计算机科学
COMPUTER SCIENCE

重新审视面向CNN模型的测试样例选取:考虑模型校准

赵通, 沙朝锋

引用本文

赵通, 沙朝锋. 重新审视面向CNN模型的测试样例选取:考虑模型校准[J]. 计算机科学, 2024, 51(6): 34-43.

ZHAO Tong, SHA Chaofeng. Revisiting Test Sample Selection for CNN Under Model Calibration[J]. Computer Science, 2024, 51(6): 34-43.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于次模函数最大化的测试用例集约简

Test Suite Reduction via Submodular Function Maximization

计算机科学, 2021, 48(12): 75-84. https://doi.org/10.11896/jsjkx.210300086

基于最优输运和*k*-近邻的离群文档检测

Outlier Document Detection via Optimal Transport and *k*-nearest Neighbor

计算机科学, 2021, 48(7): 105-111. https://doi.org/10.11896/jsjkx.200400140

基于Stack Overflow的数据库相关主题分析

Analysis of Topics on Database Systems in Stack Overflow

计算机科学, 2021, 48(6): 48-56. https://doi.org/10.11896/jsjkx.200800217

基于随机投影和主成分分析的网络嵌入后处理算法

Post-processing Network Embedding Algorithm with Random Projection and Principal Component Analysis

计算机科学, 2021, 48(5): 124-129. https://doi.org/10.11896/jsjkx.200500058

基于多峰高斯函数的直方图规定化算法

Image Specification Algorithm Based on Multi-peaks Gaussian Function

计算机科学, 2014, 41(12): 245-250. https://doi.org/10.11896/j.issn.1002-137X.2014.12.053

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230400029

重新审视面向 CNN 模型的测试样例选取:考虑模型校准

赵 通 沙朝锋
复旦大学计算机科学技术学院 上海 200433
(20210240031@fudan.edu.cn)

摘 要 深度神经网络(DNN)已被广泛应用于各种任务,而在部署前对 DNN 进行充分测试尤为重要,因此需要构建能够对 DNN 进行充分测试的测试集。由于标注成本受限,通常通过测试样例选取的方式得到测试子集。然而,人们使用基于预测不确定性的方法(该方法在发现误分类样例和提升重训练表现方面表现出卓越的能力)进行测试样例选取时,忽略了对测试样例的预测不确定性估计是否准确的问题。为了填补上述研究的空白,通过实验定性和定量地揭示了模型校准程度和测试样例选取任务中使用的不确定性指标之间的相关性。校准模型会使模型有更准确的预测不确定性估计,因此研究了不同校准程度的模型用不确定指标选取得到的测试子集质量是否不同。在 3 个公开数据集和 4 个卷积神经网络(CNN)架构模型上进行了充分的实验和分析,结果表明在 CNN 架构模型上:1)不确定指标和模型校准存在一定程度的相关性;2)校准程度好的模型所选择的测试子集质量优于校准程度差的模型选择的测试子集质量。在发现模型误分类样例的能力上,70.57%经过校准训练后的模型对应的实验结果优于未校准模型对应的实验结果。因此在测试样例选取任务中考虑模型校准十分重要,且可以使用模型校准来提升测试样例选取的表现。

关键词:卷积神经网络测试;预测不确定性;模型校准;测试样例选取
中图分类号 TP391

Revisiting Test Sample Selection for CNN Under Model Calibration

ZHAO Tong and SHA Chaofeng
School of Computer Science, Fudan University, Shanghai 200433, China

Abstract Deep neural networks are widely used in various tasks, and model testing is crucial to ensure their quality. Test sample selection can solve the issue of labor-intensive manual labeling by strategically choosing a small set of data to label. However, existing selection metrics based on predictive uncertainty neglect the accuracy of the estimation of predictive uncertainty. To fill the gaps of the above studies, we conduct a systematic empirical study on 3 widely used datasets and 4 convolutional neural networks(CNN) to reveal the relationship between model calibration and predictive uncertainty metrics used in test sample selection. We then compare the quality of the test subset selected by calibrated and uncalibrated models. The findings indicate a degree of correlation between uncertainty metrics and model calibration in CNN models. Moreover, CNN models with better calibration select higher-quality test subsets than models with poor calibration. Specifically, the calibrated model outperforms the uncalibrated model in detecting misclassified samples in 70.57% of the experiments. Our study emphasizes the importance of considering model calibration in test selection and highlights the potential benefits of using a calibrated model to improve the adequacy of the testing process.

Keywords Convolutional neural network testing, Predictive uncertainty, Model calibration, Test sample selection

1 引言

深度神经网络(Deep Neural Networks, DNN)已被广泛应用于图像分类[1]、自动驾驶[2]等领域,并取得了令人瞩目的成果。然而,由于训练数据与实际应用数据的分布不同、训练数据中存在噪声等,DNN 系统在极端或边界情况下往往表现出不正确的行为。因此,在部署 DNN 软件系统之前,对其进行充分的测试是非常有必要的,以保证其质量和可靠性。

为了测试 DNN 模型,发现 DNN 模型的错误行为,需要构建能对 DNN 进行充分测试的测试集。由于标注成本受限,研究人员以固定的标注成本从未标注的候选测试集中选取数据进行标注并最终组成测试子集。在有限的标注成本下,可从两个方面评估该测试子集,以使候选测试集能够得到最充分的利用:1)测试子集应尽可能多地发现 DNN 模型的错误行为;2)模型在测试子集上的重训练应获得较高的表现提升。将上述从候选测试集中选取测试子集并使选择的测试子集尽量满足上述两个目标的任务定义为测试样例选取。研究人员提出了一系列测试样例选取的方法,如神经元覆盖

到稿日期:2023-04-05 返修日期:2023-05-29
通信作者:沙朝锋(cfsha@fudan.edu.cn)

率[3-4]、惊讶程度[5]，以及预测不确定性[6-7]等，且基于预测不确定性的方法在发现误分类样例和提升重训练表现方面的能力优于另外两类指标。

模型校准指模型提供的预测概率能准确地反映其预测不确定性，也即 DNN 应指出它们何时可能给出不正确的预测。例如，在医疗诊断应用中，应该将模型给出低置信度预测的图像交给医生复核，自信但不正确的预测可能会带来灾难性的后果。近年来，研究人员已提出了多种校准模型的方法，如以后处理方式校准训练好的模型的方法 Temperature Scaling[8]，以及两种训练校准程度更好的模型的方法 Mixup[9]和 Label Smoothing[10]。

在基于预测不确定性进行测试样例选取时，大部分针对 DNN 模型进行测试样例选取的相关工作往往忽略了预测不确定性估计的准确性，也未考察预测不确定性与模型校准之间的关系。此外，这些工作也没有探究校准模型是否会对基于预测不确定性的测试样例选取方法的表现造成影响。

卷积神经网络(Convolutional Neural Networks，CNN)是计算机视觉领域广泛采用的一种 DNN 架构。因此，针对上述问题，本文在图像分类任务上使用 CNN 模型，主要开展了两方面的工作：1)研究任务中广泛采用的预测不确定性方法，定性分析不确定性指标和模型校准的关系，用 Kendall 相关系数定量地评估二者的关系，评估不确定性指标的估计是否准确；2)用 Temperature Scaling，Mixup，Label Smoothing 训练得到不同校准程度的模型，研究模型校准如何影响基于不确定性指标选取的测试子集的质量。

本文在 3 个常用图像数据集和 4 种 CNN 架构模型上进行了翔实的实验，实验结果显示，在原始测试集上模型校准和不确定性指标之间具有较强的相关性，在部分 OOD(Out-of-Distribution)测试集上，模型校准和不确定性指标之间的关系呈减弱的现象。在测试样例选取任务中考虑模型校准的影响，提出了评估不同校准程度模型选取测试子集质量的两个指标：相对准确率和基准模型准确率。实验发现，在发现模型误分类样例的能力上，以基准模型准确率来衡量，70.57%经过校准训练后的模型对应的实验结果优于未校准模型对应的实验结果：1)用 Mixup 和 Label Smoothing 训练的校准程度更好的模型选取的测试子集质量优于未校准的模型；2)Temperature Scaling 虽然能提高模型在测试集上的校准程度，却不会为基于预测不确定性的测试样例选取提供帮助。

2　相关工作

2.1　测试样例选取

受传统软件测试代码覆盖率的启发，研究人员提出神经元覆盖率，最大化这些神经元覆盖率往往意味着最大化测试的充分性。Pei 等[3]提出了 DeepXplore，首先引入神经元覆盖率来评估测试集覆盖 DNN 模型的程度。Ma 等[4]提出了测试覆盖率指标 DeepGauge，其包括在神经元级别和层级别不同粒度的覆盖率标准，进一步表明 DNN 模型结构的更高覆盖率意味测试数据更有效。最近，Yuan 等[11]提出了一种新的神经元覆盖率指标 Neural Layer Coverage，考虑将 DNN 的层而不是神经元作为计算单元。Kim 等[5]提出了 Surprise Adequacy，该指标衡量模型在面对新输入时的惊讶程度。他们假设一个测试子集涵盖更多样化的意外值时，更适合测试以及重训练。

Ma 等[6]提出了一组基于预测不确定性的测试样例选取指标，包括预测的最大概率(MaxP)、多次 dropout 模型预测的方差(Var)、结合了 MaxP 和 Var 的加权方差(Var_w)以及 Kullback-Leibler 得分。Feng 等[7]提出了 DeepGini，基于模型对输入样本的预测分布计算 Gini 系数。在研究工作中，他们还利用预测分布的熵(Entropy)作为其方法的替代指标。Weiss 等[12]发现，用 MaxP 或者 Entropy 评估预测不确定性在测试样例选取工作中的表现和 DeepGini 相当。Zhang 等[13]使用 Margin 来衡量预测不确定性，该指标也被称为预测-置信度分数(PCS)，是测试样例选取任务中广泛采用的一种有效方法。

最近，研究人员不仅从预测不确定性的角度考虑选择测试子集，而且还考虑了测试子集的代表性和多样性。Zhao 等[14]发现大部分测试样例选取方法不能保证测试样例选取的多样性。Chen 等[15]提出了基于多边界聚类的测试样例选取算法，算法将测试样本聚类到深度学习模型多个分类边界的边界区域，并按照一定的优先级从所有边界区域均匀地选择样本，以确保在重训练过程中，DNN 学习新的分类边界时都有足够的有用样本。Gao 等[16]提出了一种自适应测试样例选取方法，利用模型输出之间的差异来衡量 DNN 测试数据的多样性，旨在从未标记数据集中选取具有不同特征测试样本的测试子集。Hu 等[17]提出了 Aries，并且认为测试样例选取任务中仍需要手工标注测试子集，Aries 仅使用从原始测试数据中获得的信息来估计 DNN 在新的未标记数据上的性能。Arrieta 等[18]提出了一种基于元测试的多目标测试样例选取方法，该方法可以基于执行源测试用例时获得的信息，选择合适的变异测试样例来测试 DNN。

2.2　模型校准

校准模型会使模型有更准确的预测不确定性估计，因此，提出更准确的计算预测不确定性的指标是模型校准研究领域的一项重要任务。Guo 等[8]建议使用 Temperature Scaling 来校准模型，是一种有效且高效的方式。Zadrozny 等[19]提出了直方图分箱，作为一种简单的非参数校准方法。Mixup[9]训练已被证明可以改善模型的泛化和校准，防止模型对其预测过于自信[20]。Label Smoothing[10]被证明可有效提高不确定性估计的质量[21]。Qin 等[22]提出了基于对抗鲁棒性的自适应标签平滑(AR-AdaLS)，它将对抗鲁棒性和校准的相关性集成到训练中，通过训练样本基于对手攻击它的难易程度设置 Label Smoothing 的参数，即使在分布变化的情况下，AR-AdaLS 也可以对模型进行更好的校准。Li 等[23]提出了一种基于高斯过程利用模型实际应用场景的数据校准模型的方法。Li 等[24]提出可以利用模型的可解释性来提升模型的校准性能。Minderer 等[25]重新审视了最新的非 CNN 神经网络架构模型，例如 ViT，MLP-Mixer 的校准性能，发现此类模型除了在鲁棒性、准确率等方面优于 CNN 模型外，在校准性能上也有良好表现。

3　背景知识

3.1　DNN 模型架构

DNN 由多个层组成，每个层包含多个神经元，图 1 为

DNN 模型及内部神经元的一个简单样例。一个 DNN 通常至少由三(通常更多)层组成:一个输入层、一个输出层和一个或多个隐藏层。一层中的每个神经元都直接连接到下一层中的神经元。神经元是 DNN 中的单个计算单元,它在其输入上应用激活函数 σ 并将结果传递给其他连接的神经元。

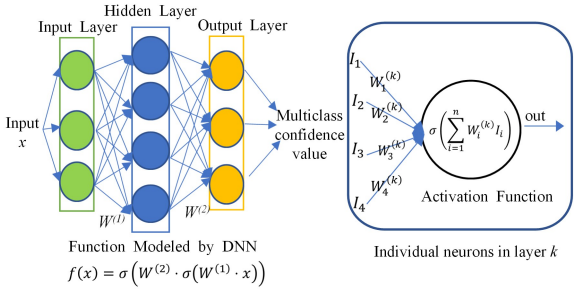


图 1 DNN 模型及内部神经元示例

Fig. 1 DNN model and its inner neuron sample

在后续介绍中,我们考虑类别为 C 的分类任务,数据集 $D=\{(x,y)\}$ 的大小为 N,其中 x 为一条样本,y 为对应的类标签。深度学习模型为 f,f 对 x 预测得到 Logit 向量为 $\boldsymbol{z}(x)$,经过 Softmax 计算得到概率向量为 $\boldsymbol{p}(x)$,$p_i(x)$ 为对类别 i 的预测概率。Softmax 的计算式如下:

$$p_i(x)=\frac{e^{z_i(x)}}{\sum_{c=1}^{C}e^{z_c(x)}}$$

训练中,用类标签 y 计算得到其对应的 one-hot 向量 $\boldsymbol{y}$,最小化 $\boldsymbol{y}$ 和网络输出之间的交叉熵:

$$\mathcal{L}(\boldsymbol{p}(x),\boldsymbol{y})=-\sum_{i=1}^{C}y_i\log(p_i(x))$$

3.2 基于预测不确定性的测试样例选取

在测试样例选取任务中,由于标注成本的限制,人们往往期望在给定的标注成本下尽可能地多发现模型预测错误的测试样例。而对于那些模型给出较大不确定预测的测试样本,模型往往会给出错误的预测,因此,预测不确定性已被广泛应用于测试样例选取任务。

DeepGini 基于模型对输入样本的预测分布计算 Gini 系数,定义 DeepGini 为:

$$DeepGini(x)=1-\sum_{i=1}^{C}p_i(x)^2$$

当 DNN 以高确定性预测一个类别时,DeepGini 的值较小,例如得到的预测概率向量为 one-hot 向量时,DeepGini 取得最小值 0。

定义熵(Entropy)为:

$$Entropy(x)=-\sum_{i=1}^{C}p_i(x)\log p_i(x)$$

当仅有一个类别的 $p_i(x)$ 较高时,则熵较小;当 DNN 对所有类别的预测较为接近时,则熵较大。

Margin 为 DNN 预测概率向量的最大概率和第二大概率之差。Margin 值较小的数据样本可能更接近 DNN 的分类边界,这意味着对 DNN 输入样本微小的扰动,可能改变 DNN 的预测类别。定义 Margin 为:

$$Margin(x)=\max_{i_1=1,C}p_{i_1}(x)-\max_{i_2=1,C,i_2!=i_1}p_{i_2}(x)$$

MaxP 只考虑预测概率最大的类别来计算不确定性。MaxP 根据最高的预测概率对样本进行排序,并选择分类置信度最低的样本。定义 MaxP 为:

$$MaxP(x)=\max_{i=1,C}p_i(x)$$

3.3 模型校准的方法和评估校准程度指标

模型校准衡量预测概率和准确率之间的一致性,经过良好校准的预测传达了我们应该在多大程度上信任模型预测。下面我们介绍 3 种常用的校准模型的方法和 1 种广泛采用的衡量模型校准程度的指标。

3.3.1 Temperature Scaling

Temperature Scaling 将 logit 向量 $\boldsymbol{z}(x)$ 调整为 $\boldsymbol{z}(x)$,其中 $T>0$ 为温度参数,DNN 新的预测为:

$$\boldsymbol{q}(x)=\mathrm{Softmax}(\boldsymbol{z}(x))$$

当 $T>1$ 时,它"软化"Softmax,当 $T\to\infty$ 时,概率 $q_i(x)$ 接近 $1/C$,预测的不确定性最大。在 $T=1$ 的情况下,恢复原始概率 $p_i(x)$。当 $T\to0$ 时,预测的概率向量坍缩到一个类别上(即 $q_i(x)=1$)。T 由在验证集上对损失进行优化得到,因为参数 T 不会改变 Softmax 预测的最大值,所以类别预测保持不变,不会影响模型的准确率。

3.3.2 Mixup

Mixup 是一种在训练过程中使用的数据增强方式,用其训练得到的模型校准程度得以提高并且对预测不确定性的估计更加准确[20]。使用 Mixup 训练时,分类器不仅在训练样本上训练,而且在每个训练样本的附近的空间训练。邻近点根据如下规则生成:

$$\tilde{x}=\lambda x_1+(1-\lambda)x_2$$

$$\tilde{y}=\lambda y_1+(1-\lambda)y_2$$

其中,x_1 和 x_2 是两个随机采样的输入样本,y_1 和 y_2 是它们关联的类标签。

确定混合比的 $\lambda\in[0,1]$ 在每次训练迭代时从对称 Beta 分布 $Beta(\alpha,\alpha)$ 中得出,其中 α 是控制样本对之间的插值强度和训练标签的相关平滑的参数。$\alpha=0$ 时恢复出原训练标签(one-hot 编码,在这种情况下生成的样本只是 x_1 或 x_2)的基本情况,α 的值越大,最终得到的输入和标签越平均。在[0.1, 0.4]中的 α 值给出了最好的分类结果。

3.3.3 Label Smoothing

Label Smoothing(LS)是一种正则化技术,可缓解使用深度学习模型进行分类任务时出现的过度拟合和过度自信问题。除了提高泛化能力外,Label Smoothing 还提高了模型校准程度[21]。对于使用参数 s 的 Label Smoothing 方法训练网络,训练过程中最小化修改后的标注 y_i^{LS} 和网络输出 $p_i(x)$ 之间的交叉熵,其中修改后的标注为:

$$y_i^{LS}=y_i(1-s)+sC$$

3.3.4 Expected Calibration Error(ECE)

本研究在后续的实验中,我们使用 ECE 来衡量模型在测试集上的校准程度。要计算模型在给定数据集上的 ECE,首先需要将数据集中的所有数据分成 K 组,并按预测类别的置信度排序。令 G_k 表示第 k 组中的所有数据样本,那么 G_k 的准确率和置信度定义为 $acc(G_k)=\frac{1}{|G_k|}\sum_{t\in G_k}\mathbf{1}(\hat{y}_t=y_t)$ 和 $conf(G_k)=\frac{1}{|G_k|}\sum_{t\in G_k}\hat{p}_t^{\hat{y}_t}$,其中 $\hat{y}$ 和 y 分别表示预测类别和真实类别,$\hat{p}^{\hat{y}}$ 表示 $\hat{y}$ 的预测概率。定义 ECE 为:

$$ECE=\sum_{k=1}^{K}\frac{|G_k|}{N}|acc(G_k)-conf(G_k)|$$

3.4　评估不同校准程度模型测试子集质量的指标

本研究和以往研究人员在测试样例选取任务上进行的研究存在不同：以往人们固定模型，比较不同测试样例选取指标之间的差别，而本文比较的是不同校准程度的模型选择测试子集之间的区别。由于不同模型在候选测试集上具有不同的准确率，因此无法继续使用测试子集在原模型上的准确率来比较测试子集的优劣。

为了量化测试子集的优劣，我们提出相对准确率来衡量不同模型选出测试子集的质量。假设模型在候选测试集上的分类准确率为A_0，在测试子集上的分类准确率为A_1，定义相对准确率R为：

$$R=A_0-A_1$$

R的值越大，表示测试子集的相对质量越高。直观上，R越大，通常意味着A_0较大，A_1较小，也就意味着在一个准确率较高的模型上选出的测试子集在该模型上的准确率较低。

模型在测试子集上重训练表现的提升也是测试样例选取任务中评估测试子集质量的一个重要方面。与测试子集在模型上的准确率不适合评估测试子集的质量类似，模型在测试子集上重训练表现的提升同样不能用于评估测试子集的质量。为了在实验中继续使用模型重训练表现上的提升来评估测试子集的质量，需额外引入一个基准模型M_0。用不同模型选择的测试子集上重训练M_0，以M_0在测试子集重训练表现的提升来评估测试子集的质量，从而对比不同模型选出测试子集的质量。同时，也可以利用测试子集在M_0上的准确率来评估发现模型误分类样例方面的能力。基准模型应取得有竞争力的表现，且其选择应具有代表性。

4　实验设置

本文实验旨在回答以下研究问题(RQ)。

RQ1　测试样例选取任务中使用的不确定指标和模型的校准程度之间是否存在相关性？

RQ2　经过校准后的模型，若校准程度在候选测试集上优于未校准模型，再用不确定指标发现误分类样例的能力是否优于未校准模型？

RQ3　经过校准后的模型选出的测试子集在提升重训练模型表现上的能力是否优于未校准模型？

4.1　数据集收集和预处理

在 3 个广泛使用的图像分类数据集上进行了实验，包括 MNIST[26]，CIFAR10[27] 和 CIFAR100[27]。直接采用 MNIST，CIFAR10，CIFAR100 对应的变换和扰动后的测试数据集 MNIST-C[28]，CIFAR10-C[29] 和 CIFAR100-C[29] 作为 OOD 测试集。这 3 个 OOD 测试集包括不同类型的扰动或变换(MNIST 有 15 种类型，CIFAR10 有 19 种类型，CIFAR100 有 17 种类型)，例如自然图像中经常遇到的噪声、模糊、天气等。对于 MNIST，每种类型扰动有 1 个级别的扰动强度，实验选择该扰动强度对应的测试集作为 MNIST 的 OOD 测试集，选择了 4 种多样且有代表性的扰动类型：Brightness，Fog，Blur(Glass)，Shear。对于 CIFAR10 和 CIFAR100，每种类型的扰动都有 5 个级别的扰动强度，级别越高，扰动程度越高，实验选取扰动级别为 3 对应的测试集作为 CIFAR10 和 CIFAR100 的 OOD 测试集，同样选择 4 种多样且有代表性的扰动类型：Brightness，Fog，Blur(Gaussian)，Noise(Gaussian)。

4.2　模型训练

本文参考以往测试样例选取工作中广泛采用的 4 种 CNN 模型[30-31]，在训练集上训练获得了准确率有竞争力的 CNN 模型。对所有的模型和数据集，训练采用广泛使用的优化器 SGD 和学习率调度器 WarmupCosineScheduler[31]。优化器学习率为 0.1；Momentum[32] 可以加速模型收敛和提升模型泛化性能，设为 0.9；Weight Decay[33] 可以控制模型复杂度且能提升模型泛化性能，设为 5×10^{-4}；WarmupCosineScheduler 可以有效避免训练模型时出现模型震荡的现象，设置预热的轮数为 2 个 epoch。模型在相应的训练集上训练足够的轮数从而保证在训练集上损失收敛。由于 MNIST 数据集较为简单，因此对 MNIST 的模型训练 50 轮；相应地，对 CIFAR10 和 CIFAR100 的模型训练 200 轮。实验设置训练过程中的一批数据的大小为 128 条样本。

如表 1 所列，训练获得的模型在原始测试集上表现出了有竞争力的准确率，这在一定程度上表明了以上训练参数设置的合理性。然而有些模型在测试集上的校准程度不佳，当切换到 OOD 测试集时，模型的准确率和校准程度出现了一定程度的下降。

表 1　模型在各测试集上的准确率和 ECE

Table 1　Accuracy and ECE of models on test set

Dataset	Model	Metric	Original	OOD			
				Brightness	Fog	Blur	Noise(Shear)
MNIST	LeNet5	ACC	99.27	96.27	85.04	94.64	97.45
		ECE	0.17	9.35	37.18	4.78	0.38
	MNISTCNN	ACC	96.08	23.75	17.99	75.71	89.28
		ECE	0.73	39.05	56.86	7.31	1.03
CIFAR10	VGG16	ACC	94.30	93.14	89.69	81.04	39.74
		ECE	3.96	4.97	7.53	14.65	48.68
	ResNet50	ACC	95.69	94.58	91.87	82.63	41.45
		ECE	3.03	3.86	6.09	13.06	48.35
CIFAR 100	VGG16	ACC	73.96	70.30	62.24	54.24	18.83
		ECE	12.98	14.79	19.02	23.65	49.16
	ResNet50	ACC	78.63	74.57	68.61	57.08	20.49
		ECE	7.07	8.90	11.96	17.53	41.35

4.3 校准模型训练

实验需要获取在同一数据集上具有不同校准程度的模型。我们用 Mixup、Label Smoothing 训练方式及 Temperature Scaling 后处理方式获取到不同校准程度的模型。同时，由于 MNIST 数据集在原始测试集上已取得较高的准确率和校准程度，因此实验仅在 CIFAR10 和 CIFAR100 上进行。

对于 Temperature Scaling(TS)，参考原论文的训练设置[8]，将原始训练集 95%的数据用于训练模型。训练阶段结束后，固定训练好的模型参数并用剩下的 5%的数据训练温度参数 T，直至温度参数拟合。

对于 Mixup(MX)，参考原论文建议的参数设置[9]，调整 Mixup 的 α 参数以及每步训练使用 Mixup 的概率 $prob$，α 从[0.2,0.4]中选取。对于 CIFAR10，α 设置为 0.2；对于 CIFAR100，α 设置为 0.3。固定两个数据集的 $prob$ 为 0.5。

对于 Label Smoothing(LS)，参考原论文推荐的参数设置[10]，调整 Label Smoothing 的 s 参数，参数在[0.05,0.2]中选取。对于 CIFAR10，s 设置为 0.05；对于 CIFAR100，s 设置为0.1。

如表 2 所列，校准后的模型相比未校准模型在原始测试集上和 OOD 测试集上取得了较高的准确率和较低的 ECE，且所有模型在 OOD 测试集上准确率下降，ECE 上升，这在一定程度上表明了以上训练方法参数设置的合理性，且成为我们下一步对比不同校准程度的模型用不确定性指标选取测试子集质量的基础。Densenet121 基准模型在各数据集上都取得了较高的准确率，因此，若其他模型选择的测试子集能够发现更多使该基准模型误分类的测试样例，或者更大地提升该模型的重训练表现，则表明模型选择的测试子集质量更高。

表 2 不同校准程度模型和基准模型的准确率和 ECE

Table 2 Accuracy and ECE of different calibrated model and baseline models on test set

	Model	Metric	Original	OOD			
				Brightness	Fog	Blur	Noise(Shear)
CIFAR10	Densenet121	ACC	95.80	94.35	92.00	84.72	46.95
		ECE	3.04	4.18	6.10	11.73	42.96
	VGG16-MX	ACC	94.37	93.36	90.32	81.97	37.89
		ECE	1.95	2.65	4.36	10.68	47.44
	VGG16-LS	ACC	94.28	93.20	89.19	81.20	41.30
		ECE	2.54	2.45	3.31	9.48	39.24
	VGG16-TS	ACC	94.30	93.14	89.69	81.04	39.74
		ECE	1.66	1.80	3.63	9.88	39.59
	ResNet50-MX	ACC	95.87	94.65	92.27	82.78	44.29
		ECE	2.36	3.24	5.03	12.20	41.48
	ResNet50-LS	ACC	95.47	94.19	91.25	80.95	41.15
		ECE	2.54	2.12	2.13	10.09	42.10
CIFAR100	Densenet121	ACC	95.69	94.58	91.87	82.63	41.45
		ECE	1.16	1.71	3.59	9.26	41.21
	Densenet121	ACC	80.00	76.99	70.64	60.17	13.94
		ECE	12.19	13.94	18.33	25.08	65.10
	VGG16-MX	ACC	73.90	71.10	63.71	55.8	20.42
		ECE	8.92	9.91	13.32	17.57	40.50
	VGG16-LS	ACC	73.85	70.33	62.04	53.89	19.93
		ECE	4.58	4.25	4.07	7.67	31.54
	VGG16-TS	ACC	73.96	70.30	62.24	54.24	18.83
		ECE	4.17	4.33	5.75	9.54	32.40
	ResNet50-MX	ACC	80.15	77.00	71.38	59.57	17.35
		ECE	3.48	4.67	7.29	12.34	34.78
	ResNet50-LS	ACC	79.11	75.39	68.20	57.02	16.62
		ECE	4.77	5.43	8.54	14.70	45.15
	ResNet50-TS	ACC	78.63	74.57	68.61	57.08	20.49
		ECE	2.89	2.82	4.54	8.62	30.91

4.4 不同校准程度模型测试样例选取

在测试样例选取的实验，我们比较了不同校准程度模型选取的测试子集发现误分类样例的表现及重训练模型的表现。实验的候选测试集和最终测试集的构建以及标注成本的设置参考以往工作中测试样例选取实验中的设置[15,30]。为了模拟现实生活中遇到的 OOD 场景，构建了原始测试样本和 OOD 测试样本混合的候选测试集和最终测试集，用对应的 OOD 样本替换原图像。例如，对有 10000 个测试样本的 CIFAR10 测试集，随机提取 5000 个原始样本，选择其对应的 5000 个 OOD 样本，并将它们添加到 CIFAR10 中未被选中的样本集合中，以替换原来的 5000 个原始样本。通过上述方式构建了候选测试集，而被替换掉的原始样本和未被替换的原始样本对应的 OOD 样本组成最终的测试集，从而得到同分布的候选测试集和最终测试集。两个测试集中原始样本和 OOD 样本的比例皆为 50%。

在构建候选测试集后，使用不确定性指标来选择测试子集，评估测试子集的相对准确率和基准模型准确率。最后重训练基准模型 M 以获得重训练后的模型M'，计算模型M'在最终测试集上的准确率，从而比较不同校准程度的模型选择的测试子集对重训练的帮助。

考虑有代表性的标注成本 2%和 4%，例如，当候选测试集大小为 10000 且标注成本为 2%时，需要选择 200 个测试样本组成测试子集来重新训练模型。用于重新训练模型的优化器同样为 SGD，学习率为 0.01，Momentum 设置为 0.9，

Weight Decay 设置为 5×10^{-4}。学习率调度器为 Warmup-CosineScheduler。在每个测试子集上，重新训练模型 10 轮，其中预热的轮数为 2。为了防止过拟合，若模型在测试子集上的损失降到 0.1 以下，则停止训练。

5　实验结果与分析

5.1　RQ1：相关性分析实验结果

为了探索基于不确定性的测试样例选取指标与模型校准之间的关系，需要观察模型对具有不同预测不确定性的测试样本预测的校准程度。然而模型的校准程度无法在单个测试样本上计算，因此需要将测试样本按预测不确定性分组，计算每组的校准程度，从而达到研究二者关系的目的。

根据每个样本的不确定性指标的值对其排序，然后根据排序结果将数据样本分成 20 个大小相等的桶。给每个桶一个不确定性级别（Uncertainty Level），更高的级别意味着排名在前，桶中样本的不确定性更大。对于每个不确定性级别桶，实验计算其中样本的分类准确率、置信度、ECE 和不确定指标的值，并观察它们随不确定性级别的变化。图 2 给出了 DeepGini 指标在原始测试集上的各不确定级别桶上得到的实验结果，其余指标对应的实验结果请参考本文 Github 仓库[1)]。从图 2 中第一行和第三行的结果可以观察到，准确率和置信度都随着不确定性水平的降低而提高，并且在 3 个数据集的每个桶上置信度始终高于准确率。当不确定性级别较高（大约大于 15）时，准确度和置信度之间的差距较大。上述观察结果表明，尽管分类模型达到了有竞争力的准确率，但往往会给出校准不佳的预测。

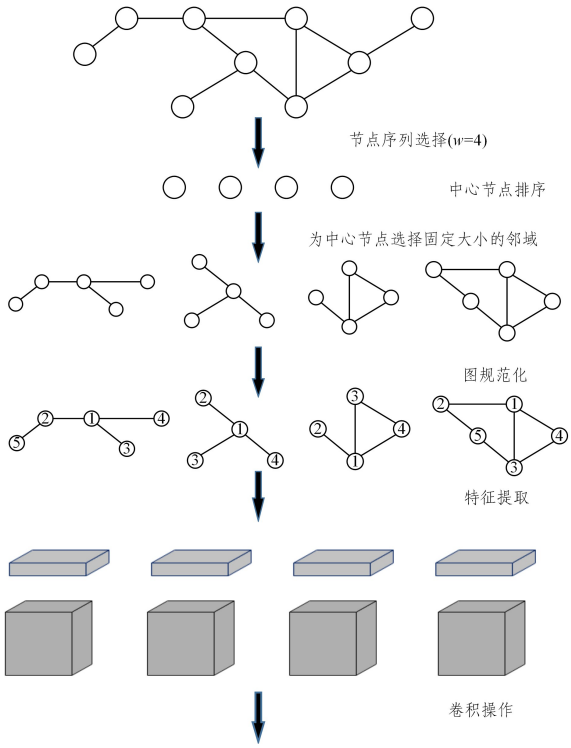


图 2　DeepGini 指标在原始测试集的各不确定级别桶上的结果

Fig. 2　Results of DeepGini indicator on original test set with different uncertainty levels

更进一步，我们计算了每个不确定性级别桶中的 ECE 和不确定指标的平均值。从图 2 的第二行和第四行可以看出，在 MNIST 和 CIFAR10 上，当不确定性水平低于 15 时，DeepGini 的值接近 0，这表明模型对测试样本的预测是确定的。在 CIFAR100 上，当不确定性等级低于 10 时，DeepGini 的值接近 0，说明对于 CIFAR100，模型给出了更不确定的预测。

一般来说，落入较高不确定等级桶中的测试样本更有可能校准得不太好。在 MNIST 和 CIFAR10 上，对于那些具有较低预测不确定性的样本，模型的预测置信度和准确率之间存在更好的对齐，并且这些样本的 ECE 接近 0，且随不确定性水平降低严格单调递减。然而，在 CIFAR100 的两个模型上，ECE 并不严格地随着不确定性水平的降低单调递减，而是随不确定性等级的降低先增大后减小。当不确定性水平大约大于 13 时，ECE 随 DeepGini 的降低而增大；当不确定性水平大约低于 13 时，ECE 随 DeepGini 的降低而减小。ECE 的最大值出现在不确定性水平较高且 DeepGini 的值较小时。

总的来说，在原始测试集上我们有以下观察结果：若模型在测试集上的预测较为确定时，具有较高预测不确定性的测试样本更有可能具有校准较差的预测，ECE 随着不确定性的降低而减小。当模型在测试集上的预测较为不确定时，具有较高预测不确定性的测试样本的校准程度虽然很差，但具有最差校准的测试样本不会出现在最高预测不确定性的测试样本上，ECE 随着不确定性的降低先增大后减小。

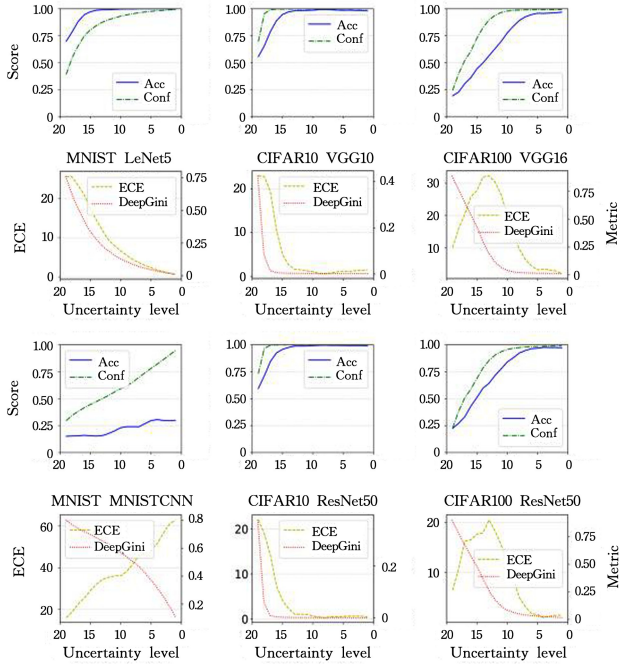


图 3　DeepGini 在 Brightness 测试集上的各不确定级别桶上的结果

Fig. 3　Results of DeepGini indicator on Brightness test set with different uncertainty levels

为了研究上述在原始测试集上的观察结果在 OOD 测试集上是否仍然成立，我们以相同的方式在 OOD 测试集上进行实验，并在图 3 中给出了 Brightness 测试集上 DeepGini 指标所对应的实验结果。

1) http://github.com/

从图 3 的第一行和第三行可以观察到，在所有测试集上，划分到每个不确定性级别桶中的预测不确定性有所提高；测试样本校准程度都变差，所有的测试集和模型上都给出了明显的校准不佳的预测。ECE 并不随着不确定性级别的降低而单调递减，而呈现先增后减的趋势(除了 MNIST-LeNet5)。

从图 3 第二行和第四行可以看出，在所有测试集和模型上，DeepGini 的值在每个不确定性级别上相比原始测试集都有所增加。例如在 CIFAR10-VGG16 上，DeepGini 的值在不确定性级别大约小于 15 时才接近为 0，而原始测试集上 DeepGini 的值在不确定性级别大约小于 17 时接近为 0。在大部分数据集和模型上(除 MNIST-LeNet5)，ECE 先增大后减小。例如，在 CIFAR10-VGG16 上，当不确定性级别大于 15 时，ECE 增大而 DeepGini 的值减小；当不确定性级别低于 15 时，ECE 和 DeepGini 的值一致下降。对于 CIFAR100，在 OOD 测试集上观察到的结果与在不确定性级别大于 10 的原始测试集上观察到的结果一致。

由于将原始测试集替换为 OOD 测试集，总体而言，模型在测试集上的预测不确定性增加，对所有测试样本的预测置信度降低，在大部分桶上的校准都较差，因此 ECE 增大。DNN 在具有高预测不确定性的样本上的校准也很差，并且在高预测不确定样本上的校准程度不随预测不确定性单调变化。因此，结合在原始测试集上的观察结果可得出结论：当模型对测试样本的预测存在较高不确定性时，ECE 并不会随着预测的不确定性而单调变化，仅当模型对测试样本的预测不确定性降低到一定程度后，ECE 开始随预测不确定性单调递减。

为了量化基于不确定性的测试样例选取指标与模型校准之间的相关性，本节利用上述实验中获得的实验结果(每个不确定性级别桶中，不确定性指标的值和 ECE)，并计算这两组数据之间的 Kendall 相关系数。使用此系数来近似模型校准程度和基于不确定性的测试样例选取指标相关性。如图 4 所示，方块的颜色越深表示二者之间的相关性越显著。深绿色表示相关性显著，而浅绿色表示相关性可忽略不计，浅黄色表示相关性为负值。其中图 4 的上半部分为 LeNet5 和 VGG16 对应的实验结果，下半部分为 MNISTCNN 和 ResNet50 对应的实验结果。

对于 MNIST 和 CIFAR10，ECE 和不确定性指标在原始测试集上具有显著的相关性。例如，在 CIFAR10 的原始测试集上，对于 ResNet50 模型，ECE 和 MaxP 之间的 Kendall 相关系数为 0.8，而在 CIFAR100 的原始测试集上，二者之间的相关性相比 MNIST 和 CIFAR10 没那么显著。

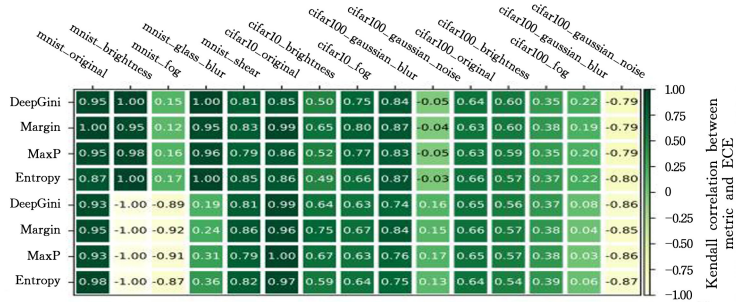


图 4 不确定指标与 ECE 之间的 Kendall 相关系数(电子版为彩图)

Fig. 4 Kendall correlation coefficients between uncertainty metrics and ECE

对于 MNIST 的 OOD 测试集，ECE 和不确定性指标之间的相关性在某些测试集和模型上有所下降。特别地，对于 LeNet5 模型，在 Fog 测试集上，二者之间的相关性相比其他数据集有显著下降。对于 MNISTCNN 模型，在 Brightness 和 Fog 测试集上，相关性为较大的负值。同样地，对于 CIFAR10 的某些 OOD 测试集，二者之间的相关性也有显著的下降，例如 ResNet50 模型，在 Noise 测试集上，DeepGini 与 ECE 之间的相关系数为 0.16，而在 Blur 测试集上，DeepGini 与 ECE 之间的相关系数为 0.74。在 CIFAR100 数据集上也观察到了类似的结果。因此可以断定 ECE 和不确定指标之间的相关性与模型和测试集有一定的关系，但关系的大小仍取决于具体的测试集和模型。

基于以上实验结果，本节得到 RQ1 的回答：1)模型在具有高预测不确定性的测试样本上校准较差，但是模型在高预测不确定性的测试样本上的 ECE 不会随着预测的不确定性而单调变化。2)当模型对测试样本的预测不确定性较高时，模型在测试样本上的 ECE 先增大后减小，当模型对测试样本的预测不确定性相对较低时，ECE 随预测不确定性单调变化。3)ECE 和不确定性指标在数据集的原始测试集上的相关性较为显著，而这种相关性在 OOD 测试集上有所下降，相关性的大小取决于数据集和模型。

5.2 RQ2：不同校准程度模型发现误分类样例的能力

在上述实验中我们发现，测试样例选取任务中的不确定指标和模型校准存在一定关系。考虑这种关系的影响以及模型校准会使模型对预测不确定的估计更为准确，我们应研究模型校准对测试样例选取任务的影响。因此本节实验首先研究校准程度较好的模型用不确定指标选择测试子集在发现模型误分类样例方面的能力是否优于校准程度较差的模型。

实验选取的基准模型为 DenseNet121，其在各测试集上的表现如表 2 所列。我们利用相对准确率 R 和测试子集在基准模型上的准确率来评估不同校准程度的模型选取测试子集的质量。在 Brightness 测试集上得到的实验结果如表 3 所列，其中相对准确率越高表示模型检测误分类样例的能力越强，基准模型准确率越低表示模型检测误分类样例的能力越强。表 4 列出了实验结果在不同 OOD 测试集和不确定指标上的平均值并统计了校准模型的实验结果优于未校准模型实验结果的百分比。

表 3　在 Brightness 测试集上发现误分类样例实验结果

Table 3　Experiment results of revealing misclassification sample on Brightness test set

	Model	Relative Acc				Baseline Acc				Relative Acc				Baseline Acc			
		Gini	Margin	MaxP	Entropy	Gini	Margin	MaxP	Entropy	Gini	Margin	MaxP	Entropy	Gini	Margin	MaxP	Entropy
CIFAR10	VGG16-NA	46.80	44.80	45.80	49.30	72.5	76.5	73.0	68.5	43.05	42.05	42.30	43.05	73.50	73.50	73.25	72.50
	VGG16-MX	52.28	48.78	48.78	51.28	66.0	71.0	67.0	67.0	44.03	43.53	44.03	44.78	67.25	67.75	67.25	67.25
	VGG16-LS	43.66	42.66	43.66	45.16	63.5	64.5	64.5	61.0	44.16	42.66	43.66	45.66	70.25	70.50	69.75	69.25
	VGG16-TS	47.80	45.30	48.80	48.80	70.0	75.0	71.0	69.5	42.80	42.30	42.80	44.05	72.25	73.25	73.5	69.50
	ResNet50-NA	42.72	41.72	42.72	43.72	61.0	62.5	61.5	59.0	40.22	39.72	39.97	39.72	62.75	63.50	63.00	62.25
	ResNet50-MX	55.25	52.75	54.75	56.25	59.0	59.0	59.0	57.0	46.00	46.00	46.00	45.50	58.75	59.00	59.00	58.75
	ResNet50-LS	51.74	50.24	50.24	52.24	55.0	55.5	56.5	56.0	45.49	45.49	45.49	44.74	58.75	59.00	58.75	59.00
	ResNet50-TS	42.72	43.22	44.72	45.22	60.5	61.5	59.0	58.5	39.47	40.22	39.22	41.47	63.00	62.50	62.75	62.25
CIFAR100	VGG16-NA	57.60	50.60	58.60	57.10	43.5	41.5	43.5	41.5	56.60	50.60	56.60	56.10	43.50	47.75	43.25	45.50
	VGG16-MX	62.50	52.50	63.00	61.00	32.5	42.0	33.5	34.5	60.00	52.00	59.50	58.75	39.50	47.00	41.25	42.00
	VGG16-LS	55.77	55.77	60.77	52.27	41.5	46.0	41.0	43.0	58.52	53.77	58.27	56.52	43.00	48.25	42.00	44.25
	VGG16-TS	57.10	51.10	58.60	54.10	42.0	42.0	42.0	42.5	55.85	52.85	56.10	54.60	45.25	43.25	48.75	46.25
	ResNet50-NA	60.88	53.88	62.38	61.38	33.5	41.5	38.5	34.0	55.63	50.63	57.88	53.88	39.25	45.00	39.00	40.00
	ResNet50-MX	58.55	57.05	57.55	61.05	31.5	37.5	33.5	30.0	58.30	56.80	58.55	57.55	31.00	41.00	33.25	30.00
	ResNet50-LS	61.44	52.44	60.94	59.94	31.5	42.5	31.5	33.0	55.44	49.69	55.69	54.94	32.25	45.25	33.00	31.25
	ResNet50-TS	61.38	53.38	61.88	60.88	34.5	42.0	36.0	35.0	54.63	50.63	56.88	52.63	39.00	44.00	40.00	40.00

表 4　发现误分类样例实验中校准模型表现占优的比率

Table 4　Ratio of calibrated models outperforming uncalibrated models in revealing misclassification sample experiments

Uncertainty	Gini	Entropy	MaxP	Margin	Gini	Entropy	MaxP	Margin
Brightness	62.50	83.33	66.67	62.50	87.50	66.67	91.67	79.17
Fog	37.50	70.83	54.17	33.33	83.33	83.33	75.00	70.83
Blur	54.17	79.17	62.50	41.67	66.67	62.50	75.00	75.00
Noise	70.83	66.67	75.00	58.33	70.83	70.83	75.00	66.67

如表 3 所列，4 种校准程度不同的模型选出的测试子集在上述两个指标的评估上有明显的区别。校准后的模型选择的测试子集在基准模型准确率和相对准确率的度量上都明显优于未校准模型选择的测试子集。例如 CIFAR10 数据集标注成本为 2%，校准程度较好的 VGG16-MX 用 DeepGini 选择的测试子集的相对准确率和基准模型准确率分别为 52.28 和 66.0，而校准程度较差的 VGG16-NA 选择的测试子集对应的结果为 46.8 和 72.5。二者之间的差距较为明显。

同时，使用 Temperature Scaling 校准的模型所选择的测试子集质量相比未校准的模型并无明显优势。例如 CIFAR10 数据集标注成本为 4%，校准程度较好的 ResNet50-TS 用 DeepGini 选择的测试子集的相对准确率和基准模型准确率分别为 39.47 和 63.0，而校准程度较差的 ResNet50-NA 选择的测试子集对应的结果为 40.22 和 62.75。二者之间差距不大，但是校准后模型的表现不如未校准模型。出现上述现象的原因主要有以下两点：1）Temperature Scaling 适合在原始测试集上校准，减小模型在原始测试集上的 ECE，若测试集的分布和原始测试集相差较大，则该方法校准效果不佳。2）该方法只改变 Softmax 的置信度，使模型的预测更加尖锐或平滑，但不会改变模型准确率，对于某些不确定指标，如 MaxP，不会改变根据该指标排序的样本的相对顺序，因而失效。

如表 4 所列，总体而言，校准程度较好的模型明显优于校准程度差的模型。使用相对准确率度量，在 13/16 组实验结果中，校准程度好的模型选择的测试子集质量优于校准程度较差的模型，平均 61.19%经过校准训练后的模型对应的实验结果优于未校准模型对应的实验结果。使用基准模型准确率度量，16/16 组实验结果显示校准程度好的模型优于校准程度差的模型，平均 70.57%经过校准训练后的模型对应的实验结果优于未校准模型对应的实验结果。基于上述实验结果，我们得到 RQ2 的回答：用 Mixup 和 Label Smoothing 训练得到的校准程度好的模型相比未校准模型能选择包含更多误分类样例的测试子集，而 Temperature Scaling 在这方面的能力并不突出。

5.3　RQ3：不同校准程度模型提升重训练模型表现的能力

为了比较使用不同校准程度模型选择的测试子集对提升重训练模型表现的影响，我们用不同校准程度模型选择的测试子集重训练基准模型。表 5 列出了在 Brightness 测试集上重训练基准模型的准确率。

表 6 列出了实验结果在不同 OOD 测试集和不确定指标上的平均值并统计了校准模型的实验结果优于未校准模型实验结果的百分比。

如表 5 所列，使用不同校准程度模型选择的测试子集来重训练模型没有太大的区别。例如，在 CIFAR10 数据集中，标注成本为 2% 和 4% 时，校准较好的 VGG16-MX 使用 DeepGini 选择的测试子集的重训练模型准确率分别为 94.38 和95.1，而校准程度较差的 VGG16-NA 对应的结果为 94.44 和 95.32，二者之间的差距不大。

如表 6 所列，总体而言，校准模型的重训练实验结果优于未校准模型的比例为 10/16，平均 51.81%的校准模型对应的实验结果优于未校准模型对应的实验结果，二者之间差距不明显。出现上述现象的原因可能有以下 3 点：1）校准模型和未校准模型之间的 ECE 差距不大，例如在 CIFAR10 数据集上，VGG16-NA 在 Brightness 测试集上的 ECE 为 4.97，而 VGG16-MX 的 ECE 为 2.65。因此，选取测试子集的质量虽然可以区分，但差距不明显，不足以给重训练过程提供较大的

帮助。2)研究表明,重训练的方式和候选测试集中的数据分布对重训练模型的表现都会有所影响[35]。因此在本文的实验设置下,使用包含更多误分类样例的测试子集来重训练模型并不一定意味着表现有更大的提升。3)选取测试子集的模型与重训练模型不一致,这导致选取的测试子集可能对重训练模型没有帮助。

表 5 Brightness 测试集上模型重训练实验结果

Table 5 Experiment results of model retraining on Brightness test set

	Model	Retrain Acc(2%)				Retrain Acc(4%)			
		Gini	Margin	MaxP	Entropy	Gini	Margin	MaxP	Entropy
CIFAR10	VGG16-NA	94.44	94.15	94.05	94.71	95.32	94.45	94.34	94.78
	VGG16-MX	94.38	94.23	94.10	94.75	95.10	95.33	95.25	94.91
	VGG16-LS	94.60	94.69	94.87	94.76	94.89	94.67	95.14	94.46
	VGG16-TS	94.61	94.33	94.27	94.99	95.03	94.48	94.40	95.39
	ResNet50-NA	94.82	94.51	94.61	94.42	95.27	95.17	95.03	95.46
	ResNet50-MX	94.05	93.93	94.25	94.03	95.54	95.21	95.26	95.45
	ResNet50-LS	94.28	94.16	93.93	94.43	95.19	95.21	95.18	94.64
	ResNet50-TS	94.65	94.67	94.69	94.75	95.29	95.21	95.25	95.32
CIFAR100	VGG16-NA	72.95	71.86	73.63	73.17	75.25	74.19	74.65	75.41
	VGG16-MX	74.12	73.15	73.13	73.51	75.01	74.79	75.17	75.19
	VGG16-LS	73.60	73.79	73.22	72.83	76.40	74.65	75.67	75.98
	VGG16-TS	73.36	71.65	73.25	72.45	74.77	74.15	75.30	75.83
	ResNet50-NA	74.39	72.77	74.76	72.61	75.73	74.47	75.19	76.33
	ResNet50-MX	74.24	74.21	72.85	74.81	75.21	73.92	73.56	75.28
	ResNet50-LS	73.45	73.29	72.58	74.61	75.58	74.13	75.77	75.96
	ResNet50-TS	73.50	72.26	73.74	72.96	75.21	74.50	75.58	75.33

表 6 重训练实验中校准模型的表现占优的比率

Table 6 Ratio of calibrated models outperforming uncalibrated models in retraining experiments

Uncertainty	Gini	Margin	MaxP	Entropy
Brightness	33.33	70.83	62.50	54.17
Fog	54.17	58.33	54.17	62.50
Blur	62.50	41.67	45.83	41.67
Noise	54.17	45.83	54.17	33.33

基于上述实验结果,本节得出 RQ3 的回答:尽管校准较好的模型能选出包含更多误分类样例的测试子集,但该测试子集并不一定对重训练有帮助。一个模型选出的测试子集可能并不会对另一个模型的重训练有帮助。此外,重训练模型需要考虑到候选测试集中样本的分布。

结束语 本文指出了现有测试样例选取工作中存在的缺陷:利用预测不确定性进行测试样例选取的工作中没有考虑模型校准的影响,忽略了模型对不确定性的估计是否准确的问题。用 Kendall 系数衡量模型校准和任务中使用的预测不确定指标相关性,可发现二者之间存在一定关系,且校准程度更高的模型选择的测试子集在发现误分类样例方面的质量更高。未来我们将继续探索在其他领域的模型和数据集上,如自然语言处理任务、RNN 架构和 Self-Attention 架构的模型上,模型校准与基于不确定性的测试样例选取指标之间的关系;考虑提出基于不确定性的测试样例选取指标,但不受模型校准程度的影响。

参 考 文 献

[1] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[2] BOJARSKI M, DEL TESTA D, DWORAKOWSKI D, et al. End to end learning for self-driving cars[J]. arXiv: 1604. 07316, 2016.

[3] PEI K, CAO Y, YANG J, et al. Deepxplore: Automated white-box testing of deep learning systems[C]// Proceedings of the 26th Symposium on Operating Systems Principles. 2017: 1-18.

[4] MA L, XU J F, ZHANG F, et al. Deepgauge: Multi-granularity testing criteria for deep learning systems[C]// Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering. 2018: 120-131.

[5] KIM J, FELDT R, YOO S. Guiding deep learning system testing using surprise adequacy[C]// 2019 IEEE/ACM 41st International Conference on Software Engineering(ICSE). IEEE, 2019: 1039-1049.

[6] MA W, PAPADAKIS M, TSAKMALIS A, et al. Test selection for deep learning systems[J]. ACM Transactions on Software Engineering and Methodology(TOSEM), 2021, 30(2): 1-22.

[7] FENG Y, SHI Q, GAO X, et al. Deepgini: prioritizing massive tests to enhance the robustness of deep neural networks[C]// Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis. 2020: 177-188.

[8] GUO C, PLEISS G, SUN Y, et al. On calibration of modern neural networks[C]// International Conference on Machine Learning. PMLR, 2017: 1321-1330.

[9] ZHANG H Y, CISSÉ M, DAUPHIN Y N, et al. mixup: Beyond Empirical Risk Minimization[C]// International Conference on Learning Representations. 2018.

[10] PEREYRA G, TUCKER G, CHOROWSKI J, et al. Regularizing Neural Networks by Penalizing Confident Output Distributions [C]// International Conference on Learning Representations. 2017.

[11] YUAN Y, PANG Q, WANG S. You Can't See the Forest for Its Trees: Assessing Deep Neural Network Testing via Neural Coverage[C]// 2023 IEEE/ACM 45st International Conference on Software Engineering(ICSE). IEEE, 2023.

[12] WEISS M, TONELLA P. Simple techniques work surprisingly

well for neural network test prioritization and active learning (replicability study)[C]// Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis. 2022:139-150.

[13] ZHANG X,XIE X,MA L,et al. Towards characterizing adversarial defects of deep learning software from the lens of uncertainty[C]// Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering. 2020:739-751.

[14] ZHAO C,MU Y,CHEN X,et al. Can test input selection methods for deep neural network guarantee test diversity? A large-scale empirical study[J]. Information and Software Technology, 2022,150:106982.

[15] SHEN W,LI Y,CHEN L,et al. Multiple-boundary clustering and prioritization to promote neural network retraining[C]// Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering. 2020:410-422.

[16] GAO X,FENG Y,YIN Y,et al. Adaptive test selection for deep neural networks[C]// Proceedings of the 44th International Conference on Software Engineering. 2022:73-85.

[17] HU Q,GUO Y,XIE X,et al. Aries:Efficient Testing of Deep Neural Networks via Labeling-Free Accuracy Estimation[C]// 2023 IEEE/ACM 45st International Conference on Software Engineering(ICSE). IEEE,2023.

[18] ARRIETA A. Multi-objective metamorphic follow-up test case selection for deep learning systems[C]// Proceedings of the Genetic and Evolutionary Computation Conference. 2022: 1327-1335.

[19] ZADROZNY B,ELKAN C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers[C]// ICML. 2001:609-616.

[20] THULASIDASAN S,CHENNUPATI G,BILMES J A,et al. On mixup training:Improved calibration and predictive uncertainty for deep neural networks[C]// Advances in Neural Information Processing Systems. 2019:13888-13899.

[21] MÜLLER R,KORNBLITH S,HINTON G E. When does label smoothing help? [C]// Advances in Neural Information Processing Systems. 2019:4696-4705.

[22] QIN Y,WANG X,BEUTEL A,et al. Improving calibration through the relationship with adversarial robustness[J]. Advances in Neural Information Processing Systems, 2021, 34: 14358-14369.

[23] LI Z,MA X,XU C,et al. Operational calibration:Debugging confidence errors for dnns in the field[C]// Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. 2020:901-913.

[24] LI D,HU B,CHEN Q. Calibration Meets Explanation:A Simple and Effective Approach for Model Confidence Estimates[C]// Conference on Empirical Methods in Natural Language Processing. 2022.

[25] MINDERER M,DJOLONGA J,ROMIJNDERS R,et al. Revisiting the calibration of modern neural networks[J]. Advances in Neural Information Processing Systems,2021,34:15682-15694.

[26] DENG L. The mnist database of handwritten digit images for machine learning research[J]. IEEE Signal Processing Magazine,2012,29(6):141-142.

[27] KRIZHEVSKY A,HINTON G. Learning multiple layers of features from tiny images[J/OL]. http:// cs. toronto. edu/~kriz/learning-features-2009-TR. pdf.

[28] MU N,GILMER J. Mnist-c:A robustness benchmark for computer vision[J]. arXiv:1906. 02337,2019.

[29] HENDRYCKS D,DIETTERICH T. Benchmarking neural network robustness to common corruptions and perturbations [C]// International Conference on Learning Representations. 2019.

[30] HU Q,GUO Y,CORDY M,et al. An empirical study on data distribution-aware test selection for deep learning enhancement [J]. ACM Transactions on Software Engineering and Methodology(TOSEM),2022,31(4):1-30.

[31] HE T,ZHANG Z,ZHANG H,et al. Bag of tricks for image classification with convolutional neural networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:558-567.

[32] SUTSKEVER I,MARTENS J,DAHL G,et al. On the importance of initialization and momentum in deep learning[C]// International Conference on Machine Learning. PMLR, 2013: 1139-1147.

[33] KROGH A,HERTZ J. A simple weight decay can improve generalization[C]// Advances in Neural Information Processing Systems. 1991:950-957.



ZHAO Tong, born in 1998, postgraduate. His main research interests include deep learning and software engineering.



SHA Chaofeng, born in 1976, Ph.D, associate professor. His main research interests include machine learning, data mining, and natural language processing.

（责任编辑：杨雪敏）