**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**[?]**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**[?]**。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）**

本工作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**[?]**，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**[?]**。

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**[?]**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**[?]**，因此本工作提出的基于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，本工作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）**

本工作中提出的Offline Signature Verification TransFormer (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，本工作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，本工作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**

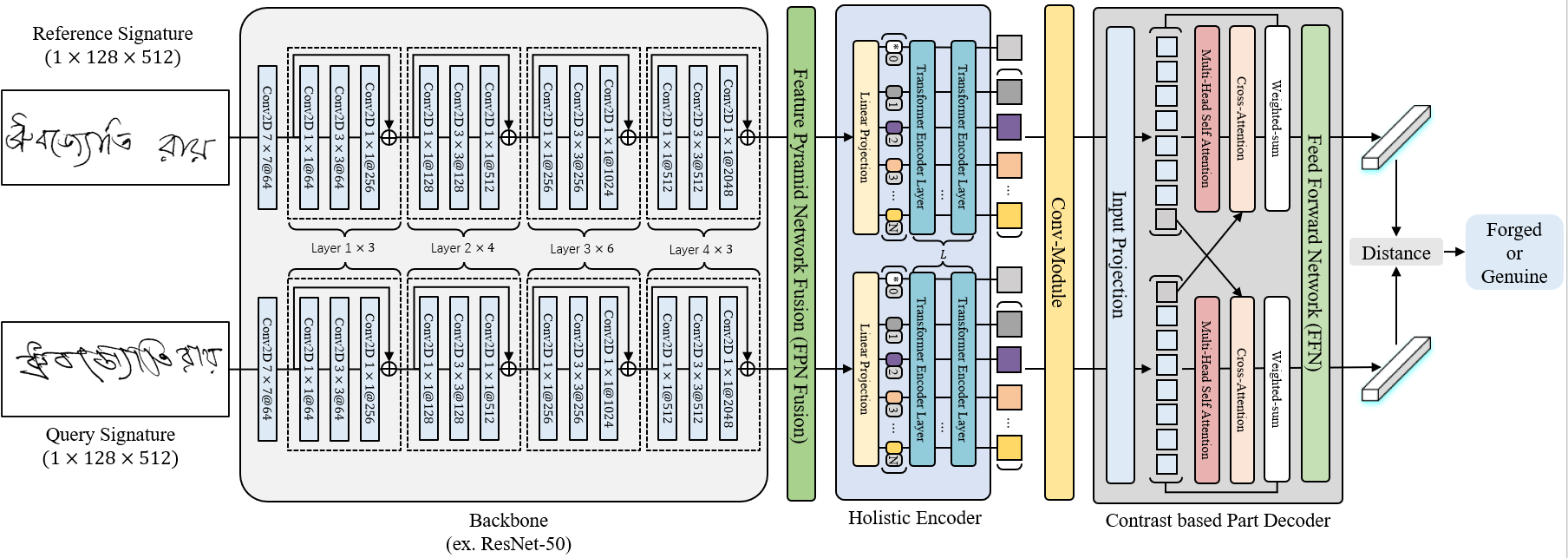
本工作将定义输出的标签类别中，forged为1，genuine为0，输入的reference签名图像定义为，query签名图像定义为

图2: OSVTF模型架构示意图

提出的多尺度特征融合的Offline Signature Verification TransFormer (OSVTF) 模型架构如图2所示，它主体是由Backbone 、FPN Fusion 、Holistic Encoder 、Conv-Module 、Contrast based Part Decoder五个部分组成。在模型中我们将输入一对签名图像样本和，初始输入图像经过共享权重的Backbone（图1中是以ResNet-50**[?]**为例子）得到多尺度特征图集合，经过FPN Fusion模块将不同尺度特征图进行融合，输出多尺度融合特征图。随即进入共享权重的Holistic Encoder得到holistic平坦特征和patch embeddings 和。为了进一步整合特征图，将上述的patch embeddings调整形状至二维形状，经过共享权重的Convolutional Module (Conv-Module)得到它的输出，它们将直接进入Contrast based Part Decoder以获得交叉注意力的对比part平坦特征。最后预测阶段，对和进行Global Average Pooling(GAP)运算得到多尺度融合平坦特征和卷积平坦特征。将全部平坦特征进行拼接得到OSVTF的总特征向量和，将基于和进行判断是否为伪造的。接下来将深入解析Backbone、FPN Fusion、Holistic Encoder、Conv-Module和Contrast based Part Decoder结构。

**3.2. Backbone**

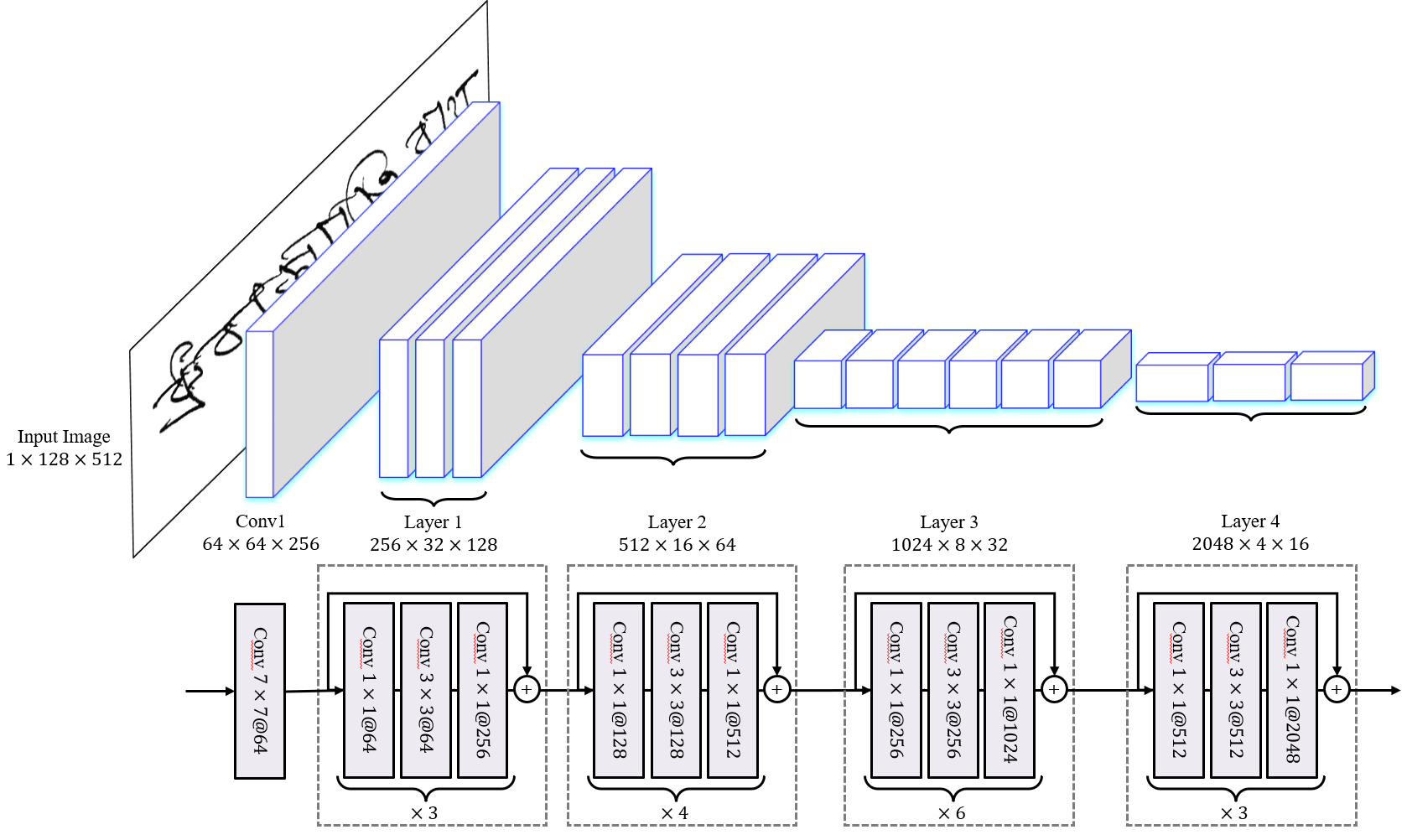


图3: Backbone (ex. ResNet-50) 架构示意图

在Backbone部分将采取舍弃输出部分中Average Global Pooling和全连接层的CNN的架构。CNN架构中大多数由输入部分的卷积层和4个定义的卷积层组成，以ResNet-50**[?]**为例如图3所示，每一层都是由若干个bottleneck块组成，单个bottleneck块由两个卷积核为的卷积2D层组成，这样安排主要有以下目的：1. 降低特征维度以达到提升计算效率的目的；2. 中间的卷积在较低维空间中提取如笔画、边缘等局部空间特征具有保持感受野的同时减少冗余，这样先减少表达通道数后再扩充至输入的表达通道能够在bottleneck中找到更好的特征学习参数，具有较好的特征抽象表达能力；3. 在实际运算当中能够搭配identity skip connection能够快速传播梯度，避免梯度消失，同时这样的设计使得中间参数量和计算负担大大降低，能够堆叠更多的层数如ResNet-101，促进更深网络的训练；4. 这种模块化设计在卷积参数上具有容易微调的优势，中间通道数可以根据环境进行调整可以控制性能-效率平衡。除此之外，ResNet系列网络引入了残差连接，解决了图像经过深度卷积运算后部分特征缺失的问题，在每个bottleneck块中引入残差连接，将输入部分的特征图与bottleneck的输出进行累加，能够一定程度上减少特征图在经过一定数量卷积层运算后特征丢失的问题。

假设卷积层输入特征图，输出特征图，卷积2D层的卷积核大小（卷积核权重），填充，步幅，偏置项。对于特征图的第个输出通道、空间位置为的像素点计算公式如下：

此外，每个bottleneck块中，除了最后一层升维的卷积核卷积2D层后只接BatchNorm (BN)**[?]**运算，每个卷积2D层后会紧接着BN和ReLU激活函数**[?]**运算，经过BN和ReLU激活函数的空间位置为的像素点值如下：

其中，是卷积2D层输出的批次多通道特征图中第通道平面的特征统计均值和方差，是可学习标度和平移参数。由此可见BN的运算思想是对某个批次数据进行归一化，具有提升模型参数训练时收敛速度的效果。在每个bottleneck块运算中，对经过三次卷积2D运算的特征图进行一个累加残差操作，即输入bottleneck块的特征图会累加到三次卷积2D运算输出的特征图上，以此传递到下一个bottleneck块。同时，backbone每个层运算后会经过一个下采样运算，以此在多通道特征图上减少冗余特征，将采取Max Pooling 2D的方式以等价下采样操作**[?]**。上述的残差链接方式，能够在保证特征图经过卷积运算后保留一定的原始图像特征，因此ResNet系列网络最适合作为深度模型的Backbone以提取图像多通道特征图。

采取backbone主要是为了提取图像多尺度的特征图，因此在最终输出上包含了除输入卷积层外的每个层的不同尺度特征图，如下：

其中，每个尺度的多通道特征图，表示layer 1输出特征图的宽度、高度和通道数，当backbone为ResNet-50时，），每一层特征图的通道和尺寸之间都是两倍的关系。

**3.3 FPN Fusion**

为了能够有效利用backbone提取的多尺度特征图，采用Feature Pyramid Network Fusion (FPN Fusion)模块以融合不同尺度的特征图，其采用的是FPN-Style**[?]**的模型架构，如图4所示。

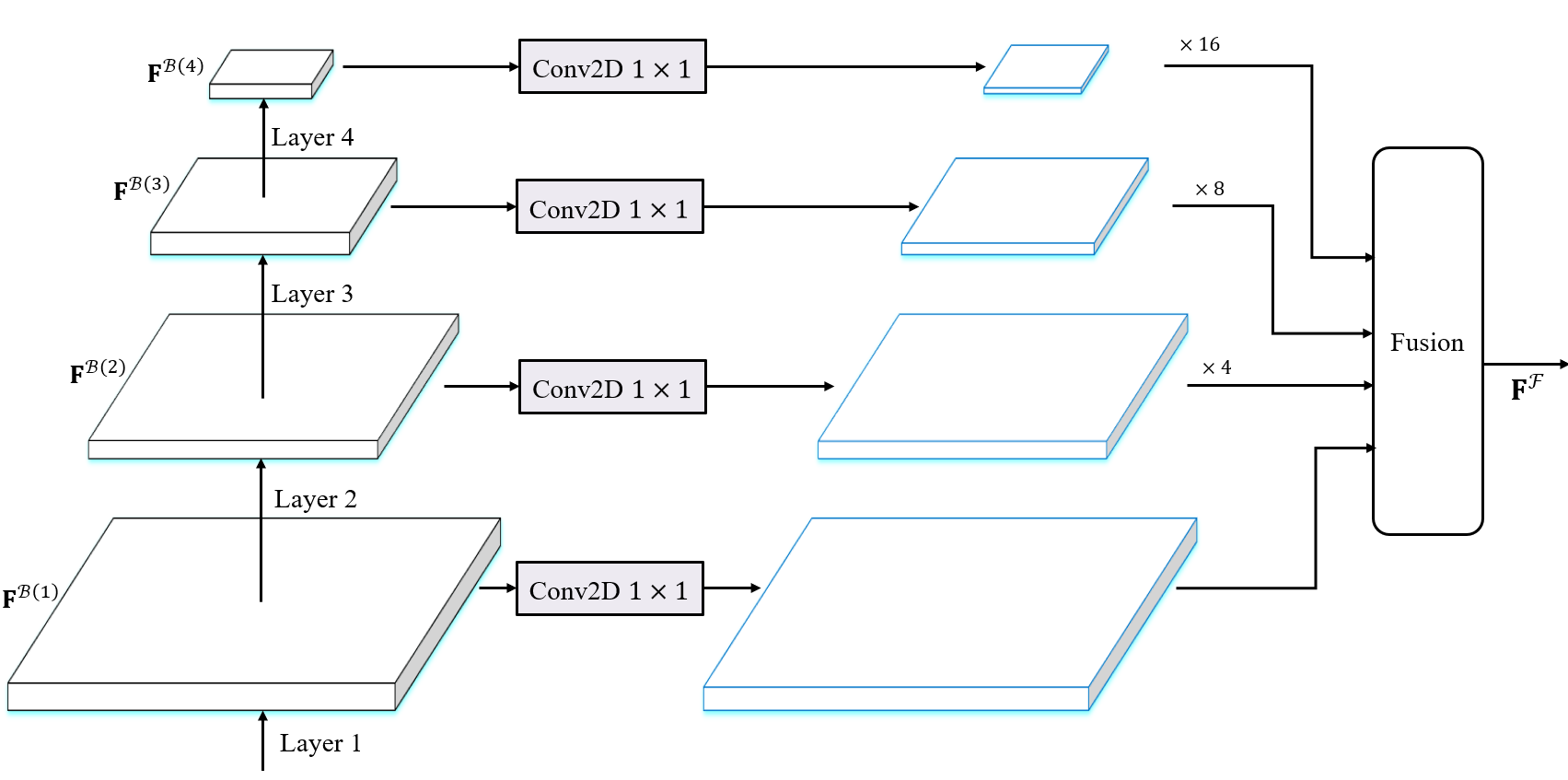


图4: FPN Fusion示意图

其中bottom-up过程为backbone推理获得不同尺度特征图的过程。在bottom-down中，对于每一层特征图将采取卷积核为的卷积2D层进行映射，将每一个尺度的不同通道数特征图映射到相同特征维度上，一定程度上达到降维的操作，从而加速模型推理速度。同时对进行上采样，增加尺寸至与相同大小，后续根据融合方式以得到多尺度融合特征图。在上采样过程中采取双线性插值**[?]**，对于第层特征图上采样输出，空间位置的值计算公式如下：

其中，表示宽度和高度方向上的小数部分，表示放大倍数，表示数值的整数部分。这种上采样方式能够很好地保留原始特征地结构，同时引入平滑过渡如图5所示。

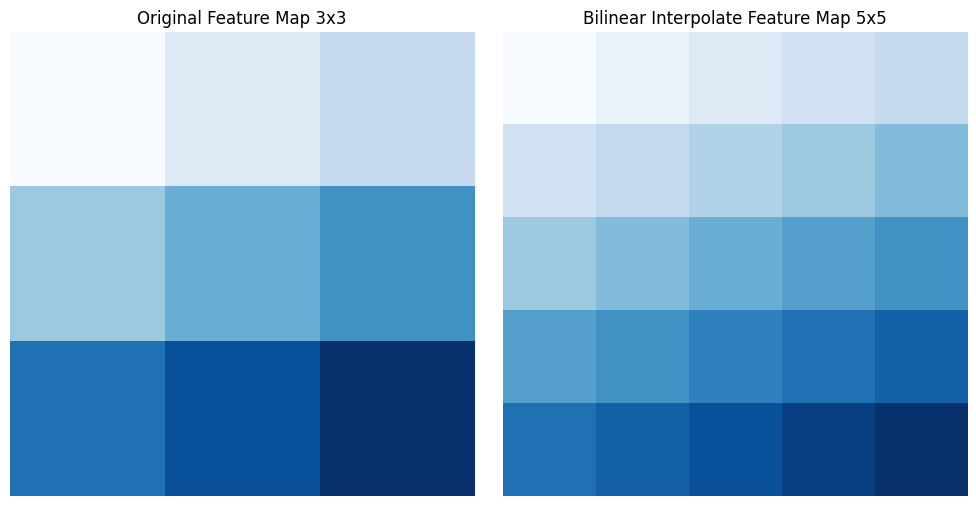


图5: 双线性插值示例

每个上采样像素点都会根据原始特征图像素点进行加权计算，不仅保留了原始特征图的结构比例，同时使得像素点之间过渡较为平滑，一定程度上能够反映不同尺度对于图像特征点的影响，比如某个笔画的特征将被放大。

在每个尺度特征图经过双线性插值地上采样后，将根据三种方式进行融合：累加、加权平均、通道拼接。第一种方式是传统的FPN-Style做法，直接对降维后的多尺度特征图直接进行累加运算，类似上述提到的残差链接。第二种方式的设计思路是为了能够保证不同场景的签名图像都能够拥有一定的学习能力，通过设定权重比例能够更注重某个尺度的特征，默认是采取的全部相等的权重。第三种方式是在通道维度上进行拼接，紧接着一个卷积核的卷积2D运算，这种方法能够最大程度保留所有尺度的特征，但是可能会导致模型计算开销过大，导致模型收敛速度变慢。

融合多尺度特征图后，最后将经过一个卷积核的卷积2D层以平滑上采样过程中带来的checkerboard artifacts，最终输出多尺度融合特征图。

**3.4. Holistic Encoder**

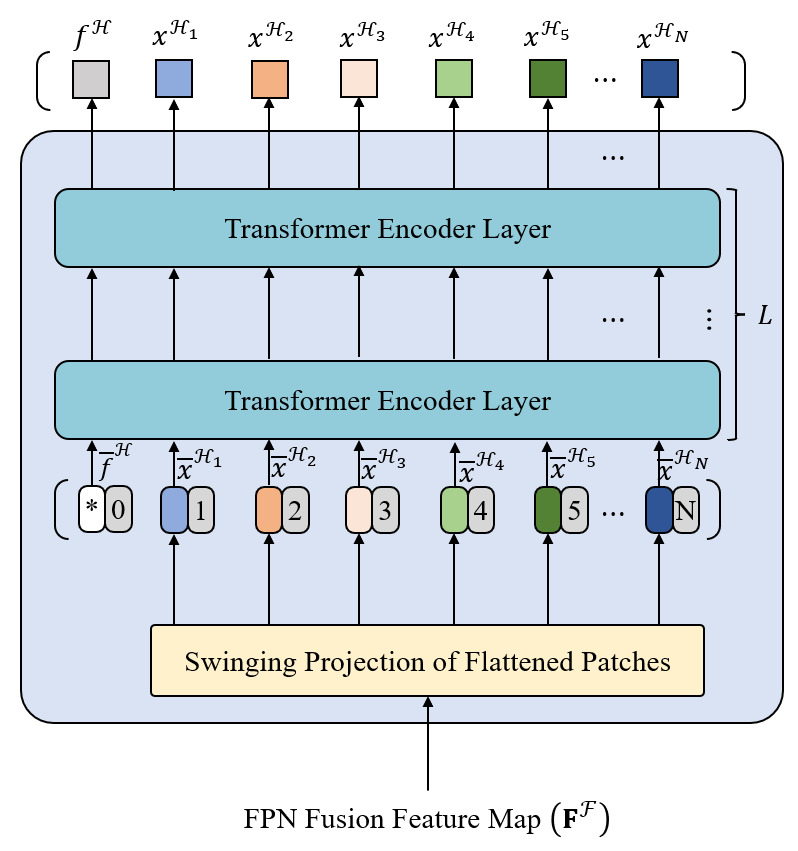


图6: Holistic Encoder结构

Holistic Encoder是参考了Vision Transformer (ViT) **[?]**设计的架构如图6。在输入阶段采取了swing embeddings**[?]**的方法代替之前的分块操作，根据设定的窗口大小和步幅，在输入的多尺度融合特征图上进行分块和映射，得到patch embeddings后进行平坦化以作为特征向量，其中表示映射特征维度数，向量长度计算如下：

其中，表示floor function。在swing embeddings中拼接一个定义为class token的可学习权重，即输入Transformer Encoder Layer的特征向量为。由于对特征图进行了平坦化运算，因此需要在特征向量上累加positional embeddings**[?]**，以此使特征向量包含其在多尺度融合特征图上的位置信息。

经过swing embeddings后，特征向量进入Transformer Encoder Layer以进行注意力特征计算。其中Transformer Encoder Layer由个堆叠形成，如图4所示。

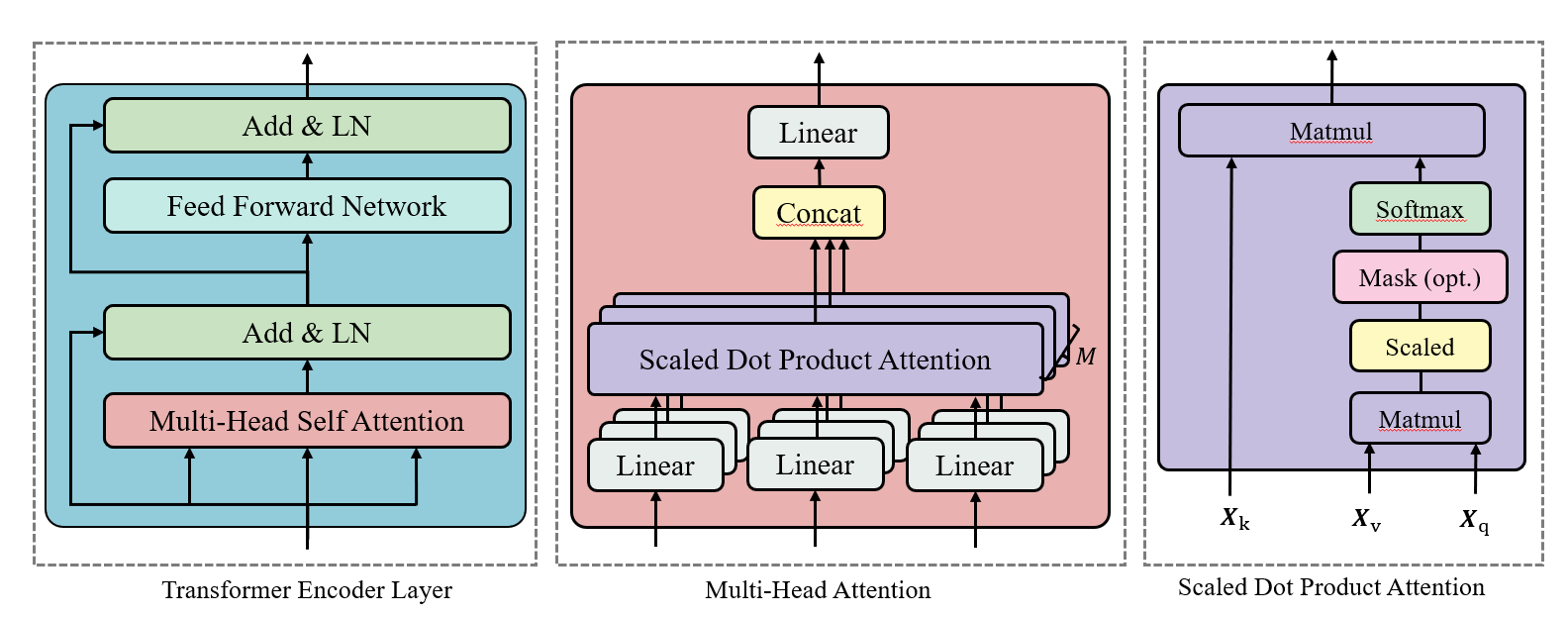


图7: Transformer Encoder Layer结构

每个Transformer Encoder Layer由一个Multi-Head Self AttentionAdd & LN和Feed Forward NetworkAdd & LN组成，其中Add&LN 表示残差链接和LayerNormalization (LN) **[?]**。输入的特征向量首先进入Multi-Head Self Attention (MHSA)以进行注意力机制计算，MHSA本质上是Multi-Head Attention (MHA)，只是在MHA运算中和映射前的来源相同，即输入特征向量将同时进入每个头的线性映射层进行降维，从而生成查询向量、键向量、值向量，因此。其中注意力机制分为多个头计算是为了节省计算资源，将特征向量输入个头的三个线性层以降维映射，从而达到并行地注意力计算，最后进行拼接和线性层映射会输入向量相同特征维度，因此MHSA的输出向量形状与输入向量形状相同。在MHSA中每个头的注意力特征向量均采取scaled dot product attention**[?]**计算，公式如下：

其中，、、之间是软查询关系**[?]**，每个查询不是选择一个具体的键-值对，而是对所有的key-value进行加权平均，权重是通过和的相似度计算而来。的每个特征token会对的每个特征token建立一个相关性分布即Softmax的注意力权重，根据这个权重来加权融合的值。该机制的优势在于能同时关注多个相关token，使得模型能更充分建模上下文信息，比如位置的注意力权重，表示在第位置上对第个位置key的权重，这个权重的值越大，说明第位置的query对第位置的key具有很强的上下文关系，反之越小则说明关系很小。根据这个注意力权重，再次乘会将自身向量注意力权重高的地方特征将进一步放大，并且token之间将包含注意力信息，从而达到每个token拥有自身全局的信息，能够更加有效地利用重要特征，忽视掉无用特征，在手写签名图像中将更有效地关注笔画部分，忽视空白部分。由此单个Transformer Encoder Layer的MHSA输出如下。

其中，表示 MHSA输入阶段的线性映射层权重，表示MHSA中最后线性映射层权重，表示每个头、、的特征维度数，因此必须满足。

在MHSA部分进行了维数较大的矩阵乘法，为了减少梯度爆炸或消失风险，同时为了减少原始特征向量信息缺失，因此在MHSA后紧接着一个残差链接和LN层。LN层的主要作用是归一化每个位置的特征向量，从而加快模型收敛的同时提高模型稳定性。与BN的均值和方差不同，LN的均值和方差来源于当前特征维度上，计算如下：

其中，表示MHSA输出特征向量的第个特征维度。后续的归一化运算与BN层一样，整个MHSAAdd & LN输出的计算公式如下：

其中，表示上一个Transformer Encoder Layer的输出，若，则。在经过注意力计算后，引入Feed Forward Network (FFN)增强整体模型对复杂模式的建模能力，整个FFN计算如下：

其中，表示FFN中第一个线性层的权重和偏置，表示FFN中第二个先行层的权重和偏置。GeLU**[?]**函数和ReLU函数作用相似，均是进行非线性转换的激活函数。整个FFN就是对模型进行一个非线性转换，与Attention的全局交互机制形成互补，联合构建强大的上下文感知表示。随后紧接着Add & LN补充特征信息和加快模型收敛，后续每一层的Transformer Encoder Layer均为堆叠效果，加深特征向量的注意力权重，整个Holistic Encoder的输出为。

**3.5. Conv-Module**

图像特征经过Holistic Encoder捕捉序列中元素的相对位置信息进行特征学习，但其存在绝对位置信息的学习局限性。签名验证任务需要精确到笔画位置信息（如笔顺、局部结构差异），因此在这之后为了加强部分小部分图像特征，将Holistic Encoder的patch token输出进行rearrange，对平坦化的token重新转换成2D形状以进行卷积运算，其中。本模块引入卷积操作对Transformer编码器输出的patch特征进行再处理，加强位置感知能力。在Conv-Module架构上，参考了TransOSV的Conv设计**[?]**，根据以往CNN的样式，调整了中间卷积2D层的输出通道数增加了下采样运算，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 |  |  |

与传统CNN的通道数逐渐变多不同，这里采取的是整体下采样运算思维，对patch特征进行一定的降维运算，在整合局部细节特征的基础上抽象表达特征图的局部特征，目的是为了提升模型的泛化能力和鲁棒性。在backbone和FPN Fusion的基础上增加该操作，进一步强化模型对图像局部特征的感知能力，强化对签名关键局部区域的关注，同时过滤背景干扰。在Conv-Module模块中，每个Conv2D层运算与backbone中的Conv2D层运算相同，均采取Conv2D+ReLU的组合，并且经过两个Conv2D+ReLU后会进行一个下采样运算，即Max Pooling 2D运算，对特征图进行整合和剔除非重要部分特征，加强特征表达性。

在模型训练过程，Conv-Module的输出在进入Contrast based Part Decoder的同时，将经过一个Global Average Pooling (GAP)运算，以此对Conv-Module输出的特征图进行全局取样，从而得到一个平坦的局部卷积特征，以完成训练和推理过程中对应的特征提取工作。

**3.6. Contrast based Part Decoder**

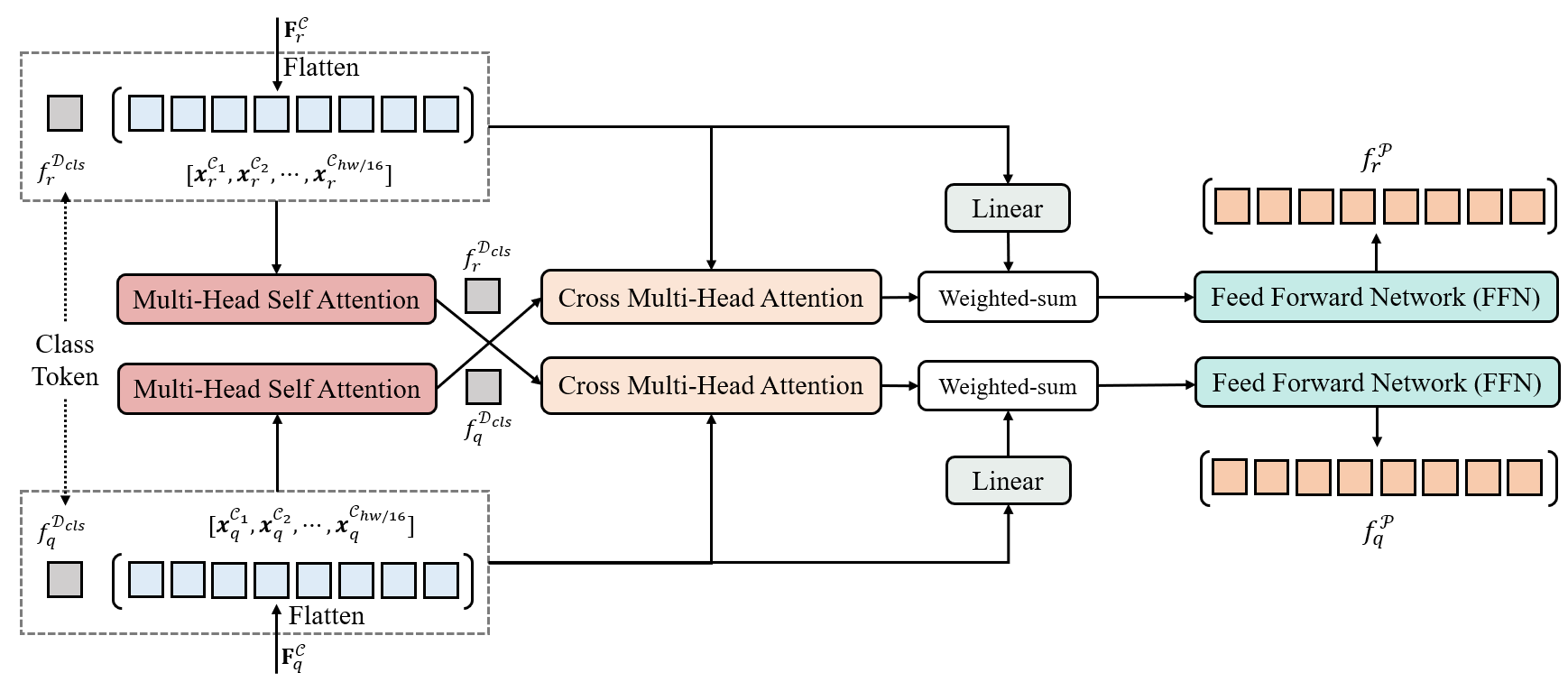
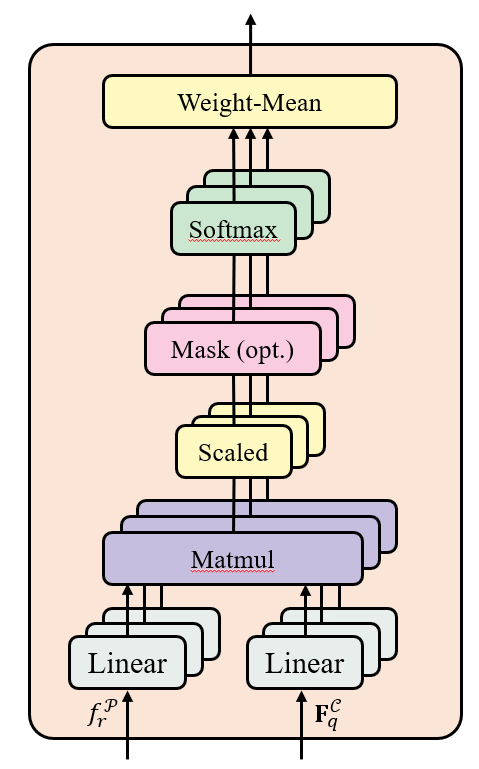
 

图8：Contrast based Part Decoder & Cross Multi-Head Attention

该部分是在TransOSV的Contrast based Part Decoder**[?]**基础上，对Cross-Attention进行了并行计算的改进，引用了Holistic Encoder中MHSA的多头注意力计算机制。与Holistic Encoder部分相似的是，在输入阶段会对reference和query的特征图映射到特征维度上，随后进行平坦化处理和增加一个可学习权重，同样定义为class token。对平坦化的一对Conv-Module特征图进行一个MHSA运算，随后将上述生成的reference和query的class token作为query矩阵输入Cross Multi-Head Attention（CMHA）。CMHA的结构如图8-b所示，与MHA的区别在于只生成query矩阵和key矩阵计算其交叉注意力权重，且最后的concat替换成加权平均以处理多个头的注意力权重。对于单个CMHA的输出交叉注意力权重计算公式如下：

其中，表示输入的query向量和key向量，表示CMHA中输入部分的线性映射层权重。由此计算过程如下：

对注意力权重和经过映射的平坦特征向量进行weight-sum运算，融合交叉注意力权重的平坦特征为后经过FFN以得到decoder的交叉注意力平坦特征，整个计算过程总结如下：

这种类似与Transformer Decoder**[?]**结构的Contrast based Part Decoder，有利于生成区分真实签名和伪造签名的局部特征信息，并且强调了一对数据样本之间的特征关系，加强模型对样本中一对数据之间的特征学习。在训练过程中，将输出的交叉注意力权重进行稀疏性损失计算，迫使交叉注意力权重分布集中，避免平均化**[?]**。

**3.7. Loss Function**

**3.7.1. Euclidean Distance**

OSVTF是一种类似孪生网络的模型架构，这种风格强调模型是共享权重的双流向形式，整体是为了对比两者之间的差距，因此将基于输入两者特征的相似度以设计损失函数。本工作将采取Euclidean Distance**[?]**以作为一对特征样本的相似度计算，假设输入一对特征，其两者之间距离计算公式如下。

**3.7.2. Focal Contrast Loss**

在手写签名样本数据处理阶段，会将作者的手写签名进行两者之间不会重复的配对操作，例如BHSig-B**[?]**数据集中一共包含了100位作者的手写签名图像，其中每位作者拥有24张真实签名图像（）和30张伪造签名图像（，真实签名图像作为reference签名，伪造签名和真实签名一起作为query签名，由此将生成positive样本对276个和negative样本对720个。因此每位作者一共拥有996个数据样本对，其中positive样本对和negative样本对之间的比例不接近50:50，因此positive样本和negative样本数量不均衡，属于hard samples范畴**[?]**。

早期算法、模型中用于二分类的对比样本相似度损失函数是Contrast Loss**[?]**，计算公式如下。

其中，表示边界值，即forged样本应至少保持的距离。对于的样本对，则直接作为损失值，使模型参数学习上对真实样本对的特征更相似；相反时，若reference和query的特征距离小于，则使用作为损失值，使模型对于伪造签名对的特征不小于。但是这种损失无法动态强调hard samples，容易出现模型参数过拟合问题。为了减少过拟合的风险，对genuine样本对同样引入一个边界值，减少genuine对之间不必要的收缩，即Double-Margin损失计算公式如下。

其中，两个边界值，对于genuine样本对，惩罚其距离超过的情况；对于forged样本对，促进距离大于。虽然double margin一定程度上优化了contrast loss容易过拟合的问题，但是在两对样本数据且的情况下，通过OSVTF计算两个样本对的距离，若，那么损失函数应该给予更大的损失，但是会平等地对待这两个样本对，在TransOSV中引入了动态权重的概念，命名为Focal Contrast Loss**[?]**，计算公式如下。

其中，表示Sigmoid激活函数，用于动态生成权重；为两个边距值；表示缩放因子，用于调节hard samples响应强度的缩放系数。当样本对的标签为genuine时，如果特征距离越大于，则当前权重值越大；当标签为forged时，如果特征距离越小于，则权重越大。由此将同时拥有动态调整训练样本的关注度、重点优化“容易混淆”的签名和提升模型泛化能力的优点，将作为OSVTF模型的主要训练损失函数之一。根据各个模块提取的平坦特征，将采取四个部分以计算focal contrast loss，整体Focal Contrast Loss计算如下。

其中，为训练超参数，用于分配各个特征的比重。

**3.7.3. Sparsity Loss**

在Contrast based Part Decoder中，对于交叉注意力计算时会出现注意力权重过于均匀分布的情况，实际上应该增强模型关注最显著的局部差异区域，由此采取交叉熵的方式以生成对比掩码的熵约束，希望交叉注意力权重能够分布稀疏，从而提升Decoder的local attention精准性**[?]**，计算公式如下。

由于该部分并不是作为主要模型训练，因此引入作为该部分的超参数参与OSVTF训练中。综上所述，OSVTF的总体损失函数构成如下：

**4. Experiments**

**4.1. Experimental Setup**

本工作中将基于basic ViT**[?]**configuration以build OSVTF。The FPN Fusion的输出特征图特征维度与第一层尺度特征卷积图通道数相同。The encoder包含8个transformer encoder layer，每个encoder layer中MHSA包含12个头，768个特征，尺寸为步幅为的滑动窗口。Decoder部分中，MHSA和CMHA的特征维度数、头的数量与encoder的各部分保持一致，在CMHA后weighted-sum的对patch embeddings映射的特征维度数设置为512。

在损失函数上，沿用TransOSV的部分设置和。The hyper-parameters设置为。在优化器上，所有实验均采取momentum为0.9、权重衰减为1e-4的SGD Optimizer。全部实验均在2张NVIDIA 3080 GPUs上进行并且设置总的batch size为16。

**4.2. Metrics**

在评价指标上，不同于常规的二分类任务，离线手写签名验证任务需要对negative样本进行特别关注，因此在分类准确率(ACC)的基础上，使用生物识别和签名验证任务中的标准指标**[?]**来评估OSVTF的性能，即False Rejection Rate (FRR)、False Acceptance Rate (FAR)和Equal Error Rate (EER)。在常规的二分类任务中，会使用混淆矩阵**[?]**的概念以统计样本标签和预测标签的数量，其格式如表2。

表2：离线签名验证混淆矩阵概念表

|  |  |  |
| --- | --- | --- |
| 预测标签/原始标签 | (Positive) | (Negative) |
|  | True Positive (TP) | False Negative (FN) |
|  | False Positive (FP) | True Negative (TN) |

由此可得ACC、FRR、FAR的计算公式如下：

EER的计算则是使用若干个阈值和距离进行对比，若样本对的特征距离大于阈值，则判断为forged，否则为genuine。本工作在训练过程中将根据批次数据样本对的模型特征计算距离，根据样本对的距离最小值和最大值以设置若干个阈值，从而自动计算当前批次下每个阈值下的FAR和FRR。对每个阈值的FRR和FAR进行计算，当FAR和FRR之间的差异最小时，则同时输出当前阈值的FRR、FAR，并且计算EER如下：

由此，实验阶段将基于ACC、FRR、FAR、EER四个评价指标来综合评估模型性能。

**4.3. Dataset**

**4.3.1. BHSig-B & BHSig-H**

这两者数据集均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。

**4.3.2. 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。

**4.4. Results Analysis**

**4.4.1 Writer-Independent Results**

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 | 9.90 | 9.90 | 9.90 |
| TransOSV[Ours] | 11.26 | 11.26 | 11.26 |
| **OSVTF [Ours]** | **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[Ours] | 7.89 | 7.89 | 7.89 |
| **OSVTF [Ours]** | **10.41** | **10.41** | **10.41** |

**4.4.2 Writer-Dependent Results**

**References (参考文献)**