基于轻量级神经网络的暴力视频分类方法研究

李 娜[1] 王迎超[1] 秦立浩[1] 黄 涛[1] 李佳乐[1]
LI Na WANG Yingchao QIN Lihao HUANG Tao LI Jiale

摘 要 随着视频类自媒体平台的迅速发展，视频内容的安全性审核需求急速增加。为提高视频内容审核的便捷性和准确性，提出结合轻量级神经网络和长短时记忆网络的暴力视频分类方法。借助轻量级神经网络提取视频帧的视觉特征，使用长短时记忆网络对视频帧之间的时序特征将进行提取，引入时空注意力机制提高视频分类准确率。实验结果表明，所提出的方法不仅在模型上表现出轻量性，而且还能提高视频分类的准确性。

关键词 轻量级神经网络；暴力视频分类；时空注意力机制；双向长短期记忆网络

doi：10.3969/j.issn.1672-9528.2024.03.049

0 引言

视频分类技术是一种通过计算机视觉和机器学习技术，将输入的视频数据分为不同的类别或标签。其目标是让计算机能够自动理解和识别视频内容，从而实现视频内容的智能分类和检索。近年来，随着自媒体和互联网的飞速发展，视频内容的复杂性使得视频分类技术也取得显著的发展。常见的视频分类技术包括深度学习[1-2]和迁移学习。前者应用较为广泛，但是需要大量的标注数据并且模型较大导致所需资源较多。后者由于使用较少的监督信息，难以泛化到其他领域，容易产生过拟合的现象。所以，视频分类过程中如何使用较少的数据进行模型的训练，同时考虑模型的轻量化具有一定的研究意义。

1 基于 2D 卷积与 BiLSTM 视频分类模型结构设计

常用的视频分类模型结构包括双流级联分类结构和卷积神经网络结合长短时记忆网络分类结构，卷积神经网络（convolutional neural network，CNN）用于从视频帧中提取空间特征，而长短时记忆网络（long short-term memory，LSTM）用于处理视频帧之间的时序关系，通过融合特征从而实现视频分类。该方法的缺点是用于提取空间特征网络通常较为复杂，参数量大，不适合嵌入在移动端使用。本研究采用轻量级 2D 卷积网络（MobileNetV3-Small 网络）作为视频帧特征提取网络，在降低模型参数量的情况下，不损失模型的分类精度。借助注意力机制降低无关信息干扰的优势，将其融入原架构以提升视频分类精度。该模型整体架构被分为三个主要部分，即数据输入层、隐藏层和输出层。其中，隐藏层由三个部分组成，分别是 MobileNetV3-Small 层、BiLSTM 层和引入注意力机制层。

该算法利用注意力机制和循环神经网络相结合，在进行实验验证时，展现出对暴力视频分类方面的良好性能特点。与其他算法相比，该算法的准确率较高，分类效果好，同时算法的参数量也相对较小，对计算资源的要求不高，具有一定的理论研究和实际应用价值。其整体框架如图 1 所示。

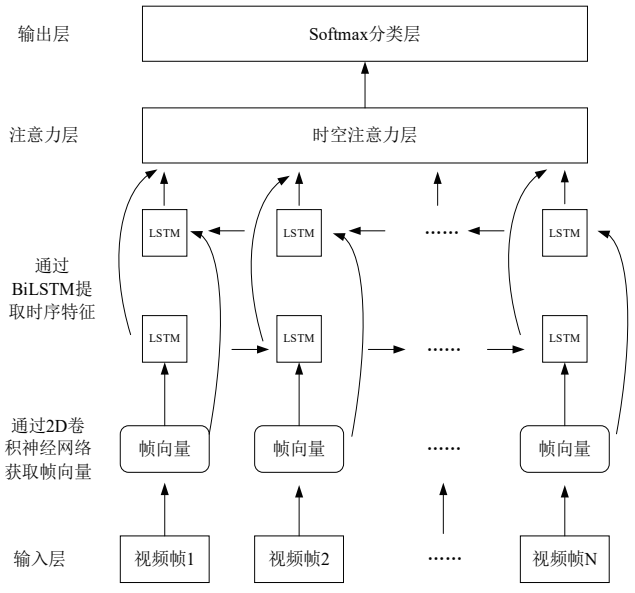


图 1 算法整体框架图

1.1 输入层

输入层的作用主要是对实验中采用的数据集进行处理，将其处理为2D卷积网络层可以接受并进行处理的数据形式。

1. 新疆科技学院信息科学与工程学院 新疆库尔勒 830091
新疆科技学院校级科研项目（2023-KYPT29）；新疆科技学院大学生创新创业训练计划项目（X2024135661102）

（1）数据读取及处理：提供三种数据读取方式：即时视频流的数据读取、指定视频文件的读取和视频数据集中的批量视频文件的读取。循环读取视频中的帧并进行预处理，预处理方式主要是通过随机旋转、翻转图像和调整图像大小等。

（2）归一化处理：使用函数 preprocess_input 对输入进行归一化处理，以适应所使用的参数的标准（RGB 色彩空间的均值减去每个像素，并将结果除以每个像素的标准差），从而有助于模型更快更好地收敛。

（3）特征编码：使用 np_utils.to_categorical 函数将视频类别标签进行 One-Hot 编码。通过 One-Hot 编码将每个类别转换为一个向量，其中向量的长度等于分类的数量。标签的值在向量中的相应位置上被设置为 1，而所有其他位置都设置为 0。此外，One-Hot 编码可以减低计算复杂度、提高模型训练速度。

（4）模型训练：数据集按比例划分，将数据经由上述操作后，就可将数据转换为可输入隐藏层的特征向量形式。

1.2 隐藏层

（1）MobileNetV3-Small 层

MobileNetV3 是一种用于移动设备和嵌入式系统的轻量级神经网络，是 MobileNet 系列中最新的一种。相对于 MobileNetV2，MobileNetV3 在准确性、速度和效率上都有一定的提升。

由于视频是由一帧一帧有序帧图像组成的，因此需要一种方法来提取每个帧的特征向量，而 MobileNetV3-Small 采用了 ECA 模块，该模块基于 SE 模块，通过对通道注意力的引入来提高模型效率，同时对计算密集型的神经网络进行了优化，使其可在移动设备上运行，提高了实时性。所以采用该模型作为视频帧的视觉特征提取器。

（2）BiLSTM 层

BiLSTM 在 LSTM 的基础上引入前向和后向机制，形成一个双向循环神经网络，能够在处理序列数据时捕捉到不同时间步间的依赖关系，从而更好地对时序信息进行建模，缓解梯度消失问题。

在视频分类任务中，每个时间步对应视频中的每一帧图像，BiLSTM 能够在视频层面上捕捉到丰富的时空结构信息，从而提升分类准确率。同时，由于视频序列的长度可能不一致，因此在实验中还实现了对序列长度的变换，使得每个视频的特征序列在长度上保持一致，增加模型的通用性。

（3）注意力机制层

注意力机制有助于加强模型对于序列数据中每个时间步的依赖关系进行建模，从而使得模型更好地理解视频帧序列中的时空结构。

在训练阶段，利用不同的处理方式，让模型更好地关注序列数据中的重要信息。其中一种机制是在 LSTM 的输出结果上进行的注意力计算，通过计算每个时间步输出结果的重要性，让模型集中于更具有判别性的部分，而忽略无关紧要的部分，从而提高分类准确率。

1.3 输出层

整体模型的输出层是全连接层（Dense layer），用于将模型学习到的特征映射到每个类别的分类概率上，最后通过 Softmax 函数对视频的每个类别进行预测，输出具有最高概率的类别作为预测结果。

2 实验结果和分析

2.1 实验数据集

为验证提出的算法在实际应用中的有效性，本研究采用两个公开的暴力视频数据集：Hockey Fight 数据集和 Violent Flows 数据集。Hockey Fight 数据集包含视频片段来自曲棍球比赛中发生的暴力事件，其特点是背景相对单一。相比之下，Violent Flows 数据集则是从 Youtube 上收集的多人参与的暴力行为视频，该数据集因为参与者和环境的多样性导致场景更加复杂。为适应实际需求，在进行实验时对数据集进行处理。实验将在两个不同类型的数据集上完成，具体信息如表 1 所示。

表 1 数据集描述

数据集	视频帧率	分辨率	视频数量	行为类别
Hockey Fight	25 帧 /s	360×288	1000	暴力 / 正常
Violent Flows	25 帧 /s	320×240	246	暴力 / 正常

2.2 结果评估指标

在将通过 BiLSTM 网络提取的特征信息经过注意力模块处理后，将融合更多信息的视频特征映射到分类空间后，通过 Softmax 分类器对输入的视频进行分类，并输出分类结果。在实验中，采用的损失函数为交叉熵损失函数，其计算公式为：

$$P_{Loss} = -\sum_{i=1}^{m} y_i \cdot \log \hat{y}_i \tag{1}$$

式中：y_i 为标签值，$\hat{y}_i$ 为预测值。

在进行分类时采用评价指标：准确率（accuracy），其计算表达式为：

$$P_{Accuracy} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \times 100\% \tag{2}$$

式中：T_P 代表模型正确地预测为正例的样本数，T_N 代表模型正确地预测为反例的样本数，F_P 代表模型错误地预测为正例的样本数，F_N 代表模型错误地预测为反例的样本数。

2.3 实验环境及参数设置

实验框架是建立在中国移动的九天深度学习平台上，主要使用 TensorFlow 2.0 的工具创建卷积神经网络架构和提取视频帧序列的特征信息。

为保证实验的公平性，将所有数据集按照 8:2 分割为训练集和测试集。训练过程中，对 Hockey Fight 和 Violent Flows 数据集分别设置 30 次的迭代次数，根据训练情况调整

批处理大小，其中 Hockey Fight 数据集的批处理大小为 64，Violent Flows 数据集上调整为 32。为降低损失值，采用交叉熵损失函数计算模型的损失值，并使用 adam 优化算法对模型参数进行优化。

2.4 消融实验结果及分析

本文使用 Xception[3]、MobileNetV2[4]、MobileNetV3-Small[5] 和分别作为视频帧序列的空间特征提取器进行对比实验。之所以选取以上几种网络是因为在精度不降低的情况下，该相比于同类型实验中常用的 VGG16 网络参数量分别减少了约 6 倍、39 倍、84 倍，计算量减少了约 3.7 倍、101 倍、189 倍。

为了验证注意力机制对整体模型中的分类效果的影响，Xception、MobileNetV2 和 MobileNetV3-Small 分别作为视频帧序列的空间特征提取器不变的条件下，在构建的网络结构中通过添加注意力机制模块进行消融实验，对比未添加注意力机制模块的网络结构，分别得出注意力机制模块对模型训练后准确率的影响。具体数值如表 2 所示。其中 × 代表未添加注意力机制，√代表添加注意力机制。

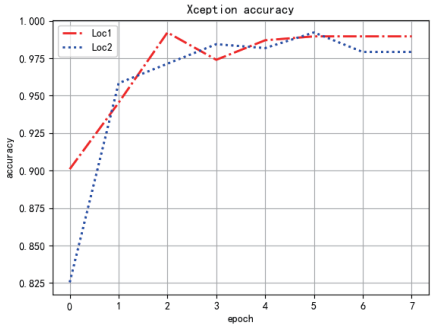
表 2 添加注意力机制前后对比

空间特征提取网络	参数量/MB	Hockey Fight/%		Violent Flows/%	
		×	√	×	√
Xception	22.8	96.22	97.7	94.2	95.9
MobileNetV2	3.5	95.39	95.9	94.2	96.0
MobileNetV3-Small	1.6	97.8	97.89	97.3	97.6

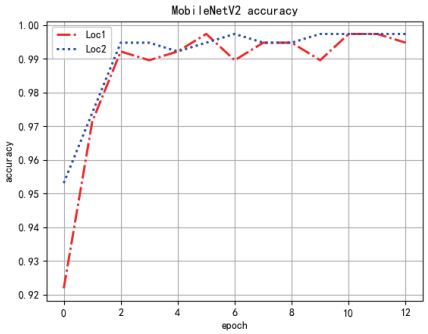
基于表 2 的数据，在不同的空间特征提取网络下，通过在整体网络结构中集成注意力机制模块，都可以在暴力视频分类任务中获得不同程度上的性能提升。在 Hockey Fight 数据集上，当网络架构中集成了注意力机制模块时，以 Xception、MobileNetV2 和 MobileNetV3-Small 为空间特征提取器的模型在准确率方面也分别提升了 0.09%~1.48% 不等。在 Violent Flows 数据集上，当网络架构中集成了注意力机制模块时，以 Xception、MobileNetV2 和 MobileNetV3-Small 为空间特征提取器的模型在准确率方面都分别提升了 0.3%、1.7%、1.8% 及 0.3%。

基于 2D 卷积与 BiLSTM 模型在基础架构的基础上作出调整，将注意力机制模块集成在不同位置进行模型的训练，其中 Loc1 代表位置在 2D 卷积神经网络与 BiLSTM 之间，Loc2 代表位置在 BiLSTM 之后。如图 2 所示，在不同 2D 卷积网络与 BiLSTM 结合的整体模型开始训练时准确率较低。随着训练的进行，模型逐渐学习到更好的特征表示，在训练的前几个时期，其准确率都有一个较大地提升。

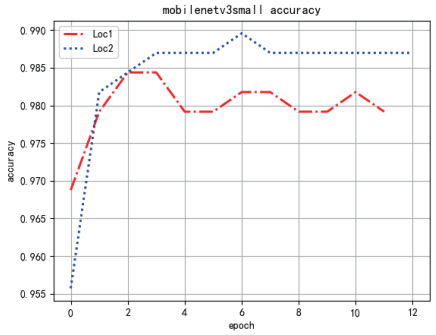
相较于 Loc2 位置，当注意力机制位于 Loc1 位置上时，其准确率到达峰值的速度普遍较快。除 Xception 作为 2D 卷积特征提取网络外，其余整体网络将注意力机制模块集成在 Loc2 位置的准确率均要高于在 Loc1 位置的准确率。



（a）Xception



（b）MobilenetV2



（c）MobilenetV3-Small

图 2 注意力机制位置对模型性能影响

根据以上分析，MobileNetV3-Small 模型表现出了在参数量大量减少的情况下仍能提高准确率的特征。在比较相近的参数量下，提出的整体网络模型取得了比其他轻量级模型更高的准确率。这表明，采用 MobileNetV3-Small 模型可在较小参数量的前提下，有效地分类暴力视频。综上所述，以 MobileNetV3-Small 网络作为空间特征提取网络的同时集成注意力机制模块，能够使得模型在实现高效率同时保持较高准确率的优势。此外，以 MobileNetV2、MobileNetV3-Small 为 2D 卷积网络的整体模型准确率均高于 Loc1 位置上的其他网络模型，其中，以 MobileNetV3-Small 为 2D 卷积特征提取网络的整体模型训练表现较为优越，在 Loc1 和 Loc2 位置上的准确率均高于其他网络。

2.5 与其他暴力分类算法对比实验

为了进一步验证 M-AttBiLSTM 网络模型的分类效果，将所提出的视频分类方法在 Hockey Fight 与 Violent Flows 数据集上与大多数现有的暴力分类算法工作进行对比实验，包括手工设计特征和深度学习方法。对于不同的数据集，用于比较的各类算法实现的准确率从 76.83% 到 98% 不等，可以看出与 Hockey Fight 数据集相比，Violent Flows 数据集更具有挑战性，因为它包含拥挤的人群和复杂的环境背景。通过对各类算法在两个数据集上的分类效果进行分析，对比结果如表 3 所示。

表 3 与现有分类方法对比

单位：%

算法名称	Hockey Fight	Violent Flows	准确率差值	平均准确率
Spatiotemporal Encoder[6]	96	98	2	97
ResNet50-LSTM[7]	89.5	91.4	1.9	90.45
Trajectory-pooled CNNs[8]	98.6	92.5	6.1	95.55
本文算法	97.89	97.6	0.29	97.745

针对表 3 中的实验结果，本文所提算法在 Hockey Fight 数据集和 Violent Flows 数据集上的平均准确率最高，与其他算法的平均准确率相比提高了 0.7%~14.68% 不等，基于包含拥挤场景和非拥挤场景的 2 个数据集训练得到的分类精度都得到了一定程度的改进。由此，可以得出结论，在不拥挤的场景下，仅通过捕获空间特征和短的时间特征就能获得较高的分类精度。

相较于现有的分类算法，采用轻量级网络 MobileNetV3-Small 作为空间特征提取器提取视频帧序列特征，使得在参数规模较小的情况下，高效实现视频分类。相较于手工提取特征的算法，本文提出的方法在 Hockey Fight 数据集上平均提高了 9% 的分类准确率，在 Violent Flows 数据集上平均提高了 14% 的分类准确率。相较于基于深度学习的算法，在 Hockey Fight 数据集上，该算法的分类准确率仅低于 Trajectory-pooled CNNs 算法 0.71%，在 Violent Flows 数据集上能到达最优。两个数据集上的准确率差距最小仅为 0.29，说明该算法对数据集的影响较低，具有很好的鲁棒性。在两个数据集上的平均准确率最高可达 97.745%，表明本文提出的算法在这两个数据集上均取得了分类效果的提升，并在资源有限的情况下能有效地进行模型的训练，验证了本文提出算法的有效性。

3 结语

暴力视频分类在迅速发展的自媒体平台中具有重要的意义，基于深度学习的视频分类模型参数量大难以在移动端中应用。本文提出基于轻量级神经网络的视频分类方法，引入注意力机制优化特征信息建模，实现轻量级的暴力视频分类。实验证明，该方法在现有数据集上具有一定的优越性。

参考文献：

[1] 孔祥魁，樊翠红．多特征融合和最小二乘支持向量机的运动视频图像分类研究 [J]. 南京理工大学学报，2022, 46(2): 164-169+176.

[2] 吴晓雨，顾超男，王生进．多模态特征融合与多任务学习的特种视频分类 [J]. 光学精密工程，2020,28(5):1177-1186.

[3] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Piscataway: IEEE, 2017: 1251-1258.

[4] SANDLER M, HOWARD A, ZHU M, et al. Mobilenetv2: inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway: IEEE,2018: 4510-4520.

[5] HOWARD A, SANDLER M, CHU G, et al. Searching for mobilenetv3[C]//Proceedings of the IEEE/CVF international conference on computer vision. Piscataway: IEEE, 2019: 1314-1324.

[6] GAO Y, LIU H, SUN X, et al. Violence detection using oriented violent flows[J]. Image and vision computing, 2016, 48: 37-41.

[7] MABROUK A B, ZAGROUBA E. Spatio-temporal feature using optical flow based distribution for violence detection [J]. Pattern recognition letters, 2017, 92(9): 62-67.

[8] MAHMOODI J, SALAJEGHE A. A classification method based on optical flow for violence detection[J]. Expert systems with applications, 2019, 127: 121-127.

【作者简介】

李娜（1997—），女，山西长治人，硕士，助教，研究方向：计算机视觉。

王迎超（1997—），通信作者，男，河南驻马店人，硕士，助教，研究方向：图像处理。

秦立浩（1995—），男，江苏宿迁人，助教，硕士，研究方向：图像处理。

黄涛（2003—），男，江西宜春人，本科，研究方向：数据分析。

李佳乐（2003—），男，陕西咸阳人，本科，研究方向：计算机应用

（收稿日期：2023-08-20）