The Hidden Uncertainty in a Neural Network Activations

师清

FDU

2022 年 12 月 29 日

Overview

摘要

- epistemic uncertainty can be identified with negative log-likelihood of observing a particular latent representation
- the output-conditional distribution of hidden representations also allows quantifying aleatoric uncertainty via the entropy of the predictive distribution
- an additional regularising loss that increases the information in the latent representations
- shallow layers yield more conservative epistemic uncertainty;the density of deeper layers behaves less conservatively and more similar to estiablished methods

回顾

- 两种不确定性:
  - epistemic uncertainty:arising from the model choice and parameter fitting
  - aleatoric uncertainty:arising from noise in the data

回顾

- **两种不确定性**:
  - epistemic uncertainty:arising from the model choice and parameter fitting
  - aleatoric uncertainty:arising from noise in the data
- **一些方法**
  - Bayesian Deep Learning:MC Dropout
  - deep Ensembles

不确定性的估计

考虑 L 层神经网络，输入为 x, 输出为 $\hat{y}$, 有 L-1 个隐藏层 $(z_0, z_1, ..., z_{L-2})$，联合概率分解

$$p_\theta(x, \hat{y}, z_0, ..., z_{L-2}) = p_\theta(\hat{y}|z_{L-2})p_\theta(z_{L-2}|z_{L-3})...p_\theta(z_0|x)p(x)$$

条件熵:

$$\begin{aligned} H(\hat{y}|x) &= E_{p_\theta(\hat{y},x)}\left[-logp_\theta(\hat{y}|x)\right] \\ &= E_{p(x)}\left[-\int p_\theta(\hat{y}|x)logp_\theta(\hat{y}|x)d\hat{y}\right] \end{aligned}$$

联合熵:

$$\begin{aligned} H(\hat{y}, x) &= H(x) + H(\hat{y}|x) \\ &= E_{p(x)}\left[-logp(x) - \int p_\theta(\hat{y}|x)logp_\theta(\hat{y}|x)d\hat{y}\right] \end{aligned}$$

$$H(\hat{y}|z_i) = H(\hat{y}|x), i \in [0, ..., L-2]$$

$$H(x) \geq H(z_0) \geq ... \geq H(z_{L-2})$$

不确定性的估计

- epistemic uncertainty

$$-logp(z_i^*) = -log\left(\int p(z_i^*|\hat{y})p(\hat{y})d\hat{y}\right)$$

不确定性的估计

- epistemic uncertainty

$$-logp(z_i^*) = -log\left(\int p(z_i^*|\hat{y})p(\hat{y})d\hat{y}\right)$$

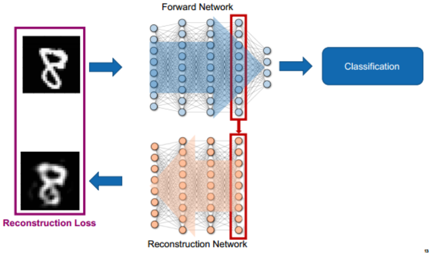
- aleatoric uncertainty

$$h(\hat{y}|z_i^*) = -\int p_\theta(\hat{y}|z_i^*)log(p_\theta(\hat{y}|z_i^*))$$

不确定性的估计

- Feature Collapse
  在原始 loss 函数上加一个重建损失，以得到 informative 的隐藏层表示

$$\hat{L} = L_{orig} + \lambda MSE(x, \hat{x})$$



- High-Dimensional Densities:PCA

实现

- output-conditional density:$p(z_i|\hat{y})$
  - for classification:GMM
  - for regression:CNF(conditional normalizing flows)
- $p(\hat{y})$ 的估计
  - for classification: 在训练集上统计预测值
  - for regression: 使用参数化分布近似，如 uniform 分布或者 Beta-prime 分布

实现

Algorithm 1 Estimating the output-conditional distribution of hidden representations.

Input: Dataset $\{X, Y\}$, Index l of layer used for uncertainty estimation, d maximum dimension of hidden representations.

Result: Trained model M, output-conditional density model M_{gen} of hidden representations at layer l, estimate of P($\hat{Y}$)

Train model M on $\{X, Y\}$

$Z_l \leftarrow$ activations at layer l on $\{X, Y\}$

$\hat{Y} \leftarrow$ M(X)

if $dim(Z_l) > d$ **then**

  $Z_l \leftarrow$ reduce dimension of Z_l using PCA

end

Initialize generative model M_{gen}

Train M_{gen} on Z_l given $\hat{Y}$ to estimate P($Z_l|\hat{Y}$)

if *classification* **then**

  Estimate marginal categorical distribution of network predictions P($\hat{Y}$) by counting frequencies

else

  Estimate marginal distribution of network predictions P($\hat{Y}$) with univariate parametric distribution (e.g. Gaussian, beta prime)

end

实现

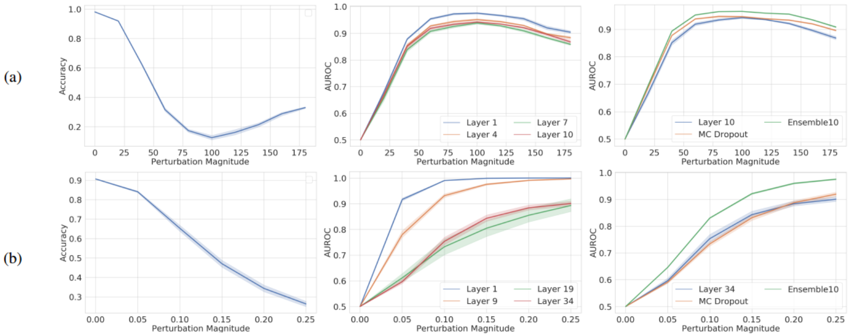
- Normalizing Flows

$$p_X(x) = p_Z(f_\theta(x))|det\left(\frac{\partial f_\theta(x)}{\partial x}\right)|$$

- coupling layers :

$$u_1^{out} = u_1^{in}$$
$$u_2^{out} = (u_2^{in} + g_t(u_1^{in}; c)) \odot g_s(u_1^{in}; c)$$

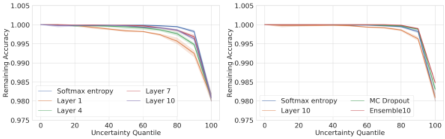
实验结果



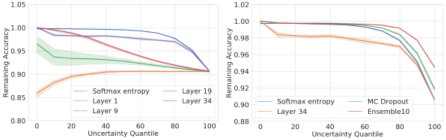
Classification accuracy. Comparison of different layers. Comparison with other methods.

- density estimates based on shallow layers yield more conservative estimates.
- density estimates based on deeper layers behave similar as other established methods

实验结果



Aleatoric uncertainty on MNIST



Aleatoric uncertainty on SVHN

- aleatoric uncertainty based on deeper layers tends to behave similar as other approaches (softmax entropy, deep ensembles, MC dropout)
- aleatoric uncertainty based on shallow layers perform poorly

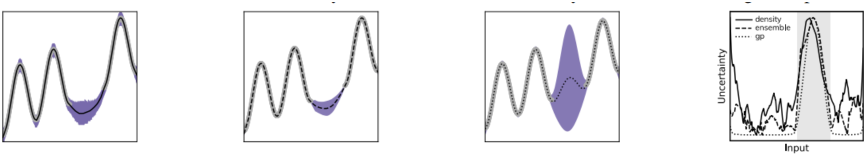
实验结果

	OOD Data	L 1	L 4	L 7	L 10	Ensemble
Trained on MNIST	FashionMNIST	**0.975**	0.922	0.855	0.811	0.896
	OMNIGLOT	0.972	0.937	0.892	0.893	**0.979**
	white noise	**1.000**	0.972	0.903	0.841	0.785
	Rotated 90°	**0.978**	0.976	0.950	0.935	0.965
	HFlip	0.902	**0.907**	0.883	0.864	0.905
	VFlip	**0.887**	0.868	0.851	0.830	0.881
Trained on FashionMNIST	MNIST	0.985	**0.991**	0.975	0.978	0.962
	OMNIGLOT	0.971	**0.987**	0.960	0.967	0.960
	white noise	**1.000**	0.985	0.971	0.930	0.840
	Rotated 90°	**0.884**	0.780	0.804	0.835	0.670
	HFlip	**0.719**	0.696	0.701	0.693	0.657
	VFlip	0.898	0.891	0.891	**0.901**	0.845

	OOD Data	L 1	L 9	L 19	L 34	Ensemble
Trained on SVHN	CIFAR10	**0.991**	0.974	0.934	0.907	0.976
	STL10	**0.999**	0.991	0.951	0.912	0.982
	white noise	**1.000**	**1.000**	0.986	0.903	0.992
	Rotated 90°	0.615	0.646	0.689	0.918	**0.957**
	HFlip	**0.500**	**0.503**	**0.495**	**0.500**	**0.501**
	VFlip	0.506	0.520	0.551	0.708	**0.736**
Trained on CIFAR10	SVHN	0.042	0.029	0.091	**0.736**	0.723
	STL10	0.790	**0.871**	0.821	0.651	0.806
	white noise	**1.000**	**1.000**	**1.000**	0.681	0.999
	Rotated 90°	0.553	0.517	0.543	0.757	**0.824**
	HFlip	**0.500**	**0.500**	**0.499**	**0.500**	**0.501**
	VFlip	0.519	0.513	0.537	0.714	**0.789**

- uncertainty estimates based on shallow layers demonstrate strong OOD performance.
- when using a convolutional architecture (ResNet18), epistemic uncertainty obtained from shallow layers fails to detect OOD data generated by globally transforming the test data.

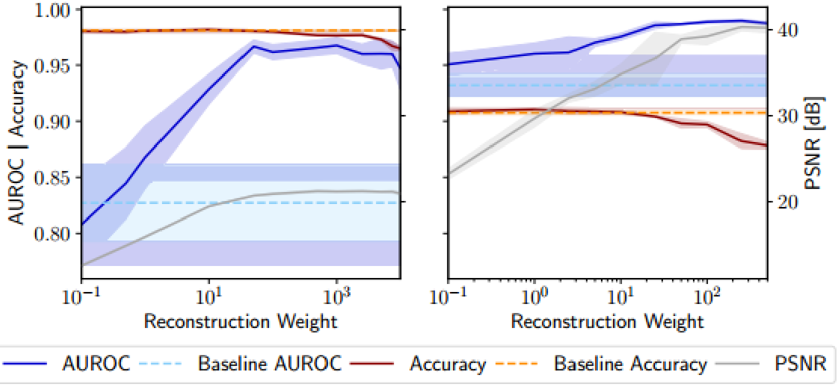
实验结果



(a) Latent Density (b) Ensemble (c) Gaussian Process (d) Epistemic Uncertainty

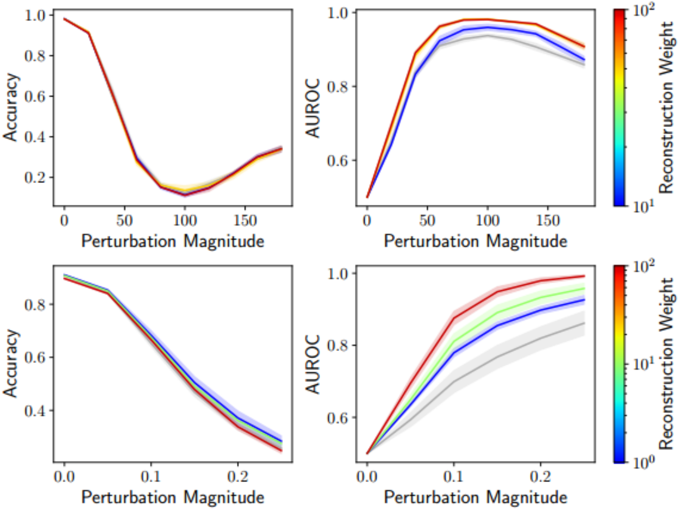
- All methods show growing uncertainty with further distance to the training data, indicating that latent densities also contain information about epistemic uncertainty in regression networks.

实验结果



- AUROC increases with the weight of the reconstruction loss and the reconstruction quality (PSNR).

实验结果



- The AUROC increases with higher reconstruction losses while the Accuracy does not deviate much from the baseline.

The End