指导教师：杨涛　　提交时间：20160315

CVPR2015 Paper Translation

No: 01
姓名：张祖德
学号：2013302489
班号：10011301

通过测量同变性和等价性实现图像的表示

Karel Lenc Andrea Vedaldi

Department of Engineering Science, University of Oxford

摘要

用梯度方向直方图和深度卷及神经网络（CNN）对于图像的表示尽管重要，但我们对于它们的理论认识仍然有限。为了填补这一空白，我们探讨三个关键的数学特性的表示：同变性，不变性，和等价性。同变性研究了编码表示的输入图像是如何转换的，不变性是一个转换对图像没有影响的特殊情况。等效性研究是否两种表示方法，例如对于一个神经网络的两种不同的参数化，会产生相同的图像信息。有许多方法可以经验性地确定这些被提出的属性，包括在 CNN 的缝合层引入变换。然后，将这些方法应用于流行的表述，用以揭示了在它们的结构方面的洞察力，包括明确了 CNN 层面的某些几何不变性的实现。本文的重点是理论性的，结构化输出的直接应用也可以相应得到证明。

1. 介绍

图像表现二十年来一直是计算机视觉的一个研究重点。显着例子包括纹理基元[11]，方向梯度直方图（SIFT[14]和 HOG[4]），视觉词袋[3][24]，稀疏[32]和本地编码[31]，超矢量编码[35]，VLAD[9]，费希尔载体[17]，以及最新的深度卷积神经网络[10，21，33]。然而，尽管其受欢迎程度很高，我们对其理论认识仍然有限。人们普遍认为一个良好的图像表示应结合不变性和辨别性，但这种描述是比较模糊的；例如，在图像表示中包含了什么不变性以及如何求得不变性往往是不清楚的。在这项工作中，我们提出了一种新的方法来研究图像表示。我们来看：表示 Φ 为一个图像的抽象函数映射，x 为矢量 Φ(x) ∈ Rd，我们凭经验建立函数的关键的数学特性。我们特别关注三个这样的属性（第 2 节）。第一个是**同变性**，

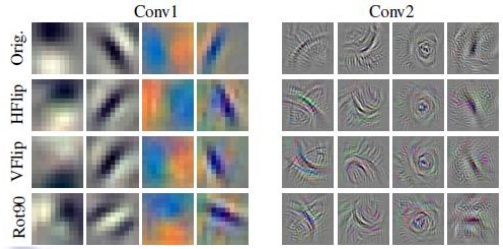

图 1：**都为 CNN 滤波器转换**。 最佳：用[23]的方法对卷积神经网络的转化率 1 和 CONV2 过滤器进行可视化。其他：几何扭曲滤波器从等变化转换网络方面重建并采用第 2 的方法输出结果，包括了水平翻转，垂直翻转和旋转 90°。

它看起来表示了输入的图像转换后是如何改变的。我们证明了，大多

数图像表示，包括 HOG 和大多数深度神经网络，输入后都通过一个容易预测的方式发生改变（图 1）。我们发现，这种等变化转换可以从数据经验中得到学习（第 2.1 节），而且，重要的是，他们等同于简单的线性变换来表示输出（第 3.1 和 3.2 节）。在这种情况下我们通过引入并学习一个新的转换层来获得卷积网络。通过分析我们学会了等变化变换也能找到和图像表示的同变性特征。我们的第二个属性**不变性**，使我们能够量化不变性并显示它如何深层次地建立深度模型。第三个属性，**等价性**，体现了被捕获的表示尽管看似异构实际上是相同的。 CNN 模型，特别是包含数百万的冗余参数[5]，由于非凸优化在研究中，重新训练时，相同的数据可能会有所不同。接下来的问题是，产生的不同结论是确实存在的或只是偶然现象。为了回答这个问题，得知我们研究了拼接层，拼接层中允许不同网络的部分发生交换。如果“Franken—神经网络”和原有方法得出的结果相等，那么等价性可以被证实（第 3.3 节）。本文的其余部分安排如下。第 2 节讨论如何经验性地学习同变性、不变性和等价性表述的方法。第 3.1 和 3.2 节分别表述了浅层和深层的实验中的同变性，第 3.3 为等价性的表述。第 3.4 演示了一个实际应用等变性交涉结构输出回归。最后，第 4 节总结我们的研究结果。

相关项目：设计不变性或等变性的问题已被广泛地应用于计算机视觉的探讨之中。例如，一个流行的策略是提取不变性局部的描述内容[13]对等变性的顶部进行（也是所谓的共变体）探测[12，13，15]。很多作者也明确地在文章中纳入同变性[20,26]。深度卷积神经网络，包括 Krizhevsky 等人构建中的一个[10]和相关国家的最先进的架构，其中每一层的增量稳定。甚至有更明确的 Sifre 和的 Mallat[22]的散射变换。在所有这些例子中，不变性是设计目的可能是由既有的体系结构来实现，也可能不是。我们的目标不是要提出另一个机制研究不变性，而是一个方法来系统地梳理出不变性，同变性，以及其他属性的一个可能有的统一表示。据我们所知，在这方面的研究分析的工作现在还很有限。也许只有神经网络的不变性特定的图像变换才能有所用处[8，33]。不过，我们相信能成为第一个在功能上表征和量化的这些性质并能研究到等价性的不同表述。

2. 图像表示的显着特性

图像表示，如 HOG，过筛，或 CNN 被认为是从 Φ 图像的 x ∈ X. 映射到向量 Φ(x) ∈ Rd. 的函数。本节将介绍三个显着的特征：同变性，不变性和等价性 - 并给出了算法的建立。根据经验，同变性。是一种等变化变换的表示， Φ 表示变换 g 当转换的输入

图像的可以被转移到的输出表述。从形式上看，发生变换 g 的同变性当存在 Mg ： Rd → Rd 例如：

$\forall x \in X: \Phi(gx) \approx Mg\Phi(x).$ (1)

其中Mg存在的充分条件是Φ可逆的，因为在这种情况下的$cMg = \Phi \circ g \circ \Phi - 1$. 已知的是如HOG是至少是约可逆[30]。因此，它不仅存在，更对于Mg的映射的结构是有用的。特别是，Mg很简单，例如一个线性函数。这在简单的预测中十分重要，如线性分类或在CNN的的情况下，有利于进一步通过线性滤波器处理。此外，通过相同的Mg映射工作，任何输入图像的固有的几何性质都会被捕获。

变换g的性质在原则上是任意的;在实践中，在本文中，我们将专注于几何变换例如仿射经线和图像翻转。

不变性：不变性是同变性的一种特殊情况

当同变性取得时的Mg（或Mg的一个子集）最简单最有可能转变为不变性，即称为身份映射。不变性往往被视为图像表示的关键属性。自从计算机视觉建立以来，目标之一就是要建立图像的不变性。例如，包含在图像中的对象类别是不变的视点。通过系统地研究不变性，就能够知道是否并如何可以表述这个视点。

等价性：虽然看上去等价性和不变性是在图像的不同转换中的一种表述，事实上等价性研究的是不同表述之间的联系。两种不同的表述Φ 和 Φ′ 是等价的当且仅当存在$E\Phi \rightarrow \Phi'$ 的映射如：

$\forall x : \Phi'(x) \approx E\Phi \rightarrow \Phi' \Phi(x).$

如果Φ是可逆的，那么$E\Phi \rightarrow \Phi' = \Phi' \circ \Phi - 1$ 满足这一条件。因此作为存在于Mg 之前，不只是存在联系，而且存在$E\Phi \rightarrow \Phi'$ 的映射结构。

示例：HOG的同变性转换。令Φ表示HOG[4]特征提取。在这种情况下，Φ（x）的可被解释为一个H×W的矢量场特征向量或元素。如果g表示图像左右翻转垂直轴，则Φ(x)和Φ(gx)由相关的明确定义为的功能组件的排列。 这个排列在水平方向替换了HOG元素。并且，每个HOG元素内，替换的组件对于梯度对称。因此，映射Mg是和一个具有完全相同Φ（GX）=MgΦ（X）的替换。相同点在于水平方向都为近似190◦旋转和旋转180◦。HOG的实现[28]的排列事实上已经有了明确规定。

示例：在卷积表述中同变性转换。HOG，密集计算的SIFT（DSIFT）和卷积网络是卷积表述的在不变性转换操作中的几个示例。除了边界和采样的影响，任何卷积表述都等同于输入图像的转换，因为这会导致此特征域的转换。

2.1. 研究稀疏结构的特征

在研究同变性和等价性时，Mg的转

[1] 大多数 HOG 实现使用 9 个方向的取向，不仅仅是旋转对称

换和E ϕ → ϕ′ 通常在闭环境下是不可用的，并且必须从数据中去估计。本节讨论的一些算法就是这样做的。讨论重点在于等变化变换Mg，处理等价变换E ϕ → ϕ’是相似的。

给定的表述 ϕ 和变换g时，目的是要找到一个映射Mg满足（1）。在最简单的情况下Mg = (Ag, bg), Ag ∈ Rd ×d, bg ∈ Rd是映射变换 ϕ (gx) ≈ Ag ϕ (x) + bg. 。这一选择并不受限制，因为在上面的例子中Mg是最初可能是置换，因此可通过一个操作进行Ag相应的矩阵置换。

估计（AG，BG）自然地是一个风险最小化的问题。鉴于数据x是从一组自然图像中采样得到，因此研究优化正规化错误的数值：

$$E(A_g, \mathbf{b}_g) = \lambda \mathcal{R}(A_g) + \frac{1}{n}\sum_{i=1}^{n} \ell(\phi(g\mathbf{x}_i), A_g\phi(\mathbf{x}_i) + \mathbf{b}_g), \tag{2}$$

其中R是正则，ℓ是回归损失其他选择在下面讨论。目标（2）可以是通过更改 ϕ（GX）以适用于等价问题 ϕ’（x）。

正则化：正则的选择作为有(d2)参数的Ag ∈ Rd×d特别

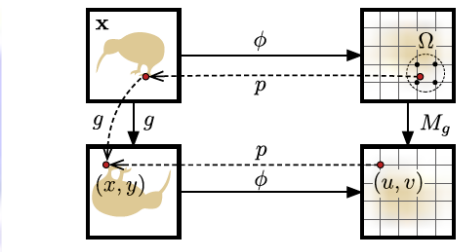

图2：稀疏结构。预测等变性特征在位置（U,V）有一个对应的邻域特征g，m（U,V）

重要。D原来可以相当大（例如，在HOG有d =
DWH），正则是必不可少的。标准的12k正则Agk2F的仍然不够;相反，稀疏诱导在这方面的工作有诸多问题，他们很鼓励Ag作类似的矩阵置换操作。我们认为有两种这样的稀疏性诱导正则。第一种正则允许Ag在每一行包含固定数量的非零项K：

$$\mathcal{R}_k(A) = \begin{cases} +\infty, & \exists i : \|A_{i,:}\|_0 > k, \\ \|A\|_F^2, & \text{otherwise.} \end{cases} \tag{3}$$

行的正则反映了一个事实，每一行都是 ϕ （GX）的特征参数的预测值。

第二种稀疏诱导正则和第一种很相似，但多次利用了卷积结构的交叉。卷积特征是来自不变性转换和个人操作运算（非线性滤波器），其中 ϕ （x）的可被解释为一个特征字段并包含空间索引（U，V）和通道索引t。 由于图像表述的代表性，ϕ (gx)的分量（U，V，T）应该相应地预测特征 ϕ (x)邻域g，m（U,V）（图2）。这导致了一个特定情况：稀疏结构可由正则化产生。

$$\mathcal{R}_{g,m}(A) = \begin{cases} +\infty, & \exists t, t', (u, v), (u', v') \notin \\ & \Omega_{g,m}(u, v) : A_{uvt,u'v't'} \neq 0 \\ \|A\|_F^2, & \text{otherwise,} \end{cases} \tag{4}$$

m表示邻接大小和A由三维向量（U，V，T）确定。邻接本身定义为M×M的输入特征网络，输出元素投影为（U，V）[2]。在计算中（3）和（4）将进行有限制

[2] 2在形式上, (x, y)表示输入图像x的像素点坐标p : (u, v) 7→ (x, y) 表示从特征索引(u, v) 到中心(x, y)在输入图像上的函数映射. Nk(u, v)表示了特征 k 的位置(u′ , v′)更接近(u, v) （后者具有分数级的坐标），根据这个定义邻接可有 (u,转换为 g,k(u, v) = Nk(p−1 ◦ g−1 ◦ p(u, v)).

的组合。在第3.2节中，是否去掉ℓ的选择是很重要的，因为HOG和类似的直方图特征如l2, Hellinger’s,或者χ2距离需要调整到很好。然而，对于更复杂的特征，在神经网络的深层，结果发现，有针对性的取舍某些情况下可以大大提高执行结果的质量。为了理解目标导向舍去的概念，需要考虑CNNΦ中训练的终端到终端的分类问题，如ILSVRC 2012图像分类任务（ILSVRC12）[19]。一种常见的方法[1,6，18]是使用第一个多层Φ1的Φ=Φ2◦Φ1作为通用特征提取。在原来的Φ1问题中，这种方法会使用一个替代的目标，其中目标的要求是保持质量是等变的：

$$E(A_g, \mathbf{b}_g) = \lambda \mathcal{R}(A_g) + \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \phi_2 \circ (A_g, \mathbf{b}_g) \circ \phi_1(g^{-1}\mathbf{x}_i)). \quad (5)$$

在这里，yi表示图像xi和ℓ底层真正的特征是都是用去掉法分类来训练Φ。注意，在此情况下（Ag，BG）被研究以补充图像变换的知识， 因此这被设置为G-1。这一观点不限于CNN的，而是适用于任何有给定目标分类或回归任务的Φ1和相应的训练预测Φ2。

2.2. 神经网络中的同变性：转换层

第2.1节中的方法大大细化了卷积表述和一定的变化下的分类情况。结构化稀疏正则（4）鼓励Ag与卷积结构表述进行匹配。如果g映射变换可以表示最高采样效果，那么等变化转换Mg就体现了平移不变性，即卷积性质。其原因是，一个映射g均匀地作用于图像域[3]，那么Mg同样如此。这有两个关键的优势：它显著减少了研究的参数，它可以有效地实现一个CNN的附加层。这样一个变换层由输入特征部位置换层（U，V，T）到输映射出特征网络（G（U，V），T）和堆线性滤波器，每个维度M×M×D。这里的M对应于邻域的大小G，M（U，V）在第2.1节。直观地，这些过滤器的主要目的是对特征频道进行重新排列和插入。在一般情况下，需要注意的是g（U，V）并不落在整数坐标上。在我们的例子中，转换层分配g（U，V）是通过舍入的方式，但是最近的整数坐标位置也可以，通过使用双线性插值近似到最近的2×2个位点[4]。

2.3. 神经网络的等价性：拼接层

上一节是研究同变性在神经网络中的作用，本节内容为等价性。在第2.1节中讨论面向任务的丢失策略以后，可以考虑两种表述Φ1和Φ’1和一个预测器Φ’2了用来解决参考性的任务Φ’1。例如，这些可以通过分解2个神经网络Φ = Φ2 ◦ Φ1和Φ′ = Φ′2 ◦ Φ′1中训练好的的ImageNet ILSVRC数据（但Φ1也可能是通过一个不同的问题学习或是人工操作）得到。目标是找到一个映射EΦ1→Φ′ 例如φ′1 ≈EΦ1→Φ′1。此映射可以被

[3] 意思是g(x + u, y + v) = g(x, y) + (u′, v′)

[4]可以通过使用图像变形技术获得更准确的结果。例如，子像素精度可通过采样转换得到，然后使该变换过滤器平移不变（或者等价地引入合适的非线性映射置换层和改造过滤器）。

看作是一个“拼接转换”让Φ’2◦EΦ1→Φ1◦Φ1在原始分类任务中和Φ’2◦Φ“1一样好。因此，这种转换中，最小化损失ℓ(yi, Φ′ 2 ◦EΦ1→Φ′ 1 ◦Φ1(xi))近似于（5）。神经网络中EΦ1→Φ“1可以理解为一个拼接层。此外，给出的卷积结构的表示该层可被视作线性堆滤波器。在此情况下不转换层，如果空间特征Φ1的尺寸和Φ’1不匹配，它可能有必要向下/上采样。

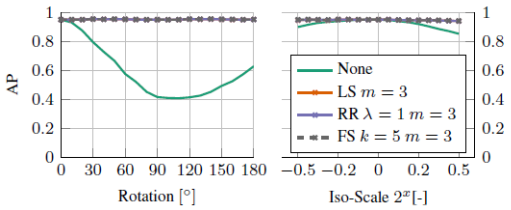

图4：使用HOG特征参数等分类。 分类旋转后的狗和猫的头，训练后的基于HOG分类器的特征测试图像，逐渐缩放得到对比图，使用LS，RR，和FS的等价性对比。

k	m	HOG size 3×3	5×5	7×7	9×9
5	∞	1.67	12.21	82.49	281.18
5	1	0.97	2.06	3.47	5.91
5	3	1.23	3.90	7.81	13.04
5	5	1.83	7.46	17.96	30.93

表1：回归成本。研究等变性的成本（以秒为单位）回归量，图4。由于HOG阵列的尺寸变化更大，最优化成本显著增加除非通过减少m使得结构稀疏。

3. 实验

实验从第3.1节开始研究该浅等变化映射。第3.2和3.3节研究深度卷积招表示，研究同变性和等价性。在第3.4节等变化映射应用于结构输出回归。

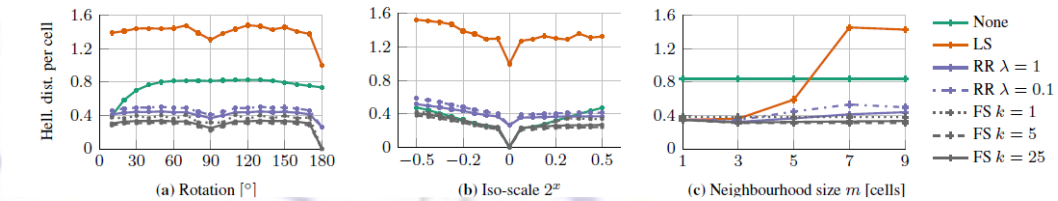

图3：回归方法。图中表示得到的HOG特征重建误差（平均每单元海林格距离）是根据映射Mg在g设置为不同的图像旋转（3a）值的和设置为不同的学习策略（3B）（见文本）。没有其他约束强加给Ag。在右侧面板（3C）的实验同样是45◦旋转，而这个时候结构化稀疏的Ag的邻域大小m不同。

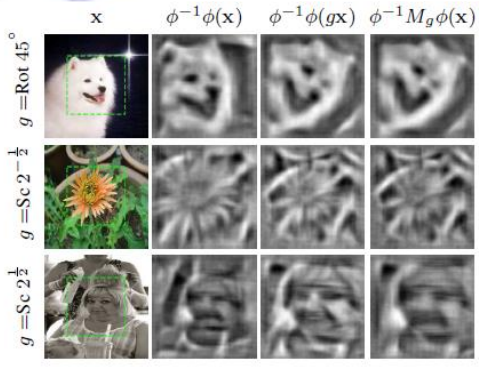

图5：HOG等变性的定性评价。可视化特征φ(x)， φ(gx) 和 Mgφ(x)使用φ-1Mgφ的HOGgle[30] HOG逆。Mg用FS来研究，其中k= 5，m =3，g设定为旋转45◦并缩放来和P2分别。虚线框表特征重建。

3.1. 同变性浅交涉

本部分利用第2.1节的方法研究浅等变化映射，特别是HOG特征。要评价的第一个方法是稀疏回归，其次是结构稀疏性。最后，是研究等变映射在识别验证方面的作用。

稀疏回归：第一个实验（图3）展现了稀疏回归的变化（2）。目标是研究映射Mg=（AG，BG），预测结果会影响变换g后的图像的HOG特征。对每个变换，映射Mg是通过最小化正则化从1000个图像[5]中训练得到，有经验性风险（5）。性能也是在测试了超过1000个图像后由平均海林格的距离Kφ（GX）-Mgφ（x）Hell衡量。图像随机从ILSVRC12的训练和验证数据集中采集样本。

本实验重点是预测一个5×5的HOG元素组，它允许训练完整的回归矩阵5×5虽然回归算法比较浅显。此外，5×5阵列是从一个较大的9×9输入矩阵预测的，因此要通过图像旋转或重新调整来避免边界问题时，这两个限制会放宽一点。图3比较以下的方法来研究Mg：选择一致转换的Mg = 1，不通过正则优化来研究Mg（2）（最小二乘法 - LS），选择不同的正则Frobenius范数（岭回归 - RR）λ值，稀疏诱导正则r（3）（正向选择 - FS[25]），回归系数k在每个输出大小不同时取不同的值。

正如在图中所示，图3A，3B，LS 过拟合严重，这并不奇怪，因为Mg包含1百万甚至更多参数，对于这些小的HOG阵列。RR执行得不错，但它很容易被FS胜过，证实了这种解决方法非常稀疏（例如对于k = 5只有0.2%一百万的系数不为零）。最好的结果是FS中k = 5时得到的。如预期的，FS的预测误差是绕0的一个180◦旋转，因为这形状是准确的（第2节），但要注意，LS和RR追不到FS。所预料的是，误差在转换时较小，尽管在FS的范围内，误差仍然较小。

结构化稀疏的回归。以前的实验结论是稀疏是实现推广结论是必须一步。然而，通过研究Ms，例如通过向前选择或L1正规化，即使是达到结构化稀疏的方法代价也很大。接下来，我们使用正则结构化稀疏（4），其中每个特征从预先设定的邻域预测中得到并依赖于图像变换g。图3c重复图3a的实验为一个45◦旋转，但为了不限于邻域的m×m个HOG矩阵。为了能够跨越较大的区间m，则使用15×15 HOG矩阵。由于空间稀疏现在实行的实验，LS，RR，和FS执行中几乎都设定M≤3，k = 5并且m = 3的邻域内的FS能得到很好

[5] 海林格距离$(P_i(\sqrt{x_i} - \sqrt{y_i})2)1/2$ 在 HOG 特征直方图中比起欧氏距离更合适

的结果。结构化的稀疏还有一个显著的优势（表1），即它限制了需要解决的回归问题的影响。我们的结论是结构化的稀疏的评价远超于通用化的稀疏。

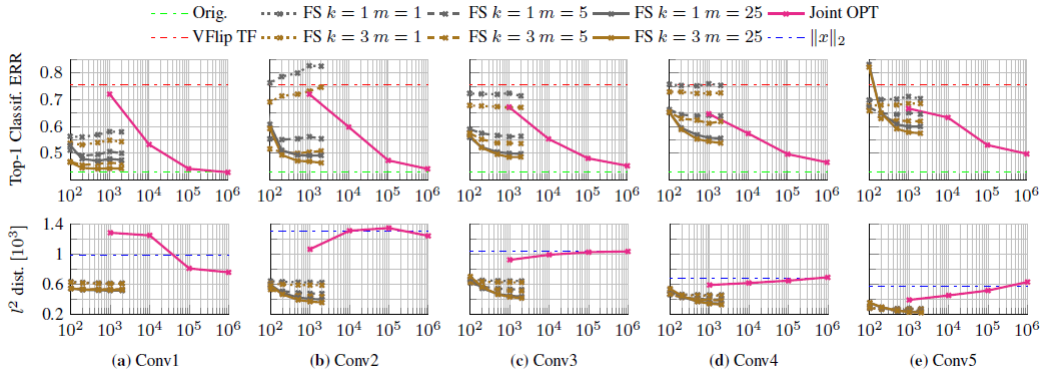

图6：CNN中的回归方法比较。研究CNN的不同层次的垂直翻转图像，得到了等变化映射Mg的回归误差。FS（灰线和棕线）和任务目标（紫色）和训练样本进行了对比评估。任务损失（顶部）和特征重构误差（底部）都在其中。在任务损失中，绿虚线是对原始图像的分类（最佳性能），红色虚线表示的是在分类器上变换图像后的情况（最坏情况）。在第二行中，每个单元的12重构误差可视作与基线 - 零向量的12平均距离在一起。

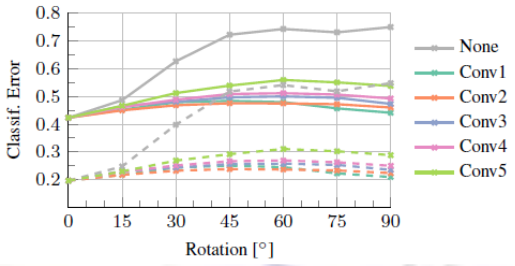

图7：学习等变化CNN映射的图像旋转。该设置是类似于图6，即扩展到几个旋转g，但仅限于面向任务的回归方法。实线和虚线分别表示了在ILSVRC12验证集中TOP1和TOP5的误差。

回归质量。迄今为止的结果都被赋予了特征的重构误差的期限;本段涉及了关于研究映射的物理特征。第一个实验是定性并使用HOGgle技术[30]以达到特征变换的可视化。如图5所示的可视化，Φ（GX）和MgΦ（X）在印证映射Mg中的确是几乎相同的，。第二个实验（图4）的计算结果不是HOG转换的特征质量，而是一个分类问题。为此，有一个SVM分类（W，Φ（x））是经过训练，可以区分狗和猫[16]（使用15×15 HOG的数据模板，400个训练数据和1000个测试数据，为平均的猫，狗图像）之间的区别。在Mg图像中进行一个逐渐增大的旋转或缩放G-1效果，得出SVM的结果（W，MgΦ（G-1X）I（通过MTG等效模型转化））。补偿分类器和原始分类器所有角度和比例都是相同的，而补偿的分类器（W，Φ（G-1X）在用于旋转时很快失败。我们总结认为是等变化变换需要有效地对视觉信息进行编码。

3.2. 深度表示中的同变性

在上一节中研究了等变化变换验证的浅度表示例如HOG。本节将延伸这些结果来进行深度表示，使用ALEXN CNN[10]作为部分参考，使用MatConvNet框架进行深层特征提取[29]。 ALEXN是一个二十个函数的组合，分成五个卷积层（包括过滤，MAX-池，规范化和ReLU）和三个完全连接层（过滤和ReLU）。实验目的是研究卷积层CONV1到Conv5在线性滤波之后的数值（在ReLU后研究线性变换层是很困难的，因为特征值是非负的）。

回归方法。第一个实验中（图6）是用不同的方法来研究CNN中的等变化映射Mg并进行比较。第一种方法是FS，计算不同邻域大小M和稀疏k。二是第2.1节中的面向任务，使用转换层。两者的特征的l2重建误差和分类错误（面向任务的损失）都表示出来了。正如第2.2节中所示，后者是补偿网络Φ2◦Mg◦Φ1（G-1X）在ImageNet ILSVRC数据（报告错误是在验证数据对比训练数据进行了优化的测量值）中的分类误差。该图表示了损失会产生更多的训练样本。为了得到结果，G被设定为垂直图像翻转，图7重复了面向任务目标和并将g从0至90度旋转（事实上，中间旋转是更困难，这表明了一个更好的Mg可以通过更仔细地研究插值和边界效应得到）的实验。

可以得到几个观察结果。首先，使用方法总比不使用方法要效果好（～75％最高-1的误差），如果不都是原分类器大多数都会恢复原状（43％）。这表明，线性等变映射Mg可以通过细胞神经网络研究得到了。其次，对于较浅的特征（直到CONV2）FS都是较好的选择，因为它需要比较少的训练样本，比起面向任务的损失，它会产生更小的重构误差与分类误差。相比较第3.1节，最优的m = 3和x = 25的设置基本上很少，但是，从Conv3起，任务导向的损失更少了，比起FS汇聚得到的分类误差要低得多。FS仍然出现了较小的重建误差，表示特征重建中并不能一直预测到分类结果。第三，匹配的分类误差在深层次下会增大，因为深层包含更多的直觉信息：正因为如此，在训练过程中完美地进行层的转换并且不经过变化（比如垂直翻转）应该是不可能的。

测试转换。接下来，我们调查一个CNN的不同层可以表示哪些几何转换（表2），特别是水平转换和垂直翻转，以及半缩放和90◦转动。首先，例如水平翻转和缩放的转换，研究等变映射不比研究特征不变更好：原因是，CNN隐含地学习了不变性等因素。对于垂直翻转和旋转，等变映射可以在基本上减小误差。特别是，开始的第几层很容易变化，需要确认其泛用性。

量化不变性。一种映射Mg的用法是用

来标识图像表示中的不变特征。这些是在变换后由它们自己预测地最好的那一部分。在实践中，在CNN的一个变换层中（第2.2节）标识出了不变性的信道，因为相同的变换过滤器均匀地施加在所有的空间位置里。在实践中，几乎从来没有精确地实现过不变性；相反，一个特征信道的不变性程度是由Mg的欧几里得范数的比率与相应行“对角线”的部分抑制组合后得出。然后，在Mg的第p行不变性的得出的最大可能由单位矩阵的（缩放的）行替换。最后，如果分类性能相对于Mg没有差到5%以上，修改后的转换Mg是可以评估和接受的。相应的对于最大可能p的特征信道可能会考虑到不变性。

表3表示了在ALEX CNN中N90◦旋转后水平和垂直翻转以及缩放的分析的结果。有几个显着的结论。首先，对于如水平翻转和重新缩放的转换，整体网络不变，主要是在Conv3或Conv4处体现了不变性。第二，不变性并不总是随着深度的增加而增加，例如CONV1比CONV2更具不变性，这是可能的，因为，即使在一个层中的特征信道是不变的，在随后的层中未必会汇集。第三，发生变换如垂直翻转和旋转90◦，不变特征数量意外地很小，有待进一步的方法来验证。

3.3. 深度表示的等价性

前两节研究的是等价性表示，本节着眼于它们的等价本身。我们的目标是要分辨是否存在异构表示即实际上可以通过和另一种表示方法替换来捕获相同的视觉信息如图2和图2.3。

Layer	Horiz. Flip		Vert. Flip		Sc. $2^{-\frac{1}{2}}$		Rot. 90°	
	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
None	0.44	0.21	0.75	0.54	0.61	0.37	0.75	0.54
Conv1	0.43	0.20	0.43	0.20	0.45	0.22	0.44	0.20
Conv2	0.45	0.22	0.46	0.22	0.48	0.24	0.46	0.22
Conv3	0.45	0.21	0.46	0.22	0.49	0.25	0.47	0.23
Conv4	0.44	0.21	0.48	0.24	0.49	0.25	0.49	0.25
Conv5	0.44	0.21	0.51	0.26	0.50	0.26	0.53	0.28

表2：CNN等变。ILSVRC12的验证性能，基于等变映射的特定转换的CNN补偿分类 。作为参考，未修改的ALEXN中最高和前五名个误差分别为0.43和0.20。

Layer	Horiz. Flip		Vert. Flip		Sc. $2^{-\frac{1}{2}}$		Rot. 90°	
	Num	%	Num	%	Num	%	Num	%
Conv1	52	54.17	53	55.21	95	98.96	42	43.75
Conv2	131	51.17	45	17.58	69	26.95	27	10.55
Conv3	238	61.98	132	34.38	295	76.82	120	31.25
Conv4	343	89.32	124	32.29	378	98.44	101	26.30
Conv5	255	99.61	47	18.36	252	98.44	56	21.88

表3：CNN不变性。通过分析相应的等变转换，确定了在ALEXN网络特征的信道中数量和比例。

Layer	IMNET → ALEXN		PLCS → ALEXN		PLCS-H → ALEXN	
	Top1	Top5	Top1	Top5	Top1	Top5
Conv1	0.43	0.20	0.43	0.20	0.43	0.20
Conv2	0.46	0.22	0.47	0.23	0.46	0.22
Conv3	0.46	0.22	0.50	0.25	0.47	0.23
Conv4	0.46	0.22	0.54	0.29	0.49	0.24
Conv5	0.50	0.25	0.65	0.39	0.52	0.27

表4：CNN等价性。性能上ILSVRC12通过设置的几个“Franken-CNN”将IMNET，PLC和PLCS-H的第一层和ALEXN的最后部分进行拼接。

为了验证这个想法，ALEXN CNN Φ = Φ’2◦Φ1的前几个层Φ1和IMNET的层Φ1替换，同时训练了ILSVRC12的数据，即训练了麻省理工学院的位置数据PLC [34]，并且将麻省理工学院的位置数据和ILSVRC12图像混合训练。这些表述有类似的部分，但不完全相同，结构和参数都完全不同。

表4显示了混合模型的最高性能φ’2◦Eφ1→φ“1◦φ1，其中，等价映射Eφ1→φ’1是作为ILSVRC12中拼接层（第2.3）的训练图像被研究。有一些显著的事实，第一，设置Eφ→φ=1的映射表示有最高> 99％的错误（表中未示出），匹配需要的直觉使得不同的参数化下使信道不能直接兼容。其次，等价性可以很好的在ALEXN和IMNET之间建立一个水平很高的Conv4和一个稍微不那么好的PLCS-H;然而，在PLCS深层处基本上相容性较差。具体来说，CONV1和CONV2在所有的情况下都可互换，而Conv5是不完全互换，特别是对的PLC。这证实了对于CONV1是CONV2一般图像编码直觉很重要，而Conv5较为具体。但是请注意，即使在最坏的情况下，性能也有显着提高的机会，这表明所有这些特征在一定程度上都是兼容的。

3.4. 结构化输出回归的应用

随着理论研究的深入，目前，该方向有一个直接的实际应用是关于第2节的等变映射，即以结构化输出回归[27]。在结构化输出回归中，图像x通过函数y（x）=argmaxy，zhφ（X，Y，Z）（直接回归）映射到y上，其中，z是可选的隐变量并且φ有共同特征的映射。如果y和/或Z包含几何参数，共同特征可以部分地完全改写为以减少φ（X，Y，Z）=Myzφ（x）hMTy,zw，φ(x)i的最大化（等变回归）。这样有两个计算上的优点：（1）φ（x）只需要计算一次（2）向量MTy,zw可以预先计算。这种想法的提出是为了论证位姿估计，其中y = g是一个类对象的可能位姿G-1∈-G中的几何变换。举个例子，考虑估算猫面部的位姿在PASCAL VOC 2007（VOC07）[7]数据来源于G经过（i）旋转或（ii）仿射变换（图9）。G的旋转是每10度和参考标注对鼻子到眼睛之间的中点通过连接确认旋转位置并进行均匀采样。这些重点在VOC07部分注解区域[2]是作为相应的重心得出的。在仿射变换G中不是由聚类向量[cTL，cTR，cTN]T得到而是从VOC007数据的300个例子中眼睛和鼻子的位置分析得出。这些集群是使用GMM-EM获得的训练数据，并用于在相同的位姿得到测试数据的映射的评估。G也包含一组仿射变换映射，重点是以每个集群为中心的规范框架中的[¯cTL，¯cTR，¯cT n]T。

矩阵Mg是预学习（从不包含猫的一般图像中）用第2节的FS，其中k =5以及m= 3。 由于VOC07中的猫的面部数据通常是垂直的。第二个更具挑战性的数据问题（用圈表示）是图像的旋转程度是随机变化的。（W，φ（GX））和同变性的（W，Mgφ（X））用300个训练样本和300个测试评估样本得到了评价函数得分函数。表5表示了在直接和等变回归情况下，HOG和CNN Conv3，Conv4和Conv5特征的计算的精度和速度。后者通常和直接回归差不

多，但会用22倍的速度再一次验证映射Mg。图8表示了对于不同的回归量累计误差曲线的形状。

$\phi(x)$	Bsln	HOG		Conv3		Conv4		Conv5	
		g	M_g	g	M_g	g	M_g	g	M_g
Rot [°]	23.8	14.9	17.0	13.3	11.6	10.5	11.1	10.1	13.4
Rot ○ [°]	86.9	18.9	19.1	13.2	15.0	12.8	15.3	12.9	17.4
Aff [-]	0.35	0.25	0.25	0.25	0.28	0.24	0.26	0.24	0.26
Time/TF [ms]	-	18.2	0.8	59.4	6.9	65.0	7.0	70.1	5.7
Speedup [-]	-	1	21.9	1	8.6	1	9.3	1	12.3

表5：等变回归。该表报告是对于猫的头部用结构化SVM回归量在直接/等变位姿情况下进行旋转/仿射误差的预测。该误差是在预期程度以内，分别以剩余旋转或平均距离作为关键点。用基线方法预测了一个持续的转换。

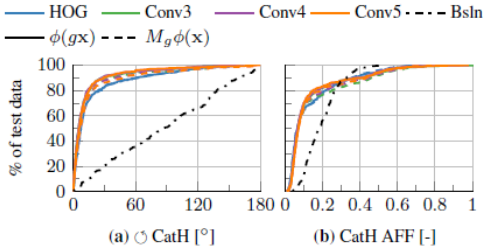

图8：等变回归误差。由旋转和仿射曲线产生的累积误差构成了表5的回归量。

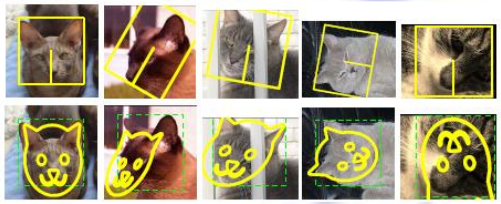

图9：等变回归的例子。旋转（上部）和仿射位姿（底部）是对猫面部通过VOC07数据进行预测。眼睛和鼻子的位置决定了仿射位姿的评估值。前四列都是回归成功的样本，最后一个是失败的样本。回归采用了CNN Conv5特征在绿色虚线框的计算结果。

4. 结论

本文介绍了研究图像表示的方法通过研究它们的等变性和等效性。它体现了用图像经线和相应的替代参照，用最容易预测的方法进行浅度表示和最开始几层的CNN深度转换，在不同的体系结构中都是有效的。更深层的一些特性程度较轻，和特定的任务有关。此外用一些分析工具，这些方法具有实用性例如通过加快结构输出回归量来用一个简单优雅的方式进行分类。

致谢。牛津工程科学DTA对Karel Lenc的支持。

参考文献

[1] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proc. BMVC*, 2014. 3

[2] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 8

[3] G. Csurka, C. R. Dance, L. Dan, J. Willamowski, and C. Bray. Visual categorization with bags of

keypoints. In *Proc. ECCV Workshop on Stat. Learn. in Comp. Vision*, 2004. 1

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005. 1, 2

[5] M. Denil, , B. Shakibi, L. Dinh, M. Ranzato, and N. de Freitas. Predicting parameters in deep learning. In *Proc. NIPS*, 2013. 1

[6] 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. *CoRR*, abs/1310.1531, 2013. 3

[7] M. Everingham, A. Zisserman, C. Williams, and L. V. Gool. The PASCAL visual obiect classes challenge 2007 (VOC2007) results. Technical report, Pascal Challenge, 2007. 8

[8] I. Goodfellow, H. Lee, Q. V. Le, A. Saxe, and A. Y. Ng. Measuring invariances in deep networks. In *Advances in neural information processing systems*, pages 646－654, 2009. 2

[9] H. J′egou, M. Douze, C. Schmid, and P. P′erez. Aggregating local descriptors into a compact image representation. In *Proc. CVPR*, 2010. 1

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012. 1, 2, 6

[11] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1), 2001. 1

[12] T. Lindeberg. Principles for automatic scale selection. Technical Report ISRN KTH/NA/P 98/14 SE, Royal Institute of Technology, 1998. 2

[13] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, 1999. 2

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2(60):91－110, 2004. 1

[15] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Proc. CVPR*, 2003.

2

[16] O. Parkhi, A. Vedaldi, C. V. Jawahar, and A. Zisserman. The truth about cats and dogs. In *Proc. ICCV*, 2011. 5

[17] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorizaton. In *Proc. CVPR*, 2006. 1

[18] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. In *CVPR DeepVision Workshop*, 2014. 3

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge, 2014.

3

[20] U. Schimdt and S. Roth. Learning rotation-aware features: From invariant priors to equivariant descriptors. In *Proc. CVPR*, 2012.

2

[21] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Le-Cun. Overfeat: Integrated recognition, localization and detection using convolutional networks. volume abs/1312.6229, 2014. 1

[22] L. Sifre and S. Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. In *Proc. CVPR*, 2013. 2

[23] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2013. 1

[24] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. ICCV*, 2003. 1

[25] K. Sj¨ostrand, L. H. Clemmensen, R. Larsen, and B. Ersbøll. Spasm: A matlab toolbox for sparse statistical modeling. *Journal of Statistical Software*, 2012. 5

[26] K. Sohn and H. Lee. Learning invariant representations with local transformations. *CoRR*,

abs/1206.6418, 2012. 2

[27] B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks.
In *Proc. NIPS*, 2003. 8

[28] A. Vedaldi and B. Fulkerson. VLFeat - An open and portable library
of computer vision algorithms. In *Proc. ACM Int. Conf. on Multimedia*, 2010. 2

[29] A. Vedaldi and K. Lenc. MatConvNet - convolutional neural networks
for MATLAB. *CoRR*, abs/1412.4564, 2014. 6

[30] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. HOGgles:
Visualizing object detection features. In *Proc. ICCV*, 2013. 2, 5

[31] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Localityconstrained
linear coding for image classification. *Proc. CVPR*, 2010. 1

[32] J. Yang, K. Yu, and T. Huang. Supervised translation-invariant sparse
coding. In *Proc. CVPR*, 2010. 1

[33] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional
networks. *CoRR*, abs/1311.2901, 2013. 1, 2

[34] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. *NIPS*,
2014. 7

[35] X. Zhou, K. Yu, T. Zhang, and T. S. Huang. Image classification using super-vector coding of local image descriptors. In *Proc. ECCV*, 2010. 1