指导教师：杨涛　　提交时间：2016-03-21

CVPR2015 Paper Translation

No: 01
姓名：袁 涛
学号：2013302508
班号：10011302

场景分类与语义费舍尔矢量

Mandar Dixit, Si Chen, Dashan Gao, Nikhil Rasiwasia and Nuno Vasconcelos

加州大学圣地亚哥分校

圣地亚哥高通公司

SnapDeal.com，印度

摘要

在卷积神经网络（CNN）的帮助下，将场景图像表示为一类语义（BOS）并练就识别物体的能力。这涉及到分类图像补丁使用网络和考虑到类的后验概率向量作为本地提取的语义描述符。图像的 BoS 是使用费舍尔总结矢量（FV）嵌入，它利用的属性这些描述符的空间。由此产生的表示是称为语义费希尔载体。二种实现语义 FV 正在进行着研究。第一个涉及建模 BoS 狄利克雷混合物和计算费舍尔梯度模型。鉴于非欧几里得的概率单纯形混合建模的难度，这种方法证明是不成功的。第二个实现派生使用语义的解释描述符作为多项分布的参数。就像指数级别的家庭的参数一样，这些可以投射到他们的自然参数空间。对于 CNN，这相当于显示使用 soft-max 层块描述符的输入。然后计算语义阵线作为高斯混合阵线在这些自然参数。这种形式呈现，优于其他方法，如阵线从中间 CNN 的特性层或分类器获得的适应(微调)CNN。该阵线是一个嵌入的对象分类概率。作为一个形象代表，因此，它是互补的特性从一个场景分类 CNN。两者的结合表示实现先进的显示结果在麻省理工学院的室内场景和 SUN 的数据集。

1. 简介

自然场景分类是一个具有挑战性的问题计算机视觉，因为大多数场景都是实体的集合(如对象)组织在一个高度可变的布局。这种高可变性的外表使灵活的视觉对于这个问题表示很受欢迎。许多作品提出了代表场景图像不整齐的集合，或者“袋”，在本地

提取视觉特征，如 SIFT 或 HoG[23, 5]。这被称为 bag-of-features(BoF) 表示。为目的的分类，这些功能集中到一个不变的图像表示称为费舍尔矢量(FV)[12, 25]然后用于判别学习。直到最近，bag-of-SIFT 阵线取得先进的场景分类的结果[30]。

最近，有很多兴奋替代图像表示，与卷积神经网络(CNNs)[19]学习，展示了令人印象深刻的大规模对象识别结果[16]。这促使许多研究者扩展 CNN 等问题行为识别[15],，对象定位[9]，场景分类[10, 39]和领域适应气候变化[8]。当前多层 CNN 可以分解为卷积层的第一阶段，第二个完全连接阶段，最终分类阶段。卷积层表现像素明智的转换，紧随其后本地化的池，可以认为是提取器的视觉特性。因此，卷积层输出是一个 BoF 表示。完全连接层然后这些特性映射到一个向量服从线性分类。这是费舍尔的 CNN 模拟向量映射。

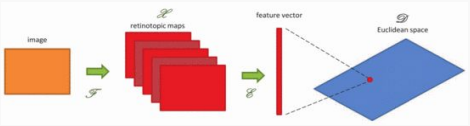

图 1 所示。袋的特性(BoF)。初步的功能映射公式,将图像映射到空间网膜代表特性公式。非线性嵌入公式用于地图这中间表示成一个特征向量的欧氏空间的公式

除了 SIFT 费舍尔向量和 CNN 层，存在一个不同的类被称为语义表示的图像映射。这些映射需要分类器的输出向量，或语义描述符，从图像中提取。几个作者认为这样表示的潜力[35, 27, 33, 17, 18, 3, 20]。例如，被用来描述语义表示对象属性[18]，表示场景的集合对象[20]和捕捉语境之间的关系类[29]。对于一些视觉任务，如哈希或大规模检索，一个全球性的语义描述符通常是首选[34, 4]。场景分类建议，另一方面，倾向于依靠本地语义提取图像描述符的集合，我们称之为袋语义(BoS)[33, 17, 20].。而基于 BoS 场景表示已优于低维 BoF 交涉，比高通常是不那么有效维 BoF-FV。这是由于这一事实[17]，1)本地或 patch-based 语义特征可以很复杂，和 2)很难组合成一个通用形象代表，类似于费舍尔向量。

在这项工作中，我们认为，高度精确的分类，比如 ImageNET 训练 CNN[16]的消除第一个问题。我们获得 BoS 图像表示使用该网络通过提取语义描述符(对象类的后验概率向量)从当地的形象补丁。然后我们考虑语义费舍尔向量的设计，这是一个扩展的标准这个 BoS 费舍尔向量。我们表明，这是很难实现直接在概率向量的空间，因为它的非欧几里得的本性。另一方面，如果从一个图像语义描述符被视为多项分布的参数，然后映射到它们的自然参数空间，一个健壮的语义阵线可以简单地使用标准的基于高斯混合编码转换的描述符[25]。对于 CNN，这个自然参数映射显示相当于 soft-max 的逆函数。因此，语义阵线可以实现为一个典型的(高斯混合)类型 pre-softmax CNN 输出。

语义阵线，代表了一个强大的嵌入的特性在本质上是相当抽象的。由于这个表示的不变性，这是一个直接结果的语义抽象，它较低的显示优于费舍尔向量层 CNN 特性[10]以及分类器通过微调 CNN 本身[9]。最后，由于对象语义是用于生产我们的形象代表，是互补的场景分类的特点提出网络(CNN)的地方[39]。实验表明，两个描述符的简单组合，产生一个先进的场景分类器在麻省理工学院室内和麻省理工学院阳光基准。

2. 图像模型

在本节中，我们简要回顾一下 BoF 和 BoS 图像分类。

2.1 袋的特性

SIFT-FV 分类器和 CNN 都是特殊情况的一般体系结构在图 1 中，俗称袋特性(转炉)分类器。我一个图像(左)，1 表示空间位置，它定义了一个初始映射 F 为一组网膜代表特征图谱颗(1)。这些地图保存图像的空间拓扑结构。映射 F 的常见例子包括密集的筛选，猪和 CNN 的卷积层。产生的转炉 F 是受到高度非线性将 E 嵌入到一个高维特征空间 d 这是欧几里得空间结构，在一个线性分类器 C 就足够良好的性能。

它可以辩称，这种架构可能总是需要场景分类。特征映射可以视为一个 F(非线性)当地的卷积过滤器的输入图像，如边缘探测器或对象的部分。这使得分类器的高度选择性，如区分行人和汽车。然而，由于其网膜代表自然，F 的输出很敏感现场布局的变化。将 E 嵌入的因此，non-retinotopic 空间 D 是必要的-方差等变化。此外，D 必须有一个空间欧氏结构与线性支持分类决定边界。

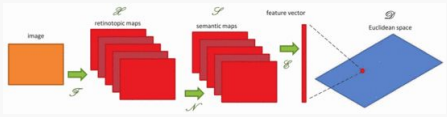

图 2。袋语义(bos)。网膜代表特征的空间公式首先映射到一个网膜代表语义空间公式,使用分类器的图像补丁。非线性嵌入公式用于地图这表示成一个特征向量的欧氏空间的公式。

CNN 最近基于分类器实现壮观结果 ImageNET 对象识别的挑战[16, 31]。他们的成功鼓励了许多研究人员使用所学到的特性和嵌入这些网络场景分类，取代传统的 SIFT-FV 基础架构[8, 32, 10, 22]。它无可置否的出现，他们的网膜代表映射 F，强烈非线性(多个迭代池吗在本质上和整流)判别(由于反向传播)[37]，一定程度的选择性不能浅相匹配的映射，如筛选。不太清楚，然而，使用嵌入的学习的优势 ImageNET 场景费舍尔向量的表示。场景图像表现出更大程度的内部类对象图像相比，变化的能力为现场平衡选择性不变性是至关重要的分类。而费舍尔向量导出使用混合物基础编码由设计不变，CNN 嵌入从几乎集中对象图像不太可能应对变化的场景。

2.2 袋的语义

语义表征是另一种体系结构图 1 的。他们只是每个图像映射到分类器的输出，使用这些作为后续的功能处理。由此产生的特征空间一般被称为语义特征空间。因为场景语义不同图像区域，场景分类需要一个空间局部语义映射。这是表示随 bag-of-semantics(BoS)表示。

如图 2 所示，BoS 类似于转炉，但基于语义描述符。其第一步是网膜代表映射 f .然而，而不是嵌入 E，这是紧随其后的是另一个网膜代表映射 N 到美国在每个位置 1，N 将转炉映射描述符提取从附近的 1 语义描述符。这个描述符的维度是发生概率的视觉类(如对象类、属性等)。BoS 是一个整体的网膜代表这些概率的地图。一个嵌入 E 用于最后 BoS 特性映射到欧几里得空间 D。

虽然整体语义表示已经成功图像检索等应用程序或散列，本地化表示，如 BoS，已被证明更少有效的场景分类，有几个原因。首先，场景语义很难定位。他们各有不同从图像补丁图像补丁和困难建立可靠的场景分类器。因此，本地语义往往是嘈杂的[28, 20]，大多数使用一个工作全球每图像语义描述符[34, 2, 3]。这可能有利于散列，但不够表达场景分类。第二，当语义提取在当地，嵌入 E 欧几里得空间很难实现[17]。这是因为语义描述符是概率向量，因此居住一个非常非欧几里得的空间，概率单工，常用的描述符数据失去有效性。在我们的结果显示，即使是复杂的费舍尔向量编码[25]，当直接实现，难以表现在这个空间。

我们认为,最近的可用性的分类器等的 CNN[16],在大型数据集,比如 ImageNET[7],有效地解决了嘈杂的语义问题。这是因为一个 ImageNET CNN,事实上,训练分类对象可能发生在局部地区或补丁一个场景图像。实现一个不变的嵌入问题 E 在语义空间,然而,还有待解决。

3. BoF 嵌入

我们第一次尝试分析,场景分类的适用性,已知的转炉嵌入,即费舍尔向量和的完全连接层 ImageNET

3.1 CNN 嵌入

CNN 的[16],映射 F 由 5 卷积的层。这些产生一个图像转炉我={ x1,x2,。xN },习近平被称为 conv5 描述符。描述符是 max 汇集的地方社区和嵌入转化的大肠嵌入实现使用两个完全连接网络阶段,每个执行线性投影和非线性 ReLu 转换{ W×()} +。由此产生的输出 7 层,我们表示 fc7,的特点空间维,如图 1 所示。

3.2 FV 嵌入

另外,一艘嵌入可以实现的转炉 conv5 描述符。这由一个初步投影到一个主成分分析(PCA)子空间

$$x = Cz + \mu, \tag{1}$$

C 是一个低维主成分分析基础和 z 系数 conv 5 描述符 x 的投影。z 的认为高斯混合分布

$$z \sim \sum_k w_k N(\mu_k, \sigma_k). \tag{2}$$

阵线的一个核心组成部分是自然梯度对参数(均值、方差和权重)模型[30]。conv5 特性,我们发现梯度的意思是[25]

$$\mathcal{G}^{\mathcal{I}}_{\mu_k} = \frac{1}{N\sqrt{w_k}} \sum_{i=1}^{N} p(k|z_i) \left(\frac{z_i - \mu_k}{\sigma_k} \right) \tag{3}$$

它能够满足良好的性能。请注意,这是一个梯度编码和池操作子。它破坏了网膜代表转炉和担保的拓扑不变性现场布局的变化。

3.3 对比

我们比较了 CNN 和阵线嵌入,在两个受欢迎对象识别(加州理工学院 256[11])和场景分类(麻省理工学院室内[26])数据,结果显示在表 1 的上半部分。CNN 嵌入,7 日连接层的特性获得了“咖啡”[14]。[8]后,这 4096 维的特征向量从每个图像提取全球。这是随后功率归一化(开方)和 L2 规范化更好的性能[32]。的分类器训练表示表中“fc 7”来标示。阵线嵌入,256 -维 conv5 描述符 PCA 减少 200 维和池(3),使用 100 年高斯混合。这是紧随其后的是一个平方根和 L2 正常化,加上第二个主成分分析降低维数 4096 年,表示表中的“conv5 +阵线”。这两个表示使用了一个线性支持向量机分类器。

这个实验的结果突出的长处和这两个嵌入的局限性。而 fc7 是优势对象识别的阵线(获得的 12%的加州理工学院),这显然是对场景分类(麻省理工学院室内损失 2%)。这表明,虽然不变的足以代表图像包含单一对象,CNN 嵌入无法应对变化场景的图像。另一方面,混合物基于编码机制阵线是相当有效的现场数据集。

阵线/ conv 5,然而,是一个低级的嵌入 CNN 的特性。原则上,一个等价嵌入 BoS 功能应该有更好的性能,因为语义描述符比 conv5 有更高层次的抽象,从而表现出更大的不变性视觉外观的变化。

Method	MIT Indoor	Caltech 256
fc 7	59.5	68.26
conv5 + FV	61.43	56.37
fc7 + FV	65.1	60.97

在某种程度上,提出图像表示龚。[10]显示这样的不变性的好处,尽管使用嵌入的中间 7 日层激活,而不是语义描述符在网络输出。他们代表一个场景一个集合的形象 fc7 激活从当地农作物中提取或补丁。这些总结了使用一种近似的(3),弗拉德[13]。由此产生的嵌入,表示表 1 中“fc7 +阵线”,是现场 classification1 非常有效。然而,由于表示不会推出从语义特征,这可能是更少的判别和更少的抽象的真正语义嵌入图 2。实现有效的语义嵌入,另一方面,并不是一蹴而就的。我们认为这问题的剩余工作。

4. 语义 FV 嵌入

我们首先简要回顾 BoS 图像表示然后提出合适的嵌入。

4.1BoS

给定一个词汇 V = { v1。,vS } S 西曼-国际资本流动的概念,一个图像可以被描述为一袋从这些概念实例,在图像局部补丁/地区。定义一个 S-dimensional 二进制指标向量 si,这样= 1 爵士和 sik = 0,k 6 = r,当第 i 个图像补丁 xi 描述了语义类r,图像可以表示成 I = { s1,s2,……sn },n 是补丁的总数。假设如果采样从多项分布参数 π i,对数似我的图像可以表示为,

$$\mathcal{L} = \log \prod_{i=1}^{n} \prod_{r=1}^{S} \pi_{ir}^{s_{ir}} = \sum_{i=1}^{N} \sum_{r=1}^{S} s_{ir} \log \pi_{ir}. \quad (4)$$

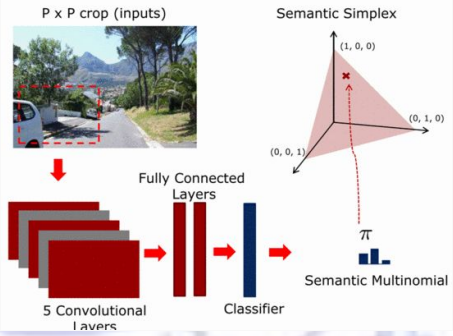

图 3。CNN 基于语义的图像表示。每个图像块在语义空间映射到一个 SMN 公式公式,通过结合卷积转炉映射公式和二次映射公式完全连接网络阶段。由此产生的 bos 网膜代表表示,即一 SMN 图像补丁

由于精确的语义标签 si 的图像区域通常不知道,是很常见的而不是依靠预期对数似

$$E[\mathcal{L}] = \sum_{i=1}^{n} \sum_{r=1}^{S} E[s_{ir}] \log \pi_{ir} \quad (5)$$

使用这一事实 π ir = E(先生)或 P(r | xi),它遵循预期的图像对数似是完全确定的 π i 多项参数向量。这是表示语义多项式(SMN)[27]。他们通常是计算 1)应用分类器,训练有素的语义 V 的图像补丁,并使用结果后 2)类概率 SMNs π i[21]。这个过程是 CNN 分类器如图 3 所示。每一个补丁因此映射到简单的概率,表示语义空间 S 如图 2 所示。图片终于代表由 SMN 收集 I = { π 1, π n}。这是 BoS 表示。在我们的实现中,我们使用 V 和 ImageNET ImageNET 类 CNN[16]估计 SMNs π i。

4.2 直接 FV 实现

自从 BoS 是转炉的家庭成员表示,可以将其映射成一个欧氏通过一个阵线嵌入空间维 E,如图 1 所示。然而,因为单纯形本身不是欧几里德,操作(1)和(3)不直接适用。另一方面一方面,可以使用“费舍尔食谱”模型适合 SMN 的描述符。狄利克雷分布是最受欢迎的多项式模型概率向量[24]。费舍尔梯度的狄利克雷的混合物(DMM),因此,更自然的嵌入图像比 GMM-FV SMNs(3),对数似图像的 BoS I = { π 1,。π n } DMM 下

$$\mathcal{L} = \log P(\{\pi_i\}_{i=1}^{n} | \{\alpha_k, w_k\}_{k=1}^{K}) \quad (6)$$

$$= \log \prod_{i=1}^{n} \sum_{k=1}^{K} w_k \frac{\gamma(\sum_l \alpha_{kl})}{\prod_l \gamma(\alpha_{kl})} e^{\sum_l (\alpha_{kl}-1) \log \pi_{il}}. \quad (7)$$

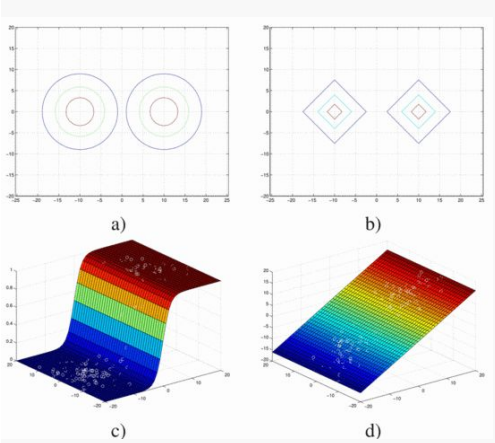

图 4。:两个分类器的欧几里得空间特征公式,与度量公式或 b)公式规范。底部:c)样本的投影)到语义空间公式(只有公式如图所示)。后表面破坏的欧氏结构公式和高斯和拉普拉斯算子非常相似样品(简洁 lapalacian 省略)。D)c)的自然参数空间的映射。

α_k,wk 是 R 分布参数,和 $\gamma(x) = \infty$

$o\ x_{t-1}e^{-x}dx$.。费舍尔得分对数似

是

$$\mathcal{G}^I_{\alpha_k} = \frac{1}{n}\frac{\partial \mathcal{L}}{\partial \alpha_k} = \frac{1}{n}\sum_{i=1}^{N} p(k|\pi_i)\left(\psi(\sum_l \alpha_{kl}) - \psi(\alpha_k) + \log \pi_i\right) \quad (8)$$

在 $\psi(x) = \partial\gamma(x)/\partial x$。使用一些常见的假设这艘文献[25],,我们近似费舍尔的信息块对角矩阵 F

$$\mathcal{F}_{lm} = E\left[-\frac{\partial^2 \log P(\pi|\{\alpha_k, w_k\}_{k=1}^K)}{\partial \alpha_{kl}\partial \alpha_{km}}\right] \approx w_k\left(\psi'(\alpha_{kl})\delta(l,m) - \psi'(\sum_l \alpha_{kl})\right) \quad (9)$$

δ(l,m) = 1 if l = m.。F 的完整推导给出了第一节的补充。数字费舍尔矢量图像我终于获得(8)和(9) $\mathcal{F}^{-1/2}\mathcal{G}^I_{\alpha_k}$.

4.3 限制

SMNs DMMis 自然模型的同时,我们的实验表明,DMM 阵线不会导致一个有效场景分类器(见 7.2 节)。这可能是由于一个非常非欧几里得的自然空间的概率向量。一般来说,数据建模的难度取决于其拓扑空间 X。大多数机器学习假设向量空间与欧氏结构,例如在哪里习的自然测量距离例子 xi ∈ X 是一个指标。这不是单纯形概率的情况下,有摘要 Kullback-Leibler 散度的学习很自然的距离测量,使模型难以建立。

为了说明这个问题,我们提出了两种二元分类问题如图 4)和 b)。情况下,两个类是高斯,其他的他们拉普拉斯算子。class-conditional 分布问题是

$$P(x|y) \propto \exp\{-d(x, \mu_y)\}$$

$Y \in \{0, 1\}$ 是标签和的类

$$d(x, \mu) = ||x - \mu||_p \quad (10)$$

当 p = 1 为高斯拉普拉斯算子和 p = 2 的数据。图 4 a)和 b)的 iso-contours 概率分布在两个场景。请注意,这两个分类器有不同的指标。

类 Y = 1 的后验分布,在这两种情况下,

$$\pi(x) = P(y = 1|x) = \sigma(d(x, \mu_0) - d(x, \mu_1)) \quad (11)$$

$\sigma(v) = (1 + e^{-v})^{-1}$ 形的函数。由于乙状结肠的非线性映射,投影 $x \to (\pi(x), 1 - \pi(x))$ 的样品 xi 语义破坏了原来的欧氏结构空间空间 x 这是如图 4 所示),我们展示的地方后表面和预测 π (xi)样本习的高斯类语义图 4 a)。空间,两点之间的最短路径是不一定一条线。乙状结肠也使的非线性非常后表面的分类问题相似的。拉普拉斯算子的后表面的问题图 4 b)视觉与图 4 c)为了简便起见,我们省略了。

这个示例显示了两个非常不同的分类器转换数据转换成高度非欧几里得的语义空间，几乎难以分辨。这表明建模直接在空间的概率是很困难的将军。这是最可能的 DMM-FV 弱点原因。

5. 语义 FV 的间接实现

在本节中,我们得到一个间接实现的费舍尔语义向量

5.1 自然参数空间

场景分类的非欧几里得的性质后表面嵌入 E 图 2 的很很难学习。例如,请注意,(1)或的 PCA 高斯编码的阵线(3)毫无意义语义空间数据,因为后的测地线表面没有行。这个问题是可以避免的注意 SMNs 多项式的参数,这是指数分布的家庭的一员

$$P_S(s;\pi) = h(s)g(\pi)\exp\left(\eta^T(\pi)T(s)\right), \qquad (12)$$

T(s)是表示一个足够的统计。在这个家庭 v = η（π),使(日志)线性充分统计量的概率分布

$$P_S(s;\nu) = h(s)g(\eta^{-1}(\nu))\exp\left(\nu^T T(s)\right). \qquad (13)$$

这就是所谓的自然参数分布。在这种参数化,多项收益率的图像的 BoS 自然参数向量 v i = η(E { si })为每一个补丁,而不是一个概率向量。语义是二进制,自然参数是通过分对数变换 v 日志 π=1−π SMNs。这个地图的高度非线性的语义空间图 4 c)的线性空间图 4 d)。同样,其自然的多项分布参数的映射空间,可以获得一个一对一的转换语义空间与欧氏空间的结构。这使得嵌入图 2 的 E 变得更为容易。事实上,它现在可以实现的主成分分析(1)(3)编码操作。

5.2 间接 FV 实现

上面的讨论中显示的实现语义阵线替代 4.2 节。这包括映射 BoS I = {π1,。 πn }到自然参数空间 BoS(NP-BoS)我= {v1。。。 vn }和计算 v i 阵线的自然参数。像以前一样,这在三个步骤:

1.使用(1)映射的 PCA v i 到他们的参数在一个低维子空间投影 ξ i

2.学习(2)的高斯混合最适合低维预测 ξ i

3.所示，计算 ξ i 阵线(3)的预测。

相比直接阵线的实现部分 4.2,这个实现有利用的优势 GMM 阵线机械已经在文献中可用。多项分布的参数向量 π=(π1,。 πS)实际上有三种可能的自然参数化

$$\nu_k^{(1)} = \log \pi_k \qquad (14)$$

$$\nu_k^{(2)} = \log \pi_k + C \qquad (15)$$

$$\nu_k^{(3)} = \log \frac{\pi_k}{\pi_S} \qquad (16)$$

v k 和 π k v 的 k 条目和 π,分别。这些参数化的性能是可能的依赖于语义分类器的实现生成 SMNs。判别分类器等 CNN, v (2)可能是最好的参数化。注意,在本例中,向量的条目 π k = 1C e v,是当且仅当 C =一个概率向量 P 我 e v i。因此,从 v 映射到 π 一般是将 softmax 变换在 CNN 实现输出。这意味着 CNN 是学习如何辨别自然中的数据多项分布的参数空间是一个泛化的自然二项空间如图 4 d)。我们在 7.2 节通过比较测试这个断言(14)-(16)的参数化的场景分类。

6. 相关工作

提出了语义阵线已经与数的关系在最近的作品。

6.1 方根嵌入

概率语义阵线是一个不变的嵌入向量编码向量,根据费雪。的 DMM-FV 和 NP-BoS 阵线是不同的实现的想法。

他们提供了另一种流行的做法编码开方概率向量[6]38 岁,17 日,即应用

$$\nu_k^{(4)} = \sqrt{\pi_k}. \tag{17}$$

利用微分几何的平方根是合理的参数(38 岁,17)和原始的嵌入,诱发 Bhattacharya 转换之间的相似性点[1]。池的平方根除(root-SIFT),而不是除(L1-SIFT)也有利于[6]。比较这个嵌入(14)-(16)和 DMM-FV 节中给出 7.2。

6.2 layer 7 激活的阵线

拟议的代表,当计算映射 v (2)(15),正如上面所讨论的,直接在行动八层的输出(fc8)ImageNET CNN[16]。在这个意义上讲,它类似于[10]的费舍尔向量,计算使用完全连接的激活 7 层(fc7)。最重要的区别然而,两个是 fc8 输出语义特性获得的结果对 fc7 判别投影。因此,他们可能会更有选择性。除了他们明确的语义性质也确保了更高级别的抽象,由于他们可以更好地推广比低 CNN 层特性。我们比较两个表示来验证这些断言在 7.3 节

6.3 微调

超出其成功 ImageNET 分类,CNN 的[16]已被证明是高度适应对方分类任务。一个受欢迎的适应策略,清楚“微调”[9],包括执行额外的迭代反向传播的新的数据集。然而,这是一个启发式和耗时的过程,需要为了防止被严密监控网络过度学习。提出了语义费舍尔向量也被视为一种适应机制,充分利用最初的 CNN,提取功能,扩充它 2979 年费舍尔向量层,使其应用到其他任务。这个过程是没有启发式和消耗更少时间比“微调”。我们比较的性能两个在 7.4 节。

6.4 CNN 的地方

最近的努力提高场景分类依靠 pre-trained imageNET CNN(8 32 10、22)。主要是因为优质的功能反应[37]。我们的工作小说使用对象语义由这个网络来获取高水平表示对场景图像。周等人提出一个更直接方法,不依赖于 ImageNET CNN。他们只是学习一种新的 CNN 的大型数据库场景图像称为“地方”数据集[39]。虽然地方 CNN 的基本架构是一样的 ImageNET CNN,学习非常的类型的特性不同。而卷积 ImageNET CNN 的单位响应对象出现,那些在 CNN 的地方选择性的风景与更多的空间特性的嵌入。因此,CNN 的地方产生一个整体补充我们的语义表示的场景阵线。我们将演示相结合的效果表示在我们的分类实验。

7. 评价

在本节中,我们报告的实验设计对语义阵线的性能进行评估。

7.1 实验装置

所有实验在麻省理工学院的 67 类室内场景[26]和 397 年类太阳场景[36]数据集。CNN 特征提取与咖啡库[14]。阵线,CNN 有关特性(fc7 或 fc8)提取 fromlocal P×图像补丁制服网格。为简单起见,进行初步实验 P = 128。最后一轮实验使用多尺度特性,与 P∈{ 96、80、96、80 }。为 GMM-FVs 当地所有功能第一次减少到 500 人维度,使用主成分分析,然后使用(3)和集中 100 组件的混合物。4.2 节的 DMM-FV 学会了 50 组件混合物在 1000 维 SMN 空间。在文献中是很常见的,所有费舍尔向量 L2 电力标准化和规范化。这导致 DMM 和 GMM 大小为 50000 艘渔船,维度利用主成分分析法(PCA)的进一步减少到 4096 年。在一些实验中,我们也评估分类器的基础上全球 fc7 和 SMN 特征提取,如[8]。的全球 fc7 特性是平方根和 L2 规范化,而全球 SMNs 只是开方。场景分类器训练实施对所有图像表示与一个线性支持向量机。

7.2 直接与间接语义 FV

我们开始的比较直接和间接实现费舍尔的语义向量。前基于 4.2 节的 DMM-FV, 后者在吗参数的映射(14)-(17),这是表示 v (i)-阵线。介绍了附加基准,表示 π-阵线,这是一个基于 GMM 的费舍尔向量(1)-(3)应用直接在 SMNs。我们进行这个实验单一麻省理工学院的培训/测试分室内和 SUN 的数据集表 3 和报告的精度。所有的间接实现的语义阵线表现更好比其余的方法,获得 10 分。DMM-FV 反映了先前的表现不佳指出在单纯形建模的困难。的平方根投影(17)不的大圆表现更好。线性映射 π-FV 整体最差的性能。在间接的实现语义阵线,(15)达到最好的结果,虽然差异是微妙的。鉴于这些结果,我们采用间接的实现语义阵线,再参量化 v (2)(15)的剩余实验这只是表示“语义阵线。

7.3 不变性的作用

来测试假设语义阵线是更多判别和不变的阵线提取低网络层,与我们相比其性能[10]的 fc7 艘渔船。在这个实验中,CNN 的特性在多尺度提取(全球以及来自哪里补丁的大小 80、96、128 和 160 像素)。表 2 显示了结果语义阵线(fc8 表示)和 fc7 阵线。值得一些言论,鉴于之前类似的实验报告(9、8、10)。首先,当相比全球 CNN 提取特征的方法[8],局部表现有更好的性能。第二,尽管全球 fc7 特征提取[10]表现不佳,使用全球第八层特性导致更糟糕的性能。这可能表明结论层 8 提取”更糟场景分类的特征”。剩下的列,然而,清楚地说明了问题。当在本地提取,语义描述符是高度有效,实现增加到 3 点对 fc7 特性。局部和全球之间的性能差距语义描述符用的局部性质来解释场景语义,从补丁,补丁。一个全球语义描述符只是没有表现得能够捕获这种多样性。第三,最近的理由使用中间 CNN 特性应该修正。相反,表的结果支持这样的结论:这些特性都不如语义判别和不变的吗描述符。当加上一个合适的编码,这样随着语义阵线,后者实现最好的场景分类结果。

最后,为了确保语义阵线的收益不仅仅是由于使用(15)的变换,我们应用转换 fc7 特性。而比增加,这导致了一个巨大的性能下降(58%相比 65.1%的 fc7-FV 麻省理工学院室内在补丁大小 128)。这是预期的,因为自然在这种情况下参数空间参数并不适用。

7.4 比较艺术的状态

连接费舍尔向量 fc7 特性计算在多个补丁可以产生巨大的尺度[10]。我们实现了这一策略的 fc7-FV 和语义阵线,结果见表 2。结合 fc7-FVs 三片鳞片在麻省理工学院的室内和分类精度为 68.8%51.8%的太阳。虽然这是一个重要的改进在任何 single-scale 分类器,连接 semantic-FVs 3 尺度(精度产生更好的结果 71.24%的麻省理工学院室内和 53.24%的太阳)。当使用 4 片鳞片,被观察到了类似收益表中的报道。

比较我们的语义阵线与其他多尺度主要来源于 ImageNET 表示 cnn 表 4 所示。正如预期的那样,脱咖啡因的先驱[8]表示是所有其他的方法,因为它大大不如描述了在全球范围内提取复杂场景图像描述符使用对象 CNN[16]。在技术依赖于局部特征提取的提议刘 et al。[22]和 Razavian et al。[32]。现场代表在[22]是一个稀疏的激活代码来自第六层(fc6)CNN。Razavian 等。[32], 使用功能从倒数第二层 OverFeat 网络[31]提取从更大空间尺度上。因为所使用的功能[22]和[32]缺乏语义的不变性,他们分类器很容易超过我们的语义阵线分类器。我们也与技术称为微调[9],适应 imageNET 直接 cnn 场景分类的任务。这个过程需要一个几个成千上万的反向传播的

迭代感兴趣的场景数据集和持续大约5 - 10 小时一次 GPU。然而，最终的分类器是显著的比我们的语义阵线分类器。

Method	MIT Indoor	MIT SUN
fc8-FV (Our)	**72.86**	**54.4 ± 0.3**
fc7-FV [10]*	69.7	53.0 ± 0.4
fc7-VLAD [10]	68.88	51.98
ImgNET finetune	63.9	48.04 ± 0.19
OverFeat + SVM [32]	69	-
fc6 + SC [22]	68.2	-
DeCaF [8]*	59.5	43.76

表 4。使用 ImageNET 与最先进的方法训练有素的特性。*表示我们的实现

Method	MIT Indoor	MIT SUN
ImgNET fc8-FV (Our)	72.86	54.4 ± 0.3
Places fc7 [39]	68.24	54.34 ± 0.14
Combined	**79.0**	**61.72 ± 0.13**

表 5 所示。与 CNN 训练场景[39]

另一个使用 pre-trained 对象分类 CNN(16 日 31 日)为场景直接学习 CNN 大规模数据集。这是最近执行的周等人使用 200 万图像数据集[39]。表 5 显示比较一个场景表示 CNN 和 ImageNET 获得的地方基于语义阵线。语义阵线的结果更好，他们的室内场景数据集，然而，太阳，描述符执行相对。更重要的是，一个简单的串联两个产生收益近 7%的准确性在两个数据集，表明嵌入的，事实上，免费。这些结果我们所知的，最先进的场景分类。

8. 结论

在本文中，我们讨论了建模场景的好处图片的袋子从 ImageNET 对象语义 CNN 代替其下层激活。利用语义描述符的优质，我们提出一个有效的方法总结他们费舍尔向量，这是不平凡的。语义阵线提供了更好的比一个阵线低级特性分类架构甚至一个调整分类器。当结合 CNN 特性从一个场景分类，我们的语义阵线产生的结果。

9. 感谢

这部分工作由 NSF 资助奖 IIS -1208522 和 ccf - 1208522 和一份礼物从雅虎教师研究卓越计划(FREP)。

References

[1] R. Arandjelovi´c and A. Zisserman. Three things everyone

should know to improve object retrieval. In *IEEE Conference*

on Computer Vision and Pattern

Recognition, 2012. 6

[2] A. Bergamo and L. Torresani. Meta-class features for largescale

object categorization on a budget. In *Computer Vision*

and Pattern Recognition (CVPR), 2012. 3

[3] A. Bergamo and L. Torresani. Classemes and other classifierbased

features for efficient object categorization. *IEEE Transactions on Pattern Analysis and*

Machine Intelligence,

page 1, 2014. 1, 3

[4] A. Bergamo, L. Torresani, and A. Fitzgibbon. Picodes:

Learning a compact code for novel-category recognition.

In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and

K.Weinberger, editors, *Advances in Neural*

Information Pro-

cessing Systems 24, pages 2088–2096. 2011. 1

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for

human detection. In *Proceedings of the 2005 IEEE Com-*

puter Society Conference on Computer Vision and Pattern

Recognition (CVPR'05) - Volume 1 -

Volume 01, CVPR '05,

pages 886–893, Washington, DC, USA, 2005. IEEE Computer
Society. 1

[6] J. Delhumeau, P. H. Gosselin, H. J´egou, and P. P´erez. Revisiting

the vlad image representation. In *ACM Multimedia*,

pages 653–656, 2013. 6

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-
Fei. Imagenet: A large-scale hierarchical image database.

In *Computer Vision and Pattern Recognition, 2009. CVPR*

2009. IEEE Conference on, pages 248–255,

June 2009. 3

[8] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang,
E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation
feature for generic visual recognition, 2014. 1, 2, 3, 7, 8

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature
hierarchies for accurate object detection and semantic

segmentation. In *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR), 2014. 1,

2, 6, 7, 8

[10] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale
orderless pooling of deep convolutional activation features.
In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors,

Computer Vision ECCV 2014, volume 8695 of

Lecture Notes

in Computer Science, pages 392–407. Springer

International
Publishing, 2014. 1, 2, 4, 6, 7, 8

[11] G. Griffin, A. Holub, and P. Perona. The caltech-256. Technical
report, caltech technical report, 2006. 3

[12] T. S. Jaakkola and D. Haussler. Exploiting generative models

in discriminative classifiers. In *Proceedings of the 1998 con-*

ference on Advances in neural information processing sys-

tems II, pages 487–493, Cambridge, MA, USA,

1999. MIT
Press. 1

[13] H. J´egou, M. Douze, C. Schmid, and P. P´erez. Aggregating

local descriptors into a compact image representation. In

IEEE Conference on Computer Vision & Pattern Recognition, pages 3304–3311, jun 2010. 4

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick,
S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 3, 7

[15] 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. 1

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet
classification with deep convolutional neural networks. In
F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 1, 2, 3, 4,
6, 8

[17] R. Kwitt, N. Vasconcelos, and N. Rasiwasia. Scene recognition
on the semantic manifold. In *Proceedings of the 12th European conference on Computer Vision - Volume Part IV*, ECCV'12, pages 359–372, Berlin, Heidelberg, 2012.
Springer-Verlag. 1, 2, 3, 6

[18] C. H. Lampert, H. Nickisch, and S. Harmeling. Attributebased
classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014. 1

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

[20] L.-J. Li, H. Su, Y. Lim, and F.-F. Li. Object bank: An
object-level image representation for high-level visual recognition. *International Journal of Computer Vision*, 107(1):20–
39, 2014. 1, 3

[21] W. Li and N. Vasconcelos. Recognizing activities by attribute
dynamics. In *Advances in Neural Information Processing Systems*, pages 1115–1123, 2012. 4

[22] L. Liu, C. Shen, L. Wang, A. Hengel, and C. Wang. Encoding
high dimensional local features by sparse coding based
fisher vectors. In Z. Ghahramani, M. Welling, C. Cortes,
N. Lawrence, and K. Weinberger, editors, *Advances*

in Neu-

ral Information Processing Systems 27,

pages 1143–1151.
Curran Associates, Inc., 2014. 2, 7, 8
[23] D. G. Lowe. Distinctive image features from scale-invariant
keypoints, 2003. 1
[24] T. P. Minka. Estimating a dirichlet distribution. Technical
report, 2000. 4
[25] F. Perronnin, J. S´anchez, and T. Mensink. Improving the
fisher kernel for large-scale image classification. In

Pro-

ceedings of the 11th European conference

on Computer vi-

sion: Part IV, ECCV'10, pages 143–156, Berlin,

Heidelberg,
2010. Springer-Verlag. 1, 2, 3, 5
[26] A. Quattoni and A. Torralba. Recognizing indoor scenes.

2012 IEEE Conference on Computer Vision

and Pattern

Recognition, 0:413–420, 2009. 3, 7

[27] N. Rasiwasia, P. Moreno, and N. Vasconcelos. Bridging the

gap: Query by semantic example. *Multimedia,*

IEEE Trans-

actions on, 9(5):923–938, 2007. 1, 4

[28] N. Rasiwasia and N. Vasconcelos. Scene classification with
low-dimensional semantic spaces and weak

supervision. In

IEEE CVPR, 2008. 3

[29] N. Rasiwasia and N. Vasconcelos. Holistic context models

for visual recognition. *Pattern Analysis and*

Machine Intel-

2982

ligence, IEEE Transactions on, 34(5):902–917,

May 2012.
1
[30] J. S´anchez, 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. 1, 3
[31] 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. 2, 8
[32] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson.
Cnn features off-the-shelf: An astounding baseline for

recognition. In *The IEEE Conference on*

Computer Vision

and Pattern Recognition (CVPR) Workshops,

June 2014. 2,
3, 7, 8
[33] Y. Su and F. Jurie. Improving image classification using semantic

attributes. *International Journal of Computer*

Vision,

100(1):59–77, 2012. 1

[34] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object

category recognition using classemes. In *European Con-*

ference on Computer Vision (ECCV), pages

776–789, Sept.
2010. 1, 3

[35] J. Vogel and B. Schiele. Semantic modeling of natural scenes

for content-based image retrieval. *Int. J. Comput.*

Vision,

72(2):133–157, Apr. 2007. 1

[36] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun
database: Large-scale scene recognition from abbey to zoo.

In *Computer Vision and Pattern Recognition*

(CVPR), 2010

IEEE Conference on, pages 3485–3492, 2010. 7

[37] M. D. Zeiler and R. Fergus. Visualizing and understanding

convolutional networks. In *Computer Vision –*

ECCV 2014 –

13th European Conference, Zurich,

Switzerland, September

6-12, 2014, Proceedings, Part I, pages

818–833, 2014. 2, 7

[38] D. Zhang, X. Chen, and W. S. Lee. Text classification with
kernels on the multinomial manifold. In

Proceedings of the

28th Annual International ACM SIGIR

Conference on Re-

search and Development in Information

Retrieval, SIGIR

'05, pages 266–273, New York, NY, USA, 2005. ACM. 6

[39] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva.
Learning Deep Features for Scene Recognition using Places

Database. *NIPS*, 2014. 1, 2, 7, 8