

基于高维特征概率密度建模的模型不确定性的研究

2024 年 12 月 13 日

1 基于梯度空间分析和输入扰动的策略

1.1 GradNorm 的实验结果

ood dataset	auroc(\uparrow)	auprc (\uparrow)
svhn	0.9480	0.9613
lsun	0.9287	0.9444
cifar100	0.9058	0.9242
mnist	0.9665	0.9770
svhn+ip	0.9493	0.9599
lsun+ip	0.8990	0.9137
cifar100+ip	0.8871	0.9027
mnist+ip	0.9481	0.9550

表 1: resnet50+cafar10,accuracy=0.9489, gradNorm

ood dataset	auroc(\uparrow)	auprc (\uparrow)
svhn	0.9293	0.9516
lsun	0.9436	0.9563
cifar100	0.9426	0.9480
mnist	0.9550	0.9680
svhn+ip	0.9467	0.9633
lsun+ip	0.9175	0.9469
cifar100+ip	0.9086	0.9317
mnist+ip	0.9113	0.9286

表 2: vit+cafar10,accuracy=0.9600, gradNorm

1.2 Cifar 加入输入扰动前后对比

ood dataset	auroc(\uparrow)	auprc(\uparrow)
svhn	0.9221	0.9432
lsun	0.9363	0.9528
cifar100	0.8861	0.9028
mnist	0.9189	0.9369
tiny-imagenet	0.9318	0.9479
svhn+ip	0.9216	0.9411
lsun+ip	0.9394	0.9533
cifar100+ip	0.8885	0.9009
mnist+ip	0.9637	0.9666
tiny-imagenet+ip	0.9397	0.9493

表 3: vgg16+cafar10,accuracy=0.9405

ood dataset	auroc(\uparrow)	auprc (\uparrow)
svhn	0.9480	0.9613
lsun	0.9365	0.9497
cifar100	0.9068	0.9257
mnist	0.9774	0.9845
tiny-imagenet	0.9469	0.9580
svhn+ip	0.9735	0.9769
lsun+ip	0.9671	0.9716
cifar100+ip	0.9171	0.9372
mnist+ip	0.9939	0.9957
tiny-imagenet+ip	0.9676	0.9731

表 4: resnet50+cafar10,D=2048,accuracy=0.9489

ood dataset	auroc(\uparrow)	auprc(\uparrow)
svhn	0.9685	0.9768
lsun	0.9720	0.9788
cifar100	0.9400	0.9514
mnist	0.9906	0.9929
tiny-imagenet	0.9747	0.9803
svhn+ip	0.9891	0.9904
lsun+ip	0.9811	0.9830
cifar100+ip	0.9497	0.9585
mnist+ip	0.9978	0.9982
tiny-imagenet+ip	0.9836	0.9861

表 5: wideResnet+cafar10,accuracy=0.9650

ood dataset	auroc(\uparrow)	auprc(\uparrow)
svhn	0.9293	0.9516
lsun	0.9436	0.9563
cifar100	0.9426	0.9480
mnist	0.9550	0.9680
tiny-imagenet	0.9330	0.9367
svhn+ip	0.9728	0.9793
lsun+ip	0.9688	0.9740
cifar100+ip	0.9463	0.9495
mnist+ip	0.9904	0.9926
tiny-imagenet+ip	0.9546	0.9545

表 6: vit+cafar10,accuracy=0.9600

1.3 几种方法的对比

Method	OOD Dataset	AUROC(\uparrow)	AUROC(\uparrow)
baseline	cifar100	0.8786	0.8774
	lsun	0.8952	0.8997
	mnist	0.9434	0.9524
	svhn	0.9271	0.9405
ensemble	cifar100	0.9244	0.9166
	lsun	0.9724	0.9685
	mnist	0.9735	0.9706
	svhn	0.9667	0.9599
ddu	cifar100	0.9068	0.9257
	lsun	0.9365	0.9497
	mnist	0.9774	0.9845
	svhn	0.9480	0.9613
ddu+ip	cifar100	0.9171	0.9312
	lsun	0.9671	0.9716
	mnist	0.9939	0.9957
	svhn	0.9735	0.9769

resnet50+cifar10,D=2048,accuracy=0.9540

1.4 mnist 加入输入扰动前后对比

ood dataset	auroc(\uparrow)	auprc(\uparrow)
cifar10	0.9949	0.9964
fashionmnist	0.9914	0.9933
fer2013	0.9952	0.9969
lsun	0.9946	0.9962
cifar100	0.9945	0.9961
svhn	0.9941	0.9960
cifar10+ip	0.9965	0.9974
fashionmnist+ip	0.9921	0.9936
fer2013+ip	0.9970	0.9978
svhn+ip	0.9963	0.9972
cifar10+ip	0.9961	0.9970
svhn+ip	0.9964	0.9973

表 8: vgg16+mnist,accuracy=0.9882

ood dataset	auroc(\uparrow)	auprc(\uparrow)
cifar10	0.9949	0.9964
fashionmnist	0.9914	0.9933
fer2013	0.9952	0.9969
lsun	0.9946	0.9962
cifar100	0.9945	0.9961
svhn	0.9941	0.9960
cifar10+ip	0.9965	0.9974
fashionmnist+ip	0.9921	0.9936
fer2013+ip	0.9970	0.9978
svhn+ip	0.9963	0.9972
cifar10+ip	0.9961	0.9970
svhn+ip	0.9964	0.9973

表 9: resnet50+mnist,accuracy=0.9870

1.5 GMM vs KDE 实验结果

ood dataset	auroc(\uparrow)	auprc(\uparrow)
svhn+gmm	0.9458	0.9564
lsun+gmm	0.9287	0.9444
cifar100+gmm	0.9058	0.9242
mnist +gmm	0.9665	0.9770
tiny- imagenet+gmm	0.8932	0.9112
svhn+kde	0.7924	0.8365
lsun+kde	0.7894	0.8264
cifar100+kde	0.7772	0.8097
mnist+kde	0.8364	0.8684
tiny- imagenet+kde	0.7653	0.7950

表 10: resnet50+cafar10,GMM vs KDE ,Dimension=2048,accuracy=0.9489

ood dataset	auroc(\uparrow)	auprc(\uparrow)
svhn+gmm	0.9320	0.9530
lsun+gmm	0.9302	0.9471
cifar100+gmm	0.9014	0.9156
mnist +gmm	0.9358	0.9563
tiny- imagenet+gmm	0.9026	0.9162
svhn+kde	0.9528	0.9620
lsun+kde	0.9179	0.9299
cifar100+kde	0.8923	0.9022
mnist+kde	0.9654	0.9718
tiny- imagenet+kde	0.8975	0.8975

表 11: resnet50+cafar10,GMM vs KDE ,Dimension=1024,accuracy=0.9504

ood dataset	auroc(\uparrow)	auprc(\uparrow)
svhn+gmm	0.9364	0.9553
lsun+gmm	0.9322	0.9474
cifar100+gmm	0.9028	0.9156
mnist +gmm	0.9378	0.9567
tiny- imagenet+gmm	0.9038	0.9133
svhn+kde	0.9484	0.9584
lsun+kde	0.9211	0.9312
cifar100+kde	0.8957	0.9019
mnist+kde	0.9662	0.9732
tiny- imagenet+kde	0.9018	0.8992

表 12: resnet50+cafar10,GMM vs KDE ,Dimension=512,accuracy=0.9520

ood dataset	auroc(\uparrow)	auprc(\uparrow)
svhn+gmm	0.9418	0.9594
lsun+gmm	0.9277	0.9417
cifar100+gmm	0.8983	0.9073
mnist +gmm	0.9450	0.9628
tiny- imagenet+gmm	0.9013	0.9066
svhn+kde	0.9569	0.9659
lsun+kde	0.9230	0.9399
cifar100+kde	0.8992	0.9090
mnist+kde	0.9730	0.9790
tiny- imagenet+kde	0.9097	0.9099

表 13: resnet50+cafar10,GMM vs KDE ,Dimension=256,accuracy=0.9540

ood dataset	auroc(\uparrow)	auprc(\uparrow)
svhn+gmm	0.9222	0.9433
lsun+gmm	0.9077	0.9204
cifar100+gmm	0.8848	0.9018
mnist +gmm	0.9141	0.9308
tiny- imagenet+gmm	0.8807	0.8847
svhn+kde	0.9230	0.9362
lsun+kde	0.9047	0.9137
cifar100+kde	0.8778	0.8845
mnist+kde	0.9377	0.9456
tiny- imagenet+kde	0.8750	0.8643

表 14: vgg16+cafar10,GMM vs KDE ,Dimension=512,accuracy=0.9489

1.6 扰动前后做差

ood dataset	auroc(\uparrow)	auprc (\uparrow)
svhn	0.9221	0.9432
lsun	0.9078	0.9205
cifar100	0.8848	0.9018
mnist	0.9140	0.9308
tiny-imagenet	0.8807	0.8847
svhn+ip	0.8884	0.8952
lsun+ip	0.8400	0.8488
cifar100+ip	0.8603	0.8606
mnist+ip	0.8715	0.8757
tiny-imagenet+ip	0.8549	0.8637

表 15: vgg16+cafar10, 扰动前概率密度 vs (扰动后概率密度-扰动前概率密度),accuracy=0.9407

1.7 关于高维特征维度的实验结果

ood dataset	auroc(\uparrow)	auprc (\uparrow)
svhn	0.9480	0.9613
lsun	0.9365	0.9497
cifar100	0.9068	0.9257
mnist	0.9774	0.9845
tiny-imagenet	0.9469	0.9580
svhn+ip	0.9735	0.9769
lsun+ip	0.9671	0.9716
cifar100+ip	0.9171	0.9372
mnist+ip	0.9939	0.9957
tiny-imagenet+ip	0.9676	0.9731

表 16: resnet50+cafar10,D=2048,accuracy=0.9489

ood dataset	auroc(\uparrow)	auprc (\uparrow)
svhn	0.9303	0.9528
lsun	0.9325	0.9515
cifar100	0.9085	0.9279
mnist	0.9358	0.9564
tiny-imagenet	0.9028	0.9167
svhn+ip	0.9777	0.9822
lsun+ip	0.9544	0.9647
cifar100+ip	0.9203	0.9343
mnist+ip	0.9837	0.9885
tiny-imagenet+ip	0.9150	0.9249

表 17: resnet50+cafar10,D=1024,accuracy=0.9504

ood dataset	auroc(\uparrow)	auprc (\uparrow)
svhn	0.9343	0.9454
lsun	0.9334	0.9501
cifar100	0.9094	0.9272
mnist	0.9387	0.9582
tiny-imagenet	0.9049	0.9146
svhn+ip	0.9786	0.9828
lsun+ip	0.9535	0.9632
cifar100+ip	0.9191	0.9301
mnist+ip	0.9840	0.9887
tiny-imagenet+ip	0.9134	0.9193

表 18: resnet50+cafar10,D=512,accuracy=0.9520

ood dataset	auroc(\uparrow)	auprc (\uparrow)
svhn	0.9418	0.9596
lsun	0.9307	0.9470
cifar100	0.9081	0.9231
mnist	0.9451	0.9629
tiny-imagenet	0.9026	0.9091
svhn+ip	0.9763	0.9819
lsun+ip	0.9434	0.9534
cifar100+ip	0.9098	0.9201
mnist+ip	0.9832	0.9886
tiny-imagenet+ip	0.9056	0.9070

表 19: resnet50+cafar10,D=256,accuracy=0.9540

1.8 各种方法的对比

Method	OOD Dataset	AUROC(\uparrow)	AUROC(\uparrow)
baseline	cifar100	0.8786	0.8774
	lsun	0.8952	0.8997
	mnist	0.9434	0.9524
	svhn	0.9271	0.9405
ensemble	cifar100	0.9244	0.9166
	lsun	0.9724	0.9685
	mnist	0.9735	0.9706
	svhn	0.9667	0.9599
ddu	cifar100	0.9068	0.9257
	lsun	0.9365	0.9497
	mnist	0.9774	0.9845
	svhn	0.9480	0.9613
ddu+ip	cifar100	0.9171	0.9312
	lsun	0.9671	0.9716
	mnist	0.9939	0.9957
	svhn	0.9735	0.9769

resnet50+cifar10,D=2048,accuracy=0.9540

1.9 对抗样本检测

Input		Result	
Method	Attack Method	AUROC(↑)	AUROC(↑)
DDU	FGSM	0.8428	0.8394
	BIM	0.8070	0.8816
	PGD	0.7420	0.8021
GMM+Input Perturbation	FGSM	0.8882	0.9737
	BIM	0.8559	0.9686
	PGD	0.8433	0.9700

VGG16 + CIFAR10, 对抗样本的检测任务

2 辅助 loss

2.1 加入辅助 Loss 前后对比

实验设置		实验结果	
训练策略	OOD Dataset	AUROC(↑)	AUPRC (↑)
CE loss	svhn	0.9195	0.9412
	lsun	0.9144	0.9317
	cifar100	0.8865	0.8966
	mnist	0.9240	0.9426
CE loss +ContrastiveCenterLoss	svhn	0.9332	0.9517
	lsun	0.9221	0.9270
	cifar100	0.8952	0.9065
	mnist	0.9395	0.9579

实验设置：Vgg16+cafar10,D=512,accuracy=0.9438

实验设置		实验结果	
训练策略	OOD Dataset	AUROC(↑)	AUPRC (↑)
CE loss	svhn	0.9260	0.9437
	lsun	0.9290	0.9445
	cifar100	0.9036	0.9124
	mnist	0.9176	0.9268
CE loss + ContrastiveCenterLoss	svhn	0.9410	0.9554
	lsun	0.9329	0.9415
	cifar100	0.9115	0.9177
	mnist	0.9501	0.9636

实验设置：ResNet18+cafar10,D=512,accuracy=0.9438

实验设置		实验结果	
训练策略	OOD Dataset	AUROC(↑)	AUPRC (↑)
CE loss	svhn	0.9007	0.9008
	lsun	0.9727	0.9716
	cifar100	0.8960	0.8966
	mnist	0.8913	0.9123
CE loss + ContrastiveCenterLoss	svhn	0.9874	0.9910
	lsun	0.9891	0.9916
	cifar100	0.9668	0.9704
	mnist	0.9930	0.9939

实验设置：VIT+cafar10,accuracy=0.9780

3 其他

3.1 一些公式

Input Perturbation:

添加输入扰动的方式如下:

$$\tilde{x} = x + \epsilon * \text{sign}(\nabla_x U(x))$$

GMM 的公式其中:

$$U(x) = \log p(x)$$
$$p(x) = \max_c N(x|\mu_c, \Sigma_c) = \max_c \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma_c|^{1/2}} \exp \left\{ -\frac{1}{2}(x - \mu_c)^T \Sigma_c^{-1} (x - \mu_c) \right\}$$

超参数 ϵ 通过 gridSearch 得到.

GMM 的公式:

$$N(x|\mu_c, \Sigma_c) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma_c|^{1/2}} \exp \left\{ -\frac{1}{2}(x - \mu_c)^T \Sigma_c^{-1} (x - \mu_c) \right\}$$
$$p(x) = \sum_{c=1}^K w_c N(x|\mu_c, \Sigma_c)$$

$$\tilde{x} = x - \epsilon \cdot \text{sign}(-\nabla_x \log p(x; D))$$

a

centerloss 的公式:

$$L_c = \frac{1}{2} \sum_{i=1}^m \|x_i - x_{y_i}\|_2^2$$

$L_s : CrossEntropyLoss$

$$L = L_s + L_c$$

修正后的 centerloss:

$$L_c = \frac{1}{2} \sum_{i=1}^m \frac{\|x_i - x_{y_i}\|_2^2}{\sum_{j=1, j \neq y_i}^m \|x_i - x_j\|_2^2 + \delta}$$

3.2 算法流程图

Algorithm 1 基于输入扰动的概率密度建模的模型不确定性算法

Require: 训练集: (X, Y)

Require: 谱归一化的高维特征提取网络: $f_\theta : x \rightarrow \mathbf{R}^d$

Require: GMM 模型: $q(z) = \sum_y q(z|y=c)q(y=c)$

1:

2: **procedure** 1. 训练阶段

3: 在训练数据集数据集上训练网络 f_θ

4: **for** 属于类别 c 的样本 **do**

5: $\mu_c = \frac{1}{|x_c|} f_\theta(x_c)$

6: $\Sigma_c = \frac{1}{|x_c|-1} (f_\theta(x_c) - \mu_c)(f_\theta(x_c) - \mu_c)^T$

7: $q(y=c) = \frac{|X_C|}{|X|}$

8: **end for**

9: **end procedure**

10:

11: **procedure** 2. 模型不确定性计算

12: $z = f_\theta(x)$

13: $q(z) = \sum_y q(z|y=c)q(y=c)$, 其中 $q(z) \sim N(\mu_c, \Sigma_c)$

14: $\tilde{x} = x + \epsilon * \text{sign}(\nabla_x \log q(z))$, 其中 ϵ 通过 grid search 调参

15: $\tilde{z} = f_\theta(\tilde{x})$

16: 计算模型不确定性 $Uncertainty(x) = \sum_y q(\tilde{z}|y=c)q(y=c)$, 其中
 $q(\tilde{z}) \sim N(\mu_c, \Sigma_c)$

17: **end procedure**

3.3 应用

神经网络的不确定性研究 (Uncertainty in Neural Networks) 在许多应用场景中都具有重要价值，特别是在需要对预测结果的可靠性、稳定性和信心进行量化的任务中。传统的神经网络通常给出一个确定的预测结果（例如，分类标签或回归值），但并没有提供关于该结果的不确定性信息。而不确定性建模则能够量化模型对其预测的信心程度，这在多个领域中具有广泛的应用。

1. 自动驾驶在自动驾驶系统中，车辆必须能够做出实时决策，基于感知模型（如卷积神经网络处理的摄像头图像、雷达、激光雷达等数据）预测和识别路况。然而，感知数据本身可能存在噪声，或者在复杂环境中不完全可靠。通过不确定性建模，自动驾驶系统可以识别哪些预测是可靠的，哪些可能需要进一步验证或修正。

- 应用：- 确定对物体检测和追踪的信心程度，如识别行人、车辆和障碍物。- 判断模型在某些环境下的可靠性（例如，低光照或恶劣天气条件下的驾驶场景）。- 决策制定时结合不确定性，优先选择对结果不确定性较低的路径规划或行动。

2. 医学影像分析在医学影像分析中，如癌症检测或器官分割，模型不仅需要做出准确的预测，还需要能够提供关于其预测结果不确定性的估计。这对于医生和临床决策者至关重要，尤其是在面对复杂或模糊的医学影像时。

- 应用：- 辅助诊断：为医生提供关于某个影像分析结果（如肿瘤存在与否）的置信度，帮助医生判断是否需要进一步的检查或其他医学手段。- 图像分割：在医学图像分割任务中，某些边界可能很模糊。通过不确定性估计，模型可以指出分割结果的可靠区域，辅助医生进行更准确的解读。- 疾病预测：基于患者历史数据和医学影像，模型预测某种疾病的风险，并通过不确定性量化预测的可信度。

3. 金融风控与预测在金融行业中，模型预测的准确性对投资决策至关重要。不确定性量化有助于提高风险管理和预测系统的可靠性，尤其是在面对市场波动和不确定环境时。

- 应用：- 信贷风险评估：通过不确定性量化，金融机构能够识别贷款申请中哪些因素对违约预测有较高不确定性，从而采取更谨慎的风险管理措施。- 市场预测：量化市场趋势预测中的不确定性，帮助投资者做出更明智的决策，尤其在市场波动较大的情况下。- 欺诈检测：在金融欺诈检测中，模型可以识别哪些交易行为具有更高的不确定性，从而引起额外的审查。

4. 自然语言处理 (NLP) 在 NLP 任务中, 神经网络通常需要处理复杂的文本和语言任务, 不仅要生成准确的输出, 还要提供预测的不确定性。例如, 在机器翻译、情感分析或问答系统中, 了解模型对结果的信心度对于理解其输出的可信度非常重要。

- 应用: - 机器翻译: 模型可以提供翻译的信心水平, 帮助用户理解哪些翻译结果可靠, 哪些可能需要进一步验证。- 情感分析: 评估情感分析模型的预测可信度, 帮助商家了解用户评论中的潜在情感, 尤其在情感极为复杂或模糊的情况下。- 问答系统: 在自动问答系统中, 不确定性建模可以提供关于某个答案的置信度, 帮助系统在遇到不确定的答案时请求进一步的信息或向用户确认。

5. 机器人与自主系统机器人和自主系统在复杂和动态的环境中执行任务时, 不确定性建模至关重要。通过量化对环境状态或控制决策的信心, 机器人可以决定是否需要进行进一步的信息, 或者采取更保守的行动。

- 应用: - 路径规划与控制: 通过不确定性量化, 机器人可以识别哪些路径或动作是更安全的, 哪些决策可能会导致更高的风险。- 环境感知: 在复杂环境中 (例如多变的室内环境或户外环境), 机器人可以基于不确定性评估来选择更可靠的传感器数据进行决策。

6. 科学模拟与预测在许多科学领域, 如气候建模、流体动力学、天文学等, 不确定性建模帮助科学家理解模型预测的信心程度, 从而做出更可靠的预测或决定。

- 应用: - 气候变化预测: 通过不确定性估计, 科学家能够量化不同气候预测模型的可靠性, 从而做出更加谨慎的决策。- 流体动力学: 在复杂的流体动力学模拟中, 通过不确定性量化, 工程师可以了解哪些模型预测结果更为可靠, 哪些可能受初始条件或假设的影响较大。- 天体物理学: 在天文学的观测数据分析中, 不确定性建模有助于研究人员理解数据的不确定性来源, 提高天文观测结果的准确性。

7. 机器学习中的模型选择与调优在机器学习中, 尤其是在深度学习和强化学习的应用中, 评估模型的不确定性可以帮助选择更稳定的模型, 或者帮助调整模型的训练过程, 减少过拟合或提高模型的泛化能力。

- 应用: - 超参数优化: 通过量化不同超参数配置下的模型不确定性, 帮助选择更适合特定任务的配置。- 模型选择: 在多模型决策中, 通过评估每个模型的预测不确定性, 选择最适合的模型或集成模型。

不确定性建模方法: 1. 贝叶斯神经网络: 通过在神经网络中引入贝叶

斯推断，建模参数的不确定性，从而获得模型的输出不确定性。2. 蒙特卡洛 dropout：通过使用 dropout 技术在推理阶段进行多次前向传播，得到输出的不确定性估计。3. 深度高斯过程：通过高斯过程建模神经网络的输出，不仅可以得到预测结果，还可以量化其不确定性。4. 生成对抗网络（GANs）：通过生成模型与判别模型之间的对抗训练，可以间接量化模型的不确定性，尤其在数据缺失或噪声大的情况下。

总结：神经网络不确定性研究为许多领域提供了更高层次的可靠性分析，帮助模型在实际应用中更好地应对复杂的环境和决策任务。通过量化不确定性，模型不仅能够提供准确的预测，还能够提供可信度评估，从而更好地支持决策过程，尤其在安全性、精确性要求高的任务中，具有不可替代的应用价值。