模式识别与机器学习研究组

On Calibration of Modern Neural Networks

汇报人：1024041141郑雯月

2024/09/30

汇报提纲

01 论文研究主题(任务)及其研究意义

02 该研究主题(任务)已有方法的不足或当前面临的挑战

03 本文解决思路

04 所提出的模型/方案

05 实验设置与结果分析

06 总结与结论
启发与思考

论文研究主题（任务）及其研究意义

研究主题：

校准问题-**置信度能真实地反映模型预测正确的可能性**：

$$P\left(\widehat{Y}_i = Y_i | \widehat{P}_{iY_i} = p\right) = p, \forall p \in [0,1]$$

研究目标：希望根据$y_i, \hat{p}_i, z_i$找到校准后的输出概率$\hat{q}_i$,提供校准后的置信度。

问题定义：

对于多分类问题，训练集$D = \{x_i, y_i\}_{i=1}^{N}$，$x_i \in R^d$，$y_i \in R^1$，$x_i$是第$i$个样本特征向量，$y^i$是第$i$个样本的类别标签。

对于每个输入x_i，模型输出一个概率向量$p_i \in R^c$和预测类别标签$\hat{y}_i$，其中C是类别总数。

$\hat{y}_i \in R^1$，且$\sum_{c=1}^{C} \hat{p}_{ic} = 1$，$\hat{p}_{ic}$表示第$i$个样本属于类别$c$的预测概率。

论文研究主题（任务）及其研究意义

01

决策可靠性

在自动驾驶、医疗诊断等高风险应用中,分类网络不仅要准确,还要能够在可能错误时(置信度较低)提供指示,以便于采取其他措施。

02

模型可解释性

良好的置信度估计有助于建立用户对模型的信任,尤其是对于那些分类决策难以解释的神经网络。

03

集成概率模型

良好的概率估计可以将神经网络结合到其他概率模型中,提升性能。例如在语音识别中结合语言模型,或在目标检测中结合相机信息。

该研究主题（任务）当前面临的挑战

神经网络的过度自信

现代神经网络在分类任务中往往会出现的“**过度自信**”，即**网络对其预测结果的置信度往往高于实际准确性**。

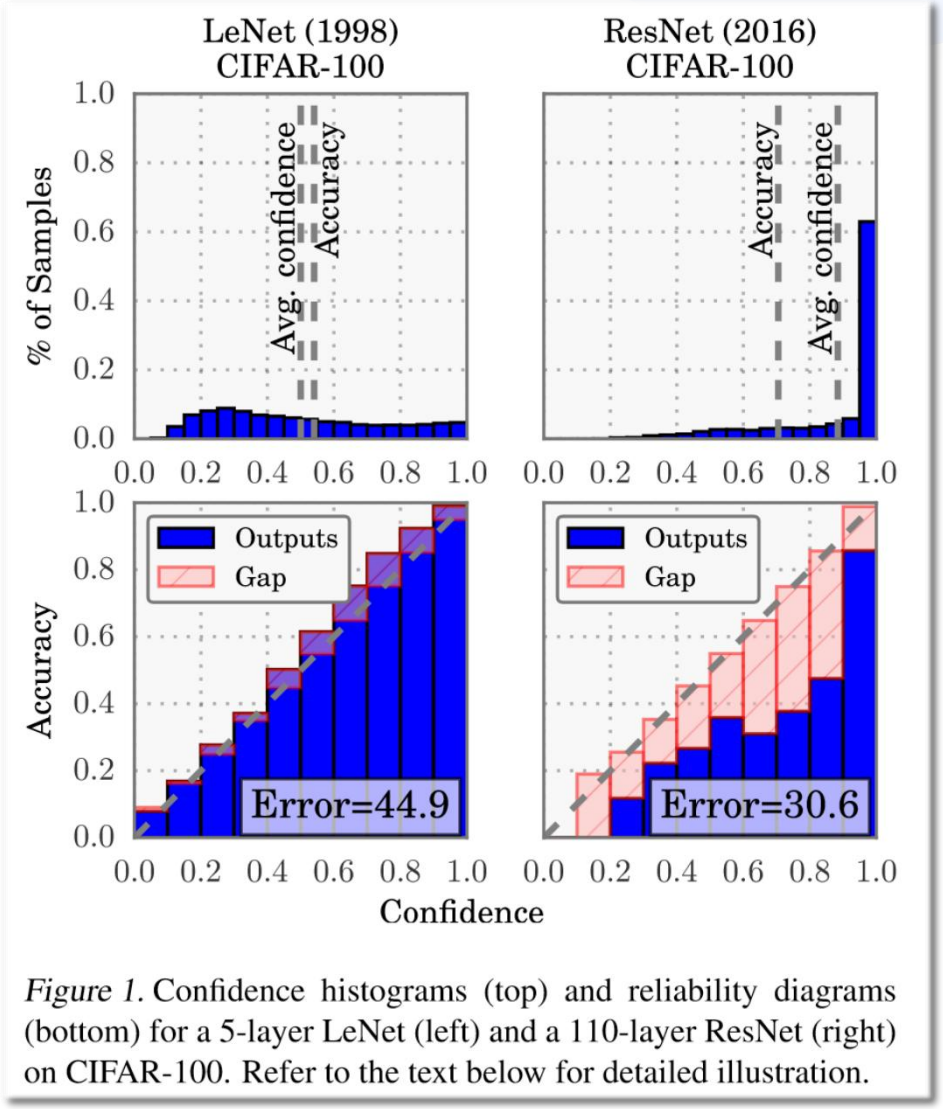


Figure 1. Confidence histograms (top) and reliability diagrams (bottom) for a 5-layer LeNet (left) and a 110-layer ResNet (right) on CIFAR-100. Refer to the text below for detailed illustration.

该研究主题（任务）当前面临的挑战

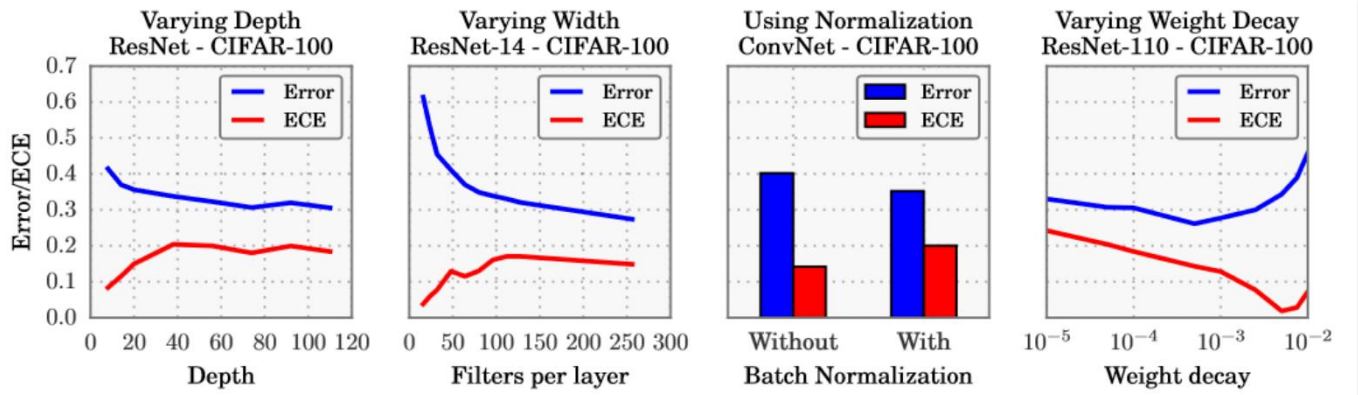


Figure 2. The effect of network depth (far left), width (middle left), Batch Normalization (middle right), and weight decay (far right) on miscalibration, as measured by ECE (lower is better).

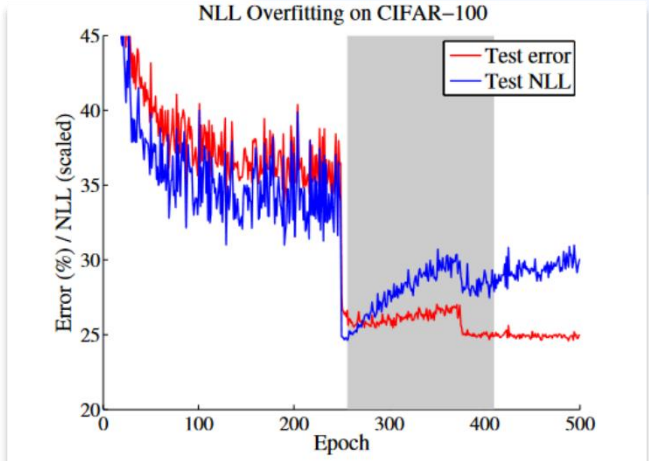


Figure 3. Test error and NLL of a 110-layer ResNet with stochastic depth on CIFAR-100 during training. NLL is scaled by a constant to fit in the figure. Learning rate drops by 10x at epochs 250 and 375. The shaded area marks between epochs at which the best validation *loss* and best validation *error* are produced.

增加模型的深度和宽度、使用Batch Normalization虽然可以**降低分类误差**，但会显著**增加模型的校准误差。**

在一定范围内**增加Weight Decay**可以**减少校准误差**，但过多的Weight Decay又会导致校准性能下降,**适当的正则化对于提高模型校准性能。**

在训练后期，模型没有提高分类正确的准确性，而是**过拟合NLL提高预测类别置信度，造成了准确率与预测置信度不匹配。**

该研究主题（任务）已有方法的不足

主要思想：调整预测类别概率分布，最小化NLL损失。

Platt Scaling（二分类）

利用逻辑回归模型将分类器的原始输出 z_i通过逻辑回归模型的sigmoid函数转换为概率：

$$\hat{q}_i = \sigma(az_i + b)$$

采用NLL损失函数：

$$loss_i = -\left(y_i \log(\hat{q}_i) + (1 - y_i)\log(1 - \hat{q}_i)\right)$$

$$L(a, b) = -\sum_{i=1}^{N}\left(y_i \log(\hat{q}_i) + (1 - y_i)\log(1 - \hat{q}_i)\right)$$

在验证集上寻找最优a, b：$(a, b) = argmin_{a,b} -\left(\sum_{i=1}^{N}\left(y_i \log(\hat{q}_i) + (1 - y_i)\log(1 - \hat{q}_i)\right)\right)$

该研究主题（任务）已有方法的不足

主要思想：调整预测类别概率分布，最小化NLL损失。

Matrix and vector scaling （多分类）

通过softmax来校准神经网络的非概率输出。其中$\mathrm{W} \in R^{C \times C}$， $\mathrm{b} \in R^{C}$

$$\hat{q}_i = \max_k \sigma_{SM}(\mathrm{W}z_i + b)^{(k)}$$

采用NLL损失函数：希望预测概率分布与真实分布一致。（$\boldsymbol{y_i}$为第i个样本标签的one-hot向量，$\boldsymbol{y_i} \in R^C$，$\hat{\boldsymbol{q}}_{\boldsymbol{i}} = \sigma_{SM}(\mathrm{W}z_i + b)$）

$$loss_i = -(\boldsymbol{y_i})^T \log(\hat{\boldsymbol{q}}_{\boldsymbol{i}})$$

$$L(\mathrm{W}, b) = -\sum_{i=1}^{N} (\boldsymbol{y_i})^T \log(\hat{\boldsymbol{q}}_{\boldsymbol{i}})$$

在验证集上寻找最优W, b:

$$(W, b) = argmin_{a,b} -\sum_{i=1}^{N} (\boldsymbol{y_i})^T \log(\hat{\boldsymbol{q}}_{\boldsymbol{i}})$$

该研究主题（任务）已有方法的不足

Matrix and vector scaling （多分类）

$$\hat{q}_i = \max_k \sigma_{SM}(\mathrm{W}z_i + b)^{(k)}$$

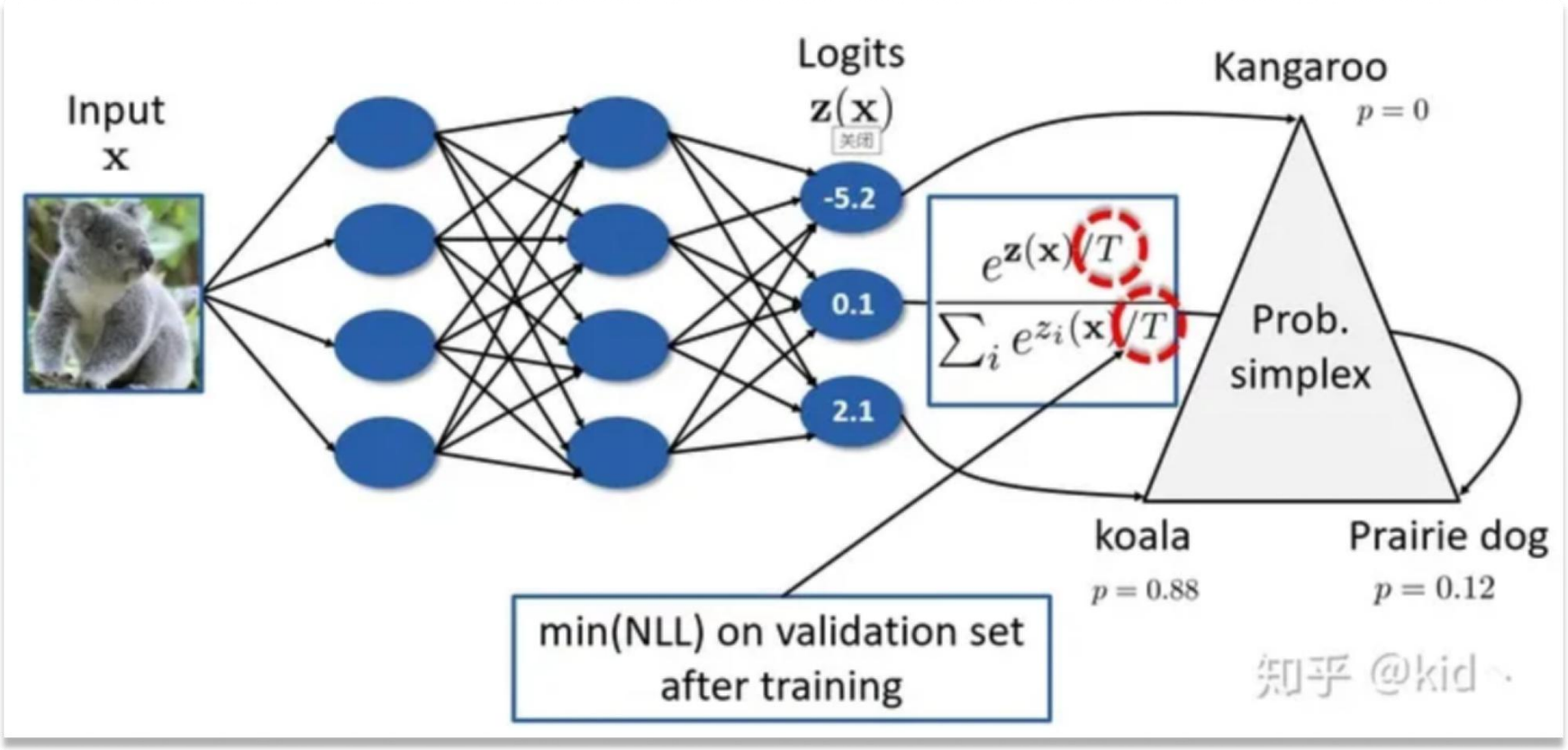
不足：

1. $\mathrm{W}z_i + \mathrm{b}$通过线性变换来调整概率分布，可能会改变预测的最大值，从而影响测试误差。
2. 参数量较大，在验证集上可能过拟合。

本文解决思路

神经网络后处理

对所有原始类别得分采用**单一参数变量**校准神经网络的**非概率输出**



所提出的模型及方案

神经网络训练后保持权重不变，利用**最大熵原理**，并**增加预测概率分布与真实分布一致的约束**，选择校准概率输出$\hat{q}_i$。

Temperature scaling

$$\max_{q} -\sum_{i=1}^{n}\sum_{k=1}^{K} q(z_i)^k \, log q(z_i)^k$$

$$subject_K to: \; q(z_i)^k \geq 0$$

$$\forall i \, , \sum_{k=1} q(z_i)^k = 1$$

$$\forall i \, , \sum_{i=1}^{n} z_i{}^{(y_i)} = \sum_{i=1}^{n}\sum_{k=1}^{K} (z_i)^k q(z_i)^k$$

利用拉格朗日乘子法进行求解得到：$\hat{q}_i = \max_k \sigma_{SM}\left(\frac{z_i}{T}\right)^{(k)}$

所提出的模型及方案关键模块

Temperature scaling

对所有原始类别得分采用**单一参数变量$\frac{1}{T}$校准神经网络的非概率输出**：

$$\hat{q}_i = \max_k \sigma_{SM}\left(\frac{z_i}{T}\right)^{(k)}$$

以NLL作为指标，选取最优温度T：（$\boldsymbol{y_i}$为第i个样本标签的one-hot向量，$\boldsymbol{y_i} \in R^C, \hat{\boldsymbol{q}}_{\boldsymbol{i}} = \sigma_{SM}(\mathrm{W}z_i + b)^{(k)}$）

$$loss_i = -(\boldsymbol{y_i})^T \log(\hat{\boldsymbol{q}}_{\boldsymbol{i}})$$

$$L(T) = -\sum_{i=1}^{N} (\boldsymbol{y_i})^T \log(\hat{\boldsymbol{q}}_{\boldsymbol{i}})$$

在验证集上寻找最优T：

$$T = argmin_{a,b} - \sum_{i=1}^{N} (\boldsymbol{y_i})^T \log(\hat{\boldsymbol{q}}_{\boldsymbol{i}})$$

所提出的模型/方案--损失函数

模型训练时采用**交叉熵**作为损失：

$\boldsymbol{y_i}$为第i个样本标签的one−hot向量，$\boldsymbol{y_i} \in R^C, \hat{\boldsymbol{p}}_{\boldsymbol{i}} = \sigma_{SM}(z_i)$

$$L(\mathrm{W}, b) = -\sum_{i=1}^{N} (\boldsymbol{y_i})^T \log(\hat{\boldsymbol{p}}_{\boldsymbol{i}})$$

实验设置与结果分析--Experimental Datasets

01 **图像分类数据集：**

Caltech-UCSD Birds： 200种鸟类。

数据量：5994/2897/2897，分别用于训练/验证/测试。

02 **ImageNet 2012**：1000个类别的自然场景图像。
数据量：130万/2.5万/2.5万，分别用于训练/验证/测试。

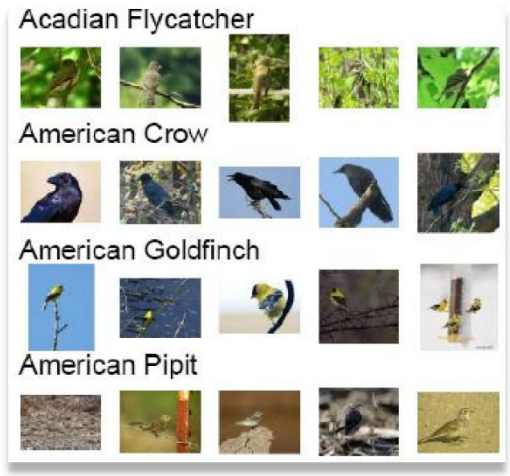
03 **CIFAR-10/CIFAR-100**： 10/100个类别的32×32彩色图像。

数据量：4.5万/5000/1万，分别用于训练/验证/测试。

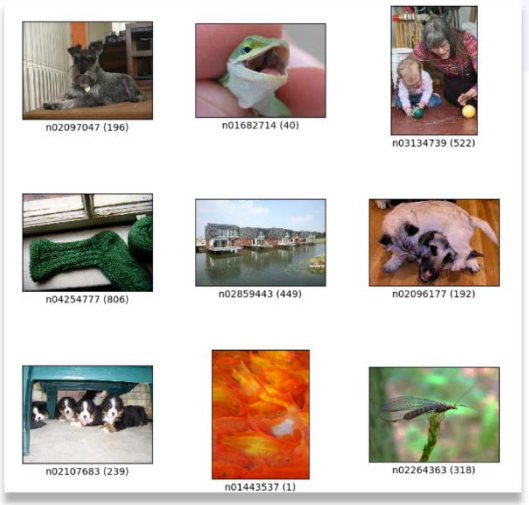
04 **Street View House Numbers (SVHN)**

从谷歌街景中裁剪出的32×32彩色房屋号码图像。

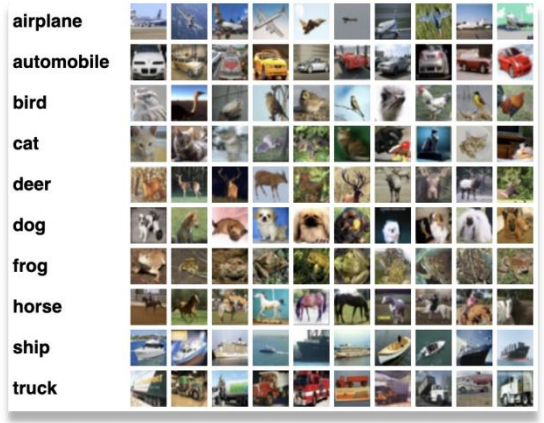
数据量：598388/6000/26032，分别用于训练/验证/测试。



Caltech-UCSD Birds



ImageNet 2012



CIFAR-10/CIFAR-100

Street View House Numbers

实验设置与结果分析--Evaluation Metrics

可靠性图：用来观察模型是否完美校准

准确性作为置信度的函数被绘制，**完美校准的模型可靠性图可以绘制出对角线。**

期望校准误差ECE：量化准确率与置信度的一致性

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} \times \left(acc(B_m) - conf(B_m)\right)$$

测试误差（Test Error）

计算时间复杂度

实验设置与结果分析--Experimental setting

图像分类：

ResNets、ResNets (SD)、Wide ResNets和DenseNets:使用了每篇论文中描述的数据预处理、训练过程和超参数。

文档分类：

20 News和Reuters上训练了具有3个前馈层和批量归一化的Deep Averaging Networks (DANs)在SST上训练了TreeLSTMs (Long Short Term Memory)。对于这两种模型,使用了作者推荐的默认超参数。

实验设置与结果分析--Experimental results

Dataset	Model	Uncalibrated	Hist. Binning	Isotonic	BBQ	Temp. Scaling	Vector Scaling	Matrix Scaling
Birds	ResNet 50	9.19%	4.34%	5.22%	4.12%	**1.85%**	3.0%	21.13%
Cars	ResNet 50	4.3%	**1.74%**	4.29%	1.84%	2.35%	2.37%	10.5%
CIFAR-10	ResNet 110	4.6%	0.58%	0.81%	**0.54%**	0.83%	0.88%	1.0%
CIFAR-10	ResNet 110 (SD)	4.12%	0.67%	1.11%	0.9%	**0.6%**	0.64%	0.72%
CIFAR-10	Wide ResNet 32	4.52%	0.72%	1.08%	0.74%	**0.54%**	0.6%	0.72%
CIFAR-10	DenseNet 40	3.28%	0.44%	0.61%	0.81%	**0.33%**	0.41%	0.41%
CIFAR-10	LeNet 5	3.02%	1.56%	1.85%	1.59%	**0.93%**	1.15%	1.16%
CIFAR-100	ResNet 110	16.53%	2.66%	4.99%	5.46%	**1.26%**	1.32%	25.49%
CIFAR-100	ResNet 110 (SD)	12.67%	2.46%	4.16%	3.58%	0.96%	**0.9%**	20.09%
CIFAR-100	Wide ResNet 32	15.0%	3.01%	5.85%	5.77%	**2.32%**	2.57%	24.44%
CIFAR-100	DenseNet 40	10.37%	2.68%	4.51%	3.59%	1.18%	**1.09%**	21.87%
CIFAR-100	LeNet 5	4.85%	6.48%	2.35%	3.77%	**2.02%**	2.09%	13.24%
ImageNet	DenseNet 161	6.28%	4.52%	5.18%	3.51%	**1.99%**	2.24%	-
ImageNet	ResNet 152	5.48%	4.36%	4.77%	3.56%	**1.86%**	2.23%	-
SVHN	ResNet 152 (SD)	0.44%	**0.14%**	0.28%	0.22%	0.17%	0.27%	0.17%
20 News	DAN 3	8.02%	**3.6%**	5.52%	4.98%	4.11%	4.61%	9.1%
Reuters	DAN 3	0.85%	1.75%	1.15%	0.97%	0.91%	**0.66%**	1.58%
SST Binary	TreeLSTM	6.63%	1.93%	**1.65%**	2.27%	1.84%	1.84%	1.84%
SST Fine Grained	TreeLSTM	6.71%	2.09%	**1.65%**	2.61%	2.56%	2.98%	2.39%

Table 1. ECE (%) (with $M = 15$ bins) on standard vision and NLP datasets before calibration and with various calibration methods. The number following a model's name denotes the network depth.

Dataset	Model	Uncalibrated	Hist. Binning	Isotonic	BBQ	Temp. Scaling	Vector Scaling	Matrix Scaling
Birds	ResNet 50	**22.54%**	55.02%	23.37%	37.76%	**22.54%**	22.99%	29.51%
Cars	ResNet 50	14.28%	16.24%	14.9%	19.25%	14.28%	**14.15%**	17.98%
CIFAR-10	ResNet 110	**6.21%**	6.45%	6.36%	6.25%	**6.21%**	6.37%	6.42%
CIFAR-10	ResNet 110 (SD)	5.64%	5.59%	5.62%	**5.55%**	5.64%	5.62%	5.69%
CIFAR-10	Wide ResNet 32	**6.96%**	7.3%	7.01%	7.35%	**6.96%**	7.1%	7.27%
CIFAR-10	DenseNet 40	**5.91%**	6.12%	5.96%	6.0%	**5.91%**	5.96%	6.0%
CIFAR-10	LeNet 5	15.57%	15.63%	15.69%	15.64%	15.57%	**15.53%**	15.81%
CIFAR-100	ResNet 110	27.83%	34.78%	28.41%	28.56%	27.83%	**27.82%**	38.77%
CIFAR-100	ResNet 110 (SD)	**24.91%**	33.78%	25.42%	25.17%	**24.91%**	24.99%	35.09%
CIFAR-100	Wide ResNet 32	**28.0%**	34.29%	28.61%	29.08%	**28.0%**	28.45%	37.4%
CIFAR-100	DenseNet 40	26.45%	34.78%	26.73%	26.4%	26.45%	**26.25%**	36.14%
CIFAR-100	LeNet 5	**44.92%**	54.06%	45.77%	46.82%	**44.92%**	45.53%	52.44%
ImageNet	DenseNet 161	22.57%	48.32%	23.2%	47.58%	22.57%	**22.54%**	-
ImageNet	ResNet 152	**22.31%**	48.1%	22.94%	47.6%	**22.31%**	22.56%	-
SVHN	ResNet 152 (SD)	**1.98%**	2.06%	2.04%	2.04%	**1.98%**	2.0%	2.08%
20 News	DAN 3	20.06%	25.12%	20.29%	20.81%	20.06%	**19.89%**	22.0%
Reuters	DAN 3	2.97%	7.81%	3.52%	3.93%	2.97%	**2.83%**	3.52%
SST Binary	TreeLSTM	11.81%	12.08%	11.75%	**11.26%**	11.81%	11.81%	11.81%
SST Fine Grained	TreeLSTM	49.5%	49.91%	48.55%	49.86%	49.5%	49.77%	**48.51%**

Table S2. Test error (%) on standard vision and NLP datasets before calibration and with various calibration methods. The number following a model's name denotes the network depth. Error with temperature scaling is exactly the same as uncalibrated.

不同校准方法的校准性能比较

实验结果：

01 Temperature scaling在校准视觉任务上的表现优于所有其他方法，并且在自然语言处理(NLP)数据集上与其他方法表现相当。

02 Vector scaling学习到的向量具有基本不变的项，几乎与Temperature scaling没有区别。

实验设置与结果分析--Experimental results

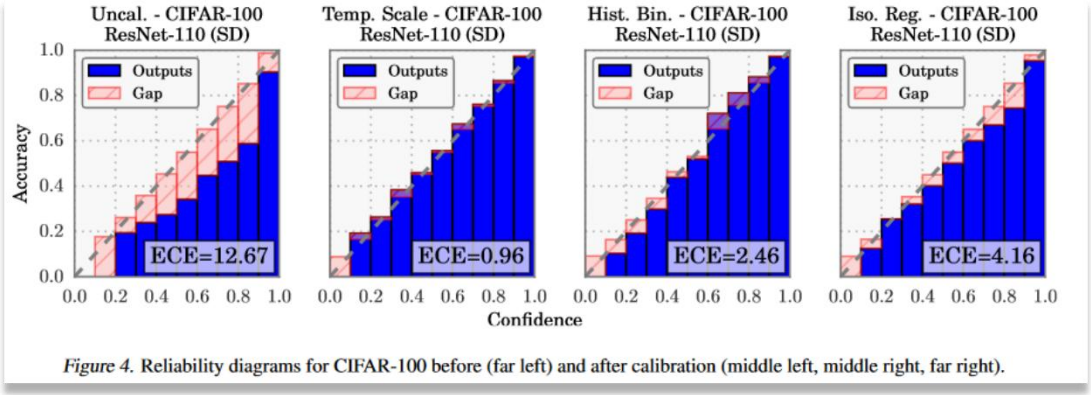


Figure 4. Reliability diagrams for CIFAR-100 before (far left) and after calibration (middle left, middle right, far right).

不同校准方法的校准性能比较

实验结果：

在CIFAR-100分类任务上，Temperature scaling相较与其他校准方法，更接近完美校准。

实验设置与结果分析--Experimental results

不同校准方法的计算性能比较

以CIFAR-100数据集为例：

1. 使用共轭梯度下降法找到Vector scaling的近似最优解所需的时间，比Temperature scaling多至少两个数量级
2. Histogram binning 和 Isotonic regression的计算时间比Temperature scaling长约一个数量级
3. BBQ方法的计算时间比Temperature scaling长约三个数量级

实验结论： Temperature scaling在计算上是最高效的校准方法

总结与结论

01

主要贡献

发现了**预测的概率估计并不代表真实的正确性可能性的校准问题**。

通过实验指出网络的深度、宽度、权重衰减和批量归一化是影响校准的重要因素。

列举了几种后校准方法,并提出了**有效高效的Temperature scaling后校准**方法。

02

本文局限

论文没有详细对超参数设置进行分析,这些超参数可能影响模型的校准性能。

所有校准方法相较于未校准的14.01%都未能显著降低MCE,是否可以表明该数据集的某些特性使得校准更加困难?论文没有继续分析

启发与思考

01

优点

采用简单而有效的后校准方法,使置信度更加可靠。

02

缺点

Temperature scaling在校准视觉任务上的表现良好,但**在不同类型和规模的数据集上的泛化能力可能还需要进一步的研究。**

03

启发

神经网络进行分类任务时不仅要考虑准确性,还应考虑校准性能