Beihang University, IAI

计算机视觉, 2023 Fall

Understanding Deep Image Representations by Inverting Them

阅读报告

姓名　魏少杭

weish @buaa.edu.cn

2023 年 12 月

目录

1 主要创新思想和学术贡献

本文提出了一套通用的框架来反转图像的表示，来研究给定图像编码情况下重建图像本身的方法。结果表明，CNN 的中间层保留了关于图像的摄影准确信息，具有不同程度的几何与光度不变性。本文在反转浅层和深层表示来重建原图的能力上具有相似性。

1.1 动机

视觉的表征方法主要包括了传统的浅层表示方法，如 SIFT[1]、HOG[2]等，以及卷积神经网络等深度神经网络。这些视觉表征方法缺少足够的理论性设计思想，且他们的各类不变性、局部性有待进一步的实验解释。本文想要通过利用这些图像表征方法所保留的图像信息，来对表征进行直接分析，具体而言，是通过不同层的编码得到的表征来重建原本的图像信息，进而对表征进行分析。

1.2 贡献

- **本文提出了反转 SIFT、HOG 和 CNNs 表征的通用方法**。作者还对该算法中所提出的正则化惩罚进行了讨论评估。
- 通过实验表明对于 HOG 表征，**本文所提出的方法也能恢复明显更好的重建结果**。
- **本文将反转技术应用于 CNN 分析，通过采样尽可能近似重建来探索 CNN 的各类不变性**。
- 本文通过从选择的神经元输出的表征重建图像，**研究了信息在像素空间和通道空间的局部性**。

2 方法

如1所示，本文使用在 ImageNet 上训练的 AlexNet 预训练模型，分别对原始图、白噪声图（从均匀分布的随机数中独立采样得到的，具有平均亮度且没有结构或模式的图像）进行前向传播，对原始图每一层卷积滤波后得到的激活图（二维三通道特征图）进行可视化。反向传播时，以原始图和白噪声图的每一层的激活图之间的损失为目标，进行梯度下降算法，更新白噪声图像素，那么白噪声图

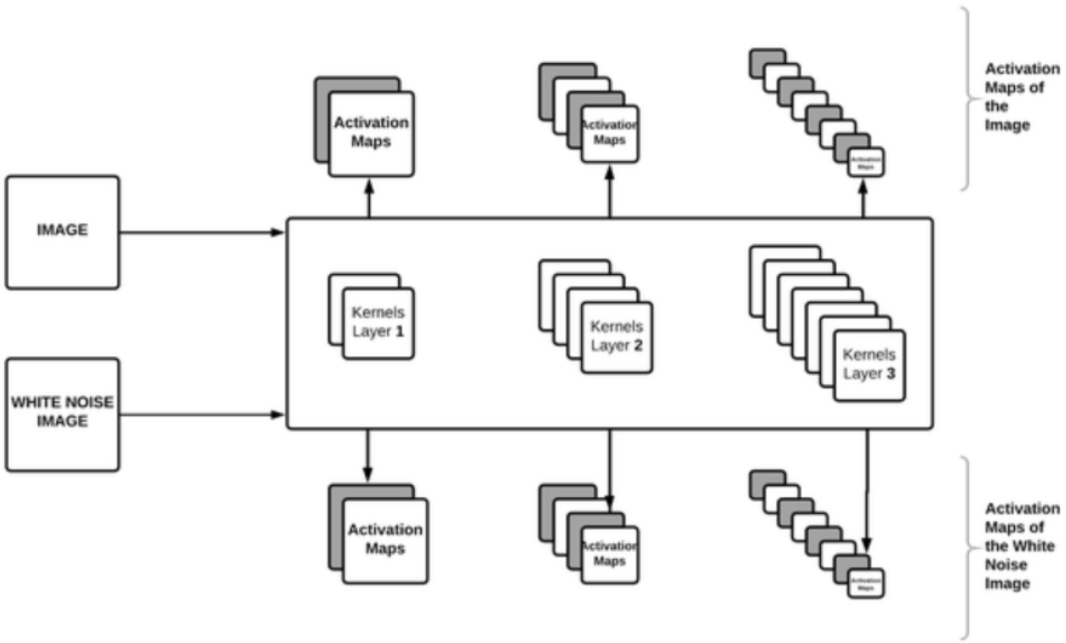

图 1: 方法示意图

最终会与原图类似。

作者可视化了每一个卷积层，提取了原始图和白噪声图的激活层，并反传激活层之间的损失。反传并不会更新卷积神经网络各层参数的权重，而是只对原始可学习的白色噪声图像进行更新，更新若干轮后白色噪声图应该与原始图至少在 CNN 模型的视角范围内相近。

2.1 算法

2.1.1 形式化定义

给定原始图 x_0，给定表征映射 $\Phi : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^d$ 和原始图的某个表征 $\Phi_0 = \Phi(x_0)$，我们希望通过重建一个可学习的图 $x \in \mathbb{R}^{H\times W\times C}$，并加上自然图像先验正则化项 $R(x)$，目标 $x^* = \underset{x\in\mathbb{R}^{H\times W\times C}}{\arg\min}\, l(\Phi(x), \Phi_0) + \lambda R(x)$，其中，$l$ 是将待更新的图 x 的表征与原始图的表征之间进行比较得到的损失函数。

2.1.2 Loss 函数设计

本文使用了表征之间的欧氏距离作为 Loss 函数：$l(\Phi(x), \Phi_0) = \|x\|_2$。

2.1.3 正则化器

使用正则化器的目的是为了解决判别式训练导致的地基图像统计信息丢失的问题，以便能够更完整地可视化所有层表征内容所必须的信息。通过引入图像先验，可以部分恢复丢失的低级图像统计信息。图像先验可以是对自然图像分布的建模，以及图像的结构、纹理、边缘等统计特性的假设。通过将图像先验融入重建过程中，可以帮助生成更加合理和自然的图像。

然而，直接对整个自然图像集合进行最小化操作是非常具有挑战性的，因为自然图像的分布非常复杂且高度多样化。因此，为了近似处理自然图像集合，可以使用适当的图像先验作为代理。这个图像先验可以是对自然图像的统计模型或分布的简化描述。

本文提出了两种正则化器。

第一种正则化器是 α-norm 的正则化器 $R_\alpha(x) = \|x\|_\alpha^\alpha$，其中 x 是向量化的或者平均减缩后的图像。通过相对较大的指数 $\alpha = 6$，可以使得图像的范围保持在目标区间内部而非发散。

第二种正则化器是全变分 (TV) 正则化器。$R_{V^\beta}(x)$，通过最小化全变分正则化项，我们鼓励图像中相邻像素的差异较小，从而使图像更平滑，并且具有分段常数的特性。最终达到了既能去除噪声，又能保留图像的边界等信息。具体而言，全变分正则化器在离散的图像中可以使用有限差分近似取代。

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

作者还讨论了，在存在子采样的情况下，当 $\beta = 1$ 时 TV 正则化器会导致子采样图像的重建尖峰问题。作者选择了 $\beta > 1$，以对大梯度进行惩罚，将变化分布到各个区域，而非集中在一个点或者曲线上。将之称为 V^β 正则化器。如图2所示，这些尖峰通过令 $\beta = 2$ 来去除。但是同时由于边缘受到更大的惩罚，图像也被洗掉了。

当重建目标是彩色图像时，需要对每一个颜色通道的正则化器进行求和。

2.1.4 正则化器和损失的平衡

第一步，对损失目标函数进行归一化：$\|\Phi(x) - \Phi_0\|_2^2/\|\Phi_0\|_2^2$ 这使得 Loss 的动态范围在归一化之后能够在 $[0, 1$ 区间内，在最优点达到 0。

虽然缩放图像范围对于 CNN 来说并不敏感，但是对于前几层 CNN 来说，不

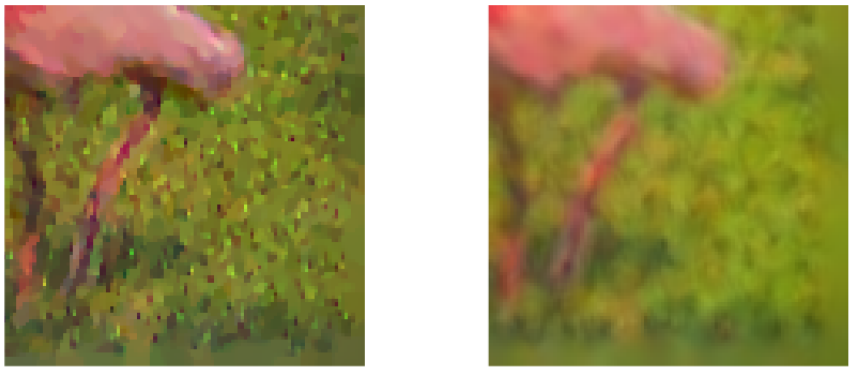

Figure 2. **Left:** Spikes in a inverse of norm1 features - detail shown. **Right:** Spikes removed by a V^β regulariser with $\beta = 2$.

图 2: 尖峰问题

同比例缩放图像得到的表征往往跟缩放的尺寸有关。其中偏差会被调整为一个自然的工作范围。这可以通过考虑目标 $\|\Phi(\sigma x) - \Phi_0\|_2^2 + \lambda_\alpha R_\alpha(x) + \lambda_{V^\beta} R_{V^\beta}(x)$ 来解决，其中缩放因子 σ 是训练数据集中自然图片的平均欧式里德范数。

第二步，应该选择 α 范数正则化器的乘数 λ_α 以鼓励重建图像 σx 包含在自然范围 [-B,B] 中。

故最终的目标函数形式是：$\|\Phi(\sigma x) - \Phi_0\|_2^2 / \|\Phi_0\|_2^2 + \lambda_\alpha R_\alpha(x) + \lambda_{V^\beta} R_{V^\beta}(x)$，由于表示函数 Φ 是非凸函数，故还需要讨论如何进行优化。

2.1.5 优化

上述最终目标函数是非凸的，简单的 Gradient Descent 程序被证明使得模型学习这些数据是非常有用的，故作者使用 GD 优化算法，并将其拓展到加入动量机制。

本文还使用了 CNN 卷积层来计算 HOG 和 DSIFT 等的导数。

2.2 图像表征函数

2.2.1 CNN-A：深度网络

本文使用的 CNN 即 AlexNet 的架构，输入图片的大小为 227。架构如图3所示。

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

Table 2. **CNN-A structure.** The table specifies the structure of CNN-A along with the receptive field size of each neuron. The filters in layers from 16 to 20 operate as "fully connected": given the standard image input size of 227×227 pixels, their support covers the whole image. Note also that their receptive field is larger than 227 pixels, but can be contained in the image domain due to padding.

图 3: AlexNet

2.2.2 CNN-DSIFT,CNN-HOG 结构

将 DSIFT 和 HOG 使用 CNN 滤波器来完成，目的是形式化 CNN 和 DSIFT、HOG 等标准表征之间的关联。并且使得这些表征的求导更加简单。

在这个部分，文章中的公式主要是使用的类似于 CNN 的 Filter 一样的矩阵计算像素值的差分近似代替 HOG 和 DSIFT 中的差分计算。

3 实验

3.1 浅层表征的实验

通过将表征逆转方法应用在 HOG 和 DSIFT 表征上进行了评估。

对于 HOG 表征逆转而言，作者介绍了一种近似的方法 HOGgle[3]作为比较。HOGgle 作为一种开源的预训练模型，将 HOG 特征的 UoCTTI 操作进行了逆转，在数值上与 CNN-HOG 是等价的，这也就使得本文可以直接做不同算法之间的对比。但是 HOGgle 方法在定性上比不过作者自己提出的 CNN-HOG，在定量上的错误率也比作者提出的逆转方法高。

修改 HOG 使得其与 SIFT 一样使用双线性梯度方向赋值，显著降低了重建误差，提高了重建质量。

作者还论证了，DSIFT 和使用了双线性梯度方向赋值的 HOG 方法的情况下，二者的重建损失相似，但是 DSIFT 的效果更好，这是因为 HOG 使用了更加精细的梯度量化，但是在相同的单元尺寸和采样下，HOG 的块归一化更重，显然比 SIFT 丢弃了更多的图像信息。

3.2 深层表征上的实验

这个部分的实验评估了应用于 CNN-A 这个深度 CNN 上的逆转算法。相比于 CNN-HOG 和 CNN-SIFT，这个网络更大、更深。

λ_{V^β}	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	conv1	relu1	pool1	norm1	conv2	relu2	pool2	norm2	conv3	relu3	conv4	relu4	conv5	relu5	pool5	fc6	relu6	fc7	relu7	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

Table 3. **Inversion error for CNN-A.** Average inversion percentage error (normalized) for all the layers of CNN-A and various amounts of V^β regularisation: $\lambda_1 = 0.5$, $\lambda_2 = 10\lambda_1$ and $\lambda_3 = 100\lambda_1$. In bold face are the error values corresponding to the regularizer that works best both qualitatively and quantitatively. The deviations specified in this table are the standard deviations of the errors and not the standard deviations of the mean error value.

图 4: CNN-A 损失

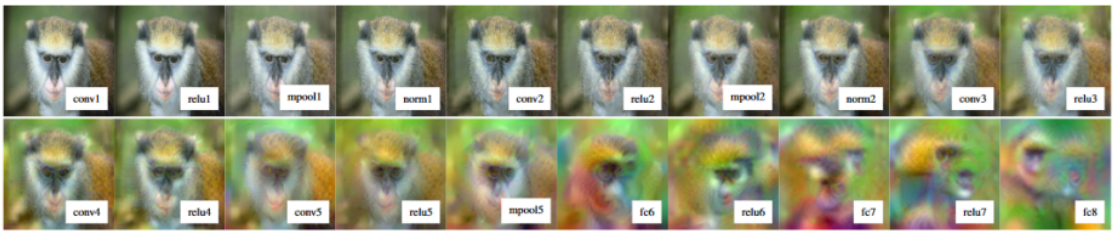

Figure 6. **CNN reconstruction.** Reconstruction of the image of Fig. 5.a from each layer of CNN-A. To generate these results, the regularization coefficient for each layer is chosen to match the highlighted rows in table 3. This figure is best viewed in color/screen.

图 5: CNN 各层的重建损失

在 CNN-A 上的各层表征的结论：在定量上看，深度 CNN 的各层表征损失与 CNN-HOG 的逆转损失相差不大，而最后一层的表征损失是非常小的，也容易逆转得到原始图。

作者基于对重建的定量和定性研究，分别选择了每个表示层的不同正则化系数 V_β。如图4，作者对这些正则化系数进行了比较分析，认为增加 V_β 会导致第一层的退化，但是对于后面的层，它有助于恢复更直观的可解释性重建。虽然这个系数可以通过对归一化的重建误差进行交叉验证来调整，但基于定性分析的选择是首选的，因为作者希望该方法能够产生视觉上有意义的图像。

作者从定性角度，如图5，对测试图片在 CNN 各层表征中进行重建，观察到几个现象：

1. 前几层实际上是图像的可逆码。

2. 所有的卷积层都保持了对图像的摄影忠实表示，尽管模糊程度越来越高。

3.4096 维的 fc6 全连接层的反演与原始图中发现的部分是相似但又不完全的组成。从 ReLU7 到 fc8，维度进一步降低到只有 1000。尽管降低维度够多，这些视觉元素中的一些仍然可以被识别。

由此，作者认为：

1. 从 CNN 模型的角度来看，所有的图像与原始图之间非常相似，几乎都不

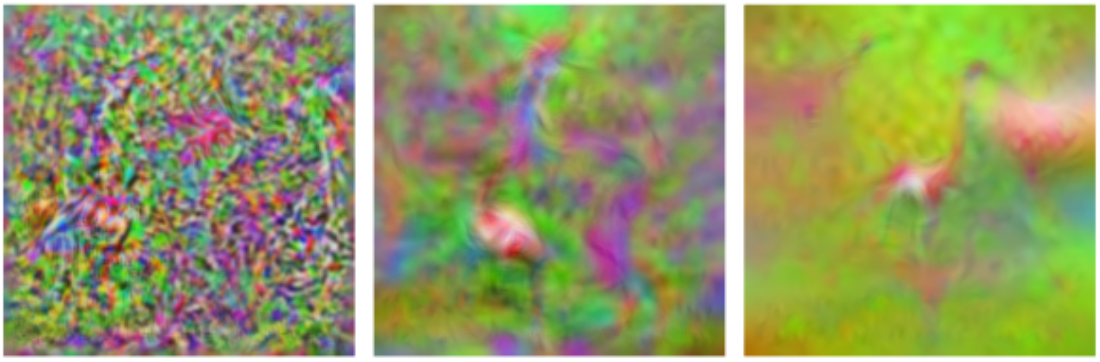

Figure 8. Effect of V^β regularization on CNNs. Inversions of the last layers of CNN-A for Fig. 5.d with a progressively larger regulariser λ_{V^β}. This image is best viewed in color/screen.

图 6: 正则化参数对比

可区分。深度 CNN 重建的结果缺乏细节，但深层的网络能够捕捉一些足以用于分类的物体草图。

2. 适当地降低正则化参数仍然可以得到非常精确的反演结果（见6中间）。但此时与自然图像几乎没有任何相似之处，这证实了 CNNs 具有较强的非自然混杂因素。

3. 一个有趣的现象是，如图7，许多反转的图片拥有大量的绿色区域，作者认为这是网络的一种属性，而非自然图像的先验。（先验的影响如图6所示，该先验仅鼓励平滑，因为它等价于惩罚重建图像的高频成分。）更重要的是，它同样适用于所有的颜色通道。

4. 局部性结论：如图8对每层的中央的 5*5 的特征片段进行反转，正则化器鼓励图像中对神经响应没有贡献的部分被关闭，特征的局部性在图形中表现明显；此外，神经元的有效感受野在某些情况下明显小于理论感受野。特征的局部性在图形中表现明显（比如 conv5、relu5 的周围像素不能有效反转出图像）。

5. 如图9，对特征通道的子集单独重建图片，结果表明，一组是对颜色信息进行调谐，另一组是对更尖锐的边缘和亮度分量进行调谐。值得注意的是，这种行为在学习到的网络中自发产生。

Figure 11. **Diversity in the CNN model.** mpool5 reconstructions show that the network retains rich information even at such deep levels. This figure is best viewed in color/screen (zoom in). More qualitative results are provided in the project web page.

图 7: 反转出现绿色

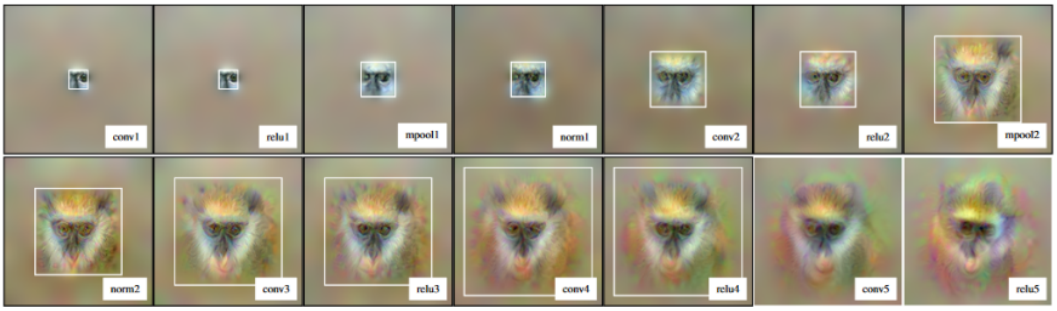

Figure 9. **CNN receptive field.** Reconstructions of the image of Fig. 5.a from the central 5×5 neuron fields at different depths of CNN-A. The white box marks the field of view of the 5×5 neuron field. The field of view is the entire image for conv5 and relu5.

图 8: 局部性探究

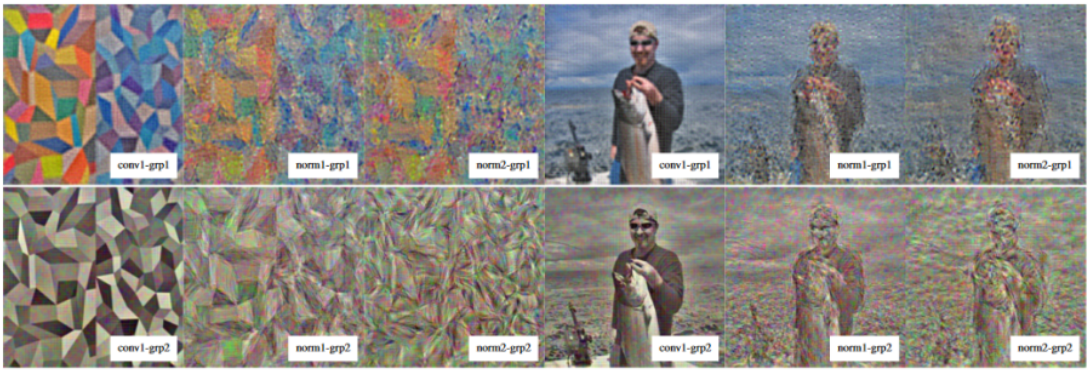

Figure 10. **CNN neural streams.** Reconstructions of the images of Fig. 5.c-b from either of the two neural streams of CNN-A. This figure is best seen in colour/screen.

图 9: 特定于通道的特征反转

4 总结

本文提出了一个翻转浅层、深层图像表征的优化方法，使用梯度下降求解最优值。相比于别的方法，理论的创新处是使用了图像先验的两个正则化器，这些正则化器是可以重建低级图像统计特征的，弥补了图像表征的缺陷。本文重点对于重建 CNN 深度表征的结果进行可视化，对每一层输出进行了量化和定性分析。本文还揭示了 CNN 的不变性、高层特征抽象性的实验结论。

参考文献

[1] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60: 91-110.

[2] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05): vol. 1. 2005: 886-893.

[3] VONDRICK C, KHOSLA A, MALISIEWICZ T, et al. Hoggles: Visualizing object detection features[C]//Proceedings of the IEEE International Conference on Computer Vision. 2013: 1-8.