北京大学

《机器学习》应用实践案例调研
——基于卷积神经网络的黑白图片着色

姓名	王奕沣
学号	1800017709
院系	元培学院
教师	王继民
助教	王世奇

二〇二〇年十二月

摘要

本文调研了一种基于卷积神经网络的将黑白图片彩色化的算法。先前的方法要么依赖于用户的交互，要么导致了不饱和的色彩化。而作者们提出了一种全自动的方法，可以产生鲜艳而逼真的色彩。他们将问题的潜在不确定性作为一个分类任务来处理，并在训练时使用基于“模拟退火算法”的分类再平衡来增加结果中颜色的多样性。他们使用 "着色图灵测试 "来评估他们的算法，要求人类参与者在生成的彩色图像和真实彩图之间进行选择，他们的方法在 32%的试验中成功地愚弄了人类，明显高于之前的方法。

关键词：黑白图片；着色；卷积神经网络；模拟退火算法

Abstract

This article investigates a convolutional neural network-based algorithm for colorizing black and white images. Previous approaches either relied on user interaction or resulted in unsaturated colorization. Instead, the authors propose a fully automated method that produces vibrant and realistic colors. They treat the potential uncertainty of the problem as a classification task and use a "simulated annealing algorithm" based classification rebalancing during training to increase the diversity of colors in the results. They evaluated their algorithm using a "coloring Turing test", asking human participants to choose between the generated color image and the real color image, and their method successfully fooled humans in 32% of trials, significantly higher than previous methods.

Keywords: Black and white images; coloring; convolutional neural network; simulated annealing algorithm

目录

1.背景

摄影发展初期，由于技术的限制，只能制造出具有单层感光物质的胶卷，由此产生的是黑白图像。随着技术的发展和进步，彩色胶卷和数码传感器依次出现，人们得以得到精致的彩色照片。由此，对大量黑白老照片进行彩色化复原就成为了一个有吸引力的课题。

乍一看，这一目标似乎令人望而生畏，因为很多信息（三维色彩空间中的两个维度）已经丢失。然而，经过更仔细的观察，我们会发现，在许多情况下，场景本身的特性和它的表面纹理为每张图像中的许多区域提供了充分的提示，如草地是绿色、天空是蓝色、瓢虫是红色等。当然，这类直觉性的对应并不是对所有事物都有效。

然而，我们的目标不一定是要恢复实际颜色，而是要产生一个可能骗过人类观察者的可信的着色。因此，任务变得更容易实现：对黑白图像的内容和纹理及其各种颜色之间的统计依赖性进行足够的建模，以产生视觉上令人信服的结果。加州大学伯克利分校的几名研究人员在 2016 年开发了一个用于黑白图像彩色化的程序[1]。

1.1 色彩空间

色彩空间是指对色彩的组织方式。借助色彩空间和针对物理设备的测试，可以得到色彩的固定模拟和数字表示。色彩空间可以只通过任意挑选一些颜色来定义，比如像 Pantone 系统就只是把一组特定的颜色作为样本，然后给每个颜色定义名字和代码；也可以是基于严谨的数学定义，比如 Adobe RGB、sRGB。

最常见的色彩空间是 RGB 色彩空间，采用加法混色法。因为它是描述三种不同的颜色通过何种比例来产生其他颜色。可想而知，光线可以从纯黑开始不断叠加产生颜色，于是 RGB 描述的是红绿蓝三色光的数值。

然而，人眼实际上对于亮度更加敏感，由此出发开发出了 Lab 色彩空间。维度 L 表示亮度，a 和 b 表示两组对立的颜色。它的 L 分量密切匹配人类亮度感知。因此可以被用来做精确的颜色平衡（通过修改 a 和 b 分量的输出色阶），或使用 L 分量来调整亮度对比。这些变换在 RGB 色彩空间中是相当困难的。

1.2 基本思路

颜色预测问题有一个很好的优点，即训练集非常易得：任何彩色照片都可以作为

训练例子，只需将图像的 L 通道作为输入，将其 ab 通道作为监督信号即可。此外，原始黑白图片可以被当作 L 通道，于是模型在预测时就不必学习如何保持光强（使用 RGB 则必须），这样，模型只需学习怎样将图像彩色化，输出 ab 通道值即可。

2.数据集

互联网上图像数据的爆炸性增长有可能促进更复杂和强大的模型和算法，以便对图像和多媒体数据进行索引、检索、组织和互动。但究竟如何利用和组织这些数据，仍然是一个关键问题。斯坦福大学的李飞飞教授在 2009 年首次开发了名为“ImageNet”的新数据库[2]，这是一个建立在 WordNet 结构基础上的大规模图像数据库。ImageNet 的目标是用平均 500-1000 张干净的全分辨率图像来对应 WordNet 的大部分单词。这将产生数千万张按 WordNet 的语义层次结构组织的注释图像。ImageNet 在刚发布时具有 12 个子树，5247 个语义类，共计 320 万张图像，并且至今仍在不断扩充中。ImageNet 在规模和多样性上比先前的图像数据集要大得多，也准确得多。ImageNet 的规模、精度、多样性和层次结构可以为计算机视觉及其他领域的研究人员提供无与伦比的机会。

在调研涉及的工作中，作者在 ImageNet 训练集中的 1.3×10^6 张图像上训练神经网络，在 ImageNet 验证集中的 1 万张图像上进行验证，并在验证集中单独的 1 万张图像上进行测试。

3.机器学习算法

“卷积神经网络”这个名字表明，该网络采用了一种叫做卷积的数学运算。卷积网络是一种特殊类型的神经网络，它至少在其中一层使用卷积来代替一般的矩阵乘法。

3.1 神经元

神经元是人工神经网络的基本处理单元，一般是多输入单输出的单元，其结构模型如图 1所示。其中，x_i表示输入信号；n个输入信号同时输入神经元j。w_{ij}表示输入信号x_i与神经元j连接的权重值，b_j表示神经元的内部状态即偏置值，y_j为神经元的输出。输入与输出之间的对应关系可用下式表示：

$$y_j = f\left(b_j + \sum_{i=1}^{n} x_i w_{ij}\right)$$

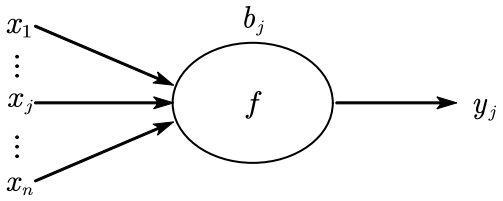

图 1 神经元模型

$f(\cdot)$为激励函数，可以是线性纠正函数、sigmoid 函数、tanh 函数等。

3.2 多层感知器

多层感知器 (Multilayer Perceptron，MLP) 是由输入层、隐含层 (一层或者多层) 及输出层构成的神经网络模型，它可以解决单层感知器不能解决的线性不可分问题。图 2是含有 2 个隐含层的多层感知器网络拓扑结构图。

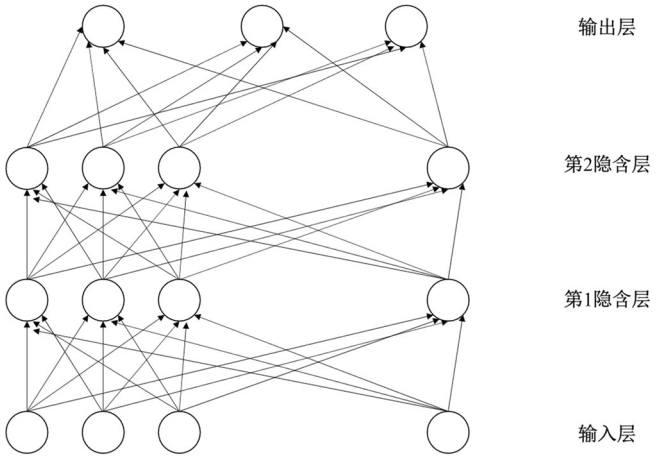

图 2 多层感知器结构图

输入层神经元接收输入信号，隐含层和输出层的每一个神经元与之相邻层的所有神经元连接，即全连接，同一层的神经元间不相连。图 2中，有箭头的线段表示神经元间的连接和信号传输的方向，且每个连接都有一个连接权值。隐含层和输出层中每一个神经元的输入为前一层所有神经元输出值的加权和。假设x_m^l是 MLP 中第l层第m个神经元的输入值，y_m^l和b_m^l分别为该神经元输出值和偏置值，w_{im}^{l-1}为该神经元与第$l-1$层第i个神经元的连接权值，则有：

$$x_m^l = b_m^l + \sum_{i=1}^{n} w_{im}^{l-1} y_i^{l-1}$$

$$y_m^l = f(x_m^l)$$

当多层感知器用于分类时，其输入神经元个数为输入信号的维数，输出神经元个数为类别数，隐含层个数及隐层神经元个数视具体情况而定。但在实际应用中，由于受到参数学习效率影响，一般使用不超过 3 层的浅层模型。相应的有监督学习算法可分为两个阶段：前向传播和后向传播，其后向传播始于 MLP 的输出层。以图 2为例，则欧氏损失函数为

$$E = \sum_{j}^{h} (y_j^l - t_j)^2$$

其中第l层为输出层，t_j为输出层第j个神经元的期望输出，对损失函数求一阶偏导，则网络权值更新公式为

$$w_{im}^{l-1} = w_{im}^{l-1} - \eta \times \frac{\partial E}{\partial w_{im}^{l-1}}$$

其中η为学习率。

3.3 卷积神经网络

典型的 CNN 由输入层、卷积层、池化层、全连接层和输出层构成。卷积层和池化层一般会取若干个，采用卷积层和池化层交替设置，即一个卷积层连接一个池化层，池化层后再连接一个卷积层，依此类推。由于卷积层中输出特征面的每个神经元与其输入进行局部连接，并通过对应的连接权值与局部输入进行加权求和再加上偏置值，得到该神经元输入值，该过程等同于卷积过程，CNN 也由此而得名。

每层卷积层由若干卷积核组成，每个卷积核的参数都是通过反向传播算法优化得到的。卷积运算的目的是提取输入的不同特征，第一层卷积层可能只能提取一些低级的特征如边缘、线条和角等层级，更多层的网络能从低级特征中迭代提取更复杂的特征。

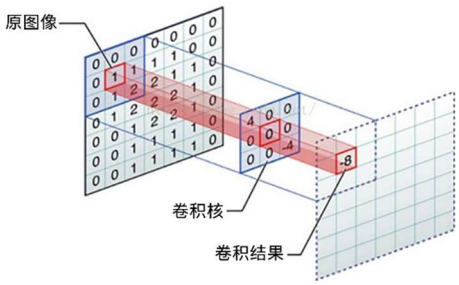

图 3 卷积层的基本结构

卷积层是池化层的输入层，池化层旨在通过降低特征面的分辨率来获得具有空间

不变性的特征，起到二次提取特征的作用，它的每个神经元对局部接受域进行池化操作，得到新的、维度较小的特征。常用的池化操作有最大池化、均值池化、高斯池化、可训练池化、随机池化等[3]。

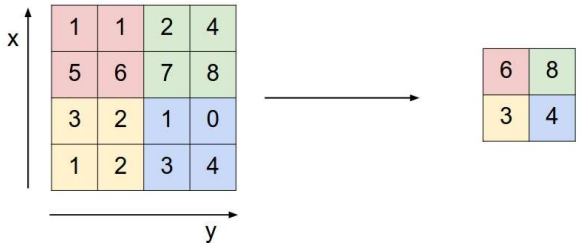

图 4 （最大）池化的过程

在 CNN 结构中，经多个卷积层和池化层后，连接着 1 个或 1 个以上的全连接层。与 MLP 类似，全连接层中的每个神经元与其前一层的所有神经元进行全连接，全连接层可以整合卷积层或者池化层中具有类别区分性的局部信息，最后一层全连接层的输出值被传递给一个输出层。

在先前的工作中，许多结果往往不令人满意，尤其是颜色不饱和现象较为突出，如图 5所示：

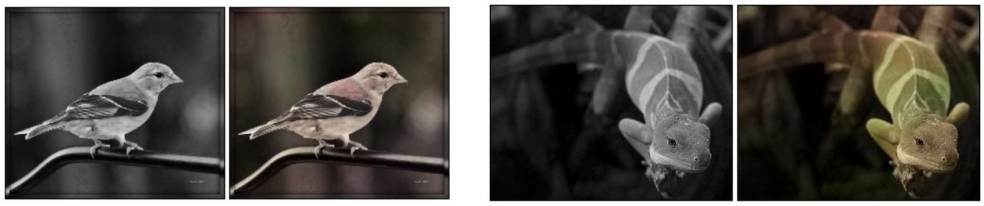

图 5 颜色不饱和的实例

作者提出的一个可能的解释是先前的工作使用常见基于欧几里得距离的损失函数，然而，这种损失函数对着色问题固有的模糊性和多模态性并不稳健。如果一个对象可以采取一组不同的 ab 值，那么欧氏损失函数的最优解将具有某些“平均值”特性。在颜色预测中，这种平均效应有利于灰度、去饱和的结果。此外，如果可信的着色集是非凸的，那么解实际上将不在集内，从而得到不合理的结果。于是，作者另辟蹊径，将着色视作多分类问题。将 ab 通道划分成 10×10 的格点，并选取 313 个 AB 对作为类的数量。对给定的输入$\mathbf{X}$，通过算法给出从原图像到各 ab 对概率分布的一个映射$\widehat{\mathbf{Z}} = \mathcal{G}(\mathbf{X})$，其中$\widehat{\mathbf{Z}} \in [0,1]^{H \times W \times Q}$，$H$和$W$是图像的长宽，$Q$是 ab 对的数量。使用的损失函数是标准的交叉熵：

$$\mathrm{L}_{cl}\left(\widehat{\mathbf{Z}},\mathbf{Z}\right)=-\sum_{h,w}v(\mathbf{Z}_{h,w})\sum_{q}\mathbf{Z}_{h,w,q}\log(\widehat{\mathbf{Z}}_{h,w,q})$$

其中$v(\cdot)$是加权项，可以根据颜色的稀有性进行加权。实际上，自然图像中的 ab 值往往较低，于是加权的因子应当与相应颜色所最接近的 ab 对相关：

$$v(\mathbf{Z}_{h,w})=\mathbf{w}_{q^*}$$

q^*代表“最接近”之意。对于权值的确定，作者提出了如下思路：

$$\mathbf{w}\propto\left((1-\lambda)\tilde{\mathbf{p}}+\frac{\lambda}{Q}\right)^{-1}$$

$$\mathbb{E}[\mathbf{w}]=\sum_{q}\tilde{\mathbf{p}}_{q}\mathbf{w}_{q}=1$$

其中$\tilde{p}$由真实图像采样并平滑化得到，$\lambda\in[0,1]$为加权系数（最终发现$\lambda=\frac{1}{2}$效果较好）。由于经过上述操作后，所得到的颜色只可能是选区的 313 个 ab 对之一，于是为了 ab 过度分立化带来的粗糙性，定义了$\mathcal{H}$，将$\widehat{\mathbf{Z}}$映射到 ab 空间的颜色值$\widehat{\mathbf{Y}}$：

$$\mathcal{H}(\mathbf{Z}_{h,w})=\mathbb{E}[f_T(\mathbf{Z}_{h,w})],\ f_T(\mathbf{z})=\frac{\exp\left(\log\left(\mathbf{z}\right)/T\right)}{\sum_q\exp\left(\log\left(\mathbf{z}_q\right)/T\right)}$$

这样的构造来源于所谓的“模拟退火算法”[4]，T是所谓的“特征温度”，这样做的目的是利用整个输出向量，而非单纯保留概率最大者，发现$T=0.38$效果较好。

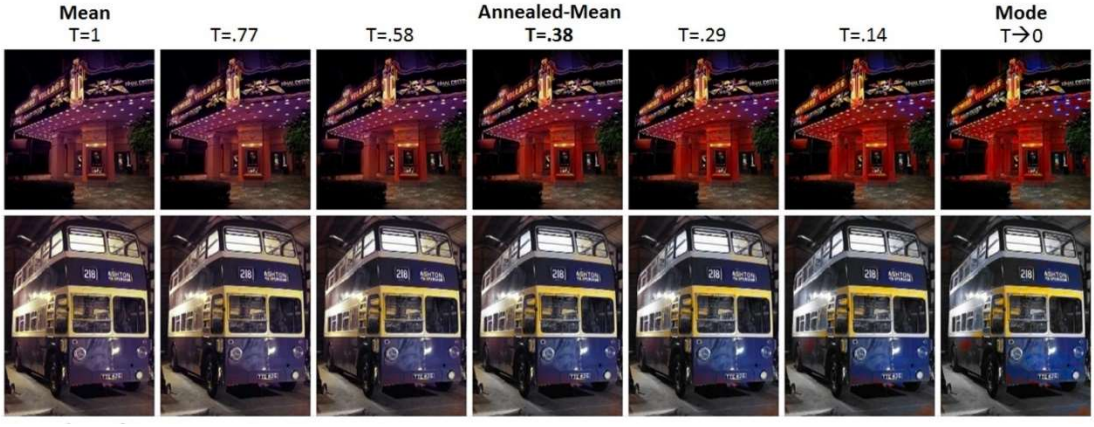

图 6 不同T下的结果

4.主要结果

虽然黑白图像着色是一项有重大意义的计算机图形学任务，但它也是计算机视觉

中一个困难的预测问题的实例。作者经过一系列对比已经证明，使用深度 CNN 和一个精心选择的损失函数进行着色，可以产生与真实彩色照片几乎难以区分的结果，一些结果如所示：

图 7 该模型的一些结果

这样的方法不仅提供了一个有用的图形输出，而且还可以被看作是表征学习的前置任务。虽然只针对颜色进行训练，但我们的网络学习的表征对于物体分类、检测和分割都是出乎意料的有用，与其他自我监督的预训练方法相比，表现十分出色。

5.小结

总之，几位作者设计出了一种高效的基于卷积神经网络的黑白图像上色技术。他们的贡献体现在两个方面：首先，他们在自动图像着色的图形问题上取得了进展：设计了一个合适的目标函数，处理了着色问题的多模态不确定性，并捕获了广泛的颜色多样性；引入了一个新的框架来测试着色算法，可能适用于其他图像合成任务；通过在 100 万张彩色照片上进行训练，为该任务设定了一个新的高标杆。其次，他们将着色任务转化为一种具有竞争力的、直接的自我监督表示学习方法，在几个基准上取得了最先进的结果。不过需要注意的是，在实际测试中发现该模型对动物（尤其是猫）的黑白照片预测效果较差，且当图像较为复杂时颜色也会失真，相信今后会有更多更强大、高效的算法出现。

6.参考文献

[1] ZHANG R, ISOLA P, EFROS A A. Colorful image colorization; proceedings of

the European conference on computer vision, F, 2016 [C]. Springer.

[2] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database; proceedings of the 2009 IEEE conference on computer vision and pattern recognition, F, 2009 [C]. Ieee.

[3] BOUREAU Y-L, LE ROUX N, BACH F, et al. Ask the locals: multi-way local pooling for image recognition; proceedings of the 2011 International Conference on Computer Vision, F, 2011 [C]. IEEE.

[4] KIRKPATRICK S, GELATT C D, VECCHI M P J S. Optimization by simulated annealing [J]. 1983, 220(4598): 671-80.