指导教师：杨 涛　　提交时间：2016/3/17

CVPR2015 Paper Translation

No: 01
姓名：陈 晨
学号：2013302519
班号：10011303

不同以往的 CNN 视频表示事件检测方法

徐 忠 文，杨 义，Alexander G.Hauptmann

悉尼技术大学 QCIS 实验室 卡内基梅隆大学计算机科学学院

zhongwen.xu@student.uts.edu.au yee.i.yang@gmail.com alex@cs.cmu.edu

摘要

在本文中,我们提出一个可识别的视频表示在有限的可利用硬件资源下,用于检测大规模视频事件数据集。本文的重点是有效地利用深度卷积神经网络（CNN）推进事件检测，帧级静态描述符可以由现有的 CNN 工具包提取。本文对卷积神经网络视频表示的推论做出两点贡献。第一点,虽然平均池和最大池一直是聚合帧级静态特性的标准方法,但我们还发现利用适当的编码方法可以显著提高性能。其次，我们建议使用一组潜在的概念描述符作为帧描述符,这丰富了视觉信息，同时保持可负担的计算量。集成这两个贡献所带来的效果是基于极大视频数据集的事件检测达到了最高水平。相比于提高了密集的轨迹,这已被公认为事件检测的最好视频表示。我们新的表现提高了 TRECVID 的平均正确率均值,14 组测试数据由 27.6%提高到 36.8%，13 组测试数据由 34.0 提高到 44.6%。

1.介绍和相关工作

复杂事件检测[1,2],它是以从Youtube上采集到的大量类似“装修房屋”缓慢变化的视频事件检测处理为目标,最近在计算机视觉领域的研究中吸取了很多关注。相比于视频的概念分析，例如动作识别，事件检测更加困难，主要是因为一个事件是更复杂的,从而有更大的内部变化。例如,一个“求婚”事件可能发生在室内或室外，并可能包括多个概念，如戒指（对象），跪（行动）和接吻（行动）。

最近的研究工作表明，结合多种功能，包括静态外观特征[9,25,41]，运动特征[23，7，43，44，33]和[28]的声学特征，事件检测中产生良好的性能,作为证明的关于顶部球队在TRECVID多媒体事件检测（MED）竞争[3，22，29，30]和研究论文[26，31，40，45]的报告，已经解决了这个问题。利用额外的数据辅助的复杂事件检测，研究者提出了利用来源于其他资源的“视频属性”促进事件检测[27]使用，或在训练样本很少[46]时利用相关的典范。在本文中，我们专注于提高视频表示，这种新的方法可以很容易地被馈送到这些框架，以进一步提高其性能。

密集的轨迹及其增强版本(IDT)[44]，由于其在其他功能如运动特征的规定[23]

和静态特征稠密SIFT [3] 的优越性能，在近年来已经成为复杂事件检测的主流。尽管性能好，沉重的计算成本极大地限制了改进版的密轨迹的使用规模。在2014年的TRECVID MED评比中 [2]，美国国家标准与技术研究所（NIST）引入了一个大视频采集技术，可以用于200000个总时间长达8000小时的视频。并联1000个内核，需要大约一个星期来提取改进版稠密轨迹在TRECVID MEDE-val14 所采集的200000个视频。即使在空间的重新调整和时间点简化下进行的采样处理，它仍然需要500个内核一周提取的特点[3]。由于不堪承受的计算成本，这对于一个相对较小的研究组与地中海有限的计算资源来说是非常困难的。对于负担得起的计算资源，它为重要的复杂事件检测提出一种有效表示。例如，一台机器，而在同一时间，试图取得更好的性能。

一个本能的想法是利用深度学习的方法，尤其是卷积神经网络（CNNs），因为它的准确图像分析能力和快速处理能力上的压倒性优势，这些能力是通过利用大规模并行处理能力的GPU的实现的[21]。然而，据TRECVID MED 2013时的报道，基于视频表示的CNN事件检测表示的性能比改进版的稠密轨迹性能要差一些[22]，[3]，如表1所示。一些技术性问题仍未解决。

	MEDTest 13	MEDTest 14
IDT [44, 3]	**34.0**	**27.6**
CNN in Lan *et al.* [22]	29.0	N.A.
CNN_{avg}	32.7	24.8

表1. 性能比较（平均精度的百分比）。Lan et al. 是TRECVID MED 2013中应用CNN特征的唯一尝试。我们CNNavg的结果来自CNN帧级描述符的平均池化。

首先，CNN技术需要大量的标记的视频数据，以培养良好的模型从零开始。大型的TRECVID MED数据集（例如，MEDTest 13 [1]和 MEDTest 14[2]）每个事件只有100的正例，还有许多无关的空视频。标记的视频的数量小于的采集的体育视频数量[20]。此外，如上所述[46]，事件视频是完全不同的活动视频，所以使用的活动类数据集训练事件检测模型是没有太大作用的。

其次，在处理一个领域内的特殊任务，只有很少的的训练数据时，微调[12]是一种有效的技术用来适应为新任务所建立的图像网络预训练模型。然而，在帧级的视频级别的事件标签是相当粗糙的，换言之，不是所有的帧必须包含事件的语义信息。如果我们对每一帧使用粗糙的视频等级标签，帧级微调的性能几乎没有提高；这是我们初步的实验1验证。

最后，给定的帧级CNN的描述符，我们需要产生一个有区别的视频级别表示。平均池化是一种标准的[32]，[3]用于静态局部特征的方法，以及用于CNN描述的方法【22】。表1显示了改进的稠密轨迹法和CNN平均池化表示法性能比较。我们提供关于Lan et al.【22】的性能以供参考。我们可以看到，CNN平均池化表示不能得到比手工制作的改进版稠密轨迹更好的性能，这是相当不同于其他视觉观测任务的[12]，[13]，[6]。

本文的贡献有三部分。首先，这是第一个利用编码技术以产生基于CNN描述符的视

频表示。其次，我们建议使用一组潜在概念描述框架的描述，进一步使输出基于更深层网络阶段的聚合化的多个空间位置多样化。一个遵循CNN描述符提取的视频帧转发方法。有了这两个贡献，在基于大量MED数据的最先进视频表示，该CNN表示实现了超过30%的相对改善，并且可以由安装着4个GPU卡的单机花费2天时间管理。此外，我们建议使产品基于CNN视频表示量化【15】以加快执行（事件搜索）时间。根据我们广泛的实验，我们表明，该方法有效的降低了输入/输出成本，从而使事件预测的速度更快，同时保持几乎相同的精度水平。

2. 预备知识

除非另有说明，本文的工作是基于由[37]发布的网络体系结构，即配置16个重要性层次的 VGG ILSVRC 2014取胜的任务分类方案。前13重要性层次是卷积层次，其中五个是最大池化层。最后的三层是完全连接层。在本文的其余部分，我们遵循符号在[6,12]:池化中。指最后池化层的激活物，fC6和fC7指的激活物分别是第一和第二的全连接层。尽管结构[37]所处的层次比经典的CNN结构层次更深[21],[6],[12],池的下标。fC6和fC7符号还是一致的，如果我们把卷积对应层至于最大池化层之间的“convo-lutionallayer”[37]。我们在纠正即线性单元之前利用激活物。(fC6和fC7),之后它们(即fC6_felu和fc7_felu),因为我们观察这两个x的性能差异。

3. 视频CNN表示

我们开始用Caffe工具包[18]与[37]共享模型中提取的帧级CNN描述符。然后，我们需要生成的基于帧级的CNN描述符的视频等级矢量表示法。

3.1基于CNN描述符的平均

在国家的最先进的复杂事件检测系统[3]，[32]，实现图像的视频表示，局部描述子提取依赖于单个帧单独的标准方式，如下：（1）获得单个帧的描述；（2）应用框架描述规范化；（3）平均池帧描述符获取视频表示，即 $x_{\text{video}} = \frac{1}{N}\sum_{i=1}^{N} x_i$, x_i 是帧级描述和公式是从视频中提取帧的总数；（4）重新规范对视频表示。

最大池的帧来生成视频表示是一种替代的方法，但它是不是典型的事件检测。我们观察到类似的性能，平均池，所以我们省略了这种方法。

3.2池在CNN视频描述符

视频池计算整个视频的视频表示，通过集中所有的视频帧的视频。36，[35]和向量的局部聚集描述符(弗拉德)[16],[17]已被证明有很大的优势，超过袋的话（弓）[38]在本地描述符编码方法。已经提出了用于图像分类和图像检索的图像分类和图像检索，编码图像的局部描述符，如密集的筛选和直方图梯度（猪）的图像分类和图像检索的建议进行了建议。人们也曾尝试应用Fisher向量和弗拉德对局部运动描述符如光流直方图(HOF)和运动边界直方图(MBH)在视频捕获的运动信息。我们的知识，这是第一个工作的美国有线电视新闻网的视频

集中的描述，我们扩大了编码方法，从局部描述符到美国有线电视新闻网的描述在视频分析。

3.2.1费舍尔向量编码

费舍尔向量编码[35],[36],高斯混合模型(GMM),其中K组件可以表示为

$\Theta = \{(\mu_k, \Sigma_k, \pi_k), k = 1, 2, \ldots, K\},$

μ_k, Σ_k, π_k分别是均值,方差和含k的参数,训练CNN帧描述符的水平,分别。鉴于CNN的$X = (x_1, \ldots, x_N)$描述符从视频中提取,我们为k均值和协方差偏移向量分量为:

$$u_k = \frac{1}{N\sqrt{\pi_k}} \sum_{i=1}^{N} q_{ki} \left(\frac{x_i - \mu_k}{\sigma_k} \right)$$
$$v_k = \frac{1}{N\sqrt{2\pi_k}} \sum_{i=1}^{N} q_{ki} \left[\left(\frac{x_i - \mu_k}{\sigma_k} \right)^2 - 1 \right], \quad (1)$$

q_{ki}是后验概率。连接的u_k和v_k的所有K组件,我们形成了费舍尔矢量视频大小的2D’K,公式在哪里PCA预处理后CNN描述符x_i的维数公式。PCA预处理是必要的为了更好的适应在对角协方差矩阵的假设[36]。功率归一化,通常签署平方根(SSR)公式符号$z = \text{sign}(z)\sqrt{|z|}$,然后和归一化$\ell_2$应用于费舍尔向量[35],[36]。

3.2.2弗拉德编码

弗拉德编码[16],[17]可以被视为一个简化版的费舍尔向量编码。与K粗中心$\{c_1, c_2, \ldots, c_K\}$生成的k - means,我们可以获得关于中心向量公式的区别:

$$u_k = \sum_{i:\text{NN}(x_i)=c_k} (x_i - c_k), \quad (2)$$

NN(x_i)公式表明x_i的距k最近的粗略中心。

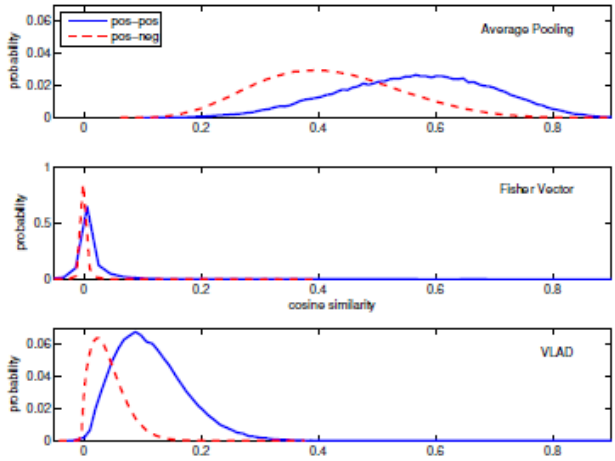

图1。在积极的正余弦相似度概率分布（蓝色和平原）和负（红色虚线）使用FC7特点视频，平均池（上），用256分量的GMM Fisher向量编码（中期），并与弗拉德用256中心编码（下）。由于不同的概率的概率的鱼矢量是非常不同的平均池和弗拉德，我们只使用一致的轴，平均池和弗拉德。这个数字是最好的颜色看。

弗拉德编码向量和尺寸连接D’K所有u_k得到的K是中心。弗拉德的另一个变体叫做VLAD-k,它扩展了最近的中心再与中心,表现出良好的性能在动作识别[19],[34]。没有规范,我们利用VLAD-k默认k=5。除了权力和归一化u_k,我们应用intra-normalization[4]弗拉德。

3.2.3定量分析

鉴于以上三种方法,我们需要找出哪一个是最合适的美国有线电视新闻网的描述。为此，我们进行了一个分析实验的medtest 14训练集[2]研究了三种类型的视频表示，

即平均分担的判别能力，视频共享与Fisher向量，和视频共享与弗拉德美国有线电视新闻网描述符。具体来说，我们计算余弦相似度在积极的典范，所有的事件中（记为POS，POS）和余弦相似性之间的正样本和负样本（记为POS NEG）。结果如图1所示。一个好的表现，积极和消极的样本数据点应远离对方，即，余弦相似度的“PES NEG”应该接近于零。此外，应该有“零售”和“PES NEG的分布之间有明显的差异。

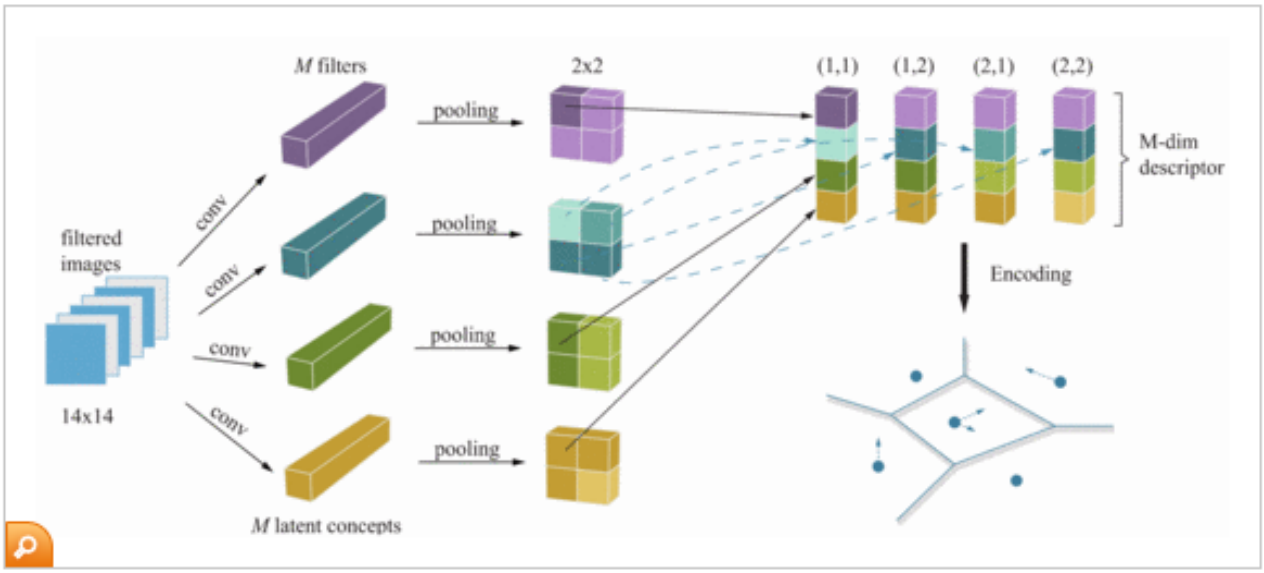

图2。编码过程的潜在概念描述。在过去的卷积层中，我们采用公式过滤公式的潜在概念分类。在最后的卷积层式过滤器（例如，一个长方体体积公式）在每一个卷积定位产生预测输出的最大池操作。然后，我们得到了不同大小的窗口的响应（在这个例子中，输出的大小是公式）为每个潜在的概念。颜色强度对应于每个滤波器的响应强度。最后，我们积累的公式过滤器在同一位置的潜在概念描述符的响应。每个维度对应一个潜在的概念。在获得所有的潜在概念的所有帧的描述符，我们然后应用编码的方法得到的最后的视频表示。这个数字是最好的颜色看。

平均池：在图1中，我们观察到，“posneg“余弦相似度分布远离零，这是高度表示阳性和阴性标本对很大一部分是相似的。此外，在大范围的[0.2，0.8]，大范围内的交叉的地区。这两种观测表明，平均池可能不是最好的选择。

费舍尔向量：虽然“pes-neg”相似分布非常接近于零，大部分“pos-pos”双也属于相同的范围。的分布没有明显区别“pos-pos”和“pos-neg”canad 被观察到。

Vzad：“pes-neg”对的分布更接近于零比一般池而“pos-pos”相似的一个相对较小的比例接近的峰“pes-neg”相似。

从上面的分析研究中，我们可以看到，弗拉德是最适合 CNN 描述符，因为弗拉德表示有最好的区别的能力，这也是符合 5.1 节的实验结果。

3.3Cnn潜在概念描述符

全层相比,pool5 包含空间信息。然而,如果我们按照标准的方式,平 pool5 成一个向量,该功能维度会很高,这将引起重计算成本。

具体来说,pool5 的特征维度是 $a \times a \times M$, a 是去年池层过滤的图像的大小和 M 是在过去的卷积层(在我们的例子中,a=7 和 M=512)。VGG 网络[37],pool5 特征向量的 25088-d,而 fc6 和 fc7 特征向量只有 4096 - d。结果,研究人员往往忽略掉 pool5[6],[13]一般特征。视频池方案的问题更严重,因为帧描述符高维度将导致不稳定问题[10]。

注意,卷积过滤器可以视为广义线性分类器在底层数据补丁,并且每个卷积滤波器对应于一个潜在的概念[24]。我们建议制定的一般特征pool5作为潜在的向量概念描述符,每个维度的潜在概念描述符代表特定的反应潜在的概念。每个过滤器在最后卷积层是独立于其他的过滤器。滤波器的响应是线性分类器的预测上的回旋的位置对应的潜在概念。那样,pool5层大小的$a \times a \times M$,可以转化为a^2潜在的概念与维度M描述符。每个潜在的概念描述符表示M过滤器为一个特定的反应池的位置。一旦我们获得潜在的概念描述符中的所有帧视频,然后我们申请一个编码方法来生成视频表示。在这种情况下,每一帧的帧包含a^2描述符,而不是一个描述符,如图2所示。

在[14],他等人声称聚合在更深的一层更兼容层次比种植在我们的大脑信息处理或包装原始输入,和他们建议使用空间金字塔池(SPP)层对象分类和检测,这不仅达到更好的性能,而且放松约束输入必须固定大小。不同于[14],我们不从头训练 SPP 的网络层,因为它需要更长的时间,特别是对一个非常深的神经网络。相反,在最后池层,我们采用多个窗口大小不同和进步没有再培训的 CNN。这样,视觉信息是丰富而只有边际计算成本补充说,我们通过网络转发帧只有一次提取潜在的概念描述符。

之后提取所有空间位置的 CNN 潜在的概念描述符中的每一帧视频,然后我们视频池适用于所有潜在的概念描述符的视频。在[14],我们运用四种不同的 CNN max-pooling 操作和获取 $(6 \times 6), (3 \times 3), (2 \times 2)$和$(1 \times 1)$输出为每个独立的卷积滤波,总共 50 空间位置为单个框架。潜在的维数概念描述符(512 - D)小于全层的描述符(4096 - D),在视觉信息丰富通过多种过滤图像上的空间位置。

3.4表示压缩

工程方面的一个快速事件搜索[2]大视频集合,我们可以利用技术,如产品量化(PQ)[15]压缩费舍尔向量或弗拉德表示。与 PQ 压缩、磁盘和内存的存储空间可以减少一个数量级以上,而性能仍然几乎相同。PQ 分解的基本思想代表 sub-vectors 成相等长度 B,然后在每个 sub-vector,K -意味着应用生成2^m中心作为代表点。所有 sub-vectors 被最近的中心和编码到近似指数最近的中心。这样,浮动原始数字位码 B 表示成为 m;因此,压缩比的$\frac{B \times 32}{m}$。例如,如果我们把 m=8 和 B=4,我们可以达到 16 倍减少存储空间。

针对预测压缩数据,而不是在原来的特性,我们可以分解学习线性分类器 w 与同等长度 B。与查表来存储 sub-vectors 之间的内积2^m中心和相应的公式,sub-vector 预测速度给你的视频可以加速$\frac{D}{B}$乘以查找操作和$\frac{D}{B}-1$添加操作为每个视频假设功能维度 D[36]。

4. 实验设置

4.1 数据集

在我们的实验中,我们利用最大的事件检测与 labels2 数据集,即 TRECVID MEDTest 13[1]和 TRECVID MEDTest 14[2]。他们介绍了 NIST TRECVID 所有参与者的竞争和社区开展实验研究。对于这两个数据集,分别有 20 个复杂事件,但 10 事件重叠。MEDTest 13 包含事件 E006-E015 和 E021-E030 E021-E040 虽然 MEDTest 14 事件。事件名称包括“生日聚会”、“自行车技巧”,等等。参考[1],[2]的事件名称的完整列表。在训练部分,大约有 100 积极的范例/事件,所有事件与约 5000 个视频分享-范本。测试部分有大约 23000 个搜索视频。视频每个集合的总时间约为 1240 小时。

4.2 特性的比较

报道[3]和与其他顶级公司的功能[30,29,22]2013 年地中海 TRECVID 竞争,我们可以看到改进的密集的轨迹相对原始茂密的轨迹有高超的优势(所使用的所有其他团队除了[3]),甚至是比方法,结合许多低级视觉特征[30],[29],[22]。提高致密轨迹提取局部描述符如轨迹,猪,霍夫,MBH,费舍尔向量然后被用于局部描述符编码成视频表示。[44],[3],我们首先减少每个描述符 2 倍的尺寸,然后利用 256 个组件生成钓鱼向量。我们评估四种类型的描述符在改善密集的轨迹,并报告描述符的最佳组合的结果,两个人描述符(猪和 MBH)有最好的表现。

此外,我们报告的结果中使用一些流行的特性TRECVID竞争参考,如煤断层[23],MoSIFT[7]和CSIFT[41],尽管他们的性能远远弱于改善茂密的轨迹。

4.3 评估的细节

在所有的实验中,我们使用线性支持向量机(SVM)与 LIBSVM 工具包[5]。我们在两个标准进行广泛的实验训练条件:100 年前,100 年积极的范例给出在每个事件和 10 例,10 给出积极的范例。我们在 100 年交货条件,利用 5 倍交叉验证选择正则化系数 C 在线性支持向量机的参数。在 10 例条件,我们遵循[22],C 在线性支持向量机设置为 1。

我们样品每 5 帧的视频和遵循预处理[21],[6]在 CNN 描述符提取。我们只从中心作物中提取的特性。CNN 描述符提取使用咖啡[18]与[37]最好的公开可用的模型,我们利用 vlfeat[42]生成费舍尔向量和弗拉德表示。

意味着平均精度(mAP)二进制分类应用于评估事件检测的性能根据 NIST 标准[1],[2]。

5. 实验结论

5.1 结果 Cnn 视频池的描述符

在本节中,我们展示了实验视频汇集 fC6 fC6。relu,fC7 fC7。relu。聚合之前,我们首先运用 PCA 与美白配方标准化 CNN 描述符。与当地的描述符,如猪,MBH 尺寸小于 200 - d,CNN 描述符有更高维度(4096 - d)。我们进行实验不同的降低维度,即。、128、256512 和 128 年,利用降低维度,最好的平衡性能和存储成本相应的特性,即。512 - d fC6 fC6_relu 和 256 - d fC7 fC7。relu。我们利用 256 组件费舍尔向量和 256 年弗拉德中心共同选择[36],[16]。我们将在 5.3 节的影响参数研究。主成分分析预测,对费舍尔向量组件 GMM,中心在 k - means 弗拉德从大约 256000 个采样帧在训练集。

因为我们观察到类似的模式在MEDTest 13和MEDTest 14 100例和10例,取MEDTest 14 100例为例,比较不同的表征,即平均池、视频池与费舍尔向量和视频池与弗拉德。从表2,我们可以看到,两个视频池与费舍尔向量和弗拉德展示伟大的优势平均池表示。CNN视频池的描述符,费舍尔向量编码并不比弗拉德表现出更好的性能。类似的观察表达[10]。我们怀疑CNN描述符的分布是完全不同于当地的描述符,如霍夫定律,我们将学习的理论性能降低的原因比弗拉德费舍尔向量在CNN视频池在未来的研究。

	fc_6	fc_6_relu	fc_7	fc_7_relu
Average pooling	19.8	24.8	18.8	23.8
Fisher vector	28.3	28.4	27.4	29.1
VLAD	**33.1**	32.6	**33.2**	31.5

表2.性能比较(地图比例)medtest 14 100例

我们比较弗拉德CNN与先进的功能描述符编码的性能改进的密集的轨迹(IDT)和平均池CNN描述符,如图3所示。我们还说明了性能最强的两个描述符在IDT(猪和MBH)。我们可以很清楚地看到,弗拉德编码CNN特性明显优于IDT和平均池在CNN描述符在所有设置。更多的引用,我们提供许多广泛使用的性能特性[29],[30],[22]MEDTest 14进行比较。MoSIFT[7]与费舍尔向量实现100年地图18.1%和5.3%在交货10例;煤断层与费舍尔向量[23]100年实现地图15.0%和7.1%在10例;CSIFT[41]与费舍尔向量实现100年地图14.7%和5.3%在交货10例。注意,与弗拉德CNN描述符编码,我们可以取得更好的性能与10例比相对贫穷MoSIFT等特性,煤断层,并与100例CSIFT !

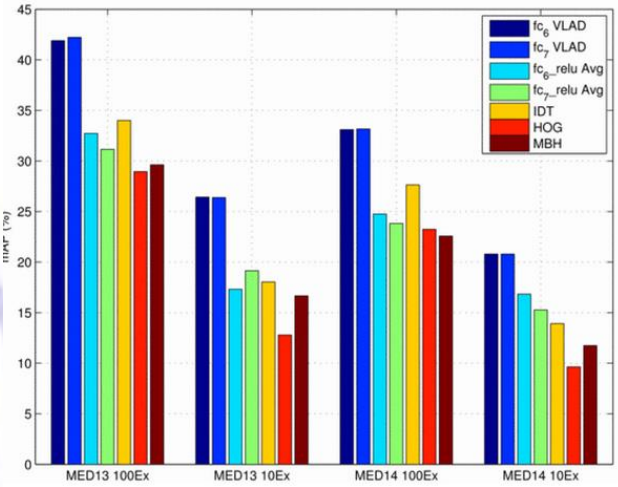

图3.在medtest 13日性能比较和medtest 14 100例和10例。这个数字是最好的颜色。

5.2 结果为 Cnn 潜在概念描述符与空间金字塔池

我们评估潜在的性能概念描述符(LCD)的原始 CNN 结构和结构与空间金字塔池(SPP)层插入验证的有效性。在描述符编码的概念,我们首先运用 PCA 与美白。进行降维的范围从 512 - d 维度如 32-D、64 d、128 d,和 256 - d,我们发现 256 - d 是最好的选择。我们看到一个类似的模式与视频池 fc 层表明费舍尔向量不如弗拉德视频池。我们省略费舍尔的结果向量将有限的空间。

我们显示的性能提出了潜在的概念描述符(LCD)在表3和表4。在100例和10例两个数据集,我们可以清楚地看到差距pool5特性平均池,这表明我们提出的优势利用 pool5小说。SPP层,弗拉德液晶编码(LCDvLAD + SPP)继续增加性能进一步从原始结构的公式。聚合在更深的阶段生成多级

	100Ex	10Ex
Average pooling	31.2	18.8
LCD_{VLAD}	38.2	25.0
LCD_{VLAD} + SPP	**40.3**	**25.6**

表3。池性能比较,medtest 13。FFFFFFFFF是弗拉德从原始CNN液晶编码结构,同时表明弗拉德编码与小型液晶层插入。

	100Ex	10Ex
Average pooling	24.6	15.3
LCD_{VLAD}	33.9	22.8
LCD_{VLAD} + SPP	**35.7**	**23.2**

表4。性能比较对pool5 medtest 14。表3 No-tations是一样的。

空间信息通过多种CNN max-pooling演示了原始CNN结构优势而只有最小的计算成本。SPP层允许一个通过转发的网络相比,多次应用空间金字塔的原始输入图像。

5.3 分析参数的影响

我们把弗拉德编码 fC7 特性下 MEDTest 14 100 例为例,见视频池过程参数的影响。

维度的主成分分析: fC7 的原始尺寸相比是相当高的局部描述符。有必要调查维度的影响在 PCA 预处理阶段,因为它是至关重要的实现更好的权衡性能和存储成本。表 5 显示,在超过 256 - D 维度,性能仍然是相似的,而编码 128 - D 显著损害性能。

Dimension	128-D	256-D	512-D	1024-D
mAP	30.6	**33.2**	33.1	33.2

表5所示。CNN的影响维度描述符后,PCA,固定在弗拉德。公式

在编码的中心: 我们探索各种数字中心的 K 在弗拉德,结果如表 6 所示。增加的 K,我们可以看到,歧视能力生成的功能改善。然而当 K=512,生成的向量可能过于稀疏,这有点不利于性能。

VLAD-k: 我们实验与传统的弗拉德,只与 k 最近的中心,而不是再中心。地图从 33.2% 下降到 32.0%。

K	32	64	128	256	512
mAP	28.7	29.7	30.4	**33.2**	32.1

表6所示。在弗拉德数量影响的中心,固定PCA尺寸256 - d。公式

功率归一化:

我们把SSR后处理在弗拉德fC7编码和测试功能。地图从33.2%下降到27.0%,从中我们可以看到SSR后处理的显著影响。

Intra-Normalization：我们把intra-normalizanon关掉。地图从33.2%下降到30.6%。

5.4产品量化压缩的结果

	original	$B=4$	$B=8$
mAP	33.2	33.5 (↑ 0.3)	33.0 (↓ 0.2)
space reduction	-	16×	32×

表7所示.性能变化分析与PQ弗拉德fc7编码压缩。F是sub-vectors PQ和长度的公式。

我们进行实验在弗拉德 fC7 编码的性能变化与产品量化(PQ)压缩。从结果在表 7 中,我们可以看到,PQ 有 B=4 保持性能甚至略有提高。当 B=8,性能略有下降。如果我们压缩 B=4,我们可以存储弗拉德编码 fC7 特性在 3.1 GB MEDEval 14,它包含 200000 个视频 8000 小时的持续时间。进一步与无损压缩技术如 Blosc3[8],我们可以存储整个集合的特点在不到 1 GB,可以读到一个正常的 SSD 硬盘在几秒钟。没有 PQ 压缩、存储功能的大小是 48.8 GB,这严重损害了执行时间由于 I / O 成本。利用压缩技术很大程度上节省了 I / O 成本预测程序,同时保护性能。

在我们的速度测试 MEDEval14 集合使用压缩数据而不是原来的特性,我们可以在 4.1 秒内完成预测 200000 个视频每个事件使用 20 个线程在 Intel Xeon e5 - 2690 v2 @ 3.00 GHz。

5.5 结果融合多个层从相同的模型

我们调查平均晚融合[39]融合不同层的预测结果与 PQ 压缩,即。,弗拉德编码 LCD SPP,fC6 fC7。从表 8 可以看出,简单的融合进一步推动性能超出了单一 MEDTest 13 和 MEDTest 14 层,并达到显著优势提高致密轨迹(IDT)。我们建议的方法进一步推动先进的性能,达到 30%以上的相对改善 100 例,65%以上的相对进步 10 例都具有挑战性的数据集。

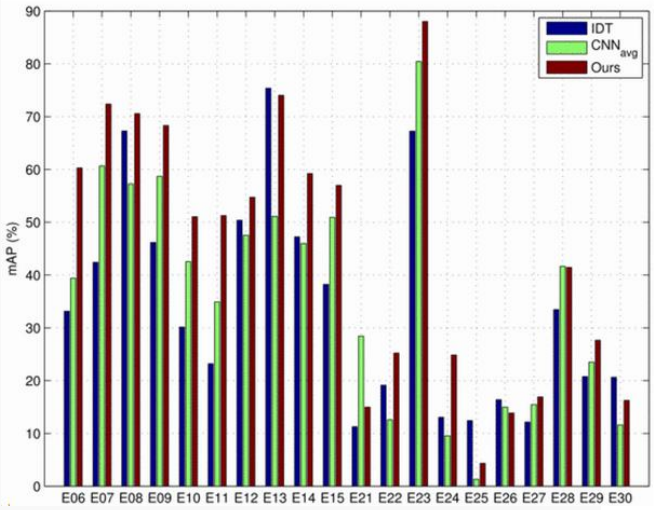

图4。Medtest 13 100例事件性能比较(在地图比例)。这个数字是最好的颜色。

	Ours	IDT	Relative Improv
MED13 100Ex	**44.6**	34.0	31.2%
MED13 10Ex	**29.8**	18.0	65.6%
MED14 100Ex	**36.8**	27.6	33.3%
MED14 10Ex	**24.5**	13.9	76.3%

表8所示。所有设置的性能比较;最后一列显示了相对改进我们的提议表示踊跃参与。

图4和图5显示了每个事件的地图比较100例设置MEDTest 13和MEDTest 14。我们提供的结果平均池CNN描述符与融合的三层后期,表示公式。我们建议的表现比其他两个强大基线在15 of20事件在MEDTest 13和14 of20事件在MEDTest 14中,分别。

5.6 比较先进的系统

我们比较 134 年 MEDTest 结果顶级表演者 TRECVID 地中海 2013 年竞争[3],[30],[22]。轴团队没有显示他们的表现在 MEDTest 13[3]。Natarajan et al . 100 例报告[30]地图 38.5%,17.9% 10 例整个视觉系统相结合的所有的低级视觉特征。局域网等。[22]报告 39.3%的地图 100 年前的整个系统包括非视觉特性,而他们在内部数据集进行了 10 例。我们的结果达到 44.6%地图地图上 100 例,29.8%在 10 例,大大优于竞争中表现最好的人将超过 10 种特性与复杂的计划。表明我们从其他形式表示是功能的补充,我们执行的平均融合后期提出表示 IDT 和 MFCC,

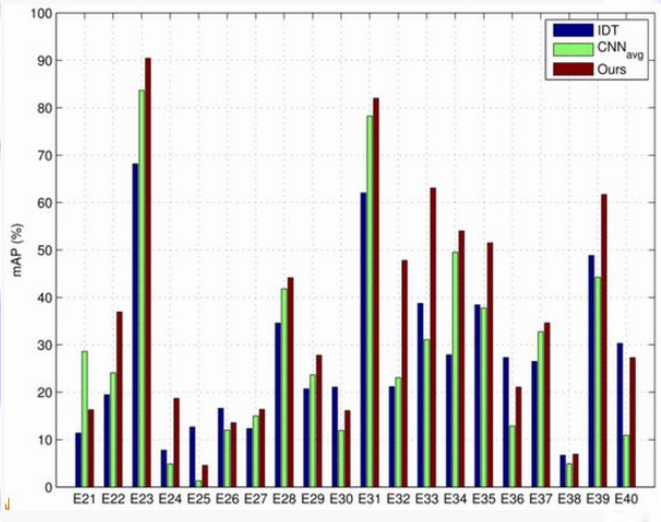

图5。Medtest 14 100例事件性能比较(在地图比例)。这个数字是最好的颜色。

并生成一个轻量级的系统与静态的,运动和声学特性,达到 48.6%地图 100 例,10 例 32.2%的地图。

6. 结论

TRECVID多媒体事件检测(地中海) 在特征提取和分类过程中已经遭受了巨大的计算成本。利用卷积神经网络(CNN)表示似乎是一个很好的解决方案,但从CNN描述符生成视频表示与图像表征有不同的特征。我们是第一个利用编码技术来生成视频表示CNN描述符的。并且我们建议潜在的概念描述符用来生成CNN描述符更恰当。对于快速事件搜索,我们利用量化的产品压缩视频表示并预测了被压缩数据大小。广泛的实验两大基于不同训练条件下的事件检测集合,展示出我们所提出的表示法的优势。我们取得了很有前景的性能, 它优于最先进的系统,结合了更多的功能。提出表示法由可拓展性,并且通过更好的CNN模型和/或适当的微调技术性能可以进一步提高。

7. 致谢

本文一部分由 973 计划 cb316400 973 项目支持,一部分由 ARC DECRA 项目支持,一部分由情报高级研究项目活动(IARPA)通过国家商业中心合同号 DIIPC20068 内政部支持。美国政府授权的复制和分发再版出于政府目的, 有版权注释免责声明:本文所包含的观点和结论是作者的,不应被解释为一定代表官方政策或支持,表示或暗示,IARPA,单位/ NBC,或美国政府。

我们十分感谢英伟达公司捐赠 GPU 用于这项研究的支持。

参考文献

[1] TRECVID MED 13. http://www.nist.gov/itl/iad/mig/med13.cfm. 1, 2, 5, 6

[2] TRECVID MED 14. http://www.nist.gov/itl/iad/mig/med14.cfm. 1, 2, 3, 5, 6

[3] R. Aly, R. Arandjelovic, K. Chatfield, M. Douze, B. Fernando, Z. Harchaoui, K. McGuinness, N. E. O'Connor, D. Oneata, O. M. Parkhi, et al. The AXES submissions at TrecVid 2013. 2013. 1, 2, 5, 8

[4] R. Arandjelovi´c and A. Zisserman. All about VLAD. In *CVPR*, 2013. 3

[5] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011. 5

[6] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014. 2, 4, 6

[7] M.-Y. Chen and A. Hauptmann. Mosift: Recognizing human actions in surveillance videos. *CMU TR*, 2009. 1, 5, 6

[8] R. G. Cinbis, J. Verbeek, and C. Schmid. Segmentation driven object detection with Fisher vectors. In *ICCV*, 2013. 7

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1

[10] M. Douze, J. Revaud, C. Schmid, and H. J´egou. Stable hyper-pooling and query expansion for event detection. In *ICCV*, 2013. 4, 6

[11] C. Gan, N.Wang, Y. Yang, D.-Y. Yeung, and A. Hauptmann. Devnet: A deep event network for multimedia event detection and evience recounting. In *CVPR*, 2015. 2

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2

[13] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *ECCV*, 2014. 2, 4

[14] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*. 2014. 4, 5

[15] H. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *TPAMI*,

33(1):117－128, 2011.
2, 5
[16] H. J′egou, M. Douze, C. Schmid, and P. P′erez. Aggregating local descriptors into a compact image representation. In *CVPR*, 2010. 3, 6
[17] H. J′egou, F. Perronnin, M. Douze, J. S′anchez, P. P′erez, and C. Schmid. Aggregating local image descriptors into compact codes. *TPAMI*, 34(9):1704－1716, 2012. 3
[18] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. *http://caffe. berkeleyvision. org*, 2013. 2, 6
[19] V. Kantorov and I. Laptev. Efficient feature extraction, encoding and classification for action recognition. In *CVPR*, 2014. 3
[20] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 2
[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1, 2, 6
[22] Z.-Z. Lan, L. Jiang, S.-I. Yu, et al. CMU-Informedia at TRECVID 2013 Multimedia Event Detection. In *TRECVID 2013 Workshop*, 2013. 1, 2, 5, 6, 8
[23] I. Laptev. On space-time interest points. *IJCV*, 64(2-3):107－123, 2005. 1, 5, 6
[24] M. Lin, Q. Chen, and S. Yan. Network in network. *CoRR*, abs/1312.4400, 2013. 4
[25] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91－110, 2004. 1
[26] Z. Ma, Y. Yang, N. Sebe, and A. Hauptmann. Knowledge adaptation with partially shared features for event detection using few exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(9):1789－1802, 2014. 1
[27] Z. Ma, Y. Yang, Z. Xu, S. Yan, N. Sebe, and A. G. Hauptmann. Complex event detection via multi-source video attributes. In *CVPR*, 2013. 1
[28] F. Metze, S. Rawat, and Y. Wang. Improved audio features for large-scale multimedia event detection. In *ICME*, 2014. 1
[29] G. K. Myers, R. Nallapati, J. van Hout, et al. The 2013 SESAME Multimedia Event Detection and Recounting system. In *TRECVID 2013 Workshop*, 2013. 1, 5, 6
[30] P. Natarajan, S.Wu, F. Luisier, et

al. BBN VISER TRECVID
2013 Multimedia Event Detection and
Multimedia Event Recounting
Systems. In *TRECVID 2013 Workshop*, 2013.
1, 5,
6, 8
[31] P. Natarajan, S. Wu, S.
Vitaladevuni, X. Zhuang, S. Tsakalidis,
U. Park, and R. Prasad. Multimodal
feature fusion for
robust event detection in web videos. In
CVPR, 2012. 1
[32] D. Oneata, M. Douze, J. Revaud, S.
Jochen, D. Potapov,
H. Wang, Z. Harchaoui, J. Verbeek, C.
Schmid, R. Aly, et al.
AXES at TRECVid 2012: KIS, INS, and MED.
In *TRECVID*
workshop, 2012. 2