基于镜头分割与空域注意力模型的视频广告分类方法

谭 凯 吴庆波 孟凡满 许林峰
(电子科技大学信息与通信工程学院 成都 611731)

摘 要 随着视频广告在检索和用户推荐等领域的广泛应用,视频广告的分类成为一个重要问题。与现有视频分类任务不同,视频广告有其自身的特点:1)在时域上,产品对象在广告视频中的出现具有非周期性和稀疏性的特点,这使得分类任务需要排除大量与视频类别不相关的视频帧的干扰,利用少数相关视频帧进行分类;2)在空域上,视频帧中除产品外,还包含复杂背景的问题,这使得有效捕捉产品信息变得困难。为了解决上述问题,文中提出了一种基于镜头分割和空域注意力模型的视频广告分类方法,简称SSSA。针对视频中存在的大量干扰帧,文中使用基于镜头切换的分割方法采样视频帧。针对视频帧中包含复杂背景,文中在网络中引入视觉注意力机制帮助网络从产品相关区域提取判别性的特征。为了验证所提方法的有效性,构建了一个包含1000多个视频广告的数据库(简称TAV)并收集了眼动数据来训练注意力模型。实验结果显示,提出的SSSA视频分类方法比现有的视频分类方法在性能上提升了10%。

关键词 分类,视频广告,注意力,标注
中图法分类号 TP391.9 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.03.019

Video Advertisement Classification Method Based on Shot Segmentation and Spatial Attention Model

TAN Kai WU Qing-bo MENG Fan-man XU Lin-feng
(School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract As video advertisement is increasingly used in some areas such as search and user recommendation, advertisement video classification becomes an important issue and poses a significant challenge for computer vision. Different from the existing video classification task, there are two challenges of advertisement video classification. First, advertised products appear in advertisement video aperiodically and sparsely. This means that most of frames are irrelevant to advertisement category, which can potentially cause interference with classification models. Second, there are complex background in advertisement video which makes it hard to extract useful information of product. To solve these problems, this paper proposed an advertisement video classification method based on shot segmentation and spatial attention model (SSSA). To address interference of irrelevant frames, a shot based partitioning method was used to sample frames. To solve the influence of complex background on feature extraction, the attention mechanism was embedded into SSSA to locate products and extract discriminative feature from the attention area which is mostly related to the advertised products. An attention predictionnetwork (APN) was trained to predict the attention map. To verify the proposed model, this paper introduced a new thousand-level dataset for advertisement video classification named TAV, and the gaze data were also collected to train the APN. Experiments evaluated on the TAV dataset demonstrate that the performance of the proposed model improves about 10% compared with the state-of-the-art video classification methods.

Keywords Classification, Video advertisement, Attention, Annotation

1 引言

随着网络多媒体技术[1-4]的发展,每天有大量的广告视频被制作出来并且投放到网络中。面对网络上不断激增的视频广告,针对广告中产品类别(如饮料、鞋子、汽车、电脑等)的准确分类成为一个必要的任务,有着广阔的应用前景。它可以帮助用户进行有效的类别检索,帮助企业对用户进行相关产品的广告推荐和广告插入等。然而,对于如此海量的视频数据,人工标注需要耗费大量的人力、物力以及时间成本。因此,利用计算机视觉技术解决视频分类问题显得尤为重要和迫切。

分类问题可以分为两大类:图像分类和视频分类。随着计算机视觉[4-7]技术的不断发展,尤其是深度学习技术的引入,图像分类方法取得了巨大的成就[8-10]。随后,人们把目光投向了更为艰巨的视频分类任务。与图像分类不同,视频分

到稿日期:2018-07-20 返修日期:2018-09-29 本文受国家自然科学基金(61601102,61502084,61871087)资助。
谭 凯(1988-),男,博士生,主要研究方向为视觉注意力、对象检测和质量评价,E-mail:kaitanuestc@gmail.com;吴庆波(1985-),男,博士,副教授,主要研究方向为图像视频编码和质量评价,E-mail:qbwu@uestc.edu.cn(通信作者);孟凡满(1984-),男,博士,副教授,主要研究方向为图像分割和对象检测;许林峰(1976-),男,博士,副教授,主要研究方向为视觉注意力、图像视频编码、视觉信号处理和多媒体通信系统。

类不仅需要处理空域信息,同时还要考虑视频特有的时域信息。现有的视频分类研究主要集中于行为识别[11],侧重于挖掘同一行为在时域上具有的规律性模式[12-14]。

与现有行为识别任务不同,广告视频有其自身的特点。1)在时域上,产品对象在视频中的出现具有非周期性和稀疏性的特点。具体来说,为了使广告具有吸引力,广告内容通常会以不同故事的形式出现,这种故事情节决定了与广告类别相关的信息(如产品)并不一直出现在视频中,甚至仅仅出现在某些特定时刻,因此产品在视频中的出现具有稀疏性的特点。同时,同一类别的广告有各种不同的故事情节,这使得与类别相关的信息在视频中的出现具有非周期性的特点。图 1 展示了 3 个广告视频序列,前三行是 3 个饮料广告的视频序列,包含产品的视频帧用灰色粗体框标示,从中我们可以清晰地看到视频中存在上述特点。因此,如何排除广告视频中大量不相关帧对分类的干扰成为一个重要的问题。2)在空域上,产品与其他对象间具有复杂的关系。广告视频帧中不仅包含产品对象,还常常包含其他无关对象甚至是复杂的背景。图 1 第四行展示了一些包含产品的视频帧。面对这些与视频类别不相关的对象和背景,分类模型需要识别产品与它们的关系,这给分类模型从单帧中捕捉有效的产品信息带来了困难。

图 1 广告视频帧

Fig. 1 Advertisement frames

为了解决上述问题,本文提出了基于镜头的视频分割方法和带有空域注意力的网络模型(简称 SSSA)用于视频广告的分类任务。针对产品对象稀疏性和非周期性的特点,本文使用基于镜头切换的分割方法采样视频帧,以减少采样帧中不相关帧的数量;同时为了从视频中提取有效信息,本文使用多支流网络来编码视频。我们观察到,在广告视频中存在一定数量的镜头切换。在这些镜头中,与类别相关的镜头的比例明显高于与类别有关的视频帧在视频序列中的比例。同时,同一镜头中的视频帧内容相同或相似,可以由一个视频帧代替镜头内所有的视频帧。通过上述观察,我们认为将视频按照镜头进行分割并采样可以降低不相关视频帧在采样样本中的比例,从而减少不相关帧对分类任务的干扰。

针对视频帧中包含复杂背景的特点,本文引入了视觉注意力机制,并将视觉注意力定位的区域定义为能够识别视频类别的区域。通过引入视觉注意力机制,模型能够抑制不相关信息的干扰,从产品相关区域提取判别性的特征。图 2 展示了 3 个不同广告的视频帧及对应的注视谱。从中可以看到,即使在包含复杂背景干扰的视频帧中,注视点仍然可以排除无关区域,关注与视频类别相关的区域。

图 2 3 个广告帧的注意力谱

Fig. 2 Visual attention maps for three advertisement frames

为了验证本文方法的有效性,构建一个包含 1200 多段广告视频的数据库(Thousand-level dataset for Advertisement Video classification,TAV)。该数据库中的视频内容包含了常见的六大类产品:电器、服装、食品、护肤、保健、交通工具。这六大类被进一步划分成 44 个日常子类,如汽车、衣服、鞋子、饮料等。同时,我们收集了眼动数据用于训练注意力模型,以预测关键区域。

综上所述,本文主要有如下 3 方面贡献。

1)提出了一种基于镜头的视频分割方法来采样视频帧。该采样方法能够降低不相关视频帧对分类任务的干扰。

2)引入视觉注意力机制,帮助模型从与类别相关(如产品)的区域提取更具判别性的特征。

3)构建了一个包含 1200 多段视频广告的数据库,用于方法验证和视频广告分类研究。

2 相关工作

对于图像分类任务,为了更加有效地表示图像,现有算法利用各种技术定位前景对象并从中提取特征,这些技术包括显著对象检测、注意力检测[10]和图像分割[22]等。与从整个图像提取特征不同,文献[9]中方法使用多个兴趣区域检测器定位潜在对象区域并从中提取特征,取得了不错的效果。随着显著对象检测方法在性能上的突破,一些工作开始尝试使用显著检测方法定位前景对象,帮助提升图像分类精度。文献[8]中提出使用非线性扩散滤波方法检测显著区域,并从中提取多尺度的信息用于图像分类。文献[10]使用多种不同的显著检测方法检测对象区域,并从中提取特征,提升了图像分类性能。

随着图像分类技术的发展以及逐渐趋向成熟,一些工作开始将注意力转移到视频分类任务上。现有的视频分类方法

主要集中于行为识别任务。文献[15]的方法从视频中提取稠密轨迹，并利用这些轨迹的特征训练 SVM 分类器，进而进行行为分类。随着深度学习在计算机视觉领域的成功，许多方法将 CNN 技术应用于行为识别任务中。文献[16]的方法将 CNN 作为特征提取器，以提取人体不同部件的特征，并使用 SVM 分类行为。Karpathy 等人[17]首先训练 CNN，以解决视频分类任务。他们使用两个网络分别从低分辨率和高分辨率的同一视频帧中提取不同尺度的特征，并使用该特征识别行为类别。Simonyan 等人[11]考虑了时域信息，提出了一种包含 2 条支流的分类网络，2 条支流分别从视频帧中提取空域特征和从光流中提取时域特征，最后将两支流的分类得分进行融合。考虑到视频帧数与网络输入不对等的问题，该方法从视频序列中随机采样几帧，并将其分别作为网络的输入，最后将这些采样帧的分类平均得分作为最后的结果。在两支流网络的基础上，为了结合空域特征与时域特征，Tran 等人[12]使用三维滤波器将空域支流和时域支流融合成为一个时空域网络，通过结合时域和空域中的抽象信息得到更加鲁棒的特征表达。另外，一些方法提出使用循环神经网络来建模视频。Ng 等人[18]将 CNN 和 LSTM 结合用于行为识别，他们使用 CNN 作为特征提取器来提取每个视频帧的特征，并将这些特征按照时间顺序依次输入到 LSTM 中，用 LSTM 来建模视频序列信息。上述方法尽管在行为识别领域取得了不错的效果，但是并不适用于广告视频分类任务，这是因为它们没有考虑到广告视频存在的特点，即非周期性、稀疏性和复杂性。

3　视频广告分类方法

本节首先介绍本文方法的整体流程，如图 3 所示。给定一个视频序列，SSSA 首先利用基于镜头的分割方法将视频划分成一组视频段（每一视频段称作一个镜头）。这些镜头之间具有独立的语义信息，在同一镜头内，连续的视频帧之间具有相同或者相似的内容。因此，镜头内的视频帧可以用来表示当前镜头内的所有视频帧。我们按时间顺序选取 K 个镜头，用镜头内的视频帧代替当前镜头，依次将 K 个视频帧输入到多支流网络（MSN）中。在 MSN 的每条支流中嵌入一个预训练的注意力预测模块（简称 APN），用于定位与视频类别相关的区域。每个支流网络提取的特征谱与 APN 预测的注意力谱进行融合，随后经过全连接层进行分类。最终，模型将所有支流的分类结果线性相加，共同预测当前视频的类别。

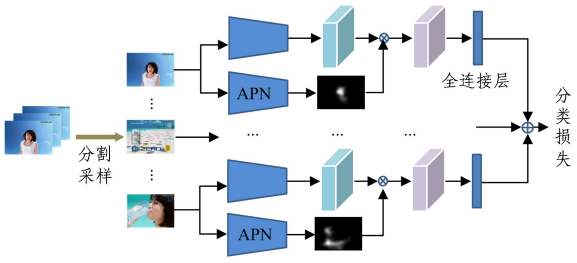

图 3　本文 SSSA 方法的流程图

Fig. 3　Flowchart of proposed SSSA

3.1　基于镜头的视频分割

给定一个视频，目的是生成一组时序上连续的、语义独立的视频段 $St=\{St_1,St_2,\cdots,St_N\}$。其中，$N$ 表示视频段的个数；$St_N=\{pt_1,pt_2\}$，表示第 N 个视频段在视频序列中的位置；pt_1 和 pt_2 分别表示起始帧和结束帧的位置。

在一个镜头中，连续的视频帧描述了相同的对象。这些对象的运动具有连续性和一致性。因此，任意相邻两帧的内容在视觉上是相似的。然而，对于任意相邻的镜头，它们可能在不同的场景中描述相同的对象或者是在相同的场景中包含不同的对象，甚至二者完全没有相关性。因此，相较于镜头内相邻帧间的差别，相邻镜头间的差别更大。因此，本文提出利用相邻帧间的差异分割镜头，从而生成视频段。具体来讲，首先计算相邻两帧在 RGB 空间上的像素差异，所有像素差异之和用来决定一个镜头的起始位置和结束位置。若该差异值大于阈值 T，则认定此处存在镜头变换，这两个相邻帧分别被当作当前相邻镜头的结束帧和起始帧。经过上述操作，所有差异值大于阈值 T 的相邻帧被找到，从而得到一组视频段 St。对于视频中所有差异都小于阈值的情况，按照差异从大到小的顺序依次定位镜头变换。假定生成了 N 个视频段 $St=\{St_1,St_2,\cdots,St_N\}$，从这 N 个视频段中采样 K 个连续的视频段，然后选取每个视频段的中间帧来代替当前视频段，并将选取的 K 个中间帧依次输入到网络模型中。

3.2　多支流网络框架（MSN）

为了从视频中提取有效信息，本文使用多支流网络来编码视频。多支流网络可以同时输入多个视频帧，从中提取与产品类别相关的信息，抑制不相关信息对模型分类的影响。

如图 3 所示，多支流网络由多个结构相同的分类网络构成。一个分类网络包含多个卷积层、max pooling 层和 Relu 层，输出一个三维的特征谱 $F\in R^{h\times w\times ch}$。特征谱 F 经过两个全连接层后，得到分类打分。对于多支流网络，将所有支流的分类打分进行线性相加，共同预测视频类别。

具体来讲，给定 K 个视频帧，它们依次被输入到多支流网络，第 i 个支流的最后一层的全连接层输出的类别概率为 S_c^i。为了预测视频的类别，将 K 个采样帧的类别概率 $\{S_c^1,S_c^2,\cdots,S_c^K\}$ 进行线性相加：

$$S_{vc}=\sum_{i=1}^{K}S_c^i \tag{1}$$

其中，S_c^i 表示第 i 帧属于第 c 类的概率；S_{vc} 表示视频属于第 c 类的概率。

3.3　空域注意力预测模型（APN）

3.3.1　空域注意力预测模型（APN）的嵌入

在 MSN 的每条支流中嵌入一个空域注意力预测模型 APN，用于帮助网络从产品相关区域提取判别性的特征。给定训练（测试）视频帧 X，它被同时输入到一条分类支流和一个注意力预测模块（APN），分别生成特征谱 $F\in R^{h\times w\times ch}$ 和注意力谱 $A\in R^{h\times w}$。在输入到全连接层之前，特征谱 F 的每个通道与注意力谱 A 进行融合：

$$\hat{F}_i=F_i\otimes A \tag{2}$$

其中，F_i 表示特征谱 F 的第 i 个通道；$\otimes$ 表示元素级相乘；$\hat{F}$ 表示 F 与注意力谱相结合后的特征谱。上述操作的动机如下：注意力谱 A 能够定位与视频类别相关的区域（如产品）。将特征谱与注意力谱相结合能够突出与类别相关区域的特征，同时弱化或者抑制不相关区域；这使得融合后的特征谱 $\hat{F}$

更加趋向于相关区域的特征，减少不相关区域对特征提取的影响。经过上述操作，融合后的特征$\hat{F}$经两个全连接层以及一个 softmax 层，输出当前帧 X 的类别概率 S_c。图 4 展示了 APN 预测得到的注意力谱，其中奇数列展示原始图片，偶数列为预测的注意力谱。在注意力谱中，亮度高的区域表示注意力区域，亮度低(黑色)的区域为背景区域。从图 4 中可以看到，APN 模型能够定位与类别相关的区域(如产品、产品与对象间关系)，即使在复杂背景下或者产品尺寸较小时，APN 仍能定位相关区域。

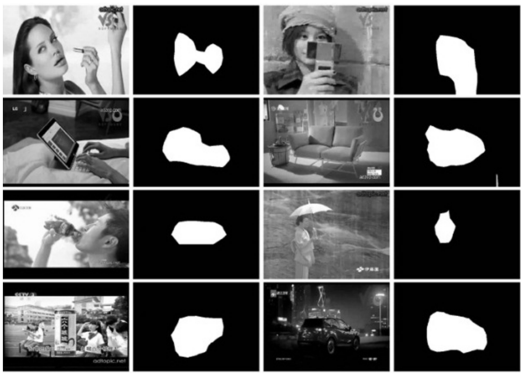

图 4 APN 预测的注意力谱

Fig. 4 Visual attention maps predicted by APN

3.3.2 空域注意力预测模型(APN)

注意力预测模型 APN 将一个视频帧作为输入，输出一张与输入尺寸相同的注意力谱。本文使用分割[19-21]中常用的全连接网络(FCN)[22]作为 APN 的基础网络。具体来说，APN 包含一系列卷积层、max pooling 层和 Relu 层，以及两个全连接层。由于多个 max pooling 层的存在，APN 输出谱的尺寸小于输入图像的尺寸。为了使输出的注意力谱与输入图片尺寸相同，我们在全连接层后加入一层转置卷积层来对注意力谱进行上采样。为了训练 APN，将问题转化为预测每个像素属于注意力区域(或背景区域)的置信度。因此，本文使用 softmax with loss 计算 APN 预测的注意力谱 p 与 ground truth y 之间的损失：

$$L=\frac{1}{T}\sum_{i=1}^{T}\sum_{j=1}^{C}y_{i,j}\cdot\log p_{i,j} \tag{3}$$

其中，T 表示像素的个数；$C=2$ 为类别数量，表示一个像素是否属于注意力区域；$y_{i,j}$ 是一个二值标签，等于 1 表示第 i 个像素属于第 j 类，否则为 0；$p_{i,j}$ 由注意力模型 APN 预测得到，表示第 i 个像素属于第 j 类的置信度。

4 数据集

4.1 数据收集

为了验证所提方法的有效性，本文构建了一个包含 1238 个广告视频的数据库(TAV)。为了构建 TAV，首先将广告分为常见的六大类：电器、服装、食品、护肤、保健、交通工具。在此基础上，每一类被进一步划分为多个日常子类，如汽车、衣服、鞋子、饮料等。对于每一个子类，按照关键词从网络中搜集与此类相关的广告视频。经过上述操作，共有大约1500 个广告视频被收集起来。人工对这些视频进行筛选，去除内容重复、主观质量太差以及不能人工识别其类别的视频，最终保留了 1238 个广告视频，并为其标记相应的广告类别。

4.2 眼动数据的采集

为了研究人类视觉注意力对视频分类的帮助，使用 Tobbi X2-60 眼动仪追踪人眼的注视点。通常情况下，视频以24 Hz 的频率展示图片。考虑到眼动仪的采样频率是 60 Hz，我们降低视频的显示频率，从而增加每一视频帧在屏幕中停留的时间。具体来说，我们将每一帧的停留时间延长到 100 ms。这样操作有两个好处：1)对于每一视频帧能够收集更多的注视点，这样可以降低噪声点的干扰，使收集的数据更加鲁棒；2)受试者可以有更多的时间观看视频帧内容，这可以使受试者更加容易地定位相关区域。

在观看每个视频之前，受试者被告知视频的类别。他们被要求在观看视频时，需尽力找到每一帧中能够表明当前广告类别的区域。我们雇佣了 5 位受试者参加此次实验，每一个受试者独立观看所有的视频广告。测试完成之后，对于每一个视频帧，所有 5 位受试者的注视点被收集起来生成一张注视点谱，再经过高斯滤波平滑后得到最终的注视谱。

5 实验

5.1 实验设置

本文将 TAV 数据库随机划分成 3 部分：训练集、验证集和测试集，三者所占比例分别为 70%，10%和 20%。同时，对于每一类，测试集必须包含至少 2 个样本，验证集至少包含 1 个样本。

5.2 网络实现

本文提出的 SSSA 网络模型使用 MatConvnet 工具箱在 MATLAB 上实现。SSSA 网络使用 VGG-16 分类模型作为基础网络，每条支流由 VGG-16 进行初始化。在多支流网络中，因为所有分类支流将视频帧作为输入以提取特征，将所有支流的对应卷积层和全连接层的参数进行共享以减小参数空间，即每个支流拥有相同的卷积层参数和全连接层参数。对于注意力预测模型 APN，前面 5 层的卷积层参数使用 VGG-16 进行初始化。对于后续的全连接层和转置卷积层，使用均值为 0、标准偏差为 0.01 的高斯核进行随机初始化，偏差设置为 0。与现有显著对象检测算法常用的像素级标签构建方法相同，将注意力谱中大于 0.5 的像素置 1，其他像素置 0，将二值化的注意力谱作为 APN 训练的像素级标签。使用随机梯度下降算法(SGD)训练该网络，初始学习率为0.0001。首先训练 APN 网络，然后固定 APN 网络参数，训练 SSSA 网络。对于 SSSA 网络，我们使用多步训练策略，初始学习率为 0.001，冲量和权重衰减分别为 0.9 和 0.05。

5.3 消去实验

5.3.1 基于镜头的视频分割

本节展示所提基于镜头的视频分割采样方法(SS)的有效性。为此，使用两种额外的分割方法作为基准比较对象：随机采样(RS)法和均匀采样(US)法。给定一个视频，RS 随机采样 K 个视频帧，并将它们按时间顺序依次作为输入；US 从视频中随机选取起始帧，以固定步长采样 K 个视频帧。在实

验中，本文将固定步长设置为 128 帧。将使用不同方法采样得到的视频帧作为输入来训练 MSN 分类网络，其分类性能如图 5 所示。从图 5 中可以看出，本文提出的 SS 分割采样方法的性能优于其他两种基准方法。原因分析如下：考虑到产品在广告视频中具有稀疏性的特点，通过 SS 方法采样的 K 帧中不相关帧的比例降低，同时包含更多与类别相关的视频帧，从而降低了不相关信息对网络模型的干扰，提升了网络的分类性能。

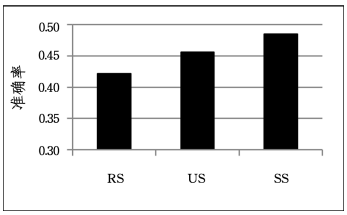

图 5　不同分割采样方法的分类性能

Fig. 5　Performance of different segmentation methods

5.3.2　方法的不同变体

本节展示本文方法中不同模块的有效性。从表 1 中可以看出，本文提出的注意力预测模块对提升模型的分类性能有一定贡献。具体来说，当多支流网络分别嵌入 APN 后（MSN＋APN），性能提升了 2%，这一结果验证了引入注意力模型对提升视频分类性能的有效性。APN 能够定位与产品类别相关的区域，嵌入分类模型后能够滤除不相关区域的干扰信息，使分类模型能够提取更有效的特征。从图 4 展示的 APN 预测的视频帧中可以看到，APN 可以定位与产品类别相关的区域。

表 1　本文方法在有/无 APN 模块下的性能

Table 1　Performance of our method with/without APN

算法	准确率/%
MSN	48.5
SSSA（MSN＋APN）	50.5

5.4　性能比较

本节将 SSSA 与行为识别领域中 3 种性能最好的方法即 TS[11]，C3D[12] 和 CF[23] 进行比较。TS 方法设计了一个包含空域支流和时域支流的两支流网络。为了更好地结合时域特征和空域特征，C3D 提出了用一个三维滤波器替代 CNN 中常用的二维滤波器。与 C3D 思想类似，CF 用一个三维滤波器将空域支流和时域支流融合为一个单路的时空域网络。除了上述算法外，实验还添加了两个分类网络：SCN 和 TCN。SCN 和 TCN 分别表示 TS 方法中的空域支流网络和时域支流网络。对于上述方法，使用原作者提供的代码和默认参数在 TAV 数据库上进行训练及测试。

所有方法的分类性能如表 2 所列。从表 2 中可以发现一个有趣的现象：TCN 性能较低时，将 SCN 和 TCN 线性融合后的 TS 方法的性能略低于 SCN。分析原因如下：与行为识别领域注重运动信息不同，在广告视频中运动信息显得并不重要。这是因为对于同一类别的广告视频，对象与产品间的行为并不一致，因此在广告视频中很难发掘规律性的时域信息。同时，上述现象也反映了外观信息即空域信息在广告视频中的重要性，因此，如何从空域中提取判别性特征显得尤为重要。上述 3 种方法在视频广告分类上具有相似的性能，它们的表现均相对较差。本文提出的 SSSA 方法比现有 3 种方法的准确率都高，尤其值得一提的是，比 3 种方法中效果最好的 TS 方法的性能提升了接近 10%。分析原因如下：1）广告视频中广告产品的稀疏性导致现有方法在选取视频帧训练网络时存在输入图片与标签类别不符的情况，本文使用基于镜头的视频分割方法采样视频帧，同时使用多支流网络，减少了输入图片与标签类别不一致的情况；2）通过引入注意力模型，可以帮助 SSSA 网络从与视频类别相关的区域提取判别性特征，同时抑制不相关区域对网络分类的干扰。

表 2　本文方法与现有方法在 TAV 数据库上的性能比较

Table 2　Performance comparison of our method and existing state-of-the-art approaches on TAV dataset

算法	准确率/%
SCN	40.8
TCN	30.4
TS	39.4
C3D	38.7
CF	38.3
SSSA	50.5

结束语　本文提出了一种基于镜头分割和空域注意力模型的广告视频分类方法。针对产品在视频中非周期性和稀疏性的特点，使用了一种基于镜头切换的分割方法对视频帧进行采样，以降低不相关帧对模型分类的影响。针对广告中产品与其他对象间关系复杂性的特点，本文在网络中引入视觉注意力机制，帮助模型从产品相关区域提取判别性的特征，同时抑制不相关区域对网络分类的影响。为了验证本文方法的有效性，构建了一个包含 1200 多个视频广告的数据库 TAV。实验结果表明，本文提出的广告视频分类方法的性能优于现有的分类方法。

参 考 文 献

[1] WU Q，LI H，WANG Z，et al. Blind image quality assessment based on rank-order regularized regression[J]. IEEE Transactions on Multimedia，2017，19(11)：2490-2504.

[2] MENG F，LI H，WU Q，et al. Seeds-based part segmentation by seeds propagation and region convexity decomposition[J]. IEEE Transactions on Multimedia，2018，20(2)：310-322.

[3] WU Q，LI H，NGAN K N，et al. Blind image quality assessment using local consistency aware retriever and uncertainty aware evaluator[J]. IEEE Transactions on Circuits and Systems for Video Technology，2018，28(9)：2078-2089.

[4] TAN K，XU L，LIU Y，et al. Small group detection in crowds using interaction information[J]. IEICE Transactions on Information and Systems，2017，100(7)：1542-1545.

[5] WU Q，LI H，MENG F，et al. A perceptually weighted rank correlation indicator for objective image quality assessment[J]. IEEE Transactions on Image Processing，2018，27(5)：2499-2513.

[6] MENG F，CAI J F，LI H. Cosegmentation of multiple image groups[J]. Computer Vision and Image Understanding，2016，

146:67-76.

[7] WU Q,LI H,MENG F,et al. Blind image quality assessment based on multichannel feature fusion and label transfer[J]. IEEE Transactions on Circuits and Systems for Video Technology,2016,26(3):425-440.

[8] HU W,HU R,XIE N,et al. Image classification using multiscale information fusion based on saliency driven nonlinear diffusion filtering[J]. IEEE Transactions on Image Processing,2014,23(4):1513-1526.

[9] ISCEN A,TOLIAS G,GOSSELINP H,et al. A comparison of dense region detectors for image search and fine-grained classification[J]. IEEE Transactions on Image Processing,2015,24(8):2369-2381.

[10] XIAO T,XU Y,YANG K,et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2015:842-850.

[11] SIMONYAN K,ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]// Advances in Neural Information Processing Systems. 2014:568-576.

[12] TRAN D,BOURDEV L,FERGUS R,et al. Learning spatiotemporal features with 3d convolutional networks[C]// Proceedings of the IEEE International Conference on Computer Vision. IEEE,2015:4489-4497.

[13] DONAHUEJ,HENDRICKSLA,GUADARRAMA S,etal. Long-term recurrent convolutional networks for visual recognition and description[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2015:2625-2634.

[14] DAVE A,RUSSAKOVSKY O,RAMANAN D. Predictive-corrective networks for action detection[C]// Proceedings of the Computer Vision and Pattern Recognition. IEEE,2017.

[15] JHUANG H,GALL J,ZUFFI S,et al. Towards understanding action recognition[C]// Proceedings of the IEEE International Conference on Computer Vision. IEEE,2013:3192-3199.

[16] CHÉRON G,IVAN L,et al. P-CNN:Pose-based CNN features for action recognition[C]// Proceedings of the IEEE International Conference on Computer Vision. IEEE,2015:3218-3226.

[17] KARPATHY A,TODERICI G,SHETTY S,et al. Large-scale video classification with convolutional neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2014:1725-1732.

[18] NG J Y H,HAUSKNECHT M J,VIJAYANARASIMHAN S,et al. Beyond short snippets:Deep networks for video classification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2015:4694-4702.

[19] MENG F,LI H,WU Q,et al. Weakly supervised part proposal segmentation from multiple images[J]. IEEE Trans. Image Processing,2017,26(8):4019-4031.

[20] MENG F,LI H,WU Q,et al. Globally measuring the similarity of superpixels by binary edge maps for superpixel clustering[J]. IEEE Transactions on Circuits and Systems for Video Technology,2018,28(4):906-919.

[21] MENG F,LI H,LIU G,et al. Object co-segmentation based on shortest path algorithm and saliency model[J]. IEEE Transactions on Multimedia,2012,14(5):1429-1441.

[22] LONG J,SHELHAMER E,DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2015:3431-3440.

[23] FEICHTENHOFER C,PINZ A,ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2016:1933-1941.