论文笔记1：Multi-label Image Recognition by Recurrently Discovering Attentional Regions


guanghuixu (/u/81fe1835dd00) + 关注
2018.01.18 21:17* 字数 2955 阅读 219 评论 2 喜欢 0
(/u/81fe1835dd00)

以下是笔者对这篇论文的粗略理解，不代表文章作者观点。
本文更多是本着学习交流（复现结果）的态度去撰写，所以并非直译，只摘取重要观点。
如果在阅读过程中觉得哪里写得不好，欢迎在下方评论指出或者私信交流。谢谢。

论文地址 (https://link.jianshu.com?t=https%3A%2F%2Farxiv.org%2Fabs%2F1711.02816%3Fcontext%3Dcs.CV)

摘要

目前解决多标签图片识别问题的方法大多依赖于额外的处理提取假设的区域（region proposals），导致多余的计算且达不到很好的效果。我们提出的多标签识别模型“记忆--注意”模块使得具有可解释性且符合上下文语境。这个模块包括两个部分：一个空间转换层，从卷积提取后的响应图中定位“注意区域”；另一部分是LSTM子网络，通过捕捉第一部分定位区域的全局依赖关系预测该区域的类别（分数）。

1. 介绍

多标签识别任务可以理解为多个单标签识别任务的集合，需要解决的困难包括：巨大的类间差距、视角、尺度变换、遮挡以及光照。同时，多标签识别任务还需准确预测出图片中存在的所有类别，这就要求模型对图片有更深的了解（如通过结合语义标签和区域去捕捉相互依赖关系）。
目前主流的多标签识别任务算法一般包括两步：

1. 产生一系列足够多的假设区域，保证所有的前景目标都在区域内。可以采取“自底向上”的策略，从图片关键特征出发；也可以通过训练一个额外的检测器。
2. 通过分类器或者神经网络的方式预测这些假设区域的类别得分

很明显，提取大量的假设区域需要巨大的计算开销，而这些假设区域中大多数又是无效的。同时，过于简单化理解目标间的依赖关系，模型在复杂场景中失效。
Wang等人在论文《A unified framework for mutil-label image classification》中采用CNN-RNN模型共同描述语义类别依赖关系和图片类别关系。然而，他们的模型中忽视了语义标签与图片内容间清晰的对应关系，同时也缺乏挖掘图片深层信息的能力。
对比之下，我们提出来的模型可以清楚地发现“注意区域”的响应是由多语义标签决定，同时基于全局视角捕捉这些区域的语义关系。
我们提出的方法主要有3点贡献：

1. 我们提出一个主流的“proposal-free”方法，能通过图片的不同比例自动发现具有语义知识的区域，同时捕捉他们的长期语境依赖关系。
2. 我们进一步提出在“空间转换层”加了三个约束（后文会详细介绍），确保学习到更多有意义的、可解释的区域，从而提高多标签分类能力。
3. 通过在VOC、COCO数据集上大量实验，证明了我们方法的精度与效率都比其他多标签分类方法要好。

2. 相关工作

目前图片分类任务取得巨大成功是因为大量被标注的数据集以及深度卷积神经网络的快

速发展。

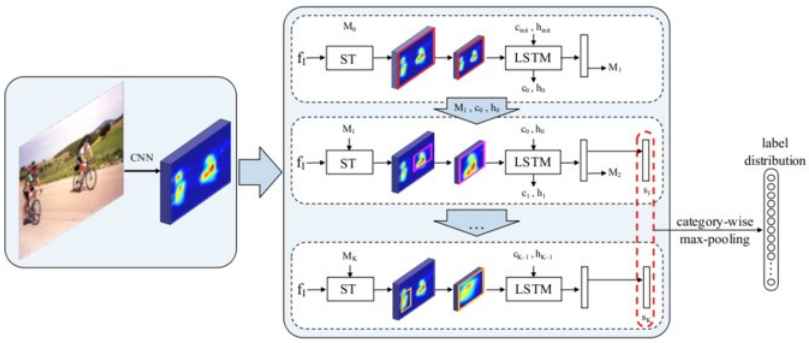
2.1 多标签图片识别

人工设计特征时期，大多采用**“词袋模型”**解决多标签识别任务。但随着复杂场景的出现及手工设计特征过于繁琐且不稳定，深度卷积神经网络提取的深层特征泛化能力强，逐渐取代“词袋模型”。为了更好地考虑标签间的相互关系，不再独立看待每个标签，许多基于深度卷积神经网络的方法还是包含了很多传统的形象化模型，如条件随机场、依赖网络、共生矩阵。
目前大多数方法都对全图提取特征，但缺乏空间转换操作。这样的处理方法，一方面不能清楚地反应区域与类别标签的对应关系；另一方面，极其容易受复杂背景的影响。同时，主流的多标签识别算法的瓶颈在预处理的“建议区域”生成过程。

2.2 可视化注意力模型

Jaderberg等人在论文《Spatial transformer networks》中提出了一个不同空间转换模块用于提取注意力区域，可克服图片比例、旋转、变换、复制等各种问题。此外，这个模型容易集成到神经网络中，基于标准的梯度反向传播算法进行优化，不需要借助强化学习技术。

3. 模型



模型结构示意图

模型结构：CNN+ST+LSTM

1. 图片经过深度卷积神经网络（如VGG-16）提取最后一层特征的响应图（如conv5_3），记为f_i
2. ST（spatial transformer，空间变换层）在每次循环中定位一个“注意区域”
3. LSTM预测该区域的类别分数，同时更新ST的参数
4. 集合所有循环产生的类别分数，得到最终的类别分布。

3.1 ST——定位注意力区域

ST（spatial transformer，空间变换层）是一个基于样本差异进行计算的模块，通过给定的尺寸，提取输入响应图的子区域作为变换后的区域。容易集成到神经网络中，通过标准反向转播算法即可实现优化。
在论文《Spatial transformer networks》中，用公式证明了

进行复制、平移、缩放操作，只需要一个映射矩阵M[2，3]
一个有趣的例子：

$$\begin{bmatrix}1 & 0 & \theta_{13}\\ 0 & 1 & \theta_{23}\end{bmatrix}\begin{bmatrix}x\\ y\\ 1\end{bmatrix}=\begin{bmatrix}x+\theta_{13}\\ y+\theta_{23}\end{bmatrix}$$

平移

$$\begin{bmatrix}\theta_{11} & 0 & 0\\ 0 & \theta_{22} & 0\end{bmatrix}\begin{bmatrix}x\\ y\\ 1\end{bmatrix}=\begin{bmatrix}\theta_{11}x\\ \theta_{22}\end{bmatrix}$$

缩放

在本文中，ST层的工作是从特征图f_l中提取注意区域（f_k），方法是由神经网络学习的映射矩阵M

$$M=\begin{bmatrix}s_x & 0 & t_x\\ 0 & s_y & t_y\end{bmatrix}$$

s_x、s_y：缩放参数
t_x、t_y：平移参数

3.2 循环“记忆——注意”模块

这是我们模型的核心模块，包括一个LSTM网络和一个空间变换层。在循环过程中，寻找最具判别力区域，并预测这些区域的类别分布。
对于第K次循环，ST层从f_l中利用M_k提取注意力区域的过程如下：

$$f_k=st(f_I,M_k),\quad M_k=\begin{bmatrix}s_x^k & 0 & t_x^k\\ 0 & s_y^k & t_y^k\end{bmatrix}$$

LSTM以f_k为输入，更新隐藏状态（hidden state）及记忆单元（memory cell）：

$$\begin{aligned}\mathbf{x}_k&=\mathrm{relu}(\mathbf{W}_{fx}\mathbf{f}_k+\mathbf{b}_x)\\ \mathbf{i}_k&=\sigma(\mathbf{W}_{xi}\mathbf{x}_k+\mathbf{W}_{hi}\mathbf{h}_{k-1}+\mathbf{b}_i)\\ \mathbf{g}_k&=\sigma(\mathbf{W}_{xg}\mathbf{x}_k+\mathbf{W}_{hg}\mathbf{h}_{k-1}+\mathbf{b}_g)\\ \mathbf{o}_k&=\sigma(\mathbf{W}_{xo}\mathbf{x}_k+\mathbf{W}_{ho}\mathbf{h}_{k-1}+\mathbf{b}_o)\\ \mathbf{m}_k&=\tanh(\mathbf{W}_{xm}\mathbf{x}_k+\mathbf{W}_{hm}\mathbf{h}_{k-1}+\mathbf{b}_m)\\ \mathbf{c}_k&=\mathbf{g}_k\odot\mathbf{c}_{k-1}+\mathbf{i}_k\odot\mathbf{m}_k\\ \mathbf{h}_k&=\mathbf{o}_k\odot\mathbf{c}_k\end{aligned}$$

image.png

其中relu、σ，tanh是三种常见的激活函数，**W**、b表示需要学习的权重和偏置，i_k、g_k、o_k、m_k、c_k、h_k分别表示输入门、遗忘门、输出门、输入调节、记忆单元和隐藏状态。本模型中采用常用的LSTM结构，不再赘述。有需要了解的读者可以参考链接 (https://www.jianshu.com/p/9dc9f41f0b29)。在理想状态下，记忆单元c_k记录（编码）了前k-1个时刻的有用信息。

1. **M的更新:**

$$\begin{aligned}\mathbf{z}_k &= \text{relu}(\mathbf{W}_{hz}\mathbf{h}_k + \mathbf{b}_z) \\ \mathbf{s}_k &= \mathbf{W}_{zs}\mathbf{z}_k + \mathbf{b}_s, k \neq 0 \\ \mathbf{M}_{k+1} &= \mathbf{W}_{zm}\mathbf{z}_k + \mathbf{b}_m\end{aligned}$$

   其中s_k表示预测的类别分数（score）分布

2. **类别预测**

   对于第K+1次循环，我们可以得到K个分数向量{s_1、s_2、s_3……s_k}；而s_k={s^1_k、s^2_k、s^3_k、……s^C_k}

   则对于第c类，有

$$s^c = max\left(s_1^c, s_2^c, s_3^c, \cdots s_K^c\right), \quad c = 1, 2, 3 \cdots C$$

4. 训练过程

4.1 分类损失

根据类别分数s^c，利用softmax函数进行归一化操作，可以得到各类别的预测概率：

$$p_i^c = \frac{\exp(s_i^c)}{\sum_{c'=1}^{C}\exp(s_i^{c'})} \; c = 1, 2, \ldots, C$$

对于每张图片 $\mathbf{x}_i$，Ground-truth $\mathbf{y}_i$ 是一个one-hot向量，y^c_i=1表示原图 $\mathbf{x}_i$ 中存在目标c；$||\mathbf{y}_i||_1$表示求向量元素绝对值之和，对于one-hot向量，即求$\mathbf{y}_i$不为0的个数。分类损失函数如下：

$$\begin{aligned}\hat{p}_i &= y_i / \left\|y_i\right\|_1 \\ \iota_{cls} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\left(p_i - \hat{p}_i\right)^2\end{aligned}$$

4.2 注意力区域约束

通过大量实验，单纯定义分类损失我们的模型也能取得一定效果，但会产生3个问题：

1. **冗余**：在每次循环中，ST层会倾向于重复提取图片中的最显著区域，这就无法保证我们的模型能够识别出全部的目标
2. **忽略微小目标**：这跟第一个问题其实是一致的，ST层会倾向于提取最显著区域（大目标）而忽略小目标
3. **空间跳跃**（Spatial flipping）：The selected attentional region may be mirrored vertically or horizontally。（这里没理解过来）

为了解决上诉问题，我们定义了关于ST层中变换矩阵**M**的约束条件：

$$M = \begin{bmatrix} s_x & 0 & t_x \end{bmatrix}$$

$$M = \begin{bmatrix} 0 & s_y & t_y \end{bmatrix}$$

1. 锚点约束

   我们认为，添加锚点约束有利于注意力区域分散到各个语义标签上，避免了只重复提取最显著区域的问题。以图片中心为圆心，1/√2(根号2分之一)为半径画圆，锚点均匀分布，个数由产生的前K-1个注意力区域决定。

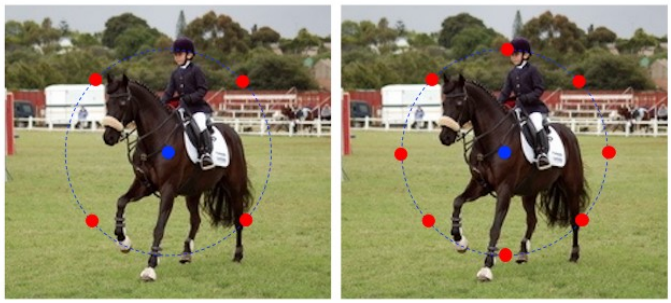


Figure 4. Anchor selection for left: K=5 and right: K=9.

锚点示意图

锚点约束公式如下：

$$\ell_A = \frac{1}{2}\{(t_x^k - c_x^k)^2 + (t_y^k - c_y^k)^2\},$$

其中(c^k_x , c^k_y)是第k个锚点的位置

关于这个约束条件的解释：

所有锚点会平均分布在圆上，所以这个约束条件并不是对第K个锚点进行约束，而是通过第K个锚点的位置使得平移参数（t_x,t_y）必须改变，从而避免了模型总是选择图片中最显著区域的情况。

2. 缩放约束

   动机是不希望注意力区域太大。

$$\ell_S = \ell_{s_x} + \ell_{s_y},$$

$$\ell_{s_x} = (\max(|s_x| - \alpha, 0))^2$$
$$\ell_{s_y} = (\max(|s_y| - \alpha, 0))^2$$

   α是一个阈值（0.5）

3. 正值约束

   此部分也是关于缩放参数的约束，也就是s_x、s_y：

$$\ell_P = \max(0, \beta - s_x) + \max(0, \beta - s_y),$$

   β也是一个阈值（0.1）

   第2和第3部分讲的是一个约束条件：

$$0.1 \leqslant s_x, s_y \leqslant 0.5$$

则关于定位误差，有：

$$\mathcal{L}_{\text{loc}} = \ell_S + \lambda_1 \ell_A + \lambda_2 \ell_P,$$

其中λ_1、λ_2是权重参数，分别设置为0.01,0.1

模型中的损失函数为：

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \gamma \mathcal{L}_{\text{loc}}.$$

其中γ=0.1

5. 实验

5.1 设置

----------**实现细节**----------

1. 采用Caffe框架，在ImageNet上预训练深度神经网络
2. 新加入的层采用Xavier初始化
3. 训练两个模型，所有图片resize到N×N大小，N分别取512,640
4. batch_size=16，momentum=0.9、0.999，
5. 初始学习速率为$1e^{-5}$，每训练30个epoch就除于10
6. 每过45个epoch，进行test数据集的测试
7. 每张resize过后的图片，分别在四个角、中心提取5个（N-64）×（N-64）区域块（patch）来增大数据集。但不应该在原图上直接提取，这样重叠区域会出现计算开销的浪费，技巧是在最后一层卷积响应图上提取对应区域。
8. 对每个区域块采用max-pooling作为该区域块的特征，将5个区域块以ave-pooling的方式可以作为原图的特征表达

讨论--来自灵魂的拷问

1. 解决了什么问题?
   多标签识别问题
2. 提出了怎样的方法?
   从全局视角提取注意力区域，再对注意力区域进行类别打分
3. 为什么会提出这个方法?
   现有的多标签识别算法，都需要额外的区域提取层：要么“自底向上”，从图片底层特征出发，提取一堆可能的区域；要么训练一个额外的检测器去检测是否有目标。这样的处理方法低效、计算开销浪费
4. 设计了怎样的模型?
   CNN+ST+LSTM
5. 建立在什么样的假设上?
   LSTM编码记录了所有有用的历史信息
6. 这篇文章的关键点是什么?
   ST+LSTM：ST提取区域，LSTM对区域打分并更新ST的参数
   LSTM所编码记录的信息需要支撑ST层提取正确的注意力区域
7. 贡献是什么?
   开启了新的多标签识别思路，采用“自顶而下”的策略，从全局角度提取注意力区域。
   (不需要总是训练额外的检测器去提取区域)
8. 如何通过这个模型验证了作者所提出的观点?
   我们认为，LSTM记忆单元能使模型从全局视角作出判断（参数更新和区域打分）
9. 这个模型为什么能work?
   所有模块采用标准的反向传播算法就能够进行参数更新

赞赏支持

论文笔记 (/nb/14273349)　　举报文章　© 著作权归作者所有

guanghuixu (/u/81fe1835dd00)
写了 11221 字，被 7 人关注，获得了 4 个喜欢
(/u/81fe1835dd00)
+关注

喜欢　更多分享
(http://cwb.assets.jianshu.io/notes/images/2243608

下载简书 App ▸
随时随地发现和创作内容
(/apps/download?utm_source=nbc)

登录 (/sign后发表评论source=desktop&utm_medium=not-signed-in-comment-form)

2条评论　只看作者　　按喜欢排序　按时间正序　按时间倒序

Helloworld_98f6 (/u/288786ab104a)
2楼 · 2018.03.17 14:26
(/u/288786ab104a)
请问一下有这篇顶会论文的源码下载地址吗?

赞　回复

guanghuixu (/u/81fe1835dd00)： 作者并没有公布源码
2018.03.19 16:29　回复

添加新评论

推荐阅读　　更多精彩内容 › (/)

机器学习库的维护方法 (/p/0410ed3dc166?utm_campaign=maleskine&ut…

1. pip install xxx --user 使用pip install时遇到权限不足时不要随便用sudo这样是方便省事，但没有管理员权限的根本无法访问（这个问题在大实验室里尤为常见，谁有事没事会给实验小白管理员权限呀）可以用下面的代码替代 2. 软件安装/卸载/安装目录查看方法 apt-get conda pip 3. deb包依赖关系没有安装

guanghuixu (/u/81fe1835dd00?
utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

End-to-End Instance Segmentation with Recurrent…

第一版 这是对End-to-End Instance Segmentation with Recurrent Attention这篇paper的初步理解，由于实习跟课程设计的关系，时间有点赶，可能有些理解不到位的地方，以后有时间会持续更新。引用部分是笔者对论文重要内容的英文摘录和翻译（意译），正文部分是笔者的理解及相关知识的补充说明。 这篇paper中作者提出了一个端对端的递归注意力网络(Recurrent Attention)架构，类似于人脑计数机制的原理，用于解决目标分割的问题。这个模型能检测出每个区域中最显著的目标并实现分割这里的翻译可能不太妥当，“human-like count...

(/p/4296d3abe1ef?
utm_campaign=maleskine&utm_content=note&utm

guanghuixu (/u/81fe1835dd00?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

推荐：有什么我用很久都不换的app(下篇) (/p/6c51c88…

(/p/6c51c882ca6e?utm_campaign=maleskine&utm_content=note&utm

修图篇 泼辣修图:手机版的LR，调整更加精细一点，一般我会在用完滤镜后要进行一些微调的时候，经常用到这个app，主要是拉高曝光，提高阴影让照片更通透，拉曲线调成胶片感，我很喜欢用这款app加锐化，会特别有质感，泼辣修图还有面部调节功能，基本上是一款很全能的app，推荐~ InstaSize:如何让自己的照片level提升好几个档次，那就是加白边！这款app是我加白边、拼图最最最爱用的app，很多模版可以选择，滤镜也很实用，还能增加文字(虽然这功能我不常用)，现在还有个更强大的功能就是给视频加白边，以下是我在成都旅游原图拍完用InstaSize加白边的效果图，总之这款app五星级推荐~ J...

那抹绿view (/u/0380c78287eb?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

再见了，我的五线小城市！ (/p/f8035127ebba?utm_c…

(/p/f8035127ebba?utm_campaign=maleskine&utm_content=note&utm

今天我的选择或许哪天也会成为你的选择！——晓多 01 此刻，我坐在新的办公室里，宽敞明亮，整理好一些交接的文件，加班完成了一篇材料，静静的坐下来喝了一口水。 打开电脑来写这篇原本几天就应该动笔的文章，为自己曾经的生活画上一个句号，然后一切清零重新开始。 而几天前我还在待了二十多年的小城市，官方的全国排名情况公布之后，这个我长大的豫北小城市已经成了五线城市。 以上学为界，大学前的十多年在这里长大，这里不是我的祖籍却是我的家。 毕业那年我回来过，在当地的报社实习，虽然没有基本工资但仅靠稿费就已经超过了不少记者，一张报纸有时会有五篇稿子是我写的，经常上头版头条。在报社招考的时候就知道能留下来，...

晓多 (/u/fee4b4b0b89e?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

承认吧，秒回信息的人并不会被珍惜 (/p/c9185bb9e0c…

(/p/c9185bb9e0c8?utm_campaign=maleskine&utm_content=note&utm

01 今天看到一个有趣的话题，讲的是那些秒回信息的人，反而不一定会得到珍惜。 所以真正会撩的高手，他也许会盼着对方秒回自己，却故意拖上一会儿才回复对方。 在心理学上，这一现象被称为“奖赏不确定性”，不确定性会提高奖赏的吸引力。 纽约大学的一位教授说，如果让实验动物按动杠杆，每按一次都能得到食物，它们按杠杆的频率会逐渐下降。 反之，如果降低提供食物的频率，它们就会积极地按动杠杆，这时它们的多巴胺水平也会升高，这令它们感到兴奋。 所以，如果将心上人的回复视作一种奖赏，那么等待对方的回复，就是享受不确定性的过程。 不确定的魅力到底有多大？《哈佛幸福课》的作者进行过一项研究，让女大学生们浏览四...

衷曲无闻 (/u/deeea9e09cbc?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)
