Material recognition in the wild with the Materials in Context Database

摘要:

对现实生活图像的材质识别是一个很有挑战性的难题。这些材料有丰富的表面纹理，几何特征，光照等，这使这个问题尤其困难。近来，深度学习与大规模图像库的结合被证明是解决这一问题的有效方法。本文我们引入 MINC(material in context database)这一新的图像库。

我们把这一图像库与 CNN 结合起来解决两个问题：基于块的材质分类和整张图像同时进行的图像分割与材质识别。基于块的材质识别在我们的图像库中最好能达到 84%的正确率。我们把训练好的 CNN 分类器转化成一个与完全连接的 CRF(条件随机场)相结合的卷积模型来对图像进行逐像素的材质判断，达到了平均 85%的正确率。

1. 引言：

材质识别对我们理解现实世界有很重要的意义，比如说判断一个冰面是否可行走，什么样的手柄能抓起物体，我们必须识别他们的材质。材质的自动识别会在很多应用中发挥作用，包括机器人，产品搜索，室内设计的图像编辑等。但现实世界中的材质识别因各种因素是个很有挑战性的问题。很多材料的种类，如织物和木材，在现实生活中随处可见，且外观表现形式各不一样。由于光照，形状等因素同一材料也会有不同的外观。

最近，大型图像库 ImageNet, SUN, SUN Attributes 与 CNNs 方法的结合对目标识别与场景分类的研究推动很大。材质识别的研究很多使用中型数据库 FMD。FMD 引入了 10 种类别的材料，每种 100 张图片。这些图片是严格挑选的，使得物体的外观各种各样。FMD 已经被利用在图像新特征和材质识别上。但由于 FMD 的图像种类和每个种类的图像都相对较少，FMD 对于现实生活的材料分类还是不够。

我们的一个重要贡献是 MINC 图像库，有超过 200 万张材料图片。MINC 从 Flickr 图像库和一些专业摄影师的图像中取材，范围更广，比现存的图像库大。

在这个数据集的基础上我们训练一个多种类的 CNN 框架。我们做了许多网络模型，图像上下文，基于块的数据训练的实验来观察效果。然后，建立在基于块的分类结果的基础上我们对每张图像同时进行材质的识别和图像分割，通过将 CNN 分类器转化成一个与 CRF(条件随机场)相结合的卷积模型来实现，这种方法比简单的滑动窗口方法计算量小得多。

2. 相关工作：

图像库：

很多之前的材质识别工作集中于特定的纹理与材料。CUReT 图像库有 61 个材料样本，每个都是在 205 种不同光照与视觉条件下的。之后，图像数量更多，更有差异性的图像库比如 KTH-TIPS, KTH-TIPS2 出现了。在材料分类领域，Sharen 发布 FMD 图像库，它包含了 10 个种类的材料。OpenSurfaces 是一个标记了 20000 多场景的图像库。OpenSurfaces 有 27 个材料分类，每个类别有 100 到 20000 个分割的材料。这些图像来源于消费者拍摄，每个材料都存在于现实中，我们也使用 OpenSurfaces 作为我们的图像来源。

材质识别：

很多之前的材质识别工作都集中在分类问题上，即把图片分到特定的材料种类中，通常使用现有的特征。Ce Liu [Exploring Features in a Bayesian Framework for Material Recognition]用很多图像的低级，中级特征来得到材料外观表现出来的各种属性。他提出一种增强的LDA模型：用贝叶斯生成模型将这些特征结合起来，并通过学习调整使这些结合达到最佳。这些低级，中级特征包括color, jet, SIFT, micro-jet and micro-SIFT等。实验证明这一方法在FMD数据库上取得很好的识别效果。Hu [Toward robust material recognition for everyday objects]引入了基于方向梯度的差异性的特征. Qi [Pairwise rotation invariant co-

occurrence local binary pattern] 引入了一种共生LBP特征。Li [Recognizing materials from virtual examples] 合成了一个基于KTH-TIPS2的图像库，并基于LBP和密集SIFT(dense SIFT)特征构造分类器。Timofte [A training-free classification framework for textures, writers, and materials]用最小参数最优化构造了一种分类模型。 Cimpoi [Describing textures in the wild] 提出一种CNN和IFV(improved Fisher vector)来对FMD和KTH-TIPS2图片库进行分类。

卷积神经网络：

尽管CNNs已经诞生了几十年，但只在最近被应用到目标分类与探测上才取得巨大进步。受大规模视觉识别挑战的激发，已经有很多成功的CNN框架，除了图像分类，CNNs还适用于目标探测与定位。[Rich feature hierarchies for accurate object detection and semantic segmentation] 一文使用选择性搜索算法来对图像中的目标区域进行分类。[Overfeat: Integrated recognition, localization and detection using convolutional networks] 中使用滑动窗口在图像间共享卷积计算来提高预测的准确率。[Very deep convolutional networks for large-scale image recognition] 中使用overfeat方法，但将深度神经网络训练到19层。CNN也可运用到语义分割中，Farabet [Learning hierarchical features for scene labeling] 使用多尺度CNN和图像的过分割来预测图像的分类。

3. MINC 图像库：

我们需要新的图像库满足一下特征：数量大，预先分类标记好，多样性，种类多。

数据来源：

我们把OpenSurfaces图像库作为MINC的数据来源，它有34个分类，大概105000个标记的材料段。但OpenSurfaces也有一定的缺点：各个种类的图像数相差太大，比如说最大的木材类有20000多个样本，最小的比如说水，只有几十个样本。而且，这些图像来自单一的数据源：Flickr。因此，我们决定增大我们的图像库，我们从houzz.com上获取一些专业的图像，这样使我们的图像库更具有普遍性，多样性。

分割片段，点(clicks)和块：

对于材料，我们将解决两个问题：1，材质分类，给定一张图片的一小块，识别其材质。2，材料分割与识别，给定一张图片，生成一个全面的逐像素的分割与标记，也就是语义分割与场景解析。分类也可以与分割同时进行，比如滑动窗口方法。

分割片段：

OpenSurfaces里有材料的分割片段，但是，收集所有的材料片段会使自己的工作量太大，速度慢。为获得更多的训练数据，我们决定收集clicks：图像中有材料类标的单一的点。

Clicks：

首先，工作人员被展示一张图像并回答图像中是否有给定的材料出现，然后工作者将图像中出现的特定的材料点击出来，最后，通过询问其他人员验证是否标记正确。

我们在OpenSurfaces和houzz图像上进行这一操作，为避免某些种类取样不足，在采集clicks的最后阶段我们只采集一些稀有的种类(砖，头发，水等)。

我们采集clicks的初衷是获得更多的训练数据，但我们发现这些clicks对训练CNN分类器很有用，但对训练图像分割的CRF没多大用处因为它们没有边界信息，为获得更好的分割效果，我们在训练的时候调节(leverage)这两种数据。

块：

标记的材料片段和clicks就形成了MINC，对训练CNN和其他分类器，有标记的，固定的块很有作用。材料片段经过分割取样，clicks通过增加一个正方形区域都可以转化成块。

4. 材质识别：

首先，我们训练一个CNN，对一个输入的图像块能生成一个单一的预测，调整各个分类的差异训练到收敛。然后，我们将这个CNN转化成一个滑动窗口来对一张图像在密集的网格上进行材质识别，我们在多尺度的完成然后聚合生成一个一元的输出。最后，密集的CRF (conditional random field) 模型将2中的输

出与其完全两两连接的推理相结合生成逐像素的材质预测。

建立材料块 (patch)的数据集

我们的数据来源于OpenSurfaces和我们clicks的块。为训练一个CNN，我们需要把图像都转化成方块的图像块。我们用中心和大小来定义图像的子区域，通过调整块的大小来使分类的正确率达到最大，5.2节我们讨论了不同块大小的分类效果并发现AlexNet图像库的最佳块大小是最小图像维数的23%。

对于每个click，块中心是click的位置，因此我们从clicks中得到1685103个块。如果我们把块逐像素地放置，我们的计算量会很大。因此，我们把块中心选定为在以那点为中心的邻居中有标记的像素点。我们也把一些太暗或重复的块排除在外。我们打算把这些块按0.85,0.05,0.1比例分为训练集，有效集，测试集。为获得更好的一个与训练集不太相关的测试集，我们把图像按相邻-重复的块分类，然后把每一个块分配到训练集，有效集，测试集中的一个，并且舍弃一些重复的块。为探测邻居的重复的块，我们对图像提取出的特征进行比较。我们使用AlexNet fc6特征，并且当两个特征的点积大于0.9时才认为两张图片是相邻-重复的。

最后，“其他”类是一些未使用的分类中的图片，加入这个分类使我们的平均分类准确率提高了0.5%。

训练过程

我们用BVLC AlexNet模型来训练，测试我们的网络。用随机梯度下降方法来微调神经网络使其达到最佳。在训练过程中，合理调节输入的图像数很重要因为每个类有不同数量的图片。相比较于我们把每个patch按顺序输入到CNN中训练，我们随机抽取一个分类，从这个分类中随机抽取一个块输入到CNN中进行训练，这种方法使我们的准确率提高了大概7.3%。同时，为降低过拟合，我们随机增大样本，空间尺度从$[\frac{1}{\sqrt{2}},\sqrt{2}]$,图像高宽比[3:4,4:3]，增大系数[0.95,1.05]。最后，由于我们是基于局部区域的，需要减去一个平均值(R:124, G: 117, B: 104)。

材质识别与分割

接着我们把训练生成的CNN转化成一个滑动窗口探测器，并且在图像上进行密集的网格分类。特别地，我们最后的完全连接层替换成卷积层，这样我们的网络模型就成了一个完全卷积的。可以分类任意形状的图像。神经网络每32像素则输出一个预测结果，我们把滑动窗口的步伐设定为16像素。Sermanet [Overfeat: Integrated recognition, localization and detection using convolutional networks] 显示了这些卷积计算可以被重复利用。这样设置滑动窗口使我们在click位置的平均准确率提升了3.3%。

尽管我们的神经网络可以处理任意大小的输入，我们还是把输入图像转换成小于1100维的图像，这被之后证明是在AlexNet数据库上的最优参数。

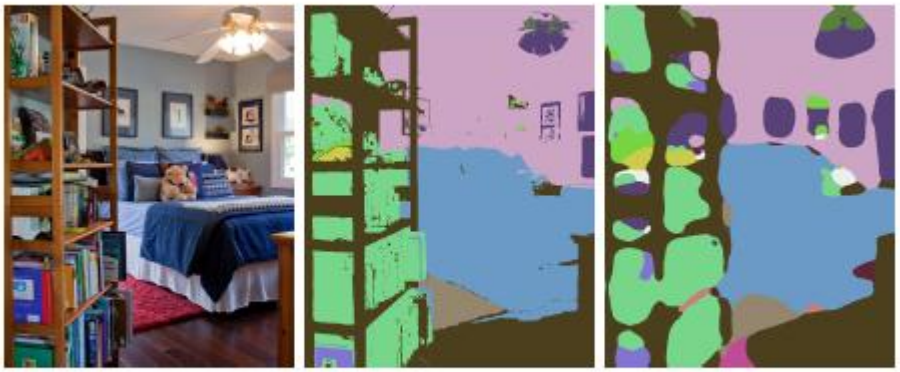
然后我们用密集的CRF来对每个像素进行一个预测：

$$E(x \mid \mathbf{I}) = \sum_i \psi_i(x_i) + \sum_{i<j} \psi_{ij}(x_i, x_j)$$

$$\psi_i(x_i) = -\log p_i(x_i)$$

$$\psi_{ij}(x_i, x_j) = w_p \, \delta(x_i \neq x_j) \, k(\mathbf{f}_i - \mathbf{f}_j)$$

其中，ψ_i是一元能量，ψ_{ij}是每两个像素点的共生概率，w_p 是权重系数，k是高斯核，δ 是Potts label compatibility term，对特征f_i ，我们把RGB图像转化成LAB图像，并用颜色和位置作为每个像素的共生特征。



Input image (test set) | (a) Labels from CRF trained on pixels | (b) Labels from CRF trained on clicks

图1：最优化dense CRF的分割效果

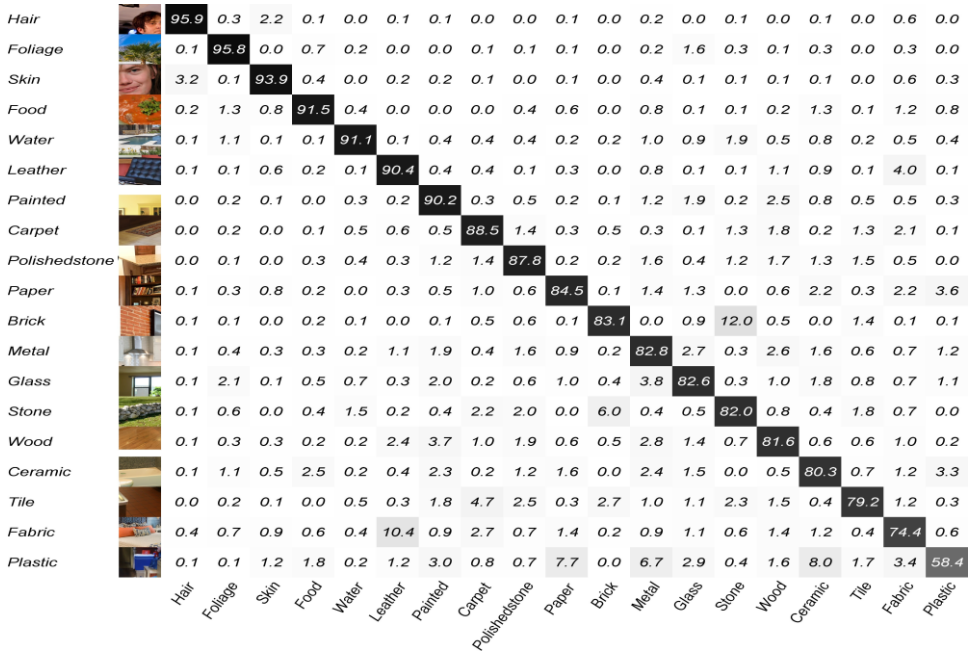


图2：基于块的在Alexnet数据库上的分类效果，平均84.9%的正确率