指导教师：杨涛　　提交时间：2016/3/15

CVPR2015 Paper Translation

No: 01

姓名：王乐文

学号：2013302517

班号：10011303

更好的利用 OS-CNNs 来更好的解决图像的事件识别问题

Limin Wang Zhe Wang Sheng Guo Yu Qiao
Shenzhen Key Lab of CVPR, Shenzhen Institutes of Advanced Technology, CAS, China
{07wanglimin, buptwangzhe2012, guosheng1001}@gmail.com, yu.qiao@siat.ac.cn

摘要

对于图像识别而言，从静态图像中识别事件是最重要的问题之一。然而，与物体识别和场景识别相比，在计算机视觉领域，事件识别受到的关注要远小于前两者。这篇论文主要解决静态图像中文化活动的识别问题，重点关注运用深度学习的方法来解决这个问题。特别的，我们成功的利用目标场景的卷积神经网络架构(OS-CNNs)进行了事件识别。OS-CNNs 网是由物体网和场景网组成的。它将预先获得的特征各自与场景识别数据集对照并转化为预先构建的模型，通过把 OS-CNNs 当作“端到端事件预测”或“通用特征提取器”，我们提出四种场景以探索 OS-CNNs 在事件识别中的应用。我们的实验结果显示用 OS-CNNs 算法检测的全国和有代表性的地方的结果一致。最后，我们提出了解决场景识别的方法，并参加了国际计算机视觉大会 2015 年的 LAP 挑战赛，我们的队伍在挑战赛中获得了第三名的成绩，而且我们的分数和第一名的分数非常接近。

1. 介绍

图像理解逐渐成为计算机视觉领域中最重要的问题之一，很多研究者一直致力于研究这一主题。然而物体识别和场景识别在图像分类的范畴内都已经被深入的研究过了，静态图像中的时间识别相比之下则受到了较少的关注，但是它在图像语义解释中扮演着重要的角色。如图 1 所示，我们可以看出事件的特点是很复杂的，它和很多因素比如物体，场景类别，人们的衣着和人们的姿势等其他因素都有关系。因此，静止图像中的事件识别对当前最先进的图像分类方法提出了更多的挑战，我们需要进一步研究计算机识别。

图 1：ICCV 举办的 LAP 的数据集中获取的文化事件图片的例子。从这些图片中，我们可以看出这些图片中显示的时间特点是很复杂的，它和很多因素比如物体，场景类别和人们的衣着都有关系。

神经卷积网络最近在大型图像分类界受到极大关注，尤其是物体识别和场景识别。对于事件识别而言，相对前两者来说为这个问题设计解决的深入学习方法则很少。我们先前的工作提出了一种新的架构，叫做 OS-CNNs，以用于文化事件的识别。

OS-CNNs 算法从包含的物体和场景类别的角度提取有用的信息进行事件分析。OS-CNNs 由两个网络组成，分别为物体网和场景网。物体网预先包含了大规模物体识别数据集，场景网是基于大规模模型场景识别数据集构建的。分解成为物体网和场景网使我们可以使用大规模注释过的图像来初始化 OS-CNNs，这可能可以进一步改善事件识别数据集。最后，事件识别的执行基于后期融合后输出的物体网和场景网。

在研究 OS-CNNs 后，在本文中，我们试图进一步探索 OS-CNNs 的不同方面以更好地利用 OS-CNNs 来进行事

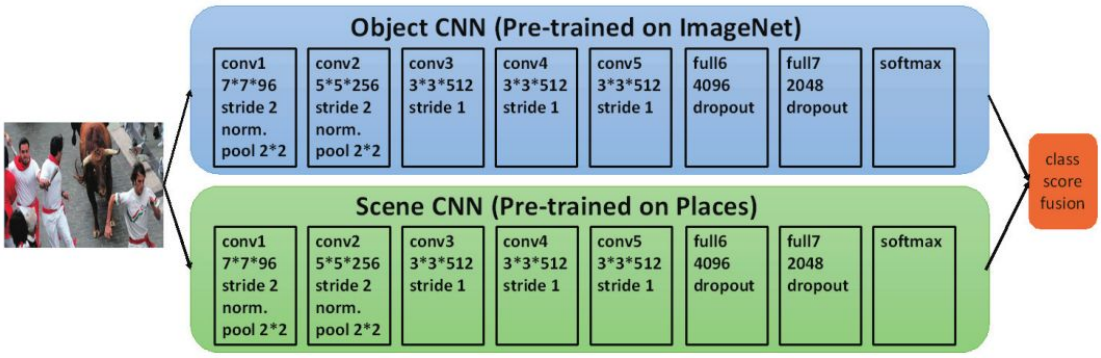

图 2：OS-CNN 在事件识别中的结构由两个网络组成：物体网和场景网。这两个网络的预训练特征分别单独存放在图像网数据集和地点 205 数据集中。

件识别。具体来说，我们设计四种类型的调查方案研究 OS-CNNs 的性能。在第一个场景中，我们直接使用 softmax 函数调用的输出结果。在接下来的三个场景中，我们将 CNN 作为特征提取器，使用它们来提取初始图像的局部和全局特征。全局特征更紧凑，这有助于捕捉整体结构；但局部特征重点描述图像细节和局部模式。我们的实验结果指出这两种特点是互补的，对于事件识别来说是必不可少的。基于我们对 OS-CNNs 的实证探索，在 ICCV-LAP 中，我们为文化活动识别提出我们的解决方案并且在比赛中获得了第三名的成绩。

本文的其余部分组织如下，在第二部分中，我们将简要介绍 OS-CNNs，包括网络架构以及实现细节。在这之后，我们将在第三部分介绍我们对 OS-CNNs 事件识别的广泛探索。然后，我们将在第四部分公布我们的实验结果。最后，我们在第五部分总结我们的方法以及介绍将来的工作。

2. 回顾 OS-CNNs

在这一部分中，首先，我们将简要介绍 OS-CNNs 的架构，提出我们先前在这方面做得工作。然后，我们将呈现实现 OS-CNNs 的细节，包括网络结构、数据扩增和学习策略。

2.1 OS-CNNs

在计算机视觉研究领域和其他两个高度相关的领域：对象识别和场景识别中事件是一个相对复杂的概

念。OS-CNNs 背后的基本思想是利用两个独立的部分来分别完成发生对象的识别和场景的识别。具体来说，OS-CNNs 网是由物体网和场景网组成的，如图二所示。

物体网。物体网是用于捕捉对象中的有用信息来帮助事件识别的。事件发生的对象可以为事件识别提供有用的信息。例如，图 1 中所示澳大利亚的文化活动——澳州国庆日，澳大利亚国旗是一个很有代表性的物体。物体网的主要目标是处理物体线索，我们创建基于最新进展的大型对象识别和预训练网络公众形象网络模型。然后，我们通过把输出结果重置为100来进一步调整数据集的模型参数（文化时间识别数据集包含了 100 个类）。

场景网。场景网预计将用于提取图像的场景信息以帮助事件理解。在一般情况下，场景的环境将对图片中的事件进行分类。例如，在札幌冰雪节的文化事件，如图 1 所示，通常户外为现场类别。具体来说，我们通过使用数据集Places205模型的学习来预训练现场网，其中包含 205 个场景类和二百五十万图像。类似于物体网，之后我们调整事件识别数据集的现场网网络权值，我们网络输出数量

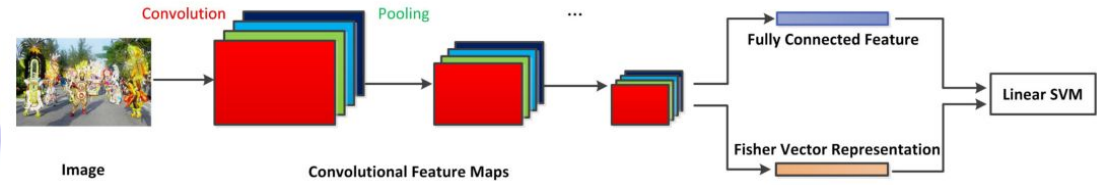

图 3：我们进一步探索 OS-CNNs 在事件识别中的应用，我们利用 OS-CNNs 来提取全局特征（通过激活完全连接层）和局部特征（通过激活卷积层），它们可以被结合起来用于静止图像中的事件识别。

设置为 100。

基于上面的分析，认识到文化活动将受益于对于对象识别和场景识别的转移学习。因此，我们将把物体网和场景网的输出融合起来作为 OS-CNNs 的预测。

2.2 实现细节

在本节中，我们将描述训练 OS-CNNs 的实现细节，包括网络结构、数据扩增，和学习策略。

网络结构。网络结构对于改善 CNNs 性能是非常重要的。在过去的几年里，许多成功的网络架构提出了对象识别，如 AlexNet[12]，ClarifaiNet[27]，OverFeat[17]，GoogLeNet[20]，VGGNet[18]，MSRANet[9]和 Inception2[10]。一些好的做法可以源自于网络体系结构的演变：较小的卷积核大小、较小的卷积跨步、更多的回旋的通道以及更深层次的网络结构。在本文中，鉴于其识别结构的良好的性能，我们选择 VGGNet-19 作为主要调查对象，它由 16 个卷积层和 3 层完全连接组成。关

于VGGNet-19的详细描述超出了本文的范围，可以在[18]中查找详细资料。

数据对应。通过数据增加，我们的意思是通过转换打乱一副图片，同时让底层的类不变。典型的转换包括角落种植、抖动规模和水平翻转。具体来说，在OS-CNNs的训练阶段，我们在图像区域(224×224)的4角和1个中心随机添加样本。与此同时将这些裁剪区域进行随机水平翻转。此外，我们使用三种不同的尺度来调整被训练图像，图像的最小大小被设置为256，384，512。

应该注意的是，数据扩增方法适用于训练图像和测试图像。在训练阶段，数据增加将产生额外的训练同时减少过度拟合带来的影响。测试阶段，数据的增加将有助于提高分类精度。扩充样本可以被视为独立的图像或通过合并或叠加的操作方式结合成一个单一的图像。在当前实现中，在测试阶段，我们使用合并的操作方式将扩充样本结合成一个单一的图像这种方法。

学习策略。有效的训练方法对于学习CNNs模型是非常重要的。相对于ImageNet[4]和Places205[28]，作为文化活动识别的训练数据集相对比较较小，我们通过使用曾经训练了ImageNet 和Places205的三个可以公开利用的模型来对OS-CNNs进行预训练。具体来说，我们使用了公共VGGNet-19模型1来预训练物体网，它在ILSVRC2014获得最佳性能。对于场景网，我们使用由[21]2创造的模型来初始化网络的权重，它至今为止在Places205数据集取得最佳性能。

网络权值学习使用带动力（设置为0.9)的小批量随机梯度下降法。在每个迭代中，一组小批量的256幅图像是由随机抽样构建的。完全连接层的回动比被设置为0.5。我们运用ImageNet 和Places205模型来预训练网络权重和，我们设置一个较小的学习速率微调OS-CNNs：学习速率从10^{-3}开始，5k次迭代后减少到10^{-4}，10k次迭代后减少到10^{-5}，经过12k次迭代后训练过程结束。为了加快培训过程，我们使用Multi-GPU扩展版本，这是在网上公开。

3．探索OS-CNNs

在前一节中我们已经介绍了OS-CNNs的架构和实现细节。在本节中，如图3所示，我们将专注于描述对OS-CNN激活不同层次的探索过程，努力提高识别性能。

3.1情况1：OS-CNN预测

直接使用CNN网络的输出(softmax层)作为最终的预测结果是利用OS-CNNs进行文化事件识别最简单的方法。具体地说，给一个图像，其识别评分计算如下:

$$s_{os}(I)=\partial_o s_o(I)+\partial_s s_s(I)$$

其中$s_o(I)$和$s_s(I)$分别为物体网和场景网的预测分数，∂_o和 ∂_s分别为物体网和场景网的融合权重。当前实现的过程中，物体网和场景网的融合权重被设为相等置。

3.2 情况 2：OS-CNNs 对全局特征值进行预训练

OS-CNNs 文化活动识别的另一种方法是将它们作为通用的特征提取，使用它们来提取图像区域的全局特征。我们通常提取完全连接层的激活部分，因为它们非常紧凑并且有识别力。在本例中，我们只使用没有微调的预训练模型。具体地说，给定一个图像区域，我们基于 OS-CNNs 提取全局特征，表示如下：

$$\Phi_{os}^{p}(I)=[\beta_o\Phi_o^p(I),\beta_s\Phi_s^p(I)]$$

$\Phi_o^p(I)$和$\Phi_s^p(I)$为预训练后激活过的物体网和场景网，β_o 和β_s为物体网和场景网的融合权重。当前实现的过程中，物体网和场景网的融合权重被设为相等置。

3.3 情况 3：OS-CNNs 对全局特征值进行预训练和微调

在以上的情况中，OS-CNNs 只是对大规模数据集对象识别和场景识别的预训练，并直接应用于小型事件识别数据集。然而，在一个预训练过的 OS-CNNs 上微调目标数据可以提高性能[8]。我们考虑在事件识别数据集中微调 OS-CNNs，结果图像的特征成为数据集中的特有的。微调的过程后，我们获得以下微调后 OS-CNNs 的全局特征的表示：

$$\Phi_{os}^{f}(I)=[\beta_o\Phi_o^f(I),\beta_s\Phi_s^f(I)]$$

$\Phi_o^f(I)$和$\Phi_s^f(I)$分别为从微调后的物体网和场景网获取的 CNN 激活，β_o和β_s为物体网和场景网的融合权重。当前实现的过程中，物体网和场景网的融合权重被设为相等置。

3.4. 情况 4：OS-CNNs 的局部特征+费舍尔向量

在先前的两种情况中，我们用 OS-CNNs 提取一个图像区域的全局特征。虽然这种全局特性是机密的以及有识别力的，但是它可能缺乏描述当地特征和细节信息的能力。受激励于最近成功的基于深度卷积描述的视频动作识别，我们调查研究了卷积层激活的有效性。卷积层特性也证明了它运用于图像任务中的有效性，例如

	Object nets	Scene nets	OS-CNNs
Scenario 1			
softmax	73.1%	71.2%	75.6%
Scenario 2			
fc7	67.2%	63.4%	69.1%
Scenario 3			
fc6	80.6%	76.8%	81.7%
fc7	81.4%	78.1%	82.3%
Scenario 4			
conv5-1	77.6%	76.6%	78.9%
conv5-2	78.6%	76.2%	79.6%
conv5-3	79.4%	76.1%	80.2%
conv5-4	78.4%	75.6%	79.7%
Fusion			
conv5-3+fc7	82.5%	79.3%	83.2%

表 1：全局特征和局部特征的实践识别性能。

对象识别等[7]，场景识别[5]和纹理识别[3]。在这种情况中，OS-CNNs首次在大规模图像中的预训练为Netand Places 205 数据集，然后在事件识别中运用微调，就像在情况3。

具体地说，给定一个图像区域，我们首先提取OS-CNNs(卷积层的激活）的卷积特性标志图，其中$C(I) \in R^{n \times n \times c}$，该试中 n是功能图的大小，C是特征信道号。卷积特性标志图中的每个激活值都对应于一个原始图像的局部区域，因此我们称这些激活卷积层为OS-CNN局部代表。

在提取了OS-CNN的局部代表后，我们利用[22]中提出的两种归一化方法，即渠道规范化方法和空间标准化方法，来预处理这些卷积特征标志图，使其转换为$\tilde{C}(I) \in R^{n \times n \times c}$卷积特征标志图。更多关于这两个归一化方法的细节不属于本文的讨论范围，可参考资料[22]。归一化的CNN激活层$\tilde{C}(I)(x,y,:) \in R^{c}$中的每一个坐标点(x,y)都被称为深度卷积描述符（TDD），结果显示，在资料[22]中，这两种标准化方法能高效改善CNN的局部性能表征。

Rank	Team	Score
1	VIPL-ICT-CAS	85.4%
2	FV	85.1%
3	**MMLAB (ours)**	84.7%
4	NU&C	82.4%
5	CVL_ETHZ	79.8%
6	SSTK	77.0%
7	MIPAL_SUN	76.3%
8	ESB	75.8%
9	Sungbin Choi	62.4%
10	UPC-STP	58.8%

表2：我们队的提交结果和其他队的提交结果的比较，2015年ICCV-LAP挑战赛中我们获得了第三名的成绩。

因此，在我们的实验探索中，我们将使用标准化方法。

最后，由于采用费舍尔向量（详见资料[16]）在物体识别和动作识别上良好的表现，我们采用它编码这些TDD为全局特征。特别是，根据我们之前对编码方法的综合性研究[15]，我们首先利用PCA使得TDD的维数减少到64。然后每个TDD都用K个组件(K设置为256)量化为一个高斯混合模型(GMM)。第一和第二顺序差异每个TDD x∈R 64及其高斯中心μk块u k和k中聚合。最后费舍尔向量表示了将这些模块进行拼接:

$$\Phi_{f_V}(I) = [u_1, v_1 \cdots u_k, v_k]$$

对于OS-CNNs而言，局部特征的费舍尔向量定义如下:

$$\phi^{f}_{os-f_V}(I) = [\beta_o \phi^{f}_{o-f_V}(I), \beta_s \phi^{f}_{s-f_V}(I)]$$

其中，$\phi^{f}_{o-f_V}(I)$表示物体网中局部特性的费舍尔向量，$\phi^{f}_{s-f_V}(I)$分别表示表示场景网中局部特性的费舍尔向量。β_o

和β_s为物体网和场景网的融合权重。当前实现的过程中，物体网和场景网的融合权重被设为相等置。

3.5 线性分类器

在之前的三个场景，所有特征$\phi(I)$是用于构造一个线性分类器$s(w,I)=w\phi(I)$，w是线性分类器的重量。在我们的实现过程中，我们选择 LIBSVM[1]作为分类器学习重量w，参数 C，平衡调整和损失，被设置为 1。值得注意的是，所有这些特征需要先被规范化，再被输入向量机进行训练。对于 OS-CNN 全局特征，我们用ℓ_2正规化，对于局部特征，我们使用内部规范化和ℓ_2正规化。

4. 实验

在这个部分中，我们首先描述在 2015 年 ICCV-LAP 挑战赛数据集的文化活动识别。然后，我们来分析实验结果，我们提出在 OS-CNNs 中用不同的特征来获取一个确定的 LAP 数据集，并进行数据验证。最后，我们在 2015 年 ICCV-LAP 挑战赛中描述了我们的解决方案。

4.1 数据集和评估协议

数据集。2015 年 ICCV-LAP 挑战赛包含了一系列的文化事件识别，同时，它也提供了一个事件识别数据集。这个数据集包含了从从两大搜索引擎（谷歌搜索和必应搜索）搜索出的图片。总共有来自不同国家的 100 个不同的事件类别（其中 99 个事件类别和 1 个背景类别），其中部分图片在图 1 中可以看到。从这些样本中，我们可以看到文化时间识别是很复杂的，因为人们的服装、姿势、物体和场景环境被同时利用来作为事件理解的线索。这个数据集被分为三个部分：发展数据（14，332 个图片）、确认数据（5，704 个图片）和评估数据（8，669 个图片）。由于我们不能利用评估数据的标签，我们只能用发展数据训练我们的模型并用确认数据公布我们的结果。

评估协议。主要的定量措施是基于精密记忆曲线的。它们利用曲线下的这片区域作为平均精度的估算，这是由数值积分计算。最后，他们平均所有事件类的每个类 AP 值并采用均值平均精度(mAP)作为最终排名标准。因此，在我们的探索实验中，我们报告结果评估时测量了每一个类的 AP 值和所有类和映射值。

4.2 结果和分析

设置。在这个探索实验中，我们使用 VGGNet-19 作为 OS-CNN 的网络结构。我们从两个完全连接层 (fc6 fc7)提取激活 OS-CNN 的全局特征，激活四个卷积层(conv5-1、conv5-2 conv5-3，conv5-4)作为 OS-CNN 的局部特征。应该注意的是，我们选择线性单元(ReLUs)修正后的激活层。我

们用 ℓ_2 正规化进程处理 OS-CNN 全局特征以进一步更好的支持向量机训练。对于 OS-CNN 局部特征的费舍尔向量表示，我们采用内部正规化和和 ℓ_2 正规化处理。

分析。我们第一次公开的数值结果如表 1 所示。从这些结果，可以得出几个结论如下:

•我们可以看到在文化事件识别方面物体网比场景网更加有优势，这可能意味着在文化活动的理解方面，物体线索比场景线索扮演更重要的角色。

•我们观察到 OS-CNNs 对于事件识别是很高效的，因为它从图像中提取物体和场景信息。无论采用什么方案，它总能实现性能优越的物体网和场景网。

•我们可以注意到，结合微调特性和线性 SVM 分类器(情况 3)相比直接使用将 softmax 函数的输出值(情况 1)能够获得更好的性能比。这一结果可

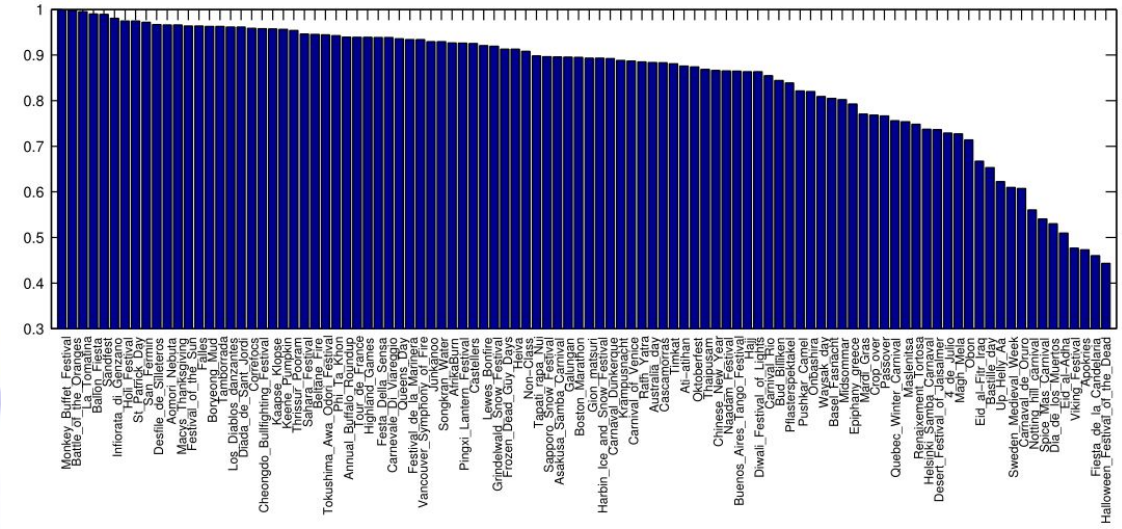

图 4：在 ICCV-LAP 比赛的数据集中，OS-CNN 结合全局特征和局部特征后测得每类的 AP 值

能归因于这样一个事实：当训练图像的数量相对较小时，在训练过程中 CNNs 很容易过度契合训练样本。

•比较微调特性(情况 3)和预训练特性(场景 2)，我们可以得出这样的结论:对于提高识别性能而言，将微调作用于目标数据集是非常有用的，这一看法和[8]中的结果一致。

•比较 OS-CNNs 的局部特征(场景 4)和全局特征(场景 3)，我们看到全局特征可以实现更高的识别精度。

•在 OS-CNNs 中，我们进一步结合全局特征(fc7)与局部特征(conv5-3)，发现这种组合能够提高最终的识别性能。这个性能改进表明，CNNs 的不同层捕获被不同级别的抽象的原始图像。这些功能在激活不同层时相辅相成。

如图 4 所示，我们还画出所有事件类的 AP 值。从这些 AP 值中，我们看到猴子自助餐节日和战斗的橙子这两个事件可以达到最高的性能(100%)。这个结果可能归因于这样一个事实:在这两个事件有具体有特征

的物体。同时，我们注意到一些事件类获得 AP 值很低，如纪念死者的节日、万圣节嘉年华、维京人的节日。这些文化事件类的 AP 值低于 50%。一般来说，这些复杂的事件类别没有特定的物体和场景，而且这些类从视觉外观的角度很容易与其他类混淆，如图 5 所示。

在图 5 中，我们想象一些识别的例子。在第 1 行我们给了八个例子，这些可以被我们的方法成功预测，比如 Pummpking， AfrikaBurn 之类的类别。与此同时，在例子的 2、3、4 行中，我们很有信心的提供了一些很容易失败的案例。从这些错误的预测实例，我们看到这些失败的原因是很合理的，因为和一些文化活动在分类时本就有很大的迷惑力。例如，亡灵节和万圣节在化妆和服装上很相似。圣火节和维京海盗节人就会有相似的节日礼服和节日物品。哈尔滨冰雪节和札幌冰雪节的场景环境比较相似和颜色外观也较为相同。中国的春节和平溪元宵节包含相似的对象。总之，这些图 5 中的例子表明事件识别是很复杂的，有些类别之间只存在细微的差别。

4.3 挑战的结果

我们对此进行最后的评估，我们合并开发数据(14332 张照片)和验证数据(5704 张照片)到一个训练数据集(20036 张照片)中，并在这个新的数据集中重新培训我们的 OS-CNN 模

图 5：我们的方法成功和失败的例子(第一行为我们成功预测的 8 幅图像，其他未我们错误预测的 24 幅图像)

型。最后，我们向 ICCV-LAP 提交的结果是基于我们的培训模式产生的模型。

根据上面的实验探索，我们得出这样的结论:OS-CNN 全局特征和局部特征相辅相成。因此，我们选择从 fc7 和 conv5-3 层结合激活，以保持性能和效率之间的平衡。与此同时，我们之前的研究表明，谷歌网是对 VGG 网[23]的补充。因此，我们也利用谷歌网的 OS-CNN 来提取全局特征。总之，我们面临的挑战有三个解决方案:(i)

VGGNet-19 的 OS-CNN 局部特征提取(ii) VGGNet-19 的 OS-CNN 全局特征提取(3) VGG 网的 OS-CNN 全局特征提取。

挑战结果如表 2 所示。我们可以看到我们的方法表现居于前列，我们的 mAP 值非常接近这一挑战赛中的最佳性能(84.7% vs 85.4%)。关于计算成本，我们的实现是基于 CUDA7.0 和 Matlab 2013 a，在配备 8 核 CPU、48 g RAM， K40 GPU 的前提下，每个图像需要大约要 1 s 的处理事件。

5. 结论

在本文中，为了找到更好的文化事件识别方式，我们全面研究 OS-CNNs 的不同层次。具体来说，我们通过设计的四种类型情况阐明了 OS-CNNs 在不同层次适应文化事件识别的高效性。从我们的实证研究中，我们证明了激活后的 CNN 卷积层和完全连接层是互补的，并且它们的结合能够提高识别性能。最后，我们想出了一个通过使用 OS-CNNs 来解决问题的方案，并在 ICCV-LAP 挑战赛上获得第三名。在未来，我们可以考虑如何利用更多的视觉线索，如人类的姿势、服装、物体和场景等，并以一种系统化的方式识别的静态图像中德事件。

声明

这项工作由英伟达公司赞助，同时这项工作的一部分被中国国家自然科学基金委(91320101，61472410)、深圳基础研究规划组(JCYJ20120903092050890，JCYJ20120617114614438，JCYJ20130402113127496)、中国科学院 100 大人才培养计划项目组和广东创新研究团队项目组(No.201001D0104648280)赞助。

参考文献

[1] C. Chang and C. Lin. LIBSVM: A library for support vector machines. ACM TIST, 2(3):27, 2011. 5

[2] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In BMVC, pages 1–12, 2011. 4

[3] M. Cimpoi, S. Maji, I. Kokkinos, and A. Vedaldi.Deep filter banks for texture recognition description and segmentation. CoRR, abs/1507.02620, 2015. 4

[4] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. ImageNet: A large-scale hierarchical image database.In CVPR, pages 248–255, 2009. 1, 3

[5] M. Dixit, S. Chen, D. Gao, N. Rasiwasia, and N. Vasconcelos. Scene classification with semantic fisher vectors. In CVPR, pages 2974–2983, 2015. 4

[6] S. Escalers, J. Fabian, P. Pardo, X. Baro, J. Gonzlez H. J. Escalante, and I. G. and. Chalearn 2015 apparent age and cultural event recognition: datasets and results. In ICCV, ChaLearn Looking at People workshop, 2015. 2, 5

[7] B. Gao, X. Wei, J. Wu, and W. Lin. Deep spatial pyramid: The devil is once again in the details. CoRR,abs/1504.05277,

2015. 4

[8] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik.Rich feature hierarchies for accurate object detection and semantic segmentation.CoRR, abs/1311.2524,2013. 4, 6

[9] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. CoRR, abs/1502.01852, 2015. 1, 3

[10] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. CoRR, abs/1502.03167, 2015. 3

[11]Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe:Convolutional architecture for fast feature embedding.CoRR, abs/1408.5093, 2014. 3

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In NIPS, pages 1106–1114, 2012. 1, 3

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324,November 1998. 1

[14] L. Li and F. Li. What, where and who? classifying events by scene and object recognition. In ICCV,pages 1–8, 2007. 1

[15] X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. CoRR, abs/1405.4506, 2014. 4, 5

[16] J. Sánchez, F. Perronnin, T. Mensink, and J. J. Verbeek. Image classification with the fisher vector: Theory and practice. International Journal of Computer Vision, 105(3):222–245, 2013. 4

[17] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. CoRR, abs/1312.6229, 2013. 3

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014. 1, 3

[19] C. Sun and R. Nevatia. Large-scale web video event classification by use of fisher vectors. In WACV, pages 15–22, 2013. 4

[20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed,D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. CoRR, abs/1409.4842, 2014. 1, 3

[21] L. Wang, S. Guo, W. Huang, and Y. Qiao. Places205VGGNet models for scene recognition. CoRR, abs/1508.01667, 2015. 1, 3

[22] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In CVPR, pages 4305–4314, 2015. 4

[23] L. Wang, Z. Wang, W. Du, and Y. Qiao. Object-scene convolutional neural networks for event recognition in images. In CVPR, ChaLearn Looking at People 2015 workshop, pages 30–35, 2015. 1, 2, 7

[24] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream convnets. CoRR, abs/1507.02159, 2015. 3

[25] X. Wang, L. Wang, and Y. Qiao. A comparative study of encoding, pooling and normalization methods for action recognition. In ACCV, pages 572–585,

2012. 4

[26] Y. Xiong, K. Zhu, D. Lin, and X. Tang. Recognize complex events from static images by fusing deep channels. In CVPR, pages 1600–1609, 2015. 1

[27] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In ECCV, pages 818–833, 2014. 1, 3

[28] B. Zhou,A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In NIPS, pages 487–495,2014. 1, 3