《Multi-Label Image Recognition with Graph Convolutional Networks》评述

刘方昊
21821216

Abstract

多标签图像识别(multi-label image recognition)是计算机视觉领域一个非常重要且比较困难的任务。任务的目标是预测一张图像中出现的多个物体标签。由于多个相关物体通常同时出现在一副图像之中，因此需要对标签依赖关系进行建模，以提高识别性能。为了捕捉和探索这些重要的依赖关系，文章提出了一种基于图卷积网络 (GCN) 的多标签分类模型。该模型通过 data-driven 方式建立标记间有向图并由 GCN 将类别标记映射为对应类别分类器。在两个多标签图像识别权威数据集上的实验表明，此方法明显优于现有的其他最先进的方法。此外，可视化分析表明，该模型学习的分类器保持了有意义的语义拓扑结构。

1. 介绍

1.1. 多标签图像识别

多标签图像识别是计算机视觉领域的一项基础性任务，其目标是预测图像中存在的一组对象。它可以应用于医学诊断识别、人类属性识别、零售识别等诸多领域。与多类别图像分类相比，输出空间的组合特性使得多标签图像识别任务更具挑战性。由于物体在物理世界中通常是共存的，多标签图像识别的关键是对标签依赖关系进行建模。

1.2. 现有方法及其局限性

解决多标签识别问题的一个朴素方法是分别考虑各个目标，通过将多标签问题转换成多个二元分类问题，预测每个目标是否存在。由于深度卷积神经网络在单标签图像分类上取得的巨大成功，二元分类的性能已得到极大提升。但是这个方法忽视了物体之间复杂的拓扑结构，因此在本质上有局限性。正是这个缺陷促使研究员寻找能够获取并从多个角度探索标签之间相关性的方法。其中的部分方法基于概率图模型或循环神经网络（RNN），可显式地对标签依赖性进行建模。前者将多标签分类问题定义为一个结构推理问题，由于计算复杂度高，可能会出现可伸缩性问题，而后者则需要根据预先定义或学习的某种顺序来预测标签。另一个研究方向是通过注意力机制来对标签相关性进行隐式建模。该方法考虑的是图像中被注意区域之间的关系（局部关联）。即便如此，该方法仍然忽略了标签之间的全局关联，这需要从单个图像之外的知识来推断。

1.3. 文章提出的方法思路

文章提出了一种新的基于 GCN 的多标签图像识别模型 (ML-GCN)，该模型能够捕获多标签图像识别的标签相关性，具有较强的可扩展性和灵活性。这种方法不把对象分类器看作一组需要学习的独立参数向量，而是建议通过一个基于 GCN 的映射函数，从先验的特征表示 (例如词嵌入) 中学习相互依赖的对象分类器。随后，生成的分类器再被应用于由另一个子网络生成的图像特征，以实现端到端训练。由于嵌入到分类器的映射参数在所有类之间共享，因此来自所有分类器的梯度都会影响这个基于 GCN 的分类器生成函数。这可以对标签关联进行隐式建模。此外，为了对分类器学习中的标签依赖关系进行显式建模，文章中设计了一个有效的标签相关矩阵来指导节点间的信息传播，有效地缓解了过拟合。

2. 算法

这一节首先介绍算法提出的动机，接着介绍一些图卷积网络初步知识，最后介绍 ML-GCN 模型以及用于相关系数矩阵构建的二次加权方法。

2.1. 动机

如何有效地捕获目标标签之间的相关性，并探索这些相关性，提高分类性能，对于多标签图像识别非常重要。文章中使用图来建模标签之间的相互依赖关系，这是一种捕获标签空间中的拓扑结构的灵活方法。具体来说，文章将图的每个节点 (标签) 表示为标签的词嵌入，并使用 GCN 将这些标签嵌入直接映射到一组相互依赖的分类器中，这些分类器可以直接应用于图像特征进行分类。基于 GCN 的模型有两个设计动机：

首先，由于嵌入到分类器的映射参数是在所有类别中共享的，学习到的分类器可以在词嵌入空间中保留弱语义结构。同时，所有分类器的梯度都会影响分类器的生成函数，从而隐含地对标签依赖关系进行建模。

其次，基于标签的共现模式，文章设计了一个全新的标签相关系数矩阵，可通过 GCN 显式地对标签依赖关系建模，使用该模式，节点特征更新时将从相关节点 (标签) 吸收信息。

2.2. 图卷积神经网络简介

图卷积网络可用于进行半监督分类任务，其基本思想是通过在节点之间传播信息来更新节点表示。与对图像中的局部欧几里德结构进行操作的标准卷积不同，GCN 的目标是学习一个图 G 的函数 $f(.,.)$。该函数的输入是特征描述 $H^l \in R^{n\times d}$ 和相关系数矩阵 $A \in R^{n\times n}$，从而把节点特征更新为 $H^{l+1} \in R^{n\times d^{'}}$。每个 GCN 层都可以写成一个非线性函数：

$$H^{l+1} = f(h^l, A)$$

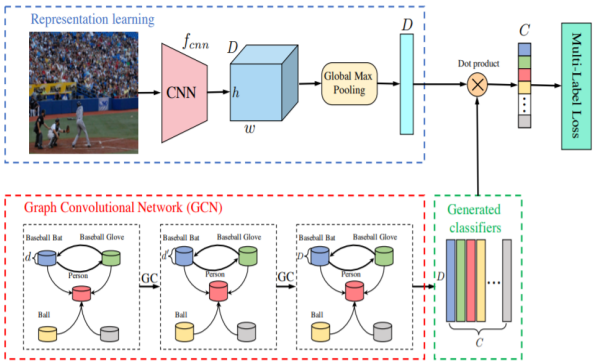
卷积运算后，$f(.,.)$ 可表示为：

$$H^{l+1} = h(\hat{A}H^lW^l)$$

这里 $W^l \in R^{d\times d^{'}}$ 表示我们要学习的转换矩阵，$\hat{A} \in R^{n\times n}$ 是相关矩阵 A 的归一化版本，$h(.)$ 表示一个非线性运算，文章中采用的是 LeakyReLU。

2.3. 用于多标签识别的 GCN

ML-GCN 是建立在 GCN 之上的。GCN 的设计初衷是半监督分类，其节点层面的输出结果是每个节点的预测得分。而在 ML-GCN 中，每个 GCN 节点的最终输出都被设计成与标签相关的分类器。此外，不同于其它任务，这里的多标签图像分类任务没有提供预定义的图结构（即相关系数矩阵）。这需要从头构建相关系数矩阵。方法的总体框架如下图所示，它由两个主要模块组成，即图像特征学习和基于 GCN 分类器学习。



图像特征学习：文章在实验中使用 ResNet-101 作为实验基础模型；然后应用全局 max-pooling 获取图像层面的特征 x：

$$x = f_{GMP}(f_{cnn}(I;\theta_{cnn}))$$

GCN 分类器学习：通过一个基于 GCN 的映射函数从标签特征学习相互依赖的目标分类器 $W = \{w_i\}_{i=1}^{C}$，并使用堆叠 GCN，其中每个 GCN 层 I 的输入都取前一层 H^l 的节点特征作为输入，然后输出新的节点特征 H^{l+1}。最后一层的输出是分类器 $W \in R^{C\times D}$。把所学分类器应用于图像特征，可以计算出预测分数：

$$\hat{y} = Wx$$

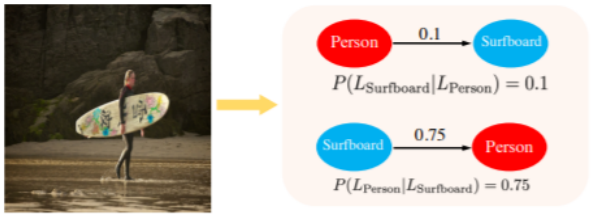
假设一张图像的真实标签是 $y \in R^C$，那么整个网络可使用传统多标签分类的损失函数来训练：

$$L = \sum_{c=1}^{C} y^c log(\sigma(\hat{y}^c)) + (1-y^c)log(1-\sigma(\hat{y}^c))$$

2.4. 相关系数矩阵

GCN 的工作原理是基于相关矩阵在节点之间传播信息。因此，如何建立相关矩阵 A 是 GCN 的一个关键问题。在大多数应用中，相关矩阵是预先定义的，但在任何标准多标签图像识别数据集中都未提供相关矩阵。文章通过数据驱动的方式建立了相关性矩阵。通过挖掘标签在数据集中的共现模式来定义标签之间的相关性。

本文以条件概率的形式 $(P(L_j|L_i))$ 对标签的相关依赖性进行建模。显然，相关系数矩阵不是对称的，即 $P(L_j|L_i)$ 未必等于 $P(L_i|L_j)$：



上图展示了一个例子，当图片中出现人的时候，同时出现冲浪板的概率并不大，但当图片中出现冲浪板的时候，同时出现人的概率却很大。

为构建相关系数矩阵，文章首先统计了训练数据集中标签对的出现次数，得到矩阵 $M \in R^{C\times C}$，然后使用这个标签共现矩阵得到条件概率矩阵：

$$P_i = M_i/N_i$$

其中 N_i 表示在训练集中 L_i 出现的次数，$P_{ij} = P(L_j|L_i)$ 表示当标签 $L-i$ 出现时标签 L_j 出现的概率。

然而，上述简单的求解相关性可能存在两个缺点。首先，标签与其他标签之间的共现模式可能呈现长尾分布，其中一些罕见的共现可能是噪声。其次，训练和测试中共现的绝对数量可能不完全一致。将相关矩阵过度拟合到训练集，会影响泛化能力。因此，文章提出对相关系数矩阵 A 进行二值化处理。具体而言，选择合适的阈值 τ 用于过滤噪声：

$$A_{ij} = \begin{cases} 0 & if\ P_{ij} < \tau \\ 1 & if\ P_{ij} \geq \tau \end{cases}$$

过度平滑问题：经过 GCN 后，一个节点的特征是其自身特征和相邻节点特征的加权和。而二值化相关系数矩阵的一个直接问题是其可能导致过度平滑。为了缓解这一问题，文章中提出了以下二次加权方法：

$$A'_{ij} = \begin{cases} p/\sum_{j=1,i\neq j}^{C} A_{ij} & if\ i \neq j \\ 1-p & if\ i = j \end{cases}$$

通过这种做法，在更新节点特征时，节点本身的权值是固定的，相关节点的权值由邻近分布决定。当 p→1 时，不考虑节点本身的特征。而另一方面，当 p→0 时，相邻信息往往被忽略。

3. 实验

文章的实验部分记录了在两个基准多标签图像识别数据集 MS-COCO 和 VOC 2007 上的结果，结果表明文章中提出的算法效果比现有方法都要好。

3.1. MS-COCO 实验结果

文中给出了基于二值相关系数矩阵与基于二次加权相关系数矩阵两个版本的结果，后者的分类表现更好，在几乎所有指标上领先其它方法，这证明了新提出的网络与二次加权法的有效性。具体实验结果见下表：

Methods	All							Top-3					
	mAP	CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1
CNN-RNN [28]	61.2	–	–	–	–	–	–	66.0	55.6	60.4	69.2	66.4	67.8
RNN-Attention [29]	–	–	–	–	–	–	–	79.1	58.7	67.4	84.0	63.0	72.0
Order-Free RNN [1]	–	–	–	–	–	–	–	71.6	54.8	62.1	74.2	62.2	67.7
ML-ZSL [15]	–	–	–	–	–	–	–	74.1	**64.5**	69.0	–	–	–
SRN [36]	77.1	81.6	65.4	71.2	82.7	69.9	75.8	85.2	58.8	67.4	87.4	62.5	72.9
ResNet-101 [10]	77.3	80.2	66.7	72.8	83.9	70.8	76.8	84.1	59.4	69.7	89.1	62.8	73.6
Multi-Evidence [6]	–	80.4	70.2	74.9	85.2	72.5	78.4	84.5	62.2	70.6	89.1	64.3	74.7
ML-GCN (Binary)	80.3	81.1	70.1	75.2	83.8	74.2	78.7	84.9	61.3	71.2	88.8	65.2	75.2
ML-GCN (Re-weighted)	**83.0**	**85.1**	**72.0**	**78.0**	**85.8**	**75.4**	**80.3**	**89.2**	64.1	**74.6**	**90.5**	**66.5**	**76.7**

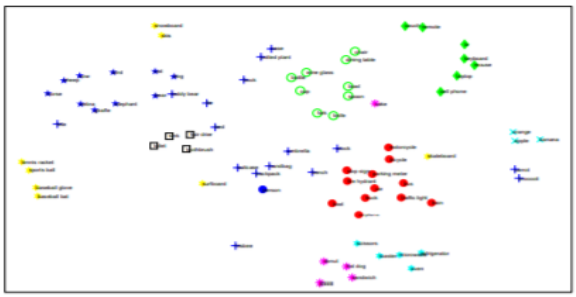
3.2. VOC 2007 实验结果

使用二次加权法的 ML-GCN 模型在 mAP 指标上得到了 94% 的分数，高出先前最优方法 2%。即使以 VGG 为基础模型，仍然超出先前最佳水平 0.8%。

Methods	*aero*	*bike*	*bird*	*boat*	*bottle*	*bus*	*car*	*cat*	*chair*	*cow*	*table*	*dog*	*horse*	*motor*	*person*	*plant*	*sheep*	*sofa*	*train*	*tv*	mAP
CNN-RNN [28]	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0	92.4	91.7	84.2	93.7	59.8	93.2	75.3	**99.7**	78.6	84.0
RLSD [34]	96.4	92.7	93.8	94.1	71.2	92.5	94.2	95.7	74.3	90.0	74.2	95.4	96.2	92.1	97.9	66.9	93.5	73.7	97.5	87.6	88.5
VeryDeep [26]	98.9	95.0	96.8	95.4	69.7	90.4	93.5	96.0	74.2	86.6	**87.8**	96.0	96.3	93.1	97.2	70.0	92.1	80.3	98.1	87.0	89.7
ResNet-101 [10]	99.5	97.7	97.8	96.4	65.7	91.8	96.1	97.6	74.2	80.9	85.0	**98.4**	96.5	95.9	98.4	70.1	88.3	80.2	98.9	89.2	89.9
FeV+LV [33]	97.9	97.0	96.6	94.6	73.6	93.9	96.5	95.5	73.7	90.3	82.8	95.4	97.7	95.9	98.6	77.6	88.7	78.0	98.3	89.0	90.6
HCP [31]	98.6	97.1	98.0	95.6	75.3	**94.7**	95.8	97.3	73.1	90.2	80.0	97.3	96.1	94.9	96.3	78.3	94.7	76.2	97.9	91.5	90.9
RNN-Attention [29]	98.6	97.4	96.3	96.2	75.2	92.4	96.5	97.1	76.5	92.0	87.7	96.8	97.5	93.8	98.5	81.6	93.7	82.8	98.6	89.3	91.9
Atten-Reinforce [2]	98.6	97.1	97.1	95.5	75.6	92.8	96.8	97.3	78.3	92.2	87.6	96.9	96.5	93.6	98.5	81.6	93.1	83.2	98.5	89.3	92.0
VGG (Binary)	98.3	97.1	96.1	96.7	75.0	91.4	95.8	95.4	76.7	92.1	85.1	96.7	96.0	95.3	97.8	77.4	93.1	79.7	97.9	89.3	91.1
VGG (Re-weighted)	99.4	97.4	98.0	97.0	77.9	92.4	96.8	97.8	80.8	93.4	87.2	98.0	97.3	95.8	98.8	79.4	95.3	82.2	99.1	91.4	92.8
ML-GCN (Binary)	99.6	98.3	97.9	97.6	78.2	92.3	97.4	97.4	79.2	94.4	86.5	97.4	97.9	97.1	98.7	84.6	95.3	83.0	98.6	90.4	93.1
ML-GCN (Re-weighted)	99.5	**98.5**	**98.6**	**98.1**	80.8	94.6	**97.2**	98.2	**82.3**	**95.7**	86.4	98.2	**98.4**	**96.7**	**99.0**	**84.7**	**96.7**	**84.3**	98.9	**93.7**	**94.0**

3.3. 分类器可视化

文章对比了采用 ML-GCN 模型习得的分类器与由 vanilla ResNet 得到的基本分类器的可视化分析。可以看到，由文章提出的方法学习到的分类器能够维持语义的拓扑结构。



(a) t-SNE on the learned inter-dependent classifiers by our model.

(b) t-SNE on the classifiers by the vanilla ResNet.

3.4. 在图像检索上的表现

文章使用 k-NN 算法执行基于内容的图像检索验证由新模型习得的图像特征的鉴别能力，实验结果表明，ML-GCN 不仅能有效地捕获标签依赖关系，学习更好的分类器，而且对图像特征学习和多标签识别都有一定的帮助。

4. 结论

标签依赖关系的捕获是多标签图像识别的关键问题之一。为了对这一重要信息进行建模和探索，文章提出了一个基于 GCN 的模型，从已有的标签表示 (如词嵌入) 中学习相互依赖的对象分类器。为了对标签依赖关系进行显式建模，我们设计了一种新的二次加权方法，通过平衡结点与其临近结点之间的权值来构造 GCN 的相关矩阵，用于节点的特征更新。该方案能够有效地解决影响 GCN 性能的两个关键问题——过拟合和过平滑。定量和定性的结果都验证了 ML-GCN 的优越性。

5. 个人总结

这篇文章的思路简明清晰，将 GCN 与 CNN 有机结合起来，达到了最佳表现。我认为，这篇论文的最大贡献和创新点在于：

1、基于 GCN 的映射函数，从先验的特征表示中学习相互依赖的分类器以实现端到端的训练。

2、设计了相关矩阵，并且设计了二次加权的方法，避免了过拟合与过平滑问题。

基于这样的思路，文中提出的方法在两个多标签的基准数据集上达到了最好的效果。这样的将现有方法有机结合起来发挥更大优势的思路非常值得我们学习。

参考文献

[1] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, Yanwen Guo. Multi-Label Image Recognition with Graph Convolutional Networks. In CVPR 2019.