, ‘Batch normalization: accelerating deep network training by reducing internal covariate shift’, in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, in ICML’15. Lille, France: JMLR.org, pp. 448–456, 2015.

[23] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1735–1742, 2006.

[24] X. Lu, L. Huang, and F. Yin, “Cut and Compare: End-to-End Offline Signature Verification Network,” in *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)*, pp. 176–183, 2021.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, 2020.

[26] A. Bhatawdekar, S. Bhattacharya, R. Khatri, and R. Tiwari, “BHSig260: A Dataset for Offline Signature Verification,” *arXiv preprint arXiv:2004.07563*, 2020.

[27] S. Pankanti, S. Prabhakar, and A. K. Jain, “Cedar: A database of handwritten signatures for benchmarking signature verification systems,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP),* pp. 5–8, 2020.