机器学习在图书馆应用初探:以 TensorFlow 为例*

□郭利敏　刘炜　吴佩娟　张磊

摘要　机器学习是人工智能的重要分支,TensorFlow 是谷歌第二代开源人工智能机器学习平台。此文重点介绍机器学习(主要是深度神经网络)的基本原理和利用 TensorFlow 进行机器学习的基本方法,探讨在图书馆领域应用的可能和场景。以《全国报刊索引》的自动分类问题作为实验对象,利用两台图形工作站,建立了 TensorFlow 深度学习模型,通过设定参数和阈值、系统调优等工作,实践了应用 TensorFlow 的完整过程,论证了其可行性。实验通过对 170 万余条题录数据进行训练和测试,克服了报刊索引数据过于简单与中国图书馆分类法的类目过于细致之间的矛盾,实现了大类近 80%和四级分类总体近 70%的准确率(其中 TP 类达到 91%),得出基本可代替人工分类流程的结论,为全国报刊索引的分类流程的半自动化提供有力工具,从而可望大大节省人力成本。下一步将继续利用 TensorFlow 的优化功能,结合更多的字段属性,进行系统调优,力争做到自动分类 90%以上的准确率。

关键词　智慧图书馆　人工智能　机器学习　TensorFlow　自动分类　神经网络
分类号　G250
DOI　10.16603/j.issn1002－1027.2017.06.004

1　引言:人工智能与机器学习

得益于芯片技术和大数据处理技术的高速发展,作为人工智能核心的机器学习近年来获得了突破性进展,依靠算法进行语音识别、机器翻译、图像描述、文本分类、对象识别等应用已经达到甚至超过了人类的准确度,在更加复杂的无人驾驶、医疗诊断、证券分析、法务助理等领域也有了巨大进步,已开始应用于实际工作中。人们突然意识到,机器全面取代人类从事各类复杂劳动的趋势已不可阻挡,一个智能化社会正在迅速到来。

人工智能是让机器能够模拟人的认知、思维、行为方式或信息过程,机器学习则是让机器具有人一样的自我学习和解决问题的能力,例如让机器学会"自动编程"就是一种最重要的机器学习。深度学习是机器学习的一种类型,是通过模拟人脑神经元对于外界刺激的感知和传导过程,即建立人工神经网络,来获取对事物的认识、解释和判断。

深度学习最早由杰夫·辛顿(Geoffrey Hinton)于 2006 年提出,随后杨乐坤(Yann LeCun)等人提出了卷积神经网络,进行了显著优化。2012 年在斯坦福大学人工智能实验室举办的 ImageNet 图像识别大赛中,深度学习算法一鸣惊人而崭露头角,近两年又由于阿法狗战胜了人类最强围棋大师而名声大噪,使之一跃成为目前各类机器学习方法中适应性最强、效果最为显著的算法流派。可以认为人工智能是计算机科学所追求的最终目的,而机器学习则是实现人工智能的基本方法,目前的深度学习是被大家看好的实现路径[1]。它们三者之间的关系如图 1 所示。

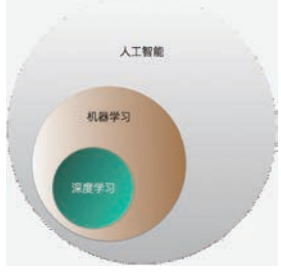

图 1　人工智能、机器学习与深度学习的关系

* 本文系国家社会科学基金重大项目"面向大数据的数字图书馆移动视觉搜索机制及应用研究"(编号:15ZDB126)的研究成果之一。
通讯作者:郭利敏. ORCID:0000-0002-8138-2762. lmguo@libnet.sh.cn.

2 人工神经网络与 TensorFlow

让机器具有智能是人类长期的梦想。人工智能发展的曲折历史告诉我们，“智能”的取得不仅需要“聪明”的算法，更需要海量数据的积累和计算能力的提高，算法、数据和芯片三者缺一不可，如今大数据时代正好带来了三者的汇聚。让机器具有某种智能，例如学会判断一张图片里的物体、识别一段语音中的文字、跟踪视频中的移动物体等，首先需要建立一定的算法模型，然后通过大批量的数据、经过高性能计算来训练模型，而最终通过不断地测试数据进行调试而实现，其中很多过程都需要创造性劳动，需要不断探索。虽然人工智能的提出已经有 60 多年历史，但目前才刚刚进入茁壮发展的早期阶段，未知的“黑箱”还有很多。

就机器学习的算法而言，迄今已形成众多流派，如符号学派、进化学派、概率学派、类推学派等，还在不断相互借鉴、融会贯通，以至于一直有人致力于“终极算法”的探索。目前以深度学习为代表的联结学派（即利用人工神经网络模拟认知过程的学派），由于杨乐坤等人在卷积神经网络（CNN）和循环神经网络（RNN）算法上的突破，取得了长足进展，一枝独秀，几乎成为机器学习的算法圣杯，被认为最有可能成为终极算法，其相应的实现平台也获得众多巨头的支持，取得长足进展。TensorFlow 就是目前最知名的支持各种神经网络算法的平台。

2.1 深度神经网络

为了说明 TensorFlow 的工作原理，必须了解深度学习的一些基本知识。深度学习又称为深度神经网络（DNN：Deep Neural Network），是由人工神经网络（Artificial Neural Network，ANN[2-3]）发展而来。神经网络是基于模拟大脑皮层的神经网络结构和功能而提出的计算模型，其原理是由感知器（perceptron）输入的信号，经由大量的神经节点（即人工神经元）构成的网络，按一定的规则进行大规模并行计算，而得到输出的完整过程。其中每个节点都需要进行一定的函数计算，称为激励函数（activation function），每两个节点之间都需要一定的加权值，用来模拟“记忆”强度，称为“权重”，这样经过一定“深度”的节点计算之后整个网络得到一系列的输出值，这个输出与期望值进行匹配之后可以用来调整函数和权重，从而更加逼近预期结果。把这个学习得来的神经网络模型应用于新的输入（即进行预测），往往也能得到类似的结果，这样就达到了“机器学习”的目的。可以看到，神经网络的输出结果依网络的连接方式、权重值和激励函数的不同而不同，训练网络的意思是通过输入和输出的对照，使某种算法中的所有函数能够无限逼近人们的预期。神经网络作为一种计算逻辑的策略表达被固定在模型中，可以反复使用，用来解决人们预期的特定问题[4]。整个模型可以形式化表达为图 2。

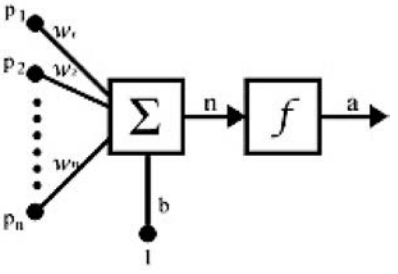

图 2 人工神经元数学模型图

其中 p 为神经元节点的输入，a 为神经元节点的输出。神经元将输入 p 加权求和，加上偏移量，经过激活函数 f，即：

$$a = f(n) = f\left(\sum_{i=1}^{n} p_i w_i + b\right) =$$

$$f\left((w_i, w_2 \cdots w_n)\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix} + b\right) = f(W^T P + b)$$

经过很多学者的长期研究和试验，目前常用的有效激活函数有线性函数、S 形函数、阈值函数、双曲正切函数、ReLu 等几种。

人工神经网络的研究早在 20 世纪 40 年代就开始了，70 年代进入低谷期，被认为是一种不切实际的幻想，到 80 年代得到一定的复苏发展[5]。神经网络算法的瓶颈并不在于其有多复杂，而在于极其巨大的并行计算量，当时的硬件还不足以支撑这类试验，在成本、技术及参数复杂度等方面的困难难以逾越，因此一直未能取得良好效果，只停留于理论研究以及小规模实验中。进入 21 世纪后，各方面条件已经成熟，特别是计算机硬件的速度可以承载上百亿节点数量的秒级计算，不仅采用专用图像处理器 GPU，而且谷歌还专门为 TensorFlow 开发了 TPU 专用处理器，可以集群进行人工神经元的模拟计算，因此大大提高了运算速度，过去需要几个月完成的计算量现在只需要几十分钟。TensorFlow 框架平

台正是在这种背景下出现并获得迅速发展。

2.2 机器学习框架的计算原理

机器学习的算法开发通常需要一定的平台软件的支持，才能够便于调试和提高效率。第一代深度学习平台大多由科研机构开发，用于算法研究目的。例如加州大学伯克利分校的 Caffe、纽约大学的 Torch、蒙特利尔大学的 Theano 等。随着机器学习得到普遍应用，需要兼顾在多种平台上快速部署和迭代开发，于是产业界逐渐成为平台研发的主力，但其基础和原型大多来自于大学和研究机构，显示了一种良好的产学研互动。例如 Facebook 接管了 Caffe 和 Torch 并开发了 Caffe2 和 PyTorch，微软推出了 CNTK，亚马逊开发了 MXNet，百度也推出了 Paddle。而 TensorFlow 则是由谷歌公司于 2016 年推出的最新平台。

谷歌的 TensorFlow 是在其第一代机器学习系统 DisBelief 基础上，结合 Theano 的优点开发的新一代系统，为了得到更多的支持而迅速占领市场，该系统采用 Apache2.0 协议开源，今年 2 月推出了具有里程碑意义的 1.0 版，目前(2017 年 9 月)已经是 1.4 版。

TensorFlow 以张量图的形式形象化地表示人工神经网络的模型和计算过程，并提供一整套开发环境，支持各类深度学习的算法研究(主要是 CNN、TNN 和 LSTM 三种算法)和应用开发，它在易用性和交互性设计方面颇为用心，且兼顾了算法研究和产业应用两个方面。

TensorFlow 的命名来自于它的运行原理，即前述人工神经网络算法的工程实现。张量(Tensor)通常表达为多维数据数组，Flow 是数据依据一定次序和规则进行计算而形成的流程，可以画成图(Graph)的形式表达，即由“结点”(nodes)和“边”(edges)表示成有向图，生动形象地表示了“张量”从图的一端流动到另一端的情况。“节点”用来表示所进行的函数操作，当然数据的输入(feed in)起点和输出(push out)终点以及在中间过程的读写操作也是一种节点，“边”表示“节点”之间的前后(输入/输出)关系，这些“边”可以传输大小可动态调整的多维数据数组。一旦模型建立起来，输入端的所有张量就已准备好，节点将被分配到各种计算设备，可执行异步分布式并行计算。

TensorFlow 作为后起之秀吸收了众多前辈平台的诸多优点，同时避免了不少不足。它支持多种机器学习常用的开发语言(如 C++/Python/Cuda)，支持几乎所有类型的深度学习算法的开发(如 CNN、RNN、LSTM 等)，能在多种硬件环境(CPU、GPU、TPU，Raspberry Pi、手机、云)下很好地利用各自的长处和特点运行，并能够进行网络分布式学习，具有本地化、领域化训练和学习模型的重用(通过 API，甚至能够支持其他平台下的模型重用)功能。由于其具有众多优点，如计算速度快、部署容易、灵活性强、可扩展等，所以一经推出就得到了人工智能界的热烈响应，开源社区迅速增长到数万人规模，已成为 GitHub 上最活跃的软件项目之一。很快地，由众多的第三方团队开发的大量工具和实验性项目，使其高速迭代，在推出 1.0 后短短半年就更新到 1.4 版，目前正在酝酿大版本更新。

谷歌公司自己是 TensorFlow 的最大用户，其众多的人工智能项目基本上都在这个平台上研发。同时它还投入了大量资源、采取了多种措施促进该平台的开放。在谷歌的强力推广下，很多高校、科研机构和第三方公司已开始使用 Tensorflow，取得令人瞩目的成果。例如谷歌利用该平台对其自动翻译服务进行了系统升级，翻译质量比过去有明显提升；在谷歌邮件系统中，通过邮件语境预测可能的回复；对视网膜影像数据进行训练，已能成功预测影像是否有糖尿病引起的视网膜病变[6]；在 AutoDraw 中开发“预测”功能，可以根据标题和用户画出的部分元素推测并继续完成一幅绘画作品等[7]。成功案例不胜枚举。

3 TensorFlow 与智慧图书馆建设

建设智慧图书馆的关键在于人工智能技术的应用，否则其是否具有“智慧”是值得商榷的。虽然国内图书馆界对于智慧图书馆的研究起步很早，但目前基本上还处于概念引入阶段，真正通过人工智能的应用赋予图书馆以“智慧”还不多见，仅有一些零星尝试，例如重庆大学图书馆利用机器学习对用户偏好进行动态测度[8]，据此提供更加精准的个性化服务；南京大学苏新宁教授团队曾经对于书目[9]和期刊文章[10]的自动分类都做过深入研究，尝试过支持向量机[SVM]和神经网络等多种模型，但未能付诸实用。国外的相关应用也不多见，这固然受到技术尚未成熟、机器学习框架平台近两年才开始建立

的限制，也与图书馆业务庞杂、数字时代的业务模型尚未建立有关。机器学习对于数据、计算资源和算法的开发都有较高的要求。

3.1 智慧图书馆的两类应用

对图书馆而言，人工智能可能应用于两个方面：图书馆内部业务和对外服务。内部业务主要指图书馆将外部资料纳入馆藏的处理流程，即从资源的采集或数字化，到编目、组织、典藏直至提供检索和存取的一整套工作；对外服务主要是指直接面向读者的一线工作，如流通、阅览、参考咨询、会议展览培训以及阅读推广等。对于前者，业务处理的实时性要求不高，机器学习只需要帮助图书馆员更加准确高效地进行知识组织工作即可，主要涉及文本处理、分类和实体对象的识别技术；对第二类应用，需要结合用户画像进行用户的聚类和资源的聚类，然后在知识组织体系内进行匹配，以提供动态的、个性化的精准服务。由此可见，在这两类应用对机器学习的要求和机器学习能起到的作用是不同的。

TensorFlow 发布迄今还不到两年，属于机器学习平台的后起之秀，虽然具有一定的普适性，但也并非灵丹妙药。从其特点和目前的成功案例来看，TensorFlow 在图书馆的应用主要集中于内部业务的智能化，如馆藏资源的自动分类、自动摘要、文本生成、主题提取、文章聚类、自动标引、图像识别、实体提取和分析预测报告的自动撰写等，在服务方面的应用具有一定的局限性，主要集中于卷积神经网络、循环神经网络能够应用的场景和领域，例如用户需求感知、自动翻译、语义理解和发现、自动参考咨询等方面。

目前 TensorFlow 在自然语言处理、图像和声音识别等领域已经取得了很大进展，以下介绍一些最新成果。

自然语言处理（NLP）和基于文本的应用是 TensorFlow 使用的一个重要领域，Google 研究发现，对于较短的文本，运用 sequence-to-sequence 模型来自动建立文本摘要（或主题）并用于 Gmail 的智能回答取得很好的效果（如图 3 所示）[11]。这一成果可应用于图书馆网上参考咨询机器人的开发。

另外金允（Yoon Kim）的研究表明，对于有限长度、结构紧凑、能够独立表达意思的句子，通过神经网络进行文本分类，也能达到令人满意的结果[12]。在图书馆服务中，文本分类是十分常见的需求。

在图像识别方面，目前机器不但能精确识别图像中的人脸以及数千种物体，而且还能理解图像所包含的内容及其相互关系，并通过自然语言表进行表达。如图 3 所示，Google 的最新研究成果表明，将计算机视觉和语言模型通过 CNN 与 RNN 网络叠加进行合并训练，所得到的系统可以自动生成一定长度的文字文本[13]，甚至能够完整讲述一张图片内所包含的故事。

语音/声音识别也是 TensorFlow 的基础应用之一[14]。通过适当的数据反馈（即采用 RNN），神经网络能够理解音频信号，进而可实现语音识别、语音搜索、情绪分析、缺陷检测等，进而能在语音与文字、图像间相互转换。这一成果可以应用于语音搜索和语音助手，除了如 Apple iOS 的 Siri，Android 的 Google Now 和 Windows Phone 的 Microsoft Cortana 那样应用于参考咨询的语音机器人之外，还可应用于为聋哑、视障或其他特殊人群服务。

计算机视觉是当前机器学习正在重点攻克的热点，TensorFlow 在视频应用方面进展神速，基于一些大学和研究机构专门建立的各类视频数据训练集，如 YouTube-8M、UCF-101 datase、Sports 1 million datase 等，在视频数据挖掘、分类和预测[15]方面已取得不少成果，这些成果可以帮助图书馆建立大型的视频数字图书馆，视频素材的标引和检索方面已基本无需人工干预。同样的技术还可以用在运动物体检测、安防保卫以及更好地与用户进行交互等方面。

图 3 模型读取短文与摘要

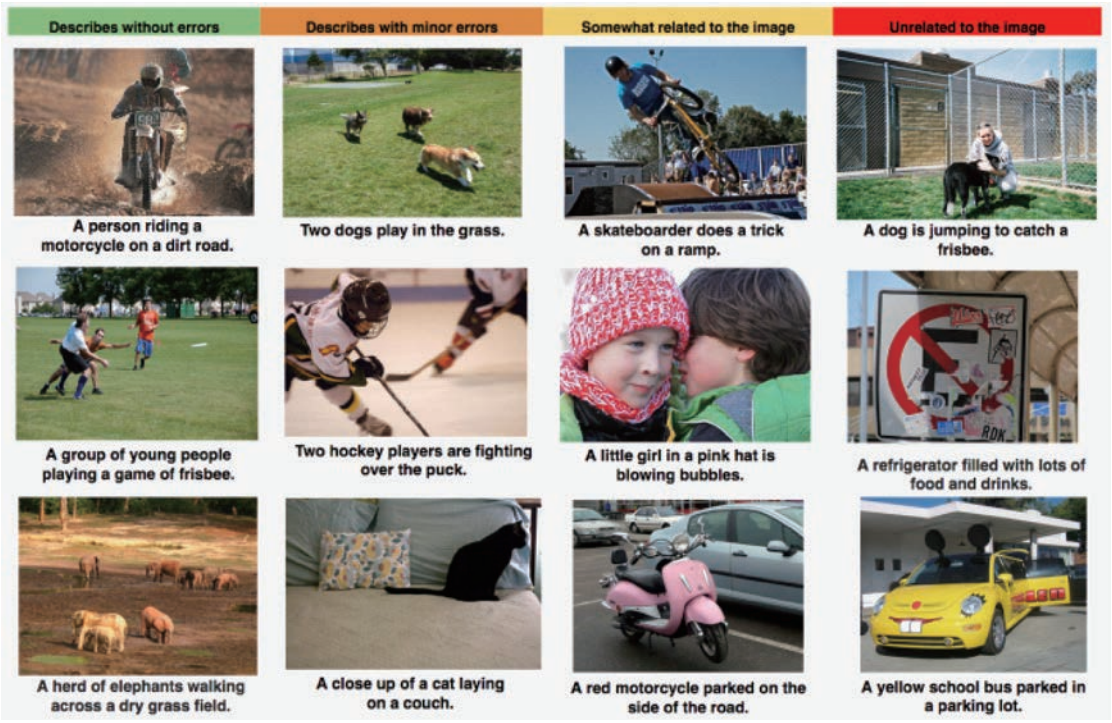

图 4　图片及机器评估结果

通过上述介绍可以得知，TensorFlow 是一个极其优秀的神经网络框架，其最大的好处是能够极大程度地降低深度学习的应用门槛，学习成本低且寻求帮助容易，这也使得深度学习在图书馆的应用成为可能。以下结合当下智慧图书馆的开发应用现状，就图书馆的参考咨询和决策咨询和两个应用场景，具体探讨 TensorFlow 可能带来的突破。

3.2　更智能的图书馆咨询服务

公共图书馆的参考咨询工作由于面对的读者类型复杂、层次跨度大，所以问题十分复杂，特别是研究型公共图书馆的读者，既有科研人员和专门领域的研究型学者，也有普通读者和大量的中小学生，其受教育情况通常呈金字塔型分布[16]。这决定了图书馆的咨询服务所涉及的问题比较庞杂，具有相当的广度和深度。图书馆的参考咨询服务不仅要以馆藏文献和馆内业务为依据、有针对性地向读者提供具体的文献知识、文献途径和服务内容（如协助读者检索书目和文献、指导读者阅读专题文献、进行文献传递等），还要解答关于活动的咨询、开闭馆时间甚至厕所位置、周边环境等问题。在新媒体环境下，读者与图书馆的连接方式更加简单直接，给在线咨询服务带来了巨大的压力，例如上海图书馆仅微信服务号渠道的月咨询量，平均就达 1 万次左右。

于是许多图书馆都想到采用咨询机器人进行问题解答，案例有清华大学的“小图”[17]、上海图书馆的“图小二”聊天机器人等。这些解决方案大多以向量空间模型使用权重计算和余弦相似度来做语句与语料的相似度判断，进行匹配，在无适当答案时，通常结合聊天记录，依据适当的推理机制提供解答。这在一定程度上解决了一般咨询所遇到的常规问题。然而这需要大量的数据加工，且对数据加工人员的要求也较高，不仅需要具有计算机专业知识，还需要对图书馆咨询服务也非常熟悉，这使得数据加工比较困难，问题答案的加工质量常常导致聊天机器人的回答准确度不高。对于研究者所需要的参考咨询，由于需要大量的专业知识背景，对参考咨询员要求更高，开发这样的自动问答系统对于传统的技术来说显然不切实际，这是目前自动参考咨询难以逾越的鸿沟。

利用神经网络的堆叠[18,19-21]，将文本从字到词、短语、句子、段落各个层级上进行特征提取，以实现文本分类、情感分析，实现机器对复杂自然语言的理解，进而对读者的问题进行细粒度分类，并结合适当的上下文语义推理，实现类似 Siri、Google Now 和 Microsoft Cortana 的智能助手的功能，相较于“小图”使用的空间向量模型的相似度计算来说，在

数据加工层面无需太多的专业知识，仅需要咨询服务人员在整理数据时，把文本划分到相应的一个或多个分类中即可，大幅降低数据加工的难度；同时对于参考咨询来说，神经网络是一个不断训练学习和不断自我完善的过程，这与参考咨询所需要的专业知识学习的需求相吻合。因此能够利用神经网络进行机器学习的参考咨询机器人将是未来的发展重点，有望得到突破。

3.3 预测与决策咨询

将大数据分析作为图书馆决策的重要参考已成为业界共识，一些大学和科研机构已有不少研究和实践，如重庆大学图书馆的大数据分析实验系统[22]，能够将资源、读者和服务三个维度的数据，通过累计、分析、归纳得出结论，对内指导业务工作，对外引导读者阅读方式，挖掘原有数据的使用价值；美国俄亥俄州立大学利用读者参与进行采购决策(PDA 项目)[23]，对所购图书的出版社分布、学科分布、图书利用率、文献类型等信息进行了详细的统计分析，不仅帮助图书馆调整资源建设标准和预设文档参数，还利用数据分析读者阅读倾向、阅读载体偏好、主题分布等信息，作为调整馆藏发展规划、构建馆藏特色资源的依据。

神经网络本身是一个适应性系统，在学习阶段可根据内外部数据流向来改变其网络结构，这使得通过神经网络进行预测与决策成为最为重要的研究与应用方向。塔库尔(Thakur)等通过分析 2000 年 1 月至 2012 年 12 月期间印度的经济数据，构建基于前馈反馈传播神经网络的通货膨胀预测模型，取得了相对满意的准确度[24]。芮丽奇(Tomislav Rolich)等对 50 个织物的经纱和纬纱的密度、单位面积的质量、厚度等数据进行建模，通过神经网络对植物的拉伸性进行预测，证明神经网络可用于织物拉伸性的预测[25]。李根永(Keun Young Lee)等收集包括温度、风速、湿度、降水等在内的气象数据，应用 ANN 和 MRL(多元线性回归)模型，实现了对首尔市永登浦区蚊子数量的预测[26]。

由于大数据分析对于统计分析相关的专业知识要求较高，TensorFlow 平台的低门槛优势就得以显现和发挥，稍经“数据科学”培训的图书馆馆员就能通过 TensorFlow 对图书馆的数据进行建模和处理，而图书馆也有大量的数据可以发挥价值，如读者数据(业务流、个体属性、群体属性、借阅数据、到馆数据、活动数据)、资源数据(流通情况、书目查询情况、出版与发行图书、期刊应用率、电子资源使用率、馆藏借阅、馆藏分布)、服务数据(网络运营数据、图书馆客流统计、网站浏览数据、网上文献检索与下载数据、专题资源数据)等，如能利用其进行深度学习，对图书馆的运营决策和精准服务将提供巨大帮助，智慧图书馆的建成将指日可待。

4 应用 TensorFlow 于全国报刊索引文本分类

受前述相关案例的启发，我们尝试使用 TensorFlow 对期刊论文依据中国图书馆分类法进行自动分类。

文献的标引编目加工是图书馆重要的业务工作之一，它工作量大、专业性强，同时又是需要多人协作的综合性工作，有自己的特点和规律。一方面，在知识爆炸时代，要对数量庞大、内容复杂、形式多样的文献进行准确的归类、标引，对工作人员的要求很高；另一方面，由于编目外包和图书馆学专业教育的转型，资深标引编目人员日趋减少，信息加工的质量和效率都呈下降趋势。

以上海图书馆全国报刊索引数据库为例，受限于人工标引，每月只能对 4 万余篇报刊篇名文献进行分类标引，无更多的人力分类更多的文献。考虑到自动文本分类是机器学习的长项之一，我们考虑完全可以通过 TensorFlow 对以往海量的分类文摘进行学习之后，实现对新入库文献的自动分类，这样能大大减少人工分类，提高效率，让更多的人进行数据核查和校对工作，还能提高数据质量，一举两得。

4.1 神经网络模型的设计

文献分类是编目人员根据文献的有关信息，如题名、作者、出处、关键词、摘要、内容等，给文献一个或多个分类号的过程。对于深度学习来说，其本质就是分类问题。本研究基于现有的 170 万条规范数据中的 90% 作为训练数据，利用自然语言处理技术，从文献题名和摘要中提取关键词，结合文章作者提供的关键词，组成训练集，首先实现了中国图书馆分类法一级类目的自动分类，而后以 TP 类等部分类目，尝试进行四级类目的自动分类，其中的 10% 作为验证数据，以测试其预测准确度。

图 5 为文献自动分类的神经网络结构。其中输入层为 20 * 20 卷积网络，隐含层由卷积核为 3 * 3 和 5 * 5 的卷基层堆叠而成，输出层为全连接层。

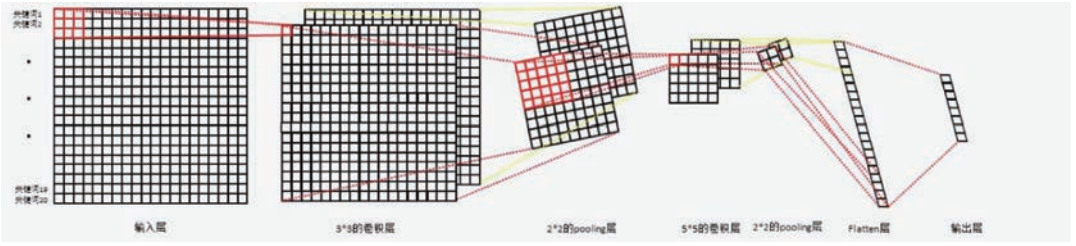

图 5　网络结构

利用 CNN 将文献中所有的词进行训练，并转换成长度为 n 的词向量[27-29]。对于一篇文章，其关键词长度为 m(长度不足 m 时做特殊词向量填充)，则其表达式为：

$$X_{1:m}=x_1\oplus x_2\oplus\cdots\oplus x_m \quad (1)$$

其中⊕为并运算。则其网络输入为 $m*n$ 的二阶张量，对输入施加卷积核大小为 $w\in R^{hk}$ 的卷积层可以得到

$$X_{1:m}=x_1\oplus x_2\oplus\cdots\oplus x_m \quad (2)$$

其中 $b\in R$ 为偏置量，f 为非线性函数。将卷积核应用于输入可以得到特征图

$$c=[c_1,c_2,\cdots,c_{m-h+1}] \quad (3)$$

对 c 加以空域信号施加最大值池化，可以得到 $\hat{c}=\max\{c\}$。

通过上述描述，可以得到一个卷积对应一个特征，模型使用了多个卷积核来提取多个特征，这些特征最终被传递到倒数第二并从二阶张量平整为一阶张量，并被全连接的 softmax 层输出为标签上的概率分布，即文章在每一个目标分类上的概率。

4.2　实验结果分析

本文模型训练和测试数据所涉及的分类总计 38 个，详细分布情况如图 6 所示，经图 5 网络模型训练的 loss 和 acc 如图 7 所示，测试集的 loss 和 acc 如图 8 所示，理论曲线与测试曲线基本一致，可见 7 万条测试集的预测准确收敛于 0.7，即网络的准确率趋于 70%。

在实际测试中，从社科(3027 条)和科技(4117 条)两大类文献共随机抽取 7144 条数据作为实测数据，如图 9－11 所示，社科大类预测正确数为 2368 条，占 78.23%，科技大类预测正确数为 3018 条，占 73.31%，整体预测正确率为 75.39%。通过对训练数据以及预测结果的抽样分析得到影响预测结果的因素如下：

(1)分词对训练数据的影响：由于采用的是通用分词，并未对专业词表做特殊处理，如："上海图书馆"这个专有名词会被切分为"上海"和"图书馆"两个词，使得其专业词汇的特征在分词后丢失，在一定程度上会影响从摘要提取关键词的准确度，进而影响训练集的准确度。可通过对分词引擎添加专业词表来解决此问题，以提高预测的准确率。

(2)通过对错误预测的结果分析发现，在部分预测集中，概率次高的结果为命中结果。这一方面是由于部分数据存在复分的情况，而训练集直接采用主分类作为输出结果，使得次分类信息丢失，进而影响该条数据在全局中的权重。另一方面，由于神经网络是通过训练数据做回归计算而得的模型，其结果与人的主观分类有所差别，通过增加训练数据，使得其结果有一定的提高。

由于文献分类在很大程度上有主观因素，本文使用的测试和训练数据量虽具有一定的统计学意义，但要达到更高的预测准确率还需要更多的训练和测试集。其次本文仅对文献大类分类进行预测，下一步将通过增加子网络的形式，以实现对文献 3－4 级类目进行自动分类。

以上实验性开发初步证明了 TensorFlow 用于自动分类的可行性，本文的讨论也展现了机器学习对于智慧图书馆建设的重要作用和美好前景。当然 TensorFlow 作为一个比较新的神经网络框架，对智慧图书馆建设是否能发挥举足轻重的作用仍有待进一步应用和观察。可以肯定的是，随着第二次机器革命的逐步靠近，图书馆也会随着人工智能的发展而变得越来越智慧化，给图书馆带来新的发展与变革的机遇。

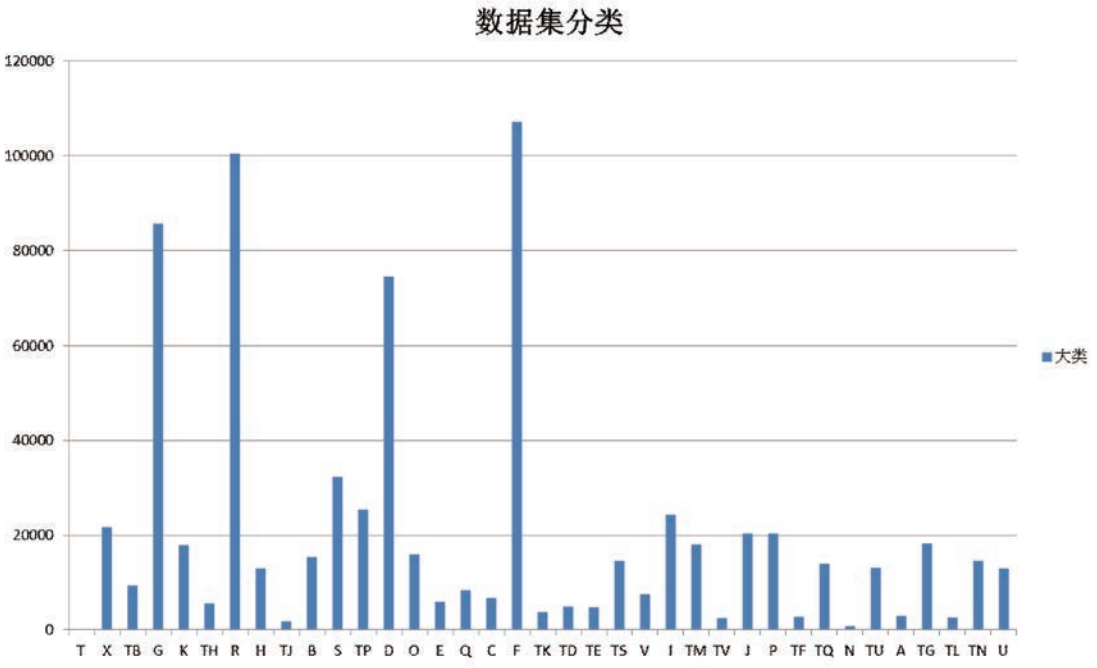

图 6　训练数据分布情况

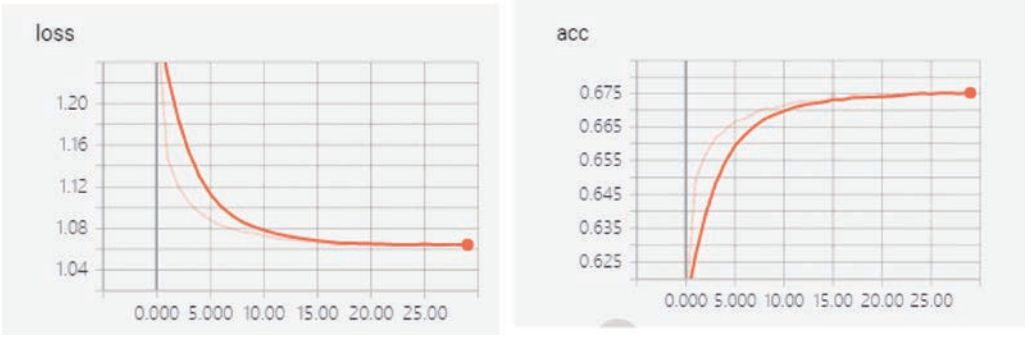

图 7　模型训练集的损失函数及准确率函数

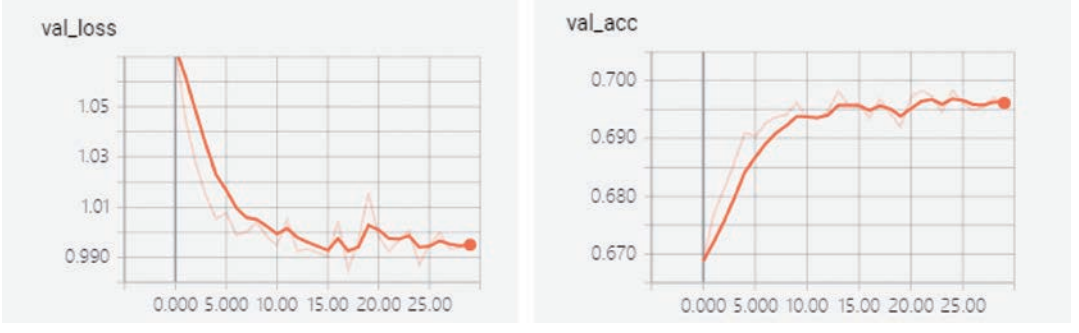

图 8　模型测试集的损失函数及准确率函数

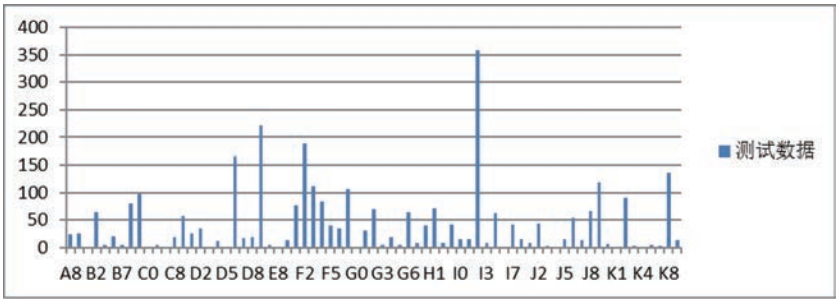

图 9　社科实测数据分布

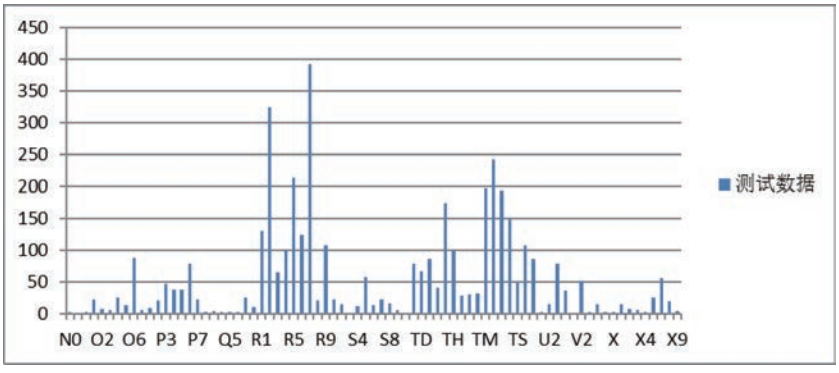

图 10　科技实测数据分布

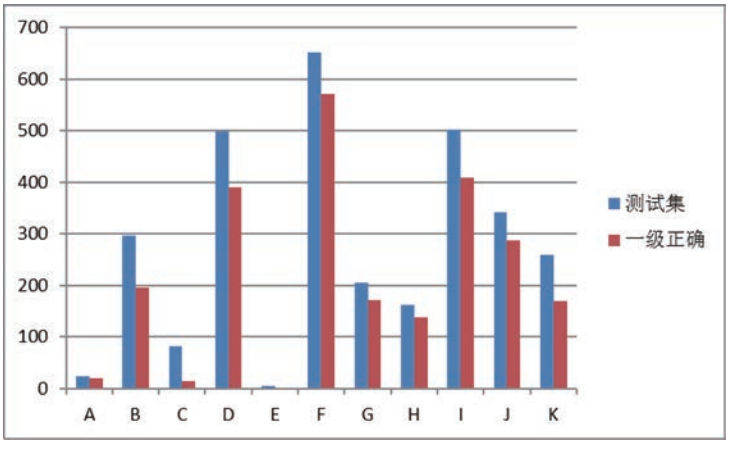

图 11　社科预测正确数据分布

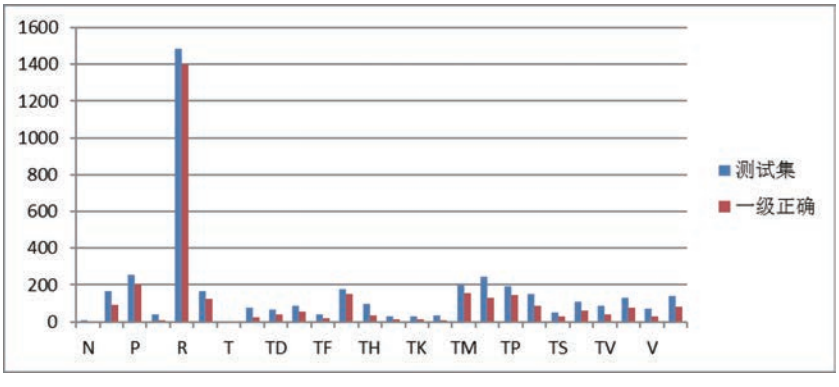

图 12　科技预测正确数据分布

参考文献

1 陈宗周. 从 GPU 到 ImageNet，两位硅谷华人改变了 AI 发展史. [2017－02－01]. https://mp. weixin. qq. com/s? __biz=MjM5NDA1Njg2MA==&mid=2651984124&idx=1&sn=ec445431989126e8c33352af54ca8b6b.

2 Hebb Donald. The Organization of Behavior a neuropsychological theory[M]. New York:John Wiley. 1949:100－136.

3 Liu M Q. Discrete－time delayed standard neural. Network and its application[J]. Sci China,2006,49(2):137－154

4 Neha Gupta, Artificial Neural Network[J]. Network and Complex Systrems, 2013 (1):24－28.

5 毛健,赵红东,姚婧婧. 人工神经网络的发展及应用[J]. 电子设计工程,2011,(24):62－65.

6 google developers blog[EB/OL]. [2017－02－01]. https://developers. googleblog. com/2017/02/announcing－tensorflow－10. html. 6 Auto Draw[EB/OL]. [2017－04－01]. https://www. autodraw. com/.

7 沈敏,杨新涯,王凯. 基于机器学习的高校图书馆用户偏好检索系统研究[J]. 图书情报工作,2015. (11):143－148.

8 王昊,严明,苏新宁. 基于机器学习的中文书目自动分类研究[J]. 中国图书馆学报,2010. (6):28－39.

9 叶鹏. 基于机器学习的中文期刊论文自动分类研究[D]. 南京大学,2013.

10 Sutskever, I., Vinyals, O., and Le, Q. (2014). Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems (NIPS 2014).

11 Kim, Yoon. 2014. Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408. 5882.

12 A Picture is Worth Thousand Coherent [EB/OL]. [2014－11－01]. https://research. googleblog. com/2014/11/a－picture－is－worth－thousand－coherent. html.

13 Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In IEEEInternationalconferenceonacoustics, speech and signal processing (pp. 6645－6649). IEEE.

14 J. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In CVPR, 2015.

15 同 12.

16 於坚秋. 公共图书馆读者信息咨询服务分析与对策[J]. 图书情报论坛,2006,(03):46－48.

17 姚飞等. 实时虚拟参考咨询服务新尝试——清华大学图书馆智能聊天机器人[J]. 现代图书情报技术,2011,(04):77－81.

18 同 12.

19 Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences[J]. arXiv preprint arXiv:1404. 2188, 2014.

20 Zhou C, Sun C, Liu Z, et al. A C－LSTM neural network for text classification[J]. arXiv preprint arXiv:1511. 08630, 2015.

21 Wen Y, Zhang W, Luo R, et al. Learning text representation using recurrent convolutional neural network with highway layers[J]. arXiv preprint arXiv:1606. 06905, 2016.

22 严轩,钟静. 从数据来源分析入手的图书馆大数据应用系统研究——以“重庆图书馆大数据分析试验系统”为例[J]. 四川图书馆学报,2016,(04):2－6.

23 李艳,吕鹏,李珑. 基于大数据挖掘与决策分析体系的高校图书馆个性化服务研究[J]. 图书情报知识,2016,(02):60－68.

24 Thakur G S M, Bhattacharyya R, Mondal S S. Artificial Neural Network Based Model for Forecasting of Inflation in India[J]. Fuzzy Information and Engineering, 2016, 8(1): 87－100.

25 Rolich T, Šajatovi A H, Pavlinic' D Z. Application of artificial neural network (ANN) for prediction of fabrics' extensibility [J]. Fibers and polymers, 2010, 11(6): 917－923.

26 Lee K Y, Chung N, Hwang S. Application of an artificial neural network (ANN) model for predicting mosquito abundances in urban areas[J]. Ecological informatics, 2016, 36: 172－180.

27 Goldberg Y, Levy O. word2vec Explained: deriving Mikolov et al.'s negative－sampling word－embedding method[J]. Eprint Arxiv, 2014.

28 Levy O, Goldberg Y. Neural word embedding as implicit matrix factorization[J]. Advances in Neural Information Processing Systems, 2014, 3:2177－2185.

29 Levy O, Goldberg Y, Dagan I. Improving distributional similarity with lessons learned from word embeddings[J]. Bulletin De La Société Botanique De France, 2015, 75(3):552－555.

作者单位：上海图书馆，上海，200031
收稿日期：2017 年 10 月 16 日

Machine Learning and Its application in Library: Take TensorFlow as an Example

Guo Limin　Liu Wei　Zhang Lei

Abstract: Machine learning (ML) is a particular approach to artificial intelligence. TensorFlow is the second generation machine learning framework of Google. This paper focuses on the basic principles of machine learning and the basic methods of machine learning by using TensorFlow. Its purpose is to explore the possibilities and scenarios of machine learning applications in library. A TensorFlow ML model is established and with the index data from National Index of Newspapers and Magazines, a complete process of automatic classification of records had been accomplished and proved feasible. Through the training process and testing of more than 170 million data records, the experiment has overcome the contradiction between the less comprehension of the index data and the trivial category labels, and reached nearly 80% of the categories and nearly 70% of the accuracy rate. It can be concluded that the approach is capable of taking into practice, at least to carry on a semi-automatic processing of classification, which is expected to significantly save labor costs. The next step will be optimizing the parameters and system tuning. We hope it can strive to achieve an accuracy of 90% by automatic classification.

Keywords: Smart Library; Artificial Intelligence; Machine Learning; TensorFlow; Automatic