Software Engineering and Applications 软件工程与应用, 2022, 11(6), 1561-1569
Published Online December 2022 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2022.116161

Hans汉斯

SA-C3D神经网络在动作识别上的应用

张宏博*，陈　胜

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2022年11月30日；录用日期：2022年12月23日；发布日期：2022年12月31日

摘　要

本文的主要目的是利用自注意力机制加强C3D网络在动作识别方面的准确率。C3D神经网络作为比较早提出的模型，在视频动作识别领域中有着重要的地位。随着各项研究的进展，C3D网络已经渐渐过时，识别准确率也较低。所以本文主要以C3D网络为基础，结合目前的自注意力机制，在C3D网络中集成了Non-Local模块，同时将固定学习率衰减替换为余弦退火学习率衰减，提高模型跳出局部最优解的能力。利用3D卷积提取动作视频的局部特征，再使用自注意力机制捕捉人体动作的全局信息，开发出新的SA-C3D网络。在没有预训练的前提下，对UCF-101数据集进行训练，识别准确率较之前的C3D网络以及一系列优秀的动作识别模型有了较大的提高，识别准确率高达95%。

关键词

C3D，3维卷积神经网络，自注意力，Non-Local，动作识别

Application of SA-C3D Neural Network in Action Recognition

Hongbo Zhang*, Sheng Chen

School of Optoelectronic Information and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Nov. 30th, 2022; accepted: Dec. 23rd, 2022; published: Dec. 31st, 2022

Abstract

The main objective of this paper is to enhance the accuracy of C3D networks for action recognition using a self-attentive mechanism. C3D neural networks, as a relatively early proposed model, have an important place in the field of video action recognition. With the progress of various researches,

*通讯作者。

文章引用: 张宏博, 陈胜. SA-C3D 神经网络在动作识别上的应用[J]. 软件工程与应用, 2022, 11(6): 1561-1569.
DOI: 10.12677/sea.2022.116161

C3D networks have gradually become obsolete and the recognition accuracy is low. Therefore, this paper focuses on the C3D network as the basis, combining the current self-attentive mechanism, integrating the Non-Local module in the C3D network, while replacing the fixed learning rate decay with the cosine annealing learning rate decay to improve the ability of the model to jump out of the local optimal solution. The new SA-C3D network is developed by using 3D convolution to extract local features of action videos, and then using a self-attentive mechanism to capture global information of human actions. Trained on the UCF-101 dataset without pre-training, the recognition accuracy has improved significantly over the previous C3D network and a series of excellent action recognition models, with recognition accuracy as high as 95%.

Keywords

C3D, 3-Dimensional Convolutional Neural Networks, Self-Attention, Non-Local, Action Recognition

1. 引言

随着社会的发展，互联网上的多媒体也在迅速发展，这也导致了在互联网上每分钟被分享的视频数量越来越多。为了应对信息爆炸，理解和分析这些视频是必要的，以达到各种目的，如搜索，推荐，排名等。而在进入了自媒体时代以后，QQ、小红书、微信、抖音等软件在社会中被越来越多的人使用，在字节跳动公布的 2022 年数据报告中，抖音软件的日活用户已经达到了惊人的数字，报告中显示，抖音的日活用户已经超过 6 亿人口。伴随着视频量的爆炸式增长，视频的商业用途也在逐渐地扩大，伴随着商业上的需求不断的增大，计算机视觉方面的技术得以快速发展，在过去十几年中，计算机视觉技术被广泛的应用于以下几个方面，1) 目标检测[1] [2]，2) 人体动作识别[3]，这其中人体动作识别方面的应用更为广泛，人体动作识别技术是计算机视觉应用的一大代表，它在办公交互、虚拟竞技、虚拟游戏、医疗诊断、视频监控等领域都有很广泛的应用[4]。尤其是近些年短视频井喷式的爆发，海量的视频信息需要人们来进行监管，但是仅仅依靠人工来检查这些巨量的视频信息是几乎不可能完成的任务，其人力资源上的消耗就足以让任何一家公司破产，在这种情况下，我们必须要研发一种高效率、高速率、稳定、可靠、智能的视频内容分析方法来满足大量视频处理的需求。视频分类技术作为视频内容分析的一种重要方法和计算机视觉领域的一个重要研究方向，是解决上述问题的有效方法，同时也受到越来越多研究者的关注[5]。

在本文的研究中，在对人体动作识别的经典模型 C3D [6]进行改进和加强，C3D 在 2015 年由 Facebook 提出，可以将视频处理后，直接输入神经网络中进行训练，由于预处理过程较少，在训练神经网络的时候消耗的时间也相对较少，而且准确率高，速度快，具有极为广阔的应用场景。但是随着技术的进步，C3D 网络已经渐渐被其他神经网络所超越，精度已经达不到现在视频处理的标准，所以在改进过程中，文中使用了现在最新的自注意力机制[7]，用 C3D 网络和注意力机制中的 Non-Local 神经网络模块进行结合，再使用余弦退火学习率衰减的方法，改进成对动作识别准确率更高的 SA-C3D 网络，大大加强了神经网络的学习能力与识别能力，文中实验所使用的数据集是来自 YouTube 官方网站的经典数据集 UCF-101，这个数据集中主要包括了人物和物体直接的一些行为，人与人之间玩耍、运动等，人使用一些乐器时的行为，还有各类乐器和肢体动作表现。目前 C3D 网络在有预训练模型的前提下，在 UCF-101

数据集上的识别准确率为 85.2%，但是在没有预训练模型的情况下，识别准确率下滑了许多。而在本文中，同样没有预训练模型的 SA-C3D 模型在识别 UCF-101 数据集中的动作上，准确率有了长足的进步，甚至还要超过有预训练模型的 C3D 模型识别准确率。

2. 算法设计与改进方法

2.1. C3D 网络

在之前的研究中，卷积神经网络 CNN [8] [9]中使用的 2D 卷积神经网络普遍用于提取单张图片的特征，但无论输入的是单张还是多张图片，都只能输出二维结果，但是其中的动作信息会被忽略，在每次卷积过后，输入的时间信息会被忽略，在这样的情况下，想要完成物体的动作识别就不能再依赖 2D 卷积，Du Tran 提出了一种简单而有效的时间特征学习方法，使用在大规模的视频数据集上训练的深度三维卷积网络(3D 卷积网络) [10]。也就是 C3D 网络架构，如图 1 C3D 网络架构图，C3D 网络有 8 个卷积层，5 个最大池化层和 2 个全连接层，然后是一个 softmax 输出层。所有的三维卷积核都是 3 × 3 × 3，在空间和时间维度上步幅为 1。在每个方框中表示过滤器的数量。所有的池化层都是 2 × 2 × 2，除了 Pool1 是 1 × 2 × 2。而全连接层有 4096 个输出单元。C3D 在一开始实验中便应用于人体动作数据集 UCF-101，效果显著，但是从 2015 年诞生到现在已经过了几年的时间，在这几年的时间里更好的深度学习模型层出不穷，尤其像今年来的 transformer 等自注意力模型，在对图像识别等方面效果非常显著，自注意力机制在二维图像上的优秀表现引起了人们的注意，经过推广到三维，也就是对视频的学习当中，效果也是非常不错的，尤其在对于全局细节上的学习，而 C3D 网络更加关注的是动作的局部信息，所以在接下来的实验中，我将 C3D 与自注意力机制进行了结合，形成了新的网络 SA-C3D，这样网络可以既学习人体动作视频的全局特征，也可以关注到局部特征，在后续的实验中，我们也得到了高于 C3D 网络识别动作准确率的结果。

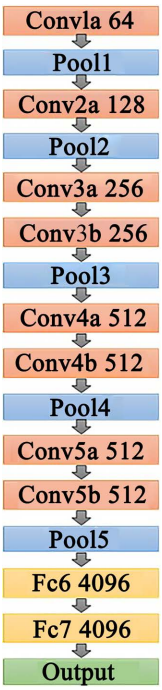

Figure 1. C3D architecture
图 1. C3D 网络结构

2.2. 自注意力机制的融入

在对人体工作进行识别的实验中，如果想要准确判断一个人正在做什么动作，那么必须关注这个人正在做什么，是用身体的某个部位或者利用什么工具在做，那么就需要关注视频中这个运动部位的特征。在之前已经介绍了，C3D 网络可以顺利的捕捉到人的运动部位或者是视频中正在完成这个动作的位置，从而完成动作的分类，但是准确度还有待提高，为了提高深度学习网络识别的准确度，我们采用了一种可以集成进入 C3D 网络中的模块，Non-Local 神经网络模块，经过实验检验，大幅度提高了深度学习网络模型的分类准确率。

Non-Local 神经网络[11]是一种基于自注意力机制[12]设计的网络模块，这个神经网络的结构更加关注视频或者动作的全局特征，这和 3D 卷积的特点可以互补，在之前的 C3D 架构图中我们可以看出，使用的卷积基本上都是 $3 \times 3 \times 3$ 的，那么这样的卷积操作只能捕捉视频的上下几帧中的特征，如果想要卷积神经网络关注长距离上的信息特征，那么必须多次重复使用卷积进行计算，这就造成了运算效率的低下。而 Non-Local 模块可以捕捉长距离上的信息依赖，而不是只关注于局部信息，非局部操作计算在某个位置上的响应，作为在输入特征图中所有位置上的特征的加权和。

Non-Local 的公式如式(1)，i 代表要计算相应的输出位置，x 表示输入信号，y 代表输出信号，他们具有相同大小，函数 f 计算出 i 和所有 j 之间的一个标量。一元函数 g 计算在位置 j 处的输入信号的表示。响应由因子 $C(x)$进行归一化处理[11]。

$$y_i = \frac{1}{C(x)} \sum_{\forall_j} f\left(x_i, x_j\right) g\left(x_j\right) \tag{1}$$

由于 Non-Local 模块输入和输出大小是一样的，故可以很容易的无缝衔接到卷积神经网络中，而且 Non-Local 并不会对深度学习网络中的每一个 block 都引入，主要原因是 non local 机制的设计初衷就是为了获取全局信息，而原来的卷积操作是为了获取局部信息，二者相辅相成才能有好的效果。Non-Local 模块大致示意图见图 2。

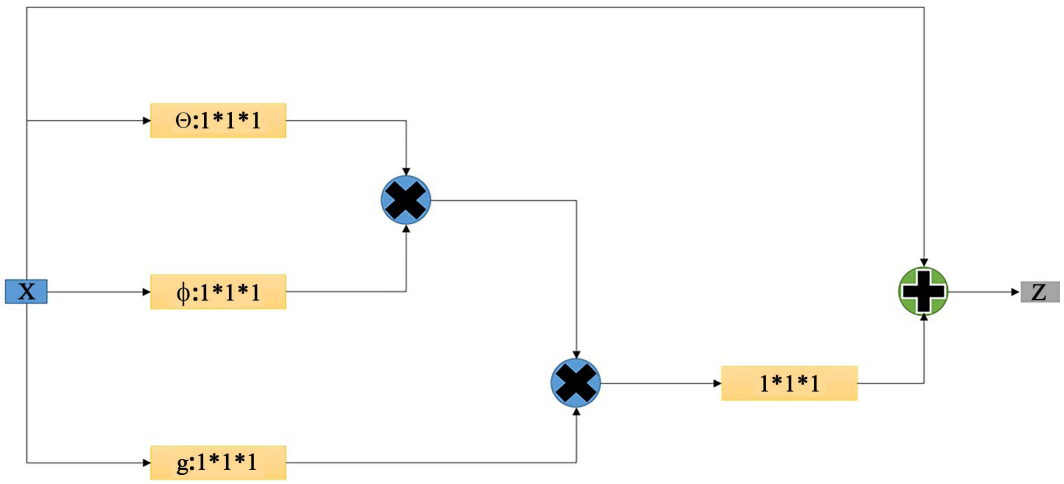

Figure 2. Non-Local Block
图 2. Non-Local 模块

2.3. 余弦退火(Cosine Annealing)学习率衰减法

因为我们的目标优化函数可能是多峰的，除了全局最优解之外还有多个局部最优解，在训练时梯度

下降算法可能陷入局部最小值，此时可以通过突然提高学习率，来“跳出”局部最小值并找到通向全局最小值的路径。这种方式称为带重启的随机梯度下降方法[12]。

余弦退火(cosine annealing)的原理如式 2

$$\eta_t = \eta_{\min}^i + \frac{1}{2}\left(\eta_{\max}^i - \eta_{\min}^i\right)\left(1+\cos\left(\frac{T_{cur}}{T_i}\pi\right)\right) \tag{2}$$

i 表示当前运行次数，$\eta_{\max}^i$ 和 $\eta_{\min}^i$ 表示学习率最大值和最小值，定义了学习率的范围，T_{cur} 表示当前执行了多少个周期，T_i 表示第 i 次运行中总的 epoch 数。通过这种方法可以提高模型对于本文中这样的小数据集的学习能力，使模型的鲁棒性更好，避免掉入局部最优的情况，如图 3。

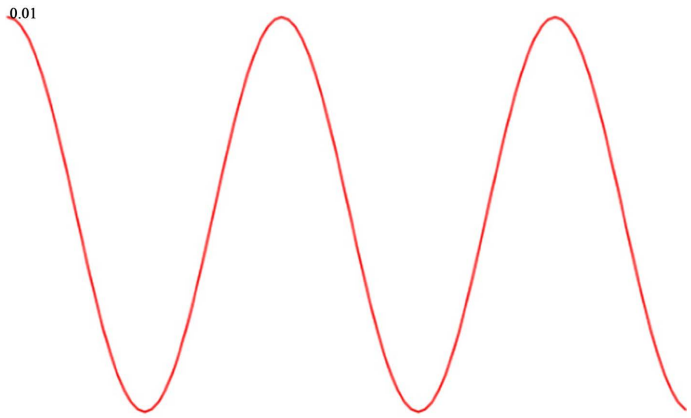

Figure 3. Cosine annealing learning rate
图 3. 余弦退火学习率

在图中可以看出，学习率衰减类似余弦曲线，不仅仅是一直减少，有时也会上升，这便可以帮助模型跳出一些局部最优点，其中纵坐标是学习率，横坐标是周期。

在本文的实验中，使用余弦退火学习率衰减法比使用固定周期衰减或不衰减的学习率，更能让深度学习模型充分的学习，具有更好的泛化能力。

2.4. SA-C3D

基于以上提出的 Non-Local 模块和 C3D 网络框架的特点，这里我们设计了一种将 C3D 和 Non-Localblock 结合的新型深度学习网络框架。如图 4，在卷积网络中，根据 Non-Local 论文中的建议，尽量将 Non-Local 模块放在较靠前的卷积层，但这样也会带来一个问题，那就是会导致巨大的参数，计算机很可能难以负担如此巨大的运算参数，所以在 Non-Local 模块中，我们又在模块输出的位置添加了池化层，以便减少一部分参数，提高计算机的运行效率，加快深度学习网络的训练速度。

3. 实验结果及分析

3.1. 实验环境

在我们准备的实验中，实验室使用的设备是 AMD 5900X CPU，内存为 64 GB，硬盘为 10 T，GPU 为 NVIDIA GeForceRTX 3090，GPU 显存为 24 GB。操作系统为 windows 11，编程语言选择 Python 3.8，深度学习框架采用 PyTorch1.7.1，主要使用的几个 python 库为 Cuda 11.1、Cudnn 8.1、OpenCV 4.4、Pillow 7.2、NumPy 1.20.1、Matplotlib 3.3.5。

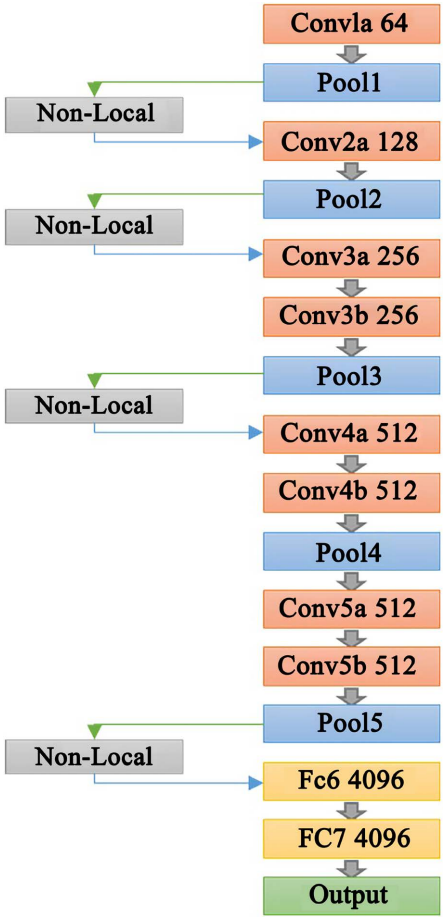

Figure 4. SA-C3D net framework
图 4. SA-C3D 网络架构

3.2. 数据集的使用

为了对比我们修改后的深度学习网络的学习能力，数据集并没有修改，仍然使用的是 C3D 论文中所使用的 UCF-101 数据集，之前提到过，该数据集收集自 YouTube 上，它是一个现实动作视频的动作识别数据集，它提供了来自 101 个动作类别的 13,320 个视频，总共时长约 27 h。UCF101 中收集的数据全面而且丰富，基本覆盖了人们日常生活中的各类运动，同时也在不同的背景下收集了大量的人体动作视频，提供了最大的多样性，它也包含了像摄像机的大量移动、杂乱的背景、亮度不一的照明条件等干扰因素，这些问题都非常考验深度学习模型的学习能力和泛化能力，对模型的训练也带来了极大的难度。

3.3. 数据的预处理

数据预处理这里采用的是先将视频转为每一帧的图像进行保存，再按照每个视频取 16 帧的方法进行取帧，确保得到的 16 帧图片可以完整的覆盖整个动作，这样便可以得到动作的完整信息，在转化为帧图像后，将整个数据集细分为训练集、测试集、还有验证集，按照 6:2:2 的比例划分，在训练时，对每一帧的大小也进行了规范限制，统一裁剪为 171 × 128 大小，然后再输入进入深度学习网络中进行训练。

3.4. 实验结果分析

实验整体流程如图 5。

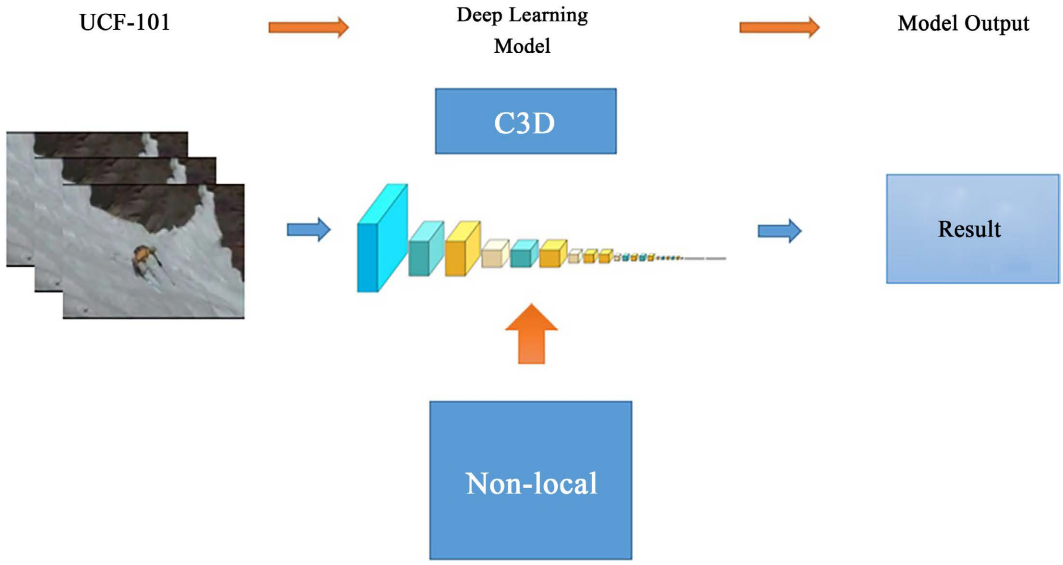

Figure 5. Flow chart of experiment
图 5. 实验流程图

本文中的实验通过 UCF-101 数据集进行训练，都是在没有使用预训练模型的前提下进行的训练(因为加入预训练模型训练的 C3D 模型论文中准确率为 85.)，进行了 150 次迭代，训练时长约 26 小时，并且每隔 10 个周期便会用测试集进行一次测试，训练集准确率如图 6 所示，函数值损失曲线图见图 7。

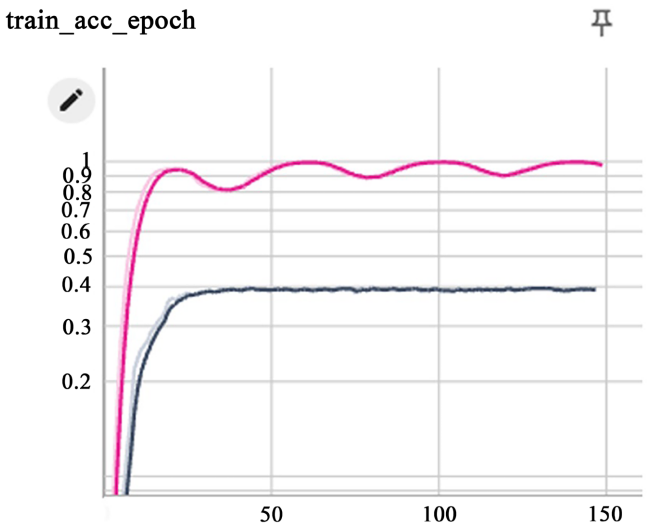

Figure 6. UCF-101 accuracy curve
图 6. UCF-101 准确率曲线图

这里可以看出，没有加入预训练模型的 C3D 网络识别准确率只达到了百分之四十便无法上升了，而同样没有加入预训练模型的 SA-C3D 网络在 UCF-101 数据集上的表现非常好，识别率在百分之九十五以上，一度达到了百分之百的识别准确率，这足以看出 Non-Local 模块和余弦退火学习率对 C3D 模型的学习能力的大幅加强。同时在 C3D 论文中，作者的实验表明，C3D 网络在有预训练模型的前提下，在 UCF-101 数据集上的识别准确率为 85.2%，这一数据也比本文中的 SA-C3D 模型结果要差，具体对比见下表 1。

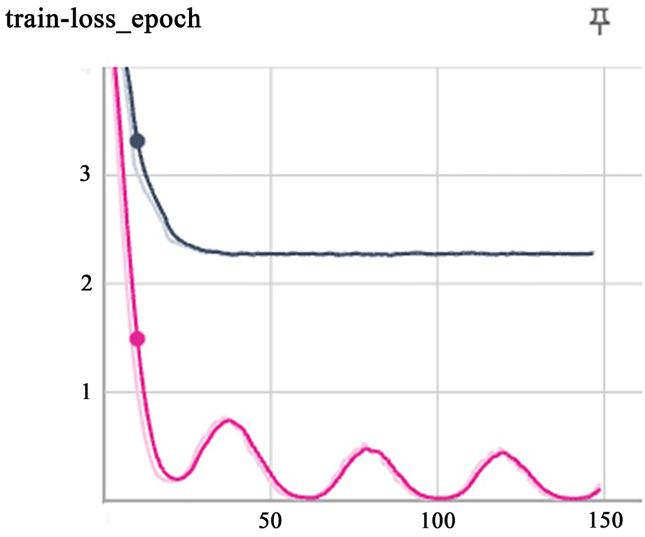

Figure 7. UCF-101 accuracy curve
图 7. UCF-101 准确率曲线图

Table 1. Comparison of the method in this paper with other methods
表 1. 本文方法与其他方法的比较

方法	识别准确率/%
No-pretrain-C3D	39.8
C3D	85.2
R3D [12]	87.6
Two-stream [13]	88.0
Network model in this paper	97.5

见上表，在表 1 中我们列举了几种不同深度学习模型，在相同实验条件下，对心梗病人进诊断的数据对比，从表中可以看出，未加入预训练模型的 C3D 网络效果不佳，但是即便是加入了预训练模型的 C3D 网络，在识别准确率上仍然弱于本文的 SA-C3D 网络，SA-C3D 网络可以很好的捕捉到帧与帧之间的运动信息以及帧的空间信息。相较于识别精度较高的 R3D、Two-stream 网络，本文的网络模型仍然具有识别准确率较高的优势，而且 Non-Local 模块可以很好的关注到全局运动信息，这保证了不会在识别人体动作时，由于前后视频跨度较长导致特征提取不完全的情况。

4. 结语

在本文的研究中，我们开发了一种快速分类的深度学习框架，通过提取运动特征，来对人体动作数据集UCF-101进行动作识别，在C3D的基础上本文集成了自注意力模块加强模型对全局信息的提取能力，同时引入余弦退火学习率衰减，取代原先的固定衰减学习率，加强了模型的学习能力，帮助模型可以跳出局部最优解，强化了模型的泛化能力。对本文深度学习网络进行测试，可以看出同样不加入预训练模型的前提下，C3D网络的泛化能力明显变弱，而本文的SA-C3D网络不仅远远超过没有预训练模型的C3D网络，就算是加入了预训练模型的C3D网络，在准确率是仍然弱于本文加强后的深度学习网络，同时也参考当前比较优秀的几种动作识别深度学习网络，可以看出本文的网络在识别准确率上仍然有着明显的优势。

本文的研究主要参考深度学习在动作识别方面的网络，并在需要的地方加以改进，改进后的网络不仅仅能用于动作识别方面，同时也可以用于在医学图像方面，在目前平行进行的实验中，利用合作医院提供珍贵数据集，在医学图像分析中表现出了卓越的性能[14]。在过去很多专家学者们研究中，医学图像中的很多特征，可以被深度学习网络所准确的捕捉到，比如在心肌梗死病人的超声心动图中的心肌运动异常，就可以通过大数据训练来准确的抓取，从而让模型具有比人类视觉检查更高的精度和灵敏度[15]。所以在未来的实验中，也许动作识别网络可以用于疾病的智能诊断上，在医学领域继续发光发热。

参考文献

[1] Wang, C., Liu, M. and Qi, F. (2018) Summary of Dynamic Target Detection and Recognition Algorithm in Intelligent Video Surveillance System. Electrical Engineering.

[2] 李坤坤, 刘正熙, 熊运余. 基于深度学习的目标检测系统性文献综述[J]. 现代计算机, 2021(16): 98-102, 117.

[3] Zhang, S., Wei, Z., Nie, J., *et al*. (2017) A Review on Human Activity Recognition Using Vision-Based Method. *Journal of Healthcare Engineering*, No. 3, 1-31. https://doi.org/10.1155/2017/3090343

[4] 钱闻卓. 基于 MA-C3D 神经网络的人体动作识别技术[J]. 现代计算机, 2021, 27(35): 70-74+94.

[5] 孙毅, 成金勇, 禹继国. 基于 C3D 模型的视频分类技术[J]. 曲阜师范大学学报(自然科学版), 2020, 46(3): 85-89.

[6] Tran, D., Bourdev, L., Fergus, R., *et al*. (2015) Learning Spatiotemporal Features with 3d Convolutional Networks. *Proceedings of the IEEE international Conference on Computer Vision*, Santiago, 11-18 December 2015, 4489-4497. https://doi.org/10.1109/ICCV.2015.510

[7] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. 31*st Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 5998-6008.

[8] Deng, J., *et al*. (2009) Imagenet: A Large-Scale Hierarchical Image Database. *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, 20-25 June 2009, 248-255. https://doi.org/10.1109/CVPR.2009.5206848

[9] Krizhevsky, A. and Hinton, G. (2009) Learning Multiple Layers of Features from Tiny Images.

[10] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L. (2014) Microsoft COCO: Common Objects in Context. 13*th European Conference*, Zurich, 6-12 September 2014, 740-755. https://doi.org/10.1007/978-3-319-10602-1_48

[11] Wang, X., Girshick, R., Gupta, A., *et al*. (2018) Non-Local Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-22 June 2018, 7794-7803. https://doi.org/10.1109/CVPR.2018.00813

[12] Loshchilov, I. and Hutter, F. (2016) Sgdr: Stochastic Gradient Descent with Warm Restarts.

[13] Hara, K., Kensho, H. and Satoh, Y. (2017) Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition. 2017 *IEEE International Conference on Computer Vision Workshops* (*ICCVW*), Venice, 22-29 October 2017, 1109-1115. https://doi.org/10.1109/ICCVW.2017.373

[14] Abdel-Aty, H., Zagrosek, A., Schulz-Menger, J., *et al*. (2004) Delayed Enhancement and T2-Weighted Cardiovascular Magnetic Resonance Imaging Differentiate Acute from Chronic Myocardial Infarction. *Circulation*, **109**, 2411-2416. https://doi.org/10.1161/01.CIR.0000127428.10985.C6

[15] Smulders, M.W., Bekkers, S.C.A.M., Kim, H.W., *et al*. (2015) Performance of CMR Methods for Differentiating Acute from Chronic MI. *JACC*: *Cardiovascular Imaging*, **8**, 669-679. https://doi.org/10.1016/j.jcmg.2014.12.030