未来发展方向：联邦学习、自监督学习与通用人工智能

2025 年 10 月 22 日

目录

1　联邦学习

联邦学习（Federated Learning, FL）在不集中原始数据的情况下对分布式数据进行协同建模。设客户端集合为 $\mathcal{K} = \{1, \dots, K\}$，第 k 个客户端拥有本地数据集 $\mathcal{D}_k$ 及目标函数 $\ell(\mathbf{w}; \mathcal{D}_k)$，全局优化问题可写为

$$\min_{\mathbf{w}} f(\mathbf{w}) = \sum_{k=1}^{K} p_k \, \ell(\mathbf{w}; \mathcal{D}_k), \qquad p_k = \frac{|\mathcal{D}_k|}{\sum_{j=1}^{K} |\mathcal{D}_j|}. \tag{1}$$

与集中式训练相比，FL 需要同时处理统计异质性、通信瓶颈、设备资源受限和隐私合规等多重挑战。

1.1　系统架构与通信模式

标准 FL 架构由中心协调器和若干客户端组成，训练过程中交替执行模型广播与本地更新聚合：

- **同步模式**：参与的客户端在每轮训练结束后统一上传更新，易受慢设备影响而出现拖尾（straggler）。
- **异步模式**：服务端随时接收并融合更新，需要对梯度时延进行建模以保证稳定性。
- **分层联邦**：将终端设备的更新先在边缘节点聚合，再上传至云端，多级结构可降低延迟并扩展规模。

客户端调度策略会考虑算力、网络带宽及数据新鲜度，必要时引入公平性约束以避免部分用户长期被忽视。

1.2 优化算法

FedAvg 是最经典的联邦优化算法，本地执行 E 次 SGD 后统一平均：

$$\mathbf{w}^{(t+1)} = \sum_{k=1}^{K} p_k \, \mathbf{w}_k^{(t+1)}, \quad \mathbf{w}_k^{(t+1)} = \mathbf{w}^{(t)} - \eta \sum_{\tau=1}^{E} \nabla \ell(\mathbf{w}_k^{(\tau)}; \xi_k^{\tau}). \tag{2}$$

为应对非独立同分布（non-IID）和通信限制，研究者提出了多种改进：

- **FedProx**：在本地目标中加入近端正则项 $\frac{\mu}{2}\|\mathbf{w} - \mathbf{w}^{(t)}\|_2^2$ 抑制漂移。
- **SCAFFOLD**：利用控制变量降低客户端漂移，提升收敛精度。
- **FedOpt 系列**：在服务端引入自适应优化器（FedAdam、FedYogi）或归一化策略（FedNova）。
- **通信压缩**：通过量化、稀疏化、误差反馈减少上传字节数。

理论分析常假设梯度光滑且客户端梯度差异有界 $\mathbb{E}\|\nabla f_k - \nabla f\|^2 \le \beta^2$，以推导全局收敛界。

1.3 隐私、安全与可信协同

为保护用户隐私，需要在模型训练与参数传输中引入安全机制：

- **安全聚合**：通过加法掩码或秘密分享让服务端仅获知各客户端更新之和。
- **差分隐私**：对梯度裁剪并加入高斯噪声，满足 (ε, δ)-DP；噪声尺度 σ 与隐私预算及迭代次数密切相关。
- **同态加密**：在暗文域执行聚合，代价是显著的计算与通信开销。

同时需防范恶意客户端发起的中毒攻击、后门攻击。常用防御包括余弦相似度检测、Robust Aggregation（Krum、Median、Trimmed Mean）以及基于影响函数的认证方法。

1.4 应用与案例

联邦学习已在多个隐私敏感场景落地：

- **移动端智能**：输入法预测、语音唤醒和个性化推荐。
- **医疗联合体**：跨医院医学影像诊断，在不交换病历的情况下提升模型性能。
- **金融风控**：多机构协同反欺诈，满足数据出境和隐私监管要求。

新兴趋势包括联邦大模型、跨企业联盟训练、星地协同（卫星-地面终端）等。

1.5 开放问题

未来研究值得关注的方向有：

- **异质性处理**：长期在线场景中数据分布漂移、设备上线率不均的问题。
- **个性化**：如何在共享全局知识的同时保留用户特定偏好，可结合元学习、模型混合或多任务学习。
- **激励机制**：设计博弈论视角下的激励与信誉体系，鼓励真实贡献。
- **合规治理**：构建可审计的训练流水线，满足数据主权与合规追溯要求。

Listing 1: 加入差分隐私与安全聚合的联邦平均算法示意。

```python
def federated_round(server_state, clients, aggregator, noise_multiplier):
    encrypted_updates = []
    for client in clients:
        local_state = client.download(server_state.model)
        local_update = client.train(local_state)
        clipped = clip_by_global_norm(local_update, max_norm=1.0)
        dp_update = clipped + gaussian_noise(scale=noise_multiplier)
        encrypted_updates.append(encrypt(dp_update, client.public_key))
    aggregate = aggregator.secure_sum(encrypted_updates)
    server_state.optimizer.step(aggregate)
    return server_state
```

2 自监督学习

自监督学习（Self-Supervised Learning, SSL）利用数据本身的结构生成监督信号，在缺乏人工标注的情况下学习可迁移表示。

2.1 对比学习目标

对比学习通过最大化不同视角之间的相似性、最小化负样本的相似度来构建鲁棒特征。给定锚点 $\mathbf{x}$、正样本 $\mathbf{x}^+$ 以及负样本集合 $\{\mathbf{x}_j^-\}$，常见的 InfoNCE 损失为

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\operatorname{sim}(g(f(\mathbf{x})), g(f(\mathbf{x}^+)))/\tau)}{\sum_{j=1}^{N} \exp(\operatorname{sim}(g(f(\mathbf{x})), g(f(\mathbf{x}_j^-)))/\tau)}, \tag{3}$$

其中 sim 通常为余弦相似度，τ 为温度系数。丰富的增强方式（裁剪、颜色扰动、混合增强）对于提升泛化至关重要。

2.2 非对比与掩码建模

BYOL、SimSiam 等方法通过动量编码器与预测头避免使用负样本，利用梯度截断（stop-gradient）与结构差异防止表示坍缩。掩码自编码器（MAE）通过重建被遮挡的补丁，目标函数为

$$\mathcal{L}_{\mathrm{MAE}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2, \tag{4}$$

其中 $\mathcal{M}$ 为掩码索引集合。类似思想已扩展至语音（HuBERT）、蛋白质序列（ESM）及多模态输入。

2.3 模型结构选择

视觉领域常采用 Vision Transformer (ViT)，其 patch 嵌入与全局注意力机制易于结合多种预训练任务；语音与文本任务则偏好卷积前端加 Transformer/Conformer 主干。投影头、批归一化及白化层可调节表示分布，避免退化。多任务预训练将聚类、旋转预测、上下文自回归等任务融合以增强稳健性。

2.4 迁移与评估

线性探测（linear probing）通过冻结编码器并训练线性分类器评估表示质量，全量微调衡量可塑性。k-NN 分类、少样本任务以及分布外鲁棒性评估提供补充视角。针对多模态模型（例如 CLIP），跨模态检索、零样本分类和文本生成等指标衡量语义对齐程度。

2.5 扩展性与高效训练

大规模自监督显示出数据量、模型规模与误差之间的幂律关系。为提升效率，可从以下方面入手：

- **存储与算力优化：**梯度检查点、LARS/LAMB 等大批量优化器、混合精度训练。
- **负样本管理：** MoCo 动态队列、记忆库、微批次内特征复用。
- **课程式掩码：**自适应地调整掩码比例，使任务难度逐渐提升。

理论分析关注互信息估计的偏差、等变性与不变性之间的平衡、以及隐式正则化对泛化的影响。

2.6 应用场景

自监督学习已成为视觉、语音、自然语言、机器人学和科学计算等领域的大模型基石。典型案例包括 CLIP 图文对齐、wav2vec 语音识别、蛋白质语言模型用于结构预测等。结合自监督世界模型的强化学习能够在低样本环境中实现规划与推理。

Listing 2: 分布式对比学习（SimCLR）训练循环示例。

```
for step, (images, _) in enumerate(loader):
    x1, x2 = augment(images), augment(images)
    z1, z2 = projector(encoder(x1)), projector(encoder(x2))
    z1 = normalize(all_gather(z1))
    z2 = normalize(all_gather(z2))
    logits = similarity_matrix(z1, z2) / temperature
    labels = torch.arange(len(z1), device=z1.device)
    loss = cross_entropy(logits, labels)
    optimizer.zero_grad()
    loss.backward()
    optimizer.step()
```

3 通用人工智能（AGI）

通用人工智能旨在构建具备跨领域迁移、规划与推理能力的系统，能够完成多样化复杂任务。其研究融合机器学习、神经科学、认知科学及哲学等多学科知识。

3.1 定义与能力分级

AGI 的定义尚无统一共识，有的强调广义问题解决能力，有的更关注灵活目标导向。Legg 与 Hutter 提出的通用智能度量为

$$\Upsilon(\pi) = \sum_{\mu \in \mathcal{E}} 2^{-K(\mu)} V_{\mu}^{\pi}, \tag{5}$$

其中 V_{μ}^{π} 表示策略 π 在环境 μ 中的期望回报，$K(\mu)$ 为环境的 Kolmogorov 复杂度。实践中通常依赖覆盖多维能力的基准集合，如推理、操作、社交与工具使用等。

3.2 体系结构范式

通往 AGI 的路径呈现多种形态：

- **大规模神经网络**：基于 Transformer 的基础模型通过扩大参数量、引入专家混合（MoE）、检索增强和工具调用等能力持续提升。
- **神经符号融合**：将可微感知与符号推理、逻辑规划、程序合成结合，以获取可解释结构。
- **具身智能**：在虚拟或真实环境中通过交互习得世界模型，使语言、动作与感知相互对齐。

- **元学习与持续学习**：通过快速适应、避免灾难性遗忘，实现终身学习。

研究热点涵盖模块化体系、世界模型、决策 Transformer 等统一规划与强化学习的框架。

3.3 安全、对齐与治理

AGI 发展伴随显著的安全风险：

- **外部对齐**：使用逆强化学习、偏好建模、宪法式 AI 等方法，使目标函数符合人类价值。
- **内部对齐**：关注模型内部目标是否与外部期望一致，防止欺骗性行为。
- **可解释性与验证**：通过机制可解释性、神经元激活分析、形式化规范提高透明度。
- **治理与政策**：建立国际合作、审计标准、事件报告机制及出口管制框架。

风险评估常以概率安全约束 $\mathbb{P}(\text{catastrophe}) \leq \delta$ 等指标表达，同时需要模拟奖励投机、规范规避等极端情境。

3.4 评估体系与基准

单一基准难以衡量通用智能，通常构建复合评估体系，包括推理（MMLU、BIG-bench）、互动（ARC、MineRL）、工具调用（代码生成、API 使用）等维度。人类在环的红队测试能够发现长尾失效模式，对模型对齐具有重要意义。

表 1: AGI 评估维度与示例基准。

评估维度	示例基准/任务	关键指标
抽象推理	逻辑推理、定理证明	样本效率、推理链完整度
具身交互	家庭操作、机器人操控	感知-动作对齐、安全约束遵守
社会认知	谈判博弈、道德困境模拟	价值一致性、心智理论能力
工具使用	代码生成、外部 API 调用	可靠性、自检能力

3.5 路线图与开放问题

面向 AGI 的关键科学问题包括：

- **高效学习**：依靠结构化先验、组合泛化与世界知识，在少量监督下获得高性能。
- **鲁棒泛化**：面对分布转移、对抗场景和未知任务时保持可靠性。

- **价值学习**：如何全面表示、多方协商并动态更新人类偏好。
- **社会影响**：评估宏观经济、就业与伦理影响，制定负责任的部署策略。

跨学科合作（认知科学、神经科学、社会科学）有助于建立统一的理论框架与评测协议。

进一步阅读

- Brendan McMahan 等. “Communication-Efficient Learning of Deep Networks from Decentralized Data.” AISTATS 2017.
- Tian Li 等. “Federated Optimization in Heterogeneous Networks.” MLSys 2020.
- Ting Chen 等. “A Simple Framework for Contrastive Learning of Visual Representations.” ICML 2020.
- Kaiming He 等. “Masked Autoencoders Are Scalable Vision Learners.” CVPR 2022.
- Brian Christian. 《The Alignment Problem》. WW Norton, 2020.
- Joseph Carlsmith. “Is Power-Seeking AI an Existential Risk?” Open Philanthropy, 2022.