计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(9):89−96 [doi: 10.15888/j.cnki.csa.009205]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于自注意力机制的亲属关系识别①

李德财[1,2], 蒋行国[1,2], 何　李[1,2], 李嘉莉[1,2]

[1](四川轻化工大学 自动化与信息工程学院, 宜宾 643002)
[2](人工智能四川省重点实验室, 宜宾 644002)
通信作者: 蒋行国, E-mail: tonny_jiang@suse.edu.cn

摘　要: 目前, 基于局部注意力机制的卷积神经网络 (CNNs) 用于亲属关系识别特征提取获得了不错的效果, 但基于卷积神经网络的主干模型提升不明显, 同时鲜有研究者使用具有全局信息捕获能力的自注意机制. 因此, 提出使用基于一种无卷积主干特征提取网络的 S-ViT 模型, 即用具有自全局注意力机制的 Vision Transformer 作为基础主干特征提取网络, 通过构建孪生网络与具有局部注意力机制的 CNN 相结合, 扩大传统分类网络, 用于亲属关系识别相关问题的研究. 最终实验结果表明, 相比 RFIW2020 挑战赛领先的方法, 所提出的方法在亲属关系识别 3 个任务上获得了良好的效果, 第 1 个任务中获得了 76.8% 验证精度排名第二, 第 2 个和第 3 个任务中排名第三, 证明了该方法的可行性和有效性, 为亲属关系识别提出了一种新的解决方法.

关键词: 亲属关系识别; 自注意机制; Vision Transformer; 卷积神经网络; 深度学习

引用格式: 李德财,蒋行国,何李,李嘉莉.基于自注意力机制的亲属关系识别.计算机系统应用,2023,32(9):89–96. http://www.c-s-a.org.cn/1003-3254/9205.html

Kinship Recognition Based on Self-attention Mechanism

LI De-Cai[1,2], JIANG Xing-Guo[1,2], HE Li[1,2], LI Jia-Li[1,2]

[1](School of Automation and Information Engineering, Sichuan University of Science & Engineering, Yibin 643002, China)
[2](Sichuan Key Laboratory of Artificial Intelligence, Yibin 644002, China)

Abstract: At present, convolutional neural networks (CNNs) based on local attention mechanism have yielded sound results in feature extraction of kinship recognition. However, the improvement of backbone models based on CNNs is not obvious, and few researchers employ self-attention mechanisms with global information capture ability. Therefore, an S-ViT model based on a convolution-free backbone feature extraction network is proposed, which is to adopt Vision Transformer with a self-global attention mechanism as the basic backbone feature extraction network. By constructing a twin network and a CNN with a local attention mechanism, the traditional classification network is expanded for research on related issues of kinship recognition. The final experimental results show that compared with the leading method of the RFIW2020 Challenge, the proposed method has performed well in the three kinship recognition tasks. The first task ranks second with verification accuracy of 76.8%, and the second and third tasks rank third. As a result, the feasibility and effectiveness of the method are improved to propose a new solution to kinship recognition.

Key words: kinship recognition; self-attention mechanism; Vision Transformer; convolutional neural network (CNN); deep learning

1　引言

亲属关系识别是计算机视觉领域的一个新研究课题, 对于许多现实的应用 (如亲属关系验证、自动照片管理、社交媒体应用等) 至关重要. 目前, 亲属关系识

① 基金项目: 四川理工学院科研基金 (2019RC12); 四川省人工智能重点实验室开放基金 (2020RZJ03)
收稿时间: 2022-12-21; 修改时间: 2023-01-09, 2023-03-02; 采用时间: 2023-03-22; csa 在线出版时间: 2023-07-14
CNKI 网络首发时间: 2023-07-17

别可以分为 3 个任务.

1) 亲属关系验证: 用于预测给定的两幅人脸图像之间是否具有亲属关系, 属于二分类问题.

2) 三主体验证: 用于确定一个孩子是否与一对父母有关.

3) 查询和检索: 目标是在图库中查找搜索对象的家庭成员, 该任务模拟了失踪人员的查找.

基于深度学习的方法[1], 是当前研究亲属关系识别的热点, 通过自动从图像中学习、理解和构建深度特征, 并对深度特征进行分析, 从而得到亲属关系识别结果. 目前, 卷积神经网络 (convolutional neural network, CNN) 用于特征提取在计算机视觉领域取得了令人瞩目的成功[2]. 基于深度学习的人脸亲属识别方法, 大多数使用具有局部注意力机制的 CNN 作为特征提取主干网络, 围绕着 VGGFace-ResNet50 架构和一些 CNN 网络来构建网络模型. 在亲属关系数据集 FIW[3] 上, 亲属关系验证精度能达到 80%[4], 要在实际生活中使用亲属关系识别算法, 目前的验证精度还有待提高. 具有局部注意力机制的 CNN 虽然能很好地提取人脸图像高频分量 (如边缘, 轮廓), 但对图像低频分量的全局语义信息关注较少. 在 2020 年, 有学者提出 Vision Transformer (ViT)[5] 把自然语言处理领域大热的 Transformer 应用到计算机视觉领域. ViT 主要特点是具有全局注意力机制, 能很好地捕获全局语义信息, 在图像分类任务上取得了很好的效果. 在亲属关系识别上, 卷积神经网络对亲属特征识别精度提升不太明显, 同时, 很少有研究者使用具有全局自注意力机制的 ViT 模型作为主干特征提取网络.

综上所述, 为拓宽用于亲属关系识别组合模型的基础分类网络, 本文提出一种基于无卷积 ViT 基础分类网络的 S-ViT 模型. 与传统使用的 CNN 模型相比, 结合自注意力机制作为特征提取主干网络预测结果的方式不同, S-ViT 模型使用了具有自注意力机制的 ViT 网络, 使得模型可以并行化训练, 而且能够拥有捕获全局信息的能力, 更好的提取亲属人脸图像语义信息. 当与具有局部注意力机制的 CNN组合成一个集成分类网络时, 能够更全面的提取亲属关系特征. 实验结果能够表明这种方法的可行性和有效性, 扩大了亲属关系识别传统基础分类网络的选择范围.

2 具有自注意力机制的 PCPVT 特征提取主干网络

Chu 等人[6] 提出了两种新的 ViT 架构, PCPVT 是其中的一种. PCPVT 特征提取主干网络架构图如图 1 所示, PCPVT 通过将 CPVT[7] 中提出的条件位置编码 (conditional positional encodings, CPE) 取代 PVT[8] 中的位置编码, CPE 通过位置编码生成器 (position encoding generator, PEG) 得到, 把 PEG 模块放在每一个 stage 的第 1 个 encoder.

在 PVT 中, stage i 中的 Transformer 编码器有 L_i 个编码器层, 每个编码器层由注意力层和前馈网络层组成. 传统自注意机制详情如图 2 所示.

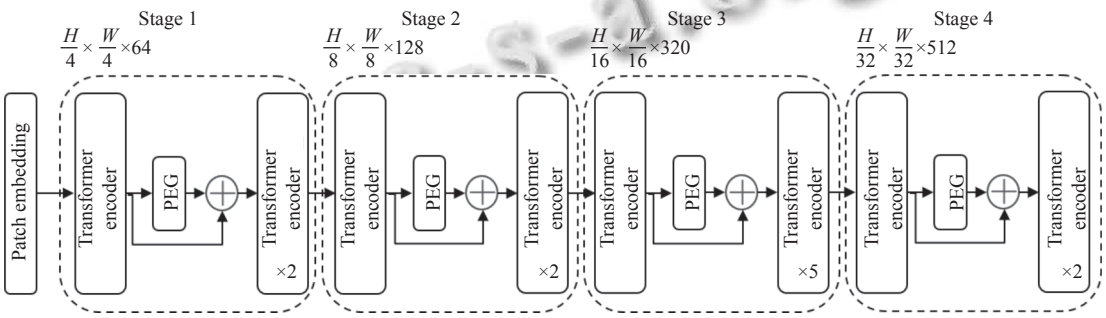


图 1 PCPVT 网络架构图

PVT 因其将传统的多头注意力层[5] 替换成了空间缩减注意力 (spatial-reduce attention, *SRA*) 层而更易于训练. 这个新设计的 *SRA* 层执行类似于多头注意力, 它接收一个查询Q、一个键K和一个值V作为输入, 并输出特性. 然而, *SRA* 在注意力操作之前降低了K和V的空间尺度, 以减少计算开销. *SRA* 的详细表述如下所示:

$$Reduce(x) = Norm\left(Reshape(x, R_i)W^S\right) \tag{1}$$

式 (1) 描述了如何对输入序列进行空间还原. 式中, x表示一个输入序列, R表示缩减比例, W是对输入

序列降维的线性投影, 将输入x调整大小为$\frac{HW}{R^2}\times\left(R^2C\right)$.

$$Attention(Q,K,V)=Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_{\mathrm{head}}}}\right)V \tag{2}$$

$$head_j=Attention(QW^Q,Reduce(K)W^K,Reduce(V)W^V) \tag{3}$$

$$SRA(Q,K,V)=Concat(head_0,\cdots,head_{Ni}) \tag{4}$$

其余的计算与原始的多头注意力相同. 首先利用 *Attention* 计算序列的两个元素与其各自的Q和K之间的两两相似度. W^Q, W^K, W^V分别是Q, K和V的线性投影参数. N是第i阶段注意力层的头数. 然后, 计算出每个头部的注意力分数后串联起来用于最终的 *SRA* 输出. 因此, *SRA* 是一种简单但有效的注意力层, 能够处理高分辨率的特征图, 同时降低计算和内存成本. PVT 的最终输出是一个特征向量, 可以输入到 Siamese 网络中, 用于亲属关系验证的下游任务.

- input $X\in R^{L\times d}$ is a sequence of embeddings of dimension d of length L
- output $Y\in R^{L\times d}$ has the same shape as input
- project X into 3 matrices of the same shape
  - query $X^Q_{m,:}=W^KX$
  - key $X^K_{:}=W^KX$
  - value $X^V_{:}=W^VX$
- calculate "soft sequence-wise nearest neighbor search"
  - "search" all $L\times L$ combinations of sequence elements of X^K and X^Q
  - for each sequence position m: output more of X^V_o the more is X^K_o similar to X^Q_m
  - this is done by weighting the value with a softmax of a dot-product and summing the values
  - in pseudo-code: Y=matmul$_L$($Softmax_L$(matmul$_d$ (X_q, X_k^T)), X_v)
  - in equation: $Y=Softmax(QK^T)V$
- results are added to the residual connection and normalized

图 2 传统自注意机制详情

另外在最后一个 stage 用全局平均池化 (GAP) 代替一般 Transformer 使用的 class token. 由于条件位置编码 CPE 支持输入可变长度, 使得 PCPVT 能够灵活处理来自不同空间尺度的特征. PCPVT 所有层都利用具有全局信息捕获能力的自注意力机制, 同时依靠空间缩减降低处理整个序列的计算成本. 最终, PCPVT 输出一个特征向量送入孪生神经网络, 用于亲属关系识别的后续任务.

3 基于自注意力机制的 S-ViT 亲属关系识别方法

3.1 S-ViT 模型

如图 3 所示, 本文提出一个基于具有自注意力机制的 PCPVT 孪生结构网络模型, 为便于叙述, 简称为 S-ViT. 本文提出使用孪生神经网络来构建模型. Bromley 等人[9] 在 1994 年首次引入了孪生神经网络, 对于把特征比较应用于更复杂的数据样本, 并且特征具有不同的维度和类型, 可能需要在处理之前进行压缩, 那么这些度量是不合适的. 在这些情况下, 孪生神经网络可能是最好的选择: 它由两个相同的神经网络组成, 每个神经网络都能够学习输入向量的隐藏表示. 这两个神经网络都是前馈感知器, 在训练过程中采用误差反向传播; 它们并行串联工作, 并在最后比较它们的输出. 孪生神经网络执行生成的输出可以被认为是两个输入向量的投影表示之间的语义相似性. 在图 3 中, 首先使用 PCPVT 特征提取孪生主干网络为每张人脸图像提取特征, 图像被编码为固定长度的向量, PCPVT 参数设置如表 1 所示. 然后, 将转换后的特征连接起来并将其送入全连接层.

在特征连接阶段, 需要对特征进行相应转换, 以提

高辨别力.

1) 使用特征融合提高网络的非线性能力. 特征融合的本质是对两个输入的人脸特征进行抽象编码, 有利于全连接网络学习更好的相似度指标. 设x和y表示提取的两个人脸的特征, 特征融合方式有$x+y$、$x-y$、$x\cdot y$、$1/2(x+y)$这 4 种, 而 S-ViT 主干特征提取网络输入为 224×224 大小的一对人脸图像, 对输出两组特征向量通过 feature fusion 模块进行融合, 得到的 4 组融合特征.

2) 将 4 组融合特征拼接成一个长向量送入全连接网络 (full connection, FC), 该 FC 由 3 个全连接层、两个 ReLU 激活函数和一个 Sigmoid 激活层构成. 为了减少 FC 层计算量, S-ViT 网络模型对第 1 层使用全局平局池化 (global max pooling, GMP) 得到第 2 层全连接网络, 然后再通过全连接得到 32 个神经元的第 3 层网络.

3) 通过 Sigmoid 函数激活得到相似度评分, 以判断人脸图像对是否有亲属关系.

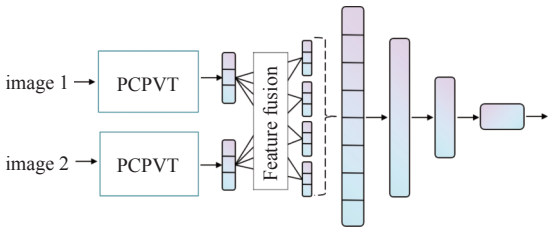


图 3 S-ViT 模型架构

表 1 PCPVT 特征提取主干网络配置细节

Stage	Output size	Layer name	PCPVT
Stage 1	$\frac{H}{4}\times\frac{W}{4}$	Patch embedding	P_1=4; C_1=64
		Transformer encoder with PEG	$\begin{bmatrix} R_1=8 \\ N_1=1 \\ E_1=8 \end{bmatrix}\times 3$
Stage 2	$\frac{H}{8}\times\frac{W}{8}$	Patch embedding	P_2=2; C_2=128
		Transformer encoder with PEG	$\begin{bmatrix} R_2=4 \\ N_2=2 \\ E_2=8 \end{bmatrix}\times 3$
Stage 3	$\frac{H}{16}\times\frac{W}{16}$	Patch embedding	P_3=2; C_3=320
		Transformer encoder with PEG	$\begin{bmatrix} R_3=2 \\ N_3=5 \\ E_3=4 \end{bmatrix}\times 6$
Stage 4	$\frac{H}{32}\times\frac{W}{32}$	Patch embedding	P_4=2; C_4=512
		Transformer encoder with PEG	$\begin{bmatrix} R_4=1 \\ N_4=8 \\ E_4=4 \end{bmatrix}\times 3$

为使 S-ViT 网络模型获得更好的性能, 采用在 ImageNet 数据集[10] 上经过预训练的 PCPVT 特征提取主干网络来构建孪生网络, ImageNet 数据集包含 1 000 个类别, 14 197 122 张图片. 数据依赖是深度学习中最严重的问题之一[11], 因为它需要大量的数据来理解数据的潜在模式. 迁移学习 (预训练模型) 不需要训练数据和测试数据是独立同分布的, 目标域中的模型也不需要从头开始训练, 可以显著减少目标域对训练数据的需求和训练时间. 迁移学习旨在通过迁移不同但相关的源域中包含的知识来提高目标学习者在目标域上的表现, 这样可以减少构建目标学习器对大量目标域数据的依赖. 预训练模型能够获得人脸更多的语义信息, 更好地提取亲属人脸信息, 并且能够极大加快模型的训练速度.

3.2 损失函数

Softmax 函数主要用于解决多分类问题, Softmax 所获得的结果代表输入图像被分到每一类的概率. Softmax loss 是 Softmax 和 cross-entropy loss[12] 组合而成的损失函数, Softmax loss 把正确类别对应的输出 Softmax 值最大化.

本文提出把 Softmax 损失与中心损失 (center loss)[13] 进行联合监督学习, 这种方式能更好的收敛, 在增大类间距离的同时减小类内距离, 使网络模型获得的特征具有更强的鉴别能力. 联合 Softmax 损失与中心损失 (center loss) 函数如式 (5) 所示:

$$L=-\frac{1}{m}\left[\sum_{i=1}^{m}\log\frac{\mathrm{e}^{W_{y(i)}^{\mathrm{T}}x(i)}}{\sum_{l-1}^{k}\mathrm{e}^{W_l^{\mathrm{T}}x(i)}}\right]+\frac{\lambda}{2}\sum_{i=1}^{m}\left\|x_i-c_{y_i}\right\|_2^2 \tag{5}$$

其中, m表示样本个数, w为网络模型权重, x_i表示第i张图片特征值, y_i表示第i类, c_{y_i}表示该第i张图片所属分类特征值的中心. λ用来平衡 2 个损失函数, 合适的λ选择有助于增强网络的特征鉴别能力, 当λ=0 时, 训练网络时仅有 Softmax 损失监督学习.

4 实验与结果

4.1 实验设置与参数

实验环境为 Windows 10、64 位操作系统、内存 16 GB、Python 编程语言、TensorFlow 深度学习框架、NVIDIA GeForce GTX 1650 和 Intel(R) Core(TM) i7-10700F CPU @ 2.90 GHz (16 CPUs). 对 S-ViT 特征提取主干网络预训练模型冻结全部权重, 仅训练全连

接层参数. 然后, 使用学习率随着训练 epoch线性下降的方法, 以防止学习率过大导致在收敛到接近全局最优点时来回震荡, 每迭代 10 个 epoch 后, 如果最大验证精度没有提升时, 学习率衰减一半. 实验参数设置如表 2 所示.

表 2 参数设置

参数	取值
Epoch	40
BatchSize	16
Optimizer	Adam
Learning rate	0.000 1

4.2 实验数据集

采用目前最大、最全面的亲属识别人脸图像 FIW 数据集, 包含的共 11 种亲属关系可分为同代关系、第 1 代关系和第 2 代关系共 3 代. 同代关系包括兄弟 (B-B)、姐妹 (S-S)、兄妹 (SIBS); 第 1 代关系包括父女 (F-D)、父子 (FS)、母女 (M-D)、母子 (MS); 第 2 代关系包括祖父孙女 (GF-GD)、祖父孙子 (GF-GS)、祖母孙女 (GM-GD)、祖母孙子 (GM-GS). 配偶之间不具有血缘关系, 不作为研究亲属关系的样本. 部分亲属关系人脸图片如图 4 所示.

4.3 结果与分析

用构建的 S-ViT 基线模型对 FG2020 挑战的 3 个任务进行全面的实验. 实验首先需要确定联合损失函数不同参数λ对 S-ViT 模型精度的影响, 结果如图 5 所示. 由图 5 可知, 一个合适的λ值有助于提高亲属关系验证精度, 参数λ的最佳结果为 0.003. 同时, 从图 4 中也可以得到, 所提 Softmax 损失和中心损失联合监督学习损失函数的亲属关系验证精度要高于仅 Softmax 损失监督学习 (当λ=0 时). λ=0.003 时, 模型损失曲线和验证精度如图 6 所示, 迁移学习极大加快了模型训练速度, 模型最终获得了 76.8% 的验证精度.

图 4 FIW 中部分亲属关系人脸图片

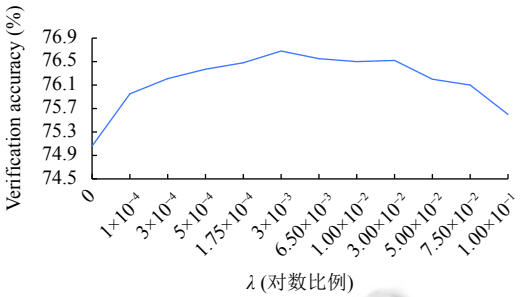
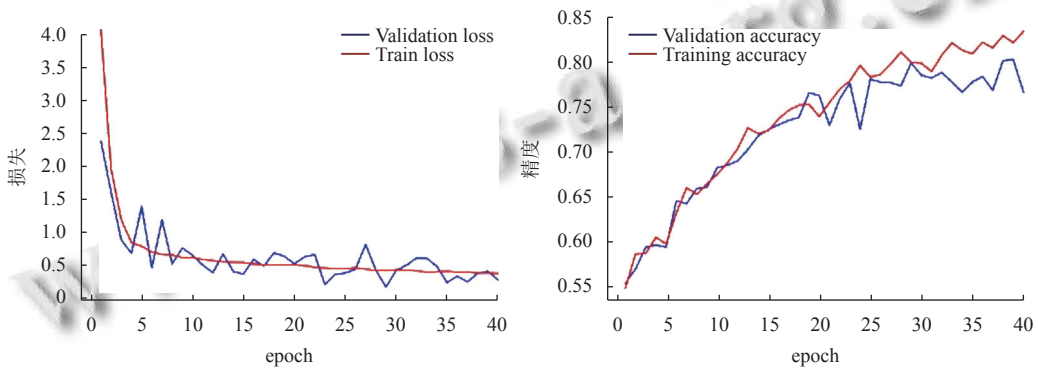


图 5 不同 λ 对应的 S-ViT 模型验证精度



(a) 损失曲线 (b) 验证精度

图 6 S-ViT 模型损失曲线和验证精度

4.3.1 亲属关系验证和分析

为了验证 S-ViT 模型性能, 与 FG2020 挑战赛领先的方法进行对比, 结果如表 3 所示. 从实验结果可知, S-ViT 验证精度比基线模型高 14%, 并列排名第二, S-ViT 方法获得不错的结果. 11 种家庭关系对的验证精度折线图如图 7 所示, 由图 7 可以看出, 所提出的 S-ViT 方法在多数关系对验证上, 相比其他方法都能够取得较好的精度. 在 GF-GS、GM-GS 关系对上, 所列举的方法精度均较低, 这是因为关系对人脸图片数量较少以及年龄等因素导致的. 同时, 从实验结果可知,

所提出的基于具有全局自注意力机制的 S-ViT 模型获得了很好的基线精度. 另外, 在以上 FG2020 挑战赛领先方法中, 均采用基于卷积神经网络 ResNet50 或 VGG 作为主干特征网络来构建模型.

表 3 与 FG2020 挑战赛领先方法亲属关系验证精度对比

Methods	F-D	F-S	M-D	M-S	GF-GD	GF-GS	GM-GD	GM-GS	B-B	S-S	SI-BS	Avg.
Baseline	0.61	0.66	0.69	0.62	0.66	0.71	0.73	0.68	0.57	0.64	0.50	0.64
Stefhoer	0.77	0.80	0.77	0.78	0.70	0.73	0.64	0.60	0.66	0.65	0.76	0.74
Ustc-nelslip	0.76	0.82	0.75	0.75	0.79	0.69	0.76	0.67	0.75	0.74	0.72	0.76
DeepBlueAI	0.74	0.81	0.75	0.74	0.72	0.73	0.67	0.68	0.77	0.77	0.75	0.76
S-ViT	0.75	0.80	0.76	0.74	0.77	0.72	0.78	0.67	0.77	0.76	0.76	0.76
Vuvko	0.75	0.81	0.78	0.74	0.78	0.69	0.76	0.60	0.80	0.80	0.77	0.78

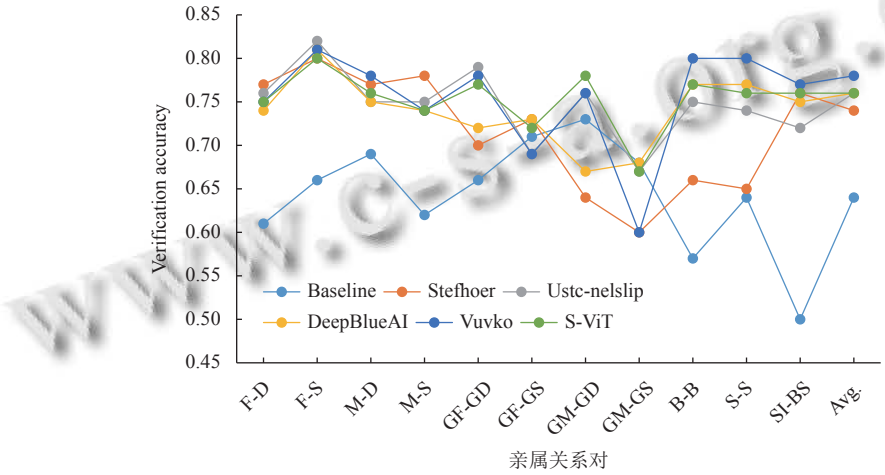


图 7 11 种亲属关系对的验证精度折线图

为进一步证明所提出方法的性能, 用提出的 S-ViT 方法, 将常见的卷积神经网络 VGG16、ResNet50 和 SENet50 和分别作为 S-ViT 的主干特征提取模型, 替换 PCPVT 特征提取主干网络, 构建仅有 CNN 的网络模型. 且所有的卷积特征提取网络模型均在 ImageNet 上进行预训练, 构建的对比孪生网络其他参数设置与 S-ViT 一致. 获得的结果如表 4 所示. 作为基线方法, 所提出的基于自注意力机制 S-ViT 网络模型也获得了最好的精度, 这进一步证明了 S-ViT 网络的良好性能. 9 种亲属关系验证方法对比如图 8 所示, 由此可知, 与 FG2020 挑战赛领先方法对比, 不同卷积主干模型所构建的孪生网络性能较差, 但所提出基于自注意力机制的 S-ViT 模型获得了较好的精度, 表明所提取的亲属人脸图像特征具有更强的语义信息.

表 4 不同主干特征提取网络验证精度对比 (%)

Backbone	Acc
VGG16	72.6
SENet50	75.4
ResNet50	75.8
PCPVT (S-ViT)	76.8

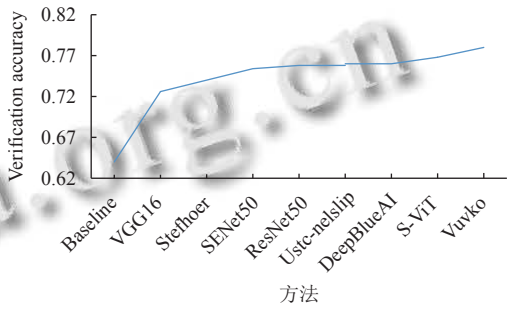


图 8 9 种亲属关系验证方法精度对比

4.3.2 三主体验证和分析

任务 2 的目标是预测一个孩子是否与一对父母有关系. 本质上, 它与任务 1 一样, 可以分为两个子问题, 即确定他们在“父-子”和“母-子”之间是否有亲属关系. 三主体 (父母-孩子) 通过 S-ViT 网络模型分别得到三主体特征表达, 对父亲-孩子和母亲-孩子的特征分别做余弦相似, 最后把得分求平均获得相似度得分. 如表 5 所示, 与 FG2020 挑战赛前三对比, 作为基于自注意力机制的基线方案, 所提 S-ViT 方法获得了令人鼓舞的评分, 获得了第 3 名. 其中, 在 FG2020 三主体验证领先

方法中, 也都是采用基于卷积神经网络 ResNet50 或 VGG 作为主干特征网络来构建模型.

表 5 与 FG2020 三主体验证领先方法性能比较

Methods (Rank)	FMS	FMD	Avg.
Baseline (5)	0.68	0.68	0.68
Stefhoer (4)	0.74	0.72	0.73
S-ViT (3)	0.76	0.75	0.75
DeepBlueAI (2)	0.77	0.76	0.77
Ustc-nelslip (1)	0.8	0.78	0.79

4.3.3 亲属关系查询和检索

任务 3 的目标是在所有家庭人脸图片 (图集) 中找到搜索对象 (即输入人脸图片) 的家族成员, 这是一个多对多的排名问题. 通过比较全类平均正确率 (mean average precision, mAP) 和 Rank@5 正确率, 如表 6 所示, S-ViT 网络模型获得了第 3 名, 这表明所提出 S-ViT 模型在亲属关系查询和检索任务上的可行性和有效性. 此处, 在 FG2020 查询和检索领先方法中, 构建网络模型采用的是卷积神经网络预训练模型.

表 6 与 FG2020 查询和检索领先方法性能比较

Methods (Rank)	mAP	Rank@5
Baseline (6)	0.02	0.10
DeepBlueAI (5)	0.06	0.32
HCMUS notweeb (4)	0.07	0.28
S-ViT (3)	0.07	0.34
Ustc-nelslip (2)	0.08	0.38
Vuvko (1)	0.18	0.60

5 结论和未来的工作

本文把具有自注意力机制的 Vision Transformer 作为特征提取主干网络并与 CNN 结合起来, 为亲属关系识别挑战提出了新的集成解决方案. 实验表明, 通过自注意机制提取的特征与 CNN 不同, 通过与 CNN 结合可以更好地提取亲属关系人脸对的特征, 在亲属关系识别上获得良好的预测性能, 同时实验也验证了 Vision Transformer 和 CNN 可以产生不同的基础分类网络. 本文虽然没有对提出的 S-ViT 方法进行微调和消融实验验证, 但是提出的方法显示了巨大的潜力, 在 FG2020 挑战赛 3 个任务上也能排名前三.

尽管利用具有自注意力机制的 PCPVT 作为主干特征提取网络发挥了很好的作用, 但仍有一些特定的模块和优化在这项工作中没有考虑.

1) 更多的具有自注意力机制的 Vision Transformer 也可以用作主干特征提取网络, 如 Swin Transformer[14]、PVT[8] 等, 同时也可以集成 Vision Transformer 和 CNNs 多种主干特征提取网络, 进一步提取更有鉴别能力的亲属关系特征.

2) 除了 S-ViT 使用的联合损失函数外, 其他的损失函数, 如对比损失[15]、三重损失[16] 也可能用到 S-ViT 中, 以便在训练中更有效地利用标签信息.

3) 在特征融合阶段可以有更多选择, 如文献 [17,18].

4) 未对人脸图像集预处理, 如图像增强、人脸对齐等.

5) CNN 的全连接层没有进行微调, 如层数、神经元个数、dropout, 优化器的选择等[19,20].

总的来说, 具有全局信息捕获能力的自注意力机制用于亲属关系识别目前仍少. 相信未来还有许多潜在的变体需要探索, 希望提出的 S-ViT 模型可以作为一个参考.

参考文献

1 Dornaika F, Arganda-Carreras I, Serradilla O. Transfer learning and feature fusion for kinship verification. Neural Computing and Applications, 2020, 32(11): 7139–7151. [doi: 10.1007/s00521-019-04201-0]

2 Li ZW, Liu F, Yang WJ, *et al*. A survey of convolutional neural networks: Analysis, applications, and prospects. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(12): 6999–7019. [doi: 10.1109/TNNLS.2021.3084827]

3 Robinson JP, Shao M, Wu Y, *et al*. Visual kinship recognition of families in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(11): 2624–2637. [doi: 10.1109/TPAMI.2018.2826549]

4 Robinson JP, Qin C, Shao M, *et al*. The 5th recognizing families in the wild data challenge: Predicting kinship from faces. Proceedings of the 16th IEEE International Conference on Automatic Face and Gesture Recognition. Jodhpur: IEEE, 2021. 1–7.

5 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.

6 Chu XX, Tian Z, Wang YQ, *et al*. Twins: Revisiting the design of spatial attention in vision transformers. Proceedings of the 35th International Conference on Neural Information Processing Systems. 2021. 9355–9366.

7 Chu XX, Tian Z, Zhang B, *et al*. Conditional positional encodings for vision transformers. arXiv:2102.10882, 2021.
8 Wang WH, Xie EZ, Li X, *et al*. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 568–578.
9 Bromley J, Guyon I, LeCun Y, *et al*. Signature verification using a “Siamese” time delay neural network. Proceedings of the 6th International Conference on Neural Information Processing Systems. Denver: Morgan Kaufmann Publishers Inc., 1993. 737–744.
10 Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database. Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009. 248–255.
11 Tan CQ, Sun FC, Kong T, *et al*. A survey on deep transfer learning. Proceedings of the 27th International Conference on Artificial Neural Networks. Rhodes: Springer, 2018. 270–279.
12 de Boer PT, Kroese DP, Mannor S, *et al*. A tutorial on the cross-entropy method. Annals of Operations Research, 2005, 134(1): 19–67. [doi: 10.1007/s10479-005-5724-z]
13 Wen YD, Zhang KP, Li ZF, *et al*. A discriminative feature learning approach for deep face recognition. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 499–515.
14 Liu Z, Lin YT, Cao Y, *et al*. Swin Transformer: Hierarchical vision transformer using shifted windows. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 10012–10022.
15 Khosla P, Teterwak P, Wang C, *et al*. Supervised contrastive learning. Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2020. 18661–18673.
16 Hoffer E, Ailon N. Deep metric learning using triplet network. Proceedings of the 3rd International Workshop on Similarity-based Pattern Recognition. Copenhagen: Springer, 2015. 84–92.
17 Yu J, Li MY, Hao XL, *et al*. Deep fusion Siamese network for automatic kinship verification. Proceedings of the 15th IEEE International Conference on Automatic Face and Gesture Recognition. Buenos Aires: IEEE, 2020. 892–899.
18 Luo ZP, Zhang ZG, Xu ZY, *et al*. Challenge report recognizing families in the wild data challenge. Proceedings of the 15th IEEE International Conference on Automatic Face and Gesture Recognition. Buenos Aires: IEEE, 2020. 868–871.
19 Dahan E, Keller Y. A unified approach to kinship verification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(8): 2851–2857.
20 Huang JY, Strome MB, Jenkins I, *et al*. Solving the families in the wild kinship verification challenge by program synthesis. Proceedings of the 16th IEEE International Conference on Automatic Face and Gesture Recognition. Jodhpur: IEEE, 2021. 1–5.

(校对责编: 孙君艳)