

深度学习

从理论基础到前沿架构

MarkZZZ WeChat: MarkZZZ20XX

课程简介

本课程系统讲授深度学习的数学基础、核心模型与前沿方法。课程以严格的数学推导为核心，从万能近似理论、反向传播的微分几何解释、到优化理论（SGD 收敛性分析、损失景观几何）；深入探讨卷积神经网络的等变性理论、循环网络的动力系统观点、Transformer 的表达能力分析；涵盖生成模型的概率论基础（VAE 的 ELBO 推导、GAN 的博弈论分析、扩散模型的 SDE/ODE 理论）；以及预训练大模型的 Scaling Law、涌现现象与对齐方法。课程要求学员具备扎实的数学基础，能够理解并推导核心定理的完整证明，掌握从理论分析到实验设计的完整研究能力。

适合人群

- 计算机科学、数学、统计学及相关专业的博士研究生
- 从事深度学习理论与方法研究的科研人员
- 希望深入理解深度学习数学原理的高级工程师
- 对 AI 基础理论（逼近论、优化理论、统计学习理论）感兴趣的研究者

前置知识

- **线性代数**：矩阵分解（SVD、特征分解）、矩阵微积分、张量运算
- **实分析/泛函分析**：测度论基础、 L^p 空间、收敛性（几乎处处、依概率、依分布）
- **概率论与随机过程**：条件期望、鞅、随机微分方程基础
- **凸优化**：凸分析基础、KKT 条件、对偶理论、梯度方法收敛性
- **统计学习理论**：PAC 学习、VC 维、Rademacher 复杂度
- **编程能力**：Python + PyTorch/JAX，能够实现论文中的算法
- **机器学习**：监督/无监督学习核心方法（SVM、核方法、EM 算法等）

1 课程内容

讲次	主题	内容概要
1	深度学习导论	统计学习框架回顾（ERM/泛化风险/偏差-方差）、神经网络起源（McCulloch-Pitts/感知机收敛定理证明）、Minsky-Papert 不可能性、发展历程与三次浪潮、三大驱动力与 Scaling Law、表示学习层次理论、深度效率定理（Telgarsky）、NTK 视角、过参数化与双下降现象、三大理论问题（表达/优化/泛化）、课程路线图
2	前馈神经网络	感知机与线性可分性、多层感知机（MLP）与计算能力、激活函数分析（ReLU 族/GELU/Swish/SiLU 的光滑性与梯度性质）、万能近似定理（Cybenko 证明/Hornik 推广/Barron 定理与逼近速率）、深度效率定理（Telgarsky 完整证明思路）、深度 ReLU 网络的线性区域计数、宽度与深度的逼近论权衡、网络表达能力的拓扑视角
3	反向传播与优化	计算图与自动微分（正向/反向模式/Jacobian-vector product）、反向传播完整推导（链式法则的张量形式）、SGD 收敛理论（凸/强凸/非凸/PL 条件）、SGD 噪声与逃逸鞍点、自适应方法（AdaGrad/RMSProp/Adam/AdamW 收敛性分析）、学习率调度（Warmup/Cosine/OneCycle）、损失景观几何（Hessian 谱分析/鞍点/平坦极小值与 PAC-Bayes）、梯度流与隐式正则化
4	卷积神经网络	卷积定理与平移等变性（群论视角）、池化与近似不变性、感受野与有效感受野分析、经典架构（LeNet/AlexNet/VGG/GoogLeNet/ResNet/DenseNet/EfficientNet）、残差学习理论（恒等映射/梯度流/集成解释）、深度可分离卷积与参数效率、反卷积与特征可视化、Vision Transformer 与 CNN 的对比
5	循环神经网络	RNN 作为离散动力系统、BPTT 完整推导、梯度消失/爆炸的 Jacobian 谱分析（Bengio 1994/Pascanu 2013）、LSTM 门控机制推导与梯度流分析、GRU 简化与 LSTM 比较、序列到序列模型与编码器-解码器框架、注意力机制引入（Bahdanau/Luong）、RNN 表达能力（图灵完备性）、状态空间模型（S4/Mamba）

讲次	主题	内容概要
6	正则化与泛化	Dropout 的贝叶斯解释与 Rademacher 复杂度分析、Batch/Layer/Group/RMS Normalization 的数学分析与梯度效应、权重衰减 vs L_2 正则化 (SGD/Adam 中的本质区别)、数据增强的理论基础 (不变性/等变性约束)、Mixup 与 Cut-Mix 的正则化效应、梯度裁剪理论、混合精度训练 (FP16/BF16 数值分析)、标签平滑与知识蒸馏、PAC-Bayes 泛化界在深度网络中的应用
7	生成模型	自编码器 (AE/DAE/Sparse AE 的正则化效应)、VAE (ELBO 完整推导/重参数化技巧/后验坍塌/ β -VAE/VQ-VAE)、GAN (极小极大博弈/最优判别器定理/Jensen-Shannon 散度/训练不稳定性分析/ Wasserstein GAN 与 Kantorovich-Rubinstein 对偶/谱归一化/渐进增长)、归一化流 (变量替换定理/可逆网络/RealNVP/Glow)、能量模型与对比散度
8	注意力机制与 Transformer	自注意力的核方法解释 (核函数视角/softmax 核)、多头注意力与子空间投影、位置编码理论 (绝对/相对/旋转 RoPE/ALiBi)、Transformer 完整架构 (Pre-Norm vs Post-Norm/残差流)、表达能力分析 (图灵完备性证明思路/逼近能力)、高效注意力 (线性注意力/稀疏注意力/Flash Attention IO 复杂度分析)、长上下文扩展策略
9	预训练与大语言模型	BERT 双向编码与掩码语言模型、GPT 自回归建模与 next-token prediction、Scaling Law (Kaplan 2020/Chinchilla 最优分配)、涌现能力与相变解释、参数高效微调 (LoRA 低秩分解/Adapter/Prefix Tuning/QLoRA)、Prompt 工程与 In-context Learning 理论、RLHF 对齐 (PPO/DPO/Constitutional AI)、推理增强 (CoT/ToT/RAG)
10	扩散模型与前沿	去噪扩散概率模型 (DDPM 前向/反向过程完整推导)、得分匹配 (Hyvärinen 2005) 与 SDE 框架 (Song 2021)、采样加速 (DDIM/DPM-Solver)、条件生成 (Classifier/Classifier-Free Guidance)、多模态学习 (CLIP 对比学习/Flamingo/LLaVA)、图神经网络 (消息传递/表达能力/WL 测试)、深度强化学习、AI for Science (AlphaFold/分子生成/PDE 求解)

2 参考资料

1. Ian Goodfellow, Yoshua Bengio, Aaron Courville. *Deep Learning*. MIT Press, 2016.
2. 邱锡鹏. 神经网络与深度学习. 机械工业出版社, 2020.
3. Aston Zhang, Zachary C. Lipton, Mu Li, Alexander J. Smola. *Dive into Deep Learning*, 2nd ed., 2023.
4. Shalev-Shwartz, Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.
5. Roman Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018.
6. Ashish Vaswani et al. "Attention Is All You Need." *NeurIPS*, 2017.
7. Kaiming He et al. "Deep Residual Learning for Image Recognition." *CVPR*, 2016.
8. Diederik P. Kingma, Max Welling. "Auto-Encoding Variational Bayes." *ICLR*, 2014.
9. Jonathan Ho, Ajay Jain, Pieter Abbeel. "Denoising Diffusion Probabilistic Models." *NeurIPS*, 2020.
10. Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers." *NAACL*, 2019.
11. Jared Kaplan et al. "Scaling Laws for Neural Language Models." *arXiv:2001.08361*, 2020.
12. Yang Song et al. "Score-Based Generative Modeling through Stochastic Differential Equations." *ICLR*, 2021.
13. Belkin, Hsu, Ma, Mandal. "Reconciling Modern Machine-Learning Practice and the Classical Bias-Variance Trade-off." *PNAS*, 2019.
14. Arora, Du, Hu, Li, Wang. "Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks." *ICML*, 2019.
15. Bubeck, Sellke. "A Universal Law of Robustness via Isoperimetry." *NeurIPS*, 2021.