课程论文题目

3D Human Action Representation Learning via Cross-View Consistency Pursuit

摘要

在这项工作中，我们提出了一个基于无监督的三维骨骼动作表示的跨视图对比学习框架 (CrosS-CLR)，利用多视图互补监督信号。CrosSCLR 由单视图对比学习 (SkeletonCLR) 和跨视图一致性知识挖掘 (CVC-KM) 模块组成，以协作学习的方式集成。值得注意的是，CVC-KM 的工作方式是，根据视图之间的嵌入相似度交换高置信度的正/负样本及其分布，确保对比上下文 (即相似分布) 方面的跨视图一致性。大量的实验表明，在无监督设置下，CrosSCLR 在 NTU-60 和 NTU-120 数据集上取得了显著的动作识别结果，观察到更高质量的动作表示。

关键词：对比学习；无监督学习；自监督学习

1 引言

人类行为识别是计算机视觉研究中一项重要但具有挑战性的任务，由于姿态估计算法比较轻量化并且性能很鲁棒，3D 骨架已经成为研究人体动作动力学的流行的特征表示。然而标注数据是很浪费时间的，并且当前的无监督方法还没有探索不同骨架模式提供的丰富的监督内部信息。本文利用多视点互补的监督信号，提出了一种基于无监督 3D 骨架的动作表示的跨视图对比学习框架 (CrosSCLR)。CrosSCLR 由单视图对比学习和跨视图一致性知识挖掘两个模块组成，以协作学习的方式集成在一起。本工作的目的是为了通过并行学习单视图的骨架数据和跨视图挖掘有用的样本，使模型能够在无监督的情况下捕获更全面的表示。[1]

2 相关工作

2.1 自监督表示学习

自监督学习是从大量无标记数据中学习特征表示，通常通过代理任务产生监督，如拼图、着色、预测旋转。对于序列数据，目前存在的生成监督的方法高度依赖于代理任务的质量。最近，基于实例区分的对比方法被提出用于表示学习。但目前的方法都无法捕获用于 3D 动作表示的跨视图知识，并且不考虑对比上下文。该论文提出的 CrosSCLR 通过鼓励跨视图一致性，同时训练所有视图中的模型，从而获得更有代表性的嵌入。

2.2 基于骨架的动作识别

为了处理基于骨架的动作识别任务，早期的方法通常基于手工艺特征，最近的方法更多地关注深度神经网络。对于骨架数据的序列结构，许多基于 RNN 的方法被用来有效地利用时间特征。但 RNN 具有梯度消失性，因此，基于 CNN 的模型引起了研究人员的关注，但它们需要将骨架数据转换为另一种形式。近年来为了更好地对骨架数据的图结构进行建模，提出了 ST-GCN，可以基于人类关节位置的时间序列表示而对动态骨骼建模，并将图卷积扩展为时空图卷积网络而捕捉这种时空的变化关系。该论文采用广泛使用的 ST-GCN 作为主干来提取骨架特征。

2.3 无监督骨架表示

许多无监督方法被提出用于捕获视频中的动作表示。对于骨架数据，前人在无深度神经网络的无监督表示学习方面取得了一定进展。而最近的深度学习方法中无监督表示都高度依赖于重建和预测的精确性，而且它们没有利用骨架数据的自然多视图知识。因此，我们引入 CrosSCLR 来表示无监督的 3D 动作。

3 本文方法

CrossSCLR 包含两个关键模块：①SkeletonCLR：用于无监督学习单视图表示的对比学习框架。以及②CVC-KM：它将最突出的知识从一个视图传达给其他视图，引入互补的伪监督约束并促进视图之间的信息共享。最后，通过合作训练可以获得更具辨别力的表示。

3.1 Single-View 3D Action Representation

它是一种用于骨架表示的记忆增强对比学习方法，它将一个样本的不同数据增强方式视为其正样本，将其他样本视为负样本。在每个训练步骤中，将批量数据都存储在先进先出的内存中，以摆脱冗余计算，作为后续步骤的负样本。正样本彼此靠近嵌入，而负样本的嵌入被推开。

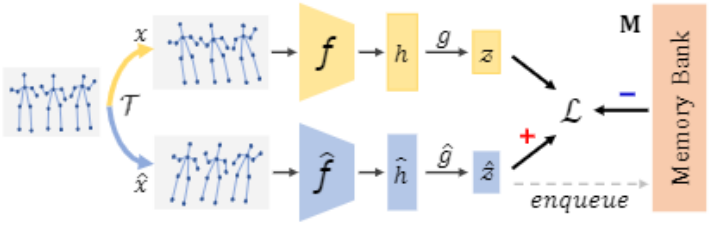

图 1: 单视图 SkeletonCLR 的架构，这是一个记忆增强的对比学习框架。

如图 1所示，SkeletonCLR 由以下主要组件组成：

- 一个数据增强模块 τ ，它随机地将给定的骨架序列转换为不同的增强版本 $x, \hat{x}$ 。它们被认为是一个正对。对于骨架数据, 我们采用 *Shear* 和 *Crop* 作为增强策略。
- 两个解码器 f 和 $\hat{f}$ 将 x 和 $\hat{x}$ 嵌入到隐藏空间：$h = f(x;\theta)$ 和 $\hat{h} = \hat{f}(\hat{x};\hat{\theta})$，$\hat{f}$ 是 f 的动量更新版本:$\hat{\theta} \leftarrow \alpha\hat{\theta} + (1-\alpha)\theta$, 其中 α 是一个动量系数。SkeletonCLR 使用 ST-GCN 作为编码器的主干。
- 一个简单的投影器 g 和它的动量更新版本 $\hat{g}$ ，将隐藏向量投影到一个低维空间: $z = g(h), \hat{z} = \hat{g}(\hat{h})$。投影器是具有 ReLU 的全连接（FC）层。
- 存储负样本的记忆库 $\mathbf{M} = \{m_i\}_{i=1}^{M}$ 以避免嵌入的冗余计算。它是一个先进先出的队列，每次通过 $\hat{z}$ 迭代更新一次。在每一个推断步骤之后, $\hat{z}$ 将入队，而 M 中最早的嵌入将出队。在对比学习的过程中，M 提供了大量的负嵌入，而新的计算 $\hat{z}$ 是正嵌入。
- 应用于实例区分的损失函数：

$$\mathcal{L} = -\log \frac{\exp(z \cdot \hat{z}/\tau)}{\exp(z \cdot \hat{z}/\tau) + \sum_{i=1}^{M} \exp(z \cdot m_i/\tau)} \tag{1}$$

其中 $m_i \in \mathbf{M}$,τ 是一个温度系数, 点积 $z \cdot \hat{z}$ 用于计算样本的两个增强版本间的的相似性。

受损失函数 $\mathcal{L}$ 的约束, 该模型被无监督地训练以区分训练集中的每个样本。最后，我们可以获得有利于提取单视图区分表示的强编码器 f。

3.2 Cross-View Consistent Knowledge Mining

我们提出了跨视图一致性知识挖掘 (Cross-View Consistent knowledge Mining, CVC-KM)，利用一个视图中样本的高相似性来指导另一个视图中的学习过程。它根据嵌入相似度挖掘跨视图的正对，促进视图之间的知识交换，从而提高每个视图中隐藏的正对的关联性，提取的骨架特征将包含多视图知识，从而形成更规则的嵌入空间，我们将跨两个视图进行对比学习挖掘一致性的损失函数定义如下：

$$\mathcal{L}_{v\to u} = -\log \frac{\exp\left(z^u \cdot \hat{z^u}/\tau\right) + \sum_{i\in N_+^v} \exp\left(s_i^u s_i^v\right)/\tau\right)}{\exp\left(z^u \cdot \hat{z}^u/\tau\right) + \sum_{i\in N} \exp\left(s_i^u s_i^v\right)/\tau\right)} \tag{2}$$

$$\mathcal{L}_{\text{cross}} = \mathcal{L}_{u\to v} + \mathcal{L}_{v\to u} \tag{3}$$

其中 $\mathcal{L}_{v\to u}$ 表示使用视图 v 中的语境监督视图 u 中的嵌入分布。

3.3 本文方法概述

我们使用并行 SkeletonCLR 模型和 CVC-KM 跨视图模块挖掘有用的样本，使模型能够在无监督的情况下捕获更全面的表示，如图 2所示：

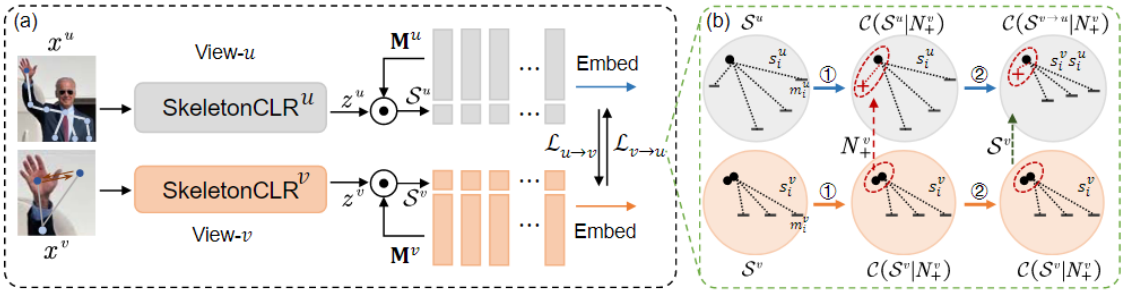

图 2: (a)CrosSCLR。给定由相同原始数据 (例如关节和运动) 生成的两个样本 x^u、x^v, SkeletonCLR 模型生成单视图嵌入，而跨视图一致性知识挖掘模块 (CVC-KM) 交换多视图互补知识。(b) $\mathcal{L}_{v\to u}$ 在嵌入空间中的作用。在步骤①，我们从相似点 S^v 中挖掘高置信知识 N_+^v，以增强视图 u 的正集，即 z^u 共享 z^v 的语境; 在步骤②，我们使用相似语境 S^v 来监督视图 u 中的嵌入分布。z^u, z^v 与其他视图共享相似关系。因此，在 $\mathcal{L}_{\text{cross}}$ 的约束下，两个嵌入空间变得相似。

大于两个视图时，CrosSCLR 的损失函数定义如下：

$$\mathcal{L}_{\text{cross}} = \sum_{u}^{U}\sum_{v}^{U} \mathcal{L}_{u\to v} \tag{4}$$

其中 U 是视图的个数并且 $u\neq v$。

4 复现细节

4.1 工作介绍

本文代码已开源，我的工作是在开源代码的基础上进行创新。

我们将目光移动到 SkeletonCLR 模块的编码器 f：论文采用 ST-GCN 作为编码器，理由是它适合于利用时空关系对图结构骨架数据进行建模，从而获得更高层的语义信息，ST-GCN 的图结构如图 3所示。

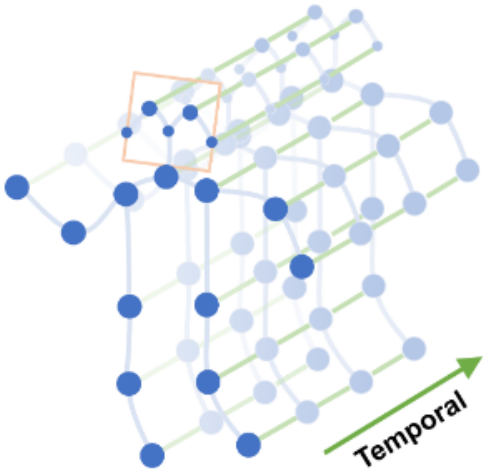

图 3: 骨骼序列的时空图，ST-GCN 在其上运行。蓝点表示身体关节。人体关节之间的身体内部边缘是基于人体的自然联系而定义的。帧间边连接连续帧之间的相同关节。关节坐标被用作 ST-GCN 的输入。

ST-GCN 分为在时间上和空间上进行图卷积，这里我们主要探讨在空间上的图卷积，在空间维度上的图卷积公式对于节点 V 其被定义为：

$$\mathbf{f}_{out} = \sum_{k}^{K_v} \mathbf{W}_k \left(\mathbf{f}_{in}\mathbf{A}_k\right) \odot \mathbf{M}_k \tag{5}$$

其中 **A** 代表由给定的人体骨架拓扑构建出来的邻接矩阵。**W** 是一个权重矩阵，在运动过程中，不同的躯干重要性是不同的。例如腿的动作可能比脖子重要，通过腿部我们甚至能判断出跑步、走路和跳跃，但是脖子的动作中可能并不包含多少有效信息。因此，ST-GCN 对不同躯干进行了加权（每个 ST-GCN 单元都有自己的权重参数用于训练）。**M** 是一个注意力图。用来表示每个节点的重要性。K_v 是空间维度里面的内核的大小。传统的 ST-GCN 采用空间型划分，将 **A** 划分为了三个子集，如图 4所示：

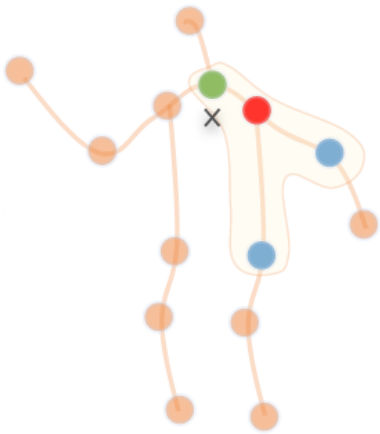

图 4: 空间构型划分，按节点与重心的距离进行划分，将节点的 1 邻域划分为 3 个子集，第一个子集连接了空间位置上比根节点更远离整个骨架的邻居节点，第二个子集连接了更靠近中心的邻居节点，第三个子集为根节点本身，分别表示了离心运动、向心运动和静止的运动特征。

我们的改进点将围绕 ST-GCN 空间维度上的图卷积公式以及邻接矩阵 **A** 进行。

4.2 实验环境搭建

本实验在 Linux 服务器上进行，所需的实验环境如下：

- Python == 3.8.2

- PyTorch == 1.7.0
- CUDA == 11.0

4.3 实验注意事项

在早期的训练过程中，如果没有标签的监督，模型不够稳定和强大，不能提供可靠的交叉视图知识。由于不可靠的信息可能会误入歧途，因此不鼓励过早启用跨视图交流。我们对 CrosSCLR 进行了两个阶段的训练：1) 模型的每个视图分别使用 SkeletonCLR 进行训练，而不需要进行跨视图通信，使用等式 (1) 作为损失函数。2) 经历 150 个 epoch，模型已经可以提供高置信度的知识，将损失函数替换为等式 (4) 开始进行跨视图的知识挖掘。

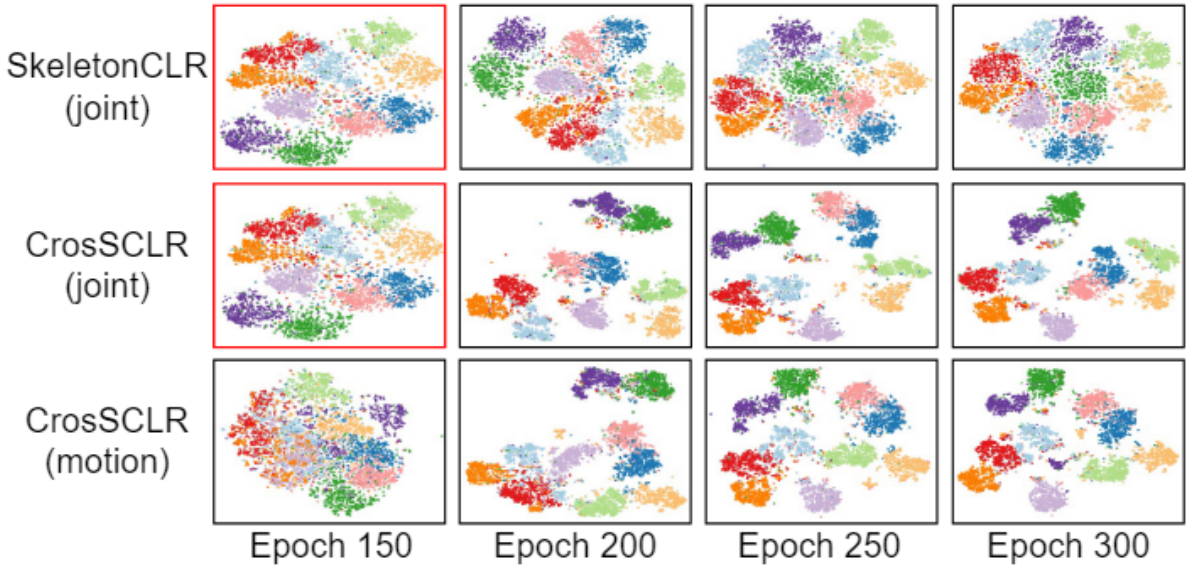

图 5: 预训练期间不同时期嵌入的 t-SNE 可视化。对来自 10 个类别的嵌入进行采样，并用不同的颜色进行可视化。对于 CrosSCLR，$\mathcal{L}_{\text{cross}}$ 在第 150 个 epoch 开始可用，因此其分布与 Epoch 150 之前的 SkeletonCLR 没有差异，如红色框所示。

4.4 改进点

经过分析，我们观察到了 ST-GCN 的一些缺点[2]：首先，ST-GCN 中使用的骨架图是预先设计好的，并且仅表示人体的物理结构。因此，不能保证它对于动作识别任务是最优的。其次，ST-GCN 上图形的拓扑在在所有的层都是固定的。所以 ST-GCN 缺乏对包含在所有层中的多级语义信息进行建模的灵活性和能力。因此，我设计出了两种方案来对 ST-GCN 进行改进：

- 我将等式 (5) 改进成如下形式[3]：

$$\mathbf{f}_{out} = \sum_{k}^{K_v} \mathbf{W}_k \mathbf{f}_{in} \left(\mathbf{A}_k + \mathbf{B}_k + \mathbf{C}_k \right) \tag{6}$$

A 代表两边之间是否连接，并且不会随训练过程更新参数，**B** 表示节点之间有无连接以及它们之间的连接强度，相对于 **A**，邻接矩阵 **B** 里面的值会和其他参数一样，随着训练而一起进行优化，让这个模型更加具有灵活性。**C** 是数据依赖图，它为每一个样本学习一个唯一的图，确认两个节点之间的相似性，因为两只手之间的关系对于识别诸如“鼓掌”和“阅读”之类的任务非常重要。然而，ST-GCN 很难捕捉到两只手之间的依赖关系，因为它们在预定义的基于人体的图中距离彼此很远。

- 从邻接矩阵 **A** 的划分入手，我们改进成中心节点自身是一个子集，从中心节点流出去的运动表示构成一个子集，流入中心节点的运动表示也构成一个子集，从全局考虑离心运动和向心运动，我们将 **A** 替换成了 $\mathbf{A}^*$，这样等式 (5) 就被我们改成如下形式，：

$$\mathbf{f}_{out} = \sum_{k}^{K_v} \mathbf{W}_k \left(\mathbf{f}_{in}\mathbf{A}_k^*\right) \odot \mathbf{M}_k \tag{7}$$

5 实验结果分析

我复现以及改进的实验结果如下：

Model	NTU 60 xview(%)	NTU 120 xview(%)
SkeletonCLR	68.3	76.4
3s-CrosSCLR	83.4	80.7
3s-CrosSCLR(Ours)	82.6	79.7
3s-CrosSCLR(A+B+C)	60.5	–
3s-CrosSCLR(A^*)	83.0	80.0

表 1: 第一行和第二行分别表示论文中 SkeletonCLR 和基于三个视图（关节，运动，骨骼）使用 CrosSCLR 的实验结果，第三行是我基于作者提供的源码对论文进行复现得到的结果，第四行是第一种改进方案得到的结果，第五行是第二种改进方案得到的结果。

我发现我的复现效果始终比不上原文给出的结果，这可能是由于训练过程中的一些细节我还未注意到。第一种改进方案失败的原因我暂时把它归咎于公式引入的参数过多导致了过拟合，因为它在训练集上的表现优于原文实验。第二种改进结果的性能较我们的复现结果是有上升的，这说明从全局考虑人体的运动对本任务是有效果的。

6 总结与展望

这次的复现工作使我受益匪浅，我对自己的研究领域有了进一步的认识，并且首次尝试了对顶会论文进行改进，这是一次很宝贵的体验。接下来我会认真思考第一种改进策略失败的原因并且争取让它成功，并且我将会思考该论文其他的改进方案，比如能不能引入更多的视图信息？为了利用骨骼数据的二阶信息，能不能考虑引入加速度视图来描述运动[4]。为了灵活地进行特征提取，能不能引入 Transformer[5]，因为 ST-GCN 中空间维度模块对全局拓扑中关节对之间的关联进行建模，放松了图卷积的约束，希望有效结合 Transformer 中的动态关注度和全局上下文，提高特征提取的灵活度。此外，合并 Transformer 允许将更多信息自然地引入我们的模型中。希望通过本次课程任务我能尽快的熟悉自己的研究领域，找到论文的方向。

参考文献

[1] LI L, WANG M, NI B, et al. 3D Human Action Representation Learning via Cross-View Consistency Pursuit[J]. computer vision and pattern recognition, 2021.

[2] YAN S, XIONG Y, LIN D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action

Recognition[J]. national conference on artificial intelligence, 2018.

[3] 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.

[4] WANG M, NI B, YANG X. Learning Multi-View Interactional Skeleton Graph for Action Recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[5] BAI R, LI M, MENG B, et al. GCsT: Graph Convolutional Skeleton Transformer for Action Recognition. [J]. international conference on multimedia and expo, 2021.