尊敬的教授们，早上好。我是2023年9月份入学的智能技术硕士学生王茗琛。我的导师是刘新教授。今天介绍的主题是Research on the structure of offline handwritten signature verification models based on transformer。该主题是2023年实习中参加的一个与珠海金智维信息科技有限公司合作项目，以此进行展开后续的研究。接下来将分为五个部分进行赘述。

首先，离线手写签名是生物特征技术中的一个具体应用。生物特征技术是一种根据指纹、面部、虹膜等个体生理特征或声纹、笔迹等行为特征进行个人身份识别或验证的技术。该技术被广泛应用于企业安全认证、金融交易、门禁系统等安全领域。例如日常生活中手机的解锁、银行卡转账等都会涉及此类的身份验证技术。本研究课题主要是根据包含个人笔迹行为特征的手写签名以展开研究。在了解手写签名之前需要先了解一下识别或验证的区别。生物特征技术主要用于识别和验证，识别是用户向系统提供个人生理或行为特征，系统将根据这些特征来判断用户是系统中登记的哪一个用户。即解决“你是谁”的问题给。而验证是在识别的基础上需要用户声明身份，系统将根据这些特征来判断用户是否属实，即解决“你是不是xxx”的问题。因此手写签名验证是通过个人书写手写签名的笔迹来判断提供的信息是否属实。在手写签名验证任务里面，将根据签名设备分为离线和在线。在线是指用户使用数字化设备书写个人签名，如银行前台所使用的数字版，日常生活学习中平板电脑和电子笔。离线是指用户仅通过笔和纸进行书写。前者在采集数据上需要严格要求数字化设备的型号、批次，并且对使用者握笔姿势、与写字板接触的角度、按压力度有很大关系，不同的个人书写习惯和设备将导致收集的手写签名样本需要额外花费时间进行数据预处理。后者在采集数据过程仅与使用者的笔相关，并不会受到其他影响，并且握笔在纸上书写个人签名的过程是不可入侵的，即中途不会因为网络或设备导致书写字迹不一致、断断续续等问题。因此学者们更多地是研究离线手写签名验证。在以往的离线手写签名验证模型或算法中，学者们大多数是采取传统机器学习的方式以完成任务。但是传统机器学习方法有个缺陷，即很大程度依赖于人工设计的特征，即图像如何转换成特征向量这个过程。该过程需要大量经验的图像处理经验才能够提取出较优质的手写签名验证。且离线手写签名验证任务本质上是一个多任务。它一共包含两个子任务:Writer-Dependent和Writer-Independent。前者定义为WD，后者定义为WI。在WD任务中，需要根据提供的参考签名，这里定义为reference签名；和提供的签名进行比对，这里定义为Query签名；从而判断提供的签名是真实的还是赝品。而WI任务中，不需要提供Reference签名，仅通过提供的query签名进行鉴别是否为赝品。因此基于上述离线手写签名采取传统机器学习方法在人工设计特征和多任务中将花费巨大的成本。

接下来将进行第二部分，阐述研究目标和研究意义。基于上述研究背景，对于离线手写签名验证模型的研究，能够在一定程度上学习深度学习如何与该方向进行创新和改进。该页右上角展示的Offline Signature Verification Transformer，缩写OSVTF即本文拟提出的基于Transformer的模型结构。在研究方面一共分为一下三个阶段：阶段1，复现目前学术界上比较新颖的深度学习离线手写签名验证模型，分析和总结它们之间的优缺点，从而总结出一套适用于OSVTF的设计、改良方案；阶段2，根据上述复现的结论，从而提出一个基于Transformer的离线手写验证模型，该模型能够结合卷积神经网络局部特征学习能力和Transformer全局特征学习能力的OSV模型结构；阶段3，对于WI和WD的多子任务，改良现用的传统机器学习方法，从而能够更优质且高效地鉴别手写签名是否为赝品。综上所述，研究该主题主要有以下四个意义：第一，提出一种更高准确率的OSV模型，一定程度上能够更清晰深度学习方法的优势，从而更好展开后续更新模型或算法的研究。第二，采取深度学习方法能够节约大部分人工设计特征的成本，能够拥有更多成本研究优质高效的模型结构或算法。第三，更少的推理成本，当模型参数量相差不大时，能够进行GPU硬件运算，从而节约更多资源出来进行平行运算其他任务，以达到更高效地完成各项任务。第四，基于手写签名图像的图像特征研究，一定程度上能够影响到OCR的相关研究工作，为OCR领域带来一定程度的方法创新。

第三部分将针对离线手写签名和Transformer的深度学习方面进行时间线相关的文献综述。首先离线手写签名验证在上个世纪80年代就已经开始研究，这个时间段是基于特征的模板匹配方法。最早于1989年Plamondon等人提出一种基于几何特征的手写签名验证方法。它们在论文中研究了基于几何特征和欧几里得距离的验证方法。这些早期方法关注于签名的整体形状、壁画粗细和斜度等基本特征。1991年Kashi等人在论文中提出了一种基于形状匹配的手写签名验证方法。该方法主要通过比较签名的几何特征来实现签名的匹配。紧接着1990年初到21世纪初是结构特征与纹理分析方法。1997年Justino等人在论文中引入隐马尔可夫模型用于签名验证，利用签名的统计特征和结构特征进行匹配。次年Sabourin等人提出一种基于小波变换的签名验证方法，通过分解签名图像的频率特征，提高了验证系统对签名细节的感知能力。21世纪到2010年初，学者们均在采取统计模型与机器方法进行离线手写签名验证研究。其中Coetzer、Vielhauer、Ferrer等人均是基于隐马尔可夫、支持向量机等统计机器学习模型进行离线手写签名验证，其中它们的方法无一例外均是使用了各种数据预处理方法来提高验证准确率，如离散Radon变换、离散小波变换与特征融合。直到2012年深度学习方法引入，Hafemann等人首次将CNN引入到离线签名验证任务中，该方法能够自动学习签名的高级特征，不再依赖于手工特征提取。由此学者们提出了基于对比损失的孪生网络，通过优化损失函数，使网络能够更精确地区分真实和伪造签名。2020年，Soleimani等人对各种深度学习模型进行了评估，其中包括CNN、Siamese Network和GAN，同时还提出了一种结合GAN和孪生网络地模型，用于生成伪造签名样本以增强训练数据集。在最新阶段即2020年到未来，逐渐往自监督学习与多模态融合发展。2021年Kim等人提出了基于自监督学习地方法，使用SimCLR模型进行预训练，以减少对大量标注数据的依赖。2022年Siddiqi等人更是提出了多模态签名验证方法，将离线和在线签名数据结合，提高了模型的鲁棒性和适应性。在另一方面深度学习也在同步发展，2012年的Alex等人于ImageNet大赛上提出的AlexNet卷积神经网络，其优质的性能使得学术界掀起一阵卷积神经网络的深度学习狂潮，越来越多的深度卷积神经网络被提出，如2016年He等人提出的残差结构CNN，能够在一定程度上保留图像的部分特征。不仅图像领域，NLP领域也在飞速发展。2017年Vaswani等人提出用于机器翻译的Transformer模型，该模型是基于多头注意力机制的Encoder-Decoder结构。该结构没有采取任何的卷积、池化操作，仅仅通过各种线性层和线性计算以实现机器学习翻译任务。其中它的多头注意力机制相较CNN的卷积运算，能够更佳注重特征向量的全局特征，从而能够在长短句不一的翻译任务上取得优质的准确率和鲁棒性。由此在CV领域Dosovitskiy等人于2021年提出Vision Transformer，简称ViT。ViT的原理是将图像进行patch和flatten，将每一个patch当作词向量的单个字符向量，整张图片作为词向量以进入Transformer Encoder。随后在后续增加一个MLP以完成ImageNet的图像分类任务。，ViT的实验证明了更佳注重全局上下文特征的模型能够在训练收敛的情况下达到更好的效果，而CNN能够在一定时间的训练下达到较好得到效果。但是基于Transformer的ViT模型在推理运算时间不如CNN快，后续深度学习发展随着硬件迭代升级Transformer的低效推理成本也随之忽视，并且更多的基于Transformer的视觉深度学习模型陆续提出，如用于目标检测的DETR和Deformable DETR，前者和ViT相似，但是DETR采取了完整的Transformer Encoder-Decoder结构，在图像经过CNN提取多通道特征图后进行Transformer推理，最终经过MLP以推理物体在图像的相对位置坐标及其类别。Deformable DETR则是在DETR的基础上增加了多尺度特征，采取了FPN-Style将CNN为backbone的推理过程中多个卷积块得到的不同尺寸特征图全部输入至Transformer Encoder中。但是得到多个尺寸的特征图将导致Transformer Encoder内部计算成本增加，因此作者在论文中提出了可变注意力机制。在多头注意力机制的基础上，采取K近邻采样，优化了Transformer Encoder多头注意力计算中整个特征向量的注意力计算，针对部分特别关注的点进行注意力计算，并以此进行部分比较关注的特征注意力计算。在论文实验中利用更短时间达到DETR及其他非端到端目标检测模型的性能。因此Deformable DETR的多尺度端到端目标检测结构将成为重要参考之一。

第四部分将介绍研究课题中OSVTF模型的基本框架，并详细介绍各个部分及数据集来源。首先是OSVTF的整体架构，这里展示的是训练推理过程。首先输入一对手写签名图像，经过共享权重的backbone和encoder，得到一对平坦token和特征。其中这对token会进行FC损失计算，特征将rearrange后进入卷积模块。经过卷积模块后将进行两个分支推理：其中一个进行全局平均池化后得到卷积平坦特征以进行FC损失计算；另一个分支则直接进入decoder计算，得到平坦decode特征，以此进行FC损失计算。即训练过程整个模型架构将收集平坦encode、卷积、decode特征以进行FC损失计算。OSVTF中的backbone部分是CNN的主干网络部分，使用卷积部分进行提取多通道特征图，这里以ResNet-50为例子。Backbone推理后接着是Encoder，在TransOSV模型结构中是直接采取了ViT的Encoder架构，将多通道特征图patch后进行flatten以得到与词向量形状相同的2D特征向量，以此进行后续的多头注意力计算。多头注意力中是采取了Scaled Dot Product Attention来计算单个头的注意力权重。在ViT Encoder中会随机生成一个与词向量形状相同的参数，以此累加在特征向量中进行注意力权重计算。这一步的目的是为了能够一个可学习参数来加快模型收敛速度。在后续实验将backbone的多个尺寸特征图拼接以进行多尺度的多头注意力计算。在decoder中是直接沿用了TransOSV的Contrast based Part Decoder。采取了cross attention的decoder与传统Transformer decoder不同，这种decoder在输入特征基础上与ViT相同新增了一个可学习参数，以此来更好地关注reference和query的对比关系。Decoder内部首先会经过一个多头注意力，先对输入特征向量进行注意力权重分配，以此来强化明显特征，弱化比较微弱的特征。随后根据上述新增的科学系参数进行交叉注意力计算，以此将reference和query的特征进行交叉对比，从而得到最终的decode特征。在模型训练的损失函数处，一共采取了两种损失函数：Sparsity loss和focal contrast loss。Sparsity loss是仅针对decoder

中线性映射提出的，能够加快decoder的收敛速度。而focal contrast loss是基于评价两个对象差异的contrastive loss和CaP中双倍边际损失改良的用于一对手写签名特征的训练损失函数。其中D()是表示计算样本之间距离，这里采取的是欧氏距离。Alpha 1和alpha2表示两个边距值，overline K和overline V表示缩放因子，上述均为训练过程超参数。数据集部分是采用了离线手写签名验证的公开数据集BHSig-B、BHSig-H和CEDAR。其中BHSig的B和H表示书写语言Bengali和Hindi，BHSig系列数据集是由印度理工学院古瓦哈提分校发布的100个孟加拉语用户和160个印地语用户手写签名数据集，这两个数据集每个用户均提供了24张真实签名和30张伪造签名。CEDAR数据集是由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。

最后第五部分将展示现有成果和预期成果，以及遇到的一些问题。首先是对部分OSV模型的复现，着重复现了基于Transformer的TransOSV，由于训练设备的差异，原作者是采取了8张NVIDIA 3090显卡在batch size为64的情况下训练，复现过程是采取了1张4070S进行复现和训练，一定数量epoch训练下BHSig-B和BHSig-H的数据集表现上差距大概在20%到30%左右。其中采取的是和原文提供的相同学习率、衰减率的SGD优化器，调整优化器方案应该能缩小差距在10%左右。后续将着重针对多尺度和encoder部分进行模型架构优化。其次，该研究课题是源于实习所在公司与本科学院的相关合作项目，在2021年末首次提交了广东省企业科技特派员研发项目立项申请书，后续于2022年加入项目组并进行中文手写签名数据采集和模型部署的工作。在2023年后续申请成功后估算累计年新增产值300万元左右，并且与该公司其他AI产品结合形成基于业务场景的自动审核解决方案，将有更广阔的经济前景。最后，在2023年3月中旬左右投稿了ICNCC-2023，并于ACM中收录相关文章：Offline Signature Verification Using a 2D Attention Encoder-Decoder Network。目前研究过程中大概有三个问题：第一个issue是针对图像多尺度问题，对于部分手写签名图像宽度和高度少于ImageNet图像分类任务的224像素，采取类似Deformable DETR的多尺度方法能否得到更有效的模型质量提升，且目前基于ViT的encoder必须规定统一尺寸大小的图像输入，因此多尺度图像特征是后续模型结构优化的主要方向。第二个issue是在encoder中采取的是flatten特征以进行多头注意力计算，能否在找到方案来弥补图像特征拉平后进行解码过程中的部分特征丢失问题。第三个issue是训练环境问题，单凭一张16G的NVIDIA 4070S显卡难以达到深度学习模型训练环境，后续需要租赁服务器以进行模型训练，从而得到更优质的模型性能。最后对于期望结果，希望能够在采取了CNN+Transformer的方式后能够得到一个优于TransOSV的模型性能，且拥有更稳定的泛化性，期望作用于中文手写签名上。对于优化后的模型架构，在特征提取部分希望能够给予OCR领域相关的图像研究一些思路，从而拓宽设计OCR模型架构的思路。同时也希望在合作项目中能够落地实现OSVTF的部署，结合其他AI产品从而更稳定地实现对应业务场景地自动审核解决方案。

以上便是我的开题报告，感谢各位教授的倾听。