北京大学

博士研究生综合考试报告

题目：云计算环境下的人工智能相关技术研究

姓　　名：李炎

学　　号：2001111305

院　　系：信息科学技术学院

专　　业：计算机软件与理论

研究方向：云计算与普适计算

导　　师：梅宏

二〇二一年四月

版权声明

摘要

21 世纪 10 年代以来, 云计算和人工智能可谓计算机科学领域最为炙手可热的两个研究方向。以虚拟化、资源管理和服务化为代表的云计算核心技术在近十余年里取得了丰硕的研究成果。当前，云计算已成为工业化社会重要的信息基础设施，支撑并推动着大数据和人工智能产业的快速发展。与此同时，人工智能技术在近十年内也相继在计算机视觉、自然语言处理等多个领域取得了突破，“智能化”已然成为现代社会的重要标签之一。

本文将以上述两大技术的蓬勃发展为背景，研究云计算与人工智能相互影响、相互支持、相辅相成的相关技术。本文将按照如下几章展开。

第一章对相关的技术背景做出介绍。首先对云计算近十年的发展做简要回顾，并介绍具有代表性的若干核心技术。其次对人工智能近十年的发展做简要概述，阐述其在计算机视觉、自然语言处理等子领域的代表性科研成果。最后分析二者在交叉领域现有的相关研究，即利用人工智能技术解决云计算中的资源配置与调度问题，以及云计算环境中在软件和硬件层面对人工智能应用的支撑情况。

第二章开始探究二者的关系。本章从“服务于云计算系统架构的 AI 技术”这一视角展开，研究人工智能对云计算的增强技术。云计算本质上是一个巨大的公共资源池，用户如何在资源池中选取资源，云厂商如何为不同的用户调度资源，是云计算领域两个重要的话题。近五年来，该领域的研究者开始尝试利用一系列基于机器学习的算法来辅助解决上述两个问题，本章将重点讨论与上述算法相关的研究。

第三章从另一角度，研究云计算环境中用以支持人工智能的系统软件技术。随着人工智能算法和系统研究的发展，其训练-测试-部署的流程愈发复杂。很多云厂商基于本地化的云原生技术，构建了一站式的人工智能开发-部署软件栈供用户使用。同时，伴随着新的云计算模式（如无服务计算 serverless）的产生，工业界和学术界也在探究将其应用在人工智能领域，使相关的应用在云上具有更高的弹性。本章将重点讨论与上述话题相关的系统软件研究工作。

第四章讨论近些年来云环境下出现的新硬件，如 GPU，AI 专用芯片，Intel-SGX 等，为人工智能技术带来的新的机遇与挑战。硬件的发展（如 GPU，AI 专用芯片等）导致的算力的提升，也是人工智能近年发展迅速的重要原因之一。同时，某些硬件层面安全机制（如 Intel-SGX）的产生，使增强人工智能应用的安全性有的新的潜在解决方案。本章将关注近十年来的新硬件为人工智能带来的新的机遇与挑战。

第五章总结了上述三个方向的重要文献和相关研究团队概况。

第六章介绍了作者下一步的研究计划。

关键词：云计算，人工智能，机器学习，新硬件

目录

第一章　引言

1.1　云计算的基本概念

云计算（Cloud Computing），根据美国国家标准技术研究所（NIST）的定义，指的是一种可以实现对可配置计算资源共享池（如网络、服务器、存储、应用和服务）进行随时随地、便捷、按需网络访问模型。这些资源可以迅速地配分配和释放，并且这个过程只需要足最低限度的资源管理工作以及与服务提供商最少的交互。美国亚马逊公司再 2006 年 3 月推出了 Amanzon Web Service（AWS），这一事件一般被认为代表着云计算时代的正式开启。经过十几年的发展，凭借着“方便易用、弹性伸缩、按需服务”的技术特征，云计算概念已被广泛接受，云计算产业取得了商业上的巨大成功，云计算平台已成为当今社会的关键信息基础设施，云计算技术为大数据、人工智能的领域的蓬勃发展特工了重要的支撑作用。

1.1.1　云计算的传统服务模型

NIST 将云计算分为了三种服务模型。

这三种服务模型分别是基础设施即服务（Infrastructure as a Service，IaaS）、平台即服务（Platform as a Service，PaaS）以及软件即服务（Software as a Service，SaaS）。IaaS 为消费者提供用来运行应用的计算资源，包括服务器、存储、网络等。其中虚拟机是云厂商提供的最核心的 IaaS 产品。与 IaaS 只提供最基础的底层资源不同，PaaS 强调为消费者提供云开发环境，除计算资源意外，PaaS 为用户提供中间件开发，运行平台及工具，帮助用户更方便地开、管理、测试和运行应用。SaaS 是厂商提供的基于云的软件，用户无需下载安装软件，通过浏览器即可访问服务。

图1.1给出了云计算三种服务模型的代表产品。亚马逊公司的 AWS EC2，谷歌公司的 Google Compute Engine 以及阿里云公司的 ECS 都是典型的 IaaS 产品。其主要服务形态是云厂商向消费者售卖虚拟机或者裸金属服务器以及连带的网络、存储等附属产品。PaaS 的代表性产品包括 AWS Beanstalk、Google App Engine、Microsoft Azure App Servics 等。此类韩品为用户提供再云中快速部署和管理应用的能力，提供包括应用扩容，负载均衡，应用监控和安全管理等功能。相比于 IaaS 仅售卖以虚拟机为主的基础设施，PaaS 降低了用户开发、管理、运维应用的成本，使得用户可以更加专注于构建应用本身。在云计算已经发展了十几年的当今时代，越来越多的 SaaS 产品涌现了出来。谷歌公司开发的 Google Docs、Google Maps 以及微软公司开发的 Microsoft Office 365

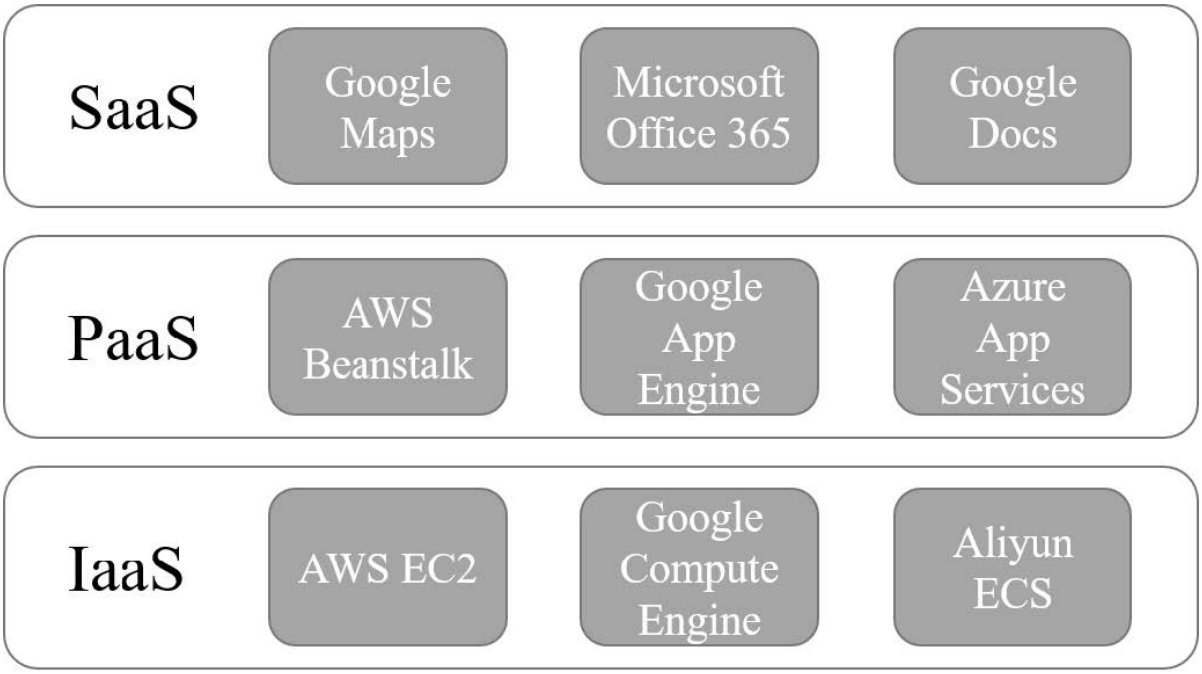

图 1.1　云计算服务模型代表产品

都是典型的 SaaS 产品。以 Google Docs 为例，与传统的文件处理办公软降相比，用户无需在本地花费大量存储空间来安装软件，只需要打开浏览器，输入 URL，即可使用 Google Docs 服务处理文件，并且所有的文件都会被及时同步保存到云端。另一个代表性的 SaaS 产品是 Google Maps。Google Maps 与 Google Docs 类似，为个人用户提供在浏览器中直接使用的地图服务。与传统软件相比，SaaS 在使用方式上具有方便灵活，跨平台的特性，同时用户存储在云端的数据经过云厂商的冗余备份也具有更高的可靠性。

1.1.2　云计算的新兴服务模型

云计算发展至今日，其服务模型已经不严格局限于 NIST 最初总结的这三种基本形态，世界各地各领域的研究者们已经提出了众多不同的 X as a Service，包括 Blockchain as a Service，Sensing as a Service，Workspace as a Service 等。与传统的三种服务形态相比，这些服务不单纯是硬件服务或者软件服务，其结合二者的特点，面向特定的领域方向进行更深度的定制，如区块链、物联网、分布式共识等。服务形态的日益丰富，服务内容的日益复杂体现了云计算更见领域化，精细化的发展趋势。而近几年来最热门的概念莫过于 FaaS，即 Fucntion as a Service。

FaaS 是一种新兴的计算模式，亦被称为 Serverless Computing。从字面理解，Serverless Computing 即为“无服务器计算”之意。然后，其并非意味着真的没有服务器，而已说开发者不用过多考虑服务器的相关问题。在传统的 IaaS 服务中，开发者需要自己

图 1.2　AWS Lambda 中示例程序：照片大小调整

惊醒服务器管理与运维，负责服务的发布，在流量变化时对服务器集群进行扩容或缩容。而在 FaaS 中，开发者只需要关注业务逻辑，至于服务的发布、管理、弹性伸缩等，则交由云厂商来完成。

FaaS 背后的机制一般是以容器技术为基础的。典型地，开发者上传自己的业务代码后，云厂商并不会直接收费。当对该服务的请求到来之时，云厂商将启动一系列容器来运行该服务，从而对用户的请求进行响应。通常而言，开发者指定的服务会与某些时间绑定（hook），在发生该事件时，立即触发开发者定义的服务。我们以 AWS 的 serverless computing 服务 Lambda 中的一个示例程序为例，讲述整个流程。图1.2所示的是一个为照片调整大小的服务。该服务与 AWS S3（AWS 的对象存储服务）的上传事件绑定，当云厂商检测到有用户向 S3 上传图片时，会立即触发开发者定义的图片大小调整函数。整个流程中，开发者需要关注的只有第四步的函数开发工作，至于该函数的横向拓展，全部由云厂商来负责。

主流的云厂商均提供了 FaaS 服务，例如 AWS 的 Lambda，阿里云的函数计算等，近年来越来越受到开发者的青睐。一方面是因为它的高弹性，易于开发。另一方面则是因为其细粒度的收费模式。通常而言，FaaS 的服务是按照请求次数进行收费。当函数闲置时，并不产生额外的费用。

1.2　人工智能技术近年的发展

TODO: 简介近些年人工智能技术的发展

1.3　云计算和人工智能交叉领域的常见研究问题

云计算和人工智能是两个息息相关的热门领域。一方面，以深度学习为典型代表的人工智能技术在当今社会被应用的越来月广泛，研发、调试、发布新的模型的需求日益增长。与这种发展趋势对应，多数主流的云厂商都提供了机器学习模型训练-测试-部

署的 pipeline。学术界也不断探索“云上机器学习”这一话题，利用新的计算模式（如 FaaS）在云上以更便捷、更经济高效地开展 ML 模型的训练和部署。同时得益于近年硬件技术的发展，多种新硬件（多体现为人工智能的加速芯片）在云环境中得到应用，进一步方便机器学习用户将整个开发流程迁移到云端。另一方面，机器学习技术也越来越多被用于解决云计算中常见的问题。例如使用推荐算法解决置优化问题和利用强化学习解云环境中的资源调度问题。下文对这几个常见问题做简单概述，具体研究将在后续几章展开。

1.3.1 基于公有云服务的机器学习平台

1. 产业界

主流的云厂商都提供了面向机器学习的平台系统，例如 AWS 和 Sagamaker[4, 6, 7]，Azure 的 Azure ML Studio[2] 等。这些基于云的平台提供了一站式调试、训练和部署 ML 模型的能力。一般而言此类平台被视为 SaaS 类的服务，因为其是基于云资源构建的上层软件栈，使得用户能够直接使用 web 的方式使用。例如 Sagemaker 就支持用户直接在浏览器中用 jupyter notebook 编写和调试模型代码。

2. 学术界

一般而言，产业界的平台系统面向的是一般性用户的普遍需求。因此，对于有特殊需求的用户，通常会有与产业界的解决方案并行的工作。例如，为了以尽可能低的成本在云上完成模型的训练，相关工作 [3, 5] 尝试利用云上的动态资源（价格低但是稳定性/可用性低）进行模型的训练，并辅以一定的策略增强其可靠性。再比如，机器学习模型的在线服务会有低延迟、高吞吐率的要求。为了实现上述需求，相关工作 [8] 利用云上的多种资源（如 IaaS，FaaS 等），根据负载的动态变化，敏捷地在不同资源之间切换，充分利用不同类型资源的优点，规避掉其缺点，实现高效、经济的模型在线服务。一般而言，学术界的此类研究构建于云厂商服务的上层，是一种 Cloud-of-Clouds 的模式。

1.3.2 利用新的计算模式在云上运行机器学习 pipeline

以 FaaS 为代表的新型云计算模式，以其高弹性、灵活的计费方式等特点吸引了众多研究者的注意。学术界开始探讨如何使得机器学习工作流享用到 FaaS 的诸多优点。

1.3.3 面向人工智能的专用芯片

1.3.4 利用机器学习算法解决配置优化问题

1.3.5 利用机器学习算法解决资源调度问题

第二章　章节

pkuthss 文档模版最常见问题：

`\cite`、`\parencite` 和 `\supercite` 三个命令分别产生未格式化的、带方括号的和上标且带方括号的引用标记：1，[9]、[1,9]。

若要避免章末空白页，请在调用 pkuthss 文档类时加入 openany 选项。

如果编译时不出参考文献，请参考 `texdoc pkuthss`“问题及其解决”一章“上游宏包可能引起的问题”一节中关于 biber 的说明。

因无法假定用户使用哪种方式排版表格，用户须自行保证表格字号符合学校规定。

第三章　结论和展望

pkuthss 文档模版最常见问题：

`\cite`、`\parencite` 和 `\supercite` 三个命令分别产生未格式化的、带方括号的和上标且带方括号的引用标记：1，[9]、[1,9]。

若要避免章末空白页，请在调用 pkuthss 文档类时加入 openany 选项。

如果编译时不出参考文献，请参考 `texdoc pkuthss`“问题及其解决”一章“上游宏包可能引起的问题”一节中关于 biber 的说明。

因无法假定用户使用哪种方式排版表格，用户须自行保证表格字号符合学校规定。

参考文献

[1] Author. “*Title*” [J]. *Journal*, 2014-04-01.

[2] Leila Etaati. “*Azure Machine Learning Studio*”. **2019**: 201–223.

[3] Aaron Harlap, Alexey Tumanov, Andrew Chung *et al.* “*Proteus: agile ML elasticity through tiered reliability in dynamic resource markets*”. In: *Proceedings of the Twelfth European Conference on Computer Systems*. **2017**: 589–604.

[4] Ameet V Joshi. “*Amazon’s Machine Learning Toolkit: Sagemaker*”. **2020**: 233–243.

[5] Yan Li, Bo An, Junming Ma *et al.* “*SpotTune: Leveraging Transient Resources for Cost-efficient Hyper-parameter Tuning in the Public Cloud*”. In: *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*. **2020**.

[6] Edo Liberty, Zohar Karnin, Bing Xiang *et al.* “*Elastic Machine Learning Algorithms in Amazon SageMaker*”. In: *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. **2020**: 731–737.

[7] Valerio Perrone, Huibin Shen, Aida Zolic *et al.* “*Amazon SageMaker Automatic Model Tuning: Scalable Black-box Optimization.*” *arXiv preprint arXiv:2012.08489*, **2020**.

[8] Chengliang Zhang, Minchen Yu, Wei Wang *et al.* “*MArk: Exploiting Cloud Services for Cost-Effective, SLO-Aware Machine Learning Inference Serving.*” In: *2019 USENIX Annual Technical Conference (USENIX ATC 19)*. **2019**: 1049–1062.

[9] 作者。“标题”[J]。期刊，2014-04-01。

附录 A　附件

pkuthss 文档模版最常见问题：

`\cite`、`\parencite` 和 `\supercite` 三个命令分别产生未格式化的、带方括号的和上标且带方括号的引用标记：1，[9]、[1,9]。

若要避免章末空白页，请在调用 pkuthss 文档类时加入 `openany` 选项。

如果编译时不出参考文献，请参考 `texdoc pkuthss`“问题及其解决”一章“上游宏包可能引起的问题”一节中关于 biber 的说明。

因无法假定用户使用哪种方式排版表格，用户须自行保证表格字号符合学校规定。

致谢

pkuthss 文档模版最常见问题：

\cite、\parencite 和 \supercite 三个命令分别产生未格式化的、带方括号的和上标且带方括号的引用标记：1，[9]、[1,9]。

若要避免章末空白页，请在调用 pkuthss 文档类时加入 openany 选项。

如果编译时不出参考文献，请参考 texdoc pkuthss“问题及其解决”一章“上游宏包可能引起的问题”一节中关于 biber 的说明。

因无法假定用户使用哪种方式排版表格，用户须自行保证表格字号符合学校规定。

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：　　日期：　　年　　月　　日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因须要延迟发布学位论文电子版，授权学校在 □ 一年 / □ 两年 / □ 三年以后在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名：　　导师签名：　　日期：　　年　　月　　日