Multi-label Image Recognition by Recurrently Discovering Attentional Regions

阅读总结

15331416 赵寒旭

**Abstract**

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. Introduction**

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. Related Works**

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 Multi-label image recognition**

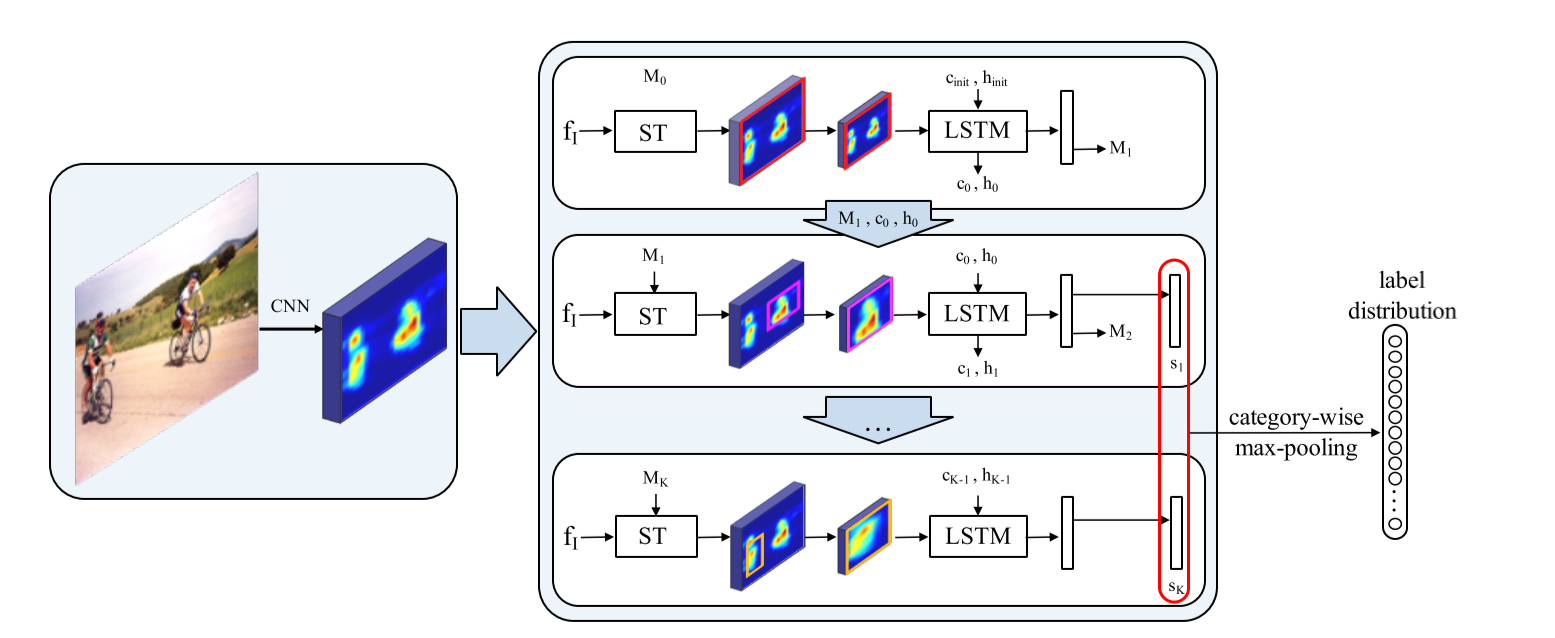
人工设计特征时期，大多采用“词袋模型”解决多标签识别任务。但随着复杂场景的出现及手工设计特征过于繁琐且不稳定，深度卷积神经网络提取的深层特征泛化能力强，逐渐取代“词袋模型”。为了更好地考虑标签间的相互关系，不再独立看待每个标签，许多基于深度卷积神经网络的方法还是包含了很多传统的形象化模型，如条件随机场、依赖网络、共生矩阵。

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 Visual attention model**

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. Model**



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

（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。

<https://www.jianshu.com/p/620acf8fcdab>

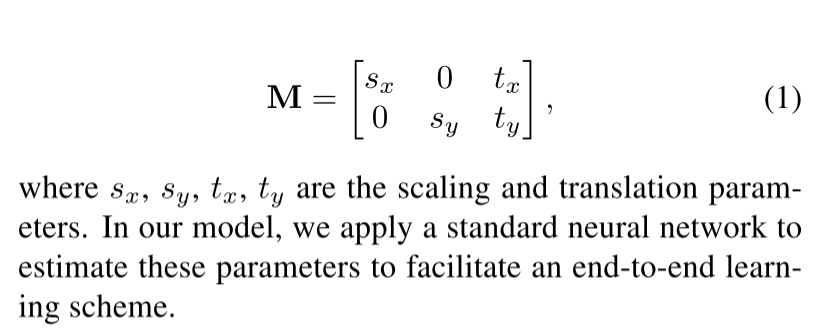
**3.1 ST for Attentional Region Localization**

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。

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]。

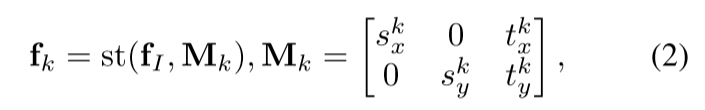
在本文中，ST层的工作是从特征图fI中提取注意区域（fk），方法是由神经网络学习的映射矩阵M。



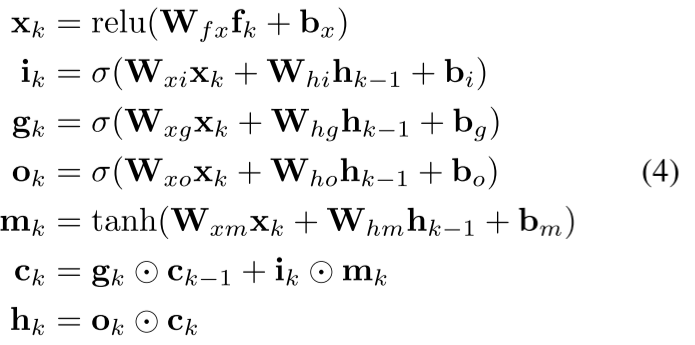
**3.2 Recurrent Memorized-Attention Module**

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fI中利用Mk提取注意力区域的过程如下：



LSTM以为输入，更新隐藏状态（hidden state）及记忆单元（memory cell）。



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，、、、、、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。在理想状态下，记忆单元ck记录（编码）了前k-1个时刻的有用信息。

1）M的更新

其中表示预测的类别分数（score）分布

2）类别预测

**4. Learning**

**4.1 Loss for Classification**

**4.2 Loss for Attentional Region Constraints**

**5. Experiments**

**5.1 Settings**

**5.2 Comparison with State-of-the-art Methods**

**5.2.1 Performance on the VOC 2007 dataset**

**5.2.2 Performance on the MS-COCO dataset**

**5.3 Ablation Study**

**5.3.1 Attentional regions v.s. object proposals**

**5.3.2 Analysis of the attentional region constraints**

**5.3.3 Multi-scale multi-view evaluation**

**6. Conclusion**