深度学习在游戏中的应用

郭潇逍 [1] 李 程 [2] 梅俏竹 [1, 2]

摘 要 综述了近年来发展迅速的深度学习技术及其在游戏 (或博弈) 中的应用. 深度学习通过多层神经网络来构建端对端的从输入到输出的非线性映射, 相比传统的机器学习模型有显见的优势. 最近, 深度学习被成功地用于解决强化学习中的策略评估和策略优化的问题, 并于多种游戏的人工智能取得了突破性的提高. 本文详述了深度学习在常见游戏中的应用.

关键词 深度学习, 博弈, 深度强化学习, 围棋, 人工智能

引用格式 郭潇逍, 李程, 梅俏竹. 深度学习在游戏中的应用. 自动化学报, 2016, **42**(5): 676−684

DOI 10.16383/j.aas.2016.y000002

Deep Learning Applied to Games

GUO Xiao-Xiao[1] LI Cheng[2] MEI Qiao-Zhu[1, 2]

Abstract In this article, we present a survey of recent deep learning techniques and their applications to games. Deep learning aims to learn an end-to-end, non-linear mapping from the input to the output through multi-layer neural networks. Such architecture has several significant advantages as compared to traditional machine learning models. There has been a flurry of recent work on combining deep learning and reinforcement learning to better evaluate and optimize game policies, which has led to significant improvements of artificial intelligence in multiple games. We systematically review the use of deep learning in well-known games.

Key words Deep learning, games, deep reinforcement learning, Go, artificial intelligence

Citation Guo Xiao-Xiao, Li Cheng, Mei Qiao-Zhu. Deep learning applied to games. *Acta Automatica Sinica*, 2016, **42**(5): 676−684

2016 年是载入人工智能史册的一年. Alphabet (原 Google) 旗下的 DeepMind 公司研发的计算机围棋程序 AlphaGo 成功地打败了近 15 年来一直被认为是世界顶尖棋手的李世石九段. 这距 IBM 的深蓝 (Deep Blue) 程序击败国际象棋棋王卡斯帕罗夫正好二十年, 也再一次在学术界和民间掀起了人工智能的热潮. 与深蓝不同的是, AlphaGo 的成功极大程度上归功于其采用了深度学习的算法. 本文从一个更广的角度来介绍深度学习在博弈中的应用.

1 深度学习 (Deep Learning)

深度学习是近年来大放异彩的一种机器学习模式. 其主要的方法是通过训练多层的神经网络 (Neural networks) 以达到更好的学习效果. 常见的多层网络结构包括多层感知器 (Multilayer perceptron, MLP)、卷积神经网络 (Convolutional neural network, CNN) 和递归神经网络 (Recurrent neural network, RNN) 等. 多层神经网络的理论在 80 年代即已被广泛研究[1−4], 但一直到最近十年, 由于训练算法与计算能力的局限, 研究者普遍只能成功地训练两层或者三层的神经网络 (卷积神经网络是一个例外). 更多层的神经网络反而让学习结果变差[5]. 2006 年, 多伦多大学的 Hinton 及其合作者提出了深度置信网络 (Deep belief networks, DBN). 其使用非监督学习对神经网络的每一层进行分别训练, 从而能够成功地训练具有多层网络结构的限制性玻尔兹曼机 (Restricted Boltzmann machine)[6]. 类似的利用非监督学习来分层训练的方法也适用于其他的深度网络结构[7−8]. 其后, 蒙特利尔大学的研究者深入分析了非监督学习对于深层结构的帮助[9] 以及原始训练方法失败的原因[10], 并提出了适用于深层结构的参数初始化方法[10] 和激活函数 (Activation function)[11]. 随着训练算法和计算能力瓶颈的突破 (尤其是对图形处理 (Graphics processing unit, GPU) 和高性能计算 (High-performance computing, HPC) 的使用), 深度学习被广泛应用于人工智能相关的领域, 并在多个研究问题上取得了巨大进展. 其典型应用场景包括图像处理中的图像分类 (Image classification)[12]、物体检测 (Object detection)[13−14]、视频分类 (Video classification)[15]、场景解析 (Sense parsing)[16] 和阴影检测 (Shadow detection)[17], 语音理解中的语音识别 (Speech recognition)[18]、韵律预测 (Prosody contour predic-

收稿日期 2016-04-22 录用日期 2016-05-10
Manuscript received April 22, 2016; accepted May 10, 2016
本文责任编委 周志华
Recommended by Associate Editor ZHOU Zhi-Hua
1. 密歇根大学电子工程与计算机系 密歇根州安娜堡市 MI 48109 美国 2. 密歇根大学信息学院 密歇根州安娜堡市 MI 48109 美国
1. Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA 2. School of Information, University of Michigan, Ann Arbor, MI 48109, USA

tion)[19] 和韵律预测 (Prosody contour prediction)[19] 和文本到语音的合成 (Text-to-speech synthesis)[20−21], 自然语言处理中的句法分析 (Parsing)[22]、机器翻译 (Machine translation)[23−24] 和上下文实体链接 (Contextual entity linking)[25] 以及数据挖掘中的情感分析 (Sentiment analysis)[26] 和信息检索 (Information retrieval)[27−28] 等. 详情可参阅关于深度学习的综述文章[5, 29−30].

总的来说, 在有大量训练样本存在的情况下, 深度学习在预测问题 (如分类和回归) 上往往有很好的表现. 为什么相比传统的有监督学习的模式, 深度学习如此有效呢? 我们认为其原因大致可以归结于三点: 端对端学习的思想、非线性学习的能力以及面对大规模数据的可扩展性.

1.1 端对端学习 (End-to-end Learning)

传统机器学习的流程往往分作多个独立的模块. 这些模块之间在训练过程中并不传递学习误差 (Error propagation), 因此前一个模块也不会根据之后模块的训练结果进行调整. 例如在一个典型的自然语言处理的问题中, 传统的方法会依次采取分词 (Tokenization)、词性标注 (POS tagging)、句法分析 (Parsing)、语义分析 (Semantic analysis) 等独立的步骤. 在这样的流程中, 每一个步骤是一个独立的学习任务, 都需要大量标注好的训练样本. 为不同的学习任务标注大量的训练样本代价高昂, 而每一个步骤作出的错误预测 (不管由什么原因产生) 也都会影响之后的任务. 即使在一个单独的学习任务中 (如句法分析), 特征提取 (Feature extraction) 也往往是一个独立于训练的预处理步骤. 而在深度学习的流程中, 可训练的层次化表示 (Trainable hierarchical representation) 取代了预定义的特征表示, 而每一层的表示都会在训练中根据之后各层传递的误差信息来进行调整, 从而有利于对目标函数的优化[31]. 另一方面, 在深度学习中每一层的训练都不依赖于额外的学习目标和训练样本 (当然, 深度学习也可以利用额外的标记给中间层的训练提供副目标函数). 这样不仅能让有限的资源优先用于为最终的学习目标标记训练样本, 也能灵活地利用现成的有标记的数据集.

图 1 展示了一个在视频游戏中使用深度学习的范例. 我们可以看到一个多层的卷积神经网络被用来端对端地学习从输入 (游戏屏幕) 到输出 (游戏控制信号) 的映射.

1.2 非线性学习 (Non-linear Learning)

深度学习通过堆叠多层的神经网络来构建从原始的数据输入到最终的预测目标的映射函数. 传统的机器学习模型擅于寻找数据到输出目标的线性变换, 但在现实中, 从数据输入到目标输出的映射往往是复杂的、非线性的. 在深层神经网络中, 每一层都是一个非线性的从输入到输出的映射. 这些多层的非线性变换形成了数据的层次化表示, 随着层数的增加, 表示也更加抽象、更加普适 (Invariant). 例如, 对于图像数据, 深度学习结构首先提取出关于物体边界的表示, 之后的网络层再从边界的表示中提取出关于物体部件的特征, 而物体部件的特征进而作为下一层神经网络的输入从而得到关于物体的特征向量[32]; 对于文本数据, 深度学习结构首先提取出词的特征表示 (Word representation), 之后的网络层再从词的特征通过非线性变换提取出语句的特征, 而语句的特征作为下一层的神经网络输入从而得到关于文章的特征向量. 虽然从理论上讲, 超过两层的神经网络结构即可表达任意函数, 但更深层的网络结构对于特定的函数集合具有更高效的表达能力, 从而能够通过更少的参数 (如少数的网络结点) 来表达复杂的函数[33−36]. 这种高效的表达能力对于复杂的人工智能任务是必须的. 深层网络的表达优势对于神经网络的学习具有明显的帮助: 底层和高层的变量可以共享统计特征, 从而提高学习效率. 因为两层的神经网络中并没有这种层次化的特征表示, 我们一般不认为两层的网络是深度学习结构. 同样, 直接建立在原始输入空间的非线性模型, 例如决策树和支持向量机 (非线性 Kernel), 也不被认为是深度学习模型[36].

1.3 可扩展性 (Scalability)

学习复杂的、非线性的函数当然需要更多的训练样本, 训练样本太少会导致严重的过拟合 (Overfitting) 问题. 这不仅让如何获取大量的标注数据成为关键, 也对机器学习面对大量训练样本的可扩展性提出了考验. 虽然传统的机器学习算法 (如 SVM (Support vector machine) 和 Logistic regression) 也能通过核方法 (Kernel method) 来学习非线性转换, 但这样的计算复杂度很高: 至少是 $\mathrm{O}(N^2)$, N 为训练数据样本数. 当数据量太大时, 这样的算法显然困难重重. 深度学习的训练方法多基于随机梯度下降 (Stochastic gradient descent), 不仅不需计算训练样本的两两关系, 甚至往往不需遍历所有的训练样本, 从而可以灵活地利用更大的数据集. 同时, 反向传播算法 (Backpropagation) 能快速地计算整个网络的梯度, 并让训练误差有效地传播到底层的特征空间[37]. 随着训练样本的增加, 传统监督学习算法的性能往往出现收益递减 (Diminishing return) 的现象. 相反, 深度学习算法的效果则可以随着可训练数据集的增加取得明显的提升. 这种优势源于深层网络结构灵活的表达能力, 让深度学习可以用简洁的结构表达复杂的函数集合[5, 38].

以上我们简单介绍了深度学习的基本思想、优点和应用, 接下来让我们看看它是如何被应用在游

戏中的.

2 博弈中的深度学习

博弈, 或称 “玩游戏” (Game playing), 是人工智能的经典问题. 能够客观判定胜负的游戏不仅为人工智能算法提供了完美的测试平台, 也让计算机玩家和人类玩家得以对战比较. 根据不同的分类方法, 游戏可以被分为单人游戏和双人 (或多人) 游戏, 棋盘游戏和视频游戏, 协作游戏和对战游戏等. 双人或更多人的游戏常被称为博弈, 而博弈又可被分为完全信息博弈 (Perfect information games, 如围棋与象棋) 和非完全信息博弈 (如扑克和军棋). 早期的人工智能算法往往依赖于搜索算法, 这在简单的、搜索空间有限的游戏 (如 Tic-Tac-Toe) 中非常有效. 随着搜索难度的增加, 更复杂的游戏往往采用搜索与机器学习相结合的算法, 尤其是强化学习 (Reinforcement learning) 的算法.

强化学习本身并非深度学习. 它是机器学习的一个重要分支, 其着重解决顺序性决策问题[39−40]. 很多游戏中目前最先进的算法都是基于强化学习. 比如在 AlphaGo 之前的效果最好的围棋算法 (如 CrazyStone 与 Zen) 即是将强化学习与蒙特卡洛树搜索 (Monte-Carlo tree search, MCTS) 相结合. 相比于非监督学习和监督学习, 强化学习关注两大特有的问题: 探索与开发的权衡 (Exploration vs. exploitation) 和有时序的信用分配 (Temporal credit assignment). 一个强化学习算法需要回答两个基本问题: 1) 如何评估一个策略 (Policy); 2) 如何找到一个问题的最优策略. 深度学习在游戏中的应用, 往往是通过协助强化学习来更好地解决这两个问题 (例如图 1 所示).

2.1 深度强化学习 (Deep Reinforcement Learning)

强化学习传统的研究重点是学习表格化表示 (Tabular representation) 或线性函数近似 (Linear function approximation)[39]. 对于现实中的大规模和复杂的顺序决策过程, 简单的表格化表示和线性的近似是不够的. 深度学习能为强化学习提供端对端的、非线性的函数近似, 从而使得强化学习能够解决更现实更复杂的问题 (比如如何表达围棋中棋盘的状态). 另一方面, 为了解决游戏中常见的部分可观察的马尔科夫决策问题 (Partially observable Markov decision problems, POMDP), 强化学习的算法需要有效地处理动作和观测数据的序列 (比如有效地将历史动作和观测数据概括为一个状态表示), 以找到最优策略. 在音频、视频和文本等领域中, 深度学习已经被证明能够成功地学习序列的表示, 因此其也被用来学习 POMDP 中的状态表示[41−43]. 以深度学习优化的强化学习算法, 被通称为深度强化学习. 深度强化学习的综述可参见文献 [30].

2.2 深度强化学习在游戏中的应用

最近几年, 深度强化学习被越来越多地应用在游戏中, 包括多种单人游戏和多人游戏. 前文已经提到深度学习的优点在于能从大量的训练数据中学习出非线性的表示, 从而达到更好的预测效果. 当深度学习被应用在一个新的领域的时候, 我们需要关注三个问题: 预测的目标是什么? 训练数据从哪里来? 学习的结果表示什么?

接下来我们以一类文献中常见的游戏为例来梳理深度强化学习在游戏中的发展. 我们发现, 当深度学习被引入到游戏的时候, 其预测目标往往是某个游戏状态下每个可能的动作所对应的值 (Value) 或者概率 (即估值和策略), 其训练数据往往来源于计算机玩游戏的过程 (准确地说是成序列的状态 – 动作 – 奖励记录), 其学习的结果往往是对游戏的状态和策略的非线性表示.

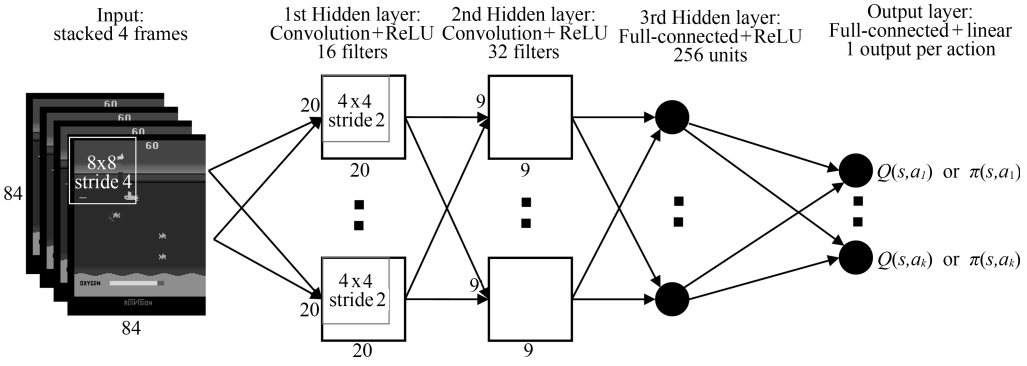

图 1 卷积神经网络学习从游戏屏幕到游戏策略的映射

Fig. 1 A convolutional neural network learns a mapping from game screens to game policy

3 雅达利游戏

Arcade 学习环境 (Arcade learning environment, ALE) 是一个人工智能研究者们公用的评测平台, 其作用类似于图像分类领域的 ImageNet 挑战[44]. ALE 提供了一个雅达利 2600 (ATARI 2600) 模拟器和大约 50 个游戏 (大多数是单人游戏)[44]. 这些游戏都建立在相同的宽 160 像素、高 210 像素的屏幕上, 每个像素有 128 种颜色. 每个游戏有不同的动作空间, 最多包括 18 个可能的动作. 要玩好这些游戏, 一个成功的算法需要同时解决游戏状态表示和策略选择两个挑战, 这几乎就是为深度强化学习度身定做的: 深度强化学习算法能从高维的、同时也是部分可观察的屏幕中学习当前游戏状态的表示, 同时也能从稀疏的、高度延迟的奖励信息 (Reward) 中学习最优策略.

经典的强化学习算法假定其值函数 (Value function) 可以用一张表来表示. 表中的每一个条目对应一个状态或者一个状态 + 动作的配对. 例如 Q-learning 算法[45] 就一张表来记录每一个状态+动作配对的值. 这些值在学习过程中使用如下公式更新:

$$Q(s_t,\, a_t) = Q(s_t,\, a_t) + \alpha(r_t + \gamma \max_b Q(s_{t+1},\, b) - Q(s_t,\, a_t))$$

其中 s_t, a_t, r_t 是在第 t 时刻的游戏状态、动作和奖励, s_{t+1} 是在第 $t+1$ 时刻的状态. 这样的表格表示适用于游戏状态和动作都不太多的情况. 当可能的游戏状态和动作数量庞大的时候, 强化学习面临的关键问题并不在于存储这张大表所需的空间, 而在于正确填写表中的目标值所需的时间和数据量. 因此这里真正的挑战在于泛化 (Generalization): 在只经历了有限的游戏状态的情况下, 如何得出对未经历的状态空间中可能的动作所对应的值. 强化学习中常用的泛化方法是函数近似. 函数近似从 (未知的) 目标函数中获得一些输入到输出的映射, 并试图将它们推广到整个函数定义域, 以构造对整个目标函数的近似. 例如, 在线性函数近似中, 一个动作的值函数被表示为:

$$Q(s,\, a;\, \theta) = \theta^{\mathrm{T}} \varphi(s,\, a)$$

其中 θ 为可被学习的参数, $\varphi(\cdot)$ 是定义在状态 + 动作的配对上的特征函数. 线性的函数近似方法从最近的游戏画面帧中提取人工设计好的特征, 再以这些特征的线性组合来表达和学习值函数. 阿尔伯塔大学的 Bellemare 等首先在 ALE 上使用了线性函数近似的 SARSA 算法和以下四种新的通用的人工设计的特征集: 1) 首先把屏幕分成交集的块, 对每个块用一个向量表示每种颜色是否出现, 并将这些向量的集合作为当前游戏画面的基本 (BASIC) 特征集; 2) 在基本特征集的基础上添加基本特征的配对组合, 这些组合构成了 BASS 特征集; 3) 首先提取屏幕上的物体, 并通过聚类对屏幕上的物体进行分类, 同时用多帧间信息来推断这些物体的位置和速度. 屏幕上所有物体的类别、位置和速度构成了当前游戏画面的 DISCO 特征集; 4) 对游戏画面使用局部敏感哈希 (Locality-sensitive Hashing), 并将所得的低维表示作为 LSH 特征集[44]. 他们在后续工作中提出了 Contingency awareness 的方法, 在原始特征集的基础上又加入了额外特征, 用于表示屏幕中哪些元素是直接受玩家输入的影响[46]. Bellemare 等随后又提出了通过使用 tug-of-war sketch 来进一步扩展特征集的方法[47]. 但总的来说, 基于线性值函数近似的强化学习算法要远远弱于人类玩家. 此外, 人工设计特征函数并不是一件容易的事情.

显然, 深度神经网络能使用非线性函数以高效地表示 $Q(s,\, a;\, \theta)$, 从而大大提高估值函数的表达能力. 可即便如此, 在 2013 年之前, 神经网络在强化学习上的成功应用却相当有限, 其中一个例外是 IBM Watson Research Center 的 Tesauro 成功地利用强化学习训练了神经网络来解决双陆棋 (Backgammon)[48]. 我们在此简单地分析深度学习在强化学习中的应用的障碍, 以及这些障碍是如何在近十年里被一一解决的.

1) 训练稳定性问题. 函数近似的思想无外乎是给没有出现过但相关联的状态 + 行动配对分配相似的值, 从而达到泛化的效果. 然而, 与监督学习不同的是, 强化学习中并没有已知的可用于训练的目标值, 它们需要通过算法的迭代更新来得到. 可问题是当某一个状态 + 动作配对被更新的时候, 这个更新所引导的参数权重的改变可能会影响到其他状态+动作配对的值. 这可能会导致这些状态+动作配对的值之前的迭代结果前功尽弃. 这个潜在问题通常导致学习时间变得很长, 甚至导致学习的失败. 解决这一稳定性问题需要两个非常重要的思想: 经验回放 (Experience replay) 以及目标值分离 (Target Q-separation). 经验回放的思想是, 当更新一个新的数据点时, 之前经历过的其他数据点也要被明确地考虑到[49]. 经验回放将所有先前经历的状态与动作 $(s,\, a,\, r,\, s')$ 存储到一个序列, 不妨称之为 D. 这个序列在每次更新 Q 函数时都会被重新使用去最小化目标函数:

$$L(\theta) = \mathrm{E}_{(s,\, a,\, r,\, s') \sim \mathrm{U}(D)}[(r + \gamma \max_b Q(s',\, b;\, \theta) - Q(s,\, a;\, \theta))^2]$$

其中 $\mathrm{U}(D)$ 表示在经验序列 D 上的一个均匀的随机分布. 基于均匀分布的经验回放打破了数据的相关性, 使学习回到了独立同分布 (i.i.d) 的状态. 经验

回放也使得强化学习算法能够从所有过去的策略中学习 (s, a, r, s'), 这使学习变得更稳定. 值得指出的是, 回放过多的经验会减缓学习进度. 如何基于学习进度确定经验回放的多少仍然是一个悬而未决的问题.

经验回放机制可以让学习更稳定, 但由于目标值 $(r+\gamma \max_b Q(s_{t+1}, b; \theta))$ 和当前估值 $Q(s, a; \theta)$ 之间的相关性, 训练中的振荡仍可能存在. 一个有效的解决方案是另用一个单独的神经网络来生成目标值. 比如深度 Q-网络 (Deep Q-network, DQN) 使用:

$$L(\theta) = \mathrm{E}_{(s, a, r, s')\sim \mathrm{U}(D)}[(r+ \gamma \max_b Q(s', b; \theta^-) - Q(s, a; \theta))^2]$$

其中 θ^- 是用于计算目标值 (Target) 的参数, 它只会周期性地与当前训练的值函数的参数 θ 同步[50]. 深度 Q-网络是经验回放和目标值分离两个思想的结合, 它通过 Q-learning 来训练卷积神经网络 (CNNs), 并在雅达利 2600 游戏上取得了巨大突破. 深度 Q-网络记录使用均匀分布来抽取训练样本, DeepMind 的 Schaul 等优化了经验回放的采样方法来提高深度 Q-网络在雅达利游戏上的成绩. 他们的想法是更多的随机选取拥有较大时间差分错误 (Temporal difference error) 的状态–动作配对[51].

另一种处理训练稳定性问题的方向是将强化学习转化为监督学习. 密歇根大学的 Guo 等使用慢速的蒙特卡洛树搜索生成少量的数据来训练快速的卷积神经网络, 卷积神经网络则通过数据集聚[52] 来模仿蒙特卡洛树搜索的行为[53]. 伯克利大学的 Schulman 等则使用信赖域策略优化 (Trust region policy optimization) 来使深度网络直接学出策略[54].

2) 估计偏差的问题. 在标准 DQN 中, 求最大值的运算符使用相同的值来选择和评估动作, 这使得它更容易选择被高估的估值, 导致过于乐观的估计, 而这种高估可能会使训练变得发散. Double DQN 的方法解耦了选择和评估, 从而降低了估计值偏差对于 DQN 的影响[55]:

$$L(\theta) = \mathrm{E}_{(s, a, r, s')\sim \mathrm{U}(D)}[(r+ \gamma Q(s', \arg\max_b Q(s', b; \theta); \theta^-) - Q(s, a; \theta))^2]$$

与 DQN 相比, Double DQN 中训练网络 θ 的权重可用于以 "贪婪法" 选择动作, θ^- 仍用于计算目标值[55]. Double DQN 是目前雅达利游戏中最先进的算法.

研究者们的另一个发现是最优的值函数会存在不一致性, 其对每个状态的任意次优动作的估值都偏大. 这是因为贝尔曼公式中 $Q^*(s, a)$ 描述的是一个不一致性策略: 当将来返回到当前游戏状态 s 时, 这一策略会选择 $\pi^*(s)$, 而不是当前所选的动作 a. 虽然保持全局一致性并不实际, Bellemare 等提出的贝尔曼公式的一致性算子 (Consistent Bellman operator), 为不一致性问题提供了一阶的解决办法. 这个一致性算子总体上对次优动作进行贬值, 但保留所有最优动作的值. 这当然会导致最优和次优动作的估值之间的差距增大. 在深度强化学习中, 增大这个差距可以减轻 Q 值估计的统计偏差对于学习的影响. Consistent Bellman operator 的使用成功的提高了 DQN 在雅达利游戏上的得分[56].

DeepMind 的 Wang 等提出了新的 "决斗" 神经网络架构 (Dueling network). 此架构分离了游戏状态的值函数和与状态相关的动作的优势函数, 这使得各个动作的估值不再独立. 不同动作共享更泛化的状态值函数对于减少对不同的动作的估计偏差很有用[57].

3) 部分可观察性问题. 有的深度强化学习算法利用递归神经网络来处理来处理雅达利游戏中的部分可观察性 (Partially observability), 比如 Deep recurrent Q network (DRQN). DRQN 的基础仍然是 DQN 框架, 其区别是在 CNN 上增加了一个长短期记忆 (Long short-term memory, LSTM) 模块. 研究发现在雅达利游戏中, DRQN 能更好地处理部分可观察性[41].

4) 如何加快训练速度. 基于 DQN 的方法在用 GPU 的情况下需要用 8 天的时间才能学习如何玩一种雅达利游戏. DeepMind 的 Mnih 等提出了异步优势演员–评论家方法 (Asynchronous advantage actor-critic, A3C) 以加快训练[58]. 该法即便使用 16 个 CPU, 学习一种游戏也只需要 4 天.

5) 多任务学习的问题. DQN 专注于单一游戏的学习, 其训练所得的网络只能用于一个游戏. 最近的一些研究考虑了如何将多个游戏中分别训练出的深度神经网络压缩为一个网络, 以使用同一网络玩多个游戏[59−60].

虽然并非耳熟能详的游戏, 雅达利游戏为深度学习在游戏中的发展提供了极佳的土壤和评测平台. 其中的很多思想和技术随即被应用在了其他的博弈问题中.

4 围棋

围棋是现存最复杂的完全信息博弈之一. 相比其他的棋盘游戏 (如国际象棋和西洋跳棋), 围棋的搜索空间极大, 局面的描述和评估都极难. 随着卷积神经网络在雅达利游戏上取得了巨大的成功, 人们期望同样的方法可以用来解决计算机围棋. 然而事实证明, 直接依样画葫芦来对计算机围棋的策略或值函数作近似是很难的. 究其原因, 还是因为围棋的策略和形势判断太过复杂. 由于每个棋盘状态真正的值是未知的, 如果没有足够好的初始策略作为强

化学习的基础, 利用不准确的动作和奖励信息训练出来的神经网络未必能做可靠的预测. 为了解决这个难题, 研究者们想到了用人类的棋谱做深度神经网络的训练集, 因为在人类棋谱中每步棋的选择和每盘棋的胜负都是已知的. 研究者们期待从人类棋谱中学习的策略更适合作为强化学习的初始策略.

爱丁堡大学的 Clark 等用人类棋手的历史棋谱来训练卷积神经网络, 从而可以预测人类棋手的策略 (即当前局面的下一手). 这样的神经网络也被称为策略网络. 它们在两个数据集上分别取得了 41.1 % 和 44.4 % 的预测准确率. 这些策略网络利用了专为围棋设计的特征集, 例如对称信息被硬编码为深度神经网络的输入, 棋盘上不合理的位置也被屏蔽了. 即便预测的准确率小于 50 %, 训练好的卷积神经网络也能打败知名的围棋程序 GnuGo (但输给了更先进的围棋程序 Fuego)[61].

多伦多大学的 Maddison 等和 DeepMind 的研究者们用一个网络对战平台 (KGS) 的历史棋谱训练了 12 层深的卷积神经网络, 其预测人类棋手的策略能达到 55 % 的准确率. 训练好的卷积神网经络对战 GnuGo 能达到 97 % 的赢率, 并能匹敌最先进的每步模拟 200 万次的蒙特卡洛树搜索算法[62]. 相似的办法也被 Tian 等用在 Facebook 研发的计算机围棋程序黑暗森林 (DarkForest) 中[63].

值得一提的是, 从人类专家棋谱中训练出来的策略网络存在偏差. 这是因为训练的目标 (预测下一手) 和实战的目标 (赢下比赛) 是不一致的. 为解决这个问题, AlphaGo 将用棋谱训练的策略网络再通过自我对局 (Self-play) 进行调整, 其思想和 Atari 游戏中用自战的序列来训练策略网络并无二致[64].

从人类棋谱中训练的策略网络让基本策略变得足够好, 这使得强化学习算法在自我对局中观察到的状态–动作–奖励信息变得更加接近最优策略的真实情况 (Ground-truth). AlphaGo 于是用自我对局的棋盘状态和胜负来训练一个价值网络, 以预测任意棋盘状态的值. 这个价值网络进一步提高了 AlphaGo 的棋力. 为了让训练数据尽可能独立, AlphaGo 从每盘自战中只选取一个局面来做训练[64]. 这个思想和 Atari 游戏中让经验回放的选取尽可能均匀随机是相似的.

在另一方面, 蒙特卡洛树搜索仍然是目前最先进的处理大型和复杂的顺序决策的算法, 包括计算机围棋. 随着随机模拟的次数变多, 搜索树变大, 对值的估计也变得更准确. 然而, 它的计算开销对于实际的围棋比赛来说太过于庞大. 因此, 研究者们也试图利用深度学习 (或者更简单的线性学习) 来帮助蒙特卡洛树搜索, 以提高对值预测的准确率或者减少搜索的计算开销. AlphaGo 推出了新的蒙特卡洛树搜索算法. 该算法结合蒙特卡罗模拟以及价值和策略网络来减少计算开销. 价值网络被用来评估当前棋盘的状态 (即预测一个给定局面的值), 从而有效地降低了搜索规划的深度. 策略网络则被用来选择下一步 (即预测一个给定局面的下一步棋的概率分布), 从而有效地降低了搜索的宽度. 这些网络的训练是人类专家动作的监督学习和自我游戏的强化学习的精密结合. AlphaGo 藉此力挫欧洲围棋冠军和世界顶尖棋手李世石[64].

5 扑克

扑克游戏的挑战来源于信息不完全, 即玩家对历史事件只能进行部分观察, 而看不到对手的信息. 阿尔伯塔大学的 Bowling 等人为德州扑克游戏提出了一个近似纳什均衡的解决方案. 此方案的基本思想是用两个 Regret 最小化算法之间反复自我对局[65]. 最近, Yakovenko 等和哥伦比亚大学的研究者们用类似的方法训练了针对扑克的卷积神经网络, 并在三种常见的扑克游戏中能和人类专家抗衡[66]. DeepMind 的 Heinrich 等和 Silver 等提出了更适用于非完全信息博弈的深度强化学习算法, 他们称之为 “神经虚拟自我对局” (Neural fictitious self-play, NFSP). NFSP 的主要思想是去近似博弈论中经典的 “虚拟对局” (Fictitious play) 模型, 其不借助于先验知识, 并能在二人零和游戏 (Zero-sum games) 或者多人势博弈 (Potential games) 中通过自我对局中收敛到纳什平衡[67].

6 其他游戏

Atari、围棋和扑克只是计算机游戏中的冰山一角. 有趣的是, 深度学习在其他的经典游戏中并没有被广泛应用. 这也许是因为计算机在多数经典游戏 (比如象棋和跳棋) 中早已能打败人类高手. 这从某种程度上让研究者们失去了进一步改善算法的兴趣. 即便如此, 深度学习或者更广泛的监督学习的思想还是散见于其他游戏中. 除了前文提到的双陆棋, 国际跳棋程序 Chinook 首先从人类专家的棋谱中提取和构建一个关于开局策略的数据库, 再通过 alpha-beta 搜索算法和以及对叶节点的策略评估函数来选取最优动作[68]. 最近的研究热点逐渐从棋牌游戏转向视频游戏, 比如微软把游戏 “我的世界 (Minecraft)” 作为测试人工智能的研究平台, 而 DeepMind 则声称他们的下一个挑战是星际争霸 (StarCraft). 相比棋盘游戏, 这些即时战略游戏的状态更复杂, 信息更不完全, 动作的选择空间也更大.

7 总结

博弈一直是人工智能的重要分支. 深度学习在其他领域的成功带给了游戏人工智能前所未有的启

发. 比如计算机围棋就经历了最早的基于规则的算法、基于启发式的局面评估的算法、基于蒙特卡洛树搜索的算法、基于强化学习的算法, 最终由深度强化学习带来质的飞跃. 在不久的将来, 我们会看到深度学习的算法和思想被应用于越来越多的游戏里. 未来的研究方向应该会从经典游戏逐渐偏向更复杂, 信息更不完全的多人游戏, 尤其是视频游戏. 游戏中的深度学习还有不少亟待解决的问题, 比如如何把已习得的知识应用到新问题中、如何更有效地利用专家知识、如何在自我对局中学到新的知识、如何根据对手改变游戏策略等. 我们也期待看到成功的游戏算法被应用在各行各业 (比如医疗和教育), 真正改善普通人的生活.

References

1 Werbos P. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences [Ph. D. dissertation], Harvard University, USA, 1974.

2 Parker D B. Learning Logic, Technical Report TR-47, MIT Press, Cambridge, 1985.

3 LeCun Y. Une procédure d′apprentissage pour Réseau à seuil assymétrique (a learning scheme for asymmetric threshold networks). In: Proceddings of the Cognitiva 85. Paris, France. 599−604 (in French)

4 Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors. *Nature*, 1986, **323**(6088): 533−536

5 Bengio Y. *Learning Deep Architectures for AI*. Hanover, MA: Now Publishers Inc, 2009.

6 Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006, **18**(7): 1527−1554

7 Ranzato M, Poultney C, Chopra S, LeCun Y. Efficient learning of sparse representations with an energy-based model. In: Proceedings of the 2007 Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2007.

8 Bengio Y, Lamblin P, Popovici D, Larochelle H. Greedy layer-wise training of deep networks. In: Proceedings of the 2007 Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2007.

9 Erhan D, Manzagol P A, Bengio Y, Bengio S, Vincent P. The difficulty of training deep architectures and the effect of unsupervised pre-training. In: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics. Clearwater, Florida, USA: AISTATS, 2009. 153−160

10 Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics. Sardinia, Italy: ICAIS, 2010.

11 Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, United States: ICAIS, 2011.

12 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 2014 International Conference on Learning Representations. Rimrock Resort Hotel, Banff, Canada: ICRR, 2014.

13 Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. In: Proceedings of the 2013 International Conference on Learning Representations. Scottsdale, Arizona: ICLR, 2013.

14 Szegedy C, Toshev A, Erhan D. Deep neural networks for object detection. In: Proceedings of the 2013 Advances in Neural Information Processing Systems. Lake Tahoe, Nevada: NIPS, 2013.

15 Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Li F F. Large-scale video classification with convolutional neural networks. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014.

16 Farabet C, Couprie C, Najman L, LeCun Y. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, **35**(8): 1915−1929

17 Khan S H, Bennamoun M, Sohel F, Togneri R. Automatic feature learning for robust shadow detection. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, OH, USA: IEEE, 2014.

18 Amodei D, Anubhai R, Battenberg E, Case C, Casper J, Catanzaro B, Chen J D, Chrzanowski M, Coates A, Diamos G, Elsen E, Engel J, Fan L X, Fougner C, Han T, Hannun A, Jun B, LeGresley P, Lin L, Narang S, Ng A, Ozair S, Prenger R, Raiman J, Satheesh S, Seetapun D, Sengupta S, Wang Y, Wang Z Q, Wang C, Xiao B, Yogatama D, Zhan J, Zhu Z Y. Deep speech 2: End-to-end speech recognition in English and Mandarin. preprint arXiv:1512.02595, 2015.

19 Fernandez R, Rendel A, Ramabhadran B, Hoory R. Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks. In: Proceedings of the 15th Annual Conference of International Speech Communication Association. Singapore: Curran Associates, Inc., 2014.

20 Fan Y C, Qian Q, Xie F L, Soong F K. TTS synthesis with bidirectional LSTM based recurrent neural networks. In: Proceedings of the 15th Annual Conference of International Speech Communication Association. Singapore: Curran Associates, Inc., 2014.

21 Sak H, Vinyals O, Heigold G, Senior A, McDermott E, Monga R, Mao M. Sequence discriminative distributed training of long short-term memory recurrent neural networks. In: Proceedings of the 15th Annual Conference of the International Speech Communication Association. Singapore: Curran Associates, Inc., 2014.

22 Socher R, Bauer J, Manning C D, Ng A Y. Parsing with compositional vector grammars. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria: ACL, 2013.

23 Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. In: Proceedings of the 2014 Advances in Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014.

24 Gao J F, He X D, Yih W T, Deng L. Learning continuous phrase representations for translation modeling. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore: ACL, 2014.

25 Gao J F, Deng L, Gamon M, He X D, Pantel P. Modeling Interestingness with Deep Neural Networks, US Patent 20150363688, December 17, 2015.

26 Socher R, Perelygin A, Wu J Y, Chuang J, Manning C D, Ng A Y, Potts C. Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP). Seattle, Washington: EMNLP, 2013.

27 Shen Y L, He X D, Gao J F, Deng L, Mesnil G. A latent semantic model with convolutional-pooling structure for information retrieval. In: Proceedings of the 23rd ACM International Conference on Information and Knowledge Management. New York, NY, USA: ACM, 2014.

28 Huang P S, He X D, Gao J F, Deng L, Acero A, Heck L. Learning deep structured semantic models for web search using clickthrough data. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. New York, NY, USA: ACM, 2013.

29 Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013. 1798−1828, DOI: 10.1109/TPAMI.2013.50

30 Schmidhuber J. Deep learning in neural networks: an overview. *Neural Networks*, 2015, **61**: 85−117

31 LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, **521**(7553): 436−444

32 Lee H, Grosse R, Ranganath R, Ng A Y. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: Proceedings of the 26th Annual International Conference on Machine Learning. New York, NY, USA: ACM, 2009.

33 Yao A C C. Separating the polynomial-time hierarchy by oracles. In: Proceedings of the 26th Annual Symposium on Foundations of Computer Science. Portland, OR, USA: IEEE, 1985. 1−10

34 Hastad J. Almost optimal lower bounds for small depth circuits. In: Proceedings of the 18th Annual ACM Symposium on Theory of Computing. New York, NY, USA: ACM, 1986.

35 Braverman M. Poly-logarithmic independence fools bounded-depth Boolean circuits. *Communications of the ACM*, 2011, **54**(4): 108−115

36 Bengio Y, Delalleau O. On the expressive power of deep architectures. *Algorithmic Learning Theory*. Berlin Heidelberg: Springer, 2011. 18−36

37 Le Cun Y, Boser B, Denker J S, Henderson D, Howard R E, Hubbard W, Jackel L D. Handwritten digit recognition with a back-propagation network. In: Proceedings of the 1990 Advances in Neural Information Processing Systems. San Francisco: Morgan Kaufmann, 1990.

38 Bengio Y, LeCun Y, DeCoste D, Weston J. Scaling learning algorithms towards AI. *Large-Scale Kernel Machines*. Cambridge: MIT Press, 2007.

39 Sutton R S, Barto A G. *Reinforcement Learning: An Introduction*. Cambridge: MIT Press, 1998.

40 Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 1996, **4**: 237−285

41 Hausknecht M, Stone P. Deep recurrent q-learning for partially observable MDPS. In: Proceedings of the 2015 AAAI Fall Symposium Series. The Westin Arlington Gateway, Arlington, Virginia: AIAA, 2015.

42 Bakker B, Zhumatiy V, Gruener G, Schmidhuber J. A robot that reinforcement-learns to identify and memorize important previous observations. In: Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. Manno-Lugano, Switzerland: IEEE, 2003

43 Wierstra D, Förster A, Peters J, Schmidhuber J. Recurrent policy gradients. *Logic Journal of IGPL*, 2010, **18**(5): 620−634

44 Bellemare M, Naddaf Y, Veness J, Bowling M. The arcade learning environment: an evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 2013, **47**: 253−279

45 Watkins C J H, Dayan P. Technical note: Q-learning. *Machine Learning*, 1992, **8**(3−4): 279−292

46 Bellemare M G, Veness J, Bowling M. Investigating contingency awareness using Atari 2600 games. In: Proceedings of the 26th AAAI Conference on Artificial Intelligence. Toronto, Ontario: AIAA, 2012.

47 Bellemare M G, Veness J, Bowling M. Sketch-based linear value function approximation. In: Proceedings of the 26th Advances in Neural Information Processing Systems. Lake Tahoe, Nevada, USA: NIPS, 2012.

48 Tesauro G. TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, 1994, **6**(2): 215−219

49 Riedmiller M. Neural fitted Q iteration–first experiences with a data efficient neural reinforcement learning method. In: Proceedings of the 16th European Conference on Machine Learning. Porto, Portugal: Springer, 2005.

50 Mnih V, Kavukcuoglu K, Silver D, Rusu A A, Veness J, Bellemare M G, Graves A, Riedmiller M, Fidjeland A K, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D. Human-level control through deep reinforcement learning. *Nature*, 2015, **518**(7540): 529−533

51 Schaul T, Quan J, Antonoglou I, Silver D. Prioritized experience replay. In: Proceedings of the 2016 International Conference on Learning Representations. Caribe Hilton, San Juan, Puerto Rico: ICLR, 2016.

52 Ross S, Gordon G J, Bagnell J A. A reduction of imitation learning and structured prediction to no-regret online learning. In: Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Ft. Lauderdale, FL, USA: AISTATS 2011.

53 Guo X X, Singh S, Lee H, Lewis R, Wang X S. Deep learning for real-time ATARI game play using offline Monte-Carlo tree search planning. In: Proceedings of the 2014 Advances in Neural Information Processing Systems. Cambridge: The MIT Press, 2014.

54 Schulman J, Levine S, Moritz P, Jordan M, Abbeel P. Trust region policy optimization. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: ICML, 2015.

55 van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, Arizona USA: AIAA, 2016.

56 Bellemare M G, Ostrovski G, Guez A, Thomas P S, Munos R. Increasing the action gap: new operators for reinforcement learning. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, Arizona USA: AIAA, 2016.

57 Wang Z Y, Schaul T, Hessel M, van Hasselt H, Lanctot M, de Freitas N. Dueling network architectures for deep reinforcement learning. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: ICML, 2016.

58 Mnih V, Badia A P, Mirza M, Graves A, Lillicrap T P, Harley T, Silver D, Kavukcuoglu K. Asynchronous methods for deep reinforcement learning. preprint arXiv:1602.01783, 2016.

59 Rusu A A, Colmenarejo S G, Gulcehre C, Desjardins G, Kirkpatrick J, Pascanu R, Mnih V, Kavukcuoglu K, Hadsell R. Policy distillation. In: Proceedings of the 2016 International Conference on Learning Representations. Caribe Hilton, San Juan, Puerto Rico: ICLR, 2016.

60 Parisotto E, Ba J L, Salakhutdinov R. Actor-mimic: Deep multitask and transfer reinforcement learning. In: Proceedings of the 2016 International Conference on Learning Representations. Caribe Hilton, San Juan, Puerto Rico: ICLR, 2016.

61 Clark C, Storkey A. Training deep convolutional neural networks to play go. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: ICML, 2015.

62 Maddison C J, Huang A, Sutskever I, Silver D. Move evaluation in Go using deep convolutional neural networks. In: Proceedings of the 2014 International Conference on Learning Representations. Rimrock Resort Hotel, Banff, Canada: ICRR, 2014.

63 Tian Y D, Zhu Y. Better computer go player with neural network and long-term prediction. In: Proceeding of the 2016 International Conference on Learning Representations. Caribe Hilton, San Juan, Puerto Rico: ICLR, 2016.

64 Silver D, Huang A, Maddison C J, Guez A, Sifre L, van den Driessche G, Dieleman S, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016, **529**(7587): 484−489

65 Bowling M, Burch N, Johanson M, Tammelin O. Heads-up limit hold′em poker is solved. *Science*, 2015, **347**(6218): 145−149

66 Yakovenko N, Cao L L, Raffel C, Fan J. Poker-CNN: a pattern learning strategy for making draws and bets in poker games. Tucson, Arizona: AIAA, 2005.

67 Heinrich J, Lanctot M, Silver D. Fictitious Self-Play in Extensive-Form Games. In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France: ICML, 2015.

68 Schaeffer J, Lake R, Lu P, Bryant M. CHINOOK the world man-machine checkers champion. *AI Magazine*, 1996, **17**(1): 21−29

郭潇逍 密歇根大学电子工程与计算机系博士研究生. 主要研究方向为深度学习和深度强化学习.
E-mail: guoxiao@umich.edu
(**GUO Xiao-Xiao** Ph. D. candidate in the Department of Electrical Engineering and Computer Science, University of Michigan. His research interest covers deep learning and deep reinforcement learning.)

李 程 密歇根大学信息学院博士研究生. 主要研究方向为数据挖掘与信息检索. E-mail: lichengz@umich.edu
(**LI Cheng** Ph. D. candidate at the School of Information, University of Michigan. Her research interest covers data mining and information retrieval.)

梅俏竹 密歇根大学信息学院和电子工程与计算机系副教授. 主要研究方向为大规模的数据挖掘, 信息检索和机器学习. 本文通信作者.
E-mail: qmei@umich.edu
(**MEI Qiao-Zhu** Associate professor at the School of Information and the Department of Electrical Engineering and Computer Science (EECS), University of Michigan. His research interest covers large-scale data mining, information retrieval, and machine learning. Corresponding author of this paper.)