研究论文

深度学习图像标注与用户标注比较研究*

陆　伟　罗梦奇　丁　恒　李　信
(武汉大学信息管理学院　武汉 430072)
(武汉大学信息检索与知识挖掘研究所　武汉 430072)

摘要:【**目的**】利用用户对图像标注的标签提出用户标签框架，并通过用户标签框架总结深度学习自动标注图像的不足。【**方法**】统计分析从 Flickr 上下载的大约 100 万张图像数据集中的用户标签，抽取高频词进行用户标签框架匹配。将用户标签与 ImageNet 数据库标签进行对比总结。对含有高频词的图像使用 MXNet 深度学习算法进行标注，分析标注结果。【**结果**】当前深度学习自动标注，在图像背景知识、总体描述以及人类感官描述等方面还存在缺陷。【**局限**】数据集的范围需要扩大，深度学习算法的种类需要增加。【**结论**】自动标注图像的发展，需要建立图像信息与背景知识、描述等的联系；并且深度学习未来发展还需要赋予计算机逻辑推理以及情境感知的能力。

关键词: 图像标注　用户标签　自动标注　机器学习　深度学习　人工智能
分类号: G255
DOI: 10.11925/infotech.2096-3467.2018.0052

1　引　言

在图像标注及图像检索领域，为了对图像进行管理和搜寻，图像标注任务是一个亟待解决的问题。传统的用户标注难以满足大数据时代的图形标注任务，利用计算机对图像进行自动标注应运而生。目前，图像标注研究主要分为基于图像概念和基于图像内容两大类[1]，前者主要侧重于利用描述词(如标题、关键词等)检索，后者侧重于图像的视觉特征(如颜色、纹理等)。如何使计算机跨越图像语义鸿沟是图像标注领域亟待解决的问题[2]。

针对这一问题，计算机、机器学习和情报学等领域学者进行了一定的探索，包括算法的提升以及精确度的提高，以及围绕视觉信息和标签信息的关联关系提取等。然而，图像自动标注不仅需要从技术层面提升精确性，还需要从语义上对图像标签进行拓展；此外，图像标注的覆盖范围也不够。因此，本文试图在理解图像的用户标注标签的基础上，归纳总结图像用户标签分类框架；接着利用数据分析的方法，对图像用户标签和深度学习自动标注标签进行统计、对比分析，从而对深度学习自动标注标签的覆盖范围进行检测，指出当前深度学习图像自动标注的不足之处；最后，提出图像自动标注需要改进的方向。

2　相关研究

2.1　图像标签分类

目前，关于图像的研究大都涉及到对图像标签的分类。比较典型的有 Flickr 和 Panofsky-Shatford 分类体系: Flickr 将图像标签细分为形容词类、复合词类(含有两个词以上的词组)、情感类、事件类、诙谐类、语言类(除英文之外的)等 18 类[3]。Panofsky-Shatford 标签矩阵作为一个比较完整的分类体系，常被用来作为分类基础，这个二维分类矩阵，将图像标签从纵向分成 who、what、where、when，视觉元素以及未知类，　从

通讯作者：罗梦奇, ORCID: 0000-0002-6762-6249, E-mail: lakeygtgz@163.com。
*本文系国家自然科学基金面上项目“面向词汇功能的学术文本语义识别与知识图谱构建”(项目编号:71473183)的研究成果之一。

横向分成总体、细节以及抽象类[4-5]。在横向分类中，总体类代表着图像的总体内容，例如人物、对象和事件等；而细节类则描述有关图像的背景知识，用户需要了解图像描述的内容，例如关于历史、地理以及艺术的知识，才能标注此类标签；抽象类是基于图像内容对图像反映的信息进行描述，例如“年轻”、“高兴”等。鉴于图像向人类传递的最直观的信息是视觉信息，文献[6]将图像描述分类为视觉、上下文、时间、命名实体、以及其他非视觉类。

有学者将图像标签分为几个大类，例如，为了对图像进行标注，文献[7]将图像内容分为对象类、物体、人物和场景等；从词性以及语义的角度，文献[8]利用WordNet 将图像标签分成 5 种类型：“where”“when”“who”“what”和“how’；从人类感知的维度，文献[9]总结了用户图像标签中比较重要的几种类型：事件和活动、特定地点、场景类、季节、图像风格、艺术与文化等；同时，也有学者对特定领域的图像标签进行分类，例如，有关艺术图像的社会标签，在文献[10]中被分类为：人物或物品、活动或状态、地点、时间、视觉元素和非主题的；文献[11]认为，用户在检索图像时，往往需要特定的检索关键词描述其图像需求，在关于艺术图像的领域将用户需求关键词描述分为目录(元数据)、内容描述、样式、图像地点、颜色、实例、简介以及情感等。在对敦煌壁画数字图像的研究中，文献[12]提出对图像的语义描述层次模型，将对图像的描述分成语义类和对象类，其中，语义类包括情感、行为活动和场景，对象类包括对象空间和对象。

2.2 用户图像标注行为意图

用户的标签不仅仅反映图像的语义信息，还能够展现出用户对图像的认知，并在一定程度上可利用作为跨越语义鸿沟的解决方法[13]。文献[14]表明，大约80%的用户会使用线上标注系统，而他们使用标注系统的目的有：管理自己的图像、方便未来搜寻、分享和展示自己的图像以及对他人提供的图像进行评论和标注。其中，“方便未来搜寻”这一目的是最显著的。图像用户标注的面向分享对象可以是自己、家人朋友或者公众[15]，也就是说，标注的目的主要是方便这三类人群对图像进行管理与分享[16]。

有的学者认为，图像标注是为了针对个人和社会两大类对图像进行管理和交流，其中管理图像的目的在于未来的检索和图像集的建立，而交流图像的目的在于记录和传递信息[17]。并且，图像用户标注的动机同时也影响到用户标签的排序[18]。为管理图像而进行标注，是方便进行浏览；为描述图像进行标注，是方便进行检索。用户群体出于管理图像的目的而标注的标签，会比仅仅为了描述图像而标注的标签显现出更大的不一致[19]。

总而言之，用户对图像标注的目的在于方便未来的管理、浏览以及检索，而图像标注的服务对象为本人、家人朋友以及公众。

2.3 图像自动标注

当前图像自动标注研究，主要利用图像视觉信息及语义信息。其中，视觉信息包括图像的颜色、纹理和形状等；语义信息包括图像的描述文字、概念等。

对于利用视觉信息进行自动标注，最常见的方法是将图像的视觉信息转化为特征向量，例如文献[20]中，训练的时候抽取人脸图像视觉信息形成特征向量，将这些特征信息进行聚类归入不同的类别中，测量待标注图像与数据库中图像的相似度，从而进行人脸图像标注；而文献[21]使用 CPAM 模型抽取图像的颜色纹理特征并转化为特征值，进而使用 SVC(Support Vector Clustering)算法利用这些特征值进行训练，即可以对图像进行聚类从而进行自动标注，其中还使用了 PSO (Particle Swarm Optimization)算法优化 SVC。标注过程使用的主要是基于机器学习和基于网络图方法的图像自动标注，例如，利用机器学习中广泛使用的 KNN(K-Nearest Neighbors)算法，得到与需要标注图像的最相近的一些图像，并使用标签传播算法为需要标注的图像分配关键词[22]。

对于利用语义信息进行自动标注，一般来说，也会结合使用图像的视觉信息。在为标签及标签的关系构建网络图的同时，利用双层 BoW 模型处理视觉显著特征，包括局部特征、图像的显著区域等[23]。文献[24]结合特征空间和概念空间，对视觉特征原型进行无监督聚类，对概念原型进行监督聚类。将待标注图像映射到与其特征对接近的原型类中，从而获得图像标签。为了计算从相似图像中得到的关键词与需标注图像之间的相似度得分，文献[25]除了使用视觉描述外，还考虑关键词之间的语义关系并将关键词及它们之间的关系构建网络连通图模型。文献[26]提出

基于核典型相关分析(Kernel Canonical Correlation Analysis, KCCA)的自动图像标注框架，将文本(包括专家标签和社会标签)特征和视觉特征融合到语义空间并利用视觉信息矫正存在噪声的用户标签。另也有将图像空间内容添加到 BoVW (Bag-of-Visual-Words)模型中的[27]。文献[28]则使用 LDA 结合自然语言处理将关键词的低层语义投射到高层语义空间而生成图像主题，并结合 ConceptNet 通过持续关联模型使用视觉特征进行聚类建立图像之间的关系。此外，利用图像分类器，可以构建一系列的图像–图像、图像–文本以及文本–文本的标注框架模型，对相似图像进行匹配并标注[29]。还有研究是利用图像和语义标签的相似度、图像与图像的相似度对标注不完整的图像加入标签[30]。此外，还有基于与图像相关联的文本而对图像进行标注，例如，文献[31]中对新闻图片的自动标注，除了利用关键词和图像类别，图像所处的新闻文本的内容也可以作为一个参考因素。对当前图像自动标注涉及到的最新的常用的方法进行概述，如表 1 所示。

表 1　图像自动标注所涉及的方法与技术

方法/技术	应用/举例
SIFT(尺度不变特征变换)/ SURF LBP(局部二值模式)	描述图像局部特征 辅助图像局部对比
VLAD	提取图像特征
CNN(卷积神经网络)/RNN(循环神经网络)/DNN(深度神经网络)/LSTM(长短期记忆网络)	生成将图像信息和文字信息对应的模型[32],图像分类[33],同时也用于图像信息的捕获与解析
NLP(自然语言处理)	对与图像相关联文本的处理
SVM(支持向量机)	图像视觉信息(语义信息)分类器[34]
Fisher Vector Encoding	将图像的视觉描述子映射为高维向量[35]
pLSA/LDA(主题模型)	对图像相关的文本、关键词进行主题模型的建立[28,36]
CCA(典型关联分析)	建立图像和文本的关联[37]
Apriori Algorithm(关联规则)	发掘图像与图像、图像与文本的关联[29]
K-Nearest Neighbor(邻近算法)/LMNN(大间隔最近邻居)	图像信息聚类
Community Detection(社区发现算法)	为图像的语义信息、视觉信息构建概念图[38]
Eigenfaces/Fisherfaces	人脸识别

总的来说，上述计算机对图像自动标注是对图像的视觉信息和语义信息进行抽取，并对这些信息进行建模分析，再进行相似度匹配，从而将已有的标签加入到待标注的图像上。现有的方法在精确度上也已经达到一定的高度。然而，计算机本身并不能理解这些图像、标签及它们关联的含义，只是单纯从训练模型中找出与图像相匹配的标签。这是由于计算机缺少相应的背景知识、感知能力以及推理能力。

3　图像用户标签分类框架与实验设计

结合用户行为动机以及前人对用户图像标签分类的总结，本文对用户图像标签进行预分类。从上述的研究成果来看，用户对图像的标注主要是方便管理及搜寻，用户对图像的认知便成为标注的主要来源。因此，时间、地点、人物和事件这 4 大类，是用户标注的基本类别；对于与人们社会活动关联性比较小的图像，直观上来说可以有颜色纹理形状和具体物品对象(如风景、动植物和建筑等)这两大类；人类感知也是用户对图像描述中必不可少的部分，可分为描述类和情感类；图像的元数据等标签可归类为图像生成设备，与之相关的还有处理图像的设备；如果图像不是自己拍摄的，用户也会标注这一类关于图像来源的标签；对图像的抽象描述，是基于用户的知识与对图像的总体感知，可归为抽象类。表 2 是对这些大类的总结与描述。

表 2　用户标签分类框架

类别	描述
时间	包括季节、年份以及早中晚等一天中的某个时段
地点	某些特定的或标志性地点，如海滩、树林等；国家、城市名和地名；东西南北等方位
人物	人的名字；某一类人群
事件	人类的社会活动；图像中描述的事情
颜色纹理形状	图像最直观的视觉信息
对象	包括风景、动植物、建筑和物品对象等
描述类	形容词和需要结合背景知识的形容词
情感类	图像表达的情感；人们看到图像的感觉；需要结合背景知识的情感
抽象类	艺术、历史和文化；图像反映的整体内容；(这一类主要是定义图像总体，需要一定的背景知识)
图像生成及处理设备	图像生成设备或参数，如拍摄的相机型号、焦距和曝光等；图像后期处理软件
图像来源	图像来源的网址等

为了进一步验证上述的用户标签框架及分类，并统计分析用户标签与深度学习自动标注标签的不同，本文进行如下实验：利用著名图片分享网站的数据(包括图像及用户标签)，总结用户标签的特征，将用户标签与深度学习自动标注标签进行比对，通过不同的角度分析统计结果。

3.1 数据源及实验工具

本实验采用Yahoo! Flickr的媒体对象数据集[39]，包括图像信息描述文件以及图像本身。选用其中大约100万张图像以及图像信息，包括图像编号、用户编号、上传时间、拍摄设备、标题、描述、用户标签、经度、纬度等。

ImageNet是一个应用广泛的大型图像集，包含大约1 400万张已标注图像。依照WordNet的语义结构，这些图像被规划到一个结构树中。这个结构树是一个从上往下、不同类别又不断分成不同子类的金字塔结构。其中，非空同义词集总数为21 841。本实验使用ImageNet数据库中的词作为“自动标注”词集，用来与用户标签进行对比。

MXNet是一个自带训练模型的深度学习框架，包含循环网络、卷积网络和动态贝叶斯网络。由于此深度学习框架能够符合本实验对图像自动标注的要求，并且其模型能够满足当前机器学习的最先进算法的实现，所以选择MXNet作为图像自动标注的算法框架。

3.2 实验过程

图像数据集中的描述文件包含图像编号和社会标签在内的各种信息，并且部分图像的某些信息栏中为空白。对本实验而言，最需要的是数据集中的图像社会标签，为了减少冗余以及保证数据完整性，首先提取在标题、描述和社会标签这三个信息栏都有描述词的图像；然后，根据实验目的，使用ImageNet数据库中的词与数据集中图像的社会标签一一对比，提取出包含深度学习自动标注词以外的用户描述标签词的图像编号并找出所对应的图像；再使用深度学习MXNet模型对提取的每张图像进行深度学习自动标注；最后，结合MXNet自动标注的词和ImageNet数据库的词，对比社会标签和深度学习自动标注词，找到每一张图像中只存在于用户标注中的标签。

4 实验结果及分析

本实验采用的原始数据集共包含1 091 312张图像。其中，拥有完整的标题、描述和社会标签的图像有1 072 938张，占比98.3%。经过对比以及筛选，得到的包含深度学习自动标注标签以外的用户标签词的图像有721 897张，包含只存在于用户标注中的标签词的数量总共有3 535 867，去重后标签总量为786 518。

4.1 用户标签分类

对数据集中的所有用户标签进行统计，计算出用户标签出现频率比较高的词，然后将它们归类到上述的用户标签初步分类框架中，同时将每一大类分成不同的子类。这样细化用户标签分类，一方面是为了给用户标签明确一个具体的框架，另一方面是通过这个框架，将自动标注的标签词进行相应的对比分类，总结出深度学习标注标签的缺漏之处，以进行深入分析。表3是对这些用户标签在分类框架中分布的描述。

表3 实验数据集的用户标签分类框架

类别	子类别及高频标签数	总频次	标签举例
时间	年份(11)	88 553	2016
	季节(5)	21 037	summer
	月份(12)	21 958	september
	一天中的某时段(3)	9 267	morning
地点	国家或地名(158)	419 308	newyork
	方位(5)	5 239	north
	标志性地点(46)	111 277	beach
人物	人名(2)	2 582	jovens
	某一类人(19)	43 547	girls
事件	活动(50)	125 493	Hiking
	结合背景知识的活动(7)	18 590	cosplay
颜色纹理形状	(11)	29 038	pink
对象	风景(14)	46 936	skyline
	动植物(17)	21 120	bird
	建筑(6)	14 103	castle
	对象(36)	62 080	building
	物品(49)	99 773	apple
描述类	形容词(18)	46 934	Beautiful
	结合背景知识的形容词(6)	8 933	National
情感类	感觉(2)	4 365	fun
	结合背景知识的情感(2)	2 346	pride
抽象类	艺术(7)	26 896	streetart
	历史(2)	3 068	historic
	文化(1)	1 050	culture
	图像反映的内容(14)	42 263	war
图像生成、处理设备	图像生成设备或参数(28)	78 053	sony
	图像处理软件(3)	17 215	fireworks
图像来源	网站(21)	24 392	www.500px.com

其中，高频词的总数为 555，总频次为 1 395 416。最高频出现次数为 15 600 次，最低频出现次数为 902 次，平均频次为 2 514 次。

4.2　高频词对比统计结果

基于上一节的用户标签高频词，将其与自动标注标签(ImageNet 数据库)进行比对，得到仅存在于用户标签中的高频词，统计结果如表 4 所示。

表 4　高频词对比统计结果

用户标签高频词数(按频次排序)	自动标注标签数	比例
100	74	74.00%
200	145	72.50%
300	215	71.67%
400	283	70.75%
500	337	67.40%
555	372	67.03%

从表 4 可以看出，以用户标注标签的高频词为准，从总体来看，自动标注能够覆盖大部分标签词。显而易见，自动标注标签的覆盖范围随着用户标签高频词数的增加而降低，而用户标签高频词的平均频次随着按频次排序的词的数量的增加而减少。因此，本实验也对高频词的平均频次与自动标注标签的覆盖范围进行统计。图 1 展示了自动标注标签范围与用户高频词的平均频次的变化。

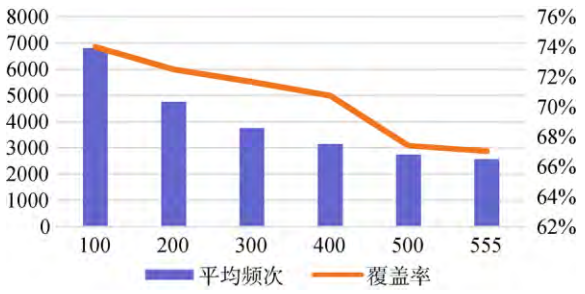

图 1　用户高频词平均频次与自动标注标签覆盖率的变化

图 1 的横轴表示用户高频词的数量，这些高频词是按照出现频次由高到低顺序排列的；条形表示每一个高频词数量级别上对应的平均词频，由图 1 中纵向右侧的数字表示；折线表示自动标注标签在每一个高频词数量级别上的覆盖率，由图 1 中纵向左侧的数字表示。可以看出，高频词的平均词频随着高频词数量的增加而下降，自动标注的覆盖范围也随着高频词数量的增加呈现出下降趋势。换言之，自动标注的覆盖范围的变化和平均词频的变化是成正比的。

4.3　仅存在于用户标签的高频词

观察标签总体，计算出上述结果里 786 518 个标签词中出现频率比较高的图像标签描述词。这些仅存在于用户标签中的高频词标签总数为 187，出现总频次为 454 855。表 5 是对比用户标签分类框架，深度学习未能标注的标签类别。

表 5　深度学习未能标注标签类别

类别及高频标签数	未能标注标签数	总频次	标签举例
年份(11)	10	76 915	2010
季节(5)	0	0	-
月份(12)	4	6 441	november
一天中的某时段(3)	0	0	-
国家或地名(158)	72	162 830	new+zealand
方位(5)	0	0	-
标志性地点(46)	3	3 224	disneyland
人名(2)	2	2 582	jovens
某一类人(19)	3	4 756	students
活动(50)	13	21 808	Carnival
结合背景知识的活动(7)	3	5 020	cosplay
颜色纹理形状(11)	1	2 125	black+and+white
风景(14)	3	7 233	sunrise
动植物(17)	3	9 181	wildlife
建筑(6)	1	1 549	buildings
对象(36)	1	3 381	clouds
物品(49)	7	10 367	cars
形容词(18)	4	18 739	Beautiful
结合背景知识的形容词(6)	0	0	-
感觉(2)	2	4 365	cold
结合背景知识的情感(2)	1	976	pride
艺术(7)	2	2 250	streetart
历史(2)	1	1 743	History
文化(1)	0	0	-
图像反映的内容(14)	1	986	lo-fi
图像生成设备或参数(28)	26	66 777	nikon
图像处理软件(3)	3	17 215	fireworks
网站(21)	21	24 392	www.500px.com

4.4　含高频词的图像举例分析

本文使用 MXNet 深度学习算法对所有图像进行自动标注，将标注结果与用户标签相比较。其中，含有仅存在用户标签的高频词的图像总数为 577 220。基于上述所分的大类，选取具有代表性的图像进行分析。

如图 2 所示，用户标注的标签为“Disneyland”

“island”“pirates”和“refurb”，而深度学习自动标注的标签仅仅为“Rope bridge”。其中，“Disneyland”为出现在仅存在于用户标签中的高频词。这表明，深度学习标注方法只通过视觉特征识别出图像中的物体，并且这个识别也存在准确性的问题；而用户标签，不仅能标注出图像中的物体，还能分析出这些物体存在于一个什么样的特定场景地点中。

图 2　用户描述场景“迪士尼”

如图 3 所示，用户标注的标签为“cosplay”“costume”“fandemonium”“Idaho”和“nampa”，深度学习自动标注的标签为“Figure skating”“Folk dancer”和“Little theater”。其中，“cosplay”是出现在仅存在于用户标签中的高频词。类似图 2 深度学习标注方法仅识别了图中的对象，而用户从图中对象分析出“角色扮演”这一活动。

图 3　用户分析为“角色扮演”

如图 4 所示，用户标注的标签为“amazing”“aquarium”“asia”“Asian”“beautiful”“beauty”“bentencho”“class”“color”“colorful”“colour”“colourful”“contrast”“feckin”“fish”“japan”“japanese”“oishisou”“orient”“oriental”“osaka”“sashimi”“sushi”和“traditional”，深度学习自动标注的标签为“Business district”“downtown”“City”“metropolis”“urban center”“Hotel-casino”和“casino-hotel”。其中，“beautiful”是出现在仅存在于用户标签中的高频词。深度学习自动标注方法仅识别了图中的对象，而用户在图中视觉对象的基础上添加了描述图像的形容词。

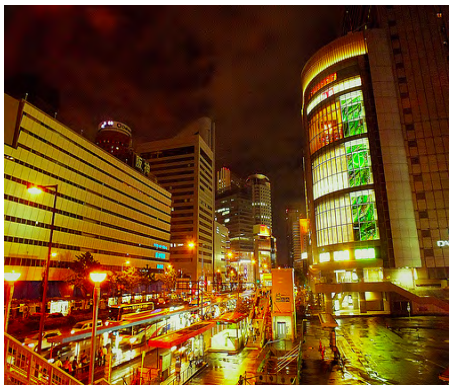

图 4　用户描述为“美”

如图 5 所示，用户标注的标签为“arbol”“blue”“canada”“clear”“cold”“cornell+community+centre+library”“red”“sky”“Toronto”“tree”和“weather”，深度学习自动标注的标签为“Row house”和“town house”。其中，“cold”是出现在仅存在于用户标签中的高频词。深度学习自动标注方法仅识别了图中的对象，而用户在图中视觉对象的基础上添加了感官描述词。

图 5　用户分析为“冷”

如图 6 所示，用户标注的标签为“art”“france”“paris”“street”和“streetart”，深度学习自动标注的标签为“Tattoo”“Sketch”“study”“Nude”和“nude painting”。其

中，“streetart”是出现在仅存在于用户标签中的高频词。深度学习自动标注方法也是仅识别了图中的对象，而用户在图中视觉对象的基础上，通过自身的背景知识，对图像所属的艺术类别进行了描述。

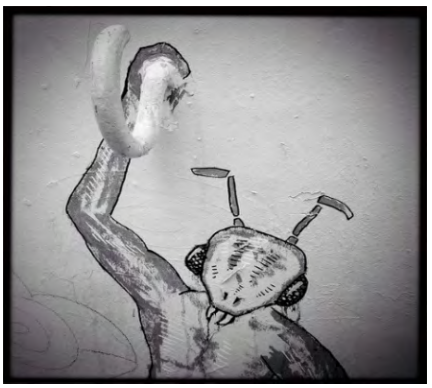

图 6　用户分析为“街头文化”

4.5　讨　论

从当前图像自动标注的研究进展来看，大多数自动标注聚焦于提取图像的视觉信息，利用相似的图像视觉描述子获得相似的图像标签。并且通过不断改进算法，提高标注的精确度。部分图像自动标注方法将语义信息作为提升精度的方向，通常是将语义特征映射到高维向量中，构建语义知识图谱。

从表 5 的实验结果可以看到，在一些方面，用户对图像标注的标签超出当前深度学习自动标注的技术能力范围。以下是对表 5 中深度学习未能标注的高频词进行的总结描述。

(1) 背景知识。例如，艺术、历史和文化类的图像标注，是基于用户已有的关于图像描述的内容的专业知识，而深度学习只是通过视觉和语义信息的相似度进行标注，并不具备图像内容的专业知识，无法进行此类专业性的标注；另外，如表 5 中的“war”一词，是用户基于图像内容对图像所反映的场景和事件的推理描述，深度学习方法中同样不具备此背景知识，也不具备推理能力。

(2) 描述词。此类词大多是用户对图像的直觉感知，如“beautiful”等，是用户在观察图像后，对图像总体或某一方面进行的评价。深度学习方法不具备感知能力，因此无法对图像内容进行准确的评价。

(3) 情感词。类似于描述词，一方面是对图像表达的情感的描述，另一方面是用户看到图像后产生的情感的描述。同上述，深度学习方法不具备感知能力，无法对图像进行情感判断。

(4) 图片来源网站。从表 5 可知，用户倾向于用一些图像来源的网站名或网址标注图像。这些关于图片的来源，深度学习方法也是无法识别出来的。

(5) 图像处理软件。例如表 5 中的“fireworks”，用户在上传图像之前，会使用一些图像处理软件对图像进行调整，因此标注中也标注了处理图像的软件名。深度学习方法具有能够识别出图像是否经过修改，但不能还原图像本身。

(6) 拍摄设备。例如表 5 中的“Nikon”是相机的品牌。类似的还有“80mm”“f3.5”等拍照时的焦距、曝光度的设置。这类词在 ImageNet 数据库中没有出现。但是这类标签词信息属于照片的元数据，计算机能够读取照片的元数据，这是可以标注的。

(7) 国家地名等。同上述，ImageNet 数据库中并没有详细的每一个国家地名的词，而这些地理信息也存在于照片的元数据中，因此同样能够被深度学习方法标注。

4.6　图像自动标注未来发展方向

针对分析得到的图像自动标注不足之处，本文对利用深度学习进行图像自动标注未来发展提出以下建议：

(1) 关于背景知识。图像内容的背景知识往往需要专业的知识库和判断，因此对于这一层面，可以考虑使用大量包含人工标注背景知识(如标签、与图像关联的文章等)的图像进行训练。不同领域的图像使用不同种类的训练数据，从而不同领域的图像有着不同的标签模型。此外，也可对不同的图像标签生成不同的背景知识主题，利用图像信息相似度为图像匹配不同的背景知识主题。

(2) 对于描述词和情感词类，基于上述的背景知识训练模型、主题模型，可以采用自然语言处理技术，做到“图像信息”、“背景知识”与情感、描述标签的联结。此外，通过自然语言处理的相关技术，还可以做到相似标签的衍生。

总的来说，计算机不具备推理能力、感知能力，无法像人类大脑一样对事物进行理解，无法在未经训练的情况下对标签语境、背景知识以及情感描述等进行推理判断。如何通过模仿人类大脑，即“仿生”和认知

计算方向，来解决这些逻辑推理以及情境感知的问题，将是未来需要努力的方向。

5 结 语

本文通过总结当前图像自动标注的发展现状，提出图像用户标注分类框架。通过对 Flickr 的上百万张用户已标注的图像数据集的统计分析，对比深度学习图像自动标注标签与图像用户标注标签的差异，并从不同角度总结深度学习图像自动标注技术的不足之处，在此基础上，对未来图像自动标注的发展方向进行展望。在未来的研究中，还需致力于解决本文中提到的问题，即在深度学习图像自动标注中，实现对图像信息与图像背景知识和图像的情感、描述的关联。并探索应该如何实现赋予计算机推理和感知的能力。

参考文献：

[1] Leung C H C, Luo M Q. Building Up of Image and Multimedia Object Index Through Continuous Usage[C]// Proceedings of International Conference on Computer Networks, E-Learning and Information Technology, Bangkok, Thailand. HongKong: ICCNEIT, 2013.

[2] Sill L A. Indexing Multimedia and Creative Works: The Problems of Meaning and Interpretation [J]. Library Collections, Acquisitions, and Technical Services, 2005, 29(4): 448-449.

[3] Beaudoin J. Folksonomies: Flickr Image Tagging: Patterns Made Visible[J]. Bulletin of the American Society for Information Science & Technology, 2007, 34(1): 26-29.

[4] Golbeck J, Koepfler J, Emmerling B. An Experimental Study of Social Tagging Behavior and Image Content[J]. Journal of the Association for Information Science & Technology, 2011, 62(9): 1750-1760.

[5] Klavans J L, Laplante R, Golbeck J. Subject Matter Categorization of Tags Applied to Digital Images from Art Museums[J]. Journal of the Association for Information Science & Technology, 2014, 65(1): 3-12.

[6] Xie L, Natsev A, Hill M, et al. The Accuracy and Value of Machine-generated Image Tags: Design and User Evaluation of an End-to-End Image Tagging System[C]//Proceedings of ACM International Conference on Image & Video Retrieval. 2010: 58-65.

[7] Ordonez V, Kulkarni G, Berg T L. Im2text: Describing Images Using 1 Million Captioned Photographs[C]// Proceedings of Conference on Neural Information Processing Systems.2011: 1143-1151.

[8] Lee S, De Neve W, Ro Y M. Image Tag Refinement along the 'What' Dimension Using Tag Categorization and Neighbor Voting [C]//Proceedings of 2010 IEEE International Conference on Multimedia & Expo.2010: 48-53.

[9] Izadinia H, Farhadi A, Hertzmann A, et al. Image Classification and Retrieval from User-Supplied Tags[OL]. arXiv Preprint. arXiv: 1411.6909.

[10] Eleta I, Golbeck J. A Study of Multilingual Social Tagging of Art Images: Cultural Bridges and Diversity[C]//Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, Seattle, Washington, USA. New York, USA: ACM, 2012: 695-704.

[11] Cunningham S J, Bainbridge D, Masoodian M. How People Describe Their Image Information Needs: A Grounded Theory Analysis of Visual Arts Queries[C]//Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries.2004: 47-48.

[12] 王晓光，徐雷，李纲.敦煌壁画数字图像语义描述方法研究[J].中国图书馆学报，2014，40(1)：50-59.(Wang Xiaoguang, Xu Lei, Li Gang. Semantic Description Framework Research on Dunhuang Fresco Digital Image[J]. Journal of Library Science in China, 2014, 40(1): 50-59.)

[13] Zhang J, Yang Y, Tian Q, et al. Personalized Social Image Recommendation Method Based on User-Image-Tag Model[J].IEEE Transactions on Multimedia, 2017, 19(11): 2439-2449.

[14] Sa N, Yuan X. What Motivates People Use Social Tagging[A]// Lecture Notes in Computer Science[M]. 2013, 8029: 86-93.

[15] Heckner M, Heilemann M, Wolff C. Personal Information Management vs. Resource Sharing: Towards a Model of Information Behavior in Social Tagging Systems[C]// Proceedings of International Conference on Weblogs and Social Media(ICWSM 2009), San Jose, California, USA. 2009.

[16] Nov O, Ye C. Why do People Tag? Motivations for Photo Tagging[J]. Communications of the ACM, 2010, 53(7): 128-131.

[17] Ames M, Naaman M. Why We Tag: Motivations for Annotation in Mobile and Online Media[C]// Proceedings of the SIGCHI Conference on Human Factors in Computing Systems(CHI 2007), San Jose, California, USA. 2007: 971-980.

[18] Nwana A O, Chen T. Who Ordered This?: Exploiting Implicit User Tag Order Preferences for Personalized Image Tagging[C]// Proceedings of the IEEE International Conference on Multimedia & Expo Workshops. 2016: 1-6.

[19] Strohmaier M, Körner C, Kern R. Why do Users Tag? Detecting Users' Motivation for Tagging in Social Tagging Systems[C]//Proceedings of International Conference on Weblogs and Social Media(ICWSM 2010), Washington, DC, USA. 2010: 23-26.

[20] Patel T, Shah B. A Survey on Facial Feature Extraction Techniques for Automatic Face Annotation[C]//Proceedings of 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA).2017: 224-228.

[21] Hao Z, Ge H, Gu T. Automatic Image Annotation Based on Particle Swarm Optimization and Support Vector Clustering[J]. Mathematical Problems in Engineering, 2017(1): 1-11.

[22] Ke X, Zhou M, Niu Y, et al. Data Equilibrium Based Automatic Image Annotation by Fusing Deep Model and Semantic Propagation[J]. Pattern Recognition, 2017, 71: 60-77.

[23] Gu Y, Xue H, Yang J. Cross-Modal Saliency Correlation for Image Annotation[J]. Neural Processing Letters, 2017, 45(3): 777-789.

[24] Bahrololoum A, Nezamabadi-Pour H. A Multi-expert Based Framework for Automatic Image Annotation[J]. Pattern Recognition, 2017, 61: 169-184.

[25] Budikova P, Batko M, Zezula P. ConceptRank for Search-based Image Annotation[J]. Multimedia Tools and Applications, 2018, 77(7): 8847-8882.

[26] Uricchio T, Ballan L, Seidenari L, et al. Automatic Image Annotation via Label Transfer in the Semantic Space[J]. Pattern Recognition, 2017, 71: 144-157.

[27] Mehmood Z, Mahmood T, Javid M A. Content-based Image Retrieval and Semantic Automatic Image Annotation Based on the Weighted Average of Triangular Histograms Using Support Vector Machine[J]. Applied Intelligence, 2017(1): 1-16.

[28] Tariq A, Foroosh H. Learning Semantics for Image Annotation[OL]. arXiv Preprint, arXiv: 1705.05102.

[29] Chien B C, Ku C W. Large-scale Image Annotation with Image-text Hybrid Learning Models[J]. Soft Computing, 2017, 21(11): 2857-2869.

[30] Verma Y, Jawahar C V. Image Annotation by Propagating Labels from Semantic Neighbourhoods[J]. International Journal of Computer Vision, 2017, 121(1): 126-148.

[31] Tariq A, Foroosh H. A Context-driven Extractive Framework for Generating Realistic Image Descriptions[J]. IEEE Transactions on Image Processing, 2017, 26(2): 619-632.

[32] Karpathy A, Li F F. Deep Visual-Semantic Alignments for Generating Image Descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 664-676.

[33] Oquab M, Bottou L, Laptev I, et al. Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2014: 1717-1724.

[34] Gong Y, Jia Y, Leung T, et al. Deep Convolutional Ranking for Multilabel Image Annotation[OL]. arXiv Preprint, arXiv: 1312.4894.

[35] Sánchez J, Perronnin F, Mensink T, et al. Image Classification with the Fisher Vector: Theory and Practice[J]. International Journal of Computer Vision, 2013, 105(3): 222-245.

[36] Tian J, Huang Y, Guo Z, et al. A Multi-Modal Topic Model for Image Annotation Using Text Analysis[J]. IEEE Signal Processing Letters, 2014, 22(7): 886-890.

[37] Yan F, Mikolajczyk K. Deep Correlation for Matching Images and Text[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015: 3441-3450.

[38] Gu Y, Qian X, Li Q, et al. Image Annotation by Latent Community Detection and Multikernel Learning[J]. IEEE Transactions on Image Processing, 2015, 24(11): 3450-3463.

[39] Thomee B, Shamma D A, Friedland G, et al. YFCC100M: The New Data in Multimedia Research[J]. Communnications of the ACM, 2016, 59(2): 64-73.

作者贡献声明:

陆伟: 提出研究思路, 设计研究方案;
罗梦奇: 设计研究方案, 进行实验, 分析数据, 论文起草和修订;
丁恒: 采集数据, 提供实验辅助;
李信: 论文修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: lakeygtgz@163.com。

[1] 罗梦奇.usertags.csv. 用户高频标签分类.

[2] 罗梦奇.onlyinusertags.csv. 深度学习未能标注标签分类.

收稿日期: 2018-01-15
收修改稿日期: 2018-01-30

Image Annotation Tags by Deep Learning and Real Users: A Comparative Study

Lu Wei Luo Mengqi Ding Heng Li Xin
(School of Information Management, Wuhan University, Wuhan 430072, China)
(Information Retrieval and Knowledge Mining Laboratory, Wuhan University, Wuhan 430072, China)

Abstract: [**Objective**] This paper proposes a user tagging framework and examines the limitations of tagging image with deep learning techniques, aiming to improve the performance of automatic annotation services. [**Methods**] We analyzed the user-added tags from one million images on flickr.com to extract the high frequency ones. Then, we mapped these tags with the proposed framework, and compared them with tags from the ImageNet database. Finally, we analyzed images with high frequency tags with the deep learning algorithm - MXNet. [**Results**] The automatic image annotation techniques based on deep learning could not effectively understand the image's background knowledge, as well as the image's descriptions from the human perceptive. [**Limitations**] Our dataset needs to be expanded and analyzed with other deep learning algorithms. [**Conclusions**] The development of automatic image annotation, requires us to establish the association between image information, background knowledge, and description, as well as cultivate deductive reasoning and context-aware abilities.

Keywords: Image Annotation User Tags Automatic Image Annotation Machine Learning Deep Learning Artificial Intelligence