Noise2Noise: Learning Image Restoration without Clean Data

Dezeming Family

2022 年 07 月 19 日

正常字体：表示论文的基本内容解释。
粗体：表示需要特别注意的内容。
红色字体：表示容易理解错误或者混淆的内容。
蓝色字体：表示额外增加的一些注释。
绿色字体：表示额外举的一些例子。

目录

abstract

我们通过机器学习将基本的统计推理应用于信号重建——学习将损坏的观测值映射到干净的信号——得出了一个简单而有力的结论：可以通过只查看损坏的示例、达到甚至超过使用干净数据的训练的性能来学习恢复图像，而无需明确的图像先验 (image priors) 或损坏 (corruption) 的可能性模型 (likelihood models)。在实践中，我们表明，单个模型仅基于噪声数据就可以来学习图像恢复，例如用于照片去噪、合成蒙特卡洛图像去噪和欠采样 MRI 扫描的重建——所有这些图像都被不同的过程破坏了。

一 Introduction

通过有噪图像和无噪图像作为 reference 之间的损失训练来得到去噪网络是常规方法。但是有些时候无噪图像很难获得，比如蒙特卡洛渲染需要大量采样来收敛到无噪声图像。

我们的方法直接从损坏图像来重建原始图像，我们既不需要损坏的显式统计似然模型，也不需要图像先验，而是从训练数据中间接学习这些。事实上，在我们的一个例子，合成蒙特卡洛渲染中，非平稳噪声不能被分析表征。

除了去噪之外，我们的观察结果还直接适用于逆问题，如从欠采样数据重建 MRI。虽然从统计学的角度来看，我们的结论几乎微不足道，但它通过提高对训练数据可用性的要求，大大简化了实际的学习信号重建。

总之，该文章提出的观点很有意思：在某些常见情况下，网络可以学习恢复信号而不用“看”到“干净”的信号，且得到的结果接近或相当于使用“干净”样本进行训练。而这项结论来自于一个简单的统计学上的观察：我们在网络训练中使用的损失函数，其仅仅要求目标信号（ground truth）在某些统计值上是“干净”的，而不需要每个目标信号都是“干净”的。

在研究该论文之前，一定要明白的是，这篇论文中仍然需要成对的数据，只是这些成对的数据中，可以都是有噪声的。比如，为了渲染一张无噪的图像，可以渲染一些有噪声的图像，然后这些图像可以散布在各个场景中，只需要保证 target 的零噪声条件，就可以训练出去噪网络。前提是数据量要足够多。而且必须要有数据对，也就是说，输入和输出都是同一个图像使用零期望噪声污染的图像（但是也未必，比如后面应用的蒙特卡洛渲染去噪，只不过【输入图像的期望】和【输出图像的期望】之前的像素关系一定是相关的）。

二 理论背景

在超分辨率算法中，由低分辨率到高分辨率图的对应是一对多的，也就是说，一张低分辨率图是可以对应多张高分辨率图的，网络直接使用 L_2 loss 去回归高分辨率的结果，实际上会倾向于回归可能对应的高分辨率图像的均值，因此预测的高分辨率图会倾向于模糊。

关键原理：让神经网络学习两张零均值噪声图的映射关系。样本数量少时，可以学习到两种零均值噪声模式的变换关系；样本数量多时，由于噪声的不可预测性，在最小化 loss 的角度，神经网络倾向于输出所有可能的输出的期望，也就是 clean 信号本身。

一句话总结 N2N：让网络强行学习从一个带噪图片到另一个带噪图片的映射。当网络有大量噪声到噪声的映射需要学习的时候，从 loss 最小化的角度来看，网络会输出所有的有噪声图像的均值。因为均值与各个目标图像的 loss 之和最小。由于假设噪声是 0 均值的，因此可以说网络学会了去噪。N2N 的缺点是需要大量的噪声图像对和噪声的 0 均值假设。

2 1 基本类比

假如对某个物理量进行多次测量，得到一系列有噪声的测量值 $(y_1, y_2, ...)$，那么一种估计真实值的方法是找一个数 z，使得它与这些测量值有最小的平均偏差，即优化：

$$\arg\min_z \mathbb{E}_y\{L(z, y)\} \tag{二.1}$$

对于 L_2 损失，其实 z 就是测量值的期望（算术平均值）；对于 L_1 损失，z 就是在测量值中值处取得；对于 L_0 损失，z 就是在测量值的众数中取得。

如果 $\hat{y}$ 的期望和 y 相同，那么对于 L_2loss 来说，求下式时得到的 z 结果不变：

$$\arg\min_z \mathbb{E}_y\{L(z,\hat{y})\} \tag{二.2}$$

2 2 推广

从统计学的角度来说，这些损失函数都是似然函数的负对数，而对它们优化的过程可以看做最大似然估计（大学概率论的点估计方法）。神经网络相当于对点估计推广，比如典型的有监督训练：

$$\arg\min_\theta \mathbb{E}_{x,y}\{L(f_\theta(x),y)\} \tag{二.3}$$

其中，θ 为网络参数，这个公式就是通过训练参数，以最小化误差。如果输入与输出是一对一的，那么这样其实并没有什么问题，就是有监督训练。但是对于一些应用，比如去噪和超分辨重建，因为有噪图像，以及低分辨图像，属于丢失了信息的图像，理论上很多原始图像都有可能通过丢失不同信息来得到有噪图。也就是说，

由于一张图对应了多个 target，所以上式相当于同时优化 x,y 两个变量。如果两个输入变量 x 之间相互独立，那么根据贝叶斯定理，就可以写为：

$$\arg\min_\theta \mathbb{E}_x\{\mathbb{E}_{y|x}\{L(f_\theta(x),y)\}\} \tag{二.4}$$

那么如何推广到神经网络呢：使用 L_2Loss 时，如果 $p(y|x)$ 替换成【条件期望相同的任意分布】，则训练得到的 θ 是不变的。$p(y|x)$ 表示确定 x 以后 y 的分布，有监督学习中，可以是一一对应的关系，但无监督学习中，可能一个 x（比如有噪图像）对应了多个 y（无噪图像）。也就是说，我们可以假设一大堆有噪的 $\hat{y}_i$，它们均值就是期望 y，那么此时的训练结果仍然 θ 不变：

$$\arg\min_\theta \sum_i L(f_\theta(\hat{x}_i),\hat{y}_i) \tag{二.5}$$

$\hat{x}_i$ 和 $\hat{y}_i$ 都是均值为 0 的噪声的分布（但没有必要是同一分布的）。且需要满足 $\mathbb{E}\{\hat{y}_i|\hat{x}_i\}=y_i$，这也是说明噪声期望为 0，也就是说对于给定的 $\hat{x}_i$，其对应的 $\hat{y}_i$ 的期望应该是我们的真实期望 y。

所以综上所述，这个论文的最大意义就是，比如我们要训练一个蒙特卡洛渲染去噪器，以前需要一大堆高噪声输入和一大堆使用几万 spp 近似收敛的 reference 输出，现在只需要一大堆高噪声图像对，就可以训练一个去噪网络了。

2 3 直观理解

优化时，如果 target 足够多，而且 target 的噪声满足 0 均值分布，那么 $y|x$ 的优化结果应该就是 clean 的 target。也就是说，由于样本非常多，但是网络学不到随机噪声映射到另一个随机噪声的映射。由于噪声的不可预测性，站在 loss 最小化的角度上，神经网络会倾向于输出所有可能的输出的期望，也就是干净的信号本身。

三 蒙特卡洛去噪实验的构建

使用的网络：一开始是 RED30，即 30 层层次残差网络。后来替换为了 shallower U-Net，训练比 RED30 快了 10 倍。

3 1 简单说明

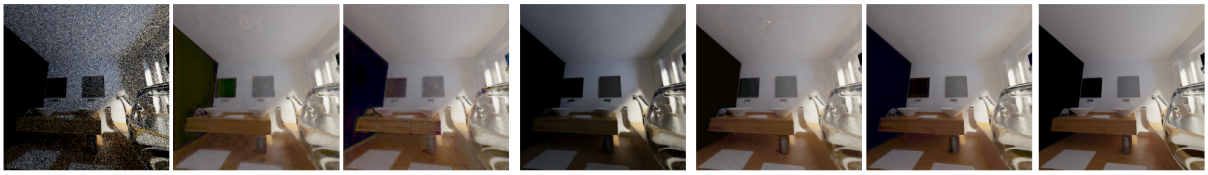
蒙特卡洛积分器被构造为使得每个像素的正确强度应当是随机路径采样过程的期望，即采样噪声是零均值。然而，尽管对重要性抽样技术进行了几十年的研究，但对分布情况却几乎没有其他说法。它因像素

而异，在很大程度上取决于场景配置和渲染参数，并且可以是任意的多模式。一些照明效果，如聚焦焦散，也会导致具有罕见明亮异常值的 long-tailed 分布。通过生成辅助信息的可能性在一定程度上缓解了该问题，该辅助信息已被经验发现与数据生成期间的干净结果相关。在我们的实验中，去噪器输入不仅包括每个像素的亮度值，还包括每个像素可见表面的平均反照率（即纹理颜色）和法向向量。

至于论文是怎么用的这些值，我个人认为是和输入的有噪图像合并，作为多通道图像。

3 2 损失函数

高动态范围在显示时需要色调映射，例如 $T(v) = (v/(1+v))^{1/2.2}$，但是对于噪声分布来说，我们必须保证的是目标图像的零噪声期望。而 $\mathbb{E}\{T(v)\} \neq T(\mathbb{E}\{v\})$，因此得到的预测结果就是错误的。



Input, 8 spp 11.32 dB | L_2 with $\hat{x}, \hat{y}$ 25.46 dB | L_2 with $T(\hat{x}), \hat{y}$ 25.39 dB | L_2 with $T(\hat{x}), T(\hat{y})$ 15.50 dB | L_{HDR} with $\hat{x}, \hat{y}$ 29.05 dB | L_{HDR} with $T(\hat{x}), \hat{y}$ 30.09 dB | Reference, 32k spp PSNR

Figure 6. Comparison of various loss functions for training a Monte Carlo denoiser with noisy target images rendered at 8 samples per pixel (spp). In this high-dynamic range setting, our custom relative loss L_{HDR} is clearly superior to L_2. Applying a non-linear tone map to the inputs is beneficial, while applying it to the target images skews the distribution of noise and leads to wrong, visibly too dark results.

将非线性色调映射应用于输入是有益的，而将其应用于目标图像会扭曲噪声的分布，并导致错误的、明显太暗的结果。关于非线性色调映射应用于输入是有益的，论文中并没有什么解释。需要注意的是，论文中是说 $\hat{x}_i$ 和 $\hat{y}_i$ 都是均值为 0 的噪声的分布，这里大概就是把 $\mathbb{E}\{T(v)\}$ 当成了神经网络的输入的期望，因为论文中说 $\hat{x}_i$ 和 $\hat{y}_i$ 的分布可以是不一致的，只需要保证噪声零期望即可。神经网络会把这种低范围到高范围的映射关系拟合到输出上。

参考文献

[1] Lehtinen J, Munkberg J, Hasselgren J, et al. Noise2Noise: Learning image restoration without clean data[J]. arXiv preprint arXiv:1803.04189, 2018.

[2] https://blog.csdn.net/u013066730/article/details/115305355

[3] https://www.cnblogs.com/aoru45/p/12250289.html

[4] https://blog.csdn.net/weixin_36474809/article/details/86535639

[5] https://zhuanlan.zhihu.com/p/563746026