指导教师：杨涛　　提交时间：

CVPR2015 Paper Translation

NO：01

姓名：吴轶成

学号：2013302606

班号：10011306

通过反演图像去了解其深层次含义

Aravindh Mahendran　　Andrea Vedaldi

牛津大学　　牛津大学

aravindh@robots.ox.ac.uk　　vedaldi@rpbots.ox.ac.uk

摘要

对于几乎任何一个图像认知系统而言，不管是尺度不变特征转换，视觉词袋模型还是像卷积神经网络，图像中的深层次含义都是一个重要的组成部分。然而，我们对其的认知还是极其有限的。在这篇文章中，我们将引导大家对表现出的视觉信息进行一个直接的理解，提出一个问题：给定一个已编码的图像，需要到什么程度才能重构它？为了回答这个问题，我们提供了一个大体的框架去反演图像含义，我们展示了这个方法去反演含义例如梯度方向直方图就表现的比同样适用于该情况的卷积神经网络要更精确，进一步，利用这种技术去研究最先进的卷积神经网络对首次表达图像含义的反演，在我们的发现中，通过不同程度的几何和光学不变性，可以展示出经过卷积神经网络算法操作后的若干个能够保持住图像精确信息的图层。

1.介绍：

绝大多数的图像认知和机器视觉方法都建立在图像的含义中，例如基元[17]面向梯度的直方图（尺度不变特征转换[20]和梯度方向直方图[4]），视觉词袋[3][27]，稀疏矩阵[37]，局部编码[34]和超矢量编码算法[40],VLAD[10],Fisher 向量[23]，以及最近的深度神经网络算法，特别是其中卷积算法的种类[15, 25, 38]。然而，尽管视觉含义的发展有了进步，设计方法仍旧发展的具有经验性，对其属性的理解仍旧缺乏。虽然当浅层特征表达的是正确的，但是特别当从数据中学习到的数以百万记的参数出现的时候，我们对于最新一代的深层含义表示更加如此。

(图.1,通过 CNN 编码出的是什么？这个图像显示了从参考图像的倒数第二层中提取的 1000 维的特征获得的参考图像的五个可能的重建

的 CNN[15](在提供 softmax 参数之前)上训练的图像网层数据。从数据的观点来看，所有这些图像几乎是等价的，该图像是彩色图像/屏幕效果最佳。)

在本文中，我们会对保留下来的描述的图像信息进行直接分析（FIG 1）.我们通过建模出一个表示图像 X 的函数$\Phi(x)$，然后计算出它的近似逆Φ^{-1}，通过$\Phi(x)$来重构图像 X，一个共同的假设是图像之间的差异性对其含义的影响无关(例子：光照或者是视角)，所以Φ应该不是独一无二的可逆的。因此,我们提出这是一个重建的问题,找到许多可能的重建,而不是一个。通过这么做，我们获得对其含义的不变性的理解。

我们的研究结果如下。首先，我们提出了一个普遍的方法去反演含义，包括 SIFT，HOG 和 CNN(第二章)。最重要的是，这种方法只使用了从图像含义和之前的通用的自然图像信息，从随机噪声作为初始启动手段，并因此捕获住仅含本身含义的信息。我们讨论和评价不同的合法化判罚作为先验的自然图像。第二，我们展示出，除去它的简单和通用性，这个方法恢复特征，相对于最近的替代方法而言,能从HOG中能够更好的重建。我们这样做了，强调的一些描述之间的含义之间的细微差别表述以及其对可逆性的影响。第三，我们利用反演技术去分析最近的深度卷积神经网络，利用抽样近似可能的表述去探索其中的不变性。我们与其深度的表示，展示了 CNN 逐渐构建一个一层层的变多的定量数目。第四，我们研究了存储在从选择的神经元组中重构出的图像含义中的信息的局部性，一部分在空间上，一部分在信道上。

本文的其余安排如下，章节 2：介绍反演方法，假设其是一个正规化的回归问题，并提出一些图像的先验知识去帮助重建。章节 3：引入了各种表达:HOG 和 SIFT 浅层交涉和利用最先进的CNN作为深层次含义表达的例子。它还显示了 HOG 和 DSIFT 如何像 CNN 一样实现，简化的计算出它的衍变。章节 4 和章节 5 利用反演技术去分析各自浅层含义(HOG 和 DSIFT)和深度 CNN 表示。最后，章节 6 汇总了我们研究结果。

关于卷积神经网络模型中对 HOG 和 DSIFT（逆 SIFT）的 matlab 代码，加上本文的代码都可以在 http://www.robots.ox.ac.uk/~vgg/research/invrep/index.htm 上获取。我们使用了 matconvnet 工具箱[32]去实现卷积神经网络。

相关的工作：

有工作通过可视化的方式表示理解一个显著量。和我们的工作最相关的是 Weinzapfel[35] 和 Vondrick[33]人所做的工作，分别利用逆稀疏 SIFT 和 HOG 的特性。我们还注意到由 Kato 等视觉词袋模型的

工作。[13]和德安杰洛等的关于局部二进制描述符。[5]。而我们的目标是类似于这些现有作品，我们的方法是基本上从技术观点来看不同，是基于这一个简单的正则回归问题的梯度下降。这样做的好处是，我们的技术同样适用于浅（SIFT，HOG）和深层（CNN）表示。 我们没有在反转视觉词袋进行任何实验功能或局部二元描述符。相比现有反演技术密集浅的表示方法，如作为 HOG[33]，它也能够从数量和质量上示出实现优异的结果。

[33，35]的一个有趣的结论是，虽然 HOG 和 SIFT 可能不完全可逆的，它们捕获的关于图像的信息的显著量。这是在明显的矛盾与 Datu 等人的结果。 [29]谁表明它有可能使任何两个图像看起来几乎相同的 SIFT 空间加入对抗噪音。对称效果证明由 Szegedy[28]等的对于 CNN 的结果。，在渐进的对抗性噪声的量足以改变所预测的类的形象。明显不一致很容易的解决，作为方法[28，29]所需要的高通结构噪声加入是不太可能多发生于自然图像中。

我们的工作也涉及到 Zeiler 和 Fergus[38]的 DeConvNet 方法，原反馈网络计算以确定哪些图像块负责某些神经激活。西蒙尼扬等[26]然而证明了 DeConvNets 可以被解释为一网络输入/输出关系的敏感性分析。一个结果便是 DeConvNets 不研究问题这里采用的意义，代表反转的这有显著方法论的后果；例如，DeConvNets 需要有关辅助信息来激活几个中间层，而我们的反转仅使用最后的图像的代码。换句话说，DeConvNets 怎么看待某些网络输出得到，而我们寻找什么样的信息是通过网络保存输出。

反转表达含义的问题，特别是基于 CNN 的问题，是关系到反转含义的神经网络，并获得显著关注的期望值。算法类似于反向传播技术。这里开发由[16，18，21，36]提出，沿基于抽样替换优化策略。然而，这些方法没有使用自然图像先验。如我们这样做，也没有被应用到当前一代的深层网络中。其他作品[11，30]专业的反相在自学习系统的情况下网络不会作进一步讨论。[1]提出了学习次一个神经网络充当原的逆网络，但是，这是通过以下事实的逆是复杂的，并且通常不是唯一的。最后，自动编码器的体系结构[9]形成的网络，它们的反作为的一种形式在一起监督;我们在这里，而不是有意可视化前馈和有区别地训练CNN来表示现在流行计算机视觉。

2.反转含义：

本节介绍我们的方法来计算的图像表示的近似逆。这被配制为找到一个图像，其表示最佳匹配给出的[36]的问题。从形式上看，鉴于一个表示函数Φ：

$R^{H*W*C} \rightarrow R^{d}$ 和一个表示 $\Phi_0 = \Phi(x_0)$

被反转，重建认定该目标最大限度地减少了图像 x∈ R^{H*W*C} 。

$$X^* = \arg\min L(\Phi(x), \Phi_0) + \lambda R(x) \quad (1)$$

其中损失 ℓ 比较图像表示Φ(X)在一个目标Φ(X)和 R：R^{H*W*C} →R 是再建前捕获的自然图像。

最小化（1）结果中的图片在 X^* 认为“酷似” X_0 从代表性的观点。虽然有可以是没有唯一的解决这个问题的，取样的可能重建的空间可用于表征图像的空间的表示认为是等效的，揭示其不变性。多个重建从初始化随机不同的起始位置的梯度下降优化得到。

接下来我们将讨论的损失和再建的选择:

损失函数。有许多可以选择的损失函数 L。当我们使用欧氏距离: $L(\Phi(x), \Phi_0) = \| \Phi(x) - \Phi_0 \|^2$ ，有可能完全改变损失的本质，例如优化选择的神经反应。后者在[6,26]中被用来生成图像，代表了所给的神经元。

调整。判别训练表示可能丢弃大量的低级图像数据因为它们对于高级任务来说通常没有意义。因为这些信息对于可视化仍然是有用的，它可以得到部分恢复通过限制反转自然图像的子集 $\chi \subset R^{H*W*C}$ 。然而，最小化 χ 需要对这种模式进行建模。作为一个代理可以在合并前重建一个适当的图像。这里我们尝试两个这样的先知先觉。第一个就是 $R_\alpha(x) = \| x \|_\alpha^\alpha$,其中 x 是矢量化和平均减去的图像。通过选择一个相对较大的指数(α = 6 用于实验)图像的范围内保持在目标区间,而不是不同的。

第二个更丰富的调整是总变量(TV) R_{V^β} （x），图像由分段常数补全。对于连续函数（或分布） f：$R^{H*W} \supset \Omega \to R$ ，TV标准有右式给出：

$$R_{V^\beta}(f) = \int_\Omega ((\frac{\partial f}{\partial v}(u,v))^3 + (\frac{\partial f}{\partial v}(u,v)^2)^{\frac{\beta}{2}}$$

$x \in R^{H*W}$ ，其中 β = 1. 这里的图像是离散的（x∈RH×W），TV标准由有限差分近似代替：

$$R_{V^\beta}(f) = \sum_{i,j} ((x_{i,j+1} - x_{i,j})^2 + (x_{i+1,j} - x_{i,j})^2)^{\frac{\beta}{2}}$$

这是观察到的经验，在抽样中TV规范化(β = 1)，也由在CNN最大pooling引起的，导致“尖峰”的重建。这是一个基于电视的图像插值中的已知问题（见例如图3［2］），如图2.1eft反转时的说明在CNN的一层。“尖峰”发生在样本的位置处，因为：（1）电视规范在两个样本之间的任何路径

仅取决于整体强度变化（不在于清晰度的变化）和（2）在二维图像上，它是最佳的集中边界附近的一个小边界的急剧变化。超拉普拉斯与β＜1经常被用来作为一个更好自然图像梯度统计的匹配[14]，但它们只会加剧这个问题。相反，我们权衡锐利度的图像，通过选择β＞ 1去除这些文物，通过限制大梯度分布变化的区域，而不是集中他们在一个点或曲线。我们称之为V调整。如图2所示（右边），尖峰处被移除β = 2，但图像被模糊因为的边缘梯度β大于1 的地方被平滑了。

图二：左：尖峰逆规范 1 功能细节所示。右：尖峰由改变为重建系数 λ_{V^β} 中的β为 2

当目标的重建是一个彩色图像时，每个颜色通道都要进行两次调整。

平衡不同的条件。平衡损失和调整（s）需要注意一点。同时优化调整可以通过交叉验证实现，重要的是从合理设置参数开始。首先，损失是由归一化的版本 $\|\Phi(x)\text{-}\Phi_0\|_2^2 / \|\Phi_0\|_2^2$。这样修复它的动态范围，为标准化后的在最佳效果附近的损失可以预计将包含在[0,1）区间，在最佳效果处接触零。为了使得调整的动态范围相媲美一个目标可以为解决X*，具有大致统一欧氏范数。虽然主要是在不敏感的缩放的图像范围，不是那么精确的对于CNN的第几层，在偏差调整到“自然”工作范围。这可以考虑客观的 $\|\Phi(\sigma x)-\Phi_0\|_2^2 / \|\Phi_0\|_2^2 + R(x)$，缩放比例σ是自然图像的平均欧氏规范。其次，α-规范调整的乘数λ应选择鼓励重建图像σX是包含在一个自然的范围[-b，b]（比如，大多数CNN实现= 128）。如果大多数像素在σX有一个大小相似的B，then R(x) ≈HWB/σ，and $\lambda_\alpha \approx \sigma^\alpha / (HWB^\alpha)$。类似的建议选择V规范调整系数。作为 $\lambda_{V^\beta} \approx \sigma^\beta / (HW(aB)^\beta)$，其中的一小部（例如，= 1%）有关的图像的动态范围的梯度。目标函数的最终形式是

$$\|\Phi(\sigma x)-\Phi_0\|_2^2 / \|\Phi_0\|_2^2 + \lambda_\alpha R_\alpha(x) + \lambda_{V^\beta} R_{V^\beta}(x) \quad (3)$$

因为它的性质，它是一般的非凸性的。接下来我们讨论如何优化它。

2.1 优选法:

找到目标的优化法看起来像是无法解决的工作，比如大多数涉及强非线性的直径表示法特别是，深度表示法链接着一些非线性层。尽管如此，简单的梯度下降过程显示出在学习数据中模型方面非常有效，按理说

这本该是更难的工作任务 所以，用 GD 算法来解决也不是不合理。扩展 GD 算法在合并动量上证明在学习深度网络上很有用，如下文所讨论:

动量:GD 算法被扩展到使用动量

$$\mu_{t+1} \leftarrow m\mu_t - \eta_t \nabla E(x),$$

$$X_{t+1} \leftarrow X_t + \mu_t$$

在这里 $E(x)= L(\Phi(x),\Phi_0)+\lambda R(x$ 是目标函数矢量 μt 是剩余的几个梯度的加权平均值，而衰减系数 m=0.9，学习将通过一个固定的学习率 ηt 继续进行几百次迭代并减少十倍，直到收敛应用 GD 需要计算由 Φ(x)表示的损失函数的导数虽然欧式平方的损失是平滑的。

但这不适合代表情况，CNNs 的一个主要特点是拥有计算每一个计算层导数的能力，构成了使用反向传播的整个函数的整体导数的后者我们 HOG 的实施和向 CNN 导入 DSIFT 让我们也能对这些表示方法运用相同的技术

表现形式

这一部分描述了在论文中研究的图片表示：此外，它展示了为了计算他们的导数，如何在一个标准的 CNN 架构中实施 DSIFT 和 HOG 方法利用章节 2.1 中的算法能够计算导数在一个标准的 CNN 架构中实施 DSIFT 和 HOG 方法让计算导数变得方便快捷。

CNN-A：深度网络作为一个参考深度网络我们考虑 Caffe-Alex 模型，此模型几乎重现了由 Krizhevsky 等人设计的网络。这个和其他很多相似的网络替换了以下的架构快：线性卷积，ReLU 门控，空间池化方法，组正常化每一个色块以一个 d 尺寸的图片作为输入并产生一个 K 尺寸的输出。另外架构块可以铺垫图像，用卷积的零点和池化算法的无穷点，或者对数据二次抽样剩余的一些层次被认为“完全连通”作为和图像的尺寸一致的线性滤波器的支持然而，他们和其他所有方面的过滤层对等，表 2 陈述了 CNN-A 的结构细节。

CNN-DSIFT and CNN-HOG

这一部分将展示 DSIFT [19, 22] 和 HOG [4]如何作为 CNNs 被实施这正式确立了 CNNs 和这些标准的表达之间的关系对于这部分的反演算法，它也使这些表达式的导数计算更简单 DSIFT 和 HOG 在 VLFeat 库中的实施被用做数字参考。这些和劳式 SIFT 还有 DPMV5 HOG 是等价的。

SIFT 和 HOG 涉及范围：计算和叠加图像梯度，对细胞直方图进行池化梯度分级，分组细胞进入色块，最后正常化色块(where K = 8 for SIFT and K = 18 for HOG).用 g 表示给定像素的梯度并考虑对此像素叠加进入 K 的取向之一。这会得到两步：定向滤波和门控。这个 K^{th} 存在于 $G_K = u_{1k}G_x + u_{2k}G_y$ 中，其中

$$u_k = \begin{bmatrix} \cos\frac{2\pi k}{K} \\ \sin\frac{2\pi k}{K} \end{bmatrix}$$

$$G_x = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

$$G_y = G_x^T$$

定向滤波的输出是(g, u_k)沿着方向的投影u_k，一个合适的门控函数对直方图元素实行像素叠加 DSIFT 使用双线性定位分级。

$$h_k = \| g \| \max\{0, 1 - \frac{K}{2\pi}\cos^{-1}\frac{< g, u_k >}{\| g \|}\},$$

鉴于 HOG 使用硬性任务，过滤是一个标准的 CNN 操作但这些分级函数不是，而它们的实施是简单的，一个有趣的替代品是近似的双线性分级。

$$h_k \approx \| g \| \max\{0, \frac{1}{1-a}\frac{< g, u_k >}{\| g \|} - \frac{a}{1-a}\} \propto \max\{0, < g, u_k > - a \| g \|\}, a = \cos 2\pi / K.$$

规范依赖 g 仍然不是标准的，但 ReLU 算子是标准的，此算子展示出程度近似分级可以实现在典型的 CNNs 中。

下一步是为了对细胞直方图池化分级梯度使用双线性空间池化，其次是从 2 × 2 (HOG) 或 4 × 4 (SIFT) 细胞中提出色块这两个操作都能由线性滤波器的堆叠实现，细胞块是正常的 1 的平方，那是一个标准的层局部反应的特殊事件，对 HOG 来说，色块被进一步分解退化为细胞，需要另一个滤波器组。最后，描述符的值被夹在在应用 y= min{x,0.2}的每个要素中。

结论是逼近 DSIFT 和 HOG 可以运用添加了不一般的梯度范数的平常的 CNN 要素被实施，然而，在研究的 CNNs 中所有涉及到的滤波器比起通用 3D 滤波器要更加系数和简单。

descriptors method	HOG HOGgle	HOG our	HOGb our	DSIFT our
error (%)	66.20 ±13.7	28.10 ±7.9	10.67 ±5.2	10.89 ±7.5

表 1：不同的表示含义平均重建误差反演方法，适用于 HOG 和 SIFT。高表示 HOG 与双线性定向分配。示出的标准偏差的误差，而不是平均误差的标准偏差。

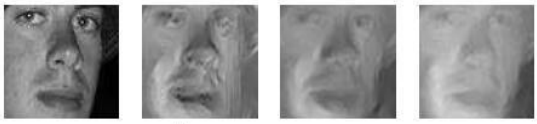

图 4. V^β 的影响。图中可见同样的反演算法。图 3（d）用于用一个较小的（V^β = 0.5），可比相同情况的（V^β = 5.0），以及较大的（V^β = 50）正系数的正规化建设更好。

在本文的其余部分，我们将使用的确切 CNN 等值 DSIFT 和 HOG，使用修改或附加组件来满足 CNN 需要。这些细胞神经网络是从 VLFeat 参考实现数字区分，但是，真到自己的 CNN 性质，允许计算功能来源于衍生物所要求的第二章的算法。

接下来，我们应用来自章节二的算法。 在 CNN-A，CNN-DSIFT 和

CNN-HOG 来分析我们的方法。

4. 浅层含义实验

本节评估表示章节 2 中反转的方法。 通过把它应用到 HOG 和 DSIFT。该分析包括一个定性的（图 3）和定量（表 1）与现有技术的比较。

$$\|\Phi(\mathrm{x}^{*})-\Phi(\mathrm{x}_{i})\|_{2} / N_{\Phi}$$

定量评价报告的标准化重构误差平均距离超过 100 幅图片在 ILSVRC 挑战 2012[24]验证数据（图像 1 至 100）规格化是必不可少的，在确定的特征所占的欧几里得距离中所占据的体积下：如果特征都紧靠在一起，然后连 0.1 欧几里德距离是非常大的，但是，如果特征被传播出去，那么即使 105 欧几里得距离可以是非常小的。我们用了 N_{Φ} 是之间的平均成对欧氏距离 $\Phi(\mathrm{X}_{i})$ 跨越 100 幅图像的。

我们在方程 3 中确定的参数，其中 $\lambda_{\alpha}=2.16*10^{8}$，$\lambda_{\nu^{\beta}}=5$，和 $\beta=2$

这需要解决几个微妙之处。在 DSIFT 梯度贡献通常是由一个高斯加权在每个描述为中心（4×4 单元块）;这里我们用 VLFeat 近似（fast 选项）的加权细胞而非梯度，可在掺入块成形滤波器。在 UoCTTI HOG，细胞中含有两种导向，未取向的梯度（在总共 27 个部件），以及 4 个纹理元件。后者被忽略为简单起见，虽然未取向的梯度面向获得于作为平均在块形成滤波器中。

奇怪的是，在 UoCITI HOG 中 l^{2} 归一化因子计算只考虑在一个块中的未取向的梯度分量，但是对于全局而言的，这需要修改规范化操作。最后，当块被分解回细胞，它们被平均，而不是叠置如原始 Dalal-Triggs HOG，它可以实施于块分解过滤器。）

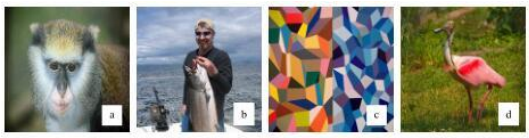

Figure 5. Test images for qualitative results.

我们的方法最接近的替代方案是 HOGgle，这是由 Vondrick 等人提出的技术[33]。用于 HOG 特征的可视化。该 HOGgle 代码是公开的，可从作者的网站得到并在整个实验中使用。重要的是，HOGgle 预先训练反转 UoCITI 实施 HOG 的，其数值相当于 CNN-HOG（章节 3），允许算法之间的直接比较。

比起我们的方法，HOGgle 快（2-3s 相比于 60s，在相同的 CPU 上），但效果不是很准确，因为它是显而易见的定性（图 3.C 相比于 d）和定量（66% 相比于 28%的重建误差，见表 1）。有趣的是，[33]提出类似的直接优化方法（1），但表明它不大于 HOGgle 更好的表现。 这个演示再建中选择的重要性和计算的表示的衍生物的能力。进一步分析了再建 $\lambda_{\nu^{\beta}}$ 的效果在图 4（后来在表 3）：没有该现有的信息，该重建呈现出一个离散假象的显著量。

在速度方面，优化的优点（1）是它可以切换到使用的 GPU 代码来立即

给出的 CNN 的基本框架;这样做会导致十倍比较的加快。另外的基于 CNN 的实现 HOG 和 SIFT 正在优化，理论上它应该是可以加速它们数倍。

也明显的是，不同的表示可以更容易或很难反转。特别是，修改 HOG 使用双线性梯度方向分配为 SIFT（章节三）显著减少了重建误差（从 28%下降至 11%），并改善重建品质（图 3.E）。更令人印象深刻是 DSIFT：它是定量在 HOG 中以类似的双线性方向，但产生显著更精细的图像（图 3.F）。由于 HOG 用在渐变的更精细的量化比较筛选中，而不是以相同的细胞的大小和采样，这样的结果可以改变 HOG 的重块的正常化，显然丢弃了比 SIFT 更多的图像信息。

5. 深层含义实验

本节中的计算结果施加在章节三的反转方法，CNN-A 中描述，相比于 CNN-HOG 和 CNN-DSIFT， 这个网络显示的更大更深。

因此，似乎该反演问题应是相当困难。CNN-A 是不是手动产生的，而是从 ImageNet ILSVRC 2012 年的 1.2M 的影像学数据[24]中训练产生。

图 3.重构的不同表示反演方法的质量，适用于 HOG 和 SIFT。高表示与 HOG 双线性定向分配。该图像在屏幕上效果最佳。

layer	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
name	conv1	relu1	mpool1	norm1	conv2	relu2	mpool2	norm2	conv3	relu3	conv4	relu4	conv5	relu5	mpool5	fc6	relu6	fc7	relu7	fc8
channels	96	96	96	96	256	256	256	256	384	384	384	384	256	256	256	4096	4096	4096	4096	1000
rec. field	11	11	19	19	51	51	67	67	99	99	131	131	163	163	195	355	355	355	355	355

表 2. CNN-A 结构。该表沿 CNN-A 的结构指定于每个神经元的感受域大小。从 16 到 20 过滤器层操作为“完全连接”：给定的 227×227 像素的标准图像输入大小，他们的支持覆盖了整个的图片。同时还指出，他们的感受域大于 227 个像素，但也可以包含在填充的图像域。

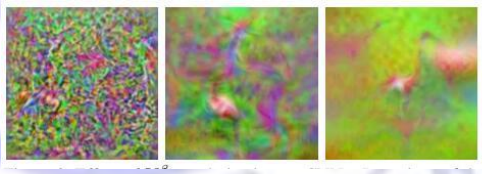

图 8. 在 CNN 的正规化中V^{β}的影响。在反演 CNN-A 的最后面一层，再建的$\lambda_{\nu\beta}$逐渐变大。该图像是彩色图像/屏

幕效果最佳。

第二章的算法用于反转从每个 CNN 层 100 张 ILSVRC 的验证图像所获得的代码（这些都不是用来训练 CNN-A 模型[15]）。类似于第四章所言：计算归一化的逆误差和在表 3 中所报告的实验是由 λ α 固定为 2.16×10 8 的固定值，并逐步增加 $\lambda_{v^{\beta}}$ 为原来的十倍，相对的从开始重复小值 λ = 0.5。该图像网层中 ILSVRC 平均图像又加回到可视化之前重建训练网络时，因为这被扣除。 有些出人意料的是，定量结果表明，CNN 显示出其实，不是比 HOG 更难反转。错误很少，没有超过 20%，这是可比的精度 HOG（章节 4）。其最后一层相对而言是比较容易与反转的平均误差为 8.5%。

我们选择在再建系数的基础上进行定量和定性研究每个代表/层重建。我们对于层 1-6 设置 λ 1 =0.5，对于层 7-12 设置 λ 2 =5.0 和对于层 13-20 设置 λ 3 =50。该相应于这些参数的误差值被标记在表 3 中增加 $\lambda_{v^{\beta}}$，这样对于第一层来说导致负面效应，但对于后者的层有助于恢复更直观的解释的重建工作。虽然这个参数可以通过交叉验证的标准化，被调谐在重建误差的基础上，首选的是定性分析一个选择，因为我们想的方法最终是来产生视觉上有意义的图像。

定性的说，图 6 示出了用于重建从 CNN-A 的每一层测试图像的编码。其进展是卓越的。前几个层是本质上的图像的可逆码。所有卷积层保持图像的内容的忠实表示，虽然随着其模糊。4,096 维的完全连接层也许更有趣，因为它们反转回相似但不相同的原始图像中发现的那些部分的组合物。从去 relu7 到 FC8 进一步降低维，到达只是 1000 维;然而一些这些视觉元素的仍可被识别。相似的效果可以在重建中被观察到在图 7. 该图除此之外还包括一个重建的抽象图案，这是不包括在任何图层网络类的;不过，这一现象清楚地表明，即使是很深层的捕捉视觉信息，所有的 CNN 代码也能获得独特的视觉原始图案的特征。

接着，图 7 检查被捕捉不变性，考虑到多个重建出每个深层的 CNN 模型。这些图像进行仔细检查，发现该代码捕捉到变大的物体逐渐变形。在“Flamingo”重建中，尤其 relu7 和 FC8 反转回的多个副本对象/在不同的位置和规模的部分。

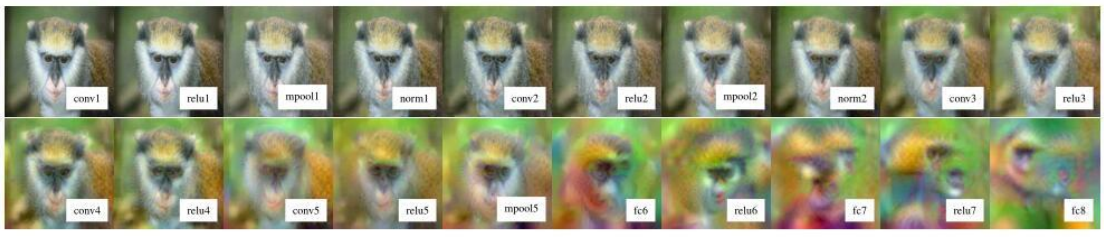

图 6. CNN 重建。图 5.a 从 CNN-A 的每一层进行图像重建。为了产生这些结果，各层的正规化系数被选择为在表 3 中突出显示的行，图在彩色/屏幕效果最佳。

λ_{V^β}	1 conv1	2 relu1	3 pool1	4 norm1	5 conv2	6 relu2	7 pool2	8 norm2	9 conv3	10 relu3	11 conv4	12 relu4	13 conv5	14 relu5	15 pool5	16 fc6	17 relu6	18 fc7	19 relu7	20 fc8
λ_1	**10.0** ±5.0	**11.3** ±5.5	**21.9** ±9.2	**20.3** ±5.0	**12.4** ±3.1	**12.9** ±5.3	15.5 ±4.7	15.9 ±4.6	14.5 ±4.7	16.5 ±5.3	14.9 ±3.8	13.8 ±3.8	12.6 ±2.8	15.6 ±5.1	16.6 ±4.6	12.4 ±3.5	15.8 ±4.5	12.8 ±6.4	10.5 ±1.9	5.3 ±1.1
λ_2	20.2 ±9.3	22.4 ±10.3	30.3 ±13.6	28.2 ±7.6	20.0 ±4.9	17.4 ±5.0	**18.2** ±5.5	**18.4** ±5.0	**14.4** ±3.6	**15.1** ±3.3	**13.3** ±2.6	**14.0** ±2.8	15.4 ±2.7	13.9 ±3.2	15.5 ±3.5	14.2 ±3.7	13.7 ±3.1	15.4 ±10.3	10.8 ±1.6	5.9 ±0.9
λ_3	40.8 ±17.0	45.2 ±18.7	54.1 ±22.7	48.1 ±11.8	39.7 ±9.1	32.8 ±7.7	32.7 ±8.0	32.4 ±7.0	25.6 ±5.6	26.9 ±5.2	23.3 ±4.1	23.9 ±4.6	**25.7** ±4.3	**20.1** ±4.3	**19.0** ±4.3	**18.6** ±4.9	**18.7** ±3.8	**17.1** ±3.4	**15.5** ±2.1	**8.5** ±1.3

表 3. CNN-A 反转误差。CNN-A 各层平均反转误差百分比(标准化)和不同数量正规化的V^β：λ1= 0.5，λ2 =10 λ1 和 λ3 =100 λ1。黑体字是对应于定性和定量效果最好的正则化误差值。在这个表中指定的偏差是误差的标准偏差，而不是标准偏差的平均误差值。

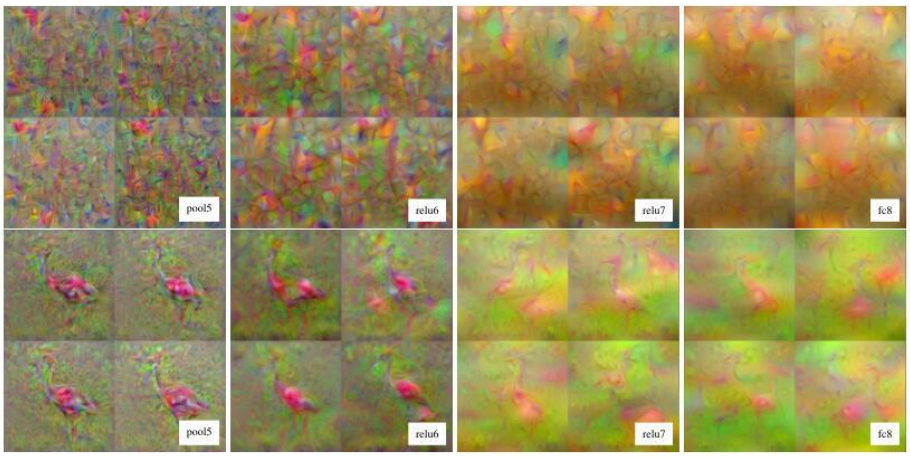

图 7. CNN 不变性。图 5.c-d 从 CNN-A 不同深度代码获得的多个图像的重建，图在彩色/屏幕效果最佳。

注意，所有这些和原始图像从 CNN 模型的视点上几乎无法区分；因此，在最深层进行重建时缺乏细节这一点值得注意，表明网络截图只是对象的一个草图，但这对于分类显然就足够了。大大降低了正则化参数仍然得到非常精确的反演（图 8），但这跟自然图像几乎没有相似之处。这证实了 CNNs 有很强的非天然的混杂因素。

有趣的是，很多反向图像有大量绿色区域(参见图 11)。我们声称，这是一个网络的属性，而不是之前的自然图像。CNN-A 的前一层的影响如图 8 所示。之前只鼓励平滑，因为它是当量（对于 β = 2），以惩罚重构图像的高频分量。更重要的是，它同样适用于所有颜色通道。逐步消除之前，随机高频分量占主导地位，难以辨别人可解释的信号。但是，绿色仍然存在；我们和麻省理工学院的 CNN[39]一样，在十张不包含绿色补丁的图像上进一步验证了这个现象。

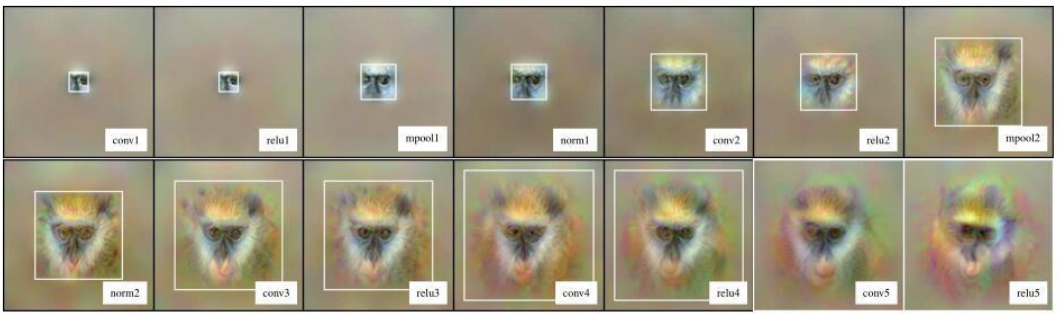

图 9. CNN 接受域。从不同深度的 CNN-A 的中央 5×5 神经元领域重建的图像。如图 5 所示。白盒标志着 5×5 的神经元领域视野。视野是 conv5 和 relu5 整体图像。

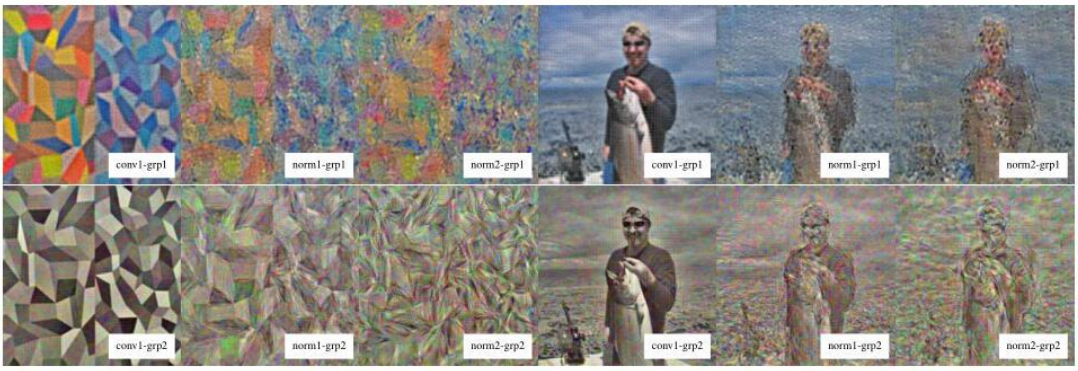

图 10. CNN 神经流。从 CNN-A 的两个神经流进行图 5.c-b 的图像重建。该图最清楚地是在彩色/屏幕看到。

我们现在检查从不同层次 CNN 的神经反应的子集获得的图像重建。图 9 通过在每一层重建中央 5×5 图像特征探讨了码的局部性。规范者鼓励要关掉不会导致神经反应的图像。图像中位置特征是显而易见的,不太明显的是,有效神经元的感受视野在某些情况下显著小于理论—图像中显示为一个白盒。

图 11.CNN 模型的多样性。mpool5 重建表明，网络甚至在如此深的水平上保留丰富的信息。这个图像是在彩色/屏幕（放大）上观看效果最佳。更多的定性结果在项目网页上提供。

最后，图 10 从特征信道的一个子集来重建图像。 CNN-A 实际上包含特征通道的两个子集，第一数层相互独立（最多范数 2）[15]。从每个子集单独重建的图像清楚地表明，一组是朝彩色信息调谐而第二组向尖锐的边缘和亮度组件调谐。值得注意的是，这种行为自发出现在学习网络中。

6. 总结

本文提出了一种优化的方法来转化基于优化的目标函数梯度的浅、深表示法。与替代方案相比，一个关键的区别是使用了图像先验，如V^{β}常态。其可以恢复由表示法除去的低级别图像。适用于 CNNs，阐明出现的每一

层的信息。尤其是，在网络中形成的逐渐不变的和抽象的图像内容。

参考文献：

[1] C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995. 2

[2] Y. Chen, R. Ranftl, and T. Pock. A bi-level view of inpainting-based image compression. In *Proc. of Computer Vision Winter Workshop*, 2014. 3

[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 *CVPR*, 2005. 1, 4

[5] E. d'Angelo, A. Alahi, and P. Vandergheynst. Beyond bits: Reconstructing images from local binary descriptors. In *ICPR*, pages 935–938, Nov 2012. 2

[6] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. Technical Report 1341, University of Montreal, 2009. 3

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 4

[8] R. B. Girshick, P. F. Felzenszwalb, and D. McAllester. Discriminatively trained deformable part models, release 5. http://people.cs.uchicago.edu/˜rbg/latent-release5/. 4

[9] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786), 2006. 2

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

[11] C. A. Jensen, R. D. Reed, R. J. Marks, M. El-Sharkawi, J.-B. Jung, R. Miyamoto, G. Anderson, and C. Eggen. Inversion of feedforward neural networks: algorithms and applications. *Proc. of the IEEE*, 87(9), 1999. 2

[12] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/, 2013. 4

[13] H. Kato and T. Harada. Image reconstruction from bag-of-visualwords. In *CVPR*, June 2014. 2

[14] D. Krishnan and R. Fergus. Fast image deconvolution using hyperlaplacian priors. In *NIPS*, 2009. 3

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1, 4, 6, 8

[16] S. Lee and R. M. Kil. Inverse mapping of continuous functions using local and global information. *IEEE Trans. on Neural Networks*, 5(3), 1994. 2

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

[18] A. Linden and J. Kindermann. Inversion of multilayer nets. In *Proc. Int. Conf. on Neural Networks*, 1989. 2

[19] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999. 4

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

[21] B.-L. Lu, H. Kita, and Y. Nishikawa. Inverting feedforward neural networks using linear and nonlinear programming. *IEEE Trans. on Neural Networks*, 10(6), 1999. 2

[22] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-offeatures image classification. In *ECCV*, 2006. 4

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

[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. 5

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

[26] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proc. ICLR*, 2014. 2, 3

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

[28] C. Szegedy,W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013. 2

[29] A. Tatu, F. Lauze, M. Nielsen, and B. Kimia. Exploring the representation capabilities of the HOG descriptor. In *ICCV Workshop*, 2011. 2

[30] A. R. V ́arkonyi-K ́oczy and A. R ̈ovid. Observer based iterative neural network model inversion. In *IEEE Int. Conf. on Fuzzy Systems*, 2005. 2

[31] A. Vedaldi. An open implementation of the SIFT detector and descriptor. Technical Report 070012, UCLA CSD, 2007. 4

[32] A. Vedaldi and K. Lenc. MatConvNet: CNNs for MATLAB. http: //www.vlfeat.org/matconvnet/, 2014. 2

[33] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. HOGgles: Visualizing object detection features. In *ICCV*, 2013. 1, 2, 5, 6

[34] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Localityconstrained linear coding for image classification.

CVPR, 2010. 1
[35] P. Weinzaepfel, H. J ́egou, and P. P ́erez. Reconstructing an image from its local descriptors. In *CVPR*, 2011. 2
[36] R. J. Williams. Inverting a connectionist network mapping by backpropagation of error. In *Proc. CogSci*, 1986. 2
[37] J. Yang, K. Yu, and T. Huang. Supervised translation-invariant sparse coding. In *CVPR*, 2010. 1
[38] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 1, 2
[39] 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. 8
[40] X. Zhou, K. Yu, T. Zhang, and T. S. Huang. Image classification using super-vector coding of local image descriptors. In *ECCV*, 2010.