# 第十七次周报

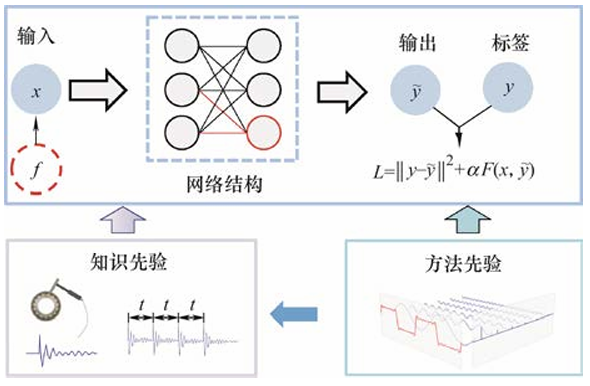
## 这两周工作内容

### 阅读的文献

* 《可解释人工智能在工业智能诊断中的挑战和机遇：先验赋能》

针对当前工业智能诊断领域人工智能模型可解释性不足的问题，本文系统综述了通过信号处理先验和物理知识先验赋能的可解释人工智能（XAI）方法，总结了相关研究进展，并分析了主要挑战与发展机遇。

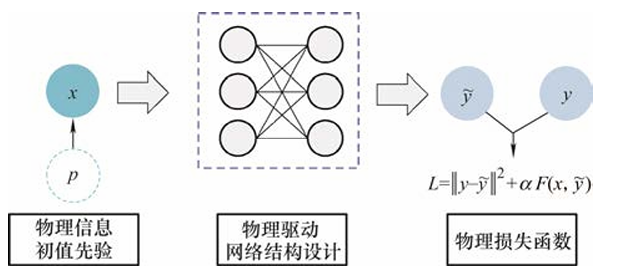
信号处理先验赋能的XAI基于故障诊断信号分析方法对深度学习算法模型结构以及优化目标进行修改，以期获得信号处理领域专家认可并能理解的模型内在的可解释性。一般将信号处理信息神经网络网络分为知识先验赋能型网络与方法先验赋能型网络。



知识先验赋能型网络根据对机械设备信号的分析与处理，可以获取设备故障信号的知识先验。其本质为结合神经网络设计合理的特征工程，此类方法使网络受到已知经验知识的约束。然而信号先验的获取严重依赖专家经验，不合理的信号先验会严重影响网络性能；部分设计中仅关注网络输入与输出，网络内部结构仍旧无法理解。此类研究目前数量较少，主要应用于故障诊断和迁移学习等领域。

方法先验的网络是近年来的研究热点，其主要范式为将常用的信号处理方法或流程作为网络特征提取模块，使用深度学习模块完成其功能或将信号处理方法参数化使其可学习。根据信号处理方法分类，可以大致分为小波变换及短时傅里叶变换等滤波器类、稀疏表征及理论展开类、降噪类及统计方法类，该类网络参数量受信号处理方法先验约束，降低了对数据量的要求，然而，由于网络结构受信号处理方法限制，通常难以通过堆叠网络深度以提升性能，无法构建大模型；并且大部分方法仅实现可解释的特征提取网络构建，从特征到决策的可解释仍旧亟待解决。此类研究已应用于数据生成、时频谱分析、信号降噪、异常检测、故障诊断、寿命预测上。

物理知识先验赋能的XAI方法则是通过引入系统的物理规律、控制方程或机理模型，对网络的输入、结构或损失函数进行约束和改造，从而提升模型的物理可解释性与泛化能力。常见实现方式包括物理信息初值先验、物理驱动的网络架构设计，以及物理损失函数等。



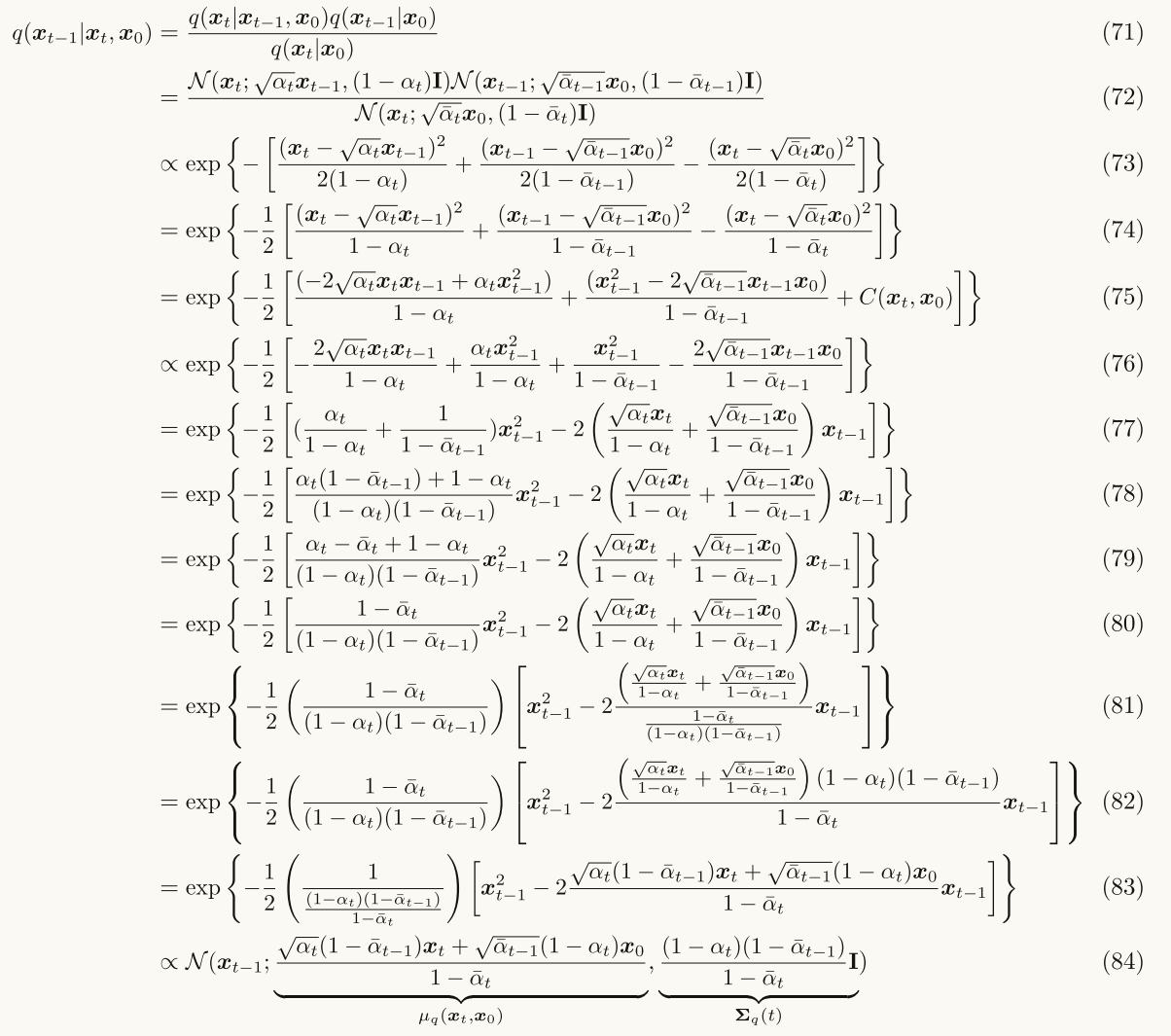
物理信息初值先验的学习范式类似于迁移学习，避免了网络从头开始学习，提高训练效率；同时具有良好的初始化权重，避免网络陷入局部最小。此方法利用物理模型产生的模拟数据来预训练深度网络模型，缓解了数据匮乏的问题。但此方法网络模型内部依然是黑箱模型；且依赖仿真软件产生仿真数据，针对复杂的物理系统，很难得到高保真的仿真数据。物理信息初值先验方法主要应用在数据生成、异常检测、故障诊断、剩余寿命预测和迁移学习等领域。

物理驱动的网络架构设计使用领域知识指定网络中的节点连接，以捕获变量之间的物理依赖关系；此外，物理指导的网络结构设计可以使之前的黑箱算法更易于解释，这是物理建模中使用深度网络模型的一个理想但通常缺失的特征；同时，在深度网络架构中添加上物理中间变量，可以帮助提取物理上有意义的、可以解释的隐藏表示；而将神经网络中的一个或多个权重固定为物理意义上的值或参数，既可以增加中间变量的可解释性，同时使得训练更加稳健；按这种方式构建的基于物理的神经网络，能够量化不同应用的不确定性和各种类型的先验知识。但这种模式下的网络结构的设计通常取决于模型结构，而不同应用场景需要特定的网络结构；物理模型的结构要求较高，使用复杂的物理模型较为困难。此类模式主要应用于异常检测、故障诊断、剩余寿命预测和物理方程求解上。

物理损失函数引入物理损失作为一种正则化项缓解过度拟合问题，减少对大训练数据集的需求，从而提高训练模型的鲁棒性，以实现更可靠的预测；同时使神经网络的输出更符合物理规律，增强了神经网络输出的可解释性；缩小了网络参数W和b的搜索空间，因此可以减少对训练样本数量的需求；并通过将物理知识整合到损失函数中，克服了使用噪声数据和高维数据建模的困难；同时在训练物理约束的神经网络时，允许使用未标记或不带标签的数据；且使得机器学习模型可以推广到训练数据以外的符合相同物理规律的场景。但建立的物理模型一般是低保真的，不同应用场景物理规律相差较大，通用性较差；且建立复杂物理现象的模型比较困难；同时网络内部结构仍然是黑箱模型。此类模式多用于异常检测、故障诊断、剩余寿命预测和数据生成上。

* 《Understanding Diffusion Models: A Unified Perspective》（正在阅读并推导公式）

本论文围绕扩散模型（Diffusion Models）的理论基础、统一视角及其与其他生成模型的联系，进行了系统的梳理和总结。目前阅读到扩散模型公式推导部分。



## 遇到的问题

### 2.1 扩散模型涉及到的概率论、随机过程的知识较多，需要再学习一下。

### 2.2 公式与物理直观的结合难度大，数学推导和物理直观之间的对应关系不容易建立，尝试在扩散模型中加入物理指导比较困难。

## 收获与启发

### 3.1 通过系统梳理扩散模型相关理论，进一步明确了生成模型之间的内在联系，特别是扩散模型、VAE、score-based models等之间的统一视角，有助于后续跨方法迁移和创新。

### 3.2 意识到当前研究需要兼顾理论创新与实际应用，下一步需重点提升自己在概率模型、深度生成模型推理方面的能力，并关注可解释性与工程应用结合的新范式。

## 下两周计划

### 4.1 继续阅读和整理扩散模型及故障预测相关文献。

### 4.2 尝试复现扩散模型的基础代码，实现简单的图像生成或信号去噪实验，加深对扩散过程、采样方法的实际理解。