MMTM: Multimodal Transfer Module for CNN Fusion

Hamid Reza Vaezi Joze

摘要

在后期融合中，每个模态在单独的单峰卷积神经网络（CNN）流中进行处理，并在最后融合每个模态的分数。由于其简单性，后期融合仍然是许多最先进的多模态应用中的主要方法。在本文中，我们提出了一个简单的神经网络模块，用于利用卷积神经网络中多种模态的知识。所提出的单元名为多模传输模块（MMTM），可以添加到特征层次的不同级别，从而实现慢模态融合。使用挤压和激励操作，MMTM 利用多种模态的知识来重新校准每个 CNN 流中的信道方向特征。与其他中间融合方法不同，所提出的模块可用于具有不同空间维度的卷积层中的特征模态融合。所提出的方法的另一个优点是，它可以在单峰分支之间添加，其网络架构的变化最小，允许使用现有的预训练权重初始化每个分支。实验结果表明，我们的框架提高了知名多模态网络的识别精度。实验上表明，在大规模数据集 NTU-RGBD 上展示了最先进的或具有竞争力的性能，这些数据集使用 RGB，depth 和身体关节的动作识别等任务领域。

关键词：动作识别；多模态；CNN

1 引言

不同的传感器可以提供关于同一上下文的互补信息。多模态融合是从不同模态中提取和组合相关信息，从而比只使用一种模态提高性能。该技术被广泛应用于各种机器学习任务中，如视频分类 [1,2]、动作识别 [3]、情感识别 [4,5] 和视听语音增强 [6,7]。其中动作识别在许多领域起着非常重要的作用。现在大多数动作识别的技术都基于单个模态的数据，准确率没有达到很好的效果。现阶段多模态技术正在新起，具有很大的应用前景。将多模态技术结合到动作识别当中是新的趋势。多模态信息的结合可以有效屏蔽单模态信息的缺点。比如视频和深度图的多模态信息结合，可以解决视频模态信息中的不同背景，光线明暗的影响，也可以解决深度图中缺少关键环境信息的问题。所以我打算复现一篇多模态的动作识别的论文，可以学习多模态动作识别的知识内容。我选择的论文是 CVPR 2020 年的文章”MMTM: Multimodal Transfer Module for CNN Fusion”。这篇文章的目的是将现在研究成果很好的单模态动作识别，通过通道的选择和抑制，将各个单模态模型提取的特征信息结合起来，形成一种通道选择，或者说是通道抑制的过程。这种选择可以很好结合多模态的特征信息，同时解决仅使用单模态的缺点。虽然只用于 CNN 模型，但也是具有很大的拓展性。

尽管模块设计是通用的，并且可能在网络层次结构中的任何级别添加，但每个应用程序的最佳位置和模块数量是不同的。论文中设计了用于动作识别任务的特定于应用程序的网络，并研究了在其架构中添加 MMTM 的好处。然后对这些应用程序进行了以下经验观察。首先，将 MMTM 添加到中间和高级特征是有益的，而低级特征也是如此。这是因为与中级和高级特征相比，低级特征中的模态间相关性较低。其次，即使在 RGB 和深度模态在空间上对齐且无需挤压操作即可进行融合的手势识别中，挤压通过提供具有全局感受野的信息显着提高了性能。最后，门控操作的激励优于残差学习中常用的求和运算，突出了强调和抑制机制的重要性。

论文以两种不同的多模态融合动作识别应用设计了不同的网络架构。通过这些实验表明，MMTM提高了后期融合方法的性能。

2 相关工作

深度图，视频Runge和骨架等单模态已广泛应用于动作识别任务。这些方法都有各自的缺点。由于缺乏明确的人体模型而且还受背景和光线的影响，基于视频的动作识别方法对背景杂波和非动作干扰的处理能力较差。当然，深度图数据没有背景和光线强弱的影响，但是它没有解决缺乏明确的人体模型的问题，所以还是具有一些难以区分相似动作的问题。另一方面，如果仅仅依靠身体姿势比如骨架数据，视频中出现的大部分上下文和全局线索将会丢失。最近也有许多方法开发了架构来融合这些模式，以进一步提高动作识别的性能。

2.1 单模态动作识别

不同的模态具有不同的优缺点。单模态的动作识别任务也有了长时间的发展，也有了许多成熟的研究成果。

2.1.1 深度图

深度图是指像素值表示从给定视点到场景中点的距离信息的图像。深度模态通常对颜色和纹理的变化具有鲁棒性，为人类受试者提供了可靠的 3D 结构和几何形状信息，因此可用于 HAR。构建深度图的本质是将 3D 数据转换为 2D 图像，并开发了不同类型的设备来获得深度图像，其中包括主动传感器（例如，飞行时间和基于结构光的相机）和被动传感器（例如立体相机）[1]。有许多科研工作者将深度图应用于动作识别任务当中，有尝试将 CNN 网络应用于深度图数据[2]，来达到提取时空特征的目的.

2.1.2 RGB 视频

RGB 模态通常是指 RGB 相机捕获的图像或视频（图像序列），旨在重建人眼看到的内容。RGB 数据通常易于收集，包含捕获的场景上下文的丰富外观信息。基于 RGB 的 HAR 具有广泛的应用。由于 RGB 视频数据非常常见，科研人员提出一种 2D 卷积膨胀到 3D 卷积的概念，增加一个时间维度的卷积, 然后利用 3D 卷积从 RGB 视频中提取出 3D 维度的空间与时间特征信息，提出 I3D 网络[3]；还有一些根据研究残差网络对卷积网络有积极性影响的结论[4]，提出在 3D 卷积当中加入残差网络，更加有效的学习模型参数, 加以实验证明 R3D 的有效性[5]。

2.1.3 骨架数据

骨架序列对人体关节的轨迹进行编码，这些轨迹表征了信息丰富的人体运动。因此，骨架数据也是 HAR 的合适模态。通过在 RGB 视频[6][7] 或深度图[8] 上应用姿态估计算法可以获得骨架数据。Lei Shi et al.[9]提出一种双流自适应 GCN，双流对应骨骼节点与骨骼向量。在模型中加入自适应矩阵，学习有联系的骨骼节点，比如拍手的两个手掌。最后再对两个流学习出来的特征进行决策级分数融合。Chao Li et al.[10]也提出一个双流网络，双流也是对应骨骼节点与骨骼向量。不过与 Lei Shi et al.[9]不同

的是，它是基于 CNN 网络的，而且还对骨骼节点的维度进行卷积，来学习提取节点之间的信息联系。论文中还拓展了动作检测模块。Hyung-Gun Chi et al.[11]提出 InfoGCN 网络，基于信息瓶颈的学习目标，引导模型学习信息丰富但紧凑的潜在表征。论文中还提出了一种新的基于自我注意的图卷积模块 SA-GC，并证明它可以有效地从数据中收集行为上下文信息，使用推断的内在拓扑。Qingqing Huang et al.[12]提出了一个通用框架 STGAT，用于建模跨时空信息流。同时为了减少局部时空特征的冗余，缩小了自注意机制的范围，沿着时间维度动态加权关节，分别将细微运动与静态特征分离。

2.1.4 多模态动作识别

在现实生活中，人类经常以多模态认知的方式感知环境。同样，多模态机器学习是一种建模方法，旨在处理和关联来自多种模式的感官信息。通过聚合各种数据模式的优点和功能，多模态机器学习通常可以提供更健壮和准确的 HAR。多模态学习方法主要有两种类型，即融合和共同学习。融合是指整合来自两个或多个模态的信息进行训练和推理，而共同学习是指不同数据模态之间的知识转移。当然在动作识别任务当中，主要研究的方向还是如何去融合多个模态的信息进行训练和推理。

Jinmiao Cai et al.[13]提出利用骨架点数据的方法，根据骨骼节点截取每个节点对应部位的 RGB 视频，并形成光流信息，和原始的骨架信息来组成一个双流 GCN 网络，以决策级融合的方式进行多模态数据融合。Jianan Li et al.[14]设计一个双流的网络结构，分别以 ST-GCN 网络提取骨架特征信息和 R(2+1)D 网络提取 RGB 视频特征信息，但不是简单地特征级融合，在提取完特征后，通过引领块的 Compact Bilinear Correlation 或者 Feature learning correlation 的方法，利用骨架特征信息处理视频特征信息，最后再简单融合。Nuno C. Garcia et al.[15]利用知识蒸馏的方法，根据 teacher 和 student 的方法进行更新，达到一种根据其他模态调整自己模态的模型的方法，也是一种非常有趣的方法。每一次迭代都选择合适的 teacher，其他则为 student。Arsha Nagrani et al.[16]提出了三种服务于 transformer 模型的多模态融合方法。第一种是将各个模态的编码级联到一起，再传入到 transformer 中。第二种是有两个 transformer 模型，根据两个模态信息更新各自模型的参数。第三种方法是在两个 transformer 模型之间加入几个信息节点，对双方的信息的提取进一步是浓缩加强。在实验当中第三种方法效果最好。

3 本文方法

3.1 本文方法概述

下面简要介绍融合的过程：以两个不相交的 CNN 流 CNN_1 和 CNN_2 之间融合的最简单情况为例。令 $\mathbf{A} \in \mathbb{R}^{N_1 \times \cdots \times N_K \times C}$ 和 $\mathbf{B} \in \mathbb{R}^{M_1 \times \cdots \times M_L \times C'}$ 分别表示 CNN_1 和 CNN_2 中某一层的特征。这里，N_i 和 M_i 代表空间维度，C 和 C' 分别代表 CNN_1 和 CNN_2 中相应特征的通道数。MMTM 接收特征 A 和 B 作为输入，从中学习全局多模态，并重新校准输入特征。这中间的过程由两步完成，分别是多模式挤压和激励过程完成的。

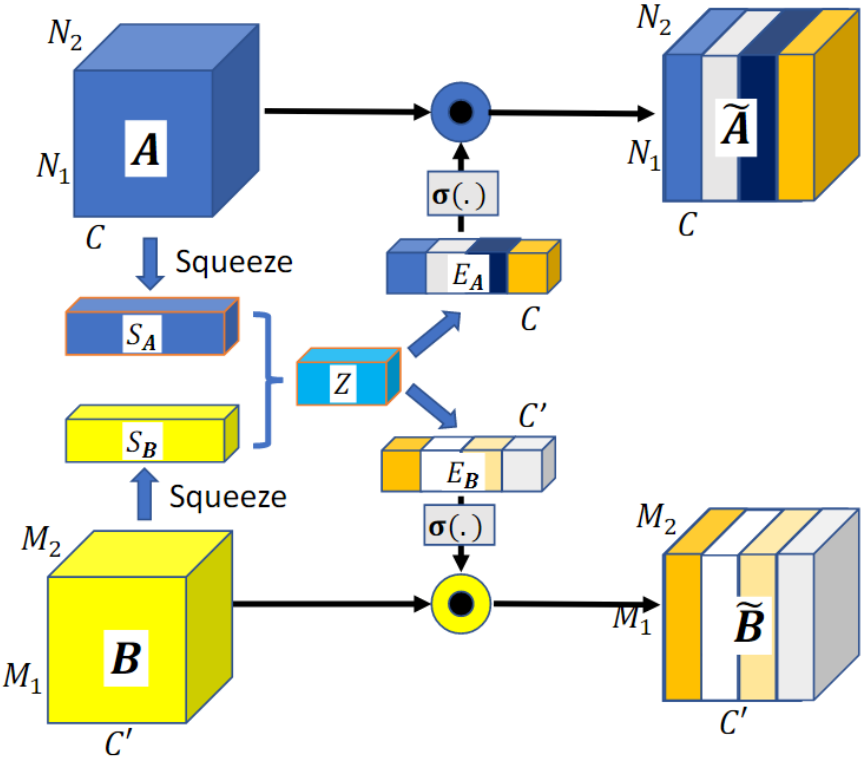


图 1: 两种模式的 MMTM 架构。A 和 B 表示两个单模态 CNN 某一层的特征，是模块的输入。为了更好的可视化，将它们的空间维数限制为 2。MMTM 使用挤压操作从每个张量生成全局特征描述符。通过使用连接和全连接层将两个张量映射到联合表示 Z。使用联合表示生成激励信号 E_A 和 E_B。最后，激励信号用于控制每个模态中的通道特征

3.2 挤压操作

首先通过对输入特征的空间维度进行全局平均池化，将空间信息压缩到通道描述符中。

$$S_{\mathbf{A}}(c) = \frac{1}{\prod_{i=1}^{K} N_i} \sum_{n_1,\ldots,n_K} \mathbf{A}\left(n_1,\ldots,n_K,c\right) (1)$$

$$S_{\mathbf{B}}(c) = \frac{1}{\prod_{i=1}^{L} M_i} \sum_{m_1,\ldots,m_L} \mathbf{B}\left(m_1,\ldots,m_L,c\right) (2)$$

挤压操作可以在具有任意空间维度特征的模态之间进行融合。虽然论文中使用简单的平均池化，但在此步骤中可以使用更复杂的池化方法。

3.3 多模态激励过程

该单元的功能是生成激励信号 $E_{\mathbf{A}} \in \mathbb{R}^{C}$ 和 $E_{\mathbf{B}} \in \mathbb{R}^{C'}$，可通过简单的门控机制用于重新校准输入特征 A 和 B:

$$\tilde{\mathbf{A}} = 2 \times \sigma\left(E_{\mathbf{A}}\right) \odot \mathbf{A} (3)$$

$$\tilde{\mathbf{B}} = 2 \times \sigma\left(E_{\mathbf{B}}\right) \odot \mathbf{B} (4)$$

其中 $\sigma(.)$ 是 sigmoid 函数，$\odot$ 是通道乘积运算。这允许抑制或激发每个流中的不同滤波器。MMTM 权重是正则化的，以便控制 E_A 和 E_B 接近于零。具体来说，增加 E_A 的正则化权重将门控信号 $2 \times \sigma(E_A)$ 推向更接近身份向量，限制了门控对特征 A 的影响。门控信号必须基于相同的输入表示将不同的校准权重应用于不同的模态。我们首先通过从压缩信号中预测联合表示 $Z \in \mathbb{R}^{Cz}$ 来实现这一点。

$$Z = \mathbf{W}\left[S_{\mathbf{A}}, S_{\mathbf{B}}\right] + b (5)$$

然后通过两个独立的全连接层预测每种模态的激励信号。

$$E_{\mathrm{A}} = \mathbf{W}_{\mathrm{A}} Z + b_{\mathrm{A}}, \quad E_{\mathrm{B}} = \mathbf{W}_{\mathrm{B}} Z + b_{\mathrm{B}} (6)$$

其中，$[\cdot, \cdot]$ 表示级联操作，$W \in \mathbb{R}^{C_Z \times (C+C')}$，$\mathbf{W_A} \in \mathbb{R}^{C \times C_Z}$，$\mathbf{W_B} \in \mathbb{R}^{C' \times C_Z}$ 为权重，$b \in \mathbb{R}^{C_Z}$，$b_A \in \mathbb{R}^{C}$，$b_B \in \mathbb{R}^{C'}$ 是全连接层的偏差。使用 $C_Z = (C + C')/4$ 来限制模型容量并增加泛化能力。为了融合两种以上的模态，我们通过连接等式 3 中所有模态的压缩特征来简单地推广这种方法，并使用等式 4 中的独立全连接层预测每种模态的激励信号。

4 复现细节

4.1 数据集与数据增强

NTU RGB + D

到目前为止，NTU RGB+D 数据集[17]是一个著名的大规模多模态数据集。它有两种推荐的评估协议，即 CrossSubject (CS) 和 Cross-View (CV)。在跨主题设置中，20 个主题的序列用于训练，其余 20 个主题的序列用于验证。在跨视图设置中，样本由相机视图分割。来自两个相机视图的样本用于训练，其余用于测试。它包含从 40 个主题捕获的 56, 880 个样本，在 80view 点执行 60 类活动。每个动作剪辑在 RGB 视频上最多包含两个人，在 3D 坐标空间上包含 25 个身体关节。我们遵循跨主题评估[17]，将 40 名受试者分成训练集和测试集。

4.1.1 深度图

NTU RGB+D 数据集[17]的作者建议使用掩码深度图，它是完整深度掩码的前景掩码版本，因此具有更好的压缩比，这有助于下载和文件管理。此外，为了能够专注于 HAR 问题并减少没有有用信息的像素数量，所有图像都被裁剪到动作发生的感兴趣区域，同时降低文件的权重。图 2 是这种图像裁剪的说明性示例。在训练时，为了实现数据增强，将裁剪后的掩码深度图调整大小为 (256×256)，然后随机裁剪为 (224×224) 大小，并且随机翻转，形成一个新的深度图序列。将处理后的深度图序列分为 8 个片段，在每个片段随机提取一个深度图。最后的数据大小就为 $(8 \times 224 \times 224)$。在测试时，进行中心裁剪, 并且取十个深度图序列片段，然后对预测结果进行平均求值。

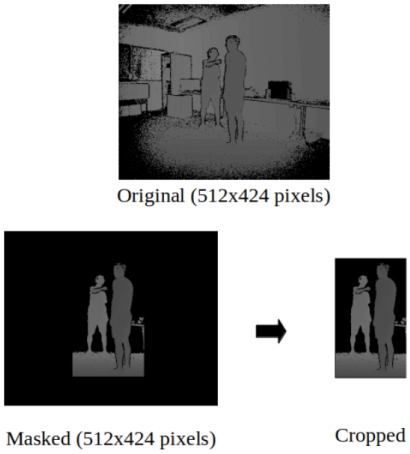


图 2: 全深度图 (上) 将掩模深度图裁剪成更小的图像 (下) 的过程。

4.1.2 骨架序列

在训练过程中，我从整个序列中随机裁剪一个子序列。种植比例来自于 [0.5,1] 之间的均匀分布。在测试过程中，我们以 0.9 的比率居中裁剪子序列。由于不同的动作持续不同的时间，输入序列被归一化为一个固定的长度 (在我的实验中为 32)，并沿着帧的尺寸进行双线性插值。

4.1.3 RGB 视频

在训练过程中，我在空间上和时间上都使用了随机裁剪——将较小的视频边调整为 256 像素，然后随机裁剪一个 224×224 的视频序列——当在足够早的时间内从这些补丁中选择开始帧以保证所需的帧数时。对于较短的视频，我们尽可能多次循环视频，以满足每个模型的输入界面。在训练期间，我们还一致地对每个视频应用随机左右翻转。在测试期间，将模型卷积应用于整个视频，居中裁剪 224×224 大小的视频序列，并对预测进行平均。

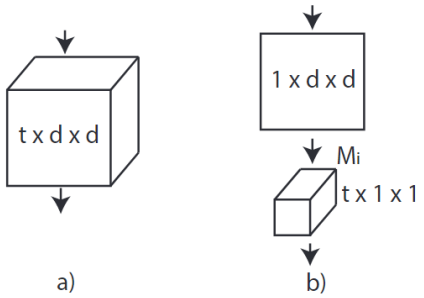
4.2 单模态动作识别 CNN 框架

4.2.1 基于深度图的 R(2+1) 模型

根据 Nuno C. Garcia et al.[15]在实验的比较中，设计了 RGB 视频,RGB 光流和深度图的残差类 3D 卷积网络，我选择其中的深度图部分作为 MMTM 实验的基线模型。深度图模态网络被实现为 R(2+1)D-18 架构模型[18]。该体系结构基于 Resnet-18 网络[4]，经过修改，使得在每个 2D 卷积之后添加一维时间卷积，从而使网络能够学习时空特征。将 3D 卷积分解为 2D + 1D 卷积的组合已被证明对视频分类任务更有效。在实验当中，我使用标准 SGD 优化器使用 0.9 的动量优化目标函数。我们从 10-2 的基本学习率开始，当损失饱和时将其降低 10 倍。r(2+1)d 模型结构如图 3 所示:

layer name	output size	R3D-18	R3D-34
conv1	L×56×56	3×7×7, 64, stride 1 × 2× 2	
conv2_x	L×56×56	$\begin{bmatrix} 3\times3\times3, 64 \\ 3\times3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3\times3, 64 \\ 3\times3\times3, 64 \end{bmatrix}\times3$
conv3_x	$\frac{L}{2}\times28\times28$	$\begin{bmatrix} 3\times3\times3, 128 \\ 3\times3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3\times3, 128 \\ 3\times3\times3, 128 \end{bmatrix}\times4$
conv4_x	$\frac{L}{4}\times14\times14$	$\begin{bmatrix} 3\times3\times3, 256 \\ 3\times3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3\times3, 256 \\ 3\times3\times3, 256 \end{bmatrix}\times6$
conv5_x	$\frac{L}{8}\times7\times7$	$\begin{bmatrix} 3\times3\times3, 512 \\ 3\times3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3\times3, 512 \\ 3\times3\times3, 512 \end{bmatrix}\times3$
	1×1×1	spatiotemporal pooling, fc layer with softmax	

(1)



(2)

图 3: (1) 卷积剩余块显示在括号中，旁边是每个块在堆栈中重复的次数。为过滤器和输出提供的维度依次为时间、高度和宽度。一系列的卷积以全局时空池化层为顶点，该层产生 512 维的特征向量。该向量被馈送到全连接层，该层通过 softmax 输出类概率。(2)(2+1)D 与 3D 卷积。给出了简化设置的说明，其中输入由具有单个特征通道的时空体积组成。(a) 使用大小为 t × d × d 的滤波器进行完整的 3D 卷积，其中 t 表示时间范围，d 是空间宽度和高度。(b) A (2+1)D 卷积块将计算拆分为空间 2D 卷积，然后是时间 1D 卷积。我们选择 2D 滤波器 (M_i) 的数量，以便我们的 (2+1)D 块中的参数数量与完整的 3D 卷积块的参数数量相匹配。

4.2.2 基于骨架序列的 HCN 模型

Chao Li et al.[10]提出一个双流网络，双流对应着骨骼节点与骨骼向量。不过与 Lei Shi et al.[9]不同的是，它是基于 CNN 网络的，而且还对骨骼节点的维度进行卷积，来学习提取节点之间的信息联系。通过研究卷积操作，将其分解为两个步骤，即跨空间域（宽度和高度）的局部特征聚合和跨通道的全局特征聚合。然后可以得出一种简单但非常实用的方法来按需调节聚合程度。将 T 表示为 $d_1 \times d_2 \times d_3$ 的 3D 张量。我们可以通过重组（转置）张量来分配不同的上下文。如果指定为通道，而其他两个编码本地上下文，则可以全局聚合来自维度 d_i 的任何信息。同时引入了骨骼运动的表示，并将其显式地输入到网络中。该框架旨在以端到端的方式联合学习联合共现和时间演化。

图 4 显示了所提出框架的网络架构。骨架序列 X 可以用 T × N × D 张量表示，其中 T 是序列中的帧数，N 是骨架中的关节数，D 是坐标维度（例如 3 用于 3D 骨架）。上述骨架运动与 X 的形状相同。它们作为输入流直接输入网络。两个网络分支共享相同的架构。但是，它们的参数不共享和单独学习。它们的特征图在 conv4 之后通过沿通道连接融合。给定骨架序列和运动输入，这些特征是分层学习的。在第 1 阶段，点级特征用 1 × 1 (conv1) 和 n × 1 (conv2) 卷积层进行编码。由于沿联合维度的内核大小保持 1，因此它们被迫独立地从每个关节的 3D 坐标中学习点级表示。之后，我们将特征图与参数 (0, 2, 1) 转置，以便将联合维度移动到张量的通道。然后在第二阶段，所有后续的卷积层从人的所有关节中提取全局共现特征。之后，特征图被展平成一个向量，并通过两个全连接层进行最终分类。

本次复现选择其骨干网络作为 MMTM 的基线模型。在实验当中，为了缓解过拟合的问题，我们在 conv4、conv5、conv6 和 fc7 之后添加 dropout, dropout 比为 0.5。fc7 的权值衰减为 0.001。我总共训练模型 1000 次迭代，最小批处理大小为 64 次。使用 Adam 优化器。学习率初始化为 0.001。

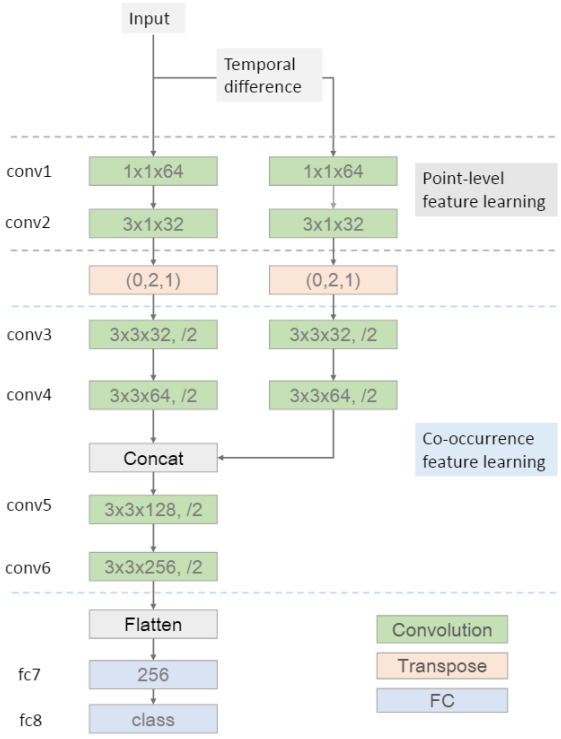


图 4: 提出的分层共现网络概述。绿色块是卷积层，其中最后一个维度表示输出通道的数量。尾随“/2”表示卷积后附加步幅为 2 的 MaxPooling 层。转置层根据顺序参数置换输入张量的维度。在 conv1、conv5、conv6 和 fc7 之后附加 ReLU 激活函数，以引入非线性。

4.2.3 基于 RGB 视频的 I3D 模型

Joao Carreira and Andrew Zisserman.[3]提出一种 2D 卷积膨胀到 3D 卷积的概念，增加一个时间维度的卷积, 然后利用 3D 卷积从 RGB 视频中提取出 3D 维度的空间与时间特征信息，提出 I3D 网络，将非常深的图像分类 ConvNet 的过滤器和池化内核扩展为 3D，从而可以从视频中学习无缝时空特征提取器。其模型结构图如图 5 。本次复现选择其 3D 卷积网络作为 MMTM 的基线模型去应用于基于 RGB 视频的动作识别任务。在实验当中，视频的训练都使用动量设置为 0.9 的标准 SGD，因为模型非常大，所以首先要在 ImageNet 和 Kinetics 数据集进行预训练，然后在 NTU-RGBD 数据集上进行训练。

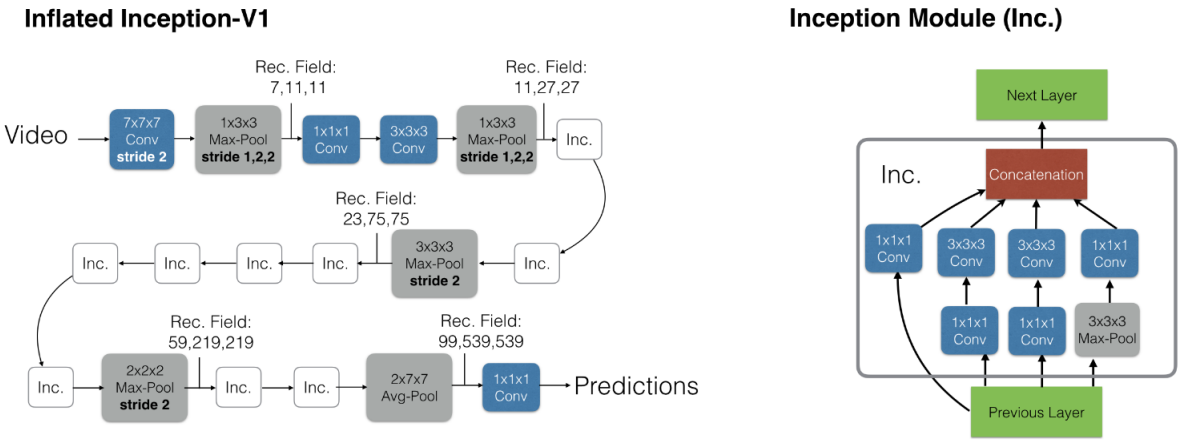


图 5: 膨胀的 Inception-V1 架构（左）及其详细的初始子模块（右）。没有指定的情况下，卷积和池化算子的步幅为 1，最后没有显示批量归一化层、ReLu 和 softmax。网络中几层的感受野大小的理论大小以“时间,x,y”格式提供——单位是帧和像素。预测是及时卷积获得的，并取平均值。

4.3 多模态框架设计

复现的论文当中仅仅设计了 RGB 视频模态的 CNN 模型和骨架序列模态的 CNN 模型的融合。深度图虽然抛弃了一部分的环境信息，但是他还是保留了一些全局的信息，避免了背景和光线的干扰。所以我还做了一个实验，将 MMTM 提到的融合方法应用到深度图的 R(2+1)D 模型和 HCN 当中，以验证 MMTM 的融合方法是否有效，而且还可以挖掘深度图与骨架序列之间有什么可以相互补充的优点。所有实验。我使用标准 SGD 优化器使用 0.9 的动量优化目标函数。从 10-2 的基本学习率开始，当损失饱和时将其降低 10 倍。其模型框架图如下图 6 所示。

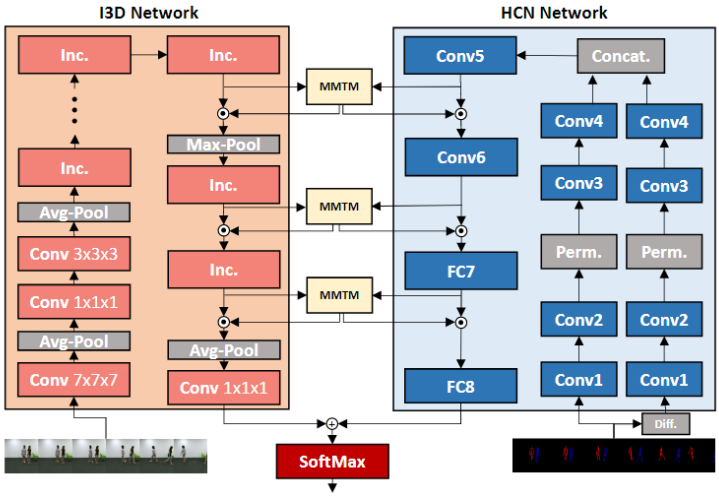


图 6: 提出的用于动作识别的多模态架构。每个“Inc.”块代表 Joao Carreira and Andrew Zisserman.[3]描述的初始模块。

4.4 与已有开源代码对比

MMTM: Multimodal Transfer Module for CNN Fusion 这篇论文作者没有公布论文相关的源代码。此处公布的源代码都是自己完成。

4.5 实验环境搭建

python3.9

pytorch1.12.1

Linux version 5.13.0-27-generic

$4 \times P100(GPU)$

4.6 创新点

尝试使用 (2+1)D 网络代替 3D 卷积；并且加入残差网络应用于深度图的动作识别任务当中；还探究 MMTM 融合模块对基于深度图的 R(2+1)D 模型和基于骨架序列的 HCN 模型的融合的影响。

5 实验结果分析

本实验要证明多模态融合带来的优势以及验证该 MMTM 模块的有效性。实验分为两个部分。

5.1 多模态的优势

多模态动作识别（multi-modal action recognition）可以利用多种输入模态（如视频、音频、加速度等）来识别动作，这样可以更准确地识别动作。相比单模态动作识别，多模态动作识别能更好的利用不同模态的信息来提高识别准确率。本次实验中包含三种模态，包括视频、深度图、骨架序列。这个实验部分主要论证多模态相对于单模态的优势。

Method	Input Modalities	Accuracy
HCN_{ours}	Pose	79.84
I3D	video	89.25
$R(2+1)D_{ours}$	depth	75.64
HCN+I3D	Pose+video	91.99
$HCN + R(2+1)D_{ours}$	Pose+depth	93.99

表 1：上部分是各个模态的 baseline。下部分是 MMTM 融合多模态的结果。

由表 1 可知，多模态的动作识别对比单模态的动作识别具有比较大的优势，进一步提高模型准确率。表 1 中的视频模态与骨架序列的结合是原论文中的结果，应为需要的时间过久，需要几年的时间，所以本地尚未运行出来。但是可以从深度图模态与骨架序列模态的结合也可以得知这个理论点是成立的。所以多模态动作识别可以利用多种输入模态——视频和骨架或深度图和骨架的结合，来识别动作，这样可以更精确地识别动作。因为多种模态中包含了不同的信息，如视频中包含了形态信息，深度图中包含了形态信息，摒弃图片背景，骨架序列中包含了动作的人体结构信息，这些信息能更好的识别动作，提高识别的准确性.

5.2 MMTM 的有效性

虽然上一个实验部分证明了多模态的融合有利于建立模型的优势，提高模型的准确率。但是并未证明 MMTM 融合模块的有效性。紧接着设计一组实验对比。就是利用两个模态的简单的决策级融合

与 MMTM 的融合方法进行比较。

Method	Input Modalities	Accuracy
$HCN + I3D + last_fusion$	Pose+video	91.56
$(HCN + R(2+1)D_{ours}) + last_fusion$	Pose+depth	90.74
HCN+I3D	Pose+video	91.99
$HCN + R(2+1)D_{ours}$	Pose+depth	91.45

表 2：决策级融合与 MMTM 融合

所以说明多模传输模块（MMTM），可以添加到特征层次的不同级别，从而实现慢模态融合。使用挤压和激励操作，MMTM 利用多种模态的知识来重新校准每个 CNN 流中的信道方向特征。与决策级融合方法不同，所提出的模块可用于具有不同空间维度的卷积层中的特征模态融合。模型的准确率比较高，证明了 MMTM 方法的有效性。

6 总结与展望

论文中提出了一个简单的神经网络融合模块，以利用卷积神经网络中多种模式的知识。该模块可以在特征层次结构的不同级别添加，允许缓慢的模态融合。对不同类型模式应用的实验表明，该模块在人类动作识别方面额多模态融合具有优秀的适用性。但是其适用于 CNN 网络，尚未证明在其他网络中具有适用性，而且其融合模块的模型结构也不一定适合所有网络。所以对于未来的发展，应该是拓宽适用的应用面，提高模型的特征选择能力。

参考文献

[1] CHEN L, WEI H, FERRYMAN J. A survey of human motion analysis using depth imagery[J]. Pattern Recognition Letters, 2013.

[2] SANCHEZ-CABALLERO A, de LÓPEZ DIZ S, FUENTES-JIMENEZ D, et al. 3DFCNN: Real-Time Action Recognition using 3D Deep Neural Networks with Raw Depth Information.[J]. arXiv: Computer Vision and Pattern Recognition, 2020.

[3] CARREIRA J, ZISSERMAN A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset [J]. computer vision and pattern recognition, 2017.

[4] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[J]. computer vision and pattern recognition, 2015.

[5] DUTA I C, LIU L, ZHU F, et al. Improved Residual Networks for Image and Video Recognition.[J]. International Conference on Pattern Recognition, 2020.

[6] SUN K, XIAO B, LIU D, et al. Deep High-Resolution Representation Learning for Human Pose Estimation[J]. computer vision and pattern recognition, 2019.

[7] GONG J, FAN Z, KE Q, et al. Meta Agent Teaming Active Learning for Pose Estimation[C]/ /. 2022.

[8] SHOTTON J, FITZGIBBON A, COOK M, et al. Real-time human pose recognition in parts from single depth images[J]. computer vision and pattern recognition, 2011.

[9] SHI L, ZHANG Y, CHENG J, et al. Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition[J]. computer vision and pattern recognition, 2019.

[10] LI C, ZHONG Q, XIE D, et al. Co-occurrence Feature Learning from Skeleton Data for Action Recognition and Detection with Hierarchical Aggregation[J]. international joint conference on artificial intelligence, 2018.

[11] CHI H G, HA H, CHI S, et al. InfoGCN: Representation Learning for Human Skeleton-based Action Recognition[J]., 2022.

[12] HUANG Q, ZHOU F, JIAKAI H, et al. Spatial–temporal graph attention networks for skeleton-based action recognition[J]. Journal of Electronic Imaging, 2020.

[13] CAI J, JIANG N, HAN X, et al. JOLO-GCN: Mining Joint-Centered Light-Weight Information for Skeleton-Based Action Recognition[J]. workshop on applications of computer vision, 2020.

[14] LI J, XIE X, PAN Q, et al. SGM-Net: Skeleton-guided multimodal network for action recognition[J]. Pattern Recognition, 2020.

[15] GARCIA N C, BARGAL S A, ABLAVSKY V, et al. Distillation Multiple Choice Learning for Multimodal Action Recognition[J]. workshop on applications of computer vision, 2021.

[16] NAGRANI A, YANG S, ARNAB A, et al. Attention Bottlenecks for Multimodal Fusion[J]. arXiv: Computer Vision and Pattern Recognition, 2021.

[17] SHAHROUDY A, LIU J, NG T T, et al. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis[J]. computer vision and pattern recognition, 2016.

[18] TRAN D, WANG H, TORRESANI L, et al. A Closer Look at Spatiotemporal Convolutions for Action Recognition[J]. computer vision and pattern recognition, 2017.