Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

初构人工智能内生安全机理体系
—智能模型的不确定性估计

硕士研究生 吴肖龙

2023年11月05日

问题回溯

- 总结反思
  - 讲解语速过快，内容引入和衔接不够自然
  - 未考虑听众的反应和接受程度，需要注重整体节奏

- 相关内容
  - 2023.04.02 夏志豪《深度神经网络鲁棒性评估方法》
  - 2022.08.23 王若辉《AI测试：历史与发展》
  - 2022.07.24 侯钰斌《神经网络模型测试方法与模型健壮性》

内容提要

- 预期收获
- 题目内涵解析
- 研究背景与意义
- 研究历史与现状
- 知识基础
- 算法原理
  - Nflows
  - DDU
- 特点总结与未来展望
- 参考文献

预期收获

- 预期收获
  - 了解模型不确定性估计的基本概念和现有体系
  - 理解模型不确定性估计两个学派四类方法的核心思想
  - 理解如何对不确定性建模和类型分离
  - 了解模型不确定性估计的前沿发展和关键挑战

目标内涵 智能模型不确定性估计

- 不确定性
  - 反映一个随机变量的离散程度
  - 模型：所输出的预测结果，在条件概率上对应的随机变量的离散程度。

Softmax为什么不行?

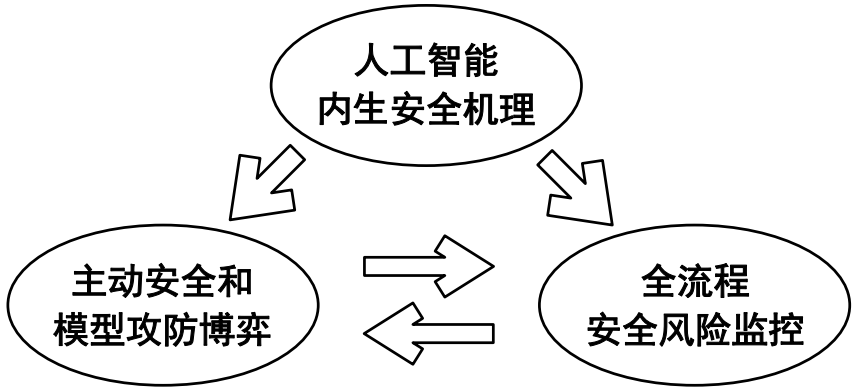
- 不确定性的来源
  - 真实世界复杂多变
  - 噪声与观测误差
  - 模型架构的缺陷
  - 模型学习程度不足

任意（aleatoric）不确定性

认知（epistemic）不确定性

预测（总体）不确定性

- 研究目标
  - 准确的量化智能模型的不确定性（各类）
  - 表述智能模型存在的内生安全问题
  - 在现有工作的基础上进行更好的校准



背景意义

- 研究背景
  - 真实世界环境下的数据缺陷
  - 模型性能的长期瓶颈
  - 高安全性需求领域的关键决策
    - 自动驾驶、智慧医疗、国防和关基设施

2021年，NHTSA报告807起自动驾驶相关案件

- 研究意义
  - 防止智能模型过度自信（学会表达不知道）
  - 表明何时人们应该放弃接受预测
  - 暴露和发掘智能模型的潜在缺陷
  - 促进可信（可靠）人工智能的构建

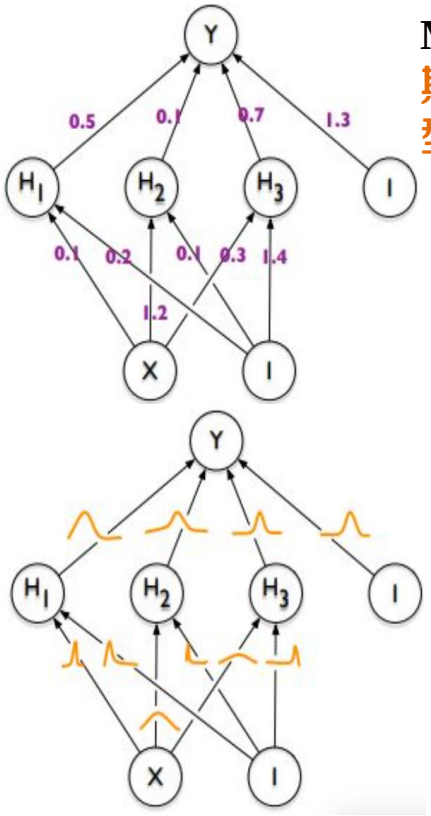
2000~2013年间，机器人手术致死患者达144人

可信人工智能体系在社会实践场景下的关键环节和基本要求！

研究历史

智能模型不确定性估计

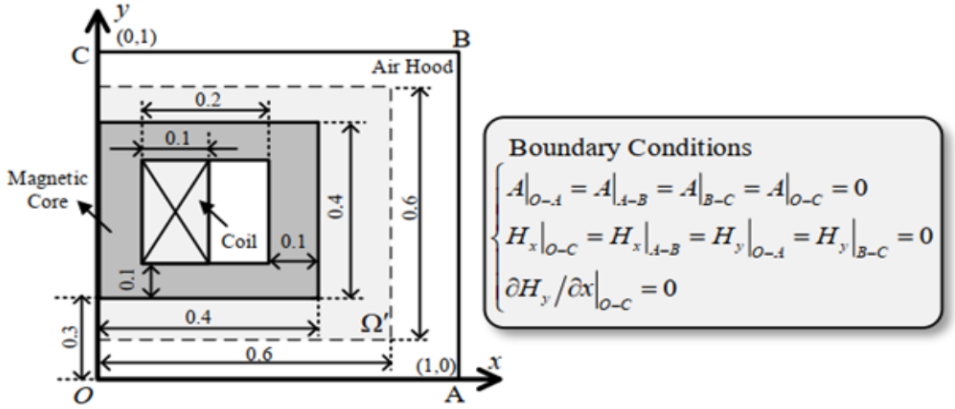
1995年，BNN的提出最早成为不确定性估计的雏形工作



2011年，SGLD和MFVI算法基于贝叶斯推断正式对智能模型进行不确定性建模

2016年，MC-dropout提出引入集成思想

2017年，基于bagging和boosting的原理正式创建深度集成的方法



2017年，将温度缩放作用于softmax，开始了对单一测度方法的研究和持续改进

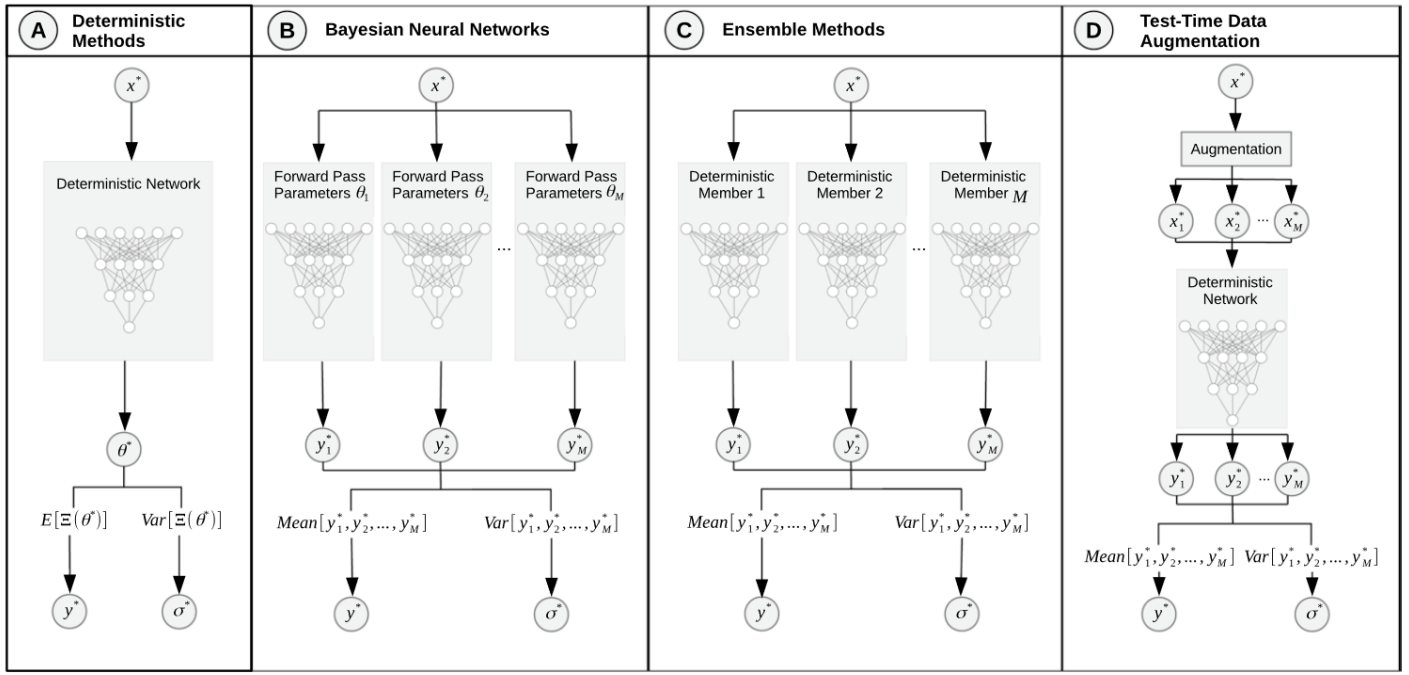
2018年，Mixup算法提出，通过线性插值做数据增广，引入了测试时间增强方法

2019年，基于熵对不确定性的任意部分和认知部分分别计算，开启了不确定性分离建模

2023～2024，面向专项领域任务的不确定性建模工作兴起
例如：ECG多标签分类心电检测、小样本化学分子建模、物理约束偏微分方程求解等

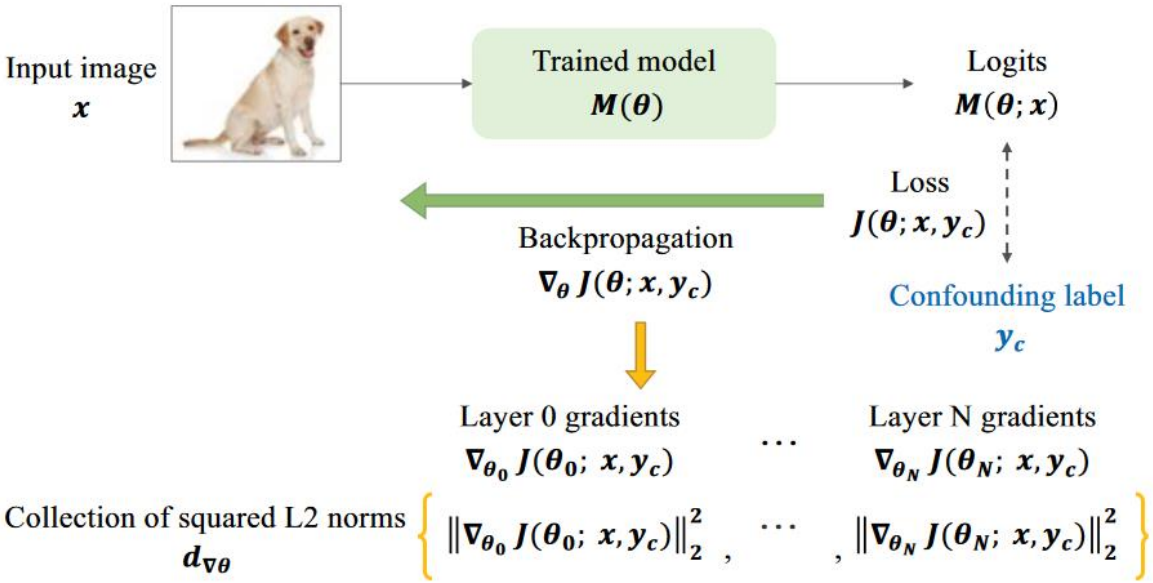
研究现状 两个学派 四类方法

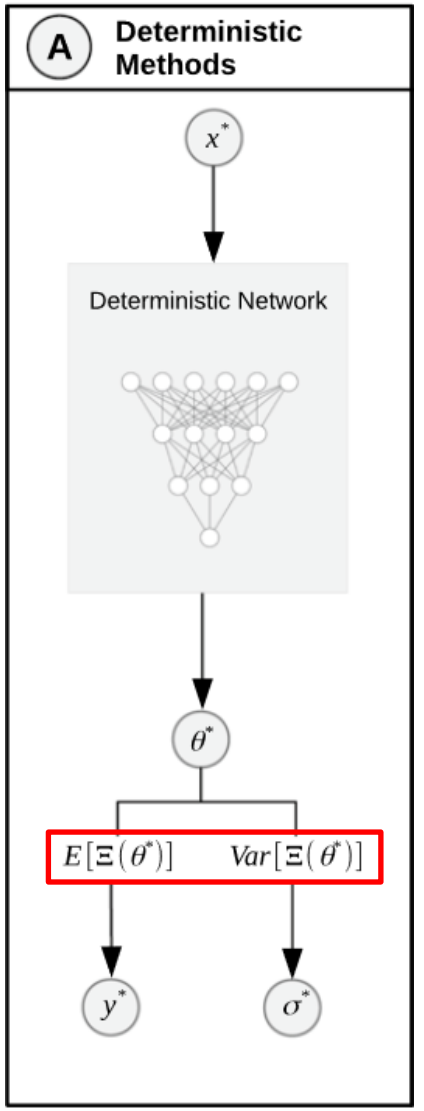
- 频率学派：在大量重复试验的情况下，观测数据会趋于真实参数的频率分布
  - 基于大数定律和中心极限定理
  - 通过统计性质和置信区间表示不确定性
- 概率学派：结合先验信息和观测数据，推断条件概率下后验概率分布
  - 基于贝叶斯定理和概率分布
  - 通过参数估计表示不确定性
- 方法类型
  - 单一测度度量
  - 贝叶斯推断
  - 深度集成
  - 测试时间增强



知识基础 单一测度

- 核心思想
  - 利用各类数理统计方法，对模型的输出或中间过程信息进行不确定性建模
- 区间神经网络、梯度更新、均值、方差、微分熵……
- 基本示例





计算方式容易实现且灵活多变 | 准确性不足且理论依据不充分

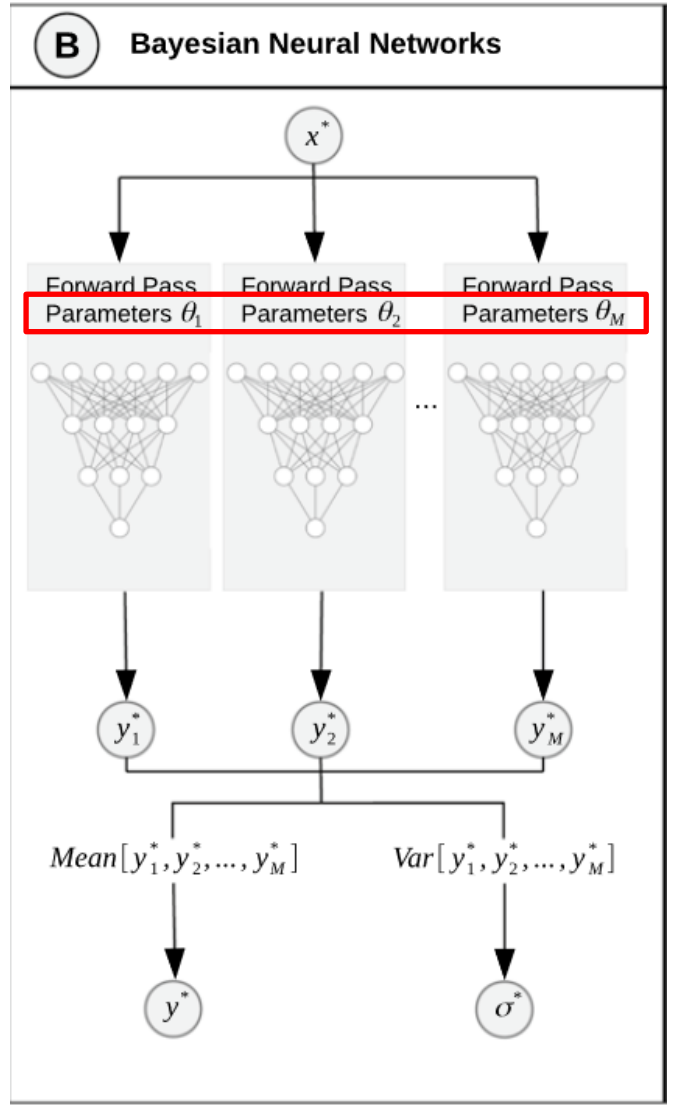
知识基础 贝叶斯推断

- 核心思想
  - 基于观察数据的似然估计推断模型参数概率，基于参数不确定程度分析模型不确定性
- BNN、变分推断、拉普拉斯近似……
- 基本示例（贝叶斯公式）

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

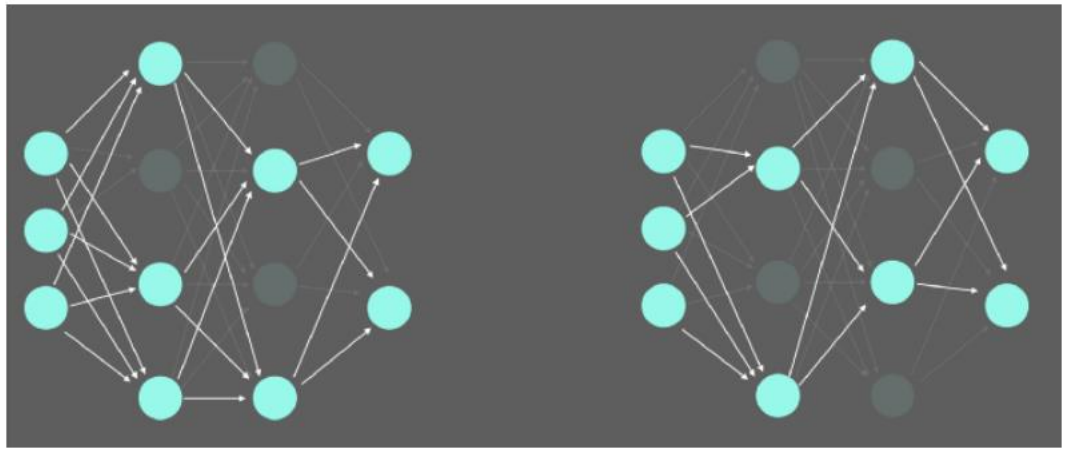
  - $P(\theta)$为参数的先验分布（附录A）
  - $P(D|\theta)$为θ条件下的模型预测的分布概率
  - $P(D)$为真实的后验概率分布，等同于$P(Y|X)$

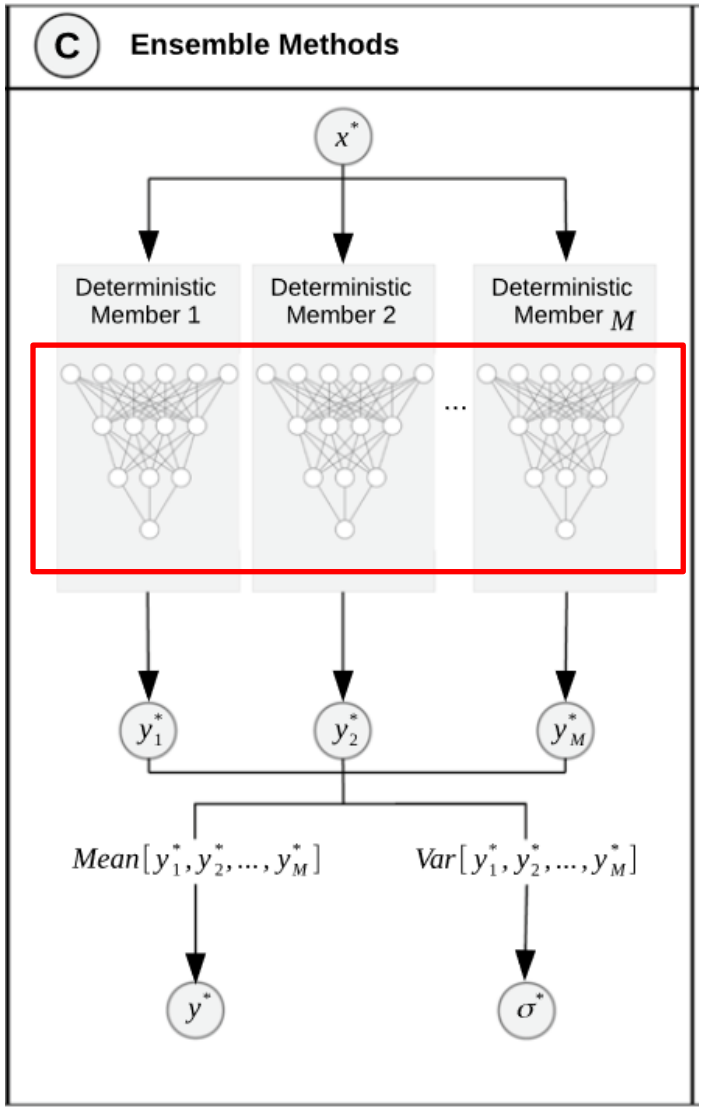
理论框架完整且能够直接推断 | 复杂度高且扩展性差



知识基础 深度集成

- 核心思想
  - 基于不同初始设置，训练多个独立的模型，通过捕捉和分析不同子模型间的差异定义模型的不确定性
- MC-dropout、MIMO、集成蒸馏……
- 基本示例

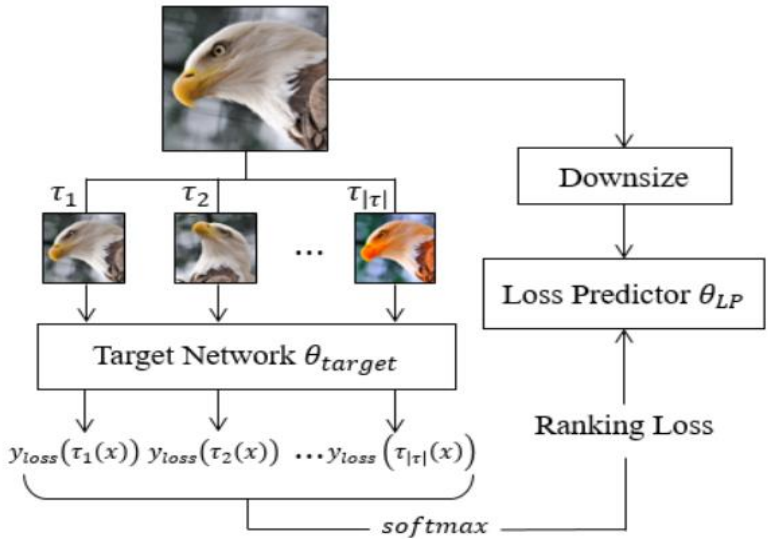




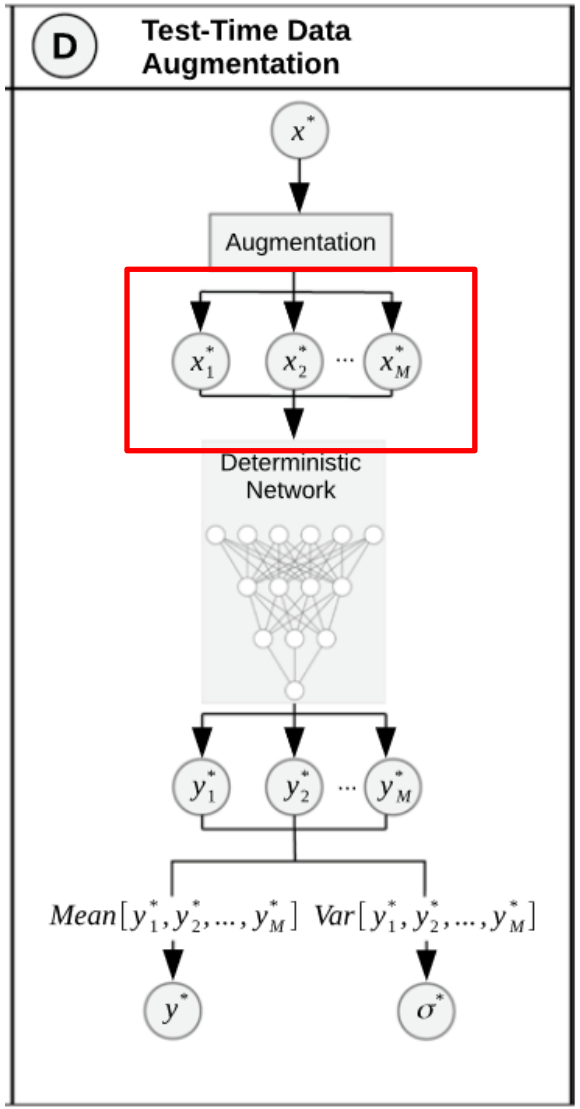
性能好且能有效降低不确定性 | 空间和算力开销

知识基础 测试时间增强

- 核心思想
  - 通过数据增广和生成，捕获和分析相似不同输入数据产生的输出，定义并计算模型的不确定性
- 线性插值、图形变换、对抗训练……
- 基本示例



兼顾对数据进行良好的评估 | 约束条件多且设计复杂



知识基础 不确定性建模和分离

- 根本任务

$$P(y^*|x^*) = \int_{\Omega} P(y^*|\omega)\, P(\omega|x^*)\, d\omega, \quad y^* = argmax\, P(y|x^*)$$

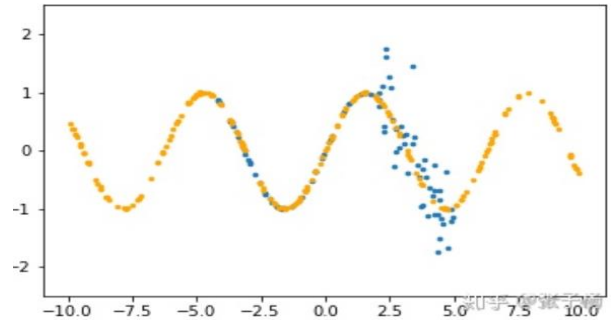
- 由潜在变量到观察数据

$$P(y^*|x^*) = \int_{D} P(y^*|D, x^*), \quad y^* = argmax\, P(y|D, x^*)$$

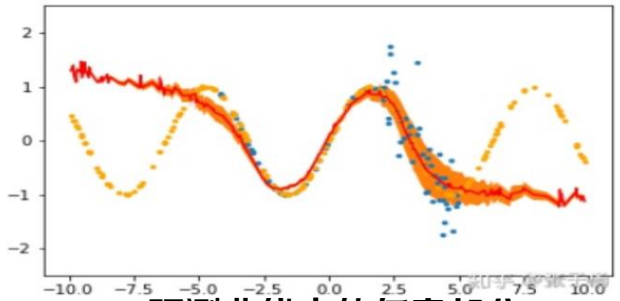
$\{x, y \mid x \in X, y \in Y\}$

- 参数化建模

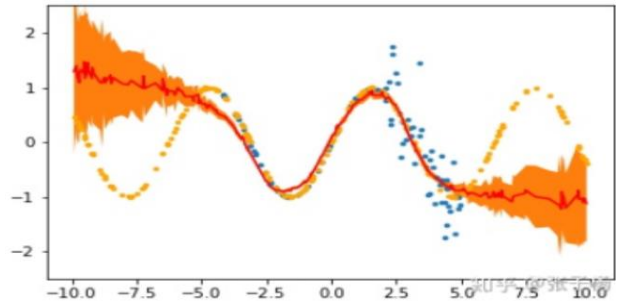
$$P(y^*|D, x^*) = \int \underbrace{P(y^*|x^*, \theta)}_{\text{aleatoric}} \underbrace{P(\theta|D)}_{\text{epistemic}}\, d\theta$$



训练样本和测试样本



预测曲线中的任意部分



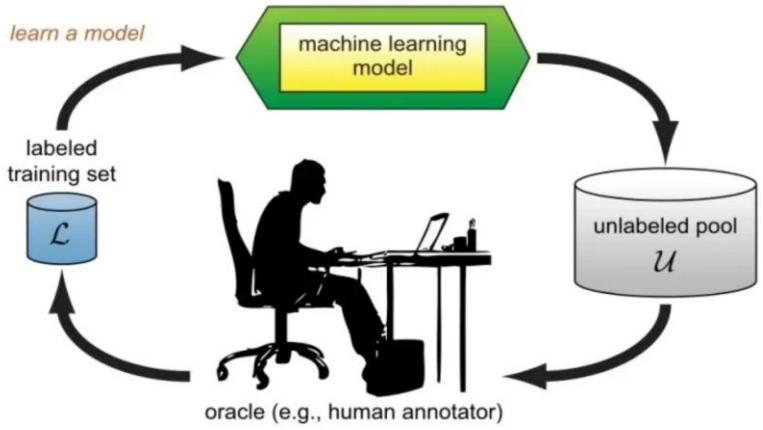
预测曲线中的认知部分

分离是为了更好的建模，而不是只在乎其中之一！！！

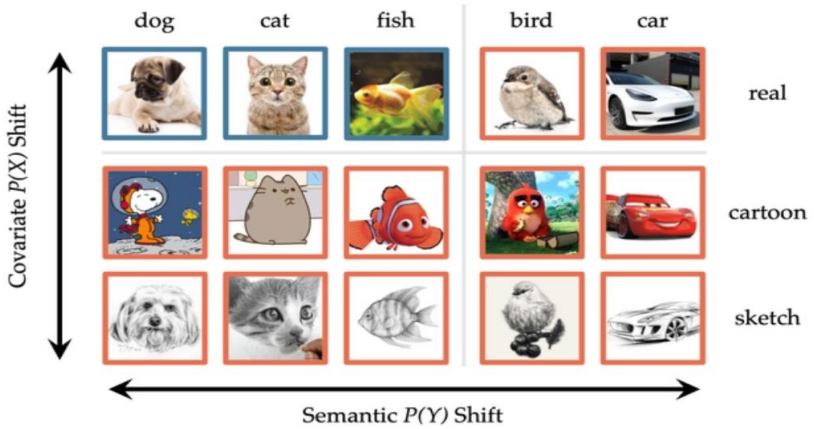
知识基础 评价体系

- 核心问题
  - 如何评估一个评估方法的好坏（实验怎么做）
- 评价指标
  - 同golden truth的距离差距（几乎没有）
  - 结合校准工作的度量（ECE，见p33）
  - 结合特定下游任务的效果
    - 主动学习
    - 分布外检测
    - 语义分割

语义分割



主动学习



语义和像素空间上的分布外数据

更高效的模型学习；更准确的数据认知

算法原理 Nflows

【AAAI】

Normalizing Flow Ensembles for Rich Aleatoric and Epistemic Uncertainty Modeling

算法原理 Nflows

T	目标	实现灵活、可靠的不确定性估计
I	输入	待测试模型*1个、训练样本*1组
P	处理	1. 利用归一化流拟合后验分布 2. 基于非线性变换和基础分布两种方式创建集成 3. 利用条件熵和差分熵定义不确定性
O	输出	模型的不确定性估计值

P	问题	1. 传统的后验近似方式仅能模拟典型的数学分布 2. 直接集成的计算和时空开销较高
C	条件	需要了解训练数据
D	难点	1. 不确定性建模和不确定性计算方式的关联性 2. 灵活性、可靠性和开销的权衡
L	水平	AAAI 2023（CCF A）【McGill University】

- 乘法归一化流（附录B）
  - 叠加简单可逆变换将简单的连续分布转换为更复杂的分布
  - 单次可逆变换$f(\cdot)$，令$Y = f(B)$

$$P_Y(y) = P_B\left(f^{-1}(y)\right)\left|det\left(J\left(f^{-1}(y)\right)\right)\right|$$

其中$J(\cdot)$为雅克比行列式

  - 叠加可逆变换

$$Z_K = f_K \circ \cdots\cdots \circ f_2 \circ f_1(Z_0),\ Z_0 \sim q_0(Z_0)$$

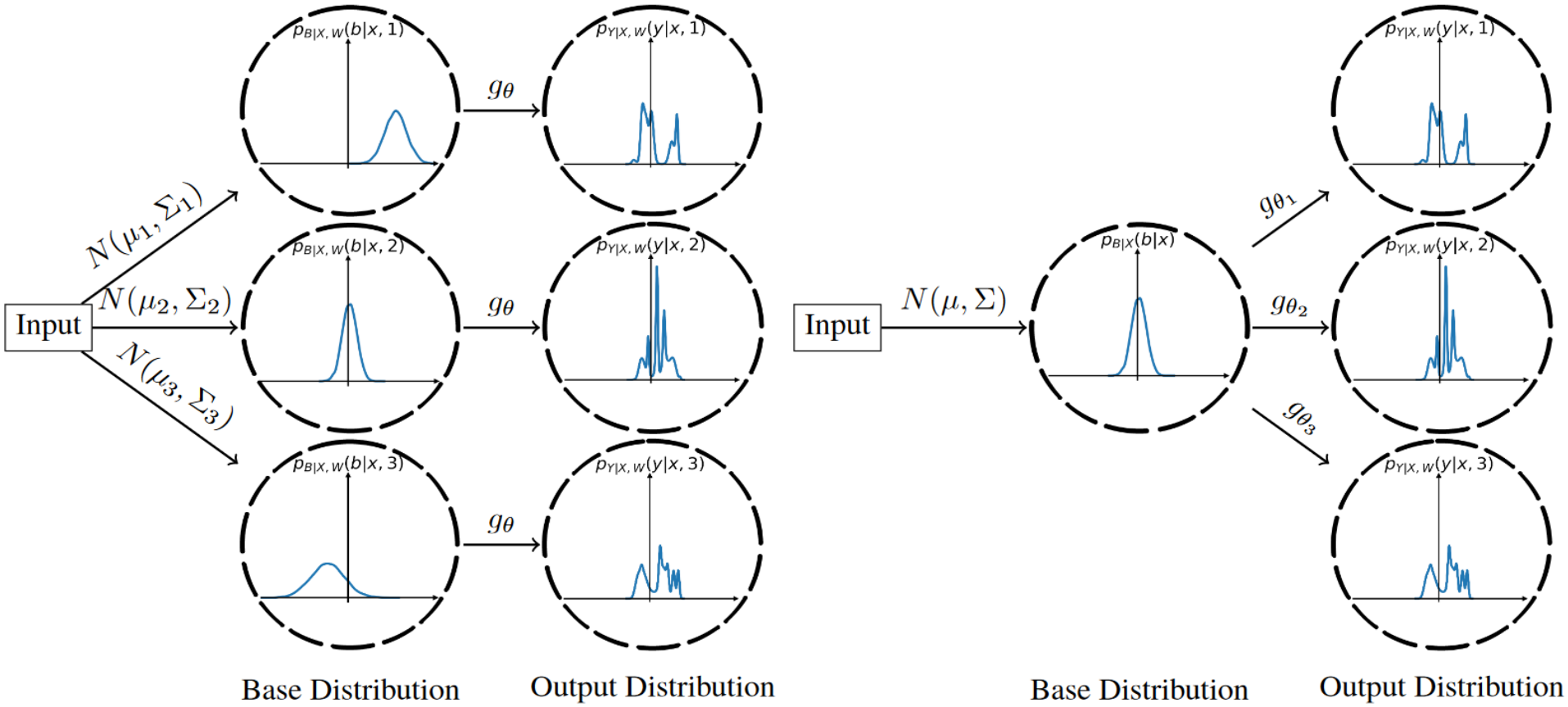
$$Z_K \sim q_K(Z_K) = q_0(Z_0)\prod_{k=1}^{K}\left(f^{-1}(y)\right)\left|det\left(J\left(f^{-1}(y)\right)\right)\right|$$

  - 通过参数化$f_\theta(\cdot)$学习对数似然

高斯分布->复数分布

算法原理 创新分析

- 利用归一化流更灵活拟合复杂后验分布
- 提出两种集成方式更丰富的进行不确定性建模



算法原理 不确定性估计

- Nflows Out：在非线性变换中创建集成（使用随机初始化和bagging）
  - 总体预测不确定性（条件熵）

$$H(y^*|x^*) = -E\left[log\left(P_{Y|X}(y|x^*)\right)\right] \approx -\frac{1}{N}\sum_{n=1}^{N} log\left(P_{Y|X}(y_n|x^*)\right)$$

  - 任意不确定性（条件熵）

$$E_{p(w)}[H(y^*|x^*,w)] = -\frac{1}{M}\sum_{w=1}^{M} E\left[log\left(P_{Y|X,W}(y|x^*,w)\right)\right]$$

$$\approx -\frac{1}{M}\sum_{w=1}^{M}\frac{1}{N_w}\sum_{N_w=1}^{N_w} log\left(P_{Y|X,W}(y_{n_w}|x^*,w)\right)$$

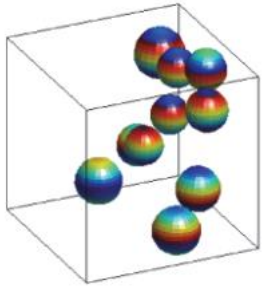
- Nflows Base：在基础分布中创建集成（固定dropout掩码）
  - 任意不确定性（差分熵）

$$\approx \frac{1}{M}\sum_{w=1}^{M}\frac{1}{2} log\left(det(2\pi\Sigma_w)\right)$$

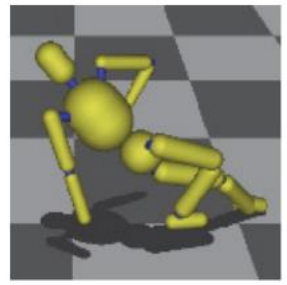
实验结果 数据资源

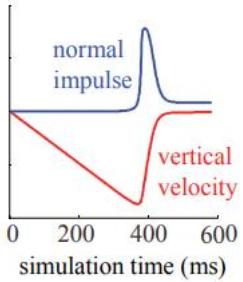
- 数据集说明

数据集	相关描述
Hetero	标准的异方差噪声采样点
Bimodal	标准的双峰分布采样点
Wet Chicken	模拟划艇接近瀑布边缘，划艇手们被吸引到瀑布边缘的物理情景，用于评估方法在多模态和异方差噪声方面的建模能力
Pendulum	模拟倒立摆的动态系统，包含了关于倒立摆的状态和动作的数据，用于测试方法对于连续动态系统的建模能力
Hopper	模拟跳跃机器人的动态系统，包含机器人的状态、动作和环境的信息，用于测试方法对于离散动态系统的建模能力



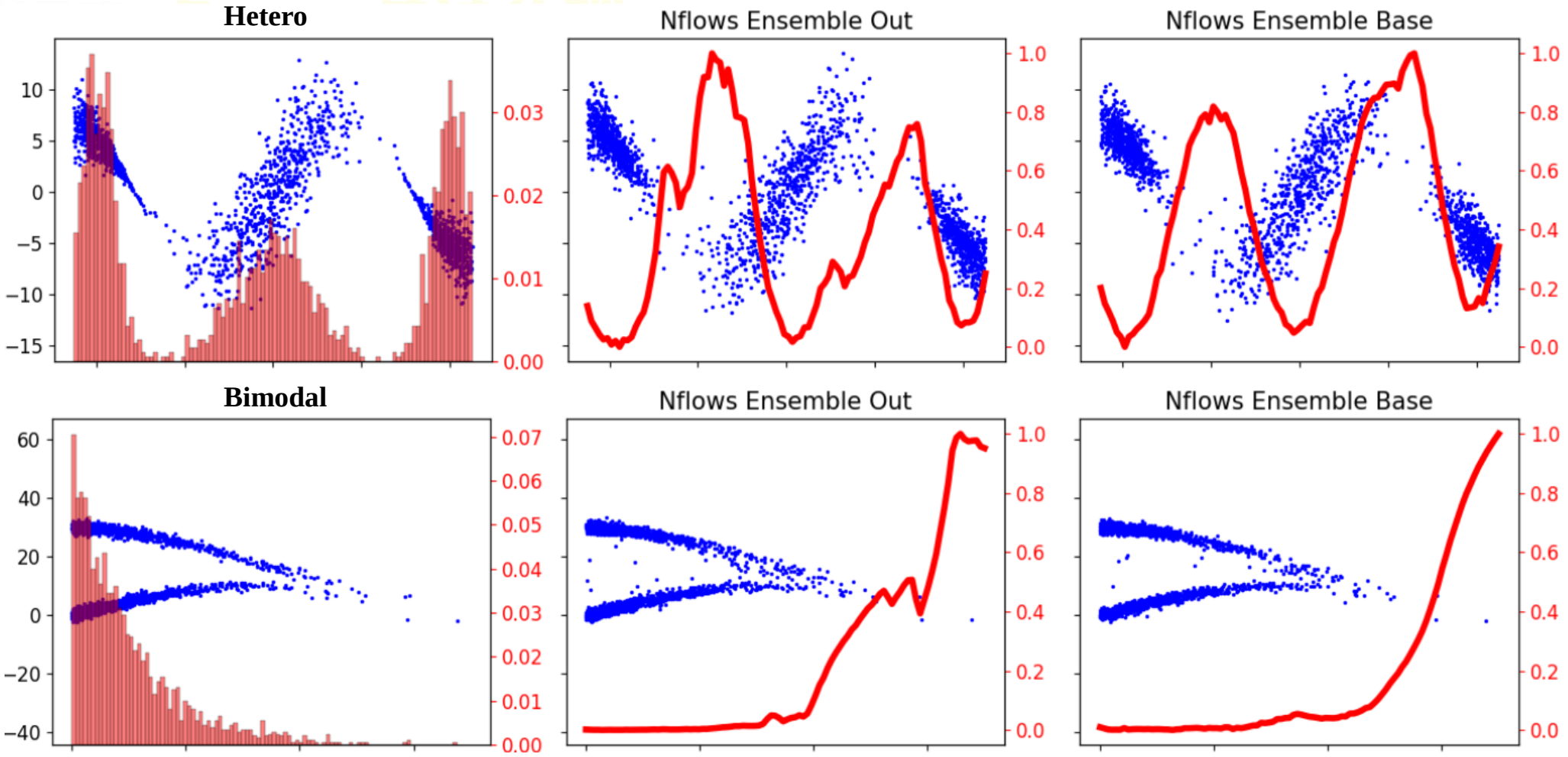
mu \ nc	7	16	27
0.1	3.8	4.8	6.5
0.5	2.6	5.3	7.5
1.0	2.8	4.9	10.2
2.0	2.9	4.6	16.2





基于典型数据分布和真实物理场景的采样点

实验结果 有效性基线实验



两种方法均有效的捕获了模型对于数据的认知不确定性

实验结果 主动学习实验

- 度量指标
  - KL散度：衡量两个概率分布间距离的非对称性度量（↓）

$$KL(P||Q) = \sum_{x \in X} P(x) log \frac{P(x)}{Q(x)} = \mathbb{E}_{x \sim P(x)} log \frac{P(x)}{Q(x)}$$

  - 均方误差（RMSE）：表示预测值与真实值之间的平均偏差程度（↓）

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}$$

- 对比方法
  - 高斯过程（GP）
  - 概率网络集成（PNEs）
  - 蒙特卡洛Dropout（MC-Dropout）

实验结果 主动学习实验-KL散度

Env	Acq Batch	GP	PNEs	MC Drop	Nflows	Nflows Out	Nflows Base
Hetero	10	1.43±0.12	1.39±0.06	1.43±0.1	1.54±0.23	**0.48±0.09**	**0.51±0.17**
	25	1.43±0.11	1.44±0.1	1.43±0.08	1.3±0.45	**0.31±0.09**	**0.43±0.11**
	50	1.41±0.07	1.39±0.07	1.46±0.1	1.39±0.21	**0.27±0.08**	**0.36±0.09**
	100	1.33±0.06	1.45±0.08	1.44±0.07	0.95±0.33	**0.3±0.08**	**0.38±0.06**
Bimodal	10	2.23±0.18	1.49±0.06	1.49±0.03	1.26±0.81	0.74±0.76	**0.36±0.07**
	25	2.02±0.07	1.46±0.05	1.47±0.04	1.21±0.6	**0.24±0.04**	**0.2±0.03**
	50	1.97±0.07	1.5±0.03	1.49±0.02	1.2±0.45	**0.22±0.03**	**0.18±0.03**
	100	2.3±0.03	1.51±0.05	1.49±0.05	1.07±0.32	**0.18±0.02**	**0.14±0.02**
Wet Chicken	10	7.61±0.21	**7.14±0.24**	7.93±0.27	8.04±1.07	7.83±1.2	7.67±1.36
	25	7.73±0.17	7.49±0.47	8.02±0.3	8.19±1.03	**6.64±0.79**	**6.25±0.98**
	50	7.81±0.12	7.61±0.44	7.95±0.18	8.12±0.78	**6.51±0.56**	**5.86±0.92**
	100	7.71±0.18	7.55±0.46	7.97±0.28	8.06±0.7	**6.73±1.06**	**5.93±1.01**
Pendulum-v0	10	**24.56±0.21**	27.29±0.73	31.04±0.41	26.45±4.61	27.62±1.87	26.07±2.22
	25	**24.52±0.3**	26.43±1.05	30.11±0.29	**24.86±3.66**	**24.13±1.18**	**23.97±2.16**
	50	24.68±0.26	26.47±1.19	29.6±0.26	24.44±3.04	**22.86±1.63**	**22.45±1.29**
	100	24.67±0.17	26.04±0.94	29.0±0.39	23.9±0.93	23.09±1.56	**21.93±1.17**
Hopper-v2	10	114.8±0.97	122.42±1.22	125.66±0.98	126.87±2.83	**112.79±1.18**	114.6±2.14
	25	113.29±0.62	120.3±1.56	125.99±0.84	123.93±1.84	**109.31±2.0**	**110.64±1.59**
	50	112.98±0.71	119.82±1.66	125.65±0.88	122.36±1.36	**108.59±1.04**	**109.44±1.87**
	100	112.27±1.0	118.4±1.21	125.21±1.36	119.97±1.91	**107.74±0.99**	**108.71±1.83**

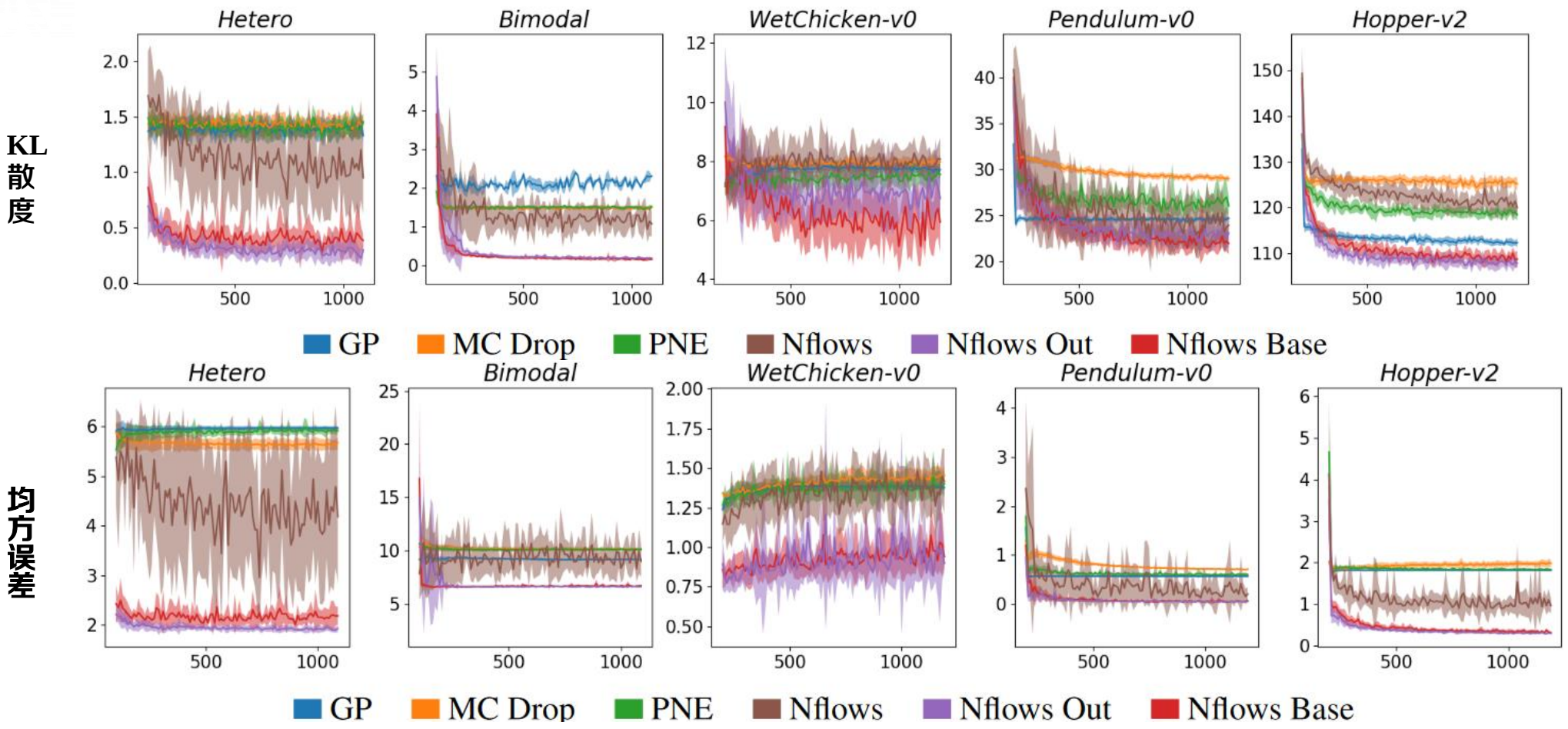
在绝大部分组合（数据集，样本量）下取得了最佳性能，Batch（样本量）达到100时论文方法均为最佳

实验结果 主动学习实验–RMSE

Env	Acq Batch	GP	PNEs	MC_drop	Nflows	Nflows_out	Nflows_base
Hetero	10	5.92±0.1	5.89±0.17	5.69±0.17	5.57±0.28	**2.01±0.12**	**2.17±0.1**
	25	5.95±0.06	5.91±0.13	5.66±0.14	4.74±1.57	**1.93±0.05**	**2.22±0.32**
	50	5.96±0.04	5.88±0.06	5.63±0.14	5.3±0.63	**1.92±0.08**	**2.15±0.16**
	100	5.96±0.04	5.91±0.09	5.67±0.13	4.18±1.37	**1.93±0.07**	**2.19±0.18**
Bimodal	10	9.19±0.24	10.22±0.24	10.3±0.27	7.69±0.87	**6.63±0.1**	**6.61±0.06**
	25	9.3±0.12	10.12±0.19	10.22±0.19	9.12±2.13	**6.62±0.08**	**6.67±0.06**
	50	9.23±0.08	10.11±0.14	10.19±0.16	9.62±2.97	**6.63±0.11**	**6.64±0.06**
	100	9.17±0.07	10.11±0.13	10.15±0.12	9.01±1.47	**6.63±0.04**	**6.72±0.1**
WetChicken	10	1.35±0.06	1.33±0.09	1.38±0.08	1.32±0.18	**0.85±0.07**	**0.89±0.09**
	25	1.38±0.03	1.36±0.07	1.39±0.04	1.37±0.2	**0.88±0.12**	**0.96±0.17**
	50	1.39±0.03	1.38±0.09	1.42±0.04	1.41±0.19	**0.88±0.09**	**0.89±0.08**
	100	1.38±0.02	1.4±0.08	1.45±0.07	1.42±0.21	**0.9±0.05**	**0.94±0.07**
Pendulum-v0	10	0.57±0.0	0.68±0.08	1.0±0.06	0.46±0.38	**0.15±0.1**	**0.17±0.14**
	25	0.57±0.0	0.61±0.02	0.84±0.05	0.28±0.39	**0.08±0.04**	**0.09±0.05**
	50	0.56±0.0	0.6±0.04	0.75±0.02	0.38±0.44	**0.05±0.03**	**0.06±0.03**
	100	0.56±0.0	0.6±0.03	0.71±0.01	0.2±0.14	**0.05±0.02**	**0.06±0.04**
Hopper-v2	10	1.81±0.01	1.87±0.05	1.86±0.05	1.44±0.76	**0.48±0.09**	**0.59±0.1**
	25	1.81±0.0	1.86±0.05	1.91±0.1	1.06±0.21	**0.39±0.06**	**0.42±0.09**
	50	1.81±0.0	1.85±0.03	1.94±0.06	1.05±0.19	**0.33±0.03**	**0.36±0.04**
	100	1.81±0.0	1.83±0.01	1.98±0.08	0.96±0.18	**0.29±0.02**	**0.31±0.03**

在所有场景下性能均为最佳

实验结果 主动学习实验-学习曲线



在初期即体现了显著的学习效果，在完成学习后性能最佳

【IEEE CVPR】

Deep Deterministic Uncertainty: A Simple Baseline

算法原理 DDU

T	目标	实现简单、高效的不确定性估计
I	输入	待测试模型*1个、训练样本*1组
P	处理	1. 利用双利普希茨约束的谱归一化构建正则化特征提取器 2. 按类别计算输出的均值、协方差等参数 3. 利用特征空间密度和分布熵分别定义2种不确定性
O	输出	模型的不确定性估计值

P	问题	1. 不加约束的模型预测输出难以保证稳定性和收敛性 2. 对两类不确定性的区分不清晰
C	条件	更适用于具有残差连接的模型结构
D	难点	1. 兼顾灵敏度和平滑度 2. 在简单实现的基础上实现高性能
L	水平	CVPR 2023（CCF A）【University of Oxford】

- 利普希茨连续
  - 满足公式的连续函数$f(x)$称为*K-Lipschitz*

$$\|f(x_2) - f(x_1)\| \leq K\|x_2 - x_1\|$$

- 双利普希茨约束

$$K_L d_I(x_1, x_2) \leq d_F\big(f_\theta(x_1), f_\theta(x_2)\big) \leq K_U d_I(x_1, x_2)$$

其中d_I, d_F分别表示输入和特征空间度量； K_L, K_U 分别表示上、下界常数

  - 下界保证对输入数据的敏感性，上界保证特征的平滑性
- 谱归一化
  - 神经网络层级间的输入输出关系

$$X_n = a_n(W_n X_{n-1} + b_n)$$

其中$a_n(\cdot)$是非线性激活函数， W_n是参数矩阵， b_n是网络的偏置

$$\Rightarrow X_n = D_n W_n X_{n-1}$$

- 谱归一化
  - 多层网络输入输出关系

$$f(X) = D_n W_n \cdots D_1 W_1 X$$

  - 在Lipschitz约束条件下的梯度要求

$$\left\|\nabla_x\left(f(x)\right)\right\|_2 = \|D_n W_n \cdots D_1 W_1\|_2 \leq \|D_n\|_2 \|W_n\|_2 \cdots \|D_1\|_2 \|W_1\|_2$$

其中$\|W\|_2$表示矩阵的谱范数，等于其最大奇异值$\sigma(W)$，即最大对角元素

  - 上式可表示为

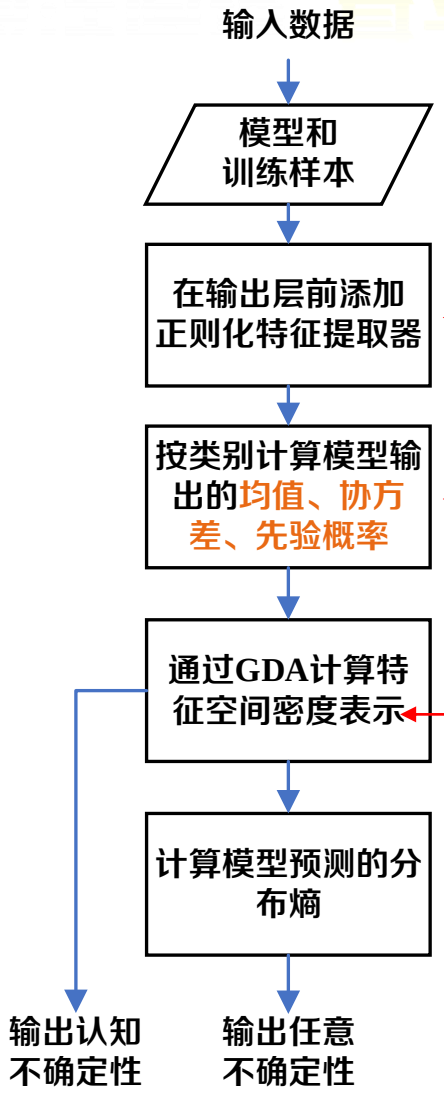
$$\left\|\nabla_x\left(f(x)\right)\right\|_2 \leq \prod_{i=1}^{N} \sigma(W_i)$$

  - 通过最大奇异值进行归一化

$$\left\|\nabla_x\left(f(x)\right)\right\|_2 = \left\| D_n \frac{W_n}{\sigma(W_n)} \cdots D_1 \frac{W_1}{\sigma(W_1)} \right\|_2 \leq \prod_{i=1}^{N} \frac{\sigma(W_i)}{\sigma(W_i)} = 1$$

基于上述方式实现了更好的特征空间正则化

算法原理 基本框架



Algorithm 1 Deep Deterministic Uncertainty

1: **Definitions:**
- Regularized feature extractor $f_\theta : x \to \mathbb{R}^d$ 双Lipschitz约束的谱归一化
- Softmax output predictions: $p(y|x)$
- GMM density: $q(z) = \sum_y q(z|y = c)\, q(y = c)$
- Dataset (X, Y)

2: **procedure** TRAIN
3: train NN $p(y|f_\theta(x))$ with (X, Y)
4: **for** each class c with samples $\mathbf{x}_c \subset X$ **do**
5: $\mu_c \leftarrow \frac{1}{|\mathbf{x}_c|} \sum_{\mathbf{x}_c} f_\theta(\mathbf{x}_c)$
6: $\Sigma_c \leftarrow \frac{1}{|\mathbf{x}_c|-1} (f_\theta(\mathbf{x}_c) - \mu_c)(f_\theta(\mathbf{x}_c) - \mu_c)^T$
7: $\pi_c \leftarrow \frac{\sum_{\mathbf{x}_c} 1}{|X|}$
8: **end for**
9: **end procedure**

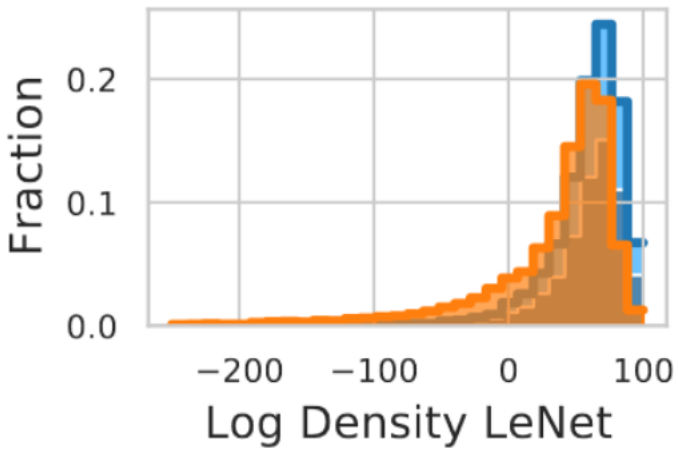
10: **function** DISENTANGLE_UNCERTAINTY(sample x)
11: compute feature representation $z = f_\theta(x)$
12: compute density under GMM: $q(z) = \sum_y \mathrm{q}(z \mid y)\, \mathrm{q}(y)$ with $\mathrm{q}(z \mid y) \sim \mathcal{N}(\mu_y; \sigma_y), \mathrm{q}(y) = \pi_y$
13: compute softmax entropy: $H_p[Y|x]$

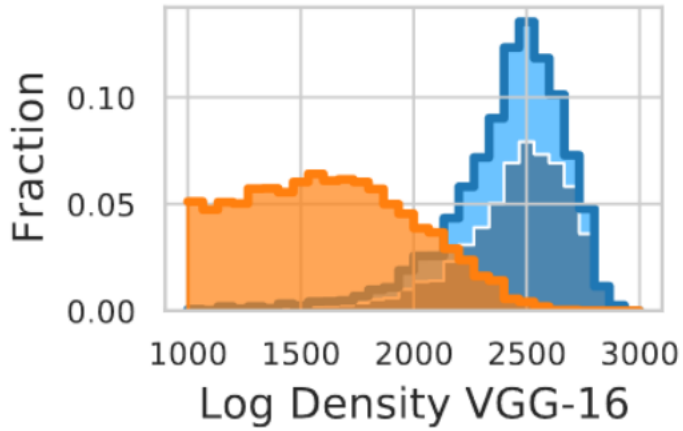
$$= -\sum P(Y = y|x) log P(Y = y|x)$$

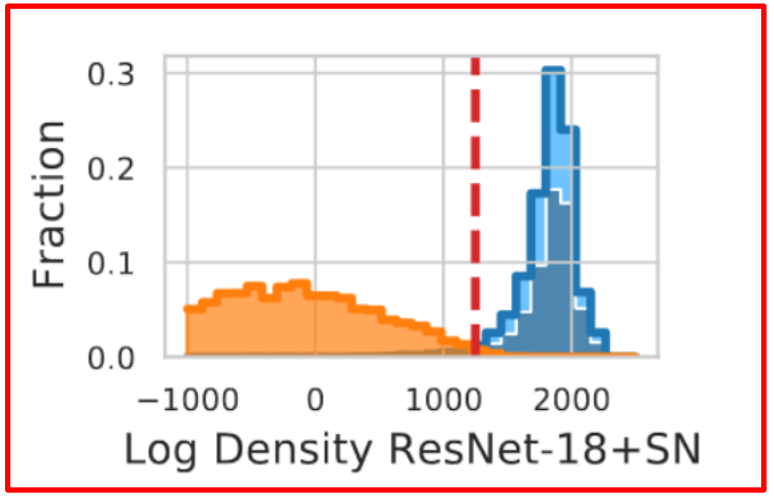
算法原理 创新分析

- 通过双利普希茨约束和谱归一化保证平滑度和灵敏度，防止特征崩塌
- 通过分别考虑特征密度估计和预测分布熵分离不确定性

- 特征崩塌
  - 特征正则化器可能将OOD输入的特征映射到特征空间中的ID区域



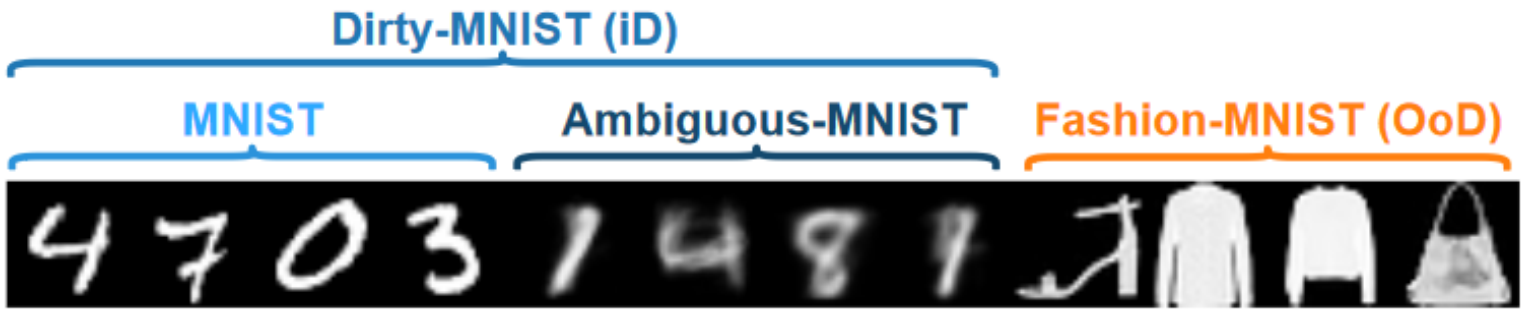




实验结果 数据资源

- 数据集说明

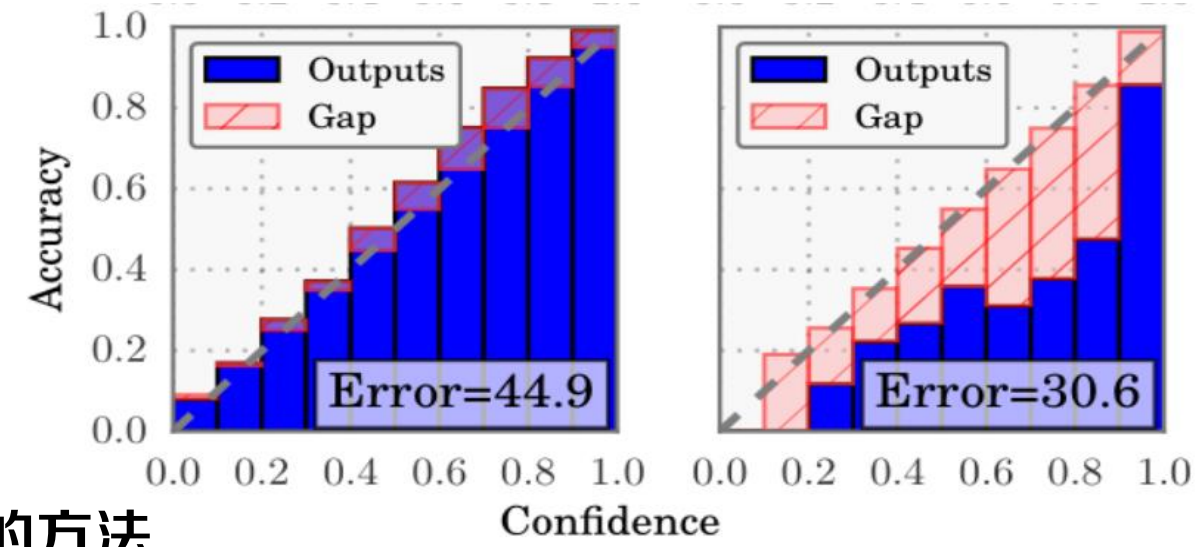
数据集	相关描述	实验中作用
MNIST	手写数字识别数据集（灰度图像）	基础数据
Ambiguous-MNIST	MNIST 的模糊化扩展	高不确定性数据
Fashion-MNIST	商品图像分类数据集，与 MNIST 数据集相似	分布外数据
CIFAR-10	来自 10 个不同类别的彩色图像	基础数据
CIFAR-10-C	对 CIFAR-10 图像应用不同类型的损坏操作来生成	高不确定性数据
CIFAR-100	CIFAR-10 的扩展，包含更细粒度的 100 个类别	分布外数据
SVHN	真实房屋门牌上的彩色数字识别	分布外数据
Tiny-ImageNet	大规模图像数据库 ImageNet 的子集，有 200 个类别	分布外数据



实验结果 评价指标和对比方法

- 评价指标
  - 准确率Accuracy（主动学习和OOD检测任务）
  - AUROC
    - 真阳率-假阳率曲线下面积
  - 平均校准误差ECE

$$ECE = \sum_{b=1}^{B} \frac{n_b}{N} |acc(b) - conf(b)|$$

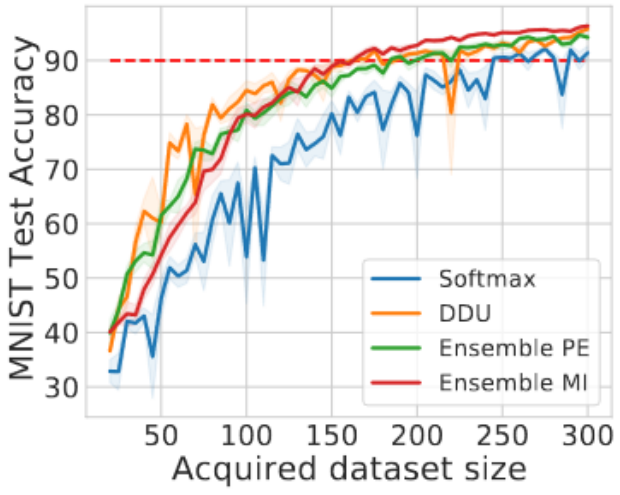


- 对比方法
  - Energy-based：基于能量分数计算的方法
  - DUQ：基于径向基函数网络的方法
  - SNGP：基于距离感知的方法
  - Deep Ensemble：深度集成基线

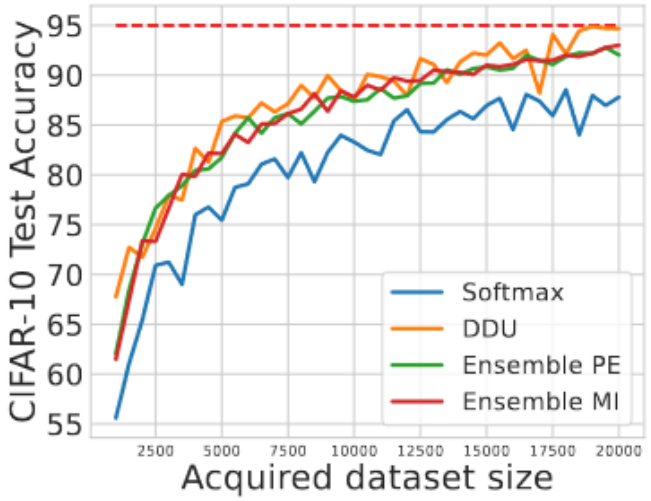
实验结果 主动学习实验

- 实验设置

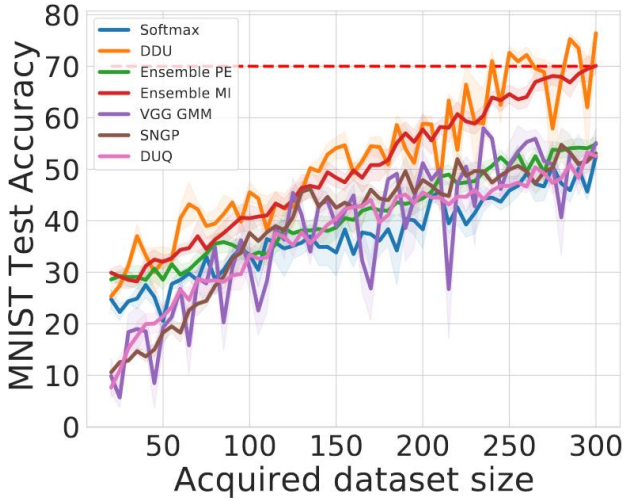
参数	实验设置（MNIST）	实验设置（CIFAR-10）
初始值	20 个随机样本点	100 个随机样本点
训练轮次	100	100
每次添加样本	5	500
截止条件	300	20000



MNIST



CIFAR-10



Dirty-MNIST

各类方法均体现出震荡上升，本方法整体位于最上端，且相对平稳

实验结果 OOD检测实验

Wide-ResNet-28-10模型架构中的实验结果

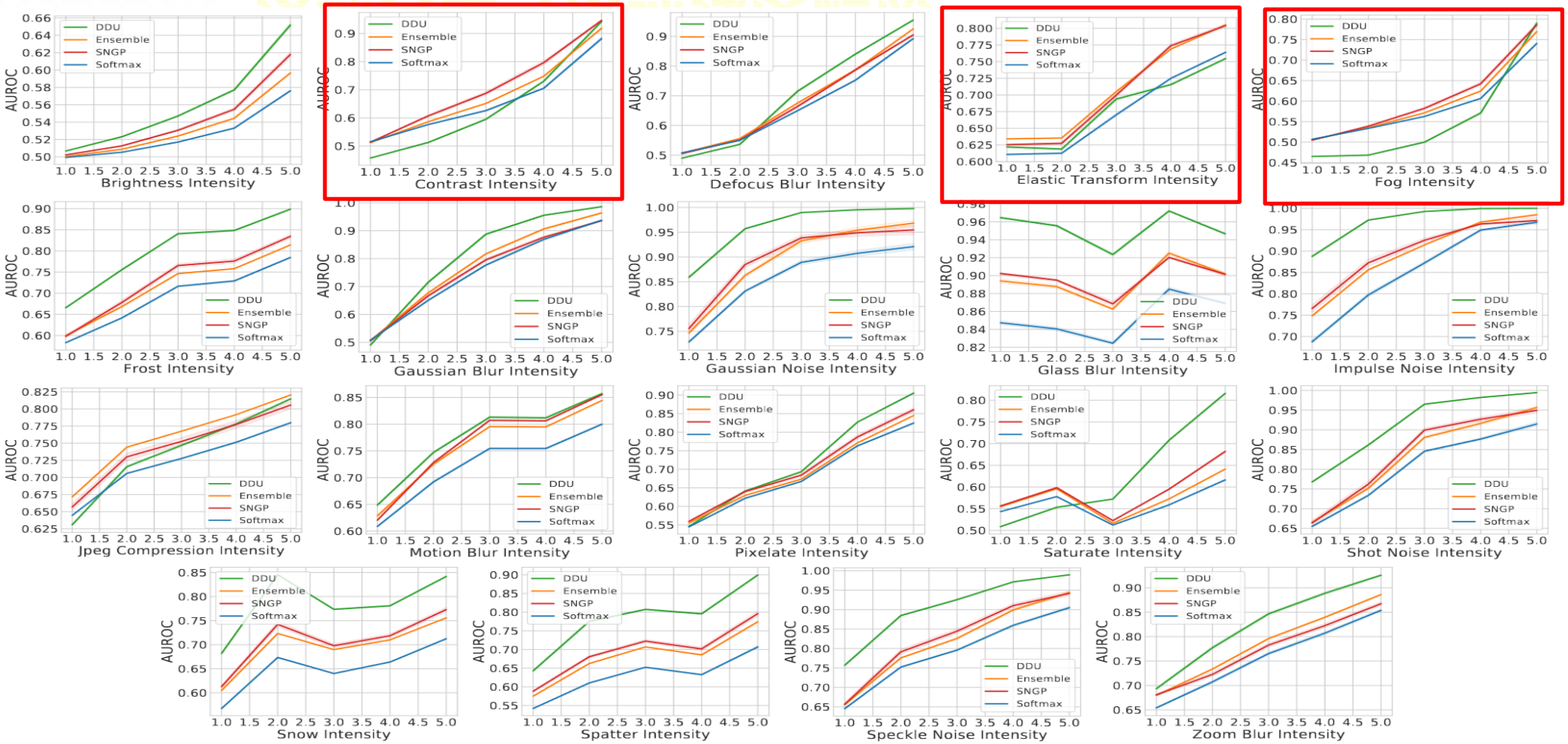
Train Dataset	Method	Penalty	Aleatoric Uncertainty	Epistemic Uncertainty	Accuracy (↑)	ECE (↓)	AUROC SVHN (↑)	AUROC CIFAR-100 (↑)	AUROC Tiny-ImageNet (↑)
CIFAR-10	Softmax	-	Softmax Entropy	Softmax Entropy	95.98 ± 0.02	**0.85 ± 0.02**	94.44 ± 0.43	89.39 ± 0.06	88.42 ± 0.05
	Energy-based (Liu et al., 2020b)	-		Softmax Density			94.56 ± 0.51	88.89 ± 0.07	88.11 ± 0.06
	DUQ (van Amersfoort et al., 2020)	JP	Kernel Distance	Kernel Distance	94.6 ± 0.16	1.55 ± 0.08	93.71 ± 0.61	85.92 ± 0.35	86.83 ± 0.12
	SNGP (Liu et al., 2020a)	SN	Predictive Entropy	Predictive Entropy	**96.04 ± 0.09**	1.8 ± 0.1	94.0 ± 1.3	91.13 ± 0.15	89.97 ± 0.19
	DDU (ours)	**SN**	**Softmax Entropy**	**GMM Density**	95.97 ± 0.03	**0.85 ± 0.04**	**97.86 ± 0.19**	**91.34 ± 0.04**	**91.07 ± 0.05**
	5-Ensemble (Lakshminarayanan et al., 2017)	-	Predictive Entropy	Predictive Entropy	**96.59 ± 0.02**	**0.76 ± 0.03**	97.73 ± 0.31	**92.13 ± 0.02**	90.06 ± 0.03
				Mutual Information			97.18 ± 0.19	91.33 ± 0.03	90.90 ± 0.03
CIFAR-100	Softmax	-	Softmax Entropy	Softmax Entropy	80.26 ± 0.06	4.62 ± 0.06	77.42 ± 0.57		81.53 ± 0.05
	Energy-based (Liu et al., 2020b)	-		Softmax Density			78 ± 0.63		81.33 ± 0.06
	SNGP (Liu et al., 2020a)	SN	Predictive Entropy	Predictive Entropy	80.00 ± 0.11	4.33 ± 0.01	85.71 ± 0.81		78.85 ± 0.43
	DDU (ours)	**SN**	**Softmax Entropy**	**GMM Density**	**80.98 ± 0.06**	**4.10 ± 0.08**	**87.53 ± 0.62**		**83.13 ± 0.06**
	5-Ensemble (Lakshminarayanan et al., 2017)	-	Predictive Entropy	Predictive Entropy	**82.79 ± 0.10**	**3.32 ± 0.09**	79.54 ± 0.91		82.95 ± 0.09
				Mutual Information			77.00 ± 1.54		82.82 ± 0.04

DenseNet-121模型架构中的实验结果

Train Dataset	Method	Penalty	Aleatoric Uncertainty	Epistemic Uncertainty	Test Accuracy (↑)	Test ECE (↓)	AUROC SVHN (↑)	AUROC CIFAR-100 (↑)	AUROC Tiny-ImageNet (↑)
CIFAR-10	Softmax	-	Softmax Entropy	Softmax Entropy	95.16 ± 0.03	1.10 ± 0.04	94 ± 0.44	87.55 ± 0.11	86.99 ± 0.12
	Energy-based (Liu et al., 2020b)	-		Softmax Density			94.07 ± 0.54	86.73 ± 0.15	86.43 ± 0.16
	DUQ (van Amersfoort et al., 2020)	JP	Kernel Distance	Kernel Distance	95.02 ± 0.14	1.08 ± 0.08	94.67 ± 0.41	87.38 ± 0.21	86.72 ± 0.14
	SNGP (Liu et al., 2020a)	SN	Predictive Entropy	Predictive Entropy	94.31 ± 0.21	1.08 ± 0.10	94.48 ± 0.34	88.86 ± 0.46	88.40 ± 0.48
	DDU (ours)	**SN**	Softmax Entropy	GMM Density	**95.21 ± 0.03**	**1.05 ± 0.03**	**96.21 ± 0.31**	**90.84 ± 0.06**	**89.70 ± 0.06**
	5-Ensemble (Lakshminarayanan et al., 2017)	-	Predictive Entropy	Predictive Entropy	**96.18 ± 0.05**	1.07 ± 0.07	95.78 ± 0.11	90.65 ± 0.03	89.62 ± 0.06
				Mutual Information			95.75 ± 0.10	90.71 ± 0.04	89.34 ± 0.06
CIFAR-100	Softmax	-	Softmax Entropy	Softmax Entropy	79.02 ± 0.08	4.11 ± 0.08	85.86 ± 0.42		81.10 ± 0.07
	Energy-based (Liu et al., 2020b)	-		Softmax Density			87.09 ± 0.49		80.84 ± 0.08
	SNGP (Liu et al., 2020a)	SN	Predictive Entropy	Predictive Entropy	79.15 ± 0.15	6.73 ± 0.10	85.00 ± 0.12		79.76 ± 0.15
	DDU (ours)	**SN**	Softmax Entropy	GMM Density	**79.15 ± 0.07**	**4.11 ± 0.06**	**88.44 ± 0.55**		**81.85 ± 0.11**
	5-Ensemble (Lakshminarayanan et al., 2017)	-	Predictive Entropy	Predictive Entropy	**81.01 ± 0.13**	4.81 ± 0.05	88.32 ± 0.61		81.45 ± 0.12
				Mutual Information			88.36 ± 0.17		81.73 ± 0.06

方法较其它单一测度方法有显著优势，能够持平乃至超越集成基线

实验结果 损坏类型与损坏强度检测 CIFAR-10-C



19类损坏中16种取得了SOTA，失利在对比度、弹性变形和雾强度（附录C）

特点总结与未来展望

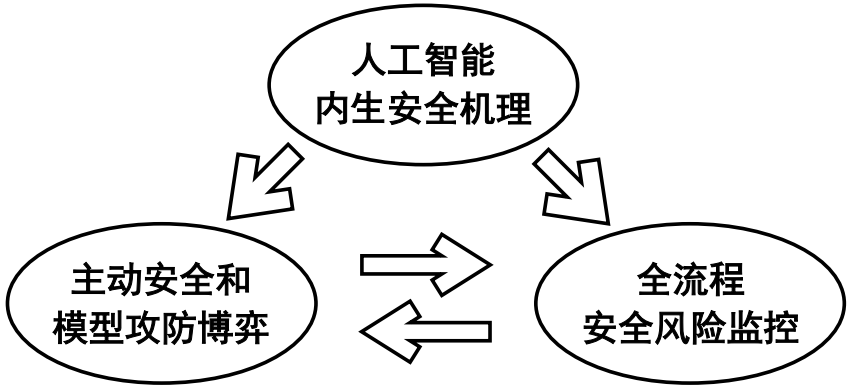
特点总结

- Nflows
  - 优点：更灵活的应对复杂的现实场景，在开销和性能上实现了较好的权衡
  - 缺点：未实现两类不确定性的良好分离，可解释性较差
- DDU
  - 优点：在简单高效的基础上实现了较好的不确定性估计，且定义比较清晰
  - 缺点：对于类别进行高斯建模难以应对更加复杂的情况
- 前沿发展与关键挑战
  - 更低的计算复杂度和时空开销（附录D）
  - 更好适应的具体应用领域中的数据
  - 结合可解释性等提供更完善的可信人工智能体系
  - 将不确定性估计任务内置到网络设计中

预期收获 回顾分析

- 预期收获
  - 了解模型不确定性估计的基本概念和现有体系
    - 预测值（随机变量）的离散程度
  - 理解模型不确定性估计两个学派四类方法的核心思想
    - 概率学派、频率学派
    - 单一测度、贝叶斯推断、深度集成、测试时间增强
  - 理解如何对不确定性建模和类型分离
    - 数据来源噪声、模型认知程度
  - 了解模型不确定性估计的前沿发展和关键挑战
    - 更低的算法开销
    - 具体的工程应用
    - 完善的智能体系
    - 内置的模型设计

- 报告价值
  - 一类评价指标或补充实验
  - 下游任务的指导思想
  - 智能模型的调优方法



参考文献 学术论文

[1] Gawlikowski J, Tassi C R N, Ali M, et al. A survey of uncertainty in deep neural networks[J]. Artificial Intelligence Review, 2023: 1-77.

[2] Berry L, Meger D. Normalizing Flow Ensembles for Rich Aleatoric and Epistemic Uncertainty Modeling. Proceedings of the AAAI Conference on Artificial Intelligence[C]. Washington, DC, USA: AAAI, 2023: 6806-6814.

[3] Mukhoti J, Kirsch A, van Amersfoort J, et al. Deep deterministic uncertainty: A simple baseline. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)[C]. Vancouver, BC, Canada: IEEE, 2023: 24384-24394.

[4] Huseljic D, Sick B, Herde M, et al. Separation of aleatoric and epistemic uncertainty in deterministic deep neural networks. 25th International Conference on Pattern Recognition (ICPR)[C]. Milano, Italy: IEEE, 2021: 9172-9179.

参考文献 补充材料

[1] https://blog.csdn.net/xys430381_1/article/details/119531335.（不确定性相关概念）

[2] https://zhuanlan.zhihu.com/p/457193790.（主动学习）

[3] https://baijiahao.baidu.com/s?id=1716146589301072672&wfr=spider&for=pc.（OOD检测）

[4] https://zhuanlan.zhihu.com/p/44304684.（归一化流）

[5] https://baijiahao.baidu.com/s?id=1780418468042486034&wfr=spider&for=pc.（科协2023）

[6] https://zhuanlan.zhihu.com/p/520107941.（利普希茨条件）

[7] https://blog.csdn.net/StreamRock/article/details/83590347.（谱归一化）

[8] https://blog.csdn.net/qq_36484003/article/details/109094107.（不确定性量化计算）

[9] https://blog.csdn.net/qq_42718887/article/details/113695473.（ECE）

[10] https://zhuanlan.zhihu.com/p/610248161.（神经网络初始化）

[11] https://blog.csdn.net/djfjkj52/article/details/130559019.（两类不确定性的区别）

[12] https://zhuanlan.zhihu.com/p/507776434?utm_id=0.（变分推断）

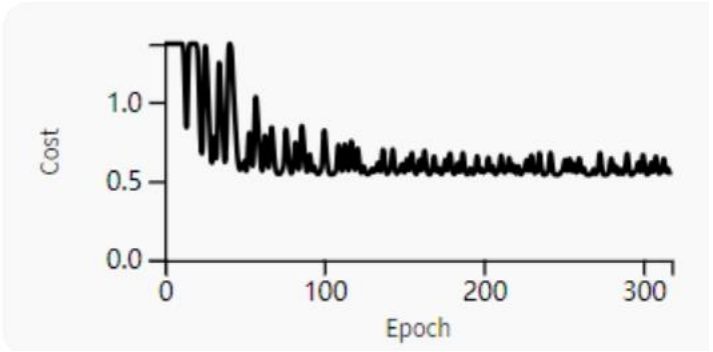
道德经

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！

附录A 神经网络参数先验初始化的方法

- 初始化方法
  - 预训练初始化（预训练模型）
  - 基于分布的初始化
    - 均匀分布、高斯分布、狄利克雷分布……
  - 预定义初始化
    - 随机初始化、零初始化
    - Xavier初始化：通过网络时，令输入和输出的方差相同【适用于sigmoid或tanh激活函数】
    - He初始化：考虑到神经元的激活性质，在Xavier基础上将输入方差除以2【适用于ReLU】
    - 正交初始化：通过保持权重矩阵的正交性来初始化参数
  - 基于规则的初始化（特定条件约束）
    - 例如，在循环神经网络（RNN）中，可以使用单位矩阵作为循环层的初始权重
    - ……



初始化导致的损失震荡

附录B 变分推断的数学证明

- 贝叶斯公式

$$p(z|x) = \frac{p(x,z)}{p(x)} = \frac{p(x,z)}{\int p(x,z)dz}$$

- 由于真实后验难以计算，将推断问题转换为一个泛函优化问题，利用简单的变分分布近似拟合真实后验，通过最小化反向KL散度求解

$$\begin{aligned}\min_{\theta} KL(q(z|x,\theta)|p(z|x,\phi)) &= \int_z q(z|x,\theta)\log\frac{q(z|x,\theta)}{p(z|x,\phi)}dz \\ &= E_{z\sim q(z|x,\theta)}[\log\frac{q(z|x,\theta)}{p(z|x,\phi)}]\end{aligned}$$

附录B 变分推断的数学证明

- 由于无法直接求解，利用证据下界ELBO进行转化

$$
\begin{aligned}
& KL(q(z|x,\theta)|p(z|x,\phi)) \\
=& E_{z\sim q(z|x,\theta)}[\log \frac{q(z|x,\theta)}{p(z|x,\phi)}] \\
=& E_{z\sim q(z|x,\theta)}[\log \frac{q(z|x,\theta)p(x|\phi)}{p(z,x|\phi)}], \text{根据} p(z|x,\phi) = \frac{p(z,x|\phi)}{p(x|\phi)} \\
=& E_{z\sim q(z|x,\theta)}[\log \frac{q(z|x,\theta)}{p(z,x|\phi)}] + E_{z\sim q(z|x,\theta)}[\log p(x|\phi)] \\
=& -\mathcal{L} + E_{z\sim q(z|x,\theta)}[\log p(x|\phi)] \\
=& -\mathcal{L} + \log p(x|\phi)
\end{aligned}
$$

附录B 变分推断的数学证明

- 证据下界$\mathcal{L}$

$$\begin{aligned}\mathcal{L} &= E_{z\sim q(z|x,\theta)}[-\log\frac{q(z|x,\theta)}{p(x|z,\phi)p(z)}] \\ &= E_{z\sim q(z|x,\theta)}[-\log\frac{q(z|x,\theta)}{p(z)} + \log p(x|z,\phi)] \\ &= E_{z\sim q(z|x,\theta)}[\log p(x|z,\phi)] - E_{z\sim q(z|x,\theta)}[\log\frac{q(z|x,\theta)}{p(z)}] \\ &= E_{z\sim q(z|x,\theta)}[\log p(x|z,\phi)] - KL(q(z|x,\theta)||p(z))\end{aligned}$$

- 最终求解

$$\max_{\theta}\mathcal{L} = \max_{\theta} E_{z\sim q(z|x,\theta)}[\log p(x|z,\phi)] - KL(q(z,\theta)||p(z))$$

简单分布近似复杂分布 → 最小化KL散度 → 最大化证据下界ELBO

附录C CIFAR-10-C的损坏类型描述

- CIFAR-10-C的损坏类型
  - 亮度强度
  - 对比度强度
  - 散焦模糊强度
  - 弹性变形强度
  - 雾强度
  - 霜冻强度
  - 高斯模糊强度
  - 高斯噪声强度
  - 玻璃模糊强度
  - 脉冲噪声强度
  - 图像压缩强度
  - 运动模糊强度
  - 像素强度
  - 饱和度
  - 散粒噪声强度
  - 降雪强度
  - 飞溅强度
  - 散斑噪声强度
  - 变焦模糊强度

附录D 中国科协2023重大科学问题难题

- 前沿科学问题（10个）
  - 如何实现低能耗人工智能
  - 如何实现飞行器在上层大气层机动飞行
  - 利用新型符合测量方式能否搜寻磁单极子和轴子暗物质的存在
  - 非线性效应会随尺度变化吗
  - 影响高性能纤维发展的基础科学问题是什么
  - 全球气候变化背景下作物如何适应土壤环境
  - 现代陆地生态系统是如何起源的
  - 生殖衰老的触发及延迟机制是什么
  - 如何实现可控核聚变的稳态燃烧
  - 如何探明更高速度轮轨系统耦合机理及能量场分布特征

其余9个工程技术难题和10个产业技术问题见参考资料[5]