硕士学位论文

基于深度学习的搜索广告点击率预测方法研究

RESEARCH ON CLICK-THROUGH RATE PREDICTION FOR SEARCH ADVERTISING BASED ON DEEP LEARNING

李思琴

哈尔滨工业大学

2015 年 6 月

国内图书分类号：TP391.3　　学校代码：10213
国际图书分类号：681.37　　密级：公开

工学硕士学位论文

基于深度学习的搜索广告点击率预测方法研究

硕士研究生：李思琴
导　　师：林磊副教授
申请学位：工学硕士
学　　科：计算机科学与技术
所在单位：计算机科学与技术学院
答辩日期：2015 年 6 月
授予学位单位：哈尔滨工业大学

Classified Index: TP391.3
U.D.C: 681.37

Dissertation for the Master Degree in Engineering

RESEARCH ON CLICK-THROUGH RATE PREDICTION FOR SEARCH ADVERTISING BASED ON DEEP LEARNING

Candidate： Li Siqin
Supervisor： Associate Prof. Lin Lei
Academic Degree Applied for： Master of Engineering
Speciality： Computer Science and Technology
Affiliation： School of Computer Science and Technology
Date of Defence： June, 2015
Degree-Conferring-Institution： Harbin Institute of Technology

摘 要

随着互联网技术的飞速发展，海量数据背景下的网络营销方式越来越受青睐。在线广告作为新的广告形式应运而生，展现出了巨大的市场潜力和商业价值，搜索广告是规模最大，增长最快的在线广告形式，它根据用户搜索的内容投放相关的广告，如今已经成为互联网行业的主要收入来源之一。搜索广告背后最为关键的技术是广告点击率的预测，它不但关系到广告投放的排名，也影响着广告点击的收费。因此，如何有效的利用海量历史数据对搜索广告的点击率进行预测是一项非常有意义的工作。目前已有的工作大多基于浅层模型进行搜索广告的点击率预测，浅层模型在特征学习方面是直接使用统计学习方法计算得到的特征，特征中每一维的含义固定并且孤立，不能表达内部之间的关系。

本文研究的目标是通过给定的信息预测搜索广告的点击率，通过使用深度学习模型，挖掘更多的特征之间的关系，从而能更有效的提高预测的结果。具体地，本文主要包含如下三方面的研究内容。

第一、本文从搜索广告点击率预测的定义出发，分析了数据集的数据的分布和特点并对数据集进行了预处理，在此基础上，本文根据对搜索广告的认识和在实际应用中的特性，提取了六类不同的特征。其次，针对深度学习在搜索广告点击率预测应用中的训练耗时和内存限制，本文设计了一种基于 GPU 计算的分块实现方案。

第二、本文首先使用了朴素贝叶斯模型、逻辑斯蒂回归模型和支持向量回归模型等主流方法对点击率进行预测，并分析了他们的不足。进而使用基于深度神经网络模型的搜索广告点击率预测的方法，我们使用 dropout 方法来降低在训练时过拟合造成的影响。实验结果表明，在特征相同的情况下，本文使用的深度神经网络模型方法能取得比主流方法更好的预测结果。

第三、本文提出了面向搜索广告点击率预测的卷积神经网络模型，通过基于局部窗口概念的卷积操作和亚采样操作，完成了从局部到整体的特征学习。在 KDD Cup 2012 中 Track 2 数据集上的实验结果表明，本文所使用的基于卷积神经网络的搜索广告点击率预测的方法能有效的提高点击率预测的结果。

关键词：点击率预测；搜索广告；深度学习；深度神经网络；卷积神经网络

Abstract

With the flourishing development of Internet Technology, it becomes more and more popular for network marketing in the background of big data. As a new form of advertising, online advertising shows a huge market potential and commercial value. Search advertising, which delivers ads according to the user's search content, is the largest and fastest-growing form of online advertising and has been the main income for Internet Industry. Predicting CTR is the most critical technology for search advertising, for it is not only related to ad rank, but also affects ad click charges. It is a meaningful work to predict CTR correctly based on massive historical search data. At present, most of the existing work predict search CTR based on shallow models. In shallow models, the meanings of features are fixed and isolated, without considering the inner relation between the features.

This paper is focused on predicting CTR of search advertising given the specific information. In order to predict the result effectively, a deep neural network is employed to dig information between the different features. The contents could be generalized into three parts:

First, after giving the definition of the CTR of search advertising, data distribution is analysised. According to the knowledge of search advertising and characteristic used in practise, six categories of features are extracted based on the processing data. To address the problem of memory limit and time consuming, this paper employs a data-chunked method, which is based on GPU.

Secondly, considering the weakness of Naïve Bayes, Logistic Regression model and Support Vector Regression Model for predicting CTR, we put forward a deep neural network-based method for predicting the CTR of search advertising, what is more, dropout method is used to prevent train from overfitting. Based on the same feature set, DNN-based method gives a better performance when compared to main method for predicting CTR of search advertising.

Thirdly, a convolution neural network-based method is employed to predict the CTR of search advertising. The operation of convolution and pooling makes it possible for the model to learn the relation between local features and global features. Experiments are conducted on the data set of KDD Cup 2012. The performance shows that CNN-based method make a big improvement for predicting the CTR of search advertising.

Keywords: click-through rate, search advertising, deep learning, deep neural network, convolution neural network

目　录

第 1 章 绪　论

1.1 本文研究的背景和意义

随着网络技术的进步，互联网在近年飞速发展。在线广告作为互联网发展的产物应运而生，是互联网企业最稳定也最有利润的商业模式，在不断的关注和研究的推动下，在线广告展现出了巨大的市场价值。

据互联网广告署（Interactive Advertising Bureau，IAB）[1]统计表明，在 2014 年全年，互联网在线广告收入在美国的总额为 495 亿，与 2013 年相比营收增加了 16%。其中，搜索广告继续领跑互联网在线广告形式，在 2014 年第四季度仅 142 亿的收入中，搜索广告收入总额达到 53 亿，与 2013 年第四季度搜索广告总额 50 亿相比，同比增长 5%。

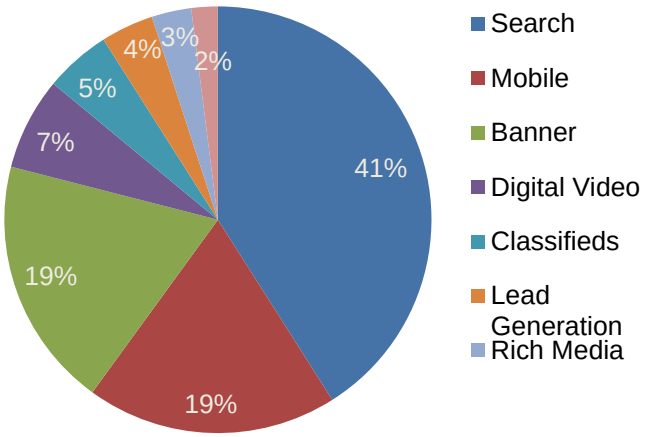

图 1-1 美国互联网广告收入 2014 年第四季度各广告形式占比图

搜索广告目前已经成为互联网行业的主要收入来源之一，也是规模最大，增长最快的广告渠道之一，它能根据用户输入的查询词，了解用户的搜索意图，在搜索的结果页面呈现出相应的广告信息，更好的完成在线广告的个性化投放和优化。对于参与搜索广告的广告商、广告媒介和用户三者来说，一方面，广告商通过支付每次点击费用（CostPerClick，CPC）的形式借助广告媒介投放广告，广告媒介的收益则来自于每次点击费用与广告点击率（Click-Through Rate，CTR）预测共同影响而得到，即 CPC*CTR，广告点击率预测的准确性与广告商

[1] http://www.iab.net/research/industry_data_and_landscape/adrevenuereport

和广告媒介的收益息息相关[1]。另一方面，用户点击广告的概率随着广告位的排放顺序呈递减趋势，对广告点击率进行预测并将预测结果高的广告投放在搜索结果页面靠前的位置，能增加用户的对广告的点击率。

搜索广告点击率预测结果的好坏直接关系到广告商与广告媒介的收益，因此，这项研究早已是工业界的热点项目之一。然而，因为点击率预测问题本身具有数据量庞大、数据非常稀疏、冷启动等困难，尽管搜索广告点击率的预测得到越来越多的研究，但目前的研究进展还比较迟缓，并且大部分研究都是基于已有的浅层学习模型来解决，并没有深入的挖掘数据和特征内在的一些关系，所以还有很大的提升空间。

本文使用深度学习的方法来预测搜索广告的点击率，主要是试图从特征学习的方面挖掘出更多特征之间隐藏的深层次的内在联系，从而提高搜索广告点击率预测的效果。准确高效的预测搜索广告的点击率不仅能增加广告媒介的收益，还能提高用户对搜索结果的满意程度，因此，我们认为本文对基于深度学习的搜索广告点击预测研究的尝试是很有意义的。

1.2 国内外研究现状

搜索广告点击率预测结果的好坏，不但对用户体验造成影响，更关系到广告商和广告媒介的经济收益，一直都是公司重点研究的内容之一，学术界里目前也涌现出越来越多对广告点击率预测的研究。此外，近年来，深度学习已经成为目前机器学习领域最热的方向之一，在多个领域上都取得了令人惊叹的成绩。本节将对现有的搜索广告点击率的研究的现状以及深度学习的研究现状进行介绍。

1.2.1 搜索广告点击率预测的研究现状

广告点击率预测是广告算法中最核心的技术，得到学者们越来越多的广泛关注。在早期的研究中，常常使用历史点击率的统计作为搜索广告点击率的预测结果，按照历史点击率的高低将广告投放在搜索结果页面中相应的位置，然而这种方法仅仅记录对历史数据的统计，很难对用户进行个性化投放，尤其是在新用户新广告问题上表现出了很大的缺陷。目前的研究主要集中在三大方向上：一是基于假设检验的方法，一是基于分类思想的方法，另外一个是基于推荐系统的方法。

1、基于假设检验的方法，这种方法在对搜索广告的点击率进行预测前，会针对问题或者模型做出某种假设，在此基础上进行预测。Chapelle 等人[2]使用动态贝叶斯网络，假设在满足贝叶斯定理[3]的情况下，通过对用户产生的点击过

程建立模型，考虑级联位置的信息模拟出特定位置与相近位置的相关性，以判断该位置上的广告是否满足用户搜索要求。Dupret 等人[4]使用用户浏览模型，假设满足相同查询下的点击都是相互独立的情况下，提出了一种消除位置偏见的点击率预测方法，根据结果列表中文档的相关性和位置关系对搜索结果的点击率进行预测。Guo 等人[5]提出了一种点击率模型，假设用户在查看搜索结果时满足依次查看的顺序，这就意味着当用户点击当前的某个搜索结果时，用户不但对当前的结果满意，并且已经浏览过位于当前结果前的所有搜索结果，在这种模型中，文档的相关性和文档间位置的关系是后续求解模型的关键。

2、基于分类思想的方法，在这种方法中，很多研究中将作为典型的预测问题的搜索广告点击率预测看作分类问题来解决，其中最常见的是应用线性模型来预测点击率。Chakrabarti 等人[6]利用点击反馈的相关性，通过提取网页和广告词等的特征，并在这些特征上使用逻辑斯蒂回归模型，用以提高广告检索和预测的效果。Wu 等人[7]则采用了模型融合的思想，先将点击率的预测使用不同的分类模型来解决，然后将不同线性模型的实验效果相结合，实现模型间的互补，来提高搜索广告点击率预测的结果。当然，在真实的场景中，点击率的预测并非简单的线性问题，因此，一些学者开始使用非线性模型来解决点击率的预测。Dave 等人[8]在搜索广告点击信息以及广告商账户信息上提取语义特征，使用基于投票思想的梯度提升决策树模型，实现对不同的特征进行投票，以方便不同特征在结果求解中的影响力，提高了点击率预测的效果。Zhang 等人[9]利用神经网络模型对影响搜索广告点击率的因素进行的探索，从特征因素方面提高点击率预测的结果，但是资源单一，数据交互的关系没有很好的利用。

3、基于推荐系统的方法，个性化推荐[10]是一种根据用户历史数据分析用户的行为，得到用户感兴趣的商品和信息等，并按得到的结果向用户进行推荐的系统。搜索广告的投放也可作为推荐系统推送的信息，因此，很多学者使用基于推荐系统的方法来解决点击率预测问题。霍晓骏等人[11]采用协同过滤算法，为页面找到与其相似的其他邻居页面，实现点击率的预测，以此作为基础进行广告推荐，但当相似页面的数量增加时，该方法的结果质量会严重下滑。文献[12-15]中使用了不同形式的矩阵分解模型：MF（Matrix Factorization）、SVD（Singular Value Decomposition）、SVD++等模型，通过已有的用户物品评分，来拟合整个用户物品矩阵，从而消除缺失值，达到预测的目的。Kanagal 等人[16]提出了一种聚焦矩阵分解模型，针对用户对具体的产品的喜好以及相关产品的信息进行学习，解决因用户-产品交互活动少而造成的数据稀疏问题。在文献[16]的基础上，Shan 等人[17]提出了一种立方矩阵分解模型，通过对用户、广告和网页三者之间关系的立方矩阵进行分解，利用拟合矩阵的值来预测 CTR，虽然立

方矩阵分解模型增加了一维交互关系，但所刻画的交互关系仍然十分局限，不能在 CTR 预测中充分挖掘广告所有特征之间的联系。

此外，很多研究还从特征工程方面入手来探索搜索广告点击率的预测。文献[18]在假设不同的关键字在本质上被点击的可能性也不用的情况下，认为类似的广告遵循类似的点击率分布，使用关键字聚类来获得特征，增加在历史数据不足时预测的有效性。文献[8]利用查询-广告点击图，来提取与查询内容和广告相关的特征。He 等人[19]认为不是所有的特征组合都是有用的，他们使用决策树对特征进行修建，以提升最后的预测效果。

1.2.2 深度学习的研究现状

深度学习[20, 21]是机器学习领域的一个新的分支，它是由人工神经网络不断发展而形成的，是一个包含多层次的复杂学习结构。深度学习的目的是通过对底层特征的层次学习和组合，得到更为深层意义更加抽象的高层特征。目前，深度学习在语音、图像和自然语言处理方向上已经得到了广泛的应用，并取得了较好的成绩。

1、在语音识别领域，对每个建模单元统计概率进行描述时，长期以来都是混合高斯模型（Gaussian mixture models，GMM）[22-24]占据主导地位，它的优点在于模型相对比较简单，能完成大数据量下模型的训练，也正是因为模型简单，使得特征的特性分布并不能被充分的学习，此外，混合高斯模型在建模时，特征的维度往往很低，在简单的模型下，特征之间的关系根本无法得到充分的挖掘，效果非常有限。

2011 年微软公司在语音识别系统中应用了深度学习方法，这一创举将改变了语音识别系统早前的框架[25]，此后，谷歌和百度等知名公司也相继对深度学习展开研究，应用到自己公司的语音产品中。使用深度神经网络后，语音样本特征原本的时序性关系得到充分的展现，它克服了以前浅层学习方法的缺陷，模拟动物的大脑架构，通过多层次对特征的学习，能更好的进行识别，提高语音识别的效果。

2、在图像识别领域[26, 27]，传统的图像特征提取方法，都是根据专业知识和实验的经验进行提取，提取过程中，主要是从颜色、纹理和形状三方面来设计提取方案，这些方案往往是源于对图像的人为了解，常常是一些简单可行、计算量低的算法，受人为思想影响大，有较强的局限性，这样的特征简单、孤立、推广型非常弱。

LeCun 教授等人[28]受动物视觉模型的启示，提出的卷积神经网络，它具有两个非常独特的优点：稀疏连接和权值共享，是一种全新的特征提取方法，它

通过在局部特征上使用卷积核，提取出了新的模糊图像，打开了图像识别领域上特征提取的新大门。而后，Hinton 教授带领他的学生使用卷积神经网络在著名的国际计算机视觉挑战赛 ImageNet 大赛中取得了第一的优秀成绩，奠定了卷积神经网络在图像识别领域深厚的基础，验证了深度学习网络模型在理解和识别自然图像上是有效的。

3、在自然语言处理领域，基于统计的模型是自然语言处理中最常用的方法，然而在这些方法中，特征的提取依然取决与经验和人为的理解，并且常常需要借助于自然语言处理的工具，如词性标注[29]等，而这些工具本身的错误率会导致整体效果的下降。此外，提取的这些特征并没有经过处理就使用，不但会增加模型计算的强度，还会带入噪音的影响。

2008 年，NEC 实验室首次在自然语言处理方向上提出了基于卷积神经网络与多任务学习的统一架构，可以使用一个语言模型解决大量的语言预测，包括词性标注，分词，命名实体识别，语法和语义标注等。此后，越来越多的学者开始将深度学习应用的自然语言处理的各个方向上，包括序列标注、情感分析等，并都取得了不错的成绩。

1.3 问题的难点和本文的研究思路

尽管搜索广告点击率预测对企业的商业价值至关重要，但是在学术界对点击率预测的研究还并没有非常成熟，大部分研究仍然集中在浅层学习模型以及基于推荐系统的模型上，并没有充分的模拟出现实中广告点击率的真实场景，更没有挖掘出特征之间的相互关系。本文采用 KDD Cup 2012 中 Track 2 提供的数据集进行研究，该数据集由腾讯公司下的搜索品牌搜搜（SOSO）搜索引擎提供。研究问题中存在的难点如下：

1）数据量大。大数据是当前研究的热门，也是广告点击率预测面临的挑战之一。本文所使用的数据集中，训练集大小为 9.87GB，包含 149639105 条样本，测试集大小为 1.26GB，包含 20297594 条样本。庞大的数据量使得完全使用全部数据存在一定的难度。

2）特征稀疏性。搜索广告的现实场景决定了在历史数据中真正点击的概率非常小，这也使得想要提高广告点击率预测的效果是有难度的。

3）浅层学习模型的效果有限。在浅层学习模型中，特征在使用上是独立的，特征之间的关系往往得不到充分的学习，如何挖掘特征之间的关系，也是本文的难点之一。

4）设备限制。广告点击率的研究通常在企业内部开展，它们集群、并行等来处理特征维度大、模型复杂度高的问题，而在学术领域中，配备的实验设备

通常不够完善，很难很好的解决高维度、高复杂度的问题。

本文使用深度学习模型对搜索广告点击率进行预测，目的是为了更好的模拟出非线性的广告场景，学习出特征之间的相互关系，从而更准确的预测搜索广告的点击率。本文的研究内容主要包括：

第一、对搜索广告背景下大数据的处理和特征的提取。在搜索广告中，针对用户的查询词，媒介往往会推送出大量的相关广告，然而用户点击的数量却寥寥无几。为此，我们在实验前对数据进行预处理，在数据分布不变的情况下缩小样本空间。在此基础上，我们提取了多类不同的特征，并提出了一种利用 L1 范式进行特征降维的方法，使得高维特征也能应用到复杂的深度学习模型中。

第二、研究了加入词向量特征的浅层学习模型的搜索广告点击率。浅层学习模型，如朴素贝叶斯、逻辑斯蒂回归模型等是用于广告点击率预测的经典模型，针对搜索广告的特性，我们使用词向量特征融入搜索广告的上下文信息。

第三、研究了基于深度神经网络和卷积神经网络的广告点击率预测。我们首先使用最基本的深度学习模型——深度神经网络，对网络参数如层数、节点数的设置进行了研究，并研究了使用 dropout 方法的深度神经网络。再者，我们研究了基于卷积神经网络的广告点击率预测，对卷积神经网络中滑动窗口的大小、卷积和亚采样的层数、各层节点数的设置进行了研究。

1.4 本文内容安排

本文共包括四章，每章内容安排如下：

第一章 本章首先介绍了搜索广告点击率预测研究的背景与意义，接着分别对广告点击率预测和深度学习的国内外研究现状进行了详细介绍，然后我们给出了本论文的研究难点，并说明了本论文的主要研究内容，在本章的最后一节，我们介绍了本论文的内容安排。

第二章 本章主要介绍了问题定义和特征表示。首先对搜索广告点击率预测问题进行了形式化的定义，考虑到搜索广告中数据集和广告日志的复杂性，我们对本论文实验中所使用的数据集进行了介绍，并列出了相关统计信息，然后对实验数据进行预处理，此后我们对本论文实验中所使用到的各类特征提取进行了详细的说明，最后给出了搜索广告点击率预测结果的评价指标。

第三章 本章主要介绍了基于深度神经网络的广告点击率预测方法。首先，我们使用朴素贝叶斯模型、逻辑斯蒂回归模型、支持向量回归模型等浅层学习模型对搜索广告的点击率进行预测，进而提出了使用深度神经网络进行广告点击率的预测。然后，我们对广告点击率中深度神经网络的应用进行了说明。最

后，我们对实验结果进行了分析。

第四章 本章主要介绍了基于卷积神经网络的广告点击率预测方法。我们先对卷积神经网络的结构和特点进行了说明，然后将卷积神经网络应用到搜索广告的点击率预测问题上，最后通过实验分析了卷积神经网络的预测效果。

第 2 章 问题定义和特征提取

2.1 引言

搜索广告的点击率预测问题是一个具有商业意义的复杂问题，预测的结果具有一定实际意义。但是广告点击的历史日志往往是数量庞大但非常稀疏的数据，本章对数据进行了预处理，而后对本论文中所使用到的特征的提取方法进行了描述，为克服实验硬件的限制困难，我们采用了分块计算来实现模型的训练和测试，本章的最后给出了广告点击率预测问题的评价指标。

2.2 问题定义

搜索广告指的是，广告商根据自己产品所具有的特性和内容等，为产品确定关键词、标题、产品描述等相关属性，并对投放的广告关键字等进行自主定价。当用户搜索的内容与广告商确定的广告属性相同，或者与广告的关键字相关时，广告媒介就会在搜索结果页面中对这些广告进行展示。此时，若用户点击了投放的广告，则广告所属的广告商需要向广告媒介缴纳相应的出价，若用户未点击所投放的广告，广告商无需向广告媒介缴费。若用户搜索的某关键字有多个广告商想要购买时，将根据竞价排名的原则对广告进行展示。

搜索广告背后的一项关键技术便是广告点击率的预测，广告点击率的定义是点击量与展示量之比，即：

$$CTR = \frac{num\_Click}{num\_Impression} \tag{2-1}$$

这里，num_Click 是点击的次数，$num\_Impression$ 是展示的次数。

在当前在线广告的经营模式中，广告点击率的预测不仅影响着广告的排名，也是广告商对广告定价的依据之一。

例如，当用户输入关键词“雨伞”时，通过竞价购买了“雨伞”关键词的广告商的相关广告会在用户搜索结果页面中得到展示，若用户点击了结果页面中投放的“雨伞”的广告，则广告商需按公式（2-2）付费给广告媒介：

$$Value = CPC * CTR \tag{2-2}$$

当然，若用户没有对结果页面中的“雨伞”广告进行点击，则无需付费。

图 2-1 搜索广告示意

整个过程中，广告点击率是主要效果的评价指标，广告商通过提高 CTR 降低 CPC 来使自己利益最大化，广告媒介通过提高 CTR 来提高公式(2-2)中的 Value 值，并且提升用户在该媒介上的用户体验。

搜索广告点击率预测的目标是通过给定的信息预测搜索网页的广告点击率。在预测搜索广告的点击率时，主要步骤可以描述为以下几步。首先确定预测的方案，包括特征的提取方案、点击率的预测算法以及后续进行的实验环境等。其次，搜集相关的广告历史日志数据集，根据特征提取的方案，结合得到的数据集，提取具体的特征。再次，使用预测算法对广告点击率进行预测，得到预测的结果值。最后，对预测的结果进行评价，并针对具体的结果分析做出相应修正，以提高后续预测的效果。

2.3 数据预处理

本文所采用的实验数据集为 KDD Cup 2012 中 Track 2 提供的数据集。该数据由腾讯公司下的搜索品牌搜搜（SOSO）搜索引擎提供，因为涉及公司商业信息，数据经过哈希处理。

完整的数据集包含训练集和测试集，此外，腾讯还提供了五个附加文件，用来对训练集和测试集以外的信息进行扩充，分别为：queryid_tokensid.txt（查询词信息），purchasedkeywordid_tokensid.txt（购买关键字信息），titleid_tokensid.txt（标题信息），descriptionid_tokensid.txt（描述信息），userid_profile.txt（用户信息）。完整数据集的规模如表 2-1 所示。

表 2-1 附加文件规模统计

文件名	实例数	大小
training.txt	149639105	9.87GB
test.txt	20297594	1.26GB
queryid_tokensid.txt	26243606	703MB
purchasedkeywordid_tokensid.txt	1249785	25.2MB
titleid_tokensid.txt	4051441	170MB
descriptionid_tokensid.txt	3171830	267MB
userid_profile.txt	23669283	280MB

在训练集中，每条样本包含 12 个属性，分别是对广告属性、用户属性、网页属性的刻画，各属性详解如表 2-2 所示。测试集与训练集的格式完全相同，除了测试集的点击次数和展示次数都需要计算，用以广告点击率的预测。前四个文件中的每行为一个 id 映射到 token 列表，分别对应于该查询，关键字，广告标题和广告描述。在每一行，id 和 token 列表由一个制表符 TAB 隔开。token 列表是自然语言中的文字。因为信息涉密，这里每个 token 是由它的散列值来表示，各 token 之间由“|”进行分割。“userid_profile.txt'文件的每一行分别是由用户 ID，性别和年龄三个字段组成，三者之间由制表符分隔。值得注意的是，并不是出现在训练集和测试集中的用户 ID 就一定存在于'userid_profile.txt'。

表 2-2 实验数据介绍

属性名	属性描述	属性类别
Click	点击次数	广告-网页属性
Impression	展示次数	广告-网页属性
DisplayURL	网址	广告属性
AdID	广告	广告属性
AdvertiserID	广告商	广告属性
Depth	广告展示个数	网页属性
Position	广告展示位置	广告-网页属性
QueryID	用户查询	用户属性
KeywordID	广告关键字	广告属性
TitleID	广告标题	广告属性
DescriptionID	广告描述	广告属性
UserID	用户	用户属性

实验目标是通过给定的信息预测搜索网页的广告点击率，由于数据量过大并且正负样本不平衡，实验中我们采用随机采样的方法，从训练集中抽取 10%作为本文实验中模型训练的训练集，即使用随机函数生成对应的样本序号，抽取出对应的样本，这样既缩小了样本空间，同时随机采样也保持了原始数据的分布信息，如图 2-2 所示。

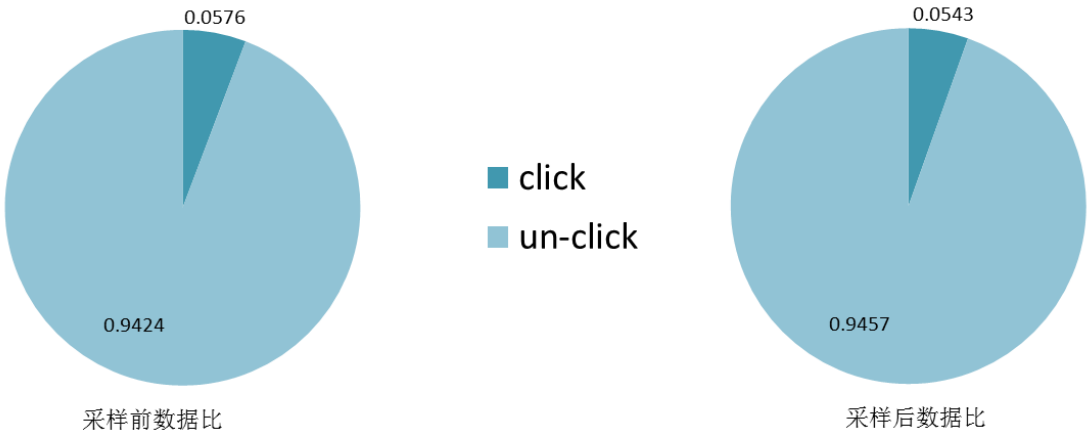

图 2-2 训练集抽样前后比例图

验证集在模型参数的调节起着至关重要的影响，实验中，我们同样采用随机采样抽取部分样本作为验证集用于参数的调节。测试集为 KDD Cup 2012 中 track 2 的全部测试数据。数据的统计信息如表 2-3 所示。

表 2-3 实验数据统计信息

数据集	样本数	点击数	展示数
KDD Cup 2012 track2 训练集	149 639 105	8 217 633	235 582 879
测试集	20 297 594	418 403	13 303 612
本文使用的训练集	15 000 000	876 389	23 652 694
验证集	500 000	25 649	609 694

2.4 特征提取

特征是作为模型的输入，它的好坏对最后的实验结果影响举足轻重，本节对特征的提取进行详细的描述。

2.4.1 类别稀疏特征

在搜索广告中，广告属性、用户属性、网页属性三者共同影响着广告的点

击率。由上一节可知，数据集提供的大部分都是 ID 类型的属性，我们采用 One-hot Representation 的思想，将这些 ID 属性都扩展成二值特征。例如，在我们使用的训练集中，广告属性中的 AdID 共有 374637 个不同的值，我们将这一属性特征扩展成一个 374638 维的向量，其中，前 374637 维对应于在训练集中出现过的 AdID，当某个 ID 出现时，就将相应位置上的值置为 1，其余都为 0。第 374638 维用来刻画在测试集中出现但未在训练集中出现的 AdID。

我们分别对 AdID，AdvertiserID，QueryID，KeywordID, TitleID, DescriptionID, UserID，DisplayURL 这 8 个属性进行了特征扩展，最终得到的特征维度超过千万，虽然维度非常高，但非零值只有 8 个，也因此，我们称之为类别稀疏特征（f_Sparse feature）。

2.4.2 历史点击率特征

在预测搜索广告的点击率时，历史点击率看似简单，但是却是十分有效的特征，因为历史点击率在一定程度上代表了类别 ID 对某个广告感兴趣程度的高低，当一个 ID 对某个广告的历史点击率高时，意味着它对这个广告更感兴趣，后续点击的概率也更大。因此，我们针对不同的类别计算它们的历史点击率作为特征。

历史点击率($pseudo-CTR$)是利用通用的点击率计算公式，通过历史数据，即训练数据计算得到的。点击率是点击数($\#click$)与展示数($\#impression$)之比，如公式(2-2)所示。

$$CTR=\frac{\#click}{\#impression} \tag{2-3}$$

在实际的计算中，很多类别在多数情况下都只有展示次数，却没有点击的次数，导致历史点击率的计算值大多为零。因此我们引入平滑方法，对公式(2-3)进行调整如下：

$$pseudo-CTR=\frac{\#click+\alpha\times\beta}{\#impression+\beta} \tag{2-4}$$

公式(2-4)中的 α 和 β 是平滑方法中的调节参数，根据(2-4)，分别计算出 AdID，AdvertiserID，QueryID，KeywordID，TitleID，DescriptiomID，UserID 的历史点击率（p_CTR feature）。

2.4.3 相似度特征

相似度特征（Ssimilar feature）用来刻画属性两两之间的相似程度，用户搜索的内容与被投放的广告属性相似度高时，广告被点击的概率更大。例如当搜

索内容 Query 与广告关键字属性 Keyword 相似度高时，意味着网页投放的广告与用户期望搜索的广告结果相似度高，更符合用户点击广告的动作。

数据集中提供了关于 Query、Keyword、Title、Description 的属性描述文件 queryid_tokensid.txt 、 purchasedkeywordid_tokensid.txt 、 titleid_tokensid.txt 、 descriptionid_tokensid.txt。尽管这些属性信息都以经过哈希后的数字形式给出，但是实质是自然语言处理中的合法的字词句，它们之间的相对含义并没有发生改变。

我们分别构造 Query、Keyword、Title、Description 中 token 的 TF-IDF 向量，计算 TF-IDF 值的公式如下：

$$tfidf_{i,j} = tf_{i,j} \times idf_i \tag{2-5}$$

其中，TF、IDF 的计算分别如下：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2-6}$$

$$idf_i = \log \frac{|D|}{|j : t_i \in d_j|} \tag{2-7}$$

这里，$n_{i,j}$ 和 $n_{k,j}$ 表示文档 j 中 token i 和 token k 出现的次数，$|D|$ 表示文档的总数，$|j : t_i \in d_j|$ 表示包含 token i 的文档总数。

得到了 TF-IDF 向量后，根据余弦相似度公式计算出它们之间的余弦相似度作为特征：

$$sim = \frac{\vec{v_1} \bullet \vec{v_2}}{|\vec{v_1}| \bullet |\vec{v_2}|} \tag{2-8}$$

其中，$\vec{v_1}$ 和 $\vec{v_2}$ 分别是需要计算相似度的两个 token 的 TF-IDF 向量。

2.4.4 位置信息特征

在信息检索中，研究表明[5]，检索结果页面中越靠前的信息越容易被点击查看。搜索广告可以看做是信息检索的一个应用，用户搜索时需求的多样化要求在对广告进行排序和投放时，在结果页面靠前的位置中尽可能地投放满足用户需求的广告，从而最大化用户的满意度、提高用户点击的欲望。

位置特征（Pos feature）描述的是指定广告在搜索结果页面中的位置信息。在实验数据集中，数据集提供有网页属性 Depth 和广告-网页属性 Position，网页属性 Depth 刻画的是搜索结果页面中广告结果展示的总数，广告-网页属性 Position 刻画的是在搜索结果页面中该广告所出现的位置，我们将两个的原始数值都加入位置特征。

此外，我们发现，单纯的使用 Depth 和 Position 属性的原始数据很难描述出广告在搜索结果页面中排序的靠前程度，因此，我们提取了相对位置 Pos，其定义如下：

$$Pos = \frac{total\_ads - ad\_position}{total\_ads} \tag{2-9}$$

这里，$total_ads$指页面投放的广告总数，即 Depth 属性值，$ad_position$指当前所预测广告的位置，即 Position 属性值。

2.4.5 高影响力特征

类别稀疏特征中，ID 属性信息通常采用 one-hot 形式的特征编码方式，在将不同的属性经过 one-hot 编码后的特征向量组合在一起，这样方式简单直观，并且在以往广告点击率的预测中表现出了不错的效果，但这样的编码方式却使得特征的维度巨大并且非常稀疏。

在深度学习模型中，由于模型的层次较多，当特征维度巨大时，将造成模型需要学习参数规模非常大，模型训练的时间迅速增加，在受限的硬件资源下，甚至会导致无法训练模型。我们发现，在这庞大且稀疏的特征中，绝大部分维度上的特征值对整个模型的预测结果贡献非常小甚至为零，只有少数维度上的特征值对预测结果有较高的影响力。

因此，我们使用可行的降维方法，对特征进行降维，提取出在模型预测中影响力较高的特征，忽略其他维度的特征。本文采用稀疏规则算子（Lasso regularization）[30, 31]，也叫 L1 范数正则化的方式，在逻辑斯蒂回归模型的代价函数中加入 L1 范数，使得模型学习得到的权值结果满足稀疏解，其中权重为 0 的值可理解为对预测结果没有影响的信息，这也是 L1 范数正则化实现特征的自动选择的应用。在本文的实验中，我们将 AdID，AdvertiserID，QueryID，KeywordID，UserID 五个属性先按照 one-hot 编码的形式组合在一起，通过使用 L1 范数正则化的逻辑斯蒂回归模型得到稀疏解，从而提取高影响力特征。

在本文的实验中，尽管使用了 L1 范数正则化得到稀疏解，但相对我们实验的硬件设备来说，想要在深度学习模型中使用这些特征，稀疏解中的非零值仍然很多。最终，我们把非零的学习参数中按大小顺序取出前 N 维权重较大的，将这 N 维权重对应位置上的特征值构建新的特征，称为高影响力特征（High_impact feature）。

2.4.6 词向量特征

词向量在近年中的自然语言领域得到了非常广泛的应用，它是将自然语言

理解问题转换成机器学习问题的纽带。自然语言中最常用的特征表示方法是 One-hot Representation 方法，在这种表示方法中，每个单词都有一个向量来代替，每个词对应着词表中的一维，但某个单词出现时，相应的维度被置 1，其它的维度都为 0，整个向量的长度是词表的长度。这样特征表示方法存在两种明显的缺陷，一个缺陷是任意两个词之间没有相互作用，都是相互孤立的；另一个缺陷是向量维度太高，不利于复杂模型使用向量，尤其是在深度学习模型中。

Word2Vec 是一款能用实数值向量表示自然语言中词的开源工具，在 2013 年时，由谷歌公司开放使用。Word2Vec 模型中，它使用了语言模型：层次化 Log-Bilinear [32]，在 Word2Vec 中包含 CBOW 模型（Continuous Bag-of-Words Model）和 Skip-gram 模型（Continuous Skip-gram Model）两种。如图 2-3 所示。

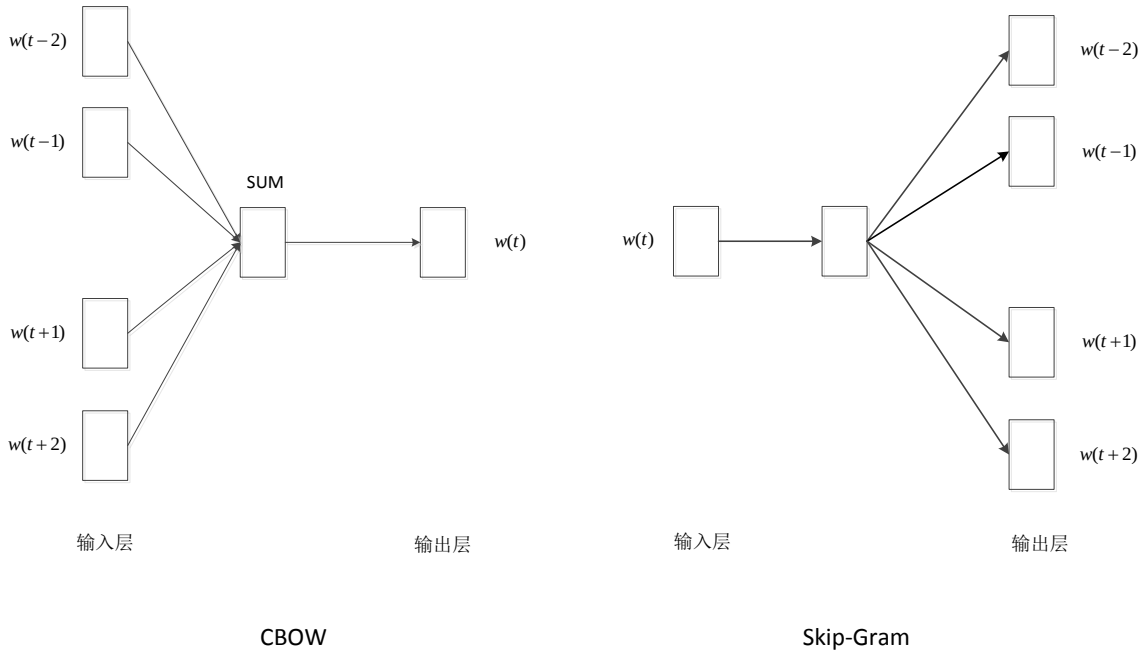

图 2-3 CBOW 模型和 Skip-gram 模型示意

无论是 CBOW 模型，还是 Skip-gram 模型，它们都是基于人工神经网络的进行了改进，去掉了耗时最长的非线性隐藏层，并且实现了所用词共享隐藏层。两者的区别在于优化目标不同，在 CBOW 模型中，给定单词出现的上下文，它的优化目标是使这个单词在给定的上下文中出现的概率最大，如公式(2-9)所示。

$$J = \sum_{w \in C} \log p(w \mid context) \tag{2-10}$$

而 Skip-Gram 模型的优化目标是给定单词，最大化其上下文出现的概率，如公

式(2-10)所示。

$$J = \sum_{w \in C} \log p(context \mid w) \tag{2-11}$$

在 CBOW 中，映射层将上下文向量相加，而在 Skip-Gram 中，模型将输入单词进行恒等映射。

在搜索广告中，广告属性 title 和 description 的描述文件分别是符合自然语言逻辑的文本，在 titleid_tokensid.txt 和 descriptionid_tokensid.txt 文件中，每一个 ID 都对应着一串描述语言（经过哈希处理的），每个单词用符号“|”隔开。我们将文件还原成空格分割的自然语言文本，为广告属性 title 和 description 描述文件中的词训练词向量(W2V feature)。

考虑到广告属性 title 和 description 的描述文件对每个 ID 描述的句子的长短并不一致，而机器学习中模型的输入通常是固定大小的，因此，我们对描述 ID 的句子中的每个词的词向量进行相应维度相加再取平均的处理，使得词向量能更为方面的放入模型中进行训练。

2.5 计算实现

图形处理器（Graphics Processing Unit，GPU）已成为当今主流的计算系统的一个重要组成部分。在过去的几年中，GPU 的性能显著增加。如今的 GPU 不仅是一个强大的图形引擎，而且是一个高度并行可编程的处理器，在复杂的数学和几何计算上已经远远超过了 CPU[33]。与 CPU 相比，GPU 在浮点运算和几何处理上能力更优，它还拥有高宽带的显存，早期主要用于图像处理和视频处理，因此强大的计算性能，并能大大降低系统成本，如今在并行计算和重复计算上得到了很好的应用[34]。

在统计机器学习领域，模型需要大量的数据以保证模型的鲁棒性，这无疑增加了计算的复杂性，尤其是近年来随着深度学习的发展，密集型的计算使得深度学习无法很容易的拓展到大数据的领域，如本文研究的内容计算广告学，目前针对大数据在存储上有分布式的 Hadoop 计算平台，而对于密集型计算，CUDA(Compute Unified Device Architecture)计算平台应运而生。

CUDA 是一种并行计算的平台，同时也是由英伟达（NVIDIA）提供的应用程序接口的简称。它允许编程者在支持 CUDA 的图形处理器上编写通用程序。CUDA 平台提供一层访问 GPU 虚拟指令单元和并行计算的单元的方法。与其他并行计算工具 Direct3D 和 OpenGL 不同，编程需要高级的图形编程技巧，CUDA 可以直接用编程语言 C、C++和 Fortran 来编写，这使得并行计算的研究者很容

易的使用 GPU 资源[2]。

如图 2-4 所示，CUDA 的处理流程可分为四步，描述如下：

1. 首先将程序运行中所需的数据从主存（Main Memory）拷贝到显卡内存（Memory for GPU）中；

2. 由中央处理器（CPU）下达向 GPU 下达指令执行程序；

3. 收到 CPU 的指令后，图形处理器（GPU）在计算单元中进行多单元并行计算，并将计算的结果存入显卡内存中；

4. 最后，将计算的结果由 GPU 拷贝到内存中。

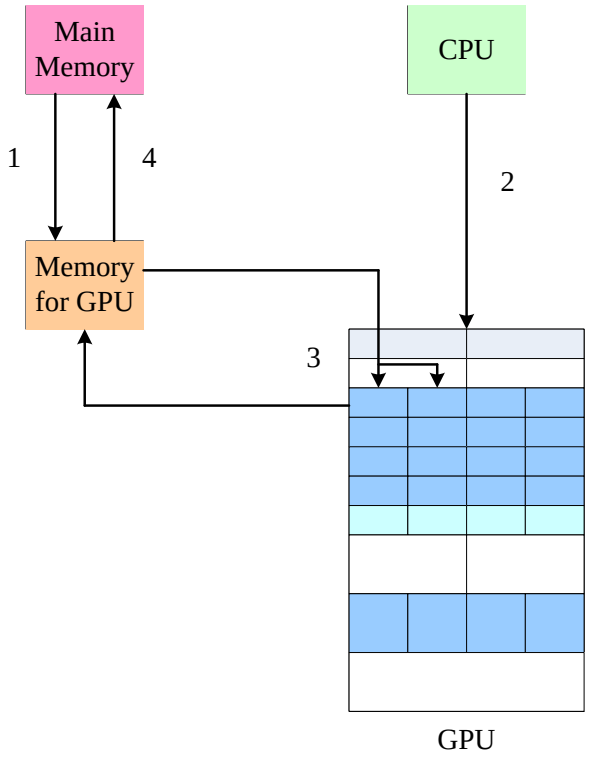

图 2-4 CUDA 处理流程示意

CUDA 相比其他 GPU 上计算工具有如下的优点：

- 可以在内存中随机读取数据；
- 统计的虚拟内存编址
- 统一的内存编址
- 多个线程之间可以共享内存
- 内存和 GPU 数据交换迅速；
- 支持整数和位操作；

[2] http://en.wikipedia.org/wiki/CUDA

在深度学习中，模型层次高，学习参数多，使得训练深度学习模型成为费时费力的工作，这也正是早起神经网络模型发展受到限制的原因之一。为了提高实验速度，我们在 CUDA 平台下，利用 GPU 来完成实验。本论文的实验所使用的操作系统为 Ubuntu 12.04 LTS OS，模型的训练在 4G RAM 的 NVIDIA GeForce GT 610 GPU 条件下完成。

GPU 上多计算单元集成的优点，使得采用 GPU 训练模型大大提高实验计算的速度，降低了实验的耗时量。然后，在解决搜索广告点点击率预测的问题时，数据量非常庞大，除了训练的时长问题，我们还面临着显存太小而无法在训练模型时将数据一次性加载到显存中的困难。为此，在本论文的实验中，我们使用分块训练的思路，将输入特征分成适当大小的块，在训练模型时，每次将一块读入显存进行计算，依次处理所有的块。具体的分块，将在各模型的实验设置中进行详细的说明。

2.6 问题的评价指标

使用 AUC(Area Under Curve)[35]作为点击率预测的评价标准。AUC 的计算过程来源于医学领域中的 ROC（Receiver Operating Characteristics）曲线，在机器学习、模式识别、数据挖掘领域中得到越来越多的使用。

与准确率等传统的评价指标相比，准确率在评价中存在一下几点不足[36]：1）在准确率的计算中，不同类别在分类中的错误代价是相同的，而在实际中，不同类别的代价往往是不一样的，尤其是在类别重要性不同时表现的更加明显。2）数据分布不均时，准确率评价不能充分反映分类的效果，对样本较少（常常是正样本）的一类关注的更多。3）当类别数增加时，准确率的值会降低。而在评价中，AUC 无需确定不同类别的分类代价，并且 AUC 的计算独立于类别的概率分布，最重要的是，当实验中样本在各类别上分布不均时，AUC 能更好的反映出分类器的效果，这使得 AUC 更适合搜索广告点击率预测中的结果的评价。例如，测试集有 1000 个样本，分别属于 A、B 两类，其中 A 类别 900 个，B 类别 100 个，分类器 1 在分类时把 1000 个样本分到 A 类别，0 个分到 B 类别，分类器 2 把 750 个样本成功的分到了 A 类别中，把 50 个样本成功分到了 B 类别中，这时，通过准确率的定义我们知道，分类器 1 的准确率是 90%，分类器 2 的准确率为 80%，然而事实上，分类器 2 的分类结果反而比分类器 1 的效果更好。AUC 在计算上则能更好的反应两个分类器的效果。

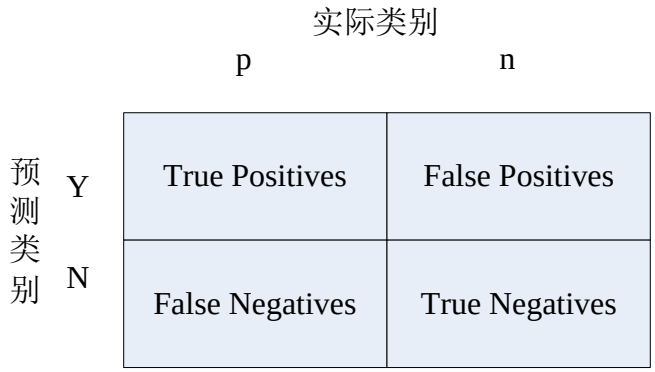

图 2-5 TPR 与 FPR 混淆矩阵图

AUC值计算中两个重要的指标分别为TPR（True Positive Rate）和FPR（False Positive Rate），它们的定义如图 2-5 所示。AUC 值等于以 TPR 为纵坐标、以 FPR 为横坐标时，实验结果所画的曲线下的面积值，其中，TPR 与 FPR 的计算定义分别如下：

$$TPR = \frac{TP}{TP + FN} \tag{2-12}$$

$$FPR = \frac{FP}{FP + TN} \tag{2-13}$$

这里，*TP*、*TN* 分别表示结果中预测对的正样本数和负样本数，*FP*、*FN* 分别表示结果中预测错的正样本数和负样本数。对于广告点击率预测问题，较大的 AUC 值代表了较好的性能。

2.7 本章小结

本章主要对搜索广告点击率预测的问题给出了形式化的定义，对特征的提取进行了详细的描述。

首先，搜索广告点击率预测问题是一个在实际应用中相对比较复杂的问题，为了能在本文后续模型的应用中能更加方便描述，我们对搜索广告点击率预测进行了形式化定义。

考虑到搜索广告中数据集和广告日志的复杂性，我们在介绍特征的提取前，先介绍了本论文中所使用到的数据集，以及对数据集进行的预处理。

然后，我们对本文中所有使用到的 6 类特征的提取方法进行了详细的描述，主要包括 f_Sparse feature（类别稀疏特征）、p_CTR（历史点击率特征）、Similar feature（相似度特征）、Pos feature（位置信息特征）、High-impact feature（高影响力特征）和 W2V feature（词向量特征）。

而后，针对本论文实验中环境的限制，我们简单的介绍了本论文实验中分

块计算的思想。

最后，我们给出了搜索广告点击率预测问题的评价指标——AUC，并分析了在对广告点击率预测进行结果评价是，AUC 比其它评价指标更适用。

第 3 章 基于深度神经网络模型的广告点击率预测

3.1 引言

在解决搜索广告点击率的预测问题中，有的研究基于统计学习模型[37, 38]，有的研究基于位置级联[5]，有的研究基于推荐系统思想[10, 11]。其中，基于统计学习模型的应用最为广泛。

基于贝叶斯定理下的朴素贝叶斯模型[39]尽管简单，但因其在分类问题上能取得不错的效果而经常被选作基线实验模型。逻辑斯蒂回归模型[6, 40]是模型工业界在解决广告点击率预估问题上最为经典的模型。支持向量回归模型[41-43]实质是利用支持向量的思想来解决回归问题，在预测问题中常能取得较好的效果。本论文分别介绍了这三种浅层学习模型，并以它们作为对比，与后续的深度学习模型实验结果进行对比分析。

深度神经网络是深度学习中最为简单的中模型，它由多层人工神经网络的堆叠而成，与浅层学习模型相比，在表达复杂函数时，深度神经网络使用的参数更少更简洁，更好的完成对特征的学习。但这样的结构也存在容易过拟合，参数比较难调，训练耗时长等问题，在面向大数据背景的搜索广告点击率的预测问题中，本论文采用 dropout 方法[44-46]来解决过拟合问题，并利用多计算单元集成的 GPU 来提高模型训练的速度。

3.2 基于浅层学习模型的广告点击率预测

3.2.1 基于朴素贝叶斯模型的广告点击率预测

朴素贝叶斯[39]是以经典贝叶斯概率理论为基础的统计学分类方法，尽管模型相对简单，但在分类问题上却能取得不错的分类效果，因此常常作为基线模型被广泛使用。

给定有 d 个样本的数据集 $D=\{x_1,x_2,...,x_d\}$，假设共有 m 个类别 $C=\{C_1,C_2,...,C_m\}$，每个样本都有 n 个属性 $A_1,A_2,A_3,...,A_n$，则有贝叶斯分类器：

$$c(x)=\arg\max_{C_i\subset C} P(C_i)P(x\mid C_i) \tag{3-1}$$

其中 $c(x)$ 表示通过计算分类所得到的标签，即选择当前样本的属性在给定条件下的后验概率最大的类别作为预测的分类类别。

但是公式(3-1)中的后验概率在实际计算中是很难计算的，于是引入“朴素

贝叶斯假设”：在给定的类别C下，样本的所有属性A_i, $i=1,2,\cdots,n.$都是相互独立不相关的。有：

$$P(A_i \mid C, A_j) = P(A_i \mid C), \forall A_i, A_j, P(C) > 0 \tag{3-2}$$

根据公式(3-1)和(3-2)，当属性之间相互独立时，联合概率等于各条件概率相乘，应用贝叶斯公式，当属性给定时，我们可以得到以下计算所属类别的后验概率：

$$P(C = c \mid A_1 = a_1 \cdots A_n = a_n)P(A) = P(C = c)\prod_{i=1}^{n} P(A_i \mid C = c) \tag{3-3}$$

朴素贝叶斯模型将同时满足(3-3)和(3-3)的样本归为c_i类：

$$p(c_i \mid a_1 a_2 \cdots a_n) = \frac{\prod_{i=1}^{n} P(a_i \mid c_i)}{p(a_1 a_2 \cdots a_n)} p(c_i) \tag{3-4}$$

$$p(c_i \mid a_1 a_2 \cdots a_n) > p(c_j \mid a_1 a_2 \cdots a_n),\ i \neq j \tag{3-5}$$

在搜索广告点击率的预测问题中，我们需要预测给定的样本点击的概率大小，而不是纯粹的将样本分为点击和不点击两类，因此在搜索广告点击率预测问题中，计算出概率值：

$$y = p(c_i \mid a_1 a_2 \cdots a_n) \tag{3-6}$$

以公式(3-6)的结果作为预测值。

3.2.2 基于逻辑斯蒂回归模型的广告点击率预测

逻辑斯蒂回归（Logistic Regression，LR）[6, 40]模型是典型的广义线性模型，用于对事件概率的预测，是广告点击率预测中最为流行的模型。

逻辑斯蒂回归模型可以用来解决二分类问题，也可以用来解决多分类问题。以最为常见的二项逻辑斯蒂回归模型为例。逻辑斯蒂回归模型概括来说，是在线性模型的基础上，加入了连接函数 logit 函数。对于给定的输入$x_i \in R^n$，$x = (x_1, x_2, x_3, ..., x_n)^T$，$Y \in \{0,1\}$为输出，我们可以得到二项逻辑斯蒂回归模型的条件概率分布如下：

$$p(Y = 1 \mid x) = \frac{\exp(w \cdot x + b)}{1 + \exp(w \cdot x + b)} \tag{3-7}$$

$$p(Y = 0 \mid x) = \frac{1}{1 + \exp(w \cdot x + b)} \tag{3-8}$$

这里，$w \in R^n$，$w = (w_1, w_2, w_3, ..., w_n)^T$是模型的回归系数，$b$是模型的偏置。

逻辑斯蒂回归模型常使用极大似然估计法对模型的参数进行估计。对于给

定输入的训练集

$$D=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$$

其中 $x_i \in R^n, y_i \in \{0,1\}$，记 $p(Y=1|x)=\psi(x)$ ， $p(Y=0|x)=1-\psi(x)$，似然函数为：

$$\prod_{i=1}^{n}[\psi(x_i)]^{y_i}[1-\psi(x_i)]^{1-y_i}$$

则有对数似然函数：

$$\begin{aligned} L(w) &= \log\{\prod_{i=1}^{n}[\psi(x_i)]^{y_i}[1-\psi(x_i)]^{1-y_i}\} \\ &= \sum_{i=1}^{N}[y_i \log \pi(x)+(1-y_i)\log(1-\pi(x_i))] \end{aligned} \tag{3-9}$$

对上面的对数似然函数 $L(w)$ 求极大值，解出 w 的估计值 $\hat{w}$，通常采用梯度下降法对模型的参数进行学习，最后学到的逻辑斯蒂回归模型为：

$$p(Y=1|x)=\frac{\exp(\hat{w}\cdot x+b)}{1+\exp(\hat{w}\cdot x+b)} \tag{3-10}$$

$$p(Y=0|x)=\frac{1}{1+\exp(\hat{w}\cdot x+b)} \tag{3-11}$$

在广告点击率预测问题中，我们使用逻辑斯蒂回归模型来预测给定的样本点击率的大小。

3.2.3 基于支持向量回归模型的广告点击率预测

支持向量机（Support Vector Machine，SVM）[41-43]是有监督学习算法中较为经典的算法之一，是由 Corinna Cortes 和 Vapnik 于 1995 年提出的基于统计学习理论的一种机器学习方法。支持向量机模型的学习目标，就是在所给定的特征空间中，有效的寻找到一个最优的超平面，用来作为样本间分类决策面，使得正负样本间隔最大化，从而更好的进行样本分类，如图 3-1 所示。

支持向量回归模型（Support Vector Regression，SVR）是支持向量机的一个扩展形式，主要是在升维以后的高维空间中构造线性决策函数来实现线性回归，输出实数值，而支持向量机则是根据这些实数值来进行分类。

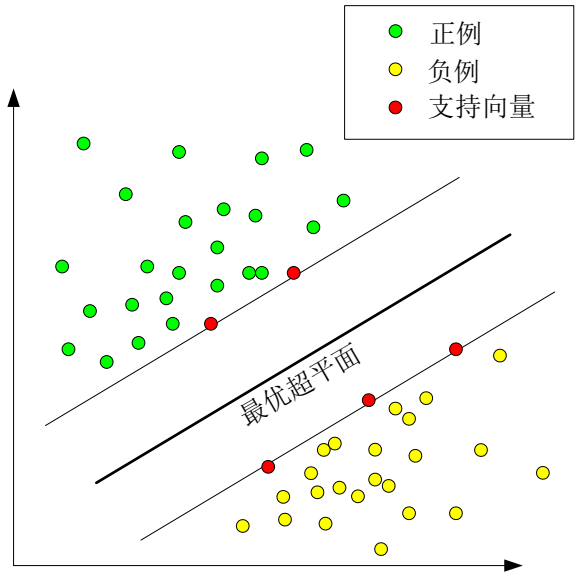

图 3-1 SVM 模型示意

假设给定训练数据集：

$$D=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$$

其中 $x_i\in\chi$， $y_i\in\{-1,1\}$， $i=1,2,\cdots,n$，则模型可表示成如下的凸二次规划问题：

$$\begin{aligned} &min \quad \frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i \\ &s.t. \quad y_i(w\cdot x_i+b)\geq 1-\xi_i \\ &\qquad \xi_i\geq 0,\ i=1,2,\cdots,n. \end{aligned} \tag{3-12}$$

为求得最优解 w^* 和 b^*，应用拉格朗日对偶性，引入拉格朗日乘子 α，将上式转化成相应的对偶问题：

$$\begin{aligned} &min \quad L(w,b,\alpha)=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(x_i\bullet x_j)-\sum_{i=1}^{n}\alpha_i \\ &s.t. \quad \sum_{i=1}^{n}\alpha_i y_i=0 \\ &\qquad 0\leq\alpha_i\leq C,\ i=1,2,\cdots,n. \end{aligned} \tag{3-13}$$

其中，$(x_i\bullet x_j)$ 表示 x_i 和 x_j 作内积。这时，原问题的最优解 w^* 和 b^* 的求解转换为求对偶问题中 α^* 的最优解，可以得到式(3-13)和式(3-13)：

$$w^*=\sum_{i=1}^{n}\alpha_i^* y_i x_i \tag{3-14}$$

$$b^*=y_j-\sum_{i=1}^{n}y_i\alpha_i^*(x_i\cdot x_j) \tag{3-15}$$

由此可得到最优超平面：

$$f(x)=\text{sign}(\sum_{i=1}^{n}\alpha^{*}y_i(x\cdot x_i)+b^{*}) \tag{3-16}$$

SVR 用于解决回归问题，输出为实数值：

$$y=\sum_{i=1}^{n}\alpha^{*}y_i(x\cdot x_i)+b^{*} \tag{3-17}$$

解决非线性问题时，常常需要引入核函数技术，可以将非线性的训练数据由低维空间到高维空间进行映射，解决在低维空间中的不可分情况。

在搜索广告点击率预测中，我们使用 SVR 来对广告点击与不点击两类情况进行拟合，将拟合的结果作为预测值进行输出。

3.3 深度神经网络模型

3.3.1 人工神经元模型

动物的大脑可以看作是一个高度复杂的、非线性的并行计算系统，它能在短时间内对外界进行学习、识别和认知。人工神经网络（Artificial Neural Networks，ANNs）是一种模拟动物神经网络的行为从而对信息进行并行处理的算法模型，常常简称为神经网络（Neural Networks，NNs）。神经网络依靠于内部大量的节点（称之为神经元）之间相互连接相互作用来实现模拟动物大脑的思维方式和组织形式的目的。

人工神经元模型如图 3-2 所示。

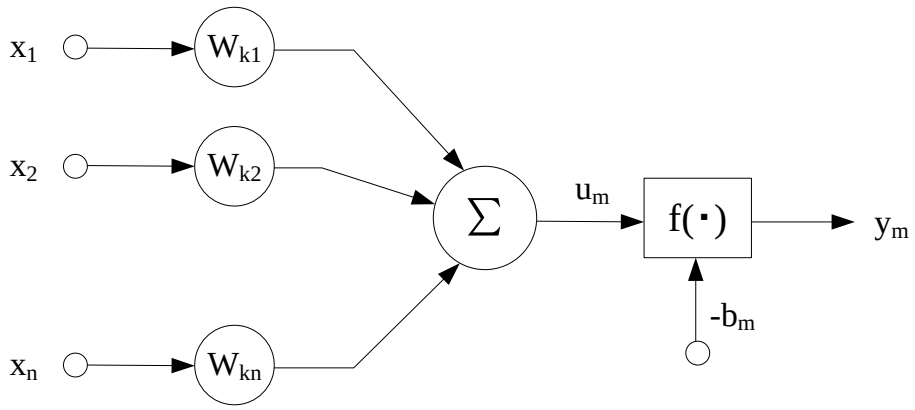

图 3-2 人工神经元模型示意

如图 3-2 所示，人工神经元模型可以描述为：

（1）输入节点的连接。输入向量 $X_m=(x_1,x_2,\cdots,x_n)^T$，其中，$x_i$,$1\le i\le n$ 表示第 i 个神经元的输入。连接强度为权值向量 $W_m=(w_{1m},w_{2m},\cdots,w_{nm})^T$，其中，$w_{im}$,$1\le m\le n$ 表示连接 i、j 两个神经元之间的权值。

（2）加权求和。以权值向量为系数对输入向量进行加权求和。

（3）激活函数。激活函数通常为一个非线性函数，以（2）中的加权和以及负偏置为参数，用来控制输出在一定的范围内。

用数学公式表示上述模型为：

$$u_m = \sum_{j=1}^{n} w_{mj} x_j \tag{3-18}$$

$$y_i = f(u_m - b_m) = f(\sum_{j=1}^{n} w_{mj} x_j - b_m) \tag{3-19}$$

其中，$f(\cdot)$ 是非线性激活函数，常采用 Sigmoid 函数和 tanh 函数。

3.3.2 BP 神经网络

BP（Back Propagation）神经网络是目前成功训练并且应用作为广泛的神经网络模型之一，由 Rumelhart 和 McCelland 等人于上个世纪 80 年代提出。BP 神经网络模型由输入层（input layer）、隐藏层（hidden layer）和输出层（output layer）组成，如图 3-3 所示。

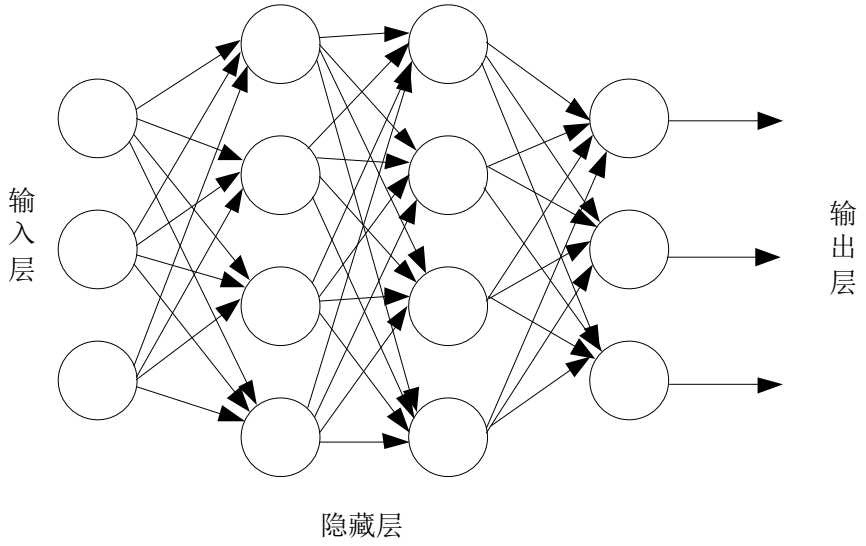

图 3-3 包含两层隐含层的 BP 网络结构示意

输入层接受外界信息，并且与中间的隐藏层相连，将信息传递个隐藏层的每个神经元。隐藏层主要对输入的信息按要求进行变换，是网络的内部处理层，可以是一层或者多层。输出层是网络与外界的连接层，将最后一层隐藏层传递来的信息经过处理后输出给外界。

对于给定的神经网络，在进行参数学习是，BP 神经网络采用的是反向传播算法。当实际的输出与预期的存在差距时，将会使误差从最末尾的输出层开始，逐层进行误差传播，并在不断的反向传播和前向传播中调整各层的权值，直至

网络的实际输出与预期之间的误差平方值达到指定的最小值范围，或达到一定的学习次数时，则停止学习[47]，传播中使用梯度下降的方法。

算法 3-1 包含两层 sigmoid 单元的 BP 神经网络算法

输入：D, η, n_{in}, n_{hidden}, n_{out}

样本集合 $D=\{(\vec{x}_1,\vec{t}_1),(\vec{x}_2,\vec{t}_2)...,(\vec{x}_{n_{in}},\vec{t}_{n_{in}})\}$，其中，$\vec{x}$ 是神经网络输入值的向量，$\vec{t}_1$ 是神经网络的目标输出值；η 是学习率；n_{in} 是输入的节点数；n_{hidden} 是隐藏层的节点数；n_{out} 是输出的节点数。

x_{ij} 表示 i 到 j 节点的输入，w_{ij} 表示 i 到 j 节点的权值。

输出：y_{out} 是网络输出的结果值

算法步骤：

- 创建具有 n_{in} 个输入，n_{hidden} 个隐藏单元，n_{out} 个输出单元的神经网络。
- 使用随机数初始化网络中所有的权值。
- 遇到终止条件前执行：
  - 对于训练样本 $D=\{(\vec{x}_1,\vec{t}_1),(\vec{x}_2,\vec{t}_2)...,(\vec{x}_{n_{in}},\vec{t}_{n_{in}})\}$ 中的每个样本 $(\vec{x},\vec{t})$：

    前向传播阶段

    1. 把输入值向量 $\vec{x}$ 输入网络，$\forall u \in net$，计算输出 o。

    反向传播阶段

    2. $\forall k \in output$，求 k 的误差项 δ_k

$$\delta_k \leftarrow o_k(1-o_k)(t_k-o_k) \tag{3-20}$$

    3. $\forall h \in hidden$，求 h 的误差项 δ_h

$$\delta_h \leftarrow o_h(1-o_h)\sum_{k\in outputs} w_{hk}\delta_k \tag{3-21}$$

    4. 更新网络的权值 w_{ij}

$$w_{ij} \leftarrow w_{ij}+\Delta w_{ij} \tag{3-22}$$

    其中

$$\Delta w_{ij}=\eta\delta_j x_{ij} \tag{3-23}$$

3.3.3 深度神经网络模型

深度神经网络与浅层学习模型相比，浅层学习模型的学习能力表达有限，尤其是在复杂的分类问题上，浅层学习模型的泛化能力非常有限，而面对复杂的函数，深度神经网络的参数则比较简洁，能更好的完成对特征的学习。深度神经网络的结构如图 3-4 所示。

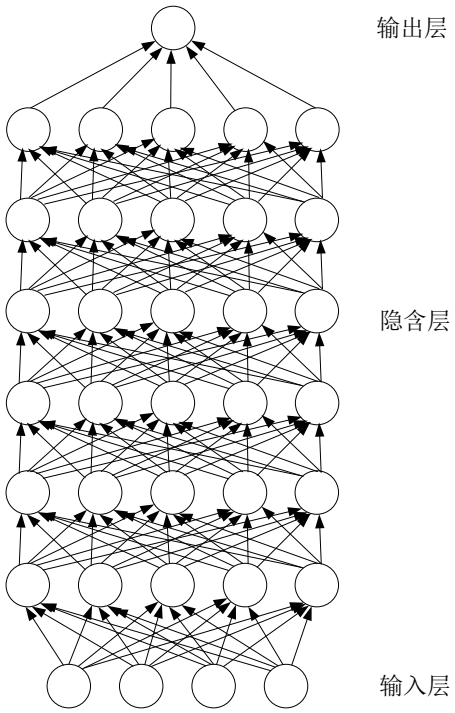

图 3-4 深度神经网络模型示意

图 3-4 所示的是由 1 层输入层、6 层隐藏层和 1 层输出层构成，在深度神经网络中，当前层的节点与前一层的所有节点、后一层的所有节点都是全连接，即对于当前层的某一个节点，它与前一层的所有节点均有相连的边，同理，它与下一层的所有节点也都有相连的边。

深度神经网络实际上是有多个人工神经元组个而成，总的来说，深度神经网络的每一层都是由前一层的所有节点求和后通过分类器或回归得到不同的节点所构成的，深度神经网络的参数学习则是通过误差逐层反向传播来修改相对应的权重完成的。对于深度神经网络，也可以分解为自底向上的输入沿网络前向传播阶段和自顶向下的误差沿网络反向传播阶段两个阶段。可以形式化表示为：

$$x \xrightarrow{\sum w_0 x+b_0} u^1 \xrightarrow{f(u^1)} o^1 \xrightarrow{\sum w_1 o^1+b_1} u^2 \xrightarrow{f(u^2)} o^2 \xrightarrow{\sum w_2 o^2+b_2} \cdots\cdots \xrightarrow{\sum w_{n-1} o^{n-1}+b_{n-1}} u^{out} \xrightarrow{f(u^{out})} o^{out} \tag{3-24}$$

$$w_1 \xleftarrow{\delta_1} w_2 \xleftarrow{\delta_2} w_3 \xleftarrow{\delta_2} \cdots\cdots \xleftarrow{\delta_{n-1}} w_{n-1} \xleftarrow{\delta_{out}} (t_{out} - o_{out}) \tag{3-25}$$

其中，x 是输入向量值，w_i 和 b_i 是第 i 层的权值和偏置，$f(\sim)$ 是网络层的选用的激活函数，δ_i 是第 i 层的误差项，误差项定义见公式(3-25)和(3-25)，t_{out} 是网络的目标输出值。在前向传播阶段，输入沿首层的输入节点进入网络，所有输入层节点求和后，经由分类器或回归的激活函数，得到隐藏层中的节点，得到的节点层又作为输入进行下一次的前向传播，计算公式参见公式(3-25)。在误差单

向传播阶段，误差从最顶端的输出层开始，计算实际的输出与预期的差距，根据 BP 神经网络模型方法，将误差由输出层至输入层反向进行传播并更新各节点对应的权值。

3.4 面向广告点击率预测的深度神经网络模型

尽管通过多层的连接，使得深度神经网络的能力更强，能表现出对特征更好的学习能力，但是深度神经网络在训练上却存在一些问题：

（1）参数的调节是一个难题，并且因为层数多节点多，训练中非常容易出现过拟合现象；

（2）训练耗时长。当层次较少时，网络的效果与其他方法相比并不突出，若网络层次增多时，随着深度神经网络层次的增加，训练网络所需要的时间将会急速增加。

在搜索广告中，训练数据非常稀疏，并且数据量非常大，使得想要利用全部的广告来训练深度神经网络从而进行广告点击率的预测是非常困难的。因此，针对（1），我们根据文献[46]的思路，使用 dropout 的方法，在模型训练阶段，以一定概率随机选择一些隐藏层的节点，使它们的权值在训练时不参与训练，如图 3-5 所示。

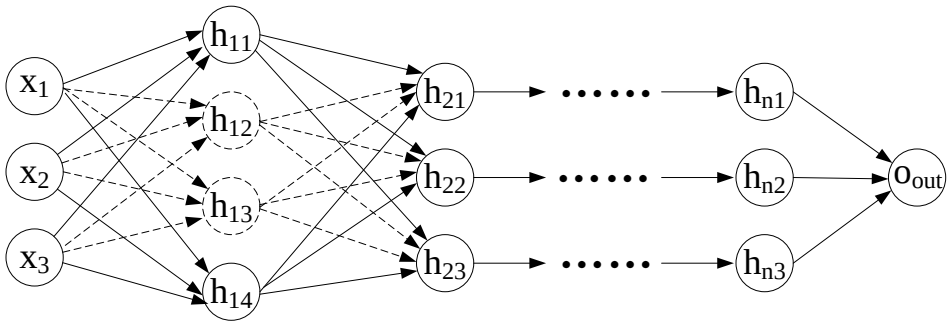

图 3-5 dropout 示意

使用 dropout 训练网络模型时，在输入沿网络前向传播过程中，隐藏层节点计算得到的输出以设定的概率被随机置零，即图 3-5 中虚线标出的边以及所连接的节点输出值被置零，这些输出值将不参与到这次传播的计算中，但是相应的权值并不会置零，这些权值会保留，只是不参与该轮的计算。当误差沿网络反向传播时，在前向传播中被保留的权值即未参与前向传播的那些权值，将不会被更新。

Dropout 之所以可以防止过拟合，一方面是因为在权值更新时是随机选择的，使得两个节点同时出现的概率降低，避免了一个节点依赖于另一个节点而造成模型泛化能力变弱，另一方面，因为每次计算都有一定比例的权值不参与更新，这些权值所对应的边被置零，每一轮的哪些边被置零是随机的，如此，

每一轮的计算都相当于不同的网络结构，而这些不同的网络结构共享原始网络的所有权值，类似于利用 bagging 的思想防止过拟合。

针对（2），由于搜索广告数据量大而稀疏，并且深度神经网络的层数较多，要克服训练速度缓慢的问题，就得提高计算的速度。GPU 往往集成了成百上千的计算单元，能够并行处理数据，具有强大的计算能力，与 CPU 相比，能大大提高深度神经网络的训练速度。

3.5 实验评测

这一节以实验为出发点来评估浅层学习模型和深度神经网络在搜索广告点击率预测问题上效果的优劣，并重点考察了深度神经网络在搜索广告点击率预测上的有效性。

3.5.1 实验设置

本论文所采用的数据集的说明在第 2 章中已经进行了介绍，本章所使用的如表 2-3 所示，随机抽取 10%的训练样本作为训练集，测试集使用 KDD Cup 2012 中 Track 2 的公共测试集，参数的调节在验证集上进行。

在进行深度神经网络实验时，我们根据第 2 章中 2.5 小结的思想，将训练样本和测试样本进行分块，依次进入显存进行实验。综合硬件设备的限制和对实验效果的影响，我们以每 500000 条训练样本组成为一块，将本章所用的训练集划分成 30 块。同理，我们对测试集进行同样的操作，共将测试集划分成 35 块。

实验中，支持向量回归模型采用 Liblinear[3]工具包来训练和预测，深度神经网络在 Deepnet[4]的基础上作改动后用于训练和预测。

为了评价深度神经网络模型与其他基线模型实验的性能，本章使用广告点击率预测领域中常用的 AUC 指标作为评价指标。AUC 的定义和计算方法在第 2 章中已经给出，这里不再赘述。

3.5.2 实验结果与分析

在本小节中，我们首先对点击率预测方法中的主流模型——浅层学习模型进行了实验，包括朴素贝叶斯模型、逻辑斯蒂回归模型和支持回归模型，对 2.4 中所提取的高影响力特征和词向量特征的实验效果进行了分析。而后，考虑到在深度神经网络中，网络的层数和每层的节点数对网络结果的影响，我们对着

[3] http://www.csie.ntu.edu.tw/~cjlin/liblinear
[4] https://github.com/nitishsrivastava/deepnet

两个参数通过实验进行选择，并在选好的参数下使用 dropout 方法进行实验，最后对实验的结果进行对比和分析。

3.5.2.1 浅层学习模型实验

（1）我们首先朴素贝叶斯模型、逻辑斯蒂回归模型和支持回归模型三个模型进行了实验，实验结果如表 3-1 所示。

表 3-1 浅层学习模型实验结果

模型	特征	AUC
朴素贝叶斯	AdID + QueryID + KeywordID + PositionID + UserID	0.7532
逻辑斯蒂回归	f_Sparse + p_CTR + Similar + Pos + W2V	0.7780
支持向量回归	f_Sparse + p_CTR + Similar + Pos + W2V	0.7674

在实验中，朴素贝叶斯模型利用 AdID，QueryID，KeywordID，PositionID，UserID 五个属性计算出它们的先验概率，计算中，考虑到数据稀疏造成的零值现象，我们使用了加一平滑的方法进行求值，而后根据公式(3-25)和(3-25)计算后验概率用以预测。逻辑斯蒂回归模型和支持向量回归模型分别使用了 2.4 小节中的类别稀疏特征、历史点击率特征、相似度特征、位置信息特征和词向量特征来训练模型。

从实验结果可以看出，逻辑斯蒂回归模型和支持向量回归模型的实验效果比朴素贝叶斯效果要好。分析原因，朴素贝叶斯模型是基于贝叶斯定理定义的，即需要满足各属性之间相互独立，而在实际应用中，搜索广告中各属性是相互影响的，比如，当搜索“雨伞”时，用户自身的偏好和广告本身的关键词属性等，都会影响到广告的点击概率，这些属性之间并非满足独立性。

（2）我们对特征在实验中产生的影响进行了分析，参考文献[7]，在同样的数据集上，类别稀疏特征、历史点击率特征、相似度特征和位置信息特征已经得到了验证，这里我们着重对高影响力特征和词向量特征的作用进行研究。

表 3-2 高影响力特征的实验效果

模型	特征	AUC
逻辑斯蒂回归	p_CTR + Similar + Pos	0.7331
	p_CTR + Similar + Pos + **High-impact**	**0.7412**
支持向量回归	p_CTR + Similar + Pos	0.7296
	p_CTR + Similar + Pos + **High-impact**	**0.7353**

我们将 AdID，AdvertiserID，QueryID，KeywordID，UserID 五个属性按照 one-hot 编码的形式组合在一起得到 13090376 维的特征，通过使用 L1 范数正则化的逻辑斯蒂回归模型得到稀疏解，从而提取高影响力特征，增加高影响力特

征时，实验结果的变化如表 3-2 所示。

结果表明，当增加高影响力特征时，在两个不同模型上的实验的效果都有不同程度提升。进一步得，针对模型得到的稀疏解中的 176771 个非零权值，我们统计了不同属性所占的数量，如图 3-6 所示。

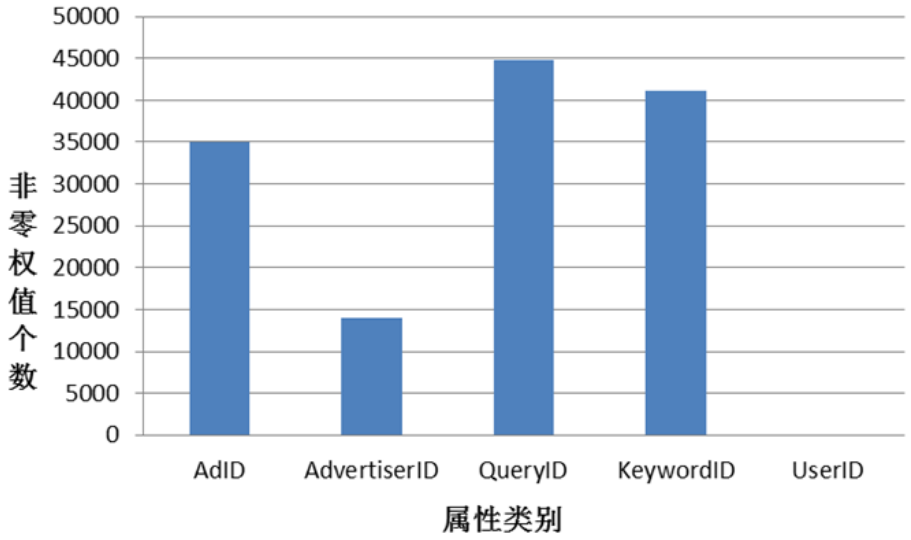

图 3-6 不同属性在稀疏解中非零权值的个数统计

由图 3-6 可以看出，相比之下，QueryID 和 KeywordID 两者的非零权值最多，也意味着在这五个属性中，QueryID 和 KeywordID 两个属性最为重要，而 UserID 个数为零，从实际场景出发，利用用户信息来进行预测时，往往利用用户与用户之间、用户与商品之间的行为交集（如相关性）来影响预测结果，如对于用户 user1，user2，user1 的 UserID 并不能直接影响对 user2 行为的预测，而是通过计算 user1，user2 的相关性等作为特征来进行预测。

对词向量特征的研究如下。我们使用谷歌的开源工具 Word2Vec[5]训练数据集中的文本文件 titleid_tokensid.txt，descriptionid_tokensid.txt，得到不同维度的词向量特征，实验结果的变化如表 3-3 所示。

表 3-3 词向量特征的实验效果

模型	特征	AUC
逻辑斯蒂回归	p_CTR + Similar + Pos	0.7331
	p_CTR + Similar + Pos + **W2V**	**0.7430**
支持向量回归	p_CTR + Similar + Pos	0.7296
	p_CTR + Similar + Pos + **W2V**	**0.7413**

从表 3-3 可以看出，加入词向量特征后，实验结果有明显的升高。因为词

[5] http://word2vec.googlecode.com/svn/trunk/

向量通过一层神经网络的学习后，将原本每一维孤立的信息融合到了一起，如对 descriptionid_tokensid 进行学习时，学到的每一个词向量综合了 tokensid 词表中的所有信息，具有一定的关联度，与其他特征在一起相互作用时，能得到更好的预测效果。

3.5.2.2 深度神经网络模型

在进行深度神经网络模型的相关实验前，我们先对本文提出的基于 GPU 的分块计算方案的可行性进行了验证。我们在训练数据中随机选取设备显存可容纳的最大数据块（500000 个训练样本）P 训练模型，实验结果记为 $result\_g$；将其均分为 n 块 $p_1,\cdots,p_n$（$n\geq 2$）后依次用来放入显存训练模型，实验结果记为 $result\_l$，比较 $result\_g$ 和 $result\_g$ 的实验效果，如表 3-4 所示。实验中使用包含 4 层隐藏层的深度神经网络模型，使用特征 p_CTR、Similar、Pos，输入层到输出层中每层节点数分别为：196，1024，1024，1024，1024，1。为了避免数据随机选择的不确定因素，我们进行多次实验，取平均值作最后的实验结果。

表 3-4 分块计算的实验效果比较

分块结果	AUC
result_g	0.6283
result_l(n=2)	0.6188
result_l(n=3)	0.6103
result_l(n=4)	0.6072

由表 3-4 的实验结果可以知道，在可接受的结果损失范围内，分块计算在设备受限的情况下是可行的，并且我们发现，对同一块数据，分块数量多将导致结果的下滑，因此，在本论文中，我们分块采取的方案是在设备可承受的情况下，尽量减少分块的数量。

根据 3.3.3 小节中深度神经网络的介绍可知，当输入固定时，网络中隐藏层的层数、各层中节点的总数等都会对实验的结果产生影响。因此，在本小节中，我们首先进行了 2 组实验，第 1 组实验是关于深度神经网络的隐藏层层数设置的实验，第 2 组实验是关于深度神经网络各层中节点数设置的实验。进行完这两组基本实验后，我们进行了面向搜索广告点击率预测的深度神经网络的实验，即使用 dropout 方法对点击率进行预测，并对 dropout 方法中对实验效果产生影响的参数进行了选择。最后，我们使用浅层学习模型做对比，分析了深度神经网络在点击率预测问题上的有效性。在这些实验中，为了保证实验的公平性，实验都使用相同的特征：历史点击率特征，相似度特征，位置特征和高影响力特征。

（1）首先，我们针对深度神经网络中隐藏层层数的设置进行实验，实验结

果如图 3-7 所示。

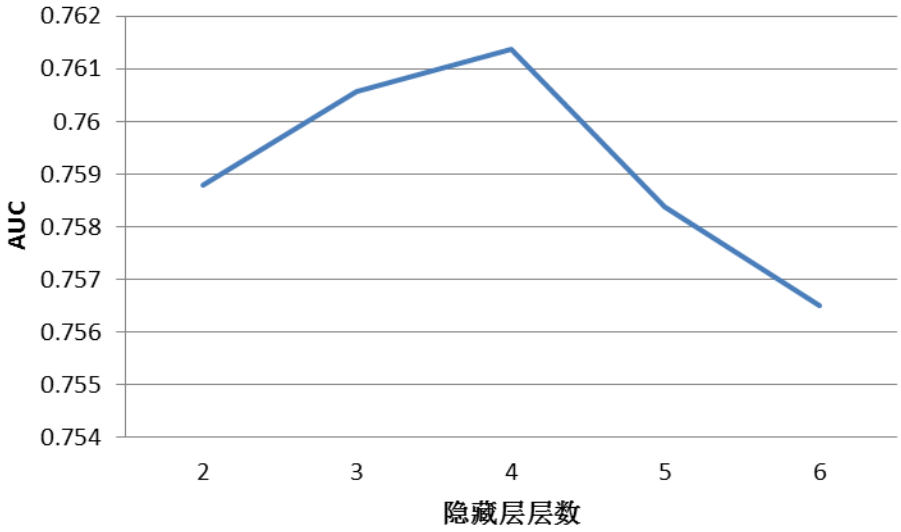

图 3-7 深度神经网络隐藏层设置结果图

由图 3-7 可以看出，随着网络中隐藏层数的增加，预测结果的 AUC 值升高了，但层数增长到一定程度（这里为 4 层隐藏层），预测结果的 AUC 值不但没有再增加，反而降低了。根据网络的结构，我们知道，网络的输入通过输入层传入网络后，随着隐藏层不断学习和传播至输出层，中间的隐藏层越多，对输入特征的学习就越多。在一定范围内，对特征的学习越多越有利于模型的训练，即在预测阶段能得到更好的结果，但是，在机器学习领域，一个很明显的问题是，当对输入数据学习的过于充分，会导致模型对训练数据过拟合，而在预测阶段，泛化能力不够，达不到好的预测效果。针对本文的训练集，当隐藏层层数超过 4 层时，会导致模型过拟合，以至于在预测阶段，对测试集数据的泛化能力太弱而降低实验效果。

（2）随后，我们针对深度神经网络中隐藏层层数的设置进行实验，根据深度神经网络隐藏层层数设置的实验，我们在后续的实验中固定隐藏层的数目为 4。在这组实验中，我们从前向后依次调节网络中节点的数量，例如，当对第 1 层隐藏层的节点数进行选择时，固定后面隐藏层的节点数不变，实验结果用以对比节点数变化时效果的好坏，从而选出第 1 层隐藏层中使得预测效果最好的节点数。而后，在已选出第 1 层隐藏层节点数的情况下，按照上述方法，固定第 1 层、第 3 层、第 4 层隐藏层节点数，对第 2 层隐藏层的节点数进行选择。剩余的隐藏层节点数按照该方法一次进行节点的选择。实验结果如图 3-8 所示。

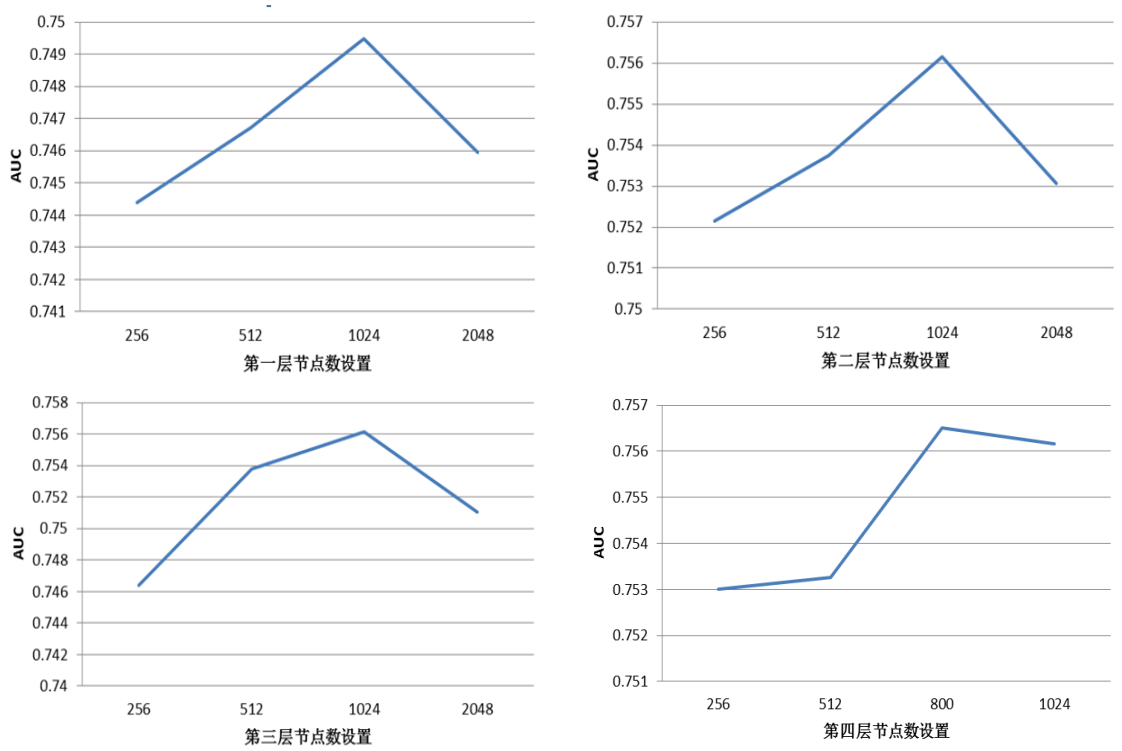

图 3-8 深度神经网络各隐藏层节点数设置实验结果图

在图 3-8 的实验中，我们首先固定第 2 层、第 3 层、第 4 层隐藏层节点数分别 512、1024、1024，对第 1 层隐藏层的节点进行调节，实验中，我们发现，在类似隐藏层层数变化的趋势，随第 1 层隐藏层的节点数的增加，预测结果会增加，当超过 1024 节点时，实验结果开始下降，对数据产生了过拟合，因此，我们选择第一层的节点数为 1024。固定第 1 层、第 3 层、第 4 层隐藏层节点数分别 1024、1024、1024，选择出第 2 层隐藏层的节点数仍是 1024，按照此方法，依次选择得到第 3 层和第 4 层隐藏层的节点数为 1024 和 800。

由实验可知，在尝试的节点范围内，预测结果的 AUC 值大部分都满足先增后减的情况，最后在局部点达到极大值。隐藏层节点过多，是使得特征学习太过细致，比如，对广告历史数据中用户 A、B、C 的信息学习刻画的过于详细，当测试集中出现新的用户 D 与 A、B、C 差异较大时，很难从对 A、B、C 的学习中预测出 D 的行为。

此外，在我们实验中选择出的最优隐藏层的节点数为 1024，1024，1024 和 800，输入层节点数为 196，输出层节点数为 1。分析这个结果，我们发现，最后一层隐藏层的节点数变少了，在特征的学习过程中起到了综合的作用，将前几层隐藏层学到的内容融合后传递给输出。

（3）Dropout 方法是在模型训练阶段，以一定概率随机选择一些隐藏层的节点，是它们的权值在训练时不参与训练，而这个概率的选择，将影响到整个

模型的实验结果。

我们选用包含 4 层隐藏层的深度神经网络，控制每层的节点数固定，为：194，1024，256，512，2048。分别选取 0.25、0.375、0.5、0.625、0.75 作为随机概率用来隐藏节点，实验结果如图 3-9 所示。

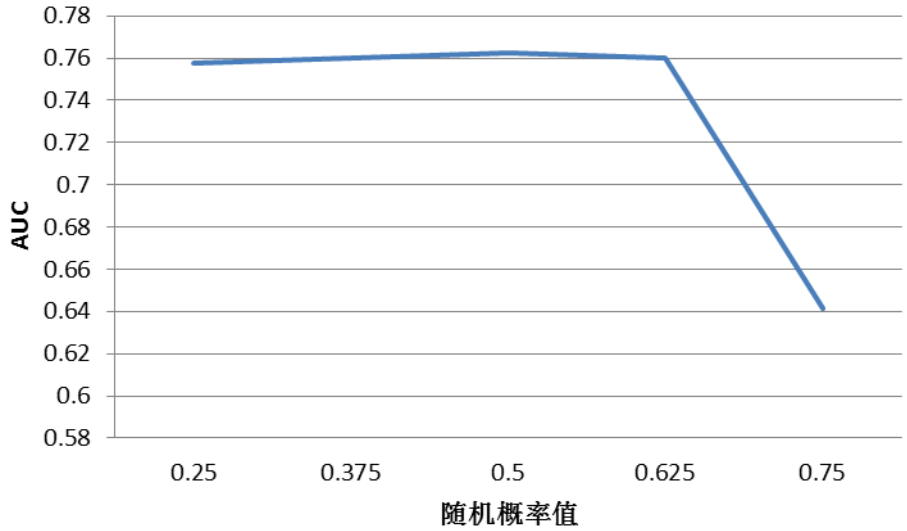

图 3-9 随机概率值实验效果图

结合 dropout 示意图 3-4，当 dropout 方法中的随机概率取 0 时，网络结构就是所有节点两两相连的深度神经网络结构；当 dropout 方法中的随机概率取 1 时，网络结构中无连接。若网络中只有 1 个节点相连，则为网络结构与基本的人工神经元模型类似。实验表明，当 dropout 方法中的随机概率取 0.5 时，效果最好，即当每次随机选择一半的节点参加训练时，预测结果最好。

（4）最后，我们在使用特征相同的情况下，将深度神经网络和逻辑斯蒂回归、支持向量回归模型对比，在测试集上的实验结果如表 3-5 所示。

表 3-5 深度神经网络实验结果对比

模型	特征	AUC
深度神经网络+dropout	p_CTR + Similar + Pos + High-impact	0.7732
深度神经网络	p_CTR + Similar + Pos + High-impact	0.7684
逻辑斯蒂回归	p_CTR + Similar + Pos + High-impact	0.7412
支持向量回归	p_CTR + Similar + Pos + High-impact	0.7353

实验结果表明，在使用相同特征的情况下，深度神经网络的效果优于逻辑斯蒂回归和支持向量回归模型，因为与这两个浅层学习模型相比，深度神经网络模型对特征进行了更多的学习。此外，通过实验说明，使用 dropout 方法确实能减少深度神经网络中过拟合的影响，提高预测的结果。

3.6 本章小结

本章首先对基于浅层学习模型的搜索广告进行了点击率预测，详细介绍了朴素贝叶斯模型，逻辑斯蒂回归模型和支持向量回归模型，作为点击率预估较为主流的基本模型，我们依次对它们进行了实验。并在此基础上，我们对每个特征在实验中的贡献分别进行了研究和分析。分析这些浅层模型，在对特征的学习中，它们都是孤立的使用每一维的特征，并没有深入的学习特征之间的相互关系，鉴于此，我们提出了基于深度神经网络的方法对搜索广告的点击率进行预测。

深度神经网络是深度学习中最为简单的中模型，在模型结构上，它由多个人工神经网络堆积组成，与浅层学习模型相比，深度神经网络在表示复杂的函数时需要的参数少，能更好地进行特征的学习。我们对深度神经网络模型结构上隐藏层层数设置和每层节点数的设置进行了实验，针对本文的数据选出了最优参数。然而，深度神经网络也正是因为层数多节点多，在训练过程中容易出现过拟合的想象，我们在面向大数据背景的搜索广告点击率的预测问题中，采用 dropout 方法来解决过拟合问题，并通过实验验证了方法的有效性。

值得注意的是，深度学习模型在训练时最大的困难之一便是训练时间的问题，我们并利用多计算单元集成的 GPU 计算来提高模型训练的速度，考虑到实验设备显存大小的限制，进而提出了分块计算的思想，最后完成了深度神经网络模型的训练。

第 4 章 基于卷积神经网络模型的广告点击率预测

4.1 引言

浅层学习模型在特征学习方面主要是利用基于统计学习的方法计算得到的，特征中的每一维的含义固定且孤立，在模型的学习中，并没有挖掘出特征之间的关系，在处理复杂问题时，往往是使用多个分类器进行特征融合，却没有反映出内部关系。

第 3 章提出的深度神经网络，在特征学习方面解决了浅层模型中特征孤立、特征关系挖掘缺失的问题，它利用全连接网络将相邻两层的节点两两相连，每一个计算的输出都与特征所有的维度相关联，每层的输出都在一定程度上刻画出了上一层所有节点之间的关系。

然而，深度神经网络虽然在一定程度上刻画出了特征之间的关系，但却比较粗糙，并没有从局部到整体的认识层次来学习特征。本章对基于卷积神经网络（Convolution neural network，CNN）[45, 48-52]的 CTR 预测进行研究，通过卷积与亚采样操作的结合，能更好地学习出数据特征之间的关系，不仅解决了线性模型无法模拟真实广告数据场景的问题，也解决了一般非线性模型无法深入挖掘特征间相互关系的问题，并且较之于传统的神经网络，卷积神经网络能更好的理解特征之间的关系。

4.2 卷积神经网络模型

卷积神经网络是人工神经网络的一种。早在 Hubel 和 Wiesel 进行猫脑皮层的相关研究的时候，发现了一种独特的可以降低神经网络复杂性的结构——局部感受的神经，因此提出了局部感受野（Receptive Field）的概念，局部感受野相当于在一个小的窗口，模拟动物在感受物体时由局部到整体的过程，即先从局部感受野的窗口中感受到物体的部分，然后通过神经网络将局部感受野的内容进行综合，得到动物大脑中完整的物体全貌。此后研究者将局部感受野用于人工神经网络中，提出了卷积神经网络模型。

卷积神经网络区别于其他神经网络，它拥有稀疏连接和权值共享的特点，使其能更好的模拟动物大脑神经的工作过程，对网络的输入能进行更为全面有效的学习。卷积神经网络是语音分析和图像识别方向上的研究热点，目前已取得了不少的成就。

卷积神经网络在结构上有两个重要的组成部分：卷积层和亚采样层。亚采

样层随卷积层的出现而出现，如图 4-1 所示。一个完整的卷积神经网络包括输入层、卷积层、亚采样层和输出层，在同一个卷积神经网络中，卷积层数和亚采样层数是相同的。

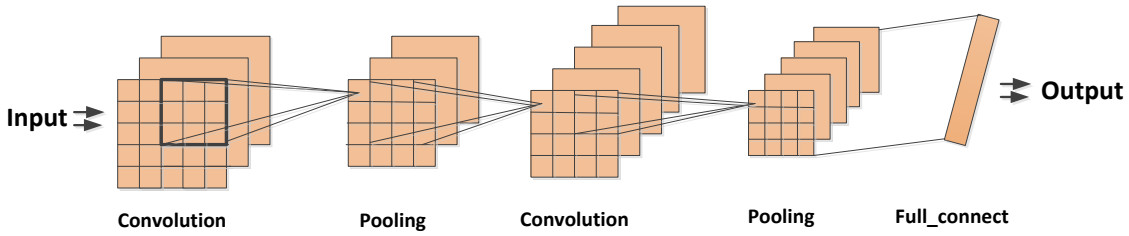

图 4-1 卷积神经网络结构示意

卷积神经网络的特性和结构将在下面的小节中进行详细介绍。

4.2.1 稀疏连接

在深度神经网络中，网络中每一层的神经元节点与相邻层之间的节点是；两两相连的。在卷积神经网络中，因为引入了局部感受野的概念，层与层之间的连接不再是两两相连，取而代之的是通过局部感受野进行连接的，即每一层的神经元节点只与相邻层的局部节点相连。如图 4-2 所示。

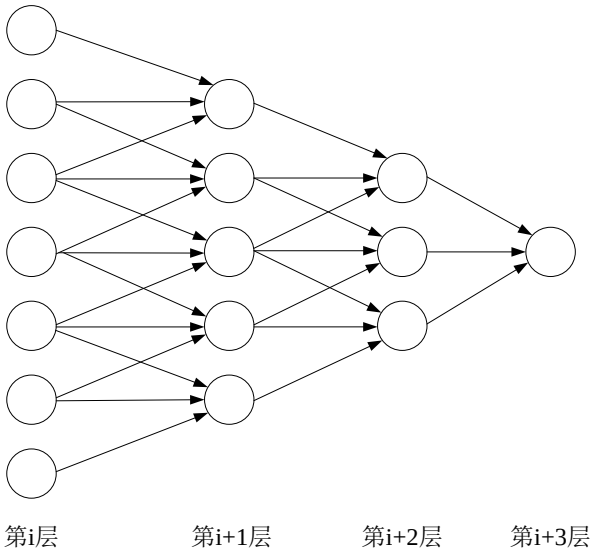

图 4-2 卷积神经网络稀疏连接方式示意

假设第 i 层是输入层，分别经过第 i+1、i+2 层，到达输出层第 i+3 层。在深度神经网络中，第 i+1 层的每个神经元节点都与第 i 层的所有节点相连，第 i+2 层、第 i+3 层同理。在卷积神经网络中，假设局部感受野的大小设定为 3 个节点，每次移动的步长设定为 1，则如图 4-2 所示，第 i+1 层中的某一节点至于第 i 层中与其相近的局部范围内的 3 个节点相连，即这个节点只是对前一层中局部的 3 个节点进行学习，如此，顺序的对第 i 层中的 7 个按照局部感受野为 3

的大小进行学习，得到第 i+1 层的 5 个节点。以此类推，网络是对局部学习到的内容层层融合传递，这样的结构大大降低了神经网络中参数的规模。

4.2.2 权值共享

权值共享能有效降低模型训练的参数，是卷积神经网络的特性之一，也是成功训练卷积神经网络的要素之一。

由稀疏连接的特性可知，在卷积神经网络中，层与层之间的计算单位是局部感受野，相应的，与局部感受野相作用的是卷积滤波器（也称卷积核）。以图像为例，在卷积神经网络中，卷积层的每一个滤波器重复作用在整个输入上的不同感受野上，从而对输入的图像进行卷积学习，提取出图像的局部特征，其结果构成输入的特征图（feature map）。如图 4-3 所示。

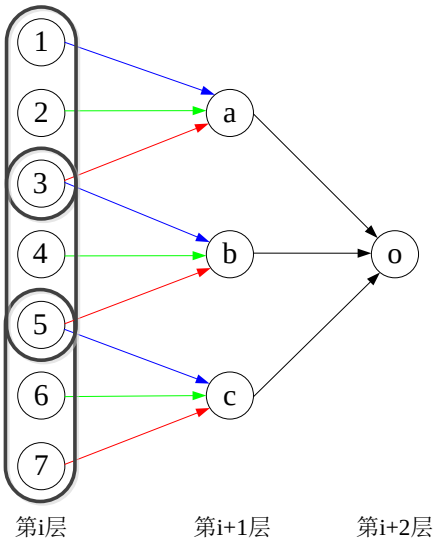

图 4-3 卷积神经网络权值共享示意

假设图 4-3 中的局部感受野大小为 3，每次移动的步长设定为 2，不同颜色的边表示不同的权值，为了方便理解，我们给每一个节点进行了编号。第 i+1 层的节点 a 由第 i 层局部感受野中的节点 1、2、3 与卷积滤波器共同作用得到，计算第 i+1 层的节点 b 时，第 i 层中的局部感受野以步长为 2 移动，由第 i 层中的节点 3、4、5 与卷积滤波器共同作用得到。计算中，虽然局部感受野进行了移动，但是卷积滤波器相应的权值不变，如图 4-3 中，节点 1 使用的权值与节点 3 使用的权值相同，节点 2 使用的权值与节点 4 使用的权值相同，节点 3 使用的权值与节点 5 使用的权值相同。

权值共享在学习特征时不用考虑局部特征的位置，每一个卷积滤波器使用相同的权值，大幅度的降低了卷积神经网络模型学习中的参数，为成功训练卷积神经网络提供了有效的保障。

4.2.3 卷积层

卷积层，顾名思义是完成卷积过程的结构层次。在卷积层中，原始特征通过卷积核进行卷积得到输出的特征，使用不同的卷积核就可以得到一系列不同的输出特征。对卷积层的计算，我们有如下公式：

$$x_j^l = f(\sum_{i \in p_j} x_i^{l-1} * k_{ij}^l + b_j^l) \tag{4-1}$$

这里，$f(\sim)$是 tanh 函数，$u = \sum_{i \in p_j} x_i^{l-1} * k_{ij}^l + b_j^l$；$p_j$代表输入特征上选定的窗口（局部感受野），即在卷积过程中当前卷积核在计算时所对应于输入特征上的位置；x_i^{l-1}是第$l-1$层输入特征上第i个窗口相应的值，x_j^l是第l层输入特征上第i个窗口相应的值；k_{ij}^l是第l层上位置(i, j)所对应的卷积核的权值；b_j^l是特征的偏置，每一层对应一个。

卷积过程，一个卷积核通过滑动会重复的作用在整个输入特征上，构建出新的特征。同一个卷积核进行卷积时，共享相同的参数，包括同样的权重和偏置，这也使要学习的卷积神经网络参数数量大大降低了。而当我们使用不同的卷积核进行卷积时，可以得到相应的不同的输出特征，这些输出特征组合到一起，构成卷积层的输出。

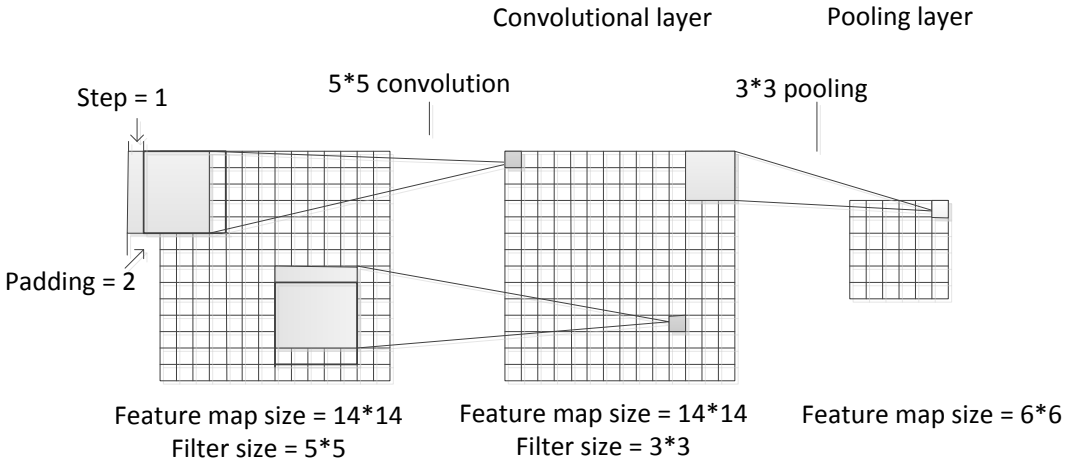

图 4-4 卷积层与亚采样层结构

如图 4-4 所示，以输入为 14×14 的输入为例，设定窗口大小为 5×5，从而卷积核（filter）的大小同样也是 5×5，填充（padding）步长设为 2，移动步长（step）设为 1。具体的卷积操作是：输入特征上以所设定窗口的大小与卷积核进行加权求和，加权求和的结果通过 tanh 函数得到一个值，这个值便是下一层的一个节点的输入值。随后，窗口按步长移动 1，重复上述计算过程。所有的计算输出值组合在一起，构成下一层节点的输入值。

4.2.4 亚采样层

亚采样层是紧随着卷积层而出现的，在这一层结构中，将对卷积层的输出结果进行亚采样，很多研究中也叫最大池采样。

在亚采样层，前一个卷积层的输出将作为该层的输入特征，首先设定亚采样窗口的大小，每次亚采样操作都是针对一个特定的窗口进行的。然后通过滑动固定大小亚采样窗口，用窗口区域中最大（或平均）的特征值来表示该窗口中的特征值，将这些特征值组合到一起，得到降维后的特征。亚采样过程可表示如下：

$$x_j^l = f(pool(x_i^{l-1}) + b_j^l) \tag{4-2}$$

这里，与卷积层类似，x_i^{l-1}是第$l-1$层输入特征上第i个窗口相应的值，x_j^l是第l层输入特征上第i个窗口相应的值，b_j^l是特征的偏置；$pool(\sim)$表示取最大值$Max(x)$（或者平均值$Avg(x)$）的函数。如图 4-5 所示为采用$Max(x)$函数的亚采样过程，设定亚采样窗口大小为 3×3，并使用不重叠的窗口滑动，即窗口滑动的步长为窗口的长度，每次取出窗口中的最大值作为结果。

图 4-5 亚采样操作示意

亚采样操作减少了对数据的依赖性，增强了模型的泛化能力，因为在亚采样操作中，使用一个值来代表一个窗口，当窗口中个别值进行变化时并不影响整个卷积神经网络的效果。此外，亚采样层在卷积神经网络中起到了降维的作用，降低来自卷积层的计算复杂度。

4.2.5 全连接层

典型的卷积神经网络除了包含输入层、n(n>=1)个卷积层和亚采样层、输出层以外，通常在网络的输出层前还连接有 m(m>=1)层的全连接层。一个亚采样层跟随在一个卷积层后出现，通过这若干卷积层和亚采样层后得到的特征，将

经过若干个全连接层与输出层相连。全连接层公式如下：

$$x^l = f(u^l)，其中 u^l = K^l x^{l-1} + b^l \tag{4-3}$$

这里，$f(\sim)$是 sigmoid 函数，$f(u) = \frac{1}{1+e^{-u}}$，$K^l$是计算第$l-1$层到第$l$层时的权值，$x^l$是第$l$层输入特征上相应的值，$x^{l-1}$是第$l-1$层输入特征上相应的值，$b^l$是第$l$层特征的偏置。

卷积层和亚采样层都应用了局部感受野的思想，在网络的传播中不断的对不同区域的局部特征进行学习。全连接层是对之前学习到的特征的一个融合，将从众多局部特征综合到一起进行学习。

4.3 面向广告点击率预测的卷积神经网络模型

根据 4.2 节中的描述，我们将搜索广告点击率的特征应用在 CNN 模型中，进行广告点击率的预测。

4.3.1 基于卷积神经网络的广告点击率预测模型结构

基于卷积神经网络的广告点击率预测的模型结构如图 4-6 所示。

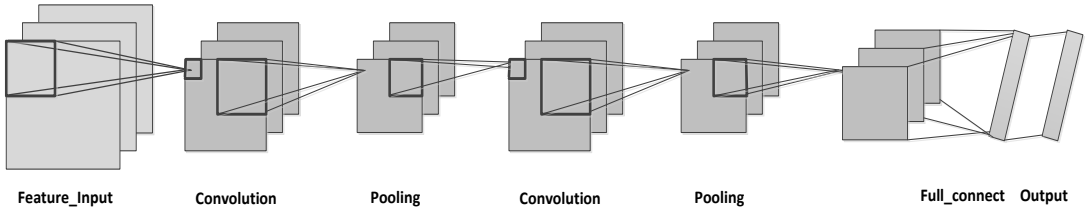

图 4-6 卷积神经网络在搜索广告点击率预估中的应用

实验模型分别设置了一层输入层、两层卷积层、两层亚采样层、一层全连接层和一层输出层。

4.3.2 基于卷积神经网络的广告点击率预测过程

本论文实验中通过实验对比，一共设置了两层卷积层、两层亚采样层以及一层全连接层。

首先从实验数据集提供的历史日志中提取相应的特征，构建得到输入（Feature_Input），对局部感受野即卷积的窗口大小进行设置，并设置好窗口滑动的步长，根据公式(4-3)对输入特征进行卷积操作。每一次卷积是将局部感受野中的所有值与卷积滤波器相进行加权求和然后通过激活函数进行求值的过程，其实质相当于对窗口内所有值的组合，因此卷积过程相当于特征融合过程。

经过卷积操作后得到的结果将作为输入传递给亚采样层。在亚采样层，同样需要先设定好亚采样窗口的大小，以及窗口滑动的步长，多数情况下，在亚采样层的窗口滑动是不重叠的，即滑动的步长等于窗口的长度，后续的实验中，我们将对这一设置进行探讨。设置亚采样的参数后，根据公式(4-3)进行最大值采样，即选取窗口中值最大节点值的代表整个窗口的输出值，可以理解为选取出窗口中的最有表达能力的特征值来表示整个窗口的特征，因此亚采样过程相当于特征的萃取过程。亚采样的输出将作为输入传递给下一层卷积层，此后的卷积操作和亚采样操作按照前面叙述的过程依次进行。

随后，我们将特征经过两层卷积层和两层亚采样层后得到的输出作为输入传递给全连接层。在全连接层，上一层亚采样的结果不再按照局部窗口进行计算，而是将它们全部展开，根据公式(4-3)进行计算。

最后，全连接层与输出层进行全连接，同样根据公式(4-3)进行计算，得到的输出便是最终的预测结果。

在整个预测过程中，一次特定的卷积操作过程中，即训练的一次迭代过程中，权值并不会随着窗口的滑动而改变，也就是说，在计算过程中，所有窗口滑过的特征享受同样的权值。这也是正是 CNN 区别于其他神经网络的特点——权值共享。这使得 CNN 更方便训练，更能多角度的对特征进行学习。

4.4 实验评测

这一节以实验为出发点来评估使用卷积神经网络来预测广告点击率的效果好坏，并将该方法与浅层学习模型和深度神经网络模型进行对比和分析。

4.4.1 实验设置

本论文所采用的数据集的说明在第 2 章中已经进行了介绍，本章所使用的如表 2-3 所示，随机抽取 10%的训练样本作为本文实验中使用的训练集，测试集使用 KDD Cup 2012 中 Track 2 的公共测试集，参数的调节在验证集上进行。

尽管已经对训练样本进行了采样，但是对于硬件配置来说，所有的样本仍不能一次性放入显卡内存中实验，因此，在训练卷积神经网络模型时，我们根据第 2 章中 2.5 小结的思想，将训练样本和测试样本进行分块，依次进入显卡内存进行实验。综合硬件设备的限制和对实验效果的影响，我们以每 500000 条训练样本组成为一块，将训练集划分成 30 块。同理，我们对测试集进行同样的操作，将测试集划分成 35 块。实验中，卷积神经网络在 Deepnet[6] 的基础上作

[6] https://github.com/nitishsrivastava/deepnet

改动后用于训练和预测。

为了评价卷积神经网络模型与其他基线模型实验的性能，本章使用广告点击率预测领域中常用的 AUC 指标作为评价指标。AUC 的定义和计算方法在第 2 章中已经给出，这里不再赘述。

4.4.2 实验结果与分析

根据 4.2 节中卷积神经网络的介绍可知，当输入固定时，卷积核的个数、卷积窗口的大小、移动的步长等因素都将影响到计算得到的节点个数，进而对实验的结果产生影响。因此，在本小节中，我们共进行了 5 组实验，第 1 组实验是关于卷积神经网络的卷积核个数设置的实验，第 2 组实验是关于卷积神经网络中层数设置的实验，第 3 组实验是关于卷积神经网络中窗口大小设置的实验。第 4 组实验我们探讨了各个特征对实验结果的贡献程度。最后，我们使用逻辑斯蒂回归模型（LR）、支持向量回归模型（SVR）和深度神经网络（DNN）作为对比方法，为了保证实验的公平性，对比实验都使用相同的特征：历史点击率特征，相似度特征，位置特征和高影响力特征。

（1）我们首先对卷积神经网络中卷积核个数的设置进行了实验。为了排除其他因素对卷积核个数设置的影响，我们将网络中的其他影响因素固定，设置网络的层数为 7 层（输入层、两层卷积层、两层亚采样层、一层全连接层和输出层），卷积窗口长宽分别为 5，滑动步长为 1，亚采样窗口长宽分别为 3，滑动步长为 3，此外，我们选用了 Dense Gaussian 对卷积层、亚采样层的边和节点进行初始化，用常数初始化输出层，学习卷积神经网络各边权值时的优化函数使用梯度下降算法，其中学习率为 0.01、动量项为 0.9，训练步数为 100。设置公式(4-3)中参数 α=0.05， β=75。我们在验证集上得到的结果如图 4-7 所示。

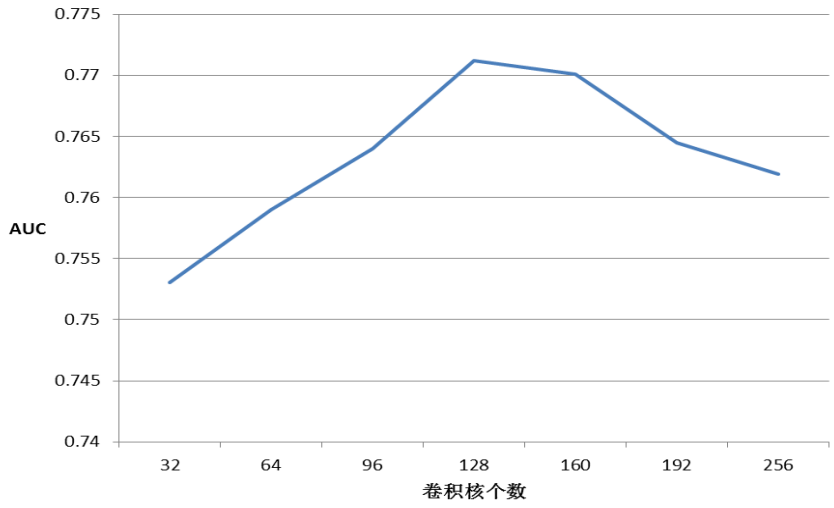

图 4-7 AUC 随卷积核个数变化的趋势图

从图 4-7 中 AUC 的值随卷积核变化的趋势图可以看出，在一定范围内，AUC 的值随着卷积核的增加而增加，在验证集上的结果显示，当卷积核个数少于 128 时，结果的 AUC 值呈上升趋势，当卷积核数多于 128 时，结果的 AUC 值开始降低，当选用 128 个卷积核的时，结果的 AUC 值达到局部最高。分析原因，通过 4.2 节卷积神经网络的介绍我们知道，使用不同的卷积核对特征进行卷积，实际上相当于从不同的方面对输入特征进行学习，使用过多的卷积核，会导致对训练样本的输入特征的学习太过细致全面，而在对测试样本的预测中表现出较差的泛化能力。

（2）在第二组实验中，我们针对卷积神经网络的层数对实验结果的影响进行了研究。在卷积神经网络中，卷积层和亚采样层是同时出现的，因此，为了方便描述，我们将一层卷积层和一层亚采样层组合在一起统称为一个隐藏层，即若我们描述网络为有两个隐藏层时，意味着网络中包含有两层卷积层和两层亚采样层。在对网络隐藏层层数进行实验的时，我们同样将其它影响的参数设为固定值，与第一组实验的方法类似。实验的结果如图 4-8 所示。

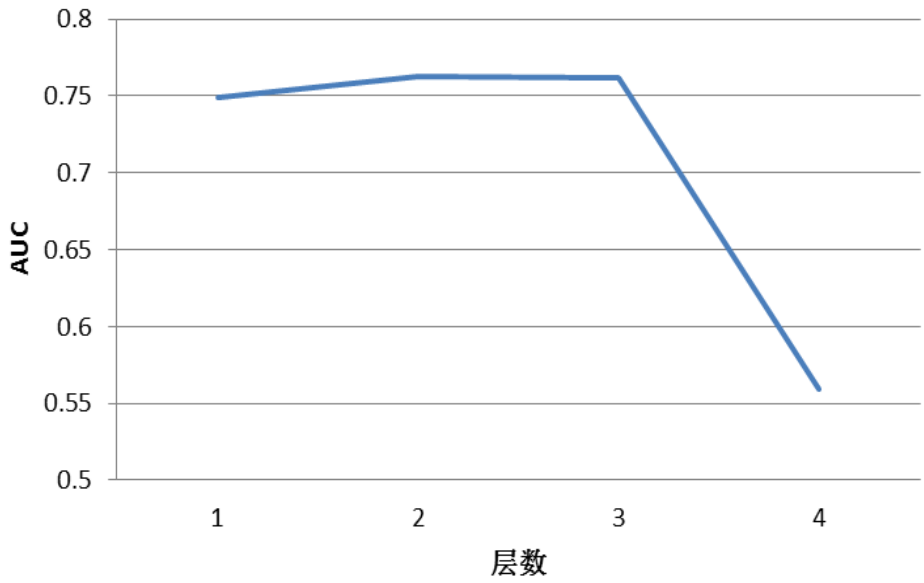

图 4-8 网络中隐藏层层数对 AUC 的影响

由图 4-8 可知，隐藏层层数对 AUC 的影响与卷积核对 AUC 的影响趋势很相像，在一定范围内，呈现先增后减的趋势，当层数为 2 时达到最大值，超过 3 层后，预测效果骤减。我们知道，在图像处理上，卷积层的作用相当于一个特征映射，而亚采样层的可以看作是模糊滤波器，作用是对特征进行二次提取，隐藏层与隐藏层之间的空间分辨率是递减的。在面向搜索广告点击率预测问题的卷积神经网络中，我们输入的特征并不具有图片所包含的内部联系，而是想要利用卷积操作和亚采样操作对特征进行局部学习后挖掘特征内在的联系，隐

藏层和隐藏层之间的特征传递其实是一个泛化的过程。因此当隐藏层过多时，会使得类别特征损失太多，而不利于预测。

（3）卷积神经网络中稀疏连接的基础是局部感受野概念的提出，也就是卷积计算中的窗口，对窗口的大小的设置不但影响着特征的分块，还影响着计算得到的后续节点数，因此，我们对卷积窗口大小的设置进行了实验，实验结果如图 4-9 所示。

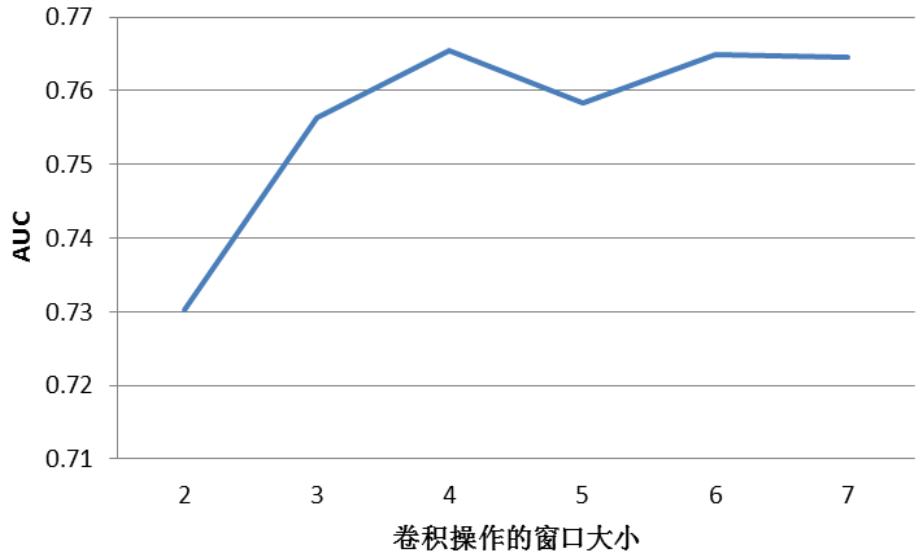

图 4-9 AUC 随卷积窗口大小变化的趋势图

卷积窗口的大小通过对输入划分的影响来影响点击率预测的效果，由图 4-9 可知，卷积窗口大小为 4 时，实验的结果最佳。我们知道，卷积过程实际上是对局部窗口内特征的学习，窗口大小为输入大小时，卷积层等用于基本的全连接层，学习到的关系太过整体化；窗口大小为 1 时，卷积层等同于所有权值都相同的全连接层。

（4）进一步地，我们探索了每一类特征对搜索广告点击率预测的贡献。在所有特征的情况下，去掉某一类特征来进行预测，实验结果如表 4-1 所示。

表 4-1　各类特征的贡献

特征	AUC
去掉历史点击率特征	0.6428
去掉相似度特征	0.7720
去掉位置特征	0.7883
去掉高影响力特征	0.7714

实验结果表明，去掉任意一类特征都将使得实验的预测效果有所下降。其中，去掉历史点击率特征是，广告点击率预测的效果下降的最明显，这说明广

告是否会被点击，与它之前的历史点击行为非常相关，历史点击率在一定程度上也刻画了一个广告被点击的概率。而去掉位置特征时，效果下降的最为不明显，分析原因是因为在实验使用的数据集中，每个页面最多仅呈现三个广告，当页面中的广告数的总数少时，位置对用户点击的影响小。此外，去掉相似度特征和去掉高影响力特征时，实验的结果变化不大。

我们还发现，无论是在深度神经网络中还是在卷积神经网络中，直接使用词向量特征不但不会提高实验的效果，反而会起到副作用，原因是词向量特征在提取上就已经经过了神经网络的处理，并且是以一个向量来表示一个词的向量特征，而其他的特征是标量形式的原始特征，将他们组合在一起，在学习中，并不能很好的融合。

（5）最后，我们在前面参数选择的实验下，设置卷积核个数为 ，层数为 ，并进行实验。在特征相同的情况下，使用逻辑斯蒂回归模型、支持向量回归模型和深度神经网络作为对比方法，与卷积神经网络的预测效果进行对比。结果如表 4-2 所示。

表 4-2　卷积神经网络对比实验结果展示

方法	AUC
逻辑斯蒂回归模型	0.7412
支持向量回归模型	0.7350
深度神经网络模型+dropout	0.7732
卷积神经网络模型	**0.7925**
KDD Cup 2012 track 2	0. 8069

卷积神经网络与各对比实验的实验结果如表 4-2 所示，可以看出，在使用相同特征的情况下，卷积神经网络的效果最佳，再一次验证了在特征学习的过程中，深度学习模型相比浅层学习模型能更好挖掘特征内部的关系。其次，卷积神经网络模型的效果要优于深度神经网络模型的实验效果，说明卷积层的特征融合和亚采样层的特征萃取过程是有效的。在表中我们可以看到，本文中 CNN 目前的实验结果略低于 KDD Cup 2012 track 2 中第一名的结果，分析原因，一方面是因为比赛队伍使用了大规模的特征，然而在我们的实验中，由于硬件的限制，我们未能将更高维度的特征放入深度学习的模型中，另一方面，比赛队伍使用了多模型融合技术，而这里，我们只使用了单个的模型分开进行实验。

4.5 本章小结

本章主要研究了基于卷积神经网络的搜索广告点击率的预测，从卷积神经网络模型的特点和结构介绍到面向搜索广告点击率预测的网络设置，以及对模

型参数进行了研究。

具体来说，我们首先从基本的卷积神经网络入手，介绍了卷积神经网络稀疏连接和权值共享的特点，并给出了构成网络结构的基本层次。然后我们针对搜索广告点击率的预测，设置了卷积神经网络的结果并阐述了工作的过程。再次，对影响预测效果较为明显的网络参数，我们通过实验进行了选择和调节，包括卷积核的个数、网络的层数、卷积窗口的大小等，并通过实验，对第二章中提取的特征在卷积神经网络中的贡献进行了分析。最后，我们通过在相同特征下的对比实验，验证了卷积神经网络在搜索广告点击率的预测上是有作用的，说明了卷积层和亚采样层对特征的学习是有效的。

结 论

搜索广告是互联网在线广告中占比最大的广告形式，因其规模大，增长快，已经成为了互联网公司的主要收入来源之一。广告的点击率影响着广告商的出价和广告媒介投放广告的位置排名。准确高效的预测搜索广告的点击率不仅能增加广告媒介的收益，还能提高用户对搜索结果的满意程度。因此，本文对如何有效的提高搜索广告点击率的预测结果的质量进行了研究，取得了如下三方面的研究成果：

（1）本文在对所用数据集进行预处理的基础上，详细的描述了本文后续实验中所使用到的特征的提取过程和方法，根据搜索广告的特征，我们在类别稀疏特征、历史点击率特征、相似度特征和位置特征的基础上增加了词向量特征，解决了因 One-hot 编码造成的维度大、单词之间孤立的缺点。此外，针对深度学习模型复杂、参数多而无法训练高维度特征的问题，我们使用了基于 L1 范数正则化的特征选择策略，可以在计算能力受限的情况下使用深度学习模型来解决广告点击率的预测问题，并取得较好的结果。

（2）针对浅层学习模型学习能力表达有限，特征关系挖掘不充分的不足，本文首先使用了基于深度神经网络模型的点击率预测方法，并使用了 dropout 方法对模型可能产生过拟合现象进行了改进。深度神经网络模型虽然学习到了特征每个维度的信息，但是所有节点是统一对待作为一个整体进行学习，因此，我们提出了基于卷积神经网络模型的点击率预测方法，通过不断的在局部特征上进行卷积和亚采样操作，完成特征的学习。实验结果表明，深度学习模型在广告点击率预测的问题上并浅层学习模型的结果更好，此外，基于卷积神经网络模型的广告点击率预测的结果要高于深度神经网络模型的结果，说明通过局部窗口进行卷积和亚采样操作是有效的。

（3）针对深度学习模型复杂导致训练时间缓慢的问题，我们设计了基于 GPU 计算的实验方法，GPU 集成众多计算单元的优点使得在大数据背景下训练深度学习模型得以完成。此外，搜索广告历史数据量非常庞大，对显存的需求较高，我们在有限显存的情况下，使用了分块计算的思路，将数据分块进行训练和测试，实验证明，在误差可接受的范围内，基于 GPU 的分块计算方法能取得一定的实验效果。

在搜索广告点击率预测的问题中，尽管本文进行了一系列的研究，并取得了一定的成功，但本文仍然存在着一些不足以及需要改进之处。

（1）本文只针对在线广告中占比最大的搜索广告进行了研究，尤其是在特

征提取方面，受在线广告形式的约束较大，想要应用到其他在线广告的点击率预测中，需要根据具体的广告形式进行改进。

（2）在主流的浅层学习模型中，词向量特征的使用提高了实验的效果，而在深度神经网络模型和卷积神经网络模型中，直接使用词向量做特征却起到了负效果，如何在面向广告点击率的深度学习模型中有效的使用词向量特征，还需要继续探索。

（3）本文仅对基于深度神经网络模型和基于卷积神经网络模型的搜索广告的点击率进行了预测，但在深度学习领域中，模型各式各样，如何确定各模型在解决广告点击率预测问题中的有效性并得到更准确的预测结果，仍值得进一步深入的研究。

参考文献

[1] Yang S, Ghose A. Analyzing the relationship between organic and sponsored search advertising: Positive, negative, or zero interdependence?[J]. Marketing Science, 2010, 29(4): 602-623.

[2] Chapelle O, Zhang Y. A dynamic bayesian network click model for web search ranking[C]// Proceedings of the 18th international conference on World wide web. ACM, 2009: 1-10.

[3] Winkler R L. An introduction to Bayesian inference and decision[M]. Holt, Rinehart and Winston New York, 1972.

[4] Dupret G E, Piwowarski B. A user browsing model to predict search engine click data from past observations[C]// Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2008: 331-338.

[5] Guo F, Liu C, Kannan A, et al. Click chain model in web search[C]// Proceedings of the 18th international conference on World wide web. ACM, 2009: 11-20.

[6] Chakrabarti D, Agarwal D, Josifovski V. Contextual advertising by combining relevance with click feedback[C]// Proceedings of the 17th international conference on World Wide Web. ACM, 2008: 417-426.

[7] Wu K-W, Ferng C-S, Ho C-H, et al. A two-stage ensemble of diverse models for advertisement ranking in KDD Cup 2012[J]. KDDCup, 2012,

[8] Dave K S, Varma V. Learning the click-through rate for rare/new ads from similar ads[C]// Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2010: 897-898.

[9] Zhang Y, Jansen B J, Spink A. Identification of factors predicting clickthrough in Web searching using neural network analysis[J]. Journal of the American Society for Information Science and Technology, 2009, 60(3): 557-570.

[10] Drachsler H, Hummel H G, Koper R. Personal recommender systems for learners in lifelong learning networks: the requirements, techniques and model[J]. International Journal of Learning Technology, 2008, 3(4): 404-423.

[11]霍晓骏，贺樑，杨燕．一种无位置偏见的广告协同推荐算法[J]．计算机工程，2014, 40(12): 39-44.

[12]Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, (8): 30-37.

[13]Brand M. Fast Online SVD Revisions for Lightweight Recommender Systems[C]// SDM. SIAM, 2003: 37-46.

[14]Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions[J]. Knowledge and Data Engineering, IEEE Transactions on, 2005, 17(6): 734-749.

[15]Rendle S. Factorization machines[C]// Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010: 995-1000.

[16]Kanagal B, Ahmed A, Pandey S, et al. Focused matrix factorization for audience selection in display advertising[C]// Data Engineering (ICDE), 2013 IEEE 29th International Conference on. IEEE, 2013: 386-397.

[17]Shan L, Lin L, Shao D, et al. CTR Prediction for DSP with Improved Cube Factorization Model from Historical Bidding Log[C]// Neural Information Processing. Springer, 2014: 17-24.

[18]Regelson M, Fain D. Predicting click-through rate using keyword clusters[C]// Proceedings of the Second Workshop on Sponsored Search Auctions. 2006,9623.

[19]He X, Pan J, Jin O, et al. Practical lessons from predicting clicks on ads at facebook[C]// Proceedings of 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, 2014: 1-9.

[20]Arel I, Rose D C, Karnowski T P. Deep machine learning-a new frontier in artificial intelligence research [research frontier][J]. Computational Intelligence Magazine, IEEE, 2010, 5(4): 13-18.

[21]孙志军，薛磊，许阳明, et al. 深度学习研究综述[J]．计算机应用研究, 2012, 29(8): 2806-2810.

[22]Pujol P, Pol S, Nadeu C, et al. Comparison and combination of features in a hybrid HMM/MLP and a HMM/GMM speech recognition system[J]. Speech and Audio Processing, IEEE Transactions on, 2005, 13(1): 14-22.

[23]息晓静，林坤辉，周昌乐，et al. 语音识别关键技术研究[J]．计算机工程与

应用, 2006, 42(11): 66-69.

[24]侯风雷, 张昆帆. 基于正交高斯混合模型的说话人识别研究[J]. 信息工程大学学报, 2002, 3(2): 43-45.

[25]Deng L, Li J, Huang J-T, et al. Recent advances in deep learning for speech research at Microsoft[C]// Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 8604-8608.

[26]高曙明. 自动特征识别技术综述[J]. 计算机学报, 1998, 21(3): 281-288.

[27]张翠平, 苏光大. 人脸识别技术综述[J]. 中国图象图形学报: A 辑, 2000, (11): 885-894.

[28]Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[29]白栓虎, 夏莹, 黄昌宁. 汉语语料库词性标注方法研究[J]. 机器翻译研究进展, P408-418, 1992,

[30]Zhao P, Yu B. On model selection consistency of Lasso[J]. The Journal of Machine Learning Research, 2006, 72541-2563.

[31]Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso[J]. Biostatistics, 2008, 9(3): 432-441.

[32]Brouhns N, Denuit M, Vermunt J K. A Poisson log-bilinear regression approach to the construction of projected lifetables[J]. Insurance: Mathematics and Economics, 2002, 31(3): 373-393.

[33]Owens J D, Houston M, Luebke D, et al. GPU computing[J]. Proceedings of the IEEE, 2008, 96(5): 879-899.

[34]钟联波. GPU 与 CPU 的比较分析[J]. 技术与市场, 2009, (9): 13-14.

[35]Fawcett T. ROC graphs: Notes and practical considerations for researchers[J]. Machine learning, 2004, 311-38.

[36]杨波, 秦锋, 程泽凯. 一种新的分类学习系统评估度量[J]. 2005 年 “数字安徽” 博士科技论坛论文集, 2005,

[37]王珏, 石纯一. 机器学习研究[J]. 广西师范大学学报: 自然科学版, 2003, 21(2): 1-15.

[38]张学工. 关于统计学习理论与支持向量机[J].自动化学报,2000, 26(1): 32-42.

[39]Rish I. An empirical study of the naive Bayes classifier[C]// IJCAI 2001 workshop on empirical methods in artificial intelligence. IBM New York,

2001,3: 41-46.

[40]李航. 统计学习方法 [M]. 北京: 清华大学出版社. 2012.

[41]Smola A J, Schölkopf B. A tutorial on support vector regression[J]. Statistics and computing, 2004, 14(3): 199-222.

[42]Burges C J. A tutorial on support vector machines for pattern recognition[J]. Data mining and knowledge discovery, 1998, 2(2): 121-167.

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

[44]Dahl G E, Sainath T N, Hinton G E. Improving deep neural networks for LVCSR using rectified linear units and dropout[C]// Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 8609-8613.

[45]Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]// Advances in neural information processing systems. 2012: 1097-1105.

[46]Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[47]Mitchell T M. 机器学习[M]. 机械工业出版社, 2003.

[48]Ma L, Lu Z, Shang L, et al. Multimodal Convolutional Neural Networks for Matching Image and Sentence[J]. arXiv preprint arXiv:150406063, 2015,

[49]Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks[C]// Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, 2014: 1725-1732.

[50]Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning[C]// Proceedings of the 25th international conference on Machine learning. ACM, 2008: 160-167.

[51]Socher R, Bauer J, Manning C D, et al. Parsing with compositional vector grammars[C]// In Proceedings of the ACL conference. Citeseer, 2013.

[52]Larochelle H, Bengio Y, Louradour J, et al. Exploring strategies for training deep neural networks[J]. The Journal of Machine Learning Research, 2009, 101-40.

攻读硕士学位期间发表的学术论文

[1] 李思琴，林磊，孙承杰.《基于卷积神经网络的搜索广告点击率预测》[J]. 智能计算机与应用，哈尔滨：哈尔滨工业大学出版社。（已录用）

哈尔滨工业大学学位论文原创性声明和使用权限

学位论文原创性声明

本人郑重声明：此处所提交的学位论文《基于深度学习的搜索广告点击率预测方法研究》，是本人在导师指导下，在哈尔滨工业大学攻读学位期间独立进行研究工作所取得的成果，且学位论文中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本学位论文的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：李思琴　　日期：2015 年 6 月 30 日

学位论文使用权限

学位论文是研究生在哈尔滨工业大学攻读学位期间完成的成果，知识产权归属哈尔滨工业大学。学位论文的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文，并向国家图书馆报送学位论文；（2）学校可以将学位论文部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）研究生毕业后发表与此学位论文研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉学位论文的使用权限，并将遵守有关规定。

作者签名：李思琴　　日期：2015 年 6 月 30 日

导师签名：林磊　　日期：2015 年 6 月 30 日

致　谢

时光飞逝，岁月无痕，两年的研究生生活转瞬即逝。回首两年的求学时光，心里满是不舍和感恩之情。舍不得哈工大，舍不得它的接受和包容；舍不得实验室，舍不得这日日夜夜奋斗的地方；舍不得老师同学，舍不得他们在这两年里所给予我的帮助和关心。在这离别之际，请允许我向所有陪伴我、关心我、指导我、帮助我的人们致以最真心的感谢。

感谢我的导师林磊老师。从研一入学的面试到如今每日的交流，从每周的周会到毕设的指导，林老师严谨的治学态度和敏锐的学术洞察力深深的影响着我。清晰的记得有一次我因为毕业设计遇到困难心情低落，是林老师跟我说：一个人的努力往往不一定都会被别人看到，但是你坚持了，你付出了，你学到的东西才是你自己的。也是他告诉我，在遇到困难的时候要学习自己走出来，要把目光放长远。两年里，林老师更像是一位父亲在关心着我们的生活和日常，让离家千里求学的我倍感温暖。

感谢孙承杰老师，尽管不是我的责任导师，但是孙老师从来不吝啬对我们的指导和帮助，在研二这一年的时间中，从遇到困难的解决方案到毕业设计每阶段的改进，处处都有孙老师细心的指导，平日里羞于感情的表达，但心里真的很感谢他。感谢单丽莉老师，虽然是师生关系，但跟单老师的相处更像是朋友，她对学术的认真和乐观的心态无时无刻不影响着我。

感谢王晓龙老师，刘秉权老师，刘远超老师，李飞秘书，感谢在这两年的学习和科研过程中给我的帮助。因为他们的付出才有实验室的今天，才有让我们如此受益的学术氛围。

感谢靳晓强，同学五年期间，尤其是研究生阶段，给我莫大的帮助和关心。感谢孙雅铭师姐，当我遇到问题和困难时，带给我快乐，教会我乐观。感谢徐振师兄，从本科到硕士一直以来的帮助，让我少走了很多弯路。

感谢师兄刘峰、王鑫平日里的关心，感谢研二的小伙伴们在我找工作期间和毕设过程中给我的鼓励和照顾，感谢师弟师妹们给我带来的快乐和感动，感谢室友在这两年里对我的包容和关爱。

感谢我的父母和哥哥，因为有他们最无私的付出和支持，才有今天的我，他们永远是我奋斗的动力，是我最值得骄傲和最爱的人。

感谢每一个陪我走过的人。爱你们。