Software Engineering and Applications 软件工程与应用, 2022, 11(3), 549-557
Published Online June 2022 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2022.113057

Hans汉斯

深度学习在大数据分析中的应用研究

贾美娟，李　欣，孔　靓，刘　春，邵国强

大庆师范学院计算机科学与信息技术学院，黑龙江　大庆

收稿日期：2022年4月12日；录用日期：2022年6月8日；发布日期：2022年6月15日

摘　要

深度学习算法通过分层学习过程提取高级、复杂的抽象作为数据表示。本文通过对现有深度学习在大数据分析中的所取得的成果，探讨了如何利用深度学习解决大数据分析中的一些典型问题，如从海量数据中提取复杂模式、语义索引、数据标记、快速信息检索及简化区分任务等问题，重点讨论的是在音视频方面的应用情况。最后，就目前研究存在的问题对未来相关工作提出见解。

关键词

大数据，深度学习，数据表示，语义索引，区分任务

Research of Deep Learning Applications in Big Data Analytics

Meijuan Jia, Xin Li, Liang Kong, Chun Liu, Guoqiang Shao

College of Computer Science and Information Technology, Daqing Normal University, Daqing Heilongjiang

Received: Apr. 12th, 2022; accepted: Jun. 8th, 2022; published: Jun. 15th, 2022

Abstract

Deep Learning Algorithm extracts high-level and complex abstractions as data representation through hierarchical learning process. Based on the achievements of existing Deep Learning in big data analysis, this paper discusses how to use Deep Learning to solve some typical problems in big data analysis, such as extracting complex patterns from massive data, semantic indexing, data labeling, fast information retrieval and simplifying differentiated tasks. The focus is on the application in audio and video. Finally, some opinions on the future related work are put forward based on the problems existing in the current research.

文章引用: 贾美娟, 李欣, 孔靓, 刘春, 邵国强. 深度学习在大数据分析中的应用研究[J]. 软件工程与应用, 2022, 11(3): 549-557. DOI: 10.12677/sea.2022.113057

Keywords

Big Data, Deep Learning, Data Representation, Semantic Index, Discriminative Task

1. 引言

现代数据密集型技术以及增加的计算和数据存储资源对大数据科学的发展做出了重大贡献[1]。谷歌、雅虎、微软和亚马逊等基于技术的公司已经收集并维护了以 EB 或更大比例衡量的数据，而 Facebook、YouTube 和 Twitter 等社交媒体组织拥有数十亿用户，目前仍然在不断生成大量数据。大数据代表了用于应用领域的问题和技术，这些应用领域收集和维护大量原始数据，用于特定领域的数据分析。从海量输入数据中挖掘和提取有意义的模式，用于决策、预测和其他推断，是大数据分析的核心。除了分析海量数据外，大数据分析还为机器学习和数据分析带来了新的挑战，如原始数据的格式变化、快速移动的流式数据、数据分析的可信度、高度分布的输入源、嘈杂和低质量的数据、高维性、算法的可扩展性、不平衡的输入数据、无监督和未分类的数据、有限的监督/标记数据等。此外，大数据分析还涉及到充分的数据存储、数据索引/标记和快速信息检索等关键问题。因此，在处理大数据时，需要创新的数据分析和数据管理解决方案。

本文主要讨论深度学习如何帮助解决大数据分析中的特定问题，这包括从海量数据中学习、语义索引、区分性任务和数据标记等。确定在大数据分析的深度学习方面需要创新的重要未来领域。

2. 数据挖掘和机器学习中的深度学习

深度学习算法中一个重要的概念是自动从数据中提取表示(抽象) [2] [3] [4]。即使用大量无监督数据自动提取复杂表示。这些算法主要受人工智能领域的推动，其总体目标是模拟人脑的观察、分析、学习和决策能力，特别是对于极其复杂的问题，该算法努力模仿人脑的分层学习方法。当试图从输入语料库中的复杂结构和关系中提取有用信息时，基于浅层学习体系结构(如决策树、支持向量机和基于案例的推理)的模型可能会失败。相比之下，深度学习体系结构能够以非局部和全局的方式进行概括，在数据中生成超越近邻的学习模式和关系[5]。事实上，深度学习是迈向人工智能的重要一步。它不仅提供了适用于人工智能任务的复杂数据表示，而且使机器独立于人类知识。

深度学习方法的另一个关键概念是数据的分布式表示。数据分布式能够为输入数据的抽象特征进行大量可能的配置，并且允许每个样本的紧凑表示。观察到的数据是通过几个已知/未知因素的相互作用生成的，因此，当通过学习因素的一些配置获得数据模式时，可能会通过学习因素和模式的新配置来描述额外的数据模式[2] [3]。与基于局部泛化的学习相比，使用分布式表示可以获得的模式数量随着学习因子的数量快速增加。

人工智能相关任务中使用的真实数据大多来自多个源的复杂交互。例如图像是由灯光、对象形状和对象材质等不同的变化源组成，而深度学习算法提供的抽象表示可以对数据变化的不同来源进行分离。深度学习算法实际上是连续层的深层架构，每一层对其输入进行非线性变换，并输出其表示。叠加非线性变换层是深度学习算法的基本思想，数据在深层架构中经过的层次越多，构建的非线性转换就越复杂。

可以将深度学习视为表示学习算法的特例，该算法在具有多个表示层次的深度体系结构中学习数据的表示，其实现的最终表示是输入数据的高度非线性函数。

3. 大数据分析

数据存储能力、计算处理能力的提高以及数据量的迅速增加促进了大数据时代的到来。如此大的数据量也是大数据的一个主要积极特征。许多公司，如 Facebook、雅虎、谷歌，已经拥有大量数据，并且最近开始利用其优势[1]。一般意义上的大数据通常指超过传统数据库和数据分析技术的典型存储、处理和计算能力的数据。无法管理的大量数据对传统计算环境构成了直接挑战，作为一种资源，大数据不仅需要能够从大规模数据中分析和提取模式的工具和方法，还需要可扩展存储和分布式策略来进行数据查询和分析。

大数据系统原始数据越来越多样化和复杂，主要由未分类/未监督的数据以及少量分类/监督的数据组成。系统需要对给定存储中各种数据表示形式的多样性进行处理，这给大数据带来了独特的挑战。除了明显的海量数据外，大数据还与其他特定的复杂性相关，如容量、多样性、速度和准确性等[6] [7] [8]。因此大数据系统需要对非结构化数据进行大数据预处理，以便提取结构化/有序表示形式，为后续研究者或者消费者使用。

在当今的数据密集型技术时代，数据速度——即数据收集和获取的增长率，与大数据的数量和多样性特征同样重要。目前，一些大型社交平台如 Twitter、Yahoo 等已经开发了用于流式数据分析的产品[6]。通常情况下虽然对流式数据不立即进行处理和分析可能会丢失数据，但可以选择将快速移动的数据保存到大容量存储中，以方便稍后进行批处理。然而，处理与大数据相关的速度的实际重要性在于反馈回路的快速性，即将数据输入转换为可用信息的过程。这对于时间敏感的信息处理尤其重要。大数据的准确性涉及数据分析结果的可信度或有用性，随着数据源和数据类型数量的增加，维持对大数据分析的信任也是一个实际挑战。

大数据分析面临着许多挑战，关键问题领域包括：数据质量和验证、数据清理、功能工程、高维性和数据缩减、数据表示和分布式数据源、数据采样，算法的可扩展性、数据可视化、并行和分布式数据处理、实时分析和决策、众包和语义输入，以改进数据分析、跟踪和分析数据来源、数据发现和集成、并行和分布式计算、探索性数据分析和解释，整合异构数据，开发新的海量数据计算模型。

4. 深度学习在大数据分析中的应用

如前所述，深度学习算法通过使用分层多级学习方法提取原始数据有意义的抽象表示。在分层多级学习方法中，高级别具有更抽象和更复杂表示，这是基于学习层次中的较低级别不太抽象的概念和表示来学习的。虽然深度学习可以应用于从足够大数量的标记数据中学习，但它主要有利于从大量未标记/无监督数据中学习[2] [4] [5]，从而有利于从大数据中提取有意义的表示和模式。

一旦通过深度学习从无监督数据中学习分层数据抽象，就可以借助相对较少的监督/标记数据点来训练更传统的判别模型，其中标记数据通常通过人工或专家输入获得。与相对较浅的学习架构相比，深度学习算法在提取数据中的非局部和全局关系及模式方面表现得更好[5]。通过深度学习能够获得的抽象表示的其他有用特征包括：1) 相对简单的线性模型可以有效地处理从更复杂和更抽象的数据表示中获得的知识；2) 从无监督数据中提取数据表示的自动化程度提高，使其能够广泛应用于不同的数据类型，如图像、纹理、音频等；3) 可以在原始数据的更高抽象和表示级别上获得关系和语义知识。尽管基于深度学习的数据表示还有其他有用的方面，但上述特定特征对于大数据分析尤其重要。

从大数据的四个特征——容量、多样性、速度和准确性考虑，深度学习算法和体系结构更适合解决与大数据分析的容量和多样性相关的问题。深度学习利用了海量数据的可用性，即大数据中的海量数据，

在大数据中，具有浅层学习层次结构的算法无法探索和理解数据模式的更高复杂性。此外，由于深度学习涉及数据抽象和表示，因此它很可能适合于分析以不同格式或来自不同来源的原始数据，即大数据的多样性，并且可以最大限度地减少对专家输入的需求，从而对大数据中的每种新数据类型进行特征提取。虽然大数据分析为更传统的数据分析方法带来了不同的挑战，但同时它也为解决与大数据相关的特定问题的新算法和模型提供了重要机会。深度学习概念为数据分析专家和从业者提供了解决方案的场所。例如，通过深度学习提取的表示可以被视为决策、语义索引、信息检索以及大数据分析中的其他用途的实用知识来源。此外，当复杂数据以更高的抽象形式表示时，可以考虑使用简单的线性建模技术进行大数据分析。

深度学习算法和体系结构领域已经完成了一些重要工作，包括语义索引、区分性任务和数据标记。本文通过研究利用深度学习进行大数据开发利用的已有成果，了解到深度学习技术在大数据分析中的新颖适用性，特别是文献中的一些应用领域涉及大规模数据。深度学习算法适用于不同类型的输入数据，本节我们将重点介绍它在图像、文本和音频数据上的应用。

4.1. 语义索引

与大数据分析相关的一项关键任务是信息检索[1]。在社交网络、安全系统、购物和营销系统、防御系统、欺诈检测和网络流量监控等领域，需要收集和提供包括文本、图像、视频和音频等大量异构数据，而以往的信息存储、检索策略和解决方案都受到大数据时代产生的海量数据和不同数据表示的挑战，因此信息的高效存储和检索是大数据中日益严重的问题。在上述这些系统中，通常是利用语义索引获得大数据而不是通过存储为数据位的字符串而获得。语义索引以更高效的方式呈现数据，并使其成为知识发现和理解的有用来源，例如使搜索引擎提供更快、更高效地工作。大数据不再使用原始输入进行数据索引，而是使用深度学习生成的高级抽象数据表示，这些抽象数据表示即为大数据服务的语义索引。这些表示可以揭示复杂的关联和因素(特别是当原始输入是大数据时)，从而产生语义知识和理解。

数据表示在数据索引中起着重要作用，例如，允许具有相对相似表示的数据点或实例存储在彼此更近的内存中，从而有助于高效的信息检索。然而，应该注意的是，高级抽象数据表示需要有意义，并证明关系和语义关联，以便更好地提供实际的语义理解和对输入的理解。

虽然深度学习有助于提供对数据的语义和关系的理解，但数据实例的向量表示(对应于提取的表示)将提供更快的搜索和信息检索。更具体地说，由于学习到的复杂数据表示包含语义和关系信息，而不仅仅是原始数据，因此当每个数据点(例如给定文本文档)由向量表示表示时，它们可以直接用于语义索引，允许基于向量的比较，这比直接基于原始数据比较实例更有效。具有相似向量表示的数据实例可能具有相似的语义。因此，在大数据分析中能够使用复杂高层数据抽象的向量表示对数据进行索引，即实现语义索引。在本节的其余部分将重点介绍基于深度学习获得的知识的文档索引。

基于从深度学习中获得的数据表示的索引的一般思想可以扩展到其他形式的数据。文档表示是许多领域信息检索的一个关键方面。文档表示法的目标是创建一种表示法，以浓缩文档的特定和独特方面，例如文档主题。文档检索和分类系统主要基于字数，表示每个单词在文档中出现的次数。各种文档检索模式都使用这种策略，例如 TF-IDF [9]和 BM25 [10]。这样的文档表示模式把单个单词认定为维度，且不同维度是独立的。在实践中单词的出现是高度相关的。使用深度学习技术提取有意义的数据表示可以从高维文本数据中获取语义特征，反过来也会降低文档数据表示的维度。

Hinton 等人描述了一种用于学习文档二进制代码的深度学习生成模型[11]。该模型中，深度学习网络的最低层表示文档的字数向量，该向量作为高维数据，而最高层表示文档的学习二进制代码，且使用 128 位代码。该文献证明了语义相似的文档的二进制代码在汉明空间中相对较近，并且文档的二进制代码可

用于信息检索。对于每个查询文档计算其与数据中所有其他文档的汉明距离，并检索前 D 个类似文档。二进制代码需要的存储空间相对较少，并且使用诸如快速比特计数等算法计算两个二进制代码之间的汉明距离，这样可以实现二进制代码相对较快的搜索。文献得出如下结论：用这些二进制代码进行文档检索比基于语义的分析更准确、更快。

深度学习生成模型还可以通过强制学习层次结构中的最高层使用相对较少的变量来生成较短的二进制代码，这些较短的二进制代码可以简单地用作内存地址。一个单词的内存用于描述一个文档，这样，围绕该内存地址的小汉明球包含语义相似的文档——这种技术称为“语义哈希”[12]。使用这种策略可以对非常大的文档集执行信息检索，且检索时间与文档集大小无关。语义哈希等技术对于信息检索非常有吸引力，可以通过查找与查询文档的内存地址相差几位的所有内存地址来检索与查询文档相似的文档。文献证明了“内存散列”比局部敏感散列快得多，而局部敏感散列是现有算法中速度最快的方法之一。此外，通过向 TF-IDF 等算法提供文档的二进制代码而不是提供整个文档，可以实现更高级别的准确性。虽然深度学习生成模型在生成用于文档检索的二进制代码方面的学习(或称之为训练)时间相对较慢，但由此产生的知识会产生快速推断，这是大数据分析的一个主要目标。更具体地说，为一个新文档生成二进制代码只需要进行一些向量矩阵计算，通过深度学习网络体系结构的编码器组件执行前馈传递。

为了更好地学习表示和抽象，可以使用一些有监督的数据来训练深度学习模型。Ranzato 等人在文献[13]中提出基于监督和非监督数据学习能够获得深度学习模型的参数。这种策略的优点是不需要完全标记大量数据(如预期的一些未标记数据)，并且模型具有一些先验知识(通过监督数据)来捕获数据中的相关标签信息。换句话说，除了提供文档类标签的良好预测外，模型还需要学习能够生成良好输入重构的数据表示。文献表明：对于学习紧凑表示，深度学习模型优于浅层学习模型。紧凑表示是有效的，因为在索引中使用它们时需要较少的计算，除此之外，当然还需要较少的存储容量。

谷歌的“word2vec”工具是一种可以从大数据中自动提取语义表示的工具。该工具以大规模文本语料库为输入，生成的词向量作为输出。它首先从训练文本数据构造词汇表，然后学习单词的向量表示，在此基础上，单词向量文件可以作为许多自然语言处理(Natural Language Processing, NLP)和机器学习应用程序的特征。Miklov 等人利用人工神经网络学习单词的分布式表示[14]，能够从包含数亿个单词和数百万个不同单词的庞大数据集中学习高质量单词向量。Dean J 等人在大规模分布式框架“DistFalse”上实现了庞大的数据集上的训练网络[15]。在大量数据上训练的词向量显示了词之间微妙的语义关系，例如城市和它所属的国家——例如，巴黎属于法国，柏林属于德国。具有这种语义关系的词向量可用于改进许多现有 NLP 应用，如机器翻译、信息检索和问题回答系统。例如，Miklov 等人演示了 word2vec 如何应用于自然语言翻译[16]。

与文本数据类似，深度学习可以用于其他类型的数据，从输入语料库中提取语义表示，从而为该数据建立语义索引。但是深度学习的出现相对较晚，所以若使用分层学习策略作为大数据语义索引方法，还需要做大量的工作。此外，当试图提取数据表示以进行索引时，使用什么标准来定义“相似”也一个急需要解决的问题。

4.2. 区分性任务

利用大数据分析技术进行区分时，可以使用深度学习算法从原始数据中提取复杂的非线性特征，然后将使用简单线性模型提取出的特征作为输入来执行区分任务。 这种方法有两个优点：1) 通过深度学习提取特征为数据分析增加了非线性，将区分性任务与人工智能紧密关联；2) 对提取的特征应用相对简单的线性分析模型，计算效率更高，这对于大数据分析非常重要。因此，从大量输入数据中可以开发非线性特征，使数据分析能够把学到的知识应用于更简单的线性模型，从而进行进一步分析，最终实现从

大量数据中可用的知识中受益。这是在大数据分析中使用深度学习的一个重要好处，允许从业者通过使用更简单的模型来完成与人工智能相关的复杂任务，例如图像理解、图像中的物体识别等。因此，在深度学习算法的帮助下，大数据分析中的判别性任务相对容易。

区分性分析是大数据分析的主要目的之一，同时也可以通过执行判别性分析对数据进行标记，从而实现搜索。Li 等人探索实现的微软研究音视频检索系统(Microsoft Research Audio Video Indexing System, MAVIS)使用基于深度学习的语音识别技术对语音和视频文件进行搜索[17]。为了将数字音频和视频信号转换为文字，MAVIS 会自动生成封闭的字幕和关键字，以增加对语音内容音频和视频文件的可访问性和发现性。

4.3. 语义标注

近年来互联网的发展和在线用户的爆炸式增长促进了数字图像收藏的规模迅速增长。数字图像的信息来源主要有社交网络、全球定位卫星、图像共享系统、医疗成像系统、军事监控和安全系统等。谷歌探索开发了提供图像搜索的系统，如谷歌图像搜索服务，包括仅基于图像文件名和文档内容的搜索系统，而不考虑涉及图像内容本身[18] [19]。图像的文本表示并不总是在大规模图像收集库中可用，为了能够利用人工智能对图像进行搜索，研究者应该超越图像的文本关系，且尽可能地收集和组织这些海量图像数据，以便能够更高效地浏览、搜索和检索这些数据。为了处理大规模的图像数据收集，一种方法是自动标记图像的过程，并从图像中提取语义信息。深度学习为构建图像和视频数据的复杂表示提供了新的前沿领域，作为相对较高的抽象级别，可用于图像注释和标记，这对图像索引和检索非常有用。由此可知，在大数据分析的背景下，深度学习将有助于区分数据的语义标记任务。

数据标记是与语义索引不同的概念，它是对输入数据语料库进行语义索引的另一种方法。在语义索引中，将深度学习抽象表示直接用于数据索引目的。而在数据标记中，抽象数据表示被视为执行数据标记的区分任务的特征，因此，这种对数据的标记也可以用于数据索引。将简单的线性建模方法应用于深度学习算法进行复杂特征的提取，可以实现深度学习对大量数据的标记。本节的其余部分主要讨论在数据标记的区分性任务中使用深度学习的一些成果。

在 ImageNet 计算机视觉竞赛中，Hinton 等人展示了一种使用深度学习和卷积神经网络的方法[20]，该方法优于其他现有的图像对象识别方法。Hinton 的团队利用 ImageNet 数据集展示了深度学习对改进图像搜索的重要性，需要注意的是，ImageNet 数据集是目前图像对象识别的最大数据集之一。Dean 等人通过使用类似的深度学习建模方法，以及用于训练人工神经网络的大规模软件基础设施，在 ImageNet 上取得了进一步的成功[15]。

文献[21] [22] [23]尝试使用受限玻尔兹曼机器(RBM)、自动编码器和稀疏编码等方法从未标记的图像数据中学习和提取特征。但是这些方法只能提取低级特征，例如边缘和斑点检测。深度学习还可以用于构建非常高级的图像检测功能，例如，谷歌和斯坦福开发了一个非常大的深度神经网络，它能够通过使用未标记的数据从零开始学习非常高级的特，如在没有任何先验知识的情况下，进行人脸检测或猫检测[24]。他们的工作是对仅使用未标记(无监督)数据的深度学习构建高级特征的可行性进行大规模调查，并清楚地证明了使用无监督数据的深度学习的好处。在谷歌的实验中，他们对从互联网上随机下载的 1000 万张 200 × 200 的图像训练了一个 9 层本地连接的稀疏自动编码器。该模型有 10 亿个连接，训练时间持续 3 天。一个由 1000 台机器和 16,000 个内核组成的计算集群用于训练具有模型并行性和异步 SGD (随机梯度下降)的网络。在他们的实验中，他们获得了功能类似于人脸检测器、猫检测器和人体检测器的神经元，基于这些特征，他们的方法也优于最新技术，并从 ImageNet 数据集中识别了 22,000 个对象类别。这证明了通过深度学习算法提取的抽象表示在新数据或不可见数据上的泛化能力，即使用从给定数据集提

取的特征在另一个数据集上成功执行区分任务。虽然谷歌的工作涉及到一个问题，即仅仅使用未标记的数据是否有可能构建人脸特征检测器，但通常在计算机视觉中，标记的图像用于学习有用的特征[25]。例如，可以使用一大组人脸图像，在人脸周围有一个边界框来学习人脸检测器功能。然而，传统上，它需要大量的标记数据才能找到最佳特征。图像数据采集中标记数据的稀缺性带来了一个具有挑战性的问题。

还有其他深度学习的成果探索了图像标记。Socher 等人介绍了递归神经网络，用于预测多模式图像的树结构，这是第一种深度学习方法在复杂图像场景的分割和注释方面取得了非常好的效果[26]。递归神经网络结构能够预测场景图像的层次树结构，优于基于条件随机场的其他方法或其他方法的组合，在分割、注释和场景分类方面也优于其他现有方法。该研究还表明，他们的算法是预测树结构的自然工具，通过使用它来解析自然语言的语义。这证明了深度学习作为从不同数据类型中提取数据表示的有效方法的优势。Kumar 等人提出，可以使用递归神经网络通过深度学习构建有意义的搜索空间，然后将搜索空间用于基于设计的搜索[27]。

Le 等人证明，通过使用独立变量分析从视频数据中学习不变性时空特征，深度学习可以用于动作场景识别和视频数据标记[28]。当与深度学习技术(如叠加和卷积)结合学习分层表示时，他们的方法优于其他现有方法。以前的作品用于将 SIFT 和 HOG 等图像的手工设计功能应用于视频领域。该项研究表明，直接从视频数据中提取特征是一个非常重要的研究方向，也可以推广到许多领域。

深度学习在提取有用的特征(即表示)以对图像和视频数据执行区分任务，以及从其他类型的数据中提取表示方面取得了显著的成果。这些具有深度学习的区别结果对于数据标记和信息检索非常有用，并且可以在搜索引擎中使用。因此，通过深度学习获得的高级复杂数据表示对于大数据分析中计算可行且相对简单的线性模型的应用非常有用。然而，仍有大量工作有待进一步探索，包括确定学习良好表征的适当目标，以便在大数据分析中执行区分性任务。

5. 结论与展望

与更传统的机器学习和特征工程算法相比，深度学习具有潜在的优势，可以提供一种解决方案来解决在大量输入数据中发现的数据分析和学习问题。更具体地说，它有助于从大量无监督数据中自动提取复杂的数据表示。这使得它成为大数据分析的一个有价值的工具，大数据分析涉及到对大量原始数据的分析，这些原始数据通常是无监督和未分类的。深度学习中不同层次的复杂数据抽象的分层学习和提取为大数据分析任务提供了一定程度的简化，特别是用于分析海量数据、语义索引、数据标记、信息检索及区分任务，如分类和预测。

在讨论文献中的成果及相关背景后，研究重点关注与深度学习和大数据相关的两个重要领域：1) 用于大数据分析的深度学习算法和架构的应用，以及 2) 大数据分析的某些特征和问题，如何为适应这些问题的深度学习算法带来独特的挑战。本文对深度学习研究和应用于不同领域的重要文献进行了有针对性的调查，以确定深度学习如何用于大数据分析的不同目的。

深度学习领域的低成熟度值得深入研究。特别是，我们需要做更多的工作，研究如何使深度学习算法适应与大数据相关的问题，包括高维度、流式数据分析、深度学习模型的可扩展性、改进的数据抽象形式、分布式计算、语义索引、数据标记、信息检索，用于提取良好数据表示的标准，以及域自适应。未来的工作应侧重于解决大数据中常见的一个或多个问题，从而有助于深入学习和大数据分析研究。

基金项目

大庆市指导性科技计划项目(ZD-2020-12)；黑龙江省自然科学基金项目(LH2021F001)。

参考文献

[1] 袁波. 大数据领域的反垄断问题研究[D]: [博士学位论文]. 上海: 上海交通大学, 2019.

[2] 孟小峰, 慈祥. 大数据管理: 概念、技术与挑战[J]. 计算机研究与发展, 2013, 50(1): 146-169.

[3] Bengio, Y. (2020) Deep Learning of Representations: Looking Forward. In: *Proceedings of the* 1*st International Conference on Statistical Language and Speech Processing*, Springer, Tarragona, 1-37. https://doi.org/10.1007/978-3-642-39593-2_1

[4] 张菊, 郭永峰. 深度学习研究综述[J]. 教学研究, 2021, 44(3): 6-13.

[5] Bengio, Y. and LeCun, Y. (2007) Scaling Learning Algorithms towards AI. In: Bottou, L., Chapelle, O., DeCoste, D. and Weston, J., Eds., *Large Scale Kernel Machines*, MIT Press, Cambridge, Vol. 34, 321-360.

[6] Dumbill, E. (2012) What Is Big Data? An Introduction to the Big Data Landscape. *O'Reilly Strata Making Data Work Conference*, Santa Clara, 28 February-1 March 2012, 315-450.

[7] Garshol, L.M. (2013) Introduction to Big Data/Machine Learning. Online Slide Show. http://www.slideshare.net/larsgaintroduction-to-big-datamachine-learning

[8] Grobelnik, M. (2013) Big Data Tutorial. European Data Forum. https://european-big-data-value-forum.eu/

[9] Salton, G. and Buckley, C. (1988) Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, **24**, 513-523. https://doi.org/10.1016/0306-4573(88)90021-0

[10] Robertson, S.E. and Walker, S. (1994) Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In: *Proceedings of the* 17*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Springer-Verlag, New York, 232-241. https://doi.org/10.1007/978-1-4471-2099-5_24

[11] Hinton, G. and Salakhutdinov, R. (2011) Discovering Binary Codes for Documents by Learning Deep Generative Models. *Topics in Cognitive Science*, **3**, 74-91. https://doi.org/10.1111/j.1756-8765.2010.01109.x

[12] 刘芳名, 张鸿. 基于多级语义的判别式跨模态哈希检索算法[J]. 计算机应用, 2021, 41(8): 48-56.

[13] Ranzato, M. and Szummer, M. (2008) Semi-Supervised Learning of Compact Document Representations with Deep Networks. In: *Proceedings of the* 25*th International Conference on Machine Learning*, ACM, New York, 792-799. https://doi.org/10.1145/1390156.1390256

[14] Mikolov, T., Chen, K. and Dean, J. (2013) Efficient Estimation of Word Representations in Vector Space. CoRR: Computing Research Repository, 1-12.

[15] Dean, J., Corrado, G., Monga, R., *et al*. (2012) Large Scale Distributed Deep Networks. In: Bartlett, P., Pereira, F.C.N., Burges, C.J.C., Bottou, L. and Weinberger, K.Q., Eds., *Advances in Neural Information Processing Systems*, Vol. 25, Curran Associates, Inc., Red Hook, 1232-1240.

[16] Mikolov, T., Le, Q.V. and Sutskever, I. (2013) Exploiting Similarities among Languages for Machine Translation. CoRR: Comput Res Repository, 1-10.

[17] Li, G., Zhu, H., Cheng, G., Thambiratnam, K., *et al*. (2012) Context-Dependent Deep Neural Networks for Audio Indexing of Real-Life Data. *IEEE Spoken Language Technology Workshop* (*SLT*), Miami, 2-5 December 2012, 143-148. https://doi.org/10.1109/SLT.2012.6424212

[18] 马腾腾, 赵宇翔, 朱庆华. 国外移动视觉搜索产品的比较分析研究[J]. 图书馆杂志, 2016, 35(9): 12-18.

[19] 魏正曦, 邱玲, 赵攀. 基于灰度分类的图像搜索引擎[J]. 四川理工学院学报(自然科学版), 2021, 27(1): 68-77.

[20] Krizhevsky, A., Sutskever, I. and Hinton, G. (2012) Imagenet Classification with Deep Convolutional Neural Networks. In: *Advances in Neural Information Processing Systems*, Curran Associates, Inc., Red Hook, Vol. 25, 1106-1114.

[21] Hinton, G.E., Osindero, S. and The, Y.-W. (2006) A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, **18**, 1527-1554. https://doi.org/10.1162/neco.2006.18.7.1527

[22] Hinton, G.E. and Salakhutdinov, R.R. (2006) Reducing the Dimensionality of Data with Neural Networks. *Science*, **313**, 504-507. https://doi.org/10.1126/science.1127647

[23] Lee, H., Battle, A., Raina, R. and Ng, A. (2006) Efficient Sparse Coding Algorithms. In: *Advances in Neural Information Processing Systems*, MIT Press, Cambridge, 801-808.

[24] Le, Q., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G., Dean, J. and Ng, A. (2012) Building High-Level Features Using Large Scale Unsupervised Learning. *Proceeding of the* 29*th International Conference on Machine Learning*, Edinburgh, 26 June-1 July 2012, 8595-8598. https://doi.org/10.1109/ICASSP.2013.6639343

[25] 朱均安. 基于深度学习的视觉目标跟踪算法研究[D]: [博士学位论文]. 北京: 中国科学院大学(中国科学院长春

光学精密机械与物理研究所), 2020.

[26] Socher, R., Lin, C.C., Ng, A. and Manning, C. (2011) Parsing Natural Scenes and Natural Language with Recursive Neural Networks. In: *Proceedings of the* 28*th International Conference on Machine Learning*, Omnipress, Madison, 129-136.

[27] Kumar, R., Talton, J.O., Ahmad, S. and Klemmer, S.R. (2012) Data-Driven Web Design. *Proceedings of the* 29*th International Conference on Machine Learning*, Edinburgh, 26 June-1 July 2012, 119-130.

[28] Le, Q.V., Zou, W.Y., Yeung, S.Y. and Ng, A.Y. (2011) Learning Hierarchical Invariant Spatio-Temporal Features for Action Recognition with Independent Subspace Analysis. *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Colorado, 20-25 June 2011, 3361-3368. https://doi.org/10.1109/CVPR.2011.5995496