arXiv:2302.02752v1 [cs.CV] 6 Feb 2023

采用注意力机制的三维卷积神经网络：MediaEval 2022 体育任务的基线方法

皮埃尔-艾蒂安马丁

马克斯·普朗克进化人类学研究所 *CCP* 部门，德国莱比锡 *D-04103*

摘要

本文介绍了为 MediaEval 2022 基准测试中的体育视频任务提出的基线方法。该任务包含两个子任务：从剪辑后的视频中进行击球分类，以及从未剪辑视频中进行击球检测。此基线方法涵盖这两个子任务。我们提出了两种三维卷积神经网络架构来解决这两个子任务。两种三维卷积神经网络均采用时空卷积和注意力机制。我们针对各自要解决的子任务，对架构和训练过程进行了专门设计。此基线方法已在线公开，旨在帮助参与者开展研究，并减轻任务中某些环节的负担，例如视频处理、训练方法、评估和提交流程。对于分类子任务，我们的 v2 模型取得了 86.4% 的准确率。对于检测子任务，我们的 v1 模型取得了 0.131 的平均精度均值和 0.515 的交并比。

1. 引言

视频动作分类是计算机视觉领域的热门课题 [1, 2, 3, 4]。为了解决这类任务，研究者最初引入了二维卷积神经网络 [5, 6]。随后，为了更好地捕捉视频中的时间信息，三维卷积方法应运而生 [7, 8]。研究者还探索了从 RGB 流计算光流，以提升性能并将 RGB 变化转化为运动信息 [9, 10]。近年来，多模态方法再次受到关注，这次将 RGB 流与音频流结合起来 [11]，在 Kintetics 等常见基准数据集上达到了最先进水平600 [12]。多视角方法与 Transformer 相结合 [13]，也是许多动作分类数据集上的当前最先进水平 [14, 15]。

在 MediaEval 2022 体育任务中，重点是对视频中的乒乓球击球动作进行分类和检测。如 [16], 所述，该任务关注视觉上类间差异较小的动作：从剪辑后的视频中对其进行分类（子任务 1），并从未剪辑视频中检测这些动作（子任务 2）。该任务基于 TTStroke-21 数据集 [17]，与其他类间差异较小的数据集类似 [18, 19, 14, 20]。

这个基线可在 GitHub 上公开获取[1]，旨在处理两个子任务，并希望帮助参赛者他们的提交内容包括视频处理、标注文件和深度学习方法等方面。

MediaEval'22：多媒体评估研讨会，*2023*年*1*月*13*日至*15*日，卑尔根（挪威）及线上
皮埃尔_艾蒂安_martin@eva.mpg.de (P. Martin)
www.eva.mpg.de/comparative-cultural-psychology/staff/pierre-etienne-martin (P. Martin) 0000-0003-0503-4580 (P. Martin)
© 2022 年本文版权归作者所有。允许依据知识共享署名 4.0 国际许可协议（CC BY 4.0）使用。
讨会论文集（CEUR-WS.org）
[1]https://github.com/ccp-eva/SportTaskME22

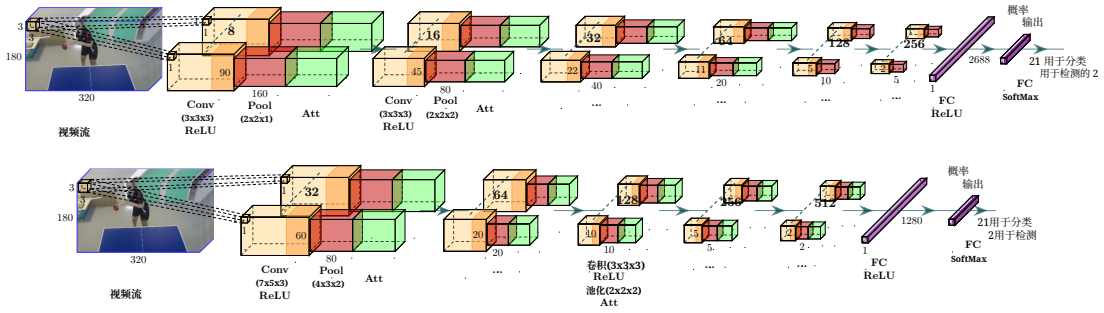


图 1： 采用注意力机制进行击球分类与检测的三维卷积神经网络 v1（上）和 v2（下）。

2. 方法

该方法保持简单，仅使用所提供视频中的 RGB信息。实现参考了 [21]。主要区别在于不使用根据光流值计算出的感兴趣区域（ROI）。数据处理十分简单：将 RGB帧调整为宽度 320，并堆叠在一起，形成长度为 96 的张量；这些帧取自剪辑后的视频，或依据 XML文件中提供的标注边界提取。通过数据增强提高数据的多样性：从不同的时间点开始，并进行空间变换（翻转和旋转）。

本文介绍了两个版本 V1 和 V2，如图 1所示。V1 由四个卷积+池化+注意力层和两个卷积+池化层组成。所有卷积层均使用 3x3x3 滤波器。第一层使用 2x2x1 池化滤波器（时间域不进行池化），其余层使用 2x2x2 池化滤波器。V2 由五个卷积+池化+注意力层组成。前两个模块的卷积滤波器尺寸为 7x5x3，池化滤波器尺寸为 4x3x2。其余模块分别使用 3x3x3 卷积滤波器和 2x2x2 池化滤波器。V2 在第二个模块之后生成近似正方形的特征图，从而能够在全连接层之前更好地融合水平、垂直和时间维度。

训练方法采用 Nesterov动量，并进行固定数量的训练轮次。学习率会根据损失的变化情况进行调整 [21]。保存验证损失表现最佳的模型。两个子任务采用相同的训练方法。目标函数是对经过 softmax函数处理的输出在批次上求和后得到的交叉熵损失：

$$\mathcal{L}(y, class) = -log(\frac{exp(y'_{class})}{\sum_i^N exp(y_i)}) \tag{1}$$

与 [22] 中的做法相同，我们为分类任务考虑 21 个类别，为检测任务考虑两个类别。我们为检测任务提取负样本，并在其测试集上构建负候选片段。检测任务的测试采用裁剪后的候选片段（以一个居中的窗口或滑动窗口，并结合若干后处理方法），或者在整个视频上运行滑动窗口。随后对最新的输出进行处理，以逐帧分割笔画。短于 30 帧的笔画不予考虑。在分类任务上训练的模型也会用于检测任务测试，且不再使用检测数据进行额外训练。我们考虑了两种方法：1）负类得分与所有其他类别得分进行比较

表1分类子任务中模型的准确率性能

模型	无窗口	Vote	Mean	高斯
V1	.847	.839	.856	.856
V2	.856	.822	.831	**.864**

表2测试集上使用候选框进行检测子任务的模型平均精度均值 |交并比

模型	无窗口	Vote	Mean	高斯
V1	.111 |.358	.114 | .360	。113 | **.365**	.113 | .361
V2	.111 |.322	**.118** |.329	.117 | .333	.117 | .331

以此进行判定；2）负类得分与所有其他类别得分之和进行比较。我们还根据时间窗口测试了几种判定方法：无窗口、投票、均值和高斯。更多详情请参见 [23]。

3. 结果

本节根据 [16] 中介绍的指标，分别展示各子任务的结果。对于这两个子任务，我们使用 0.0001 的学习率、0.5 的动量和 0.005 的权重衰减训练模型 2000 个训练轮次。

3.1. 子任务 1 - 击球分类

如表 1所示，V1 和 V2 在击球分类子任务上的表现相近，但使用高斯窗口决策的 V2 表现最佳，在测试集上的准确率为 86.4%。该模型在第 815 个 epoch 时收敛，训练集和验证集的准确率分别为 .989 和 .813。此次运行的混淆矩阵见图 2。

从混淆矩阵中可以看出，模型倾向于将一些笔画分类为非笔画（负类）。这无疑是由于负类中的差异较大，扩大了其对应的潜在空间，使未见样本更有可能被归入其中。可以通过数据增强提高这些样本的多样性，或增加对这些笔画的录制数量来解决这一问题。

3.2. 子任务 2 - 击球检测

表2 报告了使用测试集中的视频候选片段所得的结果。视频候选片段是从测试视频中提取的长度为150帧、彼此不重叠的连续样本。主要评估指标是平均精度均值，因此，使用投票决策的模型V2表现最佳。然而，以这种方式提取视频候选片段并不能高效地检测笔画。因此，表3报告了使用其他分割方法得到的结果。

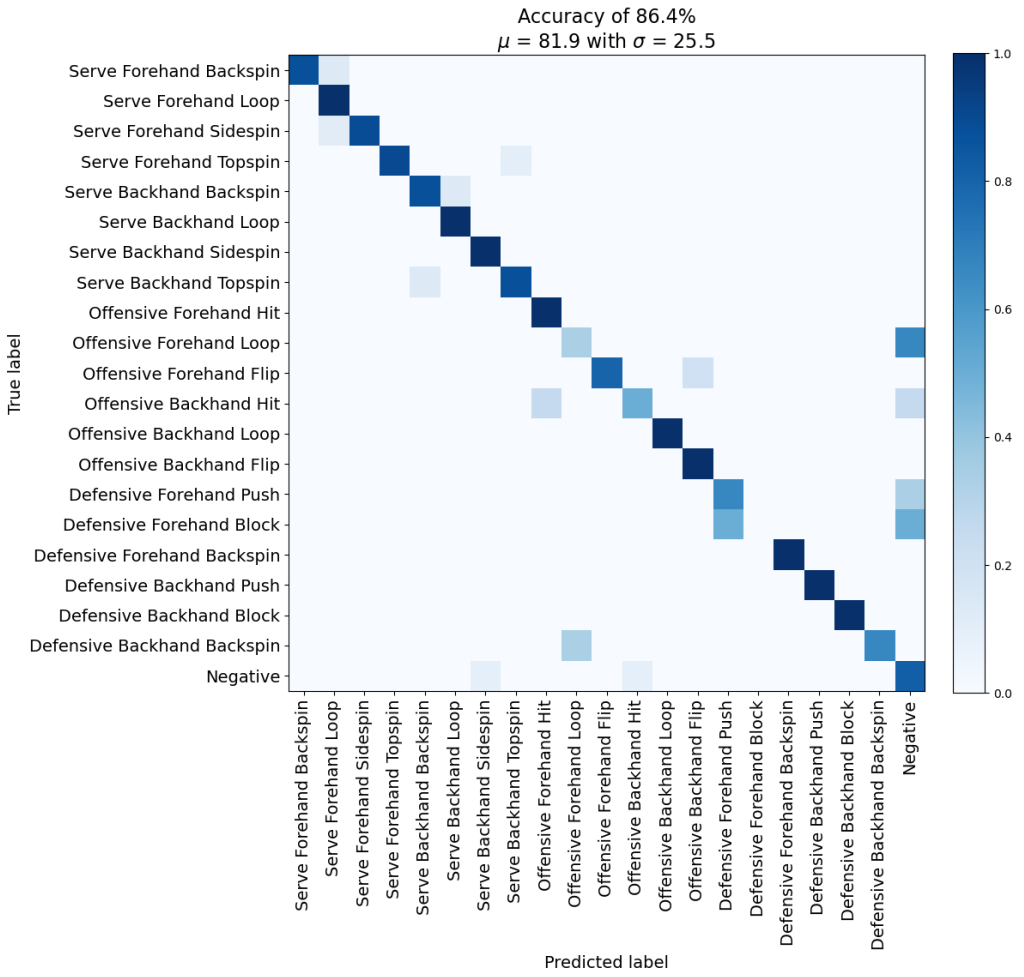


图 2： 测试集上最佳分类运行的混淆矩阵。

表 3采用滑动窗口分割的检测子任务中模型的平均精度均值 | 交并比

模型	Vote	Mean	高斯
V1	**.131 |.515**	.00201 | .341	.00227 | .33
V2	.000731 | .308	.102 | .473	.1 | .466

为了实现更好的分割，我们在测试视频上使用步长为1的滑动窗口。随后，按照前文介绍的相同窗口方法合并输出，以作出决策。子任务1中的模型也进行了测试。

如我们所见，分割方法使模型 V1 在平均精度均值和交并比方面取得了最佳表现。然而，V2 模型并非如此。

4. 结论

该基线旨在帮助参与者完成体育视频任务。这项工作延续了去年的基线 [22] ，并按照上一届的建议，实现了更多工具来帮助参与者。结合子任务 1 中的知识，可以改进子任务 2 的解决方案。此外，还可以改进数据增强和损失，以平衡样本分布不均的问题。最后，用于击球检测的分割方法仍有改进空间，从而提升该子任务的表现。这些改进方案或可在明年的基线中实现。

参考文献

[1] K. Soomro、A. R. Zamir、M. Shah，UCF101：包含野外视频中101类人类动作的数据集，CoRR abs/1212.0402（2012）。[2] C. Gu、C. Sun、D. A. Ross、C. Vondrick、C. Pantofaru、Y. Li、S. Vijayanarasimhan、G. Toderici、S. Ricco、R. Sukthankar、C. Schmid、J. Malik，AVA：时空定位原子视觉动作的视频数据集（2018），6047–6056。[3] A. Li、M. Thotakuri、D. A. Ross、J. Carreira、A. Vostrikov、A. Zisserman，ava-kinetics 局部化人类动作视频数据集，CoRR abs/2005.00214（2020）。[4] A. J. Piergiovanni、M. S. Ryoo，Avid 数据集：来自不同国家的匿名视频，载于：H. Larochelle、M. Ranzato、R. Hadsell、M. Balcan、H. Lin（编），《神经信息处理系统进展33：神经信息处理系统2020年年度会议》，NeurIPS 2020，2020年12月6日至12日，线上，2020年。网址：https://proceedings.neurips.cc/paper/2020/hash/c28e5b0c9841b5ef396f9f519bf6c217-Abstract.html。[5] H. Bilen、B. Fernando、E. Gavves、A. Vedaldi，使用动态图像网络进行动作识别，《IEEE模式分析与机器智能汇刊》40（2018）2799–2813。[6] K. Simonyan、A. Zisserman，用于视频动作识别的双流卷积网络，载于：NIPS，2014年，第568–576页。[7] H. J. Kim、J. S. Lee、H. S. Yang，使用改进的卷积神经网络进行人类动作识别，载于：ISNN（2），《计算机科学讲义》第4492卷《计算机科学讲义》，Springer，2007年，第715–723页。[8] T. Lima、B. J. T. Fernandes、P. V. A. Barros，使用三维卷积神经网络进行人类动作识别，载于：LA-CCI，IEEE，2017年，第1–6页。[9] J. Carreira、A. Zisserman，Quo vadis，动作识别？一种新模型与 Kinetics 数据集，载于：CVPR，IEEE计算机学会，2017年，第4724–4733页。[10] P. Martin、J. Benois-Pineau、R. Péteri、J. Morlier，使用孪生时空卷积神经网络进行体育动作识别：在乒乓球中的应用，载于：CBMI，IEEE，2018年，第1–6页。[11] R. Zellers、J. Lu、X. Lu、Y. Yu、Y. Zhao、M. Salehi、A. Kusupati、J. Hessel、A. Farhadi、Y. Choi，Merlot Reserve：通过视觉、语言和声音获取多模态神经脚本知识，载于：CVPR，2022年。[12] J. Carreira、E. Noland、A. Banki-Horvath、C. Hillier、A. Zisserman，关于 Kinetics-600 的简短说明，CoRR abs/1808.01340（2018）。[13] M. Dehghani、A. Gritsenko、A. Arnab、M. Minderer、Y. Tay，Scenic：用于计算机视觉研究及其他领域的 JAX 库，载于：《IEEE/CVF计算机视觉与模式识别会议论文集》（CVPR），2022年，第21393–21398页。[14] D. Damen、H. Doughty、G. M. Farinella、S. Fidler、A. Furnari、E. Kazakos、D. Moltisanti、J. Munro、T. Perrett、W. Price、M. Wray，扩展自我中心视觉：EPIC-KITCHENS 数据集，CoRR abs/1804.02748（2018）。[15] M. Monfort、C. Vondrick、A. Oliva、A. Andonian、B. Zhou、K. Ramakrishnan、S. A. Bargal、T. Yan、L. M. Brown、Q. Fan、D. Gutfreund，Moments in Time 数据集：用于事件理解的百万级视频，《IEEE模式分析与机器智能汇刊》42（2020）502–508。[16] P. Martin、J. Calandre、B. Mansencal、J. Benois-Pineau、R. Péteri、L. Mascarilla、J. Morlier，体育任务：从视频中对乒乓球击球进行细粒度动作检测与分类，用于MediaEval 2022，载于：MediaEval，CEUR 研讨会论文集，CEUR-WS.org，2022年。[17] P. Martin、J. Benois-Pineau、R. Péteri、J. Morlier，使用双路时空卷积神经网络进行细粒度体育动作识别，《多媒体工具与应用》79（2020）20429–20447。[18] D. Shao、Y. Zhao、B. Dai、D. Lin，FineGym：用于细粒度动作理解的层次化视频数据集，载于：CVPR，IEEE，2020年，第2613–2622页。

[19] Y. Li, Y. Li, N. Vasconcelos，RESOUND：迈向无表征偏差的动作识别，载于：ECCV (6)，第 11210 卷，《计算机科学讲义》，Springer，2018年，第 520–535 页。[20] S. Noiumkar, S. Tirakoat，光学运动捕捉在体育科学中的应用：高尔夫挥杆案例研究，载于：ICICM，2013年，第 310–313 页。[21] P. Martin, J. Benois-Pineau, R. Péteri, J. Morlier，双路时空卷积神经网络中的三维注意力机制：在乒乓球比赛视频动作分类中的应用，载于：ICPR，IEEE计算机学会，2021年。[22] P. Martin，MediaEval 2021 体育视频任务的时空CNN基线方法，载于：MediaEval，CEUR 研讨会论文集，CEUR-WS.org，2021年。[23] P. Martin，基于时空卷积神经网络的视频细粒度动作检测与分类：乒乓球应用。（Détection et classification fines d’actions à partir de vidéos par réseaux de neurones à convolutions spatio-temporelles. Application au tennis de table），博士论文，拉罗谢尔大学，法国，2020年。网址：https://tel.archives-ouvertes.fr/tel-03128769。