《深度学习系统》课程大纲

主责单位：计算机学院

课程编号：06113226

课程中文名称：深度学习系统

课程英文名称：Deep Learning Systems

开课学期：秋季

学分/学时：2 学分，32 学时

课程性质：专业学位核心课

先修课程：计算机系统基础、机器学习等

建议后续课程：计算机体系结构、高性能计算等

适用层次：硕博通用

适用专业：计算机、人工智能、软件工程、电子信息等

课程负责人：杨建磊

修订时间：2024 年 8 月 1 日

一、课程的性质、目的和任务

本课程是面向计算机、人工智能、电子信息等相关专业研究生开设的专业课程，为《计算机系统》与《机器学习》等课程的后续延伸，尝试将机器学习（尤其是深度学习）系统领域的最新研究前沿技术引入课堂。本课程涉及到的理论基础较为广泛，包括智能计算领域（人工智能、机器学习、数据挖掘、深度学习、计算机视觉等）与计算机系统领域（计算机组成、操作系统、编译原理、计算机体系结构、高性能计算等）的基本知识。同时，本课程亦需要较为扎实的系统工程技术实践能力（算法实现、系统设计、系统调试与优化等），因此是一门注重系统能力培养且重于研究探索的实践课程。

本课程以深度学习领域几个典型应用为研究对象，讲述深度学习系统从算法模型到硬件架构实现所涉及的算法框架、编程模型、指令集、编译器与软件栈等基本原理与系统优化方法，使学生理解从系统的角度实现深度学习的整个工作过程，培养学生在面向深度学习计算架构基础上设计和使用相关编译器和软件栈以及系统优化方法的基本技能。本课程要求学生理解目前深度学习系统设计的基本思路，采用现有的深度学习计算架构、编译器与软件栈对主流的深度学习算法进行计算效率的量化分析、系统优化等，使其掌握深度学习系统设计与优化的工程化方法，并具备软硬件异构协同设计的基本素养和学术视野，为今后更深入的理论学习和研究实践奠定良好基础。

本课程是学院规划和建设的“本研一体前沿课程”之一，并被列为“电子信息（0854）专业

学位类别核心课程”。课程实践性较强，课程教学目标高，知识内容前沿、知识面宽、知识运用综合，实验体系具有很高的难度与强度。本课程适合渴望学习前沿技术与敢于面对挑战的学生，选修本课程的学生应有充分心理准备。具体的教学目标可分解为以下几点：

（1） 了解面向深度学习领域的行业应用、软件栈、开发环境、专用硬件等几个技术层面，建立基本的深度学习系统知识体系；了解深度学习系统的工程问题、实验技术、工具资源以及该领域最新的研究前沿；

（2） 了解几类主流深度学习专用硬件架构的特征与原理，例如寒武纪架构、谷歌 TPU 架构、华为达芬奇架构、MIT Eyeriss 架构等；

（3） 掌握深度学习系统软件栈的基本原理、使用过程和相关研究方法，注重工程思维与创新思维的培养，尤其是面向专用架构将深度学习算法从系统层面进行优化的思维能力，并通过现有软件栈进行系统层面的性能测试、分析及优化等；

（4） 具备利用主流深度学习编程框架等工具设计相关算法并运行在专用平台上的能力，包括相关软件栈开发调试、模型轻量化、网络架构搜索等，以及进行相关性能分析和系统优化等方面的能力；

（5） 掌握深度学习系统设计与性能测试以及系统优化的工程化实验方法，获得实验技能的基本训练；

（6） 掌握深度学习系统模型建立、分析求解和设计方案论证的理论和方法，培养学生分析和解决深度学习系统应用等工程实践问题的创新设计能力；

（7） 具有获取和利用标准、规范、手册、图册等有关技术资料的能力。

课程目标对毕业要求的支撑关系

毕业要求	课程目标对毕业要求的支撑关系
毕业要求 1：工程知识 能够将数学、自然科学、工程基础和专业知识用于解决复杂计算机工程问题。	**课程目标：2、3、4、5**
毕业要求 2：问题分析 能够应用数学、自然科学基本原理，并通过文献研究，识别、表达、分析复杂计算机工程问题，以获得有效结论。	**课程目标：3、4、6**

毕业要求	课程目标对毕业要求的支撑关系
毕业要求 3：设计/开发解决方案 能够设计针对复杂计算机工程问题的解决方案，设计满足特定需求的计算机系统，并能够在设计环节中体现创新意识，考虑法律、健康、安全、文化、社会以及环境等因素。	**课程目标：3、4、6、7**
毕业要求 4：研究 能够基于科学原理并采用科学方法对复杂计算机工程问题进行研究，包括设计实验、分析与解释数据、并通过信息综合得到合理有效的结论。	**课程目标：4、6、7**
毕业要求 5：使用现代工具 能够在计算机工程实践中开发、选择与使用合理有效的技术、资源、现代工程工具和信息技术工具，并了解其局限性。	**课程目标：4、6**
毕业要求 6：工程与社会 具有追求创新的态度和意识，掌握基本的创新方法，以及综合运用理论和技术手段设计复杂计算机系统与过程的能力；设计过程中能够综合考虑社会、经济、文化、环境、法律、安全、健康、伦理等制约因素。	**课程目标：1**
毕业要求 7：环境和可持续发展 了解与本专业相关的职业和行业的生产、设计、研究与开发、环境保护和可持续发展等方面的方针、政策和法津、法规；能够正确认识专业工程实践对环境和社会可持续发展的影响，合理评价专业工程实践和复杂工程问题解决方案对社会、健康、安全、法律及文化的影响。	**课程目标：1**
毕业要求 8：职业规范 具有坚定正确的政治方向，良好的思想品德、社会公德和职业道德；具有人文社会科学素养、社会责任感；具有良好的身体素质和心理素质，能履行建设祖国和保卫祖国的神圣义务。	**课程目标：5、6、7**
毕业要求 9 ：个人和团队 具有在多学科团队中发挥重要作用的能力。	**课程目标：4**

毕业要求	课程目标对毕业要求的支撑关系
毕业要求 10：沟通 能够就复杂计算机工程问题与业界同行及社会公众进行有效沟通与交流，包括撰写报告和设计文稿、陈述发言、清晰表达个人见解等，并具备一定的国际视野，能够在跨文化背景下进行沟通和交流。	**课程目标：4**
毕业要求 11：项目管理 具有一定的组织与工程管理能力、表达与人际交往能力以及在多学科背景下的团队中发挥作用的能力。	**课程目标：4**
毕业要求 12：终身学习 具有自主学习和终身学习的意识，有不断学习和适应发展的能力。	**课程目标：5、6、7**

二、课程内容、基本要求及学时分配

本课程介绍深度学习系统领域的知识和工程实践方法，主要包括系统设计原理、编程模型、计算图、编译器、硬件加速、模型轻量化、模型部署与优化、典型系统应用等。

序号	教学内容	基本要求及重点和难点	学时	教学方式	对应的教学目标	支持毕业要求指标点
1	**深度学习系统概述** 深度学习系统； 硬件层基本架构； 软件栈基本架构； 计算效率优化原理等。	**基本要求：**了解深度学习系统所涉及到的软硬件架构。 **重点：**进行充分的文献调研和阅读，理解软硬件系统协同的原理与重要性。 **难点：**理解新型深度学习系统如何提高计算效率、解决系统性能瓶颈等问题。	2	课堂讲授+课后调研+答疑	1	6，7
2	**主流深度学习算法** 深度学习、机器视觉、大模型等。	**基本要求：**掌握深度学习所涉及主流算法的数学模型和理论分析方法。 **重点：**理解并掌握深度学习领域等相关算法。	4	课堂讲授+课后作业+答疑	1，3	1，2，3，4

序号	教学内容	基本要求及重点和难点	学时	教学方式	对应的教学目标	支持毕业要求指标点
		难点：编程实现及调试，理解如何从系统层面优化算法的效率。				
3	**昇腾 AI 处理器** 芯片架构、编程模型、开发平台等。	**基本要求：**了解昇腾 AI 处理器涉及的技术栈，掌握昇腾 AI 计算系统开发过程。 **重点：**理解全栈系统思维在深度学习系统中的原理与技术视图。 **难点：**编程开发实践，理解如何从全栈技术层面构建深度学习系统。	6	课堂讲授+课后作业+答疑	2，4，5	1，2，3，4
4	**AI 芯片架构** 基于 GPU/FPGA/ASIC 等架构的深度学习专用平台；架构特点与指令集；计算效率分析与优化途径；基本的设计与工程化开发方法等。	**基本要求：**了解目前主流基于 GPU/FPGA/ASIC 等平台的专用架构与指令集，掌握计算效率分析与优化的基本思路。 **重点：**理解如何通过专用的架构设计来提升深度学习算法运行的效率。 **难点：**采用量化分析方法对架构性能与效率进行评估。	4	课堂讲授+课后作业+答疑	2，4，5	1，2，3，4
5	**适配软件栈** 编程模型；专用计算架构编译器；计算图抽象；系统建模与分析工具；数据流映射、调度与优化工具；异构平台	**基本要求：**了解面向专用架构指令集的编译器基本原理和方法流程，并掌握目前主流成熟的软件栈使用方法。 **重点：**理解编译器与调度器的原理以及算子优化技巧。	4	课堂讲授+课后实践+答疑	3，5，6	1，2，3，4

序号	教学内容	基本要求及重点和难点	学时	教学方式	对应的教学目标	支持毕业要求指标点
	算法部署方法等。	**难点：**以深度学习专用架构平台为对象，采用专用编译器与软件栈优化深度学习计算效率，具有一定挑战性。				
6	**模型轻量化** 深度学习模型（尤其是深度神经网络）压缩、稀疏化、量化等轻量化方法。	**基本要求：**掌握基本的模型压缩、量化方法，并在计算资源受限情况下对计算效率和精度进行折衷。 **重点：**掌握模型轻量化方法，在嵌入式移动平台上进行模型部署与优化。 **难点：**模型轻量化处理后在端侧进行实测，考查系统调试与优化等动手能力。	4	课堂讲授+课后作业+答疑	4，5，6，7	1，2，3，4
7	**模型架构搜索** 深度学习模型（尤其是深度神经网络）架构自动搜索方法；考虑硬件计算与软件调度开销的模型搜索方法等。	**基本要求：**了解模型架构搜索（NAS）基本原理，掌握主流的 NAS 方法。 **重点：**面向特定硬件平台，采用 NAS 方法对典型神经网络进行结构搜索与优化。 **难点：**如何量化建模硬件平台的约束是有效采用 NAS 优化网络结构的主要问题。	2	课堂讲授+课后作业+答疑	4，5，6，7	1，2，3，4

序号	教学内容	基本要求及重点和难点	学时	教学方式	对应的教学目标	支持毕业要求指标点
8	**大模型技术** 大模型基本原理；主流大模型结构设计；大模型系统设计与优化等。	**基本要求：**了解大模型基本原理，掌握大模型基础框架。 **重点：**掌握大模型推理加速主流方法。 **难点：**如何针对大模型计算瓶颈，探索大模型加速技术，需要进行实践探索与迭代。	2	课堂讲授+课后实践+答疑	4，5，6，7	1，2，3，4
9	**实践驱动案例** 华为昇腾 AI 芯片全栈平台；树莓派 ARM 计算平台；TinyML 系统等。	**基本要求：**了解目前主流且相对成熟的 AI 专用计算平台与系统，包括软硬件与相关适配环境、工具。 **重点：**掌握一套完整的 AI 专用计算系统，并付诸实践。 **难点：**实践涉及诸多新架构思维、新平台环境等因素，需要不断探索、调试与优化。	4	课堂讲授+课后实践+答疑	3，4，5，6，7	5，6，7，8，9，10，11，12

三、课内外教学环节及基本要求

在教学过程中体现“学生主体、教师主导”的教学思想，提倡启发式、讨论式教学，突出对学生逻辑思维、工程创新及实践能力的培养。在讲授过程中做到由浅入深、由表及里、循序渐进，同时注重举例和类比，并加入该领域最新研究进展，活跃课堂，使课堂讲授生动有趣。

本课程课堂授课 32 学时，周学时为 2 学时，教学方式包括课内教学和课外教学两部分，授课语言为中文。课内教学采用多媒体讲义，包括课堂讲授、课堂讨论、课堂展示等。课外教学包括课下自学、课外作业、随堂测试、系统开发、实验报告等环节。

部分教学内容采用项目式教学模式，注重学生实践动手能力，为学生提供丰富的软硬件实验开发平台及实验指导书，提高学生利用现代工具解决工程实践问题的能力，并将前沿研究带入课堂，给学生提供接触最新研究成果的机会，以及亲自参与最新科技进展的实践活动中来，培养其

浓厚的学术兴趣和优良的科研素养。

四、考核方式及成绩评定

本课程以平时成绩与课程实验相结合的方式进行考核，总分为 100 分。

平时成绩占 40%：包含平时作业与交流讨论等。

平时作业：课堂随机测试，预计 5 至 6 次，共计占 30 分。

文献调研汇报：阅读深度学习系统领域经典论文，分组展示 PPT 并讨论，占 10 分。

课程实验占 60%：共 6 次课程实验，包括算法、系统实现、测试及报告、展示等。

实验 1：简单神经网络（用 C 或者 Python 语言等，手写一个简单的神经网络），占 10 分；

实验 2：深度神经网络（使用深度学习框架 Caffe/TensorFlow/PyTorch 等），占 10 分；

实验 3：昇腾 Atlas200DK 开发环境与模型部署，要求掌握其基本开发流程，占 10 分；

实验 4：昇腾 Atlas200DK 基础实验，包括图像识别、目标检测任务，占 10 分；

实验 5：昇腾 Atlas200DK 进阶实验，包括算子定制、模型压缩任务，占 10 分；

实验 6：深度学习模型轻量化，基于 Intel Distiller 工具，占 10 分。

五、教材和参考资料

参考教材：

[1] **机器学习系统：设计和实现**，麦络、董豪，清华大学出版社，https://openmlsys.github.io，2023 年。

[2] **智能计算系统**，陈云霁，机械工业出版社，2020 年。

[3] **AI System & AI Infra**，https://chenzomi12.github.io/，2024 年。

[4] **Large Language Model Systems**，https://llmsystem.github.io/llmsystem2024spring/，2024 年。

[5] **Deep Learning Systems**: Algorithms, Compilers, and Processors for Large-Scale Production，Andres Rodriguez，https://deeplearningsystems.ai/，2021 年。

[6] **Deep Learning Systems**: Algorithms and Implementation，Tianqi Chen and Zico Kolter (CMU), https://dlsyscourse.org/，2022 年。

相关资料：

[7] 昇腾 AI 处理器：架构与编程，梁晓峣著，清华大学出版社，2019 年。

[8] 当计算机体系结构遇到深度学习，杨海龙、王锐译，机械工业出版社，2019 年。

[9] AI For Systems and Systems For AI，UCB CS294 课程，2019 年。

[10] System for ML，华盛顿大学 CSE 599W 课程，2019 年。

[11] TinyML：基于 TensorFlow Lite 在 Arduino 和超低功耗微控制器上部署机器学习，魏兰，卜杰，王铁震（翻译），机械工业出版社，2020 年。

[12] 计算机体系结构基础，胡伟武，https://foxsen.github.io/archbase/，2022 年。

[13] Computer Architecture: A Quantitative Approach，John L. Hennessy 著，Morgan Kaufmann Publishers，2011 年。

[14] 深入理解计算机系统，Randal E. Bryant 和 David O'Hallaron 著，机械工业出版社，2011 年。

[15] 机器学习，周志华著，清华大学出版社，2016 年。

[16] Learning Deep Architectures for AI，Yoshua Bengio 著，Now Publishers Inc.，2009 年。

[17] 统计学习方法，李航著，清华大学出版社，2012 年。

[18] Machine Learning in Silicon，UIUC ECE 598 NS 课程，Naresh R Shanbhag，2016 年。

[19] Hardware acceleration for machine learning and big data analytics，哥伦比亚大学 E6895 Advanced Big Data Analytics 课程，Ching-Yung Lin，2015 年。

[20] High Performance Computing with a Focus on Hardware Acceleration Technologies，University of Western Ontario 大学 CS4435b-CS9624b 课程，Marc Moreno Maza，2010 年。

[21] High-Performance Hardware for Machine Learning，NIPS Tutorial，William Dally，Stanford University & NVIDIA Corporation，2015 年。

六、课程中文简介

《深度学习系统》面向计算机等专业研究生，讲授深度学习系统的前沿技术。课程涵盖智能算法与计算机系统的核心知识，强调系统工程技术实践能力的培养。通过深度学习系统的典型应用案例，介绍模型优化、算法框架、编程模型、硬件架构、编译器及软件栈等原理与优化方法，以及深度学习系统的整体工作流程。本课程旨在使学生了解主流深度学习硬件架构和软件栈，掌握深度学习系统的软硬件协同设计、量化分析、以及性能优化方法，培养学生在深度学习系统方面的全栈技术能力。

七、其他

本课程提华为昇腾、树莓派等软硬件开发平台，包括深度学习 GPU 服务器，以及深度学习框架等工具环境支持。