DOI:10.13878/j.cnki.jnuist.2017.06.001

俞俊[1,2]　谭敏[1,2]　张宏源[1,2]　张海超[1,2]

基于用户点击数据的细粒度图像识别方法概述

摘要

近年来,细粒度图像识别逐渐成为计算机视觉领域的研究热点.由于不同类别图像间的视觉差异小、语义鸿沟问题严重,传统的基于视觉特征的细粒度图像识别性能往往不尽人意.针对这些挑战,目前许多学者都在研究基于用户点击数据的图像识别.本文围绕点击数据在图像识别中数据预处理、特征提取和模型构建3大模块中的应用,总结了已有的基于点击数据的识别算法及最新的研究进展.

关键词

用户点击;图像识别;度量学习;深度学习;语义鸿沟

中图分类号 TP391.413

文献标志码 A

收稿日期 2017-07-28

资助项目 国家自然科学基金优秀青年基金(61622205);国家自然科学基金青年基金(61602136)

作者简介

俞俊,男,博士,教授,研究方向为机器学习、多媒体分析与图像处理.yujun@hdu.edu.cn

谭敏(通信作者),女,博士,讲师,主要研究方向为人工智能、计算机视觉与机器学习.tanmin@hdu.edu.cn

1 杭州电子科技大学 计算机学院,杭州,310018

2 杭州电子科技大学 复杂系统建模与仿真教育部B类重点实验室,杭州,310018

1 引言

1.1 背景

细粒度视觉分类(Fine-Grained Visual Categorization,FGVC)是目标分类的一个子领域.与Pascal VOC竞赛[1]等对船、自行车和汽车进行分类的任务不同,细粒度分类是对于视觉上非常相似的目标进行区分的过程,如鸟、狗、花的种类等,这些子类图像在视觉上差距甚小.

传统的图像识别技术大多借助于视觉特征,如颜色、纹理、形状、轮廓等.然而,图像的视觉特征仅能刻画视觉信息,忽略了它们所包含的语义信息,与人类对图像的理解存在一定的差异.这种在计算机图像理解与人类图像理解之间存在着的客观区别,即图像低层视觉特征与高层语义特征之间存在着的较大距离,被称为“语义鸿沟”[2-4].

计算机视觉和人类视觉的“语义鸿沟”使得人们在图像识别领域一直面临巨大挑战,尤其是对于细粒度的图像识别而言.近年来,许多从事图像视觉研究的人员已经逐渐认识到语义信息在图像理解中的重要性,并在图像识别的过程中引入了用户点击数据表征图像的语义特征从而解决“语义鸿沟”问题.

1.2 点击数据

点击数据是依托搜索引擎(如Google、百度、Bing等)收集的用户对图像与文本间相关性的反馈数据.如图1所示[5],针对任意查询文本,搜索引擎会检索到一组可能相关的图像集,用户会基于查询文本与候选图像的相关性点击更为“相关”的图像,从而产生大量点击数据.利用点击数据,查询文本被图像集表征.类似地,任一图像也可以被其对应的点击文本集合表示.

目前,点击数据已被广泛应用在网页检索、商品推荐等领域,它在图像识别领域中的应用还相对较少[6-10].如图2所示,在基于点击数据的图像识别中,输入的样本除图像本身 x 外,还有其对应的在文本 q 下的点击次数向量.图像识别大多是通过融合图像视觉与点击特征实现的.

近年来,世界各地的研究人员根据用户点击数据设计模型、计算新数据被点击的概率,以此更新该网页放置在返回结果中的位置.微软亚洲研究院[11-12]、谷歌研究院[13]、雅虎研究院[14]等机构在用户点击数据方面均做了深入的研究.其中典型的代表是微软亚洲研究院根据点击数据建立了一个基于点击数据的数据集——Clickture[5],该数



a. “backsplash designs”



b. “bunk beds”



c. “lifted trucks”

图 1　查询文本下点击的图像集

Fig. 1　The clicked image set under one query

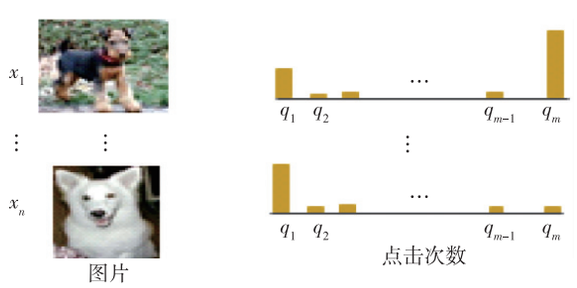


图 2　图像-文本点击次数矩阵

Fig. 2　Click count matrix for image-query clicks

据集的定义为

$$C_{\text{lickture}}=\{\langle x,q,c\rangle\} \tag{1}$$

每一个三元组表示在文本 q 下,图片 x 被点击了 c 次.本文中提到的相关方法均是围绕 Clickture 数据集展开的.

本文将总结现有的基于点击特征的图像识别技术,并详细介绍点击数据在数据预处理、特征提取、模型构建中的应用.

2　数据预处理

尽管点击数据在图像识别中能提供丰富的语义信息帮助解决语义鸿沟,但它本身的噪声信息也会给识别带来很大的负作用.因而,在基于点击数据的识别中,点击数据的预处理是一个关键步骤.

点击数据中的噪声包括 3 个方面.一是查询文本可能存在的拼写错误;二是图片本身质量过低、目标不明显、图像重复等;三是文本-图像点击的缺失和不一致性.

直接使用这些“脏”的点击数据必然会影响后续的图像识别.因此,对数据进行预处理的重要性不言而喻.针对这些有噪声的点击数据,处理方法一般分为两类:一类是通过数据清洗去除一定量的噪声样本,另一类是通过可靠性建模赋予样本权重从而达到对样本去噪的目的.

2.1　数据清洗

数据清洗作为数据预处理的关键一步,是指通过一定的手段对原始数据进行删除等处理,从而提高数据的有效性,进而提高图像识别的精度.

用户点击数据主要包含 3 个部分,即:查询文本、图片以及对应的图片-文本点击次数.利用用户的点击数据对图片数据进行数据清洗一般分为 3 种方法:基于点击次数先验的清洗、基于文本-图片相似度的清洗以及融合视觉检测器的数据清洗,3 种方

法分别利用了点击数据中的一个或者多个部分.

2.1.1 基于点击次数先验的清洗

一种最简单的数据清洗方法就是直接剔除那些点击次数少的数据[3],然而这种方法过于启发式,并不一定可靠.在一些数据中,点击次数少的数据可能比点击多的噪声更加珍贵.例如在一个被点击了3次的图片中,文本"狗(dog)"被点击了2次,而文本"吉娃娃(chihuahua)"被点击了1次,很显然对于细粒度分类任务而言,"吉娃娃(chihuahua)"的重要性大于"狗(dog)",而若是按照点击次数清洗数据的方法则很可能只会留下文本数据"狗(dog)".因此只按照点击次数来清洗数据的方法并不是一个好的解决方案.

2.1.2 基于文本-图片相似度的清洗

基于文本-图片相似度的清洗方法的主要思想是利用点击数据先学习一个图像视觉与文本特征之间的相关性模型,然后利用此模型剔除掉点击数据中"图"-"文"特征相似度过低的数据项.目前基于该思想的数据清洗方法相对较少.近年来,Bai 等[15]提出了用深度学习框架融合词嵌入的方法学习视觉-文本相似度模型,取得了很好的效果.

2.1.3 融合视觉检测器的数据清洗

如前文所述,一些研究人员利用额外的带标签在图像训练集学习视觉目标检测器,然后用图片检测器对点击数据集进行清洗,然而标注数据往往依赖大量的人力、物力和财力,因此并不是一个实用的方法.本节介绍的融合视觉检测器的数据清洗方法,同时利用了点击数据中的图-文点击次数和图片的视觉特征,在不使用额外的数据集的条件下构建了图片的检测器.

融合视觉检测器的数据清洗主要分为3步:第1步是基于图片的点击次数选择相对可靠的图像构成训练集;第2步是使用挑选出来的图片集训练出一个基于视觉特征的图片检测器;第3步是同时考虑图片的可靠性和视觉特征,筛选出视觉检测器认为概率相对较高且点击次数较大的图片,同时清洗(剔除)掉剩下的图片.

2.1.4 小结

由于点击数据的高噪声性,目前基于点击数据的图像识别很大一部分的工作内容都集中在数据清洗上.其中基于点击次数的筛选最为直接,但它过于启发式容易误筛选掉正常样本;基于文本-图片相似度的清洗方式最为合理,但由于涉及多模态特征空间的相似度模型的构建使得算法复杂度过高;同样,融合视觉检测器的数据清洗的方式也涉及繁琐的模型训练过程,但与文本-图片相似度模型相比,由于不涉及跨模态建模,它训练的视觉模型复杂度相对较低,可是模型的单一性却影响了它的清洗有效性.

2.2 数据可靠性建模

在噪声数据处理中,除了直接将噪声数据剔除外,还有一种常用的策略是对数据进行加权,使得噪声数据在识别模型学习和决策中权重相对较低.为估计样本权重,一种常见的方法是构建样本可靠性模型,从而用可靠性值来量化样本权重.

已有的样本可靠性模型大致分为两类:一是直接利用样本的质量特征(模糊程度、角度、目标区域位置等)来量化.Zheng 等[7]提出图像的用户点击量能在一定程度上反映图像的质量,因此他们利用用户点击量来估计样本可靠性.另一类则是利用二类分类模型(如可支持向量机(Support Vector Machine,SVM)[16]等)训练一个可靠性模型,再利用可靠性分类器的输出值给样本加权.Tan 等[17]提出构建样本的"可靠性特征",并基于此特征用 SVM 分类模型学习可靠性分类器.其中"可靠性特征"由样本质量、样本分类正确的概率及其分到各类概率的信息熵等构成.在训练可靠性分类器时,被分类正确/错误的样本被视为正/负样本.

在目前的研究中,利用点击数据构建样本可靠性模型的研究还比较稀少.直观来讲,作为用户反馈数据,用户点击能很大程度上反映样本可靠性.因而,利用点击数据构建可靠性模型将是一个很有潜力的解决方案.最近,Zheng 等[7]提出利用点击次数数据作为样本权重先验,并基于点击次数对权重模型进行光滑性建模.同时他们构建了如下的深度学习框架联合优化样本的深度视觉特征和权重模型:

$$\begin{aligned}(\boldsymbol{\theta}^*,\boldsymbol{w}^*)=&\arg\min_{\boldsymbol{\theta},\boldsymbol{w}}\frac{1}{2}\|\boldsymbol{\theta}\|_2^2+\frac{C}{n}\sum_{i=1}^{n}w_i l(\boldsymbol{y}_i,\boldsymbol{o}_i)+\\&\alpha P(\boldsymbol{w})+\beta S(\boldsymbol{G},\boldsymbol{w}),\\ \text{s.t.}\begin{cases}\sum_{i=1}^{n}w_i=n,\\ l(\boldsymbol{y}_i,\boldsymbol{o}_i)=-\log\left(\mathrm{e}^{o_{y_i}}/\sum_{j=1}^{N}\mathrm{e}^{o_j}\right),\\ w_i>0,\forall i,\end{cases}\end{aligned}\tag{2}$$

其中 $\boldsymbol{w}_i$ 是样本 i 的可靠性权重,$l(\boldsymbol{y}_i,\boldsymbol{o}_i)$ 是第 i 个样本的 softmax 损失函数,$\boldsymbol{o}_i$ 是视觉特征和语义特征融合后得到的特征,$\boldsymbol{\theta}$ 是深度视觉模型的参数,C、α、β

分别是损失项、先验项和平滑项的参数.

权重先验项 $P(\boldsymbol{w})$ 用来构建点击先验模型.它的定义公式如下:

$$P(\boldsymbol{w}) = \| \boldsymbol{w} - \bar{\boldsymbol{w}}^c \|_2^2, \tag{3}$$

其中

$$\bar{\boldsymbol{w}}^c = \boldsymbol{w}^c / \| \bar{\boldsymbol{w}}^c \|, \quad \boldsymbol{w}^c = T(\boldsymbol{u}), \tag{4}$$

$T(\boldsymbol{u})$ 是用来控制 $\boldsymbol{u}$ 的尺度变换函数,其作用是解决图像点击次数极度不均衡的问题.

$S(\boldsymbol{G},\boldsymbol{w})$ 是基于相似性邻接图 $\boldsymbol{G}$ 构建的平滑项公式,可使相似的图片拥有相似的权重.图片的相似度用它们之间深度特征 $\boldsymbol{z}$ 的距离来衡量:

$$\begin{cases} S(\boldsymbol{G},\boldsymbol{w}) = \sum\limits_{\forall i,j \in \chi_k} g_{i,j}(w_i - w_j)^2/2, \\ g_{i,j} = \mathrm{sim}(\boldsymbol{x}_i,\boldsymbol{x}_j) = \exp(- \| \boldsymbol{z}_i - \boldsymbol{z}_j \|). \end{cases} \tag{5}$$

3 特征提取

鉴于点击信息的强语义性,点击数据除了能有效运用于数据预处理外,也被广泛用于图像特征的提取,包括单一点击特征的构建及其与传统视觉特征的融合.

3.1 点击特征

在现存的利用点击数据构建图像特征的方法中,图像往往被表征成它在查询文本空间下的点击次数向量.由于点击数据中的文本是由一个或多个单词构成的,因此利用点击数据的特征构建通常分为两类,分别是基于查询文本(即原始的文本空间)和基于查询关键词(即查询文本中的独立单词)的构建方法.下面将分别介绍这 2 种特征构建的方法.

3.1.1 基于查询文本

采用 Clickture 数据集标准的表示方法表征点击数据,可知基于查询文本的点击次数向量的核心问题在于查询文本的数量巨大,使得用户点击特征过于稀疏、维度过高.针对这些问题,一些研究者致力于查询文本合并的研究,其核心就是文本聚类.

传统的文本合并是基于查询本身的文本特征展开的.最近,Wu 等[8]提出了基于点击数据的文本合并方法.该方法首先将文本表征为图像点击向量(点击的图像次数向量),再利用稀疏编码的技术实现分类.其中,为了解决原始图像点击特征稀疏与不光滑的特性,他们提出了利用基于图的相似度的点击传播模型,使得传播后的点击特征相对稠密且具有视觉相似一致性;在文本被表征为更有效的传播点击特征后,由于点击数据的类间极其不均衡的性质,他们又提出了基于稀疏编码的聚类.稀疏编码的字典通过热门词汇“Hot-query”构建,模型框架如图 3 所示.

实验表明,基于编码稀疏的聚类方法在处理这种极度不平衡的点击数据上的结果优于使用传统的基于 K 均值的聚类算法的结果;同时,基于热门词汇的字典构建也优于基于 K-SVD(K-Singular Value Decomposition,K-SVD)这种传统字典学习方法.

为了进一步提高文本特征的表征能力,笔者正在尝试利用深度学习模型构建深度文本特征,进而利用深度文本特征实现文本聚类.然而,点击数据的过度稀疏是深度文本模型中亟待解决的一个问题.

3.1.2 基于查询关键词

图片的点击文本,如“chihuahua with soda”是由一个一个单词构成的,在基于查询关键词的特征构建过程中,常用的方法是首先对每一个词组进行分词和词性还原操作,并在得到的单词集合里去掉标准的停顿词,进而得到处理后的点击数据,即单词、图片和对应的图片/单词点击次数(词频矩阵).

通过词频矩阵,将图片集看成一个文档,每张图片作为文档的一个段落.利用 tf-idf 算法,可以将每张图片表示为一组与词频相关的向量.tf-idf 算法是目前最常用的特征权重算法,该算法由 Salton 等[18]

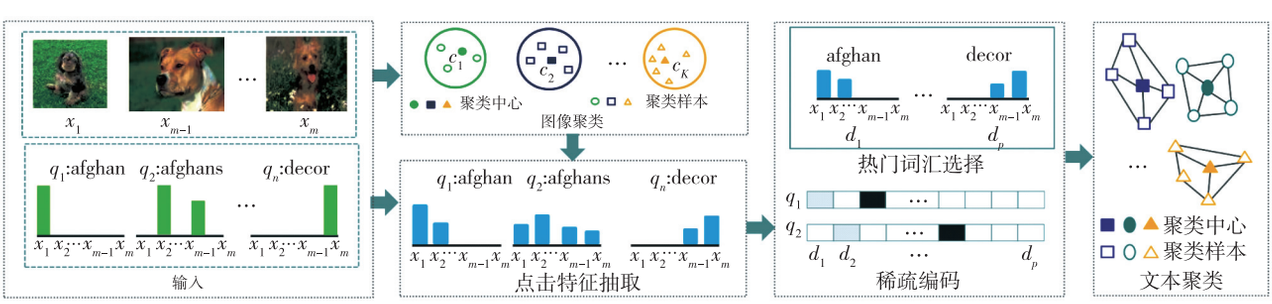


图 3 文本合并过程

Fig. 3 Query modeling pipeline

提出.它由2部分组成:1)基于文档内容的词语频率(tf),即词语在当前文档中出现的次数;2)基于文档空间的文档频率(df),即在文档空间中出现过该词语的文档数.词语频率体现了特征对当前文档的表现力,词频越高,越能表示文档的内容,对文档的表现力越强.文档频率体现了特征对文档的区分力,在越多的文档中出现的特征,对文档的区分力越弱.特征的区分力与文档频率成反比,因此在计算时采用的是逆文档频率(idf).tf-idf的经典计算公式为

$$W(t_i,d_j)=f_{\mathrm{t},ij}\times f_{\mathrm{id},i}=f_{\mathrm{t},ij}\times\log\frac{D}{d(t_i)},\qquad(6)$$

其中$f_{\mathrm{t},ij}$为特征词t_i在文档d_j中的词频;$f_{\mathrm{id},i}$为t_i的逆文档频率;D为文档空间中的文档总数;$d(t_i)$为出现t_i的文档数,即t_i的文档频率.

3.2 融合的点击与视觉特征

鉴于卷积神经网络在视觉特征提取方面的优势[19-20],近年来,一些研究人员开始使用卷积神经网络提取出的特征与点击特征进行融合,从而提高图像检索的性能[15,21].

类似的,在图像识别领域,也可采取同样的手段,将图像的点击特征和视觉特征融合,以增强图像特征的区分能力.如何将图像的点击特征和图像的视觉特征融合在一起也是一个新的挑战,本节将介绍2种将点击特征和视觉特征融合的方法:直接融合方法和词嵌入方法.

3.2.1 直接融合

将由卷积神经网络特征提取的图像视觉特征(一般用的是卷积神经网络的某一层全连接层)与图像的点击特征直接融合指的是将2个特征向量直接拼在一起作为融合后的特征向量.例如,若图像的视觉特征是4 096维的向量,图像的点击特征是1 000维的向量,那么融合后的向量为5 096维.即融合图像特征表示为$\boldsymbol{o}_i=[\boldsymbol{z}_i,\mu\boldsymbol{u}_i]$,其中$\boldsymbol{z}_i$和$\boldsymbol{u}_i$分别代表了第$i$张图像的视觉特征和点击特征,$\mu$代表特征权重,同时为了保证2个特征的尺度相同,在拼接前应该要对视觉特征和点击特征做标准化操作[6].

3.2.2 词嵌入

由于图像的视觉特征和图像的点击特征并非在一个特征空间上,将两者直接拼接在一起显然并不是一个合理的方法,因此将2个特征转换到同一空间再进行拼接是有必要的.

词嵌入指的是对于给定的文档,将文档中的每一个单词转为对应的向量表示.传统的词嵌入模型有one-hot模型、向量空间模型、word2vec模型[22]等.融合方法里的词嵌入,指的是在得到图像的视觉特征后,通过线性或者非线性(如sigmod、relu等)的转换,使得其投影到点击特征空间中,再将两者进行拼接.因为通过变换,两者已经在同一个特征空间中,拼接的操作也显得合理而有效.

3.3 小结

本节介绍了基于查询文本和基于查询关键词2种构建图像点击特征的方法.比较而言,基于文本的点击特征构建方式直观有效,但它涉及复杂的文本合并过程,而基于查询词构建的点击特征更紧凑,大大提高了算法效率.除了单一点击特征外,本节还介绍了它与视觉特征的2类融合方式,其中以词嵌入方法融合的结果较好.

4 分类模型构建

特征提取完成后,接下来就是针对特征构建分类器的过程.本文总结的分类模型主要针对融合视觉特征与点击特征的分类模型构建.

4.1 度量学习

由于视觉与点击特征在不同的子空间中,因而需要构建深度学习模型为融合特征学习可靠的度量空间.一个度量是一个定义集合中元素之间距离的函数[23].在度量确定后,则可基于新的距离度量,通过在训练样本空间中的KNN搜索实现分类.

与传统特征相比,在基于点击特征的度量学习的分类算法中至少存在2大难点:一是点击数据中的强大噪声可能影响度量学习的性能;二是点击特征维度过高,往往导致基于样本空间的搜索效率过低.

针对样本噪声,谭敏[24]提出的弱监督度量学习算法可以帮助在带噪声的样本中自动筛选相对“干净”的数据学习度量;同时,为了克服点击特征匹配效率低的缺陷,Tan等[25]还提出了联合度量及模板学习的算法,通过学习判别性强的模板实现在模板中的1-NN搜索的分类.

4.2 深度学习

深度学习[26]的概念最早由多伦多大学Hinton等于2006年提出,它利用训练样本通过一定的训练方法得到包含多层的深度网络结构,然后学习图像的深度视觉特征.深度学习中的“深度”是相对SVM、提升方法(boosting)、最大熵方法等浅层学习方法而

言的,深度学习所学得的模型中,非线性操作更多,学到的特征表征力、不变性更强.浅层学习依靠人工经验抽取样本特征,获得的是没有层次结构的单层特征;而深度学习通过对原始信号进行逐层特征变换,将样本在原空间的特征表示变换到新的特征空间中,自动地学习得到层次化的特征表示,从而更有利于分类或特征的可视化[27].

本节主要介绍用户点击数据在卷积神经网络和双线性差值卷积神经网络的应用.

4.2.1 卷积神经网络

卷积神经网络(Convolutional Neural Network,CNN)是一种前馈神经网络,它的人工神经元可以响应一部分覆盖范围内的周围单元.在图像识别任务中,传统的卷积神经网络优势十分突出.其工作原理是对输入的图片,经过卷积层、下采样、全连接等模型层后得到图像的深度视觉特征.如图 4 所示.

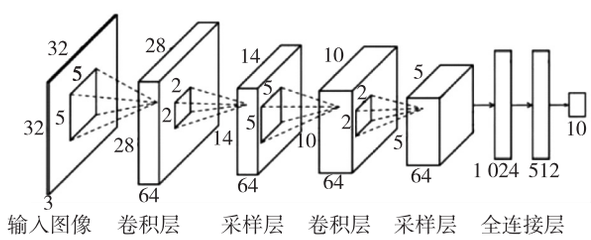


图 4 卷积神经网络结构

Fig. 4 Structure of convolutional neural network

通过深度模型提取出的点击特征可以直接用于图像分类工作,也可以和其他特征进行融合后完成图像识别.在构建融合用户点击数据的卷积神经网络时,将最后全连接层提取出来的图片视觉特征,与点击特征融合(参考第 3.2 节,融合的点击与视觉特征).最后通过融合后的特征做图像识别[6].融合特征如图 5 所示.

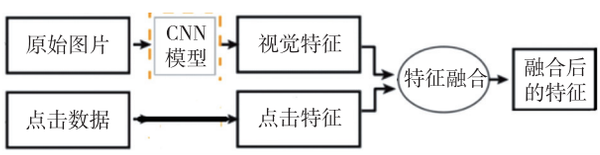


图 5 引入点击数据的卷积神经网络流程

Fig. 5 Pipeline of convolutional neural network with user click data

4.2.2 双线性插值卷积神经网络

双线性插值卷积神经网络(Bilinear Convolutional Neural Network,BCNN)[28],是将输入图片通过 2 个卷积神经网络提取出 2 个视觉特征,再将 2 个视觉特征向量通过双线性插值(Bilinear)的方式组合在一起作为最后的双线性插值图像特征,再用该图像特征进行图像识别任务[8].识别过程如图 6 所示.

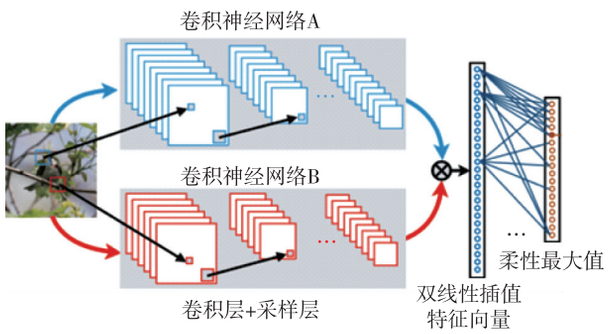


图 6 基于双线性插值卷积神经网络的图像识别

Fig. 6 Image recognition via bilinear convolutional neural network

鉴于 BCNN 模型的优势,Zheng 等[7]提出了融合点击特征的 BCNN 模型,并在细粒度分类上达到了很好的性能.如图 7 所示(图 7a 为传统的 BCNN 网络,图 7b 为引入了用户点击数据的 BCNN 网络),采用用户点击数据的双线性插值卷积神经网络在构建时,与传统的双线性插值卷积神经网络最大的区别是,在得到图片的双线性插值特征后,将与用户的点击特征融合(参考第 3.2 节,融合的点击与视觉特征)来作为图像的最终特征.同时为了保证双线性插值特征与用户点击特征尺度相同,在双线性插值特征后加入了 L2 正则化操作[7].

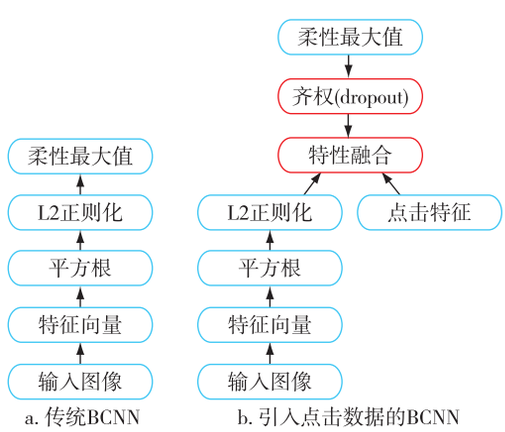


图 7 引入点击数据的卷积神经网络流程

Fig. 7 Framework of bilinear convolutional neural network with user click data

5 总结

本文介绍了用户点击数据在图像识别领域的相关研究和成果.第 2 章主要介绍了用户点击数据在数据预处理方面的工作,包括了清洗数据的 3 种方法以及数据可靠性建模的相关知识.第 3 章以特征提取为主题,详细讲述了利用点击数据,构建单一点击特征和融合点击与视觉特征的方法.第 4 章针对融合的点击特征,介绍了基于度量学习和深度模型框架的分类方法.

总体而言,现存的基于点击数据的图像识别工作相对较少,点击数据的高噪声量是影响其发展的一个主要因素.在不久的将来,此领域中仍有许多非常值得研究的问题,如:

1) 弱监督深度学习.数据集的标签是不可靠的,如针对图像数据 x,它的点击信息(点击文本集及其对应的点击次数)很可能是不可靠的;此外,数据的类别标签也可能存在大量噪声,因此弱监督的学习模型在基于户点击数据的深度学习方面潜力巨大.

2) 迁移学习.本文中介绍的方法都是针对数据的点击信息已知的情况.然而,在大多数分类任务中,图像的点击信息是没有标注的,因而,利用迁移学习实现对没有点击标注的数据集的分类将是很重要的方向.

3) 基于深度学习的文本特征构建.目前用来表征文本都是扁平的一维特征向量,如何利用深度模型框架,构建结构化的深度文本特征模型也具有重大研究意义.

参考文献

References

[1] The pascal visual object classes homepage[EB/OL].[2017-07-28].http://host.robots.ox.ac.uk/pascal/VOC

[2] 朱蓉.基于语义信息的图像理解关键问题研究[J].计算机应用研究,2009,26(4):1234-1240
ZHU Rong. Research on key problems of image understanding based on semantic information[J]. Application Research of Computers,2009,26(4):1234-1240

[3] Yu J,Yang X K,Gao F,et al.Deep multimodal distance metric learning using click constraints for image ranking[J].IEEE Transactions on Cybernetics,2016,PP(99):1-11

[4] Yu J,Tao D C,Wang M,et al.Learning to rank using user clicks and visual features for image retrieval[J].IEEE Transactions on Cybernetics,2015,45(4):767-779

[5] Hua X S,Yang L J,Wang J D,et al.Clickage:Towards bridging semantic and intent gaps via mining click logs of search engines[C]// ACM International Conference on Multimedia,2013:243-252

[6] Tan M,Yu J,Zheng G J,et al. Deep neural network boosted large scale image recognition using user click data[C]// International Conference on Internet Multimedia Computing & Service,2016:118-121

[7] Zheng G J,Tan M,Yu J,et al.Fine-grained image recognition via weakly supervised click data guided bilinear CNN model[C]// IEEE International Conference on Multimedia and Expo, 2017, DOI: 10.1109/ICME.2017.8019407

[8] Wu W C,Tan M,Yu J.Query modeling for click data based image recognition using graph based propagation and sparse coding[C]//International Conference on Internet Multimedia Computing and Service,2017(accepted)

[9] Yu J,Rui Y,Tao D C,et al.Click prediction for web image reranking using multimodal sparse coding[J].IEEE Transactions on Image Processing,2014,23(5):2019-2032

[10] Yu J,Rui Y,Chen B,et al.Exploiting click constraints and multi-view features for image re-ranking[J].IEEE Transactions on Multimedia,2014,16(1):159-168

[11] Zhao Q K,Hoi S C H,Liu T Y,et al.Time-dependent semantic similarity measure of queries using historical click-through data[C]// International Conference on World Wide Web,2006:543-552

[12] Wang T F, Bian J, Liu S S, et al. Psychological advertising:Exploring user psychology for click prediction in sponsored search[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2013:563-571

[13] Liu J H,Dolan P,Pedersen E R.Personalized news recommendation based on click behavior[C]//International Conference on Intelligent User Interfaces,2010:31-40

[14] Chapelle O,Zhang Y.A dynamic Bayesian network click model for web search ranking[C]//International Conference on World Wide Web,2009:1-10

[15] Bai Y L,Yang K Y,Yu W,et al.Automatic image dataset construction from click-through logs using deep neural network[C]// ACM Conference on Multimedia Conference,2015:441-450

[16] Chang C C,Lin C J.LIBSVM:A library for support vector machines[J].ACM Transactions on Intelligent Systems and Technology,2011,2(3):1-27

[17] Tan M,Wang B Y,Wu Z H,et al.Weakly supervised metric learning for traffic sign recognition in a lidar-equipped vehicle[J].IEEE Transactions on Intelligent Transportation Systems,2016,17(5):1415-1427

[18] Salton G,Buckley C.Term-weighting approaches in automatic text retrieval[J]. International Journal of Information Processing & Management, 1988, 24(5):513-523

[19] Wang W,Yang X Y,Ooi B C,et al.Effective deep learning-based multi-modal retrieval[J].The VLDB Journal,2016,25(1):79-101

[20] Zhang Y T,Sohn K,Villegas R,et al.Improving object detection with deep convolutional networks via Bayesian optimization and structured prediction[C]// IEEE Con-

ference on Computer Vision and Pattern Recognition, 2015:249-258
[21] Song Q,Yu S X,Leng C,et al.Learning deep features for MSR-bing information retrieval challenge[C]//ACM International Conference on Multimedia,2015:169-172
[22] 吴禀雅,魏苗.从深度学习回顾自然语言处理词嵌入方法[J].电脑知识与技术,2016,12(36):184-185
WU Bingya,WEI Miao.A review of natural language processing word embedding from deep learning [J]. Computer Knowledge and Technology, 2016, 12 (36): 184-185
[23] Nehemiah_Li.度量学习(Metric Learning)(一)[EB/OL].(2015-03-12).http://blog.csdn.net/nehemiah_li/article/details/44230053
[24] 谭敏.面向智能车的物体检测与识别[D].杭州:浙江大学计算机科学与技术学院,2015
TAN Min.Visual object detection and recognition for intelligent vehicles [D]. Hangzhou: College of Computer Science and Technology,Zhejiang University,2015
[25] Tan M, Hu Z F, Wang B Y, et al. Robust object recognition via weakly supervised metric and template learning[J].Neurocomputing,2016,181:96-107
[26] Bengio Y.Learning deep architectures for AI[J].Foundations and Trends in Machine Learning, 2009, 2(1): 1-127
[27] 尹宝才,王文通,王立春.深度学习研究综述[J].北京工业大学学报,2015,41(1):48-59
YIN Baocai,WANG Wentong,WANG Lichun.Review of deep learning[J].Journal of Beijing University of Technology,2015,41(1):48-59
[28] Lin T Y,Roychowdhury A,Maji S.Bilinear CNNs for fine-grained visual recognition [J]. arXiv e-print, 2015, arXiv:1504.07889

A survey of fine-grained image recognition based on user click data

YU Jun[1,2] TAN Min[1,2] ZHANG Hongyuan[1,2] ZHANG Haichao[1,2]
1 School of Computer Science and Technology,Hangzhou Dianzi University,Hangzhou 310018
2 Key Laboratory of Complex Systems Modeling and Simulation,Hangzhou Dianzi University,Hangzhou 310018

Abstract In recent years,fine-grained image recognition has become a hotspot in computer vision area.Due to the subtle visual differences among different image categories and the serious semantic gap,the performance of traditional image recognition algorithms for fine-grained images recognition is mostly unsatisfactory.To overcome these challenges,many researchers have been concentrating on image recognition with user click data.This paper focuses on the three key modules of the fine-grained recognition system with user click data:data pre-processing,feature extracting and model construction.Also,existing algorithms for click data based image recognition are summarized,and the related latest progresses are demonstrated.

Key words user click data;image recognition;metric learning;deep learning;semantic gap