计算机科学
COMPUTER SCIENCE

基于两阶段算法的多媒体有害信息识别方法

史晓苏, 李欣, 简玲, 倪华健

引用本文

史晓苏, 李欣, 简玲, 倪华健. 基于两阶段算法的多媒体有害信息识别方法[J]. 计算机科学, 2024, 51(6A): 231000052-6.

SHI Xiaosu, LI Xin, JIAN Ling, NI Huajian. Multimedia Harmful Information Recognition Method Based on Two-stage Algorithm [J]. Computer Science, 2024, 51(6A): 231000052-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231000052

基于两阶段算法的多媒体有害信息识别方法

史晓苏[1,2]　李　欣[1]　简　玲[2]　倪华健[3]
1 中国人民公安大学信息网络安全学院　北京 100091
2 上海市公安局网络安全保卫总队　上海 200025
3 上海闪马智能科技有限公司　杭州 310000
(903323721@qq.com)

摘　要　在互联网安全监管和网络违法犯罪打击整治的应用场景中,现有多媒体有害信息识别方法普遍存在运算效率不高、无法准确识别局部敏感信息,以及识别检测局限于单一的网络违法犯罪类型等问题。针对以上问题,文中提出了一种基于两阶段算法的多媒体有害信息识别模型。该模型将信息过滤与内容检测分阶段处理,将场景识别和元素目标检测分任务并行处理,第一阶段采用 EfficientNet-B2 构建高吞吐的前置过滤模块快速筛选掉 80%正常内容的数据;第二阶段基于 Meal-V2,Faster RCNN,NetVLAD 网络构建 3 种不同网络结构的模块,适应多维度场景、多特征元素的识别要求。结果表明,模型运算效率在 T4 卡上达到 57FPS,多媒体有害信息的识别准确率、召回率均超过 97%;与传统模型相比,在 NPDI 和自建测试集上识别准确率分别最高提升 3.09%和 19.26%。

关键词:两阶段算法;多媒体;有害信息识别
中图分类号　TP391.41

Multimedia Harmful Information Recognition Method Based on Two-stage Algorithm

SHI Xiaosu[1,2],LI Xin[1],JIAN Ling[2] and NI Huajian[3]
1 Information Network Security Academy,People's Public Security University of China,Beijing 100091,China
2 Network Security Corps,Shanghai Public Security Bureau,Shanghai 200025,China
3 Shanghai SUPREMIND Technology Co.,Ltd,Zhejiang 310000,China

Abstract　In the application scenarios of Internet content security supervision and combating and rectifying Internet crimes,existing multimedia harmful information identification methods generally have problems such as low computational efficiency,inability to accurately identify local sensitive information,and identification capabilities are limited to a single type of cyber crimes. In order to solve the above problems,the paper proposes a multimedia harmful information recognition model based on a two-stage algorithm. This method processes information filtering and content detection in stages,and splits the tasks of scene recognition and element target detection. The first stage uses EfficientNet-B2 to build a high-throughput pre-filter model to quickly filter out 80% of images and short videos with normal content. In the second stage,three modules with different network structures are built based on Meal-V2,Faster RCNN,and NetVLAD networks to adapt to the recognition requirements of multi-dimensional scenes and multi-feature elements. The results show that the model's computing efficiency reaches 57FPS(frames per second) on the T4 card,and the recognition accuracy and recall rate of multimedia harmful information exceed 97%. Compared with traditional models,the recognition accuracy rate on the NPDI dataset and the self-built test dataset increases by 3.09% and 19.26% respectively.

Keywords　Two-stage algorithm,Multimedia,Harmful information recognition

随着互联网、网络流媒体和多媒体信息检索技术的发展,人类社会进入了多媒体大数据时代,图像、短视频已经成为当前信息交流和服务的主流信息媒体。互联网在给网民带来便利的同时,其内容安全也成为了政府部门、行业主管单位、互联网企业面临的一大棘手问题。这些有害的图像视频煽动性极强,危害性极大,若不及时处理,互联网将会成为滋生违法犯罪的温床,对国家安全、社会安全以及青少年的健康成长均有较大的影响。随着计算机图像技术的飞速发展,基于机器学习技术提取特征进行识别检测的方法引起了学术界的广泛重视,此类方法需要手动设定图像特征,利用机器学习分类方法进行检测,但此方法对提取特征的重要性和准确性要求较高,一旦重要特征未被提取到,将直接影响识别的性能和结果,识别的效率和准确率难以保证。

随着深度学习技术在互联网内容安全检测领域的广泛应用,多媒体有害信息识别能力有了较大的提升,广大学者主要采取多任务多特征融合和多网络结合两种方式进行有害信息

基金项目:公安部应用创新计划(2020YYCXSHSJ019)
This work was supported by the Project of Applied Innovation Plan of the Ministry of Public Security of China(2020YYCXSHSJ019).
通信作者:李欣(lixin@ppsuc.edu.cn)

识别。Huang 等[1]提出了融合多任务卷积网络，每个任务分别检测一个特定的敏感部位，但其不足之处在于无法识别非敏感部位的隐蔽信息。Connie 等[2]构建了 8 个不同结构的神经网络分别对图像分类，通过对分类结果计算权重来进行融合，该模型在成人内容的识别准确率方面优于单一的 CNN 模型和 CNN 模型的平均和，但其因结构复杂导致运算效率较慢，模型吞吐量偏低。Moustafa[3]结合经典的 AlexNet 和 GoogleNet 两个网络进行有害信息识别，但该方法在识别局部敏感信息方面有待提高。为解决局部检测效果较差的问题，Xie[4]提出了一种基于全局分类及局部敏感信息分类的图像检测方法，在一定程度上提高了检测的准确率，但其缺点在于训练的时间成本和劳动成本很高，难以快速识别经过剪辑加工变种后的敏感信息文件。Zhang[5]提出了一种基于三维卷积网络和极限学习机的暴力视频检测算法，虽然将三维卷积网络和特征分类器结合进行暴力内容识别具有一定的可行性，但整体检测准确率仍有待提高，且检测主要集中在肢体动作方面，无法准确识别出含有危险元素的有害信息文件。

在互联网安全监管和网络违法犯罪打击整治的应用场景中，上述传统深度学习多媒体有害信息识别方法存在着运算效率不高、无法准确识别局部敏感信息，进而导致整体吞吐量较低、准确率不足的问题。同时，随着网络违法犯罪日趋多样化，有害信息数据也表现出类型发散、二义性高等特点，上述传统方法一般仅适用于识别某一类违法犯罪的有害信息，难以同时覆盖多类型的有害信息识别需求。

Gu 等[6]提出了一种互联网内容安全检测过滤系统，包含网络层、信息识别层、信息流过滤层、内容检测层 4 层，在信息流过滤层通过多特征融合判定对信息进行过滤和格式标准化，在内容检测层通过模式匹配检测是否含有特定的内容。借鉴这一思路，本文基于多媒体内容识别应用场景，简化了网络层和信息识别层，改进了信息过滤层和内容检测层的结构，在信息过滤层利用轻量级网络加快整体运行速率，在内容检测层融合多个子模块结果进行识别判断，根据数据的吞吐量、数据分布特性和性能指标要求来设计算法模型的层级连接和模型容量，最终提出一种基于两阶段算法的多媒体有害信息识别模型，整体结构设计如图 1 所示。第一阶段实现信息过滤，基于 EfficientNet-B2 构建的高性能、高召回的前置过滤模块过滤掉大部分正常内容的多媒体数据，降低整体训练成本，提升识别效率。由于前置过滤模块的输入是图像流，无法直接处理短视频数据，按照恒定量采样策略对视频截帧采样，即每个短视频均随机采样 20 帧，不足 20 帧的使用重复补全进行处理，再同图像数据一起进入前置过滤模块进行过滤。第二阶段实现内容检测，精准捕获关键局部信息并做分类，提高识别的准确率。内容检测的视频和图像流的数据管道是相互独立的。图像检测算法模型包含有害图像场景识别模块和敏感元素识别模块，经过前置过滤后的图像数据并行进入这两个模块。短视频分类算法模型包含敏感元素识别模块和有害视频场景识别模块，其中敏感元素识别与图像数据共用同一个模块，场景识别采用视频分类算法，最后对两个模块逻辑融合后输出有害信息识别结果。

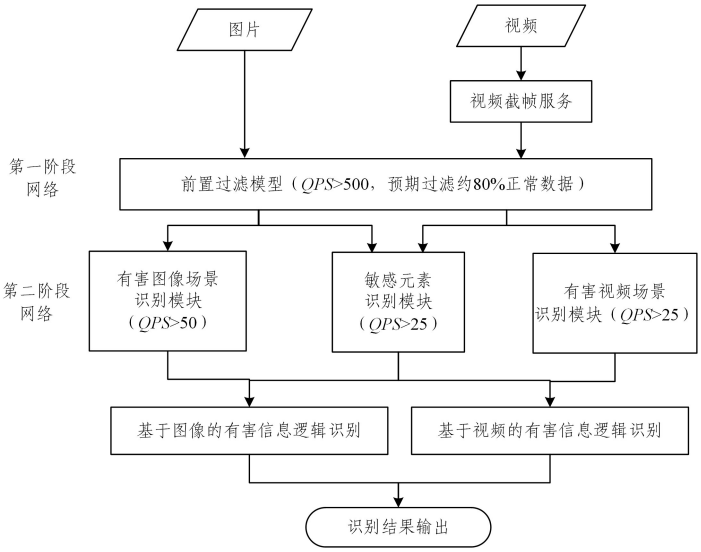

图 1 整体模型结构设计图

Fig. 1 Overall model structure design drawing

1 第一阶段前置过滤

前置过滤模块要尽可能过滤掉正常内容，召回疑似有害信息的内容。由于互联网图像的形态分布极为分散，需要采用一种能灵活调整的网络结构，来适应模型训练中数据量不断累加的一个过程。ResNet[7]具有多种不同深度的网络结构，EfficientNet[8]也是一种可扩展的网络结构，适合作为第一阶段过滤算法的主框架。ResNet 网络主要通过残差网络的堆叠增加网络深度，来达到提升卷积神经网络精度的目的。ResNet50 网络不仅小巧、复杂度低，而且在精度和速度上效果较好，常用于计算机视觉任务。EfficientNet 主要以 MB-Conv 结构为最小搜索单元，通过混合缩放的方式，合理化配置网络深度、宽度和图像输入分辨率这 3 个维度，提高网络整体的效果，不断放大网络的深度、宽度和分解率，最终从 B0 拓到 B7 结构中选取了 EfficientNet-B2。

将 ResNet50 网络和 EfficientNet-B2 网络进行微调训练，并设置对照实验。为了防止第一阶段漏检含有有害信息的数据，在训练过程中，不断下调 softmax 下的域值，牺牲一定的正常内容召回率，以保证有害信息的召回。当有害信息召回率达到 99.68%时，比较正常内容的召回率和有害信息识别

的准确率(见表 1)，最终确定以 EfficientNet-B2 网络构建前置过滤模块，期望在保证有害信息数据召回率不低于 99%的前提下，过滤超过 80%的正常数据，降低第二阶段模块识别的性能压力。

表 1　ResNet50 与 EfficientNet-B2 模型对照训练结果

Table 1　Comparison of training results between ResNet50 and EfficientNet-B2 models

Method	LR SCHEDULE	MIXUP	CUTMIX	(PRE,RECALL)
SE-RESNET50	[15,25,35]	False	False	(18.69%,77.56%)
SE-RESNET50	[15,25,35]	True,$alpha$=0.2	False	(19.53%,78.75%)
SE-RESNET50	[15,25,35]	False	True,$alpha$=0.5	(19.81%,79.12%)
EFFICIENTNET-B2	[15,25,35]	False	True,$alpha$=1.0	(20.06%,79.44%)
EFFICIENTNET-B2	cosine	False	True,$alpha$=1.0	(21.22%,80.84%)

2　第二阶段内容检测

2.1　有害图像场景识别

该模块采用神经网络结构搜索的方式，将搜索出的最佳网络作为 baseline 基准网络，并在基准网络基础上融入知识蒸馏[9]的训练方案(见图 2)，利用 3 种表达能力差异较大的 Teacher 模型做集成，使得 Student 模型能学习到不同特性的 Teacher 模型的表达能力，从而既提升算法的泛化能力，又提升推理性能。

首先，采集、定义、整理训练的有害信息场景数据集。分别用 DeiT-B 384[10]，FixEfficientNet-B8[11]，ResNeXt-101，32×48d[12]3 个算法框架在场景数据集上训练 Teacher 网络。DeiT-B 384 是 Transformer 结合蒸馏的方式训练得到的一种图像分类 Transformer 网络，FixEfficientNet-B8 通过增大分辨率对网络最后几层进行微调，可以解决数据增强导致训练和测试的物体分辨率不一致的问题，ResNeXt-101，32x48d 采用弱监督学习方法，用 hashtag 作为标签进行预训练，然后用学到的 Representation 在目标任务上微调，从而提升模型泛化能力。上述 3 个模型能分别从注意力机制、训练的后处理和弱监督预训练 3 个方向来提升 Teacher 网络的模型表达能力。

然后，采用 ProxylessNAS[13]算法在基于类 MobileNet-V2 的搜索空间超网中搜索出 Student 网络。ProxylessNAS 能够直接在目标数据集上进行网络结构搜索而不需要使用小的代理数据集，大大缩减了搜索时间成本，搜索出来的网络架构如表 2 所列。

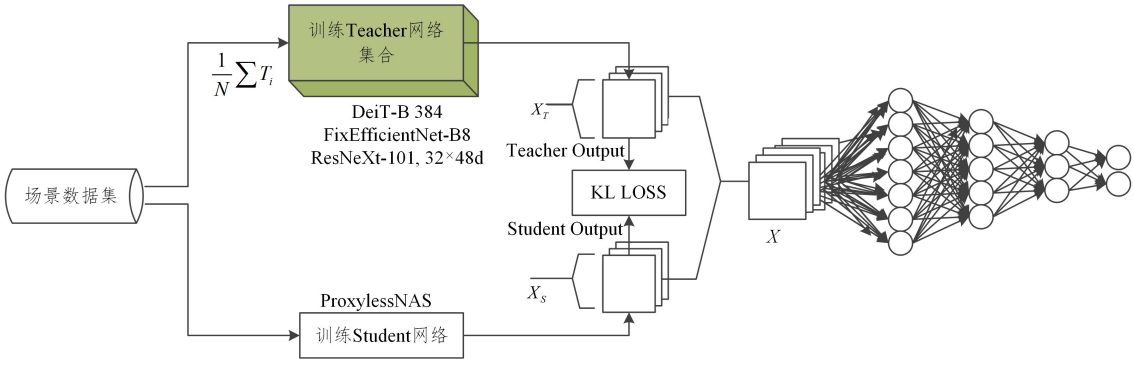

图 2　知识蒸馏算法训练有害图像场景识别模块

Fig. 2　Using knowledge distillation algorithm to train harmful image scene recognition module

表 2　Student(proxylessnas)网络结构

Table 2　Student (proxylessnas) network structure

Stage	Operator	Resolution	Channels	Layers
1	Conv3 * 3	224 * 224	3	1
2	MBConv3 * 3 MBConv5 * 5	112 * 112	16	2
3	MBConv3 * 3 MBConv5 * 5	56 * 56	32	4
4	MBConv5 * 5	28 * 28	40	6
5	MBConv7 * 7	14 * 14	80	8
6	MBConv7 * 7	7 * 7	192	8
7	fc	1 * 1	1024	1

最后，采用 Meal-V2[14]知识蒸馏方法对 Student 网络进行 finetune。Meal-V2 方法采用了所有 Teacher 模型的平均概率作为监督信息进行蒸馏，用 KL 散度度量 Student 模型的预测概率与 Teacher 模型的平均预测概率之间的相似性，最终得到 Student-Net(MealV2)网络。训练精度相比蒸馏前提升了 4%(见表 3)。

表 3　蒸馏前后训练精度对比

Table 3　Comparison of training accuracy before and after knowledge distillation

(%)

网络结构	训练精度
DeiT-B 384	96.98
FixEfficientNet-B8	95.30
ResNeXt-101,32x48d	95.81
Student-Net(proxylessnas)	90.98
Student-Net(MealV2)	94.95

2.2　敏感元素识别

敏感元素识别属于目标检测任务，由于有害多媒体数据中的敏感元素较为隐蔽且占据面积较小，因此，正负样本存在一定的不均衡性，相对于识别效率来说，对模型的识别精度要求更高。基于此，敏感元素识别模块在两阶段的 RCNN 上改造实现，选取 Faster RCNN[15]和 CascadeRCNN 作为 basline 模型进行改造，主干网络块结构采用 ResNet34[16]的残差结构，同时考虑到敏感元素的形态、尺度、姿态差异巨大，又加入

了 Deformable 卷积来有效学习目标的形态和姿态变换，并配合模型结构化剪枝、模型量化等方法，提升整体的推理效率。由于图标、旗帜、字符等敏感元素往往面积较小，容易隐藏在图像的角落中，不易被发现，小物体占据像素比例有限，分辨率较低，提取到的信息会在下采样过程中逐渐减弱。因此引入了 FPN 网络[17]，通过自上而下和横向连接的网络结构把深层的语义信息带入到浅层特征图，可以有效增强浅层特征图的语义信息，提升小物体检测效果，保证能够识别不同尺度的敏感元素目标。

最终经过对照组实验（见表 4），确定了敏感元素识别模块的网络结构为 Faster_rcnn_dconv_c3-c5_ResNet34_ fpn_1x（见图 3）。

表 4 改进版 CascadeRCNN 模型与改进版 Faster R-CNN 模型的对照训练结果

Table 4 Comparative training results of the improved CascadeRCNN model and the improved Faster RCNN model

BACKBONE 模型	TRAIN SIZE	LR STEP	ANCHOR RATIOS	FPN STAGE	MAP @ IOU=0.50
CASCADE_RCNN_C3-C5_R34_FPN_1X	1333×800	[8,11]	[0.5,1.0,2.0]	3	87.59
CASCADE_RCNN_DCONV_C3-C5_R34_FPN_1X	1333×800	[8,11]	[0.5,1.0,2.0]	3	88.93
FASTER_RCNN_C3-C5_R34_FPN_1X	1333×800	[8,11]	[0.5,1.0,2.0]	3	89.33
FASTER_RCNN_DCONV_C3-C5_R34_FPN_1X	1333×800	[8,11]	[0.5,1.0,2.0]	3	90.65

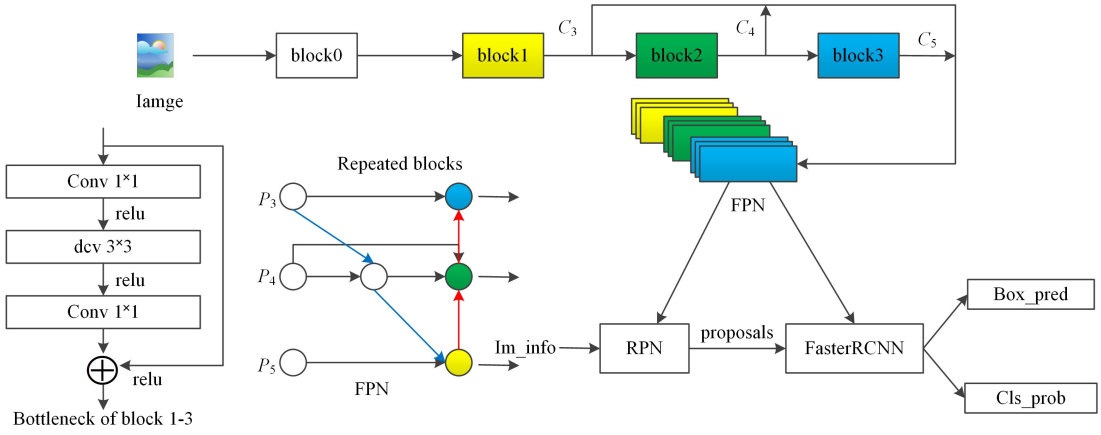

图 3 敏感元素识别模块网络结构图

Fig. 3 Network structure diagram of sensitive element recognition module

2.3 有害视频场景识别

有害视频场景识别采用的是视频分类算法。NetVLAD[18]提供了一种可训练的端到端的卷积神经网络框架直接用于场景识别，NetVLAD 是 VLAD 的改进版，VLAD（Vector of Locally Aggregated Descriptors）是一种将若干局部特征压缩为一个特定大小全局特征的方法，通过聚类实现了特征降维。Arandjelovic 等[19]将 VLAD 改造成可微可导的结构，变成了可训练的函数，这样就可以将 NetVLAD 层接在 CNN 模型提取得到的图像帧后面进行集成。同时，鉴于在识别含有隐蔽性暗示的短视频内容时，极易受到截取的关键帧影响出现混淆判断，在检测每一个关键帧的基础上又增加了对前后两帧的辅助判断，利用经典 BP 算法来学习获得 NetVLAD 计算过程中核心的 K 个 2048 维的聚类中心，并通过卷积计算出来的权重获得了当前特征与每一个 2048 维特征在 K 个聚类中心残差的累计和，记作一个新的 2048 维的特征向量。经过这样的一个特征融合技术，可以将一个任意视频长度的 $N*2048$ 维的特征聚合为一个固定的 2048 维特征，最后利用这个聚合后的特征进行最终分类器的训练，达到利用整个短视频内容来识别视频中出现的有害信息场景的目的。

3 实验

3.1 实验环境及数据准备

本文实验使用的处理器为 Intel Xeon(R)E5-2630 10 核 20 线程 @ 2.20GHz*2，操作系统版本为 Ubuntu 16.04.5，GPU 显卡为 Nvidia Tesla T4 *4，显存为 64GB(16GB*4)，深度学习框架为 pytorch1.9.0，Python 版本为 python3.6，Cuda 版本为 cuda 10.0，Java 运行环境为 JDK 1.8。为使整个识别过程运行不出现堵塞，需要对低吞吐的模块配备更多的计算资源，即分别为视频分类算法模型和图像检测算法模型各自配备两张计算卡，并通过专门的数据分流服务来实现两卡并行计算。

本文采用公开数据集和自建数据集结合的方式进行验证和测试。公开数据集采用 NPDI 数据集[20]，它由近 80h 的视频组成，NPDI 数据集将完全用作测试集（见表 5）。

表 5 NPDI 数据集

Table 5 NPDI data set

类别	视频数	小时数
Non-porn	200	11.5
Non-porn	200	8.5
Porn	400	57
总数	800	77

自建数据集是利用网络爬虫程序从百度贴吧、新浪微博等热门应用程序采集到的。由于有害信息内容留存时间很短，需要不停爬取最新的 50 个帖子和超过 10 万粉丝的微博账号，以确保相关内容在被发现和删除之前被抓取，爬取后对这些数据进行数据清洗、去重、筛选、标注，得到自建数据集如表 6 所列。

表 6　自建数据集

Table 6　Self-built data set

数据集	标签	验证集	测试集
1	Normal	968292	113327
	Sexy	756020	145744
	Porn	533742	52729
2	Normal	12898432	84564
	Terrorism	44036	32000
总数		15597322	428364

考虑到实际应用中数据在文件格式、大小等方面呈现出的多样性和随机性，本文的图像数据集包含 jpg，jpeg，png，bmp 等图像格式，图像分辨率最大为 4999 * 4999，图像文件大小可支持 8MB；短视频数据集包含 H264 编码、mp4、rmvb、mov 等视频格式，短视频时长最长支持 18min，短视频文件大小最大支持 760MB。

3.2　实验结果

3.2.1　消融实验结果

为了验证本文所提出模型结构的有效性，在测试集上进行了以下 6 个消融实验：

(1)仅包括 EfficientNet-B2 前置过滤；

(2)包括前置过滤、有害图像场景识别和有害视频场景识别，不包含敏感元素识别模块；

(3)包括前置过滤、有害图像场景识别、敏感元素识别和有害视频场景识别，有害图像场景识别使用蒸馏前的 Student-Net(proxylessnas)网络结构；

(4)包括前置过滤、有害图像场景识别、敏感元素识别和有害视频场景识别，有害视频场景识别采用图像模型做抽帧计算，抽帧计算的聚合方式为多帧平均(Avg Pooling)；

(5)包括前置过滤、有害图像场景识别、敏感元素识别和有害视频场景识别，有害视频场景识别采用图像模型做抽帧计算，抽帧计算的聚合方式为得分取最大值(Max Pooling)；

(6)本文模型结构：包括前置过滤、有害图像场景识别、敏感元素识别和有害视频场景识别等模块。

消融实验结果如表 7 所列，通过第一阶段的前置过滤模块可以识别过滤掉 80%以上的正常内容，在增加第二阶段内容检测的相关模块后，检测效率虽有一定的降低，但整体识别效果有了大幅提升。如果仅增加有害图像场景识别和有害视频场景识别两个模块，有害信息识别准确率即使提高了 48.8%，仍达不到理想的识别效果，误检率较大。加入敏感元素识别模块后，图像视频数据识别的准确率和召回率均超过 85%，图像和视频数据在逻辑融合后输出了较为理想的结果。相比于蒸馏前，蒸馏后的 Student-Net(MealV2)在整体模型结构中适配度更好，有助于提高整体识别效果。同样，相较于直接用图像模型做抽帧计算的结果，经过 NetVLAD 算法进行聚合之后得到的视频分类识别输出结果表现更为优异。与前置过滤模块相比，整体模型的有害信息识别精确率从 21.5%提高到 99.91%，有害信息召回率和正常内容召回率提高了近 19%，正常内容识别准确率提高了 2%左右，无论是在客观效果指标还是性能指标上都达到了设计需求。

表 7　不同结构模型对结果的影响

Table 7　Impact of different structural models on model results

方法	正常内容 recall/%	正常内容 precision/%	有害信息 recall/%	有害信息 precision/%	性能@T4/fps
1	80.84	97.92	80.48	21.50	511
2	86.23	98.15	82.46	70.31	115
3	94.50	99.97	85.80	93.74	74
4	97.52	99.85	88.67	95.45	42
5	95.72	99.67	89.90	90.74	46
6	99.996	99.92	99.68	99.91	57

3.2.2　与开源算法对比

将本文算法模型和其他开源算法模型进行对比实验，参与实验的开源模型包括 Google Cloud Vision API 和 Yahoo Open NSFW 模型，两个模型可通过 API 进行调用，实验数据集包含 NPDI 和自建测试集。如表 8 所列，本文的整体模型与开源模型相比准确率有明显提高，在 NPDI 公开数据集上准确率最高提升 3.09%，在自建测试集上准确率最高提升 19.26%。

表 8　NPDI 和自建测试集上不同模型实验比较

Table 8　Experimental comparison of different models on NPDI and self-built data sets

方法	数据集	准确率/%
本文模型	NPDI	94.36
Google Cloud Vision API	NPDI	91.27
Yahoo Open NSFW Model	NPDI	92.62
本文模型	自建测试集	99.92
Google Cloud Vision API	自建测试集	82.53
Yahoo Open NSFW Model	自建测试集	80.66

3.2.3　应用场景实验

为检验本文所提出的基于两阶段算法的多媒体有害信息识别模型实际应用效果，将模型与某生产系统对接，运行结果如表 9 所列，多媒体有害信息数据召回率＞98%，准确率＞97%，检测的漏检和误报率均在 5%以内，图像、视频处理效率达到了预期效果。在应用场景下，所提方法能够满足对于图像、短视频的处理性能要求以及多媒体有害信息精准识别需求。

表 9　生产环境运行结果

Table 9　Running results in production environment

类别	识别效率	召回率/%	准确率/%
有害图像	244800 张图像/小时	98.42	98.42
有害视频	6000 分钟短视频/小时	99.03	97.14

结束语　快速准确地识别多媒体有害信息不仅有利于互联网企业完善平台内容安全管理，而且对于打击网络违法犯罪、治理网络空间环境具有十分重要的意义。针对互联网安全监管和网络犯罪打击整治应用场景中，传统识别方法存在的吞吐量偏低、准确率不足、识别类型单一等问题，设计了基于两阶段算法的多媒体有害信息识别模型。总体来说，经多个数据集测试，本文方法相比传统方法在识别准确率和性能指标上表现较好，得到了更为精准的有害信息识别结果，证明了本文模型设计思路的可行性。未来将探索如何让多媒体有害信息识别方法能够快速识别出有害信息的变种文件并通过对比溯源刻画出有害信息的互联网传播链，同时结合文本信息、音频信息综合分析，满足更多领域和业务方向的多媒体有

害信息识别需求，力争在识别效率和准确率上寻找到最佳的平衡点以达到更好的识别检测效果。

参 考 文 献

[1] HUANG Y, KONG A W K. Using a CNN ensemble for detecting pornographic and upskirt images[C]// 2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems(BTAS). IEEE, 2016: 1-7.

[2] CONNIE T, AL-SHABI M, GOH M. Smart content recognition from images using a mixture of convolutional neural networks [M]// IT Convergence and Security 2017: Volume 1. Singapore: Springer Singapore, 2017: 11-18.

[3] MOUSTAFA M. Applying deep learning to classify pornographic images and videos[J]. arXiv: 1511. 08899, 2015.

[4] XIE X. Research on detection technology of pornographic information based on deep learning [D]. Chengdu: University of Electronic Science and Technology of China, 2020.

[5] ZHANG D L. Research on special video content detection algorithm based on deep features [D]. Beijing: Minzu University of China, 2019.

[6] GU Y, LI J, JING B, et al. Internet content information detection and filtration system [J]. Application Research of Computers, 2008(9): 2834-2835, 2862.

[7] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[8] TAN M, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]// International Conference on Machine Learning. PMLR, 2019: 6105-6114.

[9] GOU J, YU B, MAYBANK S J, et al. Knowledge distillation: A survey[J]. International Journal of Computer Vision, 2021, 129: 1789-1819.

[10] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[C]// International Conference on Machine Learning. PMLR, 2021: 10347-10357.

[11] TOUVRON H, VEDALDI A, DOUZE M, et al. Fixing the train-test resolution discrepancy[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019: 8252-8262.

[12] MAHAJAN D, GIRSHICK R, RAMANATHAN V, et al. Exploring the limits of weakly supervised pretraining[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018: 181-196.

[13] CAI H, ZHU L, HAN S. Proxylessnas: Direct neural architecture search on target task and hardware[J]. arXiv: 1812. 00332, 2018.

[14] SHEN Z, SAVVIDES M. Meal v2: Boosting vanilla resnet-50 to 80%+ top-1 accuracy on imagenet without tricks[J]. arXiv: 2009. 08453, 2020.

[15] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[16] SHAFIQ M, GU Z. Deep residual learning for image recognition: a survey[J]. Applied Sciences, 2022, 12(18): 8972.

[17] DONG S. The research on object detection algorithm based on improved SSD and FPN algorithm[D]. Chongqing: Southwest University, 2023.

[18] ZHANG Z H. Study on scene recognition algorithm based on NetVLAD [D]. Chongqing: Chongqing University, 2022.

[19] ARANDJELOVIC R, GRONAT P, TORII A, et al. NetVLAD: CNN architecture for weakly supervised place recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 5297-5307.

[20] AVILA S, THOME N, CORD M, et al. Pooling in image representation: The visual codeword point of view[J]. Computer Vision and Image Understanding, 2013, 117(5): 453-446.

SHI Xiaosu, born in 1996, postgraduate. Her main research interests include image processing and big data analysis.

LI Xin, born in 1977, Ph. D, associate professor, is a professional member of CCF(No. 51691M). His main research interests include cyber security and big data analysis.