Calibration Bottleneck: Over-compressed Representations are Less Calibratable

Deng-Bao Wang [1 2] **Min-Ling Zhang** [1 2]

Abstract

Although deep neural networks have achieved remarkable success, they often exhibit a significant deficiency in reliable uncertainty calibration. This paper focus on *model calibratability*, which assesses how amenable a model is to be well recalibrated post-hoc. We find that the widely used weight decay regularizer detrimentally affects model calibratability, subsequently leading to a decline in final calibration performance after post-hoc calibration. To identify the underlying causes leading to poor calibratability, we delve into the calibratability of intermediate features across the hidden layers. We observe a U-shaped trend in the calibratability of intermediate features from the bottom to the top layers, which indicates that over-compression of the top representation layers significantly hinders model calibratability. Based on the observations, this paper introduces a *weak classifier hypothesis*, i.e., given a weak classification head that has not been over-trained, the representation module can be better learned to produce more calibratable features. Consequently, we propose a progressively layer-peeled training (PLP) method to exploit this hypothesis, thereby enhancing model calibratability. Our comparative experiments show the effectiveness of our method, which improves model calibration and also yields competitive predictive performance.

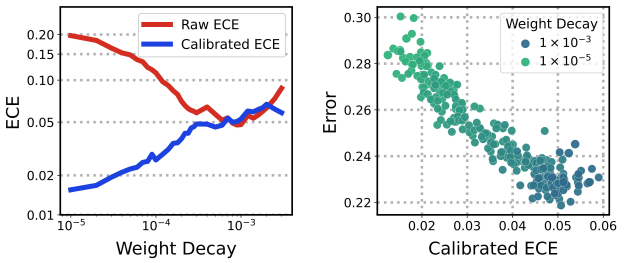
1. Introduction

In machine learning systems, a reliable predictive models should not only yield high accuracy but also offer heightened uncertainty when their predictions are prone to inaccuracy. While modern deep neural networks (DNNs) have achieved remarkable success in high-dimensional prediction tasks like computer vision, speech recognition, and natural language processing, they have exhibited a notable deficiency in reliably estimating uncertainty (Guo et al., 2017; Minderer et al., 2021; Wang et al., 2024). This uncertainty issue can lead to adverse consequences, particularly in scenarios involving safety-critical decision-making, and therefore, it has been the subject of extensive research in recent years (Gupta et al., 2021; Ashukha et al., 2019; Wang et al., 2021; 2023). To systematically investigate DNNs' uncertainty estimation problem, Guo et al. (2017) conducted a comprehensive study within this context and made two significant observations: (i) The miscalibration of model confidence, which can be used to reflect the uncertainty degree, is closely associated with large model capacity and the absence of regularization in training. And (ii) simple post-hoc methods like temperature scaling (TS) (Platt et al., 1999) and histogram binning (HB) (Zadrozny & Elkan, 2001) can effectively address the miscalibration issue.

Taking inspiration from the study of (Guo et al., 2017), two predominant strategies have been extensively investigated to improve the calibration performance of DNNs: Train-time calibration and post-hoc calibration. For the former strategy, some well-known methods which initially designed to improve generalization have been found also beneficial for calibration, including label smoothing (Müller et al., 2019), mixup (Thulasidasan et al., 2019), self-distillation (Guo et al., 2021) and focal loss (Mukhoti et al., 2020). The latter strategy aims to recalibrate predictions in a post-hoc fashion by establishing a mapping from raw outputs to well-calibrated confidences. Typically, these approaches usually incorporate extra parameters which necessitate tuning on further validation data during the post-hoc calibration. There is a surge of studies aimed at improving calibration by designing new post-hoc calibration approaches (Müller et al., 2019; Mukhoti et al., 2020; Joo & Chung, 2021; Kull et al., 2019; Rahimi et al., 2020; Gupta et al., 2021). Although numerous post-hoc calibration approaches have been proposed in recent years, TS remains the most widely used one due to its simplicity and effectiveness. It is worth noting that TS only needs one single parameter, the temperature of softmax layer, to be adjusted in post-hoc calibration phase. Moreover, different from binning-based methods like HB, TS does not compromise the dense output confidences in post-hoc calibration phase.

[1]School of Computer Science and Engineering, Southeast University, Nanjing, China [2]Key Lab. of Computer Network and Information Integration (Southeast University), MOE, China. Correspondence to: Min-Ling Zhang <zhangml@seu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.



(a) Raw and calibrated ECE (b) Error and calibrated ECE.

Figure 1. Results of ResNet-18 on CIFAR-100. (a) Raw and calibrated ECE (based on TS) with varied weight decay strengths. (b) Calibrated ECE and classification error of a bunch of models trained with varied weight decay coefficients. Please refer to the Appendix A.1 for details on all the experiments in this paper.

As the aforementioned two types of methods can independently enhance calibration performance in different phases, it seems logical to combine them for superior performance. However, previous studies have shown that while methods like label smoothing, mixup and self-distillation, are effective in enhancing calibration performance during training, they do not necessarily result in superior result than models by standard training after post-hoc calibration (Wang et al., 2021; Zhu et al., 2023; Wang et al., 2023). As Ashukha et al. (2019) suggested, comparison between different models might be inequitable if the post-hoc calibration is not taken into consideration. Therefore, in this study, rather than addressing the calibration challenge encountered during training, we shift our focus to the concept of **calibratability**. Informally, model calibratability can be defined as:

> *The calibratability of model $\boldsymbol{f}$ in terms of a specific post-hoc calibrator ψ, indicates this model's ability of leveraging ψ to gain further improvement on its calibration performance.*

Specifically, we say that model $\boldsymbol{f}_1$ is more calibratable than model $\boldsymbol{f}_2$, if the calibration performance of the composite model $\boldsymbol{f}_1 \circ \psi$ is superior to that of $\boldsymbol{f}_2 \circ \psi$. In this work, we particularly observe that weight decay, the de-facto standard regularization technique for training modern DNNs, tends to adversely make models less calibratable, despite it has been identified by Guo et al. (2017) as a significant factor in improving calibration. To ground our motivation, we first make the following observations.

Observation 1. As is shown in Figure 1(a), we find that *raw* Expected Calibration Error (ECE) would decrease by increasing the regularization strength, well after a specific point it would be negatively impacted by having too much weight decay. However, *calibrated* ECE would continue to increase when more regularization is added and it is almost always lower than raw ECE. This observation suggests that carefully selecting a regularization strength that minimizes ECE may not be as effective as applying standard training and then performing post-hoc calibration.

Observation 2. As is shown in Figure 1(b), we observe a trade-off between calibrated ECE and classification error among a bunch of models trained with varied regularization strengths. This observation raises a particularly thorny question: Are calibratability and discriminability inherently at odds within deep neural networks? Moreover, Figure 9 shows that calibrated ECE worsens over training epochs, signaling a decline in calibratability during learning dynamics. As the classification error usually decreases continuously during training, this also raises the concern about the conflict between calibratability and discriminability.

Given the above observations, this paper aims to investigate why deep neural networks are less calibratable, particularly when applying regularization during training. To this end, we delve into the calibratability of the intermediate features of the hidden layers, and illustrate that the calibrated ECE of the linear classifiers build on these intermediate features across layers from the bottom to the top exhibit a trend of initial decline, followed by a remarkable increase at a certain layer. We demonstrate through empirical evidences that this phenomenon is related to the information bottleneck principle of deep learning, which results in an unexpected side effect of the information compression on model calibration. We suggest that the cause of calibrability degradation may be the over-training of a few top layers and the subsequent over-compression of hidden features, introducing the *weak classifier hypothesis*: Learning more calibratable representations that requires a weaker classification head. Building upon this, we propose an efficient and effective training strategy called Progressively Layer-Peeled (PLP) training to overcome the calibration bottleneck, thereby enhancing model calibratability. Intuitively, our approach gradually freeze the parameters of top layers during the training process, to ensure that the top layers of a neural network do not excessively perform information compression. We conduct experiments on several image classification datasets, and the results demonstrate that our method improves the calibrated performance without significantly compromising the predictive performance.

Our contributions can be summarized as four-fold:

- We find that weight decay actually diminishes DNNs' calibratability, which raises concerns about the inherent relationship of calibratability and discriminability, both of which are crucial to achieve reliable classification.
- We delve into the calibratability of intermediate features, revealing that the calibratability of features exhibits a U-shaped trend, with an initial decline and subsequent increase, from the bottom to the top layers.
- We suggest that the information bottleneck principle is responsible for the decline in model calibratability, and empirically demonstrate that over-compression of representations leads to less calibratable models.

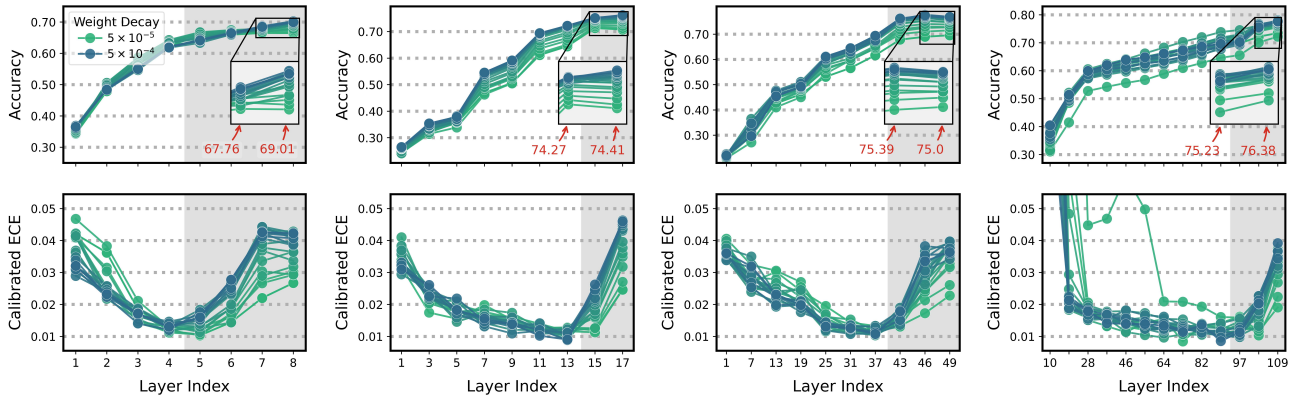
- We introduce a hypothesis stating that learning calibratable representations necessitates the use of weak heads. Inspired by this hypothesis, we propose a simple yet effective method called PLP. Experimental results show that PLP achieves superior balance between model calibratability and predictive performance.

2. Background

Train-time Calibration. There has been a surge of research on improving model calibration by utilizing implicit or explicit regularization techniques during the training. Most of these techniques follow the principle of *confidence penalty* to address the overconfidence issue of DNNs (Guo et al., 2017; Müller et al., 2019; Patra et al., 2023; Tao et al., 2023a; Wei et al., 2022). Label Smoothing, a technique widely used to mitigate overfitting, shown to improve model calibration by preventing the networks from producing overconfident predictions (Szegedy et al., 2016; Müller et al., 2019). Wang et al. (2021) suggest that label smoothing can be viewed as equivalent to applying maximum-entropy regularization on model training. Patra et al. (2023) and Joo & Chung (2021) proposed to directly leverage maximum-entropy-based regularization terms in loss to penalize the overconfident outputs. There are also some implicit regularization methods that are initially proposed to improve generalization, have also been found beneficial for calibration, such as mixup training, self-distillation and focal loss (Thulasidasan et al., 2019; Zhang et al., 2022; Guo et al., 2021; Mukhoti et al., 2020; Tao et al., 2023b). For example, Mukhoti et al. (2020) found that focal loss can be viewed as an upper bound on the regularized KL-divergence loss, where the regularizer is the negative entropy of model outputs, and hence replacing cross-entropy with focal loss has the effect of adding a maximum-entropy regularization in model training. In addition to the methods mentioned above, weight decay, which is a simpler method and now the standard de-facto regularization technique in deep learning, has also been found to significantly impact model calibration in the pioneering study by (Guo et al., 2017). It is shown that by setting appropriate weight decay coefficients, it can greatly improve model calibration performance.

Post-hoc Calibration. Post-hoc calibration approaches work by recalibrating the model outputs after training, and this kind of approaches has been widely studied for many years (Platt et al., 1999; Zadrozny & Elkan, 2001; Naeini et al., 2015; Patel et al., 2021; Tomani et al., 2021). Among these approaches, scaling-based approaches is widely considered the simplest and most effective. Platt Scaling (PS) is a classical parametric approach to calibrate binary classifiers (Platt et al., 1999). It learns scalar parameters $a, b \in \mathbb{R}$ and outputs $\hat{q} = s\left(az + b\right)$ as the calibrated probability, where z_i is a non-probabilistic model output s is the Sigmoid function. Learnable parameters a and b can be optimized using the Negative Log-Likelihood (NLL) loss over the validation set. Temperature Scaling (TS) is an extension of PS for multi-class classification tasks and neural network models. It only requires learning a single learnable parameter, the temperature of the Softmax layer, to recalibrate the model outputs. Given a logit vector $\boldsymbol{z}$ with C dimensions, the recalibrated confidence after TS can be expressed as: $\hat{q} = \max_i \frac{\exp(z_i/T)}{\sum_{c=1}^{C} \exp(z_c/T)}$. By scaling the logit vector into a new vector wich same dimension, the sharpness of output probabilities can be changed. Specifically, TS softens the output probabilities with temperature $T > 1$ and sharpens the probabilities with temperature $T < 1$. The scaling operation can be instantiated with other forms that involve more learnable scaling parameters. Vector scaling and matrix scaling learn a specific scalar for each class with linear transformation $\boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}$, where $\boldsymbol{W}$ is a full matrix and restricted to be a diagonal matrix in matrix scaling and vector scaling respectively. In addition to scaling-based approaches, other forms of post-hoc calibration approaches such as binning-based and isotonic regression-based approaches have also been used for calibrating deep models (Zadrozny & Elkan, 2001; Patel et al., 2021; Zadrozny & Elkan, 2002). However, these methods produce discrete confidences and usually require larger validation sets, while their effectiveness is often inferior to scaling-based methods (Guo et al., 2017).

Model Calibratability. In general, model calibratability considers the interaction of training-time calibration and post-hoc calibration methods. Several studies have focused on this setting and found that these two types of methods may not necessarily be mutually beneficial (Wang et al., 2021; Bouniot et al., 2023; Tao et al., 2024). The preliminary investigation in (Wang et al., 2021) shown that although label smoothing and maximum entropy regularization do help model calibration by penalizing overconfident outputs, they do not work well in conjunction with post-hoc calibration techniques. They further found that the confidence penalty effect of mixup also harm model calibratability and proposed to mitigate this issue by avoiding the implicit label smoothing operation in mixup (Wang et al., 2023). Zhu et al. (2023) found that data distillation tends to discard semantically meaningful information and models trained on distilled datasets are not calibratable. Currently, there are only a few works studied model calibratability, and these studies primarily consider mitigating the negative impact of regularization to improve existing train-time calibration methods. It is unclear why these regularization methods can diminish model calibratability. In this paper, we mainly focus on weight decay, a standard regularization technique in modern deep learning, and attempt to delve deeper into the underlying reasons for the decline in calibratability.



(a) VGG-11 (b) ResNet-18 (c) ResNet-50 (d) ResNet-110

Figure 2. The accuracy (top) and calibrated ECE (bottom) of linear probing over hidden layers on CIFAR-100 with different model architectures. The models are trained with different weight decay coefficients, varying in the range of 5×10^{-5} to 5×10^{-4} with intervals of 2.5×10^{-5}. The highlighted values in the first row represents the average accuracy of linear probing across all models.

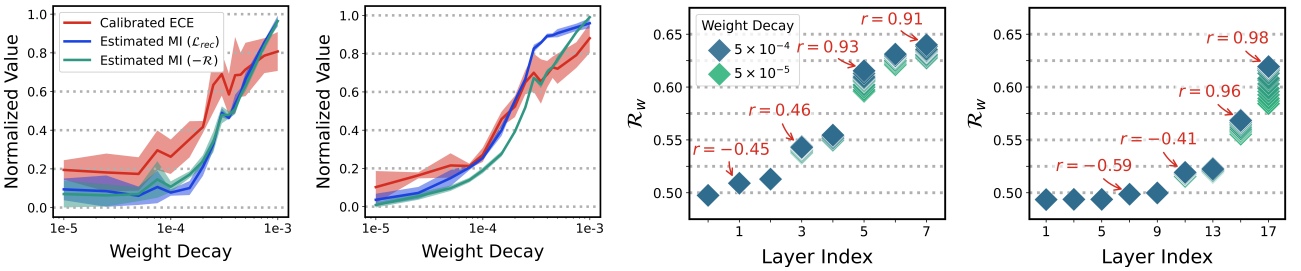
3. What Makes Models Less Calibratable?

3.1. On Calibratability of Intermediate Features

Given the significance of calibratability, it is imperative to understand why regularized training can undermine model calibrability, especially considering that regularization is recognized as crucial for the effective training of DNNs. In the above experiments, we observed a gradual deterioration of calibratability over training time, especially when strong regularization is involved. To gain further insight, it would be interesting to explore how the forward propagation process affects calibratability. Therefore, this section initiates the investigation by evaluating the calibratability of intermediate features across hidden layers of the neural networks. We build linear classifiers on the intermediate features of well trained neural networks, and examine the calibratability of these linear classifiers. Specifically, given a pretrained neural network, we first freeze its bottom k layers, then perform linear probing on these frozen layers by stochastic gradient descent optimization. Despite the feature dimensions are different, the linear classifiers upon different layers are trained with the same learning scheme. For the implementation details such as optimization policies and frozen layer numbers, please refer to Appendix A.1.

The U-shaped calibratability. Figure 2 shows the results of three different models on CIFAR-100. Each curve represents the linear probing performance over the hidden layers of the same well trained model. We can observe that as the layer depth increases, the accuracy of linear probing gradually increases, which aligns with our intuition. However, the calibrated ECE over the hidden layers follows an intriguing trend: The bottom layers generally yield relatively large calibrated ECE, which tend to decrease as we move deeper; Then, the deepest few layers, those near the top of neural networks, exhibit a striking raising on calibrated ECE, while these layers contribute little to the predictive performance. The U-shaped calibratability collapse is observed across various network architectures, while the numbers of those specific top layers at which the calibration performance starts to deteriorate may vary depending on architectures. This phenomenon indicates that although models trained with different regularization strengths can have significantly different performance, their calibrated ECE at a certain intermediate hidden layer is very close. Interestingly, the optimal calibrated ECE achieved by different curves is almostly converge to the same level. As we discussed above, there is a negative correlation between the accuracy and calibrated ECE when employing different regularization strengths. Here, if we focus on the hidden layers, the relationship between the two becomes more complex. By dividing a neural network into two parts, the correlation between these two properties can be decoupled. For the bottom (top) layers, accuracy and calibrated ECE exhibit a clear negative (positive) correlation.

The role of top layers. As is highlighted in Figure 2, the most top layers usually offer minimal gains in terms of accuracy while inflicting significant damage to calibrated ECE. For instance, the average accuracy of linear probing on the 17th layer of ResNet-18 is very close to that on the 15th layer, but at the cost of a significant increase in calibrated ECE. It suggests that discarding certain feature extraction layers at the top and replacing them with simple linear layers can enhance calibration performance without compromising much accuracy. This leads us to rethink the role of the top layers of neural networks as they induce an undesired calibration bottleneck. Moreover, the experiments demonstrate that this phenomenon exists across models with different depths. In comparison to deeper models like ResNet-18 and ResNet-50, the proportion of layers impairing calibratability is even higher for VGG-11.



(a) VGG-11 (b) ResNet-18 (c) VGG-11 (d) ResNet-18

Figure 3. (a-b) The trend of the calibrated ECE and mutual information estimated by two methods with varying regularization strength. (c-d) Decoder-based reconstruction loss of models trained with different weight decay coefficients. The highlighted red values indicate the Pearson correlation coefficient between the reconstruction loss and the calibrated ECE for each layer.

3.2. Over-compression May Hurt Calibratability

It is widely recognized that that the layer-wise compression of data by neural networks is crucial for their generalization performance. Shwartz-Ziv & Tishby (2017) and Saxe et al. (2019) revisited the information bottleneck principle (Tishby et al., 2000) to explain deep neural networks. They found that during training, deep neural networks first fit the data and then compress the information carried by the hidden layer features, and the compression intensity increases with the layer depth. Given that information compression, as noted in prior research, and the calibratability collapse we have observed both occur in the top layers, it naturally leads us to hypothesize that the information compression may be responsible for the diminished calibratability performance. In this subsection, we attempt to establish a connection between these two aspects through analytical experiments.

Specifically, the mutual information $\boldsymbol{I}(\boldsymbol{x}, \boldsymbol{z})$ between the original feature $\boldsymbol{x}$ and the intermediate feature $\boldsymbol{h}$ can be use to measure the amount of retained information in intermediate feature $\boldsymbol{h}$. Due to the intractability of the mutual information in high dimensional space[1], we adopt the following two methods to estimate the mutual information between $\boldsymbol{x}$ and $\boldsymbol{h}$. For more implementation details and a discussion on the rationale behind this strategy for estimating mutual information, please refer to the description in Appendix A.1.

Rate distortion-based estimation. Rate distortion is a concept in the field of lossy data compression, which measures the *compactness* of a random distribution. Given a random variable $\boldsymbol{z}$ and a specified value $\epsilon > 0$, the rate distortion $R(\boldsymbol{z}, \epsilon)$ denotes the minimal number of binary bits required to encode $\boldsymbol{z}$ such that the expected decoding error is less than ϵ, i.e., the reconstructed $\hat{\boldsymbol{z}}$ satisfies $\mathbb{E}\left[\|\boldsymbol{z} - \hat{\boldsymbol{z}}\|_2\right] \leq \epsilon$. Here, we use the estimation method proposed in (Ma et al., 2007) to calculate the rate distortion, which has been successful used in deep learning (Yu et al., 2020).

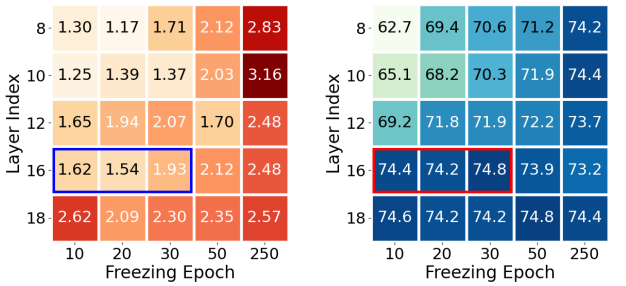
Decoder-based estimation. As is discussed in (Wang et al., 2020), the mutual information can be bound by: $\boldsymbol{I}(\boldsymbol{x}, \boldsymbol{h}) = H(\boldsymbol{x}) - H(\boldsymbol{x} \mid \boldsymbol{h}) \geq H(\boldsymbol{x}) - \mathcal{R}(\boldsymbol{x} \mid \boldsymbol{h})$, where $\mathcal{R}(\boldsymbol{x}|\boldsymbol{h})$ denotes the expected reconstruction error and $H(\boldsymbol{x})$ denotes the marginal entropy of $\boldsymbol{x}$, as a constant. Therefore, we can estimate $\mathcal{R}(\boldsymbol{x}|\boldsymbol{h})$ by training a decoder to measure the minimal reconstruction loss: $I(\boldsymbol{h}, \boldsymbol{x}) \approx \max_{\boldsymbol{w}} \left[H(\boldsymbol{x}) - \mathcal{R}_{\boldsymbol{w}}(\boldsymbol{x} \mid \boldsymbol{h})\right]$. Similar to the linear probing experiments in the previous subsection, we build decoder upon the intermediate layers and fit it to the training data to measure the minimal reconstruction loss it can achieve.

Figure 3(a) and 3(b) illustrate the calibrated ECE of linear probing on the features of the last representation layer with different regularization strengths, along with the mutual information estimated by the above two methods. It can be observed that there is a high correlation between the degree of information compression and model calibratability. When training with stronger regularization, the model tends to exhibit a more pronounced data compression effect, which often leads to an improvement in the model's representational capacity. However, information compression tends to compromise the model's calibratability to some extent, which explains the observed negative correlation between accuracy and calibrated ECE in Figure 1.

Figure 3(c) and 3(d) present the mutual information between original features and intermediate features estimated by decoder-based method. It can be observed that as the depth increases, the reconstruction loss becomes larger, indicating an increase in the degree of information compression. More importantly, the information compression in the last few layers exhibits a sudden increase, consistent with the rising of calibrated ECE in Figure 2. Additionally, the degree of information compression in the last few layers is greatly influenced by the regularization strength, while the bottom

[1] Shwartz-Ziv & Tishby (2017) use binning-based approximation to estimate the mutual information, however, it was found that binning-based method is highly sensitive to the choice of activation functions used in the hidden layers, and some of the claims in (Shwartz-Ziv & Tishby, 2017) do not hold true in the general case (Saxe et al., 2019; Lorenzen et al., 2021).

hidden layers show relatively little variation. Taking ResNet-18 as an example, if we focus on the features from the 11th and 13th layer, all models achieve similar calibrated ECE and reconstruction loss. However, the models trained with larger weight decay yield more string raising for both calibration and reconstruction loss in the 15th and 17th layer. This demonstrate an unexpected negative impact of over-compression of neural networks for model calibratability. As prior works on information compression of deep neural networks highlighted its benefits for improving predictive performance, this results further underscore concerns about the trade-off between calibratability and discriminability.



(a) Calibrated ECE (b) Accuracy

Figure 4. The heatmaps of calibrated ECE and accuracy of ResNet-18 on CIFAR-100 with a series of top-layer early stopping strategies. **Warmer colors** indicate worse calibratability, while **cooler colors** signify better predictive performance.

4. Weak Classifiers Achieve Good Calibrataility

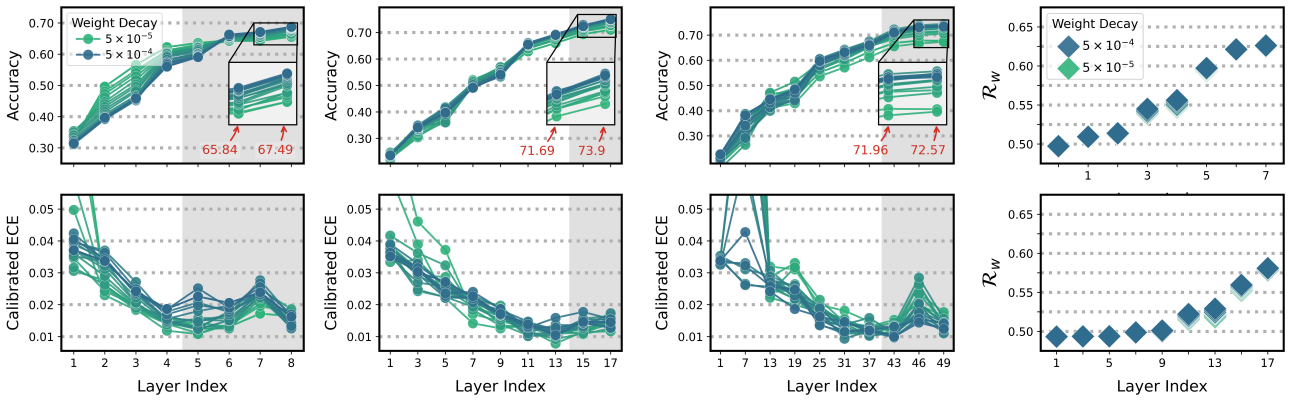
4.1. An Observation with Frozen Top Layers

After identifying that the top layers over-compress the information, thereby reducing model calibrability, we now explore how to prevent these layers from excessive information compression during training. As the compression and calibration degradation often occurs in the later stage of training, we firstly adopt a simple top-layer early stopping strategy to see if we can mitigate the harm of calibrability by avoiding the overtraining of top layers. Given a neural network, we select some specific layers as the cut points, and freeze the parameters of the layers upon these cut points after the model has been trained for a certain number of epochs. We conduct experiments on CIFAR-100 with ResNet-18 and VGG-11. The layer indices of the cut points and freezing epochs is presented in Figure 4 and 10. Each cell in the heatmaps displays the corresponding calibrated ECE and accuracy at the end of training with a specific top-layer early stopping policy.

As we can observed, the cut points of the hidden layers and the freezing epoch have significant impacts on both the calibrated ECE and accuracy. In general, freezing the parameters of the top layers in later training stage improves accuracy but reduces calibrability. On the other hand, an earlier freezing epoch enhances model calibrability but decreases accuracy. It is shown that some specific top-layer early stopping policies can achieve quite good trade-off between accuracy and calibrated ECE that even surpasses the performance of full training. As is highlighted with a red and blue boxes, freezing the parameters of the 16th and above layers of ResNet-18 before the 30th epoch, and the 5th and above layers of VGG-11 at the 50th epoch can yield impressive results. If we consider the top layers as the classification module and the bottom layers as the representation module, this empirically justify a *weak classifier hypothesis*: Given a weak classifier that has not been overtrained, the representation module can be learned better in terms of producing more calibratable features. While the performance of this top-layer early stopping strategy is very sensitive to the choices of layer index and freezing epochs, thus far, we can design more advanced methods based on this observation.

4.2. Progressively Layer-Peeled Training

We propose a straightforward method called Progressively Layer-Peeled (PLP) training to better exploit the weak classifier hypothesis, which is expected to improve calibration without sacrificing much accuracy. Our experiments suggest that finding an optimal results by freezing a specific portion of parameters at a given frozen epoch is challenging, given the myriad of possible combinations of layer cut points and freezing epochs. To address this, PLP gradually frozen the hidden layers from top to down during the whole training procedure. Specifically, for a neural network with L layers, we first divide it into K ($K<=L$) parts, for which the cut points can lies between every adjacent layers or two adjacent parts containing several layers, such as two adjacent blocks in ResNets. Then, we partition the training duration into K phases and gradually freeze the parameters of the exposed top layers as the training progresses. The number of training phases is the same with the number of the layer groups, thereby all groups can be exposed as the top trainable part for a certain epochs. For the number of epochs in each training phase T_k, $k \in \{1, 2, ..., K\}$, we can evenly partition the total number of epochs into K phases, or set other partitioning strategies to adjust the timing of freezing the top layers. Here, we adopted a simple heuristic partitioning strategy: For the kth phase, we determine its training epochs as: $T_k = \left\lfloor \frac{k^\gamma - (k-1)^\gamma}{K^\gamma} \cdot T \right\rfloor$, where γ is a hyperparameter used to control the distribution of training epochs across all phases. With γ equal to 1, the total training epochs will evenly partitioned into all phases. With γ greater than 1, starting freezing epoch for all groups during training will occur earlier. We set γ to 1 in the comparative experiment and conduct empirical analysis on the impact of γ. Figure 7 shows a simple example of PLP when $K = 3$ and $\gamma = 1$.



(a) VGG-11 (b) ResNet-18 (c) ResNet-50 (d) VGG-11 / ResNet-18

Figure 5. (a-c) The accuracy (top) and calibrated ECE (bottom) of linear probing over hidden layers of PLP-trained models on CIFAR-100. (d) Decoder-based reconstruction loss of models trained with different weight decay coefficients (top: ResNet-18; bottom: VGG-11).

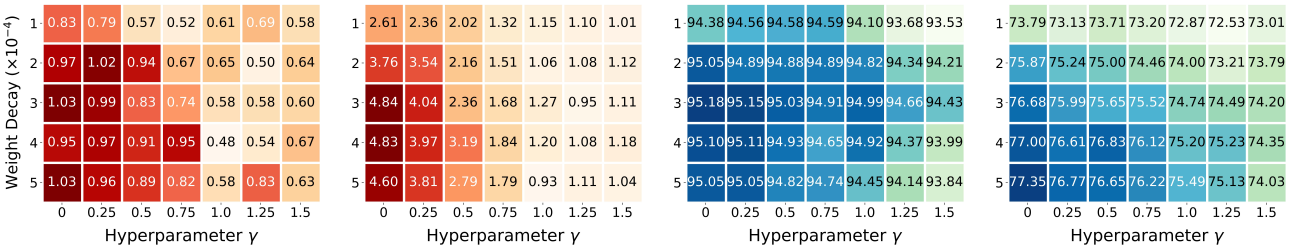
There are several existing works employing the top-down training strategy. Fang et al. (2021) introduced the layer-peeled model, from which we actually drew inspiration for the name of our method, as a nonconvex yet analytically tractable optimization framework to better understand deep neural networks. Based on this model, Yang et al. (2022) proposed to learn a neural network from class-imbalanced datasets, with the classifier (i.e., the last fully connected layer of a neural network) randomly initialized as an equiangular tight frame and fixed during training. Zhang et al. (2019) proposed a progressive top-down training method to alleviate the undertraining of the bottom layers. However, there is a fundamental distinction between our method and theirs: their methods are based on the *good classifier hypothesis*, which posits that, assuming a fixed classifier is sufficiently well-trained, the bottom layers can be further enhanced to align with this strong classifier. Moreover, they employ the retraining strategy with reinitialized bottom layers.

Calibratability of intermediate features. Following the linear probing experiment in Figure 2, we present in Figure 5 the calibrated ECE and accuracy over hidden layers of models trained with PLP strategy under different weight decay coefficients. As is shown, the top layer features of PLP-trained models exhibit better calibratability without the U-shaped trend observed in Figure 2. Interestingly, the results of ResNet-50 and VGG-11 show a double descent phenomenon. Specifically, after a slight increase in the calibrated ECE at a few top layers, the last layer reaches a notably low level of calibrated ECE. Another interesting observation is that, unlike the limited improvement of accuracy of the top layers in Figure 2, the top layers of the PLP-trained model, despite having fewer training epochs, consistently contribute to the enhancement of the model's accuracy. It shown that for ResNet-18, in Figure 2, the accuracy improvement from the 15th to the 17th layer is only 0.14%, while in Figure 5(b) the improvement from the 15th to the 17th layer is 2.36%. We also assess the information compression of the intermediate features from the model trained with PLP in Figure 5(d). It's important to note that the y-axis range here is the same as in Figure 3, but we can see that the the results depicted in these figures are quite different. As the layers deepen, the intermediate features of PLP-trained models also exhibit a trend of gradual compression, but there is no significant difference among models trained with different weight decay coefficients. It suggest that PLP can avoid the excessive information compression as observed in Figure 3.

Results with varying γ. Now, we show the improvement in calibratability brought by the PLP under different coefficients. PLP includes a hyperparameter γ, which determines the starting epoch at which the top layers begin to be frozen. When γ is smaller, top layers will start freezing later, and when it approaches 0, PLP becomes equivalent to standard training. As γ increases, the number of epochs for training the top layers reduces. Here, we selected seven parameter values within the range of 0.75 to 1.25 to see the impact of this parameter. Taking our experimental setup of 350 epochs as an example, when γ is set to 0.75 and 1.25, the training epochs counts for the outermost layer are 17 and 57, respectively. When gamma is set to 0, as illustrated in the leftmost column of the figure, PLP degenerates into standard training. Overall, as shown in Figure 6, our method demonstrates consistent performance within a reasonable parameter range on calibrated ECE across all weight decay coefficients. Specifically, PLP exhibits greater improvement in calibratability under strong regularized training. Moreover, on CIFAR-100, where standard training obtains poor calibrated ECE, a larger γ (i.e., earlier top layer freezing) yields better results.

Table 1. The comparative results of several metrics on four datasets. The **boldface** and underline denote the best and the second best results of each row. Due to limited space, the terms ECE, AECE and NLL here refer to the calibrated results.

	Method	SVHN				CIFAR-10				CIFAR-100				Tiny-ImageNet				Avg.
		ECE	AECE	NLL	ACC	ECE	AECE	NLL	ACC	ECE	AECE	NLL	ACC	ECE	AECE	NLL	ACC	Rank
ResNet-18	Standard	**0.44**	**0.43**	.166	95.5	0.83	0.81	.181	94.4	2.61	2.51	0.99	73.8	1.43	1.28	1.99	54.1	4.69
	Weight Decay	0.77	0.79	.182	95.0	1.03	1.42	.169	95.1	4.60	4.49	0.92	**77.3**	1.66	1.81	2.06	53.1	5.88
	Brier	0.55	0.49	.166	95.6	1.03	1.28	.189	94.4	3.31	3.23	1.00	74.4	2.60	2.66	2.24	49.2	6.31
	MMCE	0.93	1.00	.189	95.1	1.06	0.93	.197	94.2	2.28	2.26	1.03	72.5	1.19	1.22	1.98	54.1	6.19
	Distillation	1.06	1.80	.186	95.4	1.54	3.03	.212	94.4	6.99	6.93	1.07	75.6	1.14	**0.89**	1.86	56.0	6.81
	Label Smoothing	0.96	1.71	.173	**95.7**	1.46	2.99	.216	94.7	4.41	4.41	1.05	76.2	1.88	1.74	2.11	54.7	6.75
	Focal Loss	0.59	0.65	.164	95.7	1.13	1.42	.192	93.9	1.19	1.08	0.92	74.2	2.50	2.49	1.97	53.2	5.44
	Mixup	1.95	1.96	.232	94.5	1.13	1.43	.168	**95.9**	1.18	1.30	0.92	76.9	1.59	1.69	2.03	54.4	5.75
	MIT	0.83	1.03	.306	91.7	0.68	0.54	**.156**	94.9	2.03	1.73	0.89	75.6	1.36	1.59	2.22	49.1	5.25
	PLP	0.55	0.45	**.161**	95.6	**0.58**	**0.49**	.165	94.5	**0.93**	**0.91**	**0.86**	75.5	**0.96**	0.98	**1.82**	**56.1**	1.94
ResNet-50	Standard	0.64	0.79	.164	95.7	0.76	0.73	.180	94.4	2.73	2.70	0.96	74.6	1.66	1.70	1.70	60.2	4.44
	Weight Decay	1.01	1.03	.249	93.6	1.03	1.21	.167	94.9	4.04	4.00	0.87	**78.1**	2.68	2.69	1.76	59.6	5.94
	Brier	1.32	1.47	.179	95.8	0.70	0.75	.182	94.5	3.03	2.99	1.06	72.4	3.37	3.44	1.68	61.9	5.88
	MMCE	1.17	1.37	.193	95.6	1.56	1.53	.198	94.2	2.10	1.94	0.97	73.5	1.30	1.12	1.68	60.3	5.69
	Distillation	1.58	2.82	.200	95.7	1.51	3.04	.221	94.3	5.82	6.29	1.18	72.9	1.43	1.26	1.62	61.3	7.12
	Label Smoothing	1.71	2.96	.201	95.8	1.54	3.62	.250	94.5	5.08	5.27	1.15	75.1	3.88	4.00	1.79	**62.2**	7.94
	Focal Loss	1.22	1.46	.170	95.7	1.12	1.20	.203	93.5	1.13	1.25	0.92	73.9	2.62	2.76	1.67	59.5	5.81
	Mixup	2.11	2.12	.196	95.5	1.48	1.56	.165	**96.1**	2.77	3.06	0.93	77.5	2.50	2.50	1.70	62.1	5.75
	MIT	0.93	1.20	.262	93.2	0.66	**0.47**	**.152**	95.2	2.21	2.17	**0.85**	76.7	2.37	2.35	1.70	60.1	4.44
	PLP	**0.63**	**0.45**	**.159**	**95.8**	**0.48**	0.48	.163	94.7	**1.09**	**1.06**	0.89	74.2	**1.10**	**0.92**	**1.57**	61.3	2.00



(a) CIFAR-10, Calibrated ECE (b) CIFAR-100, Calibrated ECE (c) CIFAR-10, Accuracy (d) CIFAR-100, Accuracy

Figure 6. The calibrated ECE and accuracy of PLP-trained ResNet-18 under different regularization coefficients and varying γ. The results without using PLP are indicated within the blue box.

Comparison with other methods. We conduct comparative experiments with other calibration methods on four image classification datasets. All the reported results are based on the average of 5 random trials. The implementation details including dataset splitting, training policies and the introduction of comparison methods can be found in Appendix A.1. Table 1 shows the accuracy as well as the calibrated ECE, calibrated Adaptive Expected Calibration Error (AECE) and calibrated NLL. It can be observed that PLP achieves the best average results on these metrics and is highly robust to different datasets and architectures. Where standard training already achieves satisfactory calibration, our method largely maintains the performance, while in cases where standard training falls short in calibration, our approach offers notable enhancements. Conversely, the performance of the regularization methods is extremely unstable and even worse than standard training in general. Among these comparison methods, label smoothing and distillation, which are based on label softening mechanism, exhibit the poorest calibratability, despite their usual gains in accuracy. While methods like mixup and focal loss perform well on certain datasets, their overall performance does not match ours. Remarkably, the trainable calibration loss MMCE, which is specifically designed to enhance calibration in training, significantly harms model calibratability.

5. Conclusions

In this work, we refocused our attention from train-time calibration performance to explore model calibratability. To investigate the reasons for the decline in neural networks's calibratability, we delve into the calibratability of the intermediate features of the hidden layers of neural networks. Our findings indicate that overtraining in the network's top layers is the primary challenge, implying that a more conservative discriminability at these layers could improve calibratability. This insight correlates with the information compression inherent in deep neural networks and highlights the negative impact of over compression on uncertainty calibration. To address this, we introduced a progressive layer-peeled training strategy aimed at reducing the adverse effects of overtraining on model calibratability.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgments

The authors wish to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Science Foundation of China (623B2023, 62225602), the Fundamental Research Funds for the Central Universities (2242024K30035), and the Big Data Computing Center of Southeast University.

References

Ashukha, A., Lyzhov, A., Molchanov, D., and Vetrov, D. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *Proceedings of the International Conference on Learning Representations*, 2019.

Bouniot, Q., Mozharovskyi, P., and d'Alché Buc, F. Tailoring mixup to data using kernel warping functions. *arXiv:2311.01434*, 2023.

DeGroot, M. H. and Fienberg, S. E. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Fang, C., He, H., Long, Q., and Su, W. J. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning*, pp. 1321–1330, 2017.

Guo, H., Pasunuru, R., and Bansal, M. An overview of uncertainty calibration for text classification and the role of distillation. In *Proceedings of the 6th Workshop on Representation Learning for NLP*, pp. 289–306, 2021.

Gupta, K., Rahimi, A., Ajanthan, T., Mensink, T., Sminchisescu, C., and Hartley, R. Calibration of neural networks using splines. In *Proceedings of the International Conference on Representation Learning*, 2021.

Joo, T. and Chung, U. Revisiting explicit regularization in neural networks for well-calibrated predictive uncertainty. *arXiv:2006.06399*, 2021.

Krizhevsky, A. Learning multiple layers of features from tiny images. *Tech Report*, 2009.

Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., and Flach, P. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*, pp. 12295–12305, 2019.

Kumar, A., Sarawagi, S., and Jain, U. Trainable calibration measures for neural networks from kernel mean embeddings. In *Proceedings of the International Conference on Machine Learning*, pp. 2805–2814, 2018.

Lorenzen, S. S., Igel, C., and Nielsen, M. Information bottleneck: Exact analysis of (quantized) neural networks. In *Proceedings of the International Conference on Learning Representations*, 2021.

Ma, Y., Derksen, H., Hong, W., and Wright, J. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1546–1562, 2007.

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, pp. 15682–15694, 2021.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., and Dokania, P. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems*, pp. 15288–15299, 2020.

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pp. 4696–4705, 2019.

Naeini, M. P., Cooper, G., and Hauskrecht, M. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 2015.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *Advances in Neural Information Processing Systems Workshops*, 2011.

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

Patel, K., Beluch, W., Yang, B., Pfeiffer, M., and Zhang, D. Multi-class uncertainty calibration via mutual information maximization-based binning. In *Proceedings of the International Conference on Representation Learning*, 2021.

Patra, R., Hebbalaguppe, R., Dash, T., Shroff, G., and Vig, L. Calibrating deep neural networks using explicit regularisation and dynamic data pruning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1541–1549, 2023.

Platt, J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3): 61–74, 1999.

Rahimi, A., Shaban, A., Cheng, C.-A., Hartley, R., and Boots, B. Intra order-preserving functions for calibration of multi-class neural networks. In *Advances in Neural Information Processing Systems*, pp. 13456–13467, 2020.

Saxe, A. M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B. D., and Cox, D. D. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, 2019.

Shwartz-Ziv, R. and Tishby, N. Opening the black box of deep neural networks via information. *arXiv:1703.00810*, 2017.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Tao, L., Dong, M., Liu, D., Sun, C., and Xu, C. Calibrating a deep neural network with its predecessors. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 4271–4279, 2023a.

Tao, L., Dong, M., and Xu, C. Dual focal loss for calibration. In *Proceedings of the International Conference on Machine Learning*, pp. 33833–33849, 2023b.

Tao, L., Zhu, Y., Guo, H., Dong, M., and Xu, C. A benchmark study on calibration. In *Proceedings of the International Conference on Learning Representations*, 2024.

Thulasidasan, S., Chennupati, G., Bilmes, J. A., Bhattacharya, T., and Michalak, S. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 13888–13899, 2019.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. *arXiv:physics/0004057*, 2000.

Tomani, C., Gruber, S., Erdem, M. E., Cremers, D., and Buettner, F. Post-hoc uncertainty calibration for domain drift scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10124–10132, 2021.

Wang, D.-B., Feng, L., and Zhang, M.-L. Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. In *Advances in Neural Information Processing Systems*, pp. 11809–11820, 2021.

Wang, D.-B., Li, L., Zhao, P., Heng, P.-A., and Zhang, M.-L. On the pitfall of mixup for uncertainty calibration. In *Proceedings of the International Conference on Learning Representations*, 2023.

Wang, S., Wang, J., Wang, G., Zhang, B., Zhou, K., and Wei, H. Open-vocabulary calibration for vision-language models. *arXiv:2402.04655*, 2024.

Wang, Y., Ni, Z., Song, S., Yang, L., and Huang, G. Revisiting locally supervised learning: an alternative to end-to-end training. In *Proceedings of the International Conference on Learning Representations*, 2020.

Wei, H., Xie, R., Cheng, H., Feng, L., An, B., and Li, Y. Mitigating neural network overconfidence with logit normalization. In *Proceedings of the International Conference on Machine Learning*, pp. 23631–23644, 2022.

Yang, Y., Chen, S., Li, X., Xie, L., Lin, Z., and Tao, D. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? In *Advances in Neural Information Processing Systems*, pp. 37991–38002, 2022.

Yu, Y., Chan, K. H. R., You, C., Song, C., and Ma, Y. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. In *Advances in Neural Information Processing Systems*, pp. 9422–9434, 2020.

Zadrozny, B. and Elkan, C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the International Conference on Machine Learning*, pp. 609–616, 2001.

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002.

Zhang, L., Deng, Z., Kawaguchi, K., and Zou, J. When and how mixup improves calibration. In *Proceedings of the International Conference on Machine Learning*, pp. 26135–26160, 2022.

Zhang, S., Do, C.-T., Doddipatla, R., Loweimi, E., Bell, P., and Renals, S. Top-down training for neural networks. *Tech Report*, 2019.

Zhu, D., Lei, B., Zhang, J., Fang, Y., Xie, Y., Zhang, R., and Xu, D. Rethinking data distillation: Do not overlook calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4935–4945, 2023.

A. Appendix

A.1. Experimental Details

Datasets. The experiments in main text are based on four widely used image classification datasets: SVHN (Netzer et al., 2011), CIFAR-10, CIFAR-100 (Krizhevsky, 2009) and Tiny-ImageNet (Deng et al., 2009). SVHN is an image dataset which consists of 32×32 colored images of 0∼9 digits. CIFAR-10 and CIFAR-100 consist of 32×32 colored natural images arranged in 10 and 100 classes, respectively. For Tiny-ImageNet, the experiments are based on images resized as 64x64. We split the original training dataset into training set and validation set for main training and post-hoc calibration with the following ratios: 68257/5k for SVHN, 45k/5k for CIFAR-10/100 and 90k/10k for Tiny-ImageNet.

Training details. We conduct experimental analysis based on three models ResNet-18, ResNet-50 and VGG-11 in the main paper. To learn these models, we use SGD as the opimizer with a momentum of 0.9 and a weight decay of 10^{-4} unless otherwise specified. We train on SVHN/CIFAR-10/CIFAR-100 by total 350 epochs with the initial learning rate as 0.1, and divide it by a factor of 10 after 150 epochs and 250 epochs respectively. For Tiny-ImageNet, we conduct training based on the open sourced pretrained models. Based on the pretrained models, we train 200 epochs with the initial learning rate as 0.01, and divide it by a factor of 2 after every 30 epochs. We set the batch size as 128 on SVHN/CIFAR-10/CIFAR-100, and 64 on Tiny-ImageNet. In Table 2, we set the weight decay of PLP as 0.0005 for SVHN/CIFAR-10/CIFAR-100, and 0.0001 for Tiny-ImageNet. `Code is available at https://github.com/dengbaowang/PLP.`

Post-hoc Calibration. For post-hoc calibration, we adopt TS as the post-hoc calibration method in the experiments of the main paper. We also conduct experiment with Vector Scaling (VS) and present the results in Appendix A.2. For both TS and VS, we use the LBFGS to optimize the parameters on validation set.

Linear probing details. In Figure 2, we conduct linear probing on the features extracted from hidden layers. For hidden layer features with different dimensions, we simply flatten the feature tensors of each layer into one-dimensional vectors and train fully connected layers as classifiers based on them. Here, we train the linear classifier by using SGD with a momentum of 0.9 and a weight decay of 10^{-4}. We train the classifier for 10 epochs with the initial learning rate as 0.1, and decay it by a factor of 10 after 6 and 8 epochs respectively. To access the calibratability of hidden features, we also apply temperature scaling to the linear classifier after linear probing.

The choices of cut points. As we discussed in Section 3.2, the PLP method needs to first divide all the layers into multiple parts, for which the cut points can lies between the adjacent groups containing several adjacent layers. Here, we divide ResNet-18 and ResNet-50 into 11 parts with the cut point sets $\{2, 4, 6, 8, 10, 12, 14, 16, 17, 18\}$ and $\{11, 17, 23, 26, 29, 32, 38, 41, 44, 47\}$, respectively, and divide VGG-11 into 6 parts with the cut point set $\{2, 3, 4, 6, 9\}$.

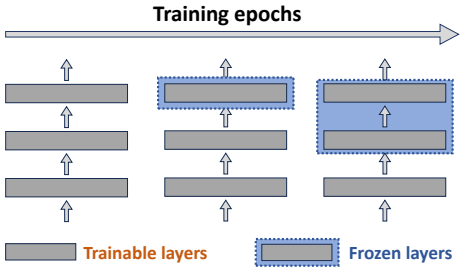
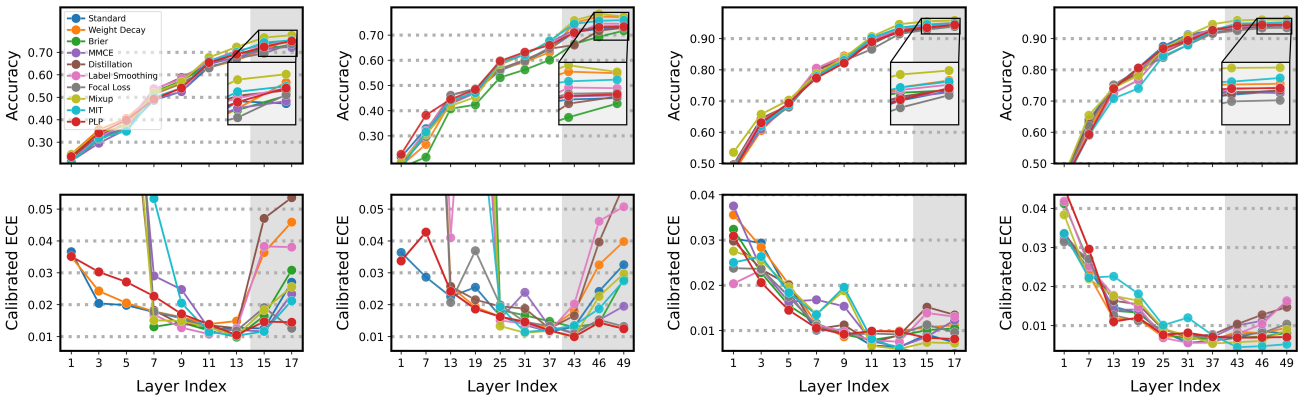


Figure 7. An illustration of PLP training.

Rate distortion-based mutual information estimation. Rate distortion measures the *compactness* of a random distribution. Given a random variable $\boldsymbol{z} \in \mathbb{R}_d$ and a specified value $\epsilon > 0$, the rate distortion $R(\boldsymbol{z}, \epsilon)$ denotes the minimal number of binary bits required to encode $\boldsymbol{z}$ such that the expected decoding error is less than ϵ, i.e., the reconstructed $\hat{\boldsymbol{z}}$ satisfies $\mathbb{E}\left[\|\boldsymbol{z} - \hat{\boldsymbol{z}}\|_2\right] \leq \epsilon$. However, we lack knowledge about the distribution of intermediate features and can only access the specific feature tensors corresponding to each sample. Ma et al. (2007) proposed an analytical expression for calculating the number of binary bits needed to encoded finite samples: $\mathcal{L}(\boldsymbol{Z}, \epsilon) \doteq \left(\frac{m+d}{2}\right) \log\det\left(\boldsymbol{I} + \frac{d}{m\epsilon^2}\boldsymbol{Z}\boldsymbol{Z}^\top\right)$, where $\boldsymbol{Z}$ denotes the feature matrix of a batch of samples $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_m]$, m denotes the batch size. We employe the extended version of this analytical expression for multi-class data to estimate mutual information in our experiments. For more implementation details, we recommend referring to (Yu et al., 2020).

Decoder-based mutual information estimation. In Figure 3, we estimate the mutual information $\boldsymbol{I}(\boldsymbol{x}, \boldsymbol{h})$ by training a decoder parameterized by $\boldsymbol{w}$ to obtain the minimal reconstruction loss on training data (Wang et al., 2020). We directly use the binary cross-entropy loss in original feature space to estimate the mutual information. Specifically, we train a light-weight decoder with two convolutional layers using Adam optimizer for 30 epochs with a constant learning rate 0.01. We are primarily focus on the comparisons of information across intermediate layers and between different models rather than obtaining the exact values of $\boldsymbol{I}(\boldsymbol{x}, \boldsymbol{h})$. Therefore, the utilization of the same training policy among all layers or models makes the comparisons fair.

Evaluation metrics. A perfectly calibrated classifier is expected to satisfy $\mathbb{P}(\hat{y} = y \,|\, \hat{p} = p) = p$ for $p \in [0, 1]$. Given this, calibration performance could be measured



(a) CIFAR-100, ResNet-18 (b) CIFAR-100, ResNet-50 (c) CIFAR-10, ResNet-18 (d) CIFAR-10, ResNet-50

Figure 8. The accuracy (top) and calibrated ECE (bottom) of linear probing over hidden layers of models trained with different methods.

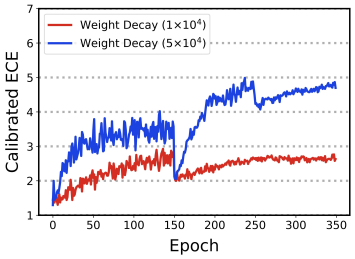
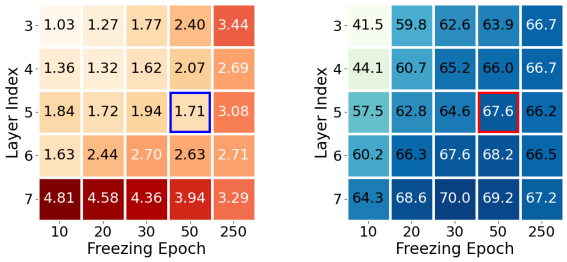


Figure 9. The calibrated ECE over training epochs of ResNet-18 on CIFAR-100.



(a) Calibrated ECE (b) Accuracy

Figure 10. The heatmaps of calibrated ECE and accuracy of VGG-11 on CIFAR-100 with a series of top-layer early stopping strategies. **Warmer colors** indicate worse calibratability, while **cooler colors** signify better predictive performance.

by the difference between accuracy and confidence in expectation, i.e., $\mathbb{E}_{\hat{p}}\left[\mathbb{P}(\hat{y}=y \mid \hat{p}=p)-p\right]$. In practice, this can be approximated by first grouping all the samples into M equally spaced bins $\{B_m\}_{m=1}^{M}$ with respect to their confidence scores, and taking a weighted average of the accuracy/confidence difference between these bins. Formally, ECE is defined as $\mathrm{ECE}=\sum_{m=1}^{M} \frac{|B_m|}{N}\left|\mathrm{acc}\left(B_m\right)-\mathrm{avg}\,\mathrm{Conf}\left(B_m\right)\right|$, where N denotes the total number of samples in testing set. Different from ECE that bins with equal confidence interval, AECE adaptively groups the samples into intervals with same sample size. In this way, each bin B_r in $\{B_r\}_{r=1}^{R}$ has $\frac{N}{M}$ samples and the metric can be formally defined as (Nixon et al., 2019): $\mathrm{AECE}=\frac{1}{R}\sum_{r=1}^{R}\left|\mathrm{acc}\left(B_r\right)-\mathrm{avg}\,\mathrm{Conf}\left(B_r\right)\right|$. In our experiments, we set the bin number as 15 for both ECE and AECE.

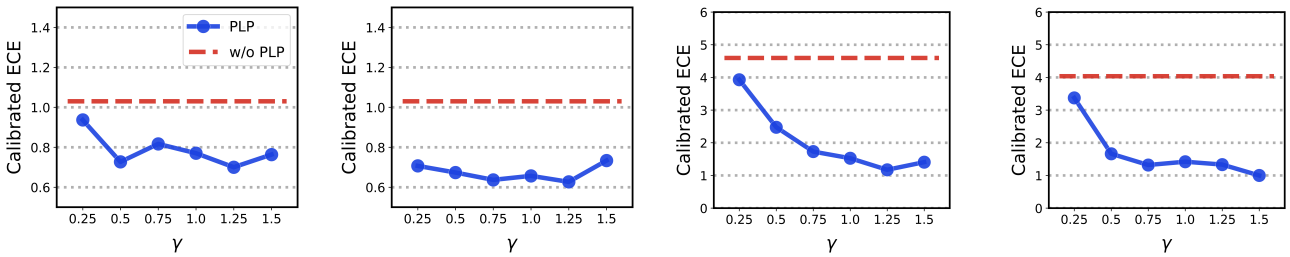
A.2. Complementary Results

Feature calibratability. In the main text, we present the empirical study on the feature calibratability of models tained with different weight decay coefficients. Similarly, we now conduct the experiments for other comparative methods in Figure 8. On CIFAR-100, the pronounced U-shaped trend in feature calibratability is also shown with these methods. Most comparative methods induce the increase in calibratability in the last few layers, while our method largely mitigates this U-shaped trend. It should be noted that due to the differences in dataset difficulty, the calibration performance on CIFAR-10 is much better than that on CIFAR-100. However, as we can see, the detrimental impact of the top layers on calibratability observed in the main text still remains. Furthermore, the number of layers causing a decline in calibratability is consistent with the that on CIFAR-100, but the degree of degradation is somewhat less pronounced on CIFAR-10.

Results with block-wise PLP. We have conducted additional experiments with K set to 4 for ResNet-18 and ResNet-50 on CIFAR-10 and CIFAR-100. By treating each block as a layer group for parameter freezing, we have observed in Figure 11 that PLP with $K = 4$ also consistently yields better calibratability than standard training. These results are in line with our original results, showing that increasing the γ can enhance the calibrated ECE achieved by PLP. The results suggest that the hyperparameter γ, which controls the distribution of training epochs of different layers, plays a more critical role in determining the overall performance than K.

Table 2. The comparative results of several metrics on four datasets with **Vector Scaling**. The **boldface** and underline denote the best and the second best results of each row. Due to limited space, the terms ECE, AECE and NLL here refer to the calibrated results.

	Model & Method	SVHN				CIFAR-10				CIFAR-100				Tiny-ImageNet				Avg.
		ECE	AECE	NLL	ACC	ECE	AECE	NLL	ACC	ECE	AECE	NLL	ACC	ECE	AECE	NLL	ACC	Rank
ResNet-18	Standard	**0.38**	**0.44**	.168	95.5	0.89	0.84	.178	94.4	2.96	2.74	1.00	73.8	1.67	1.59	1.99	54.1	4.56
	Weight Decay	0.68	0.73	.183	95.0	1.01	1.25	.166	95.1	4.86	4.78	0.93	**77.3**	2.26	2.17	2.05	53.1	6.19
	Brier	0.51	0.53	.166	95.6	1.06	1.18	.185	94.4	3.46	3.38	1.01	74.4	3.41	3.40	2.21	49.2	6.50
	MMCE	0.88	0.96	.189	95.1	1.01	0.82	.187	94.2	2.69	2.64	1.04	72.5	**1.21**	**1.25**	1.97	54.1	5.81
	Distillation	1.02	1.54	.186	95.4	1.52	2.10	.207	94.4	6.86	6.70	1.07	75.6	1.77	1.79	1.86	56.0	7.06
	Label Smoothing	0.92	1.46	.173	**95.7**	1.52	2.31	.210	94.7	4.58	4.57	1.05	76.2	2.11	2.06	2.11	54.7	6.75
	Focal Loss	0.62	0.63	.164	95.7	1.13	1.33	.189	93.9	1.68	1.56	0.93	74.2	2.03	2.08	1.96	53.2	5.50
	Mixup	1.76	1.77	.228	94.5	0.94	1.19	.164	**95.9**	1.54	1.55	0.91	76.9	1.94	1.80	2.00	54.4	5.00
	MIT	0.77	0.82	.303	91.7	**0.56**	**0.55**	**.152**	94.9	2.10	2.04	0.88	75.6	2.06	2.09	2.17	49.1	5.44
	PLP	0.50	0.45	**.163**	95.6	0.60	0.57	.166	94.5	**1.50**	**1.41**	**0.88**	75.5	1.63	1.55	**1.81**	**56.1**	2.19
ResNet-50	Standard	**0.65**	0.74	.164	95.7	0.82	0.65	.178	94.4	3.12	3.05	0.96	74.6	2.08	2.03	1.70	60.2	4.69
	Weight Decay	0.82	0.81	.244	93.6	0.95	1.04	.163	94.9	4.37	4.36	0.87	**78.1**	3.38	3.33	1.75	59.6	5.88
	Brier	1.31	1.45	.179	95.8	0.68	0.70	.178	94.5	3.45	3.34	1.07	72.4	4.03	4.01	1.67	61.9	5.88
	MMCE	1.02	1.20	.193	95.6	1.26	1.16	.190	94.2	2.42	2.42	0.98	73.5	1.49	1.51	1.70	60.3	5.81
	Distillation	1.63	2.27	.200	95.7	1.63	2.27	.216	94.3	6.18	6.18	1.18	72.9	1.97	1.90	1.60	61.3	7.25
	Label Smoothing	1.67	2.37	.202	95.8	1.21	2.36	.241	94.5	5.08	5.01	1.15	75.1	4.21	4.20	1.77	**62.2**	7.81
	Focal Loss	1.21	1.33	.170	95.7	1.19	1.30	.200	93.5	1.65	1.57	0.93	73.9	2.30	2.30	1.66	59.5	5.69
	Mixup	1.71	1.72	.191	95.5	1.31	1.36	.160	**96.1**	2.64	2.73	0.91	77.5	2.44	2.49	1.69	62.1	5.44
	MIT	0.77	1.03	.262	93.2	**0.54**	**0.57**	**.148**	95.2	2.24	2.09	**0.84**	76.7	2.95	2.92	1.68	60.1	4.31
	PLP	0.68	**0.44**	**.158**	**95.8**	0.59	0.59	.163	94.7	**1.32**	**1.25**	0.91	74.2	**1.48**	**1.50**	**1.57**	61.3	2.25



(a) SVHN (b) CIFAR-10 (c) CIFAR-100 (d) Tiny-ImageNet

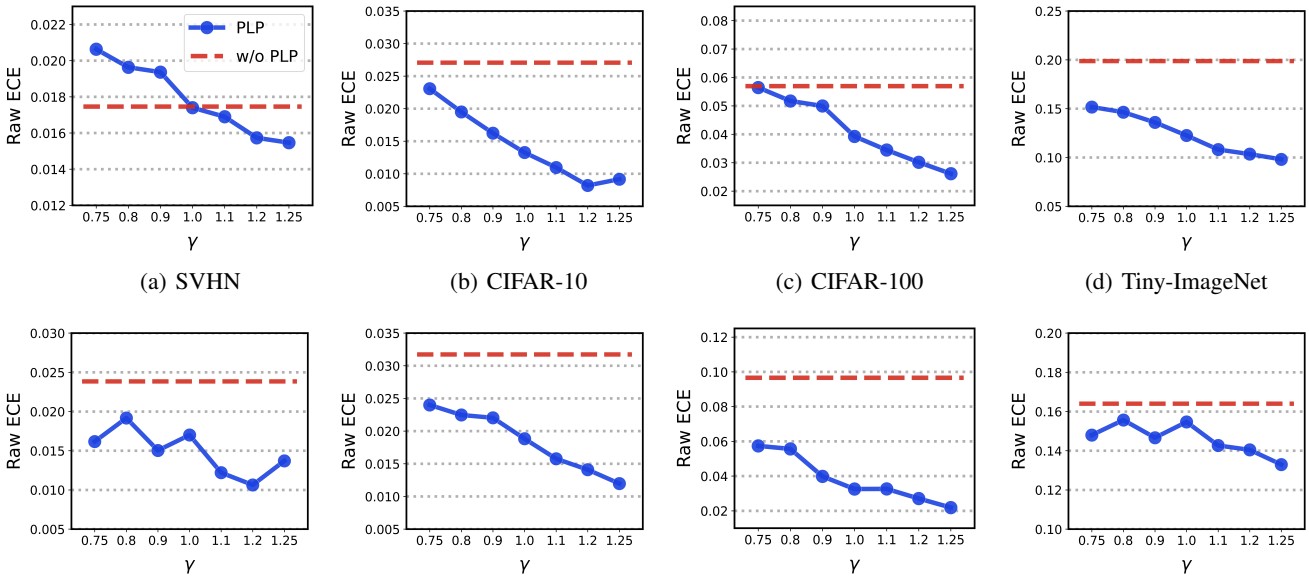
Figure 11. Calibrated ECE of PLP-trained ResNet-18 and ResNet-50 with $K = 4$ (block-wise PLP training) and varying γ on four datasets. The blue solid lines represents the results of models trained using the PLP strategy, while the red dashed line represents the results without employing the PLP strategy.

Results with vector scaling. In all the experiments above, we employed TS as the post-hoc calibration method. This choice is due to the fact that TS is considered the simplest and most effective method while preserving the dence model outputs. We conducted similar comparative experiments by replacing it with vector scaling. The results shown in Table 2 indicate that when using vector scaling as the post-hoc calibration method, the comparative results of between models trained with different methods are generally consistent to that in Table 1. Our method achieved the best results in general.

Calibration performance of raw outputs. Although this work focuses on model calibratability rather than raw calibration performance, we unintentionally discovered that the PLP is also very helpful in improving the calibration performance for the model's raw outputs. The experimental results in Figure 12 demonstrate that increasing γ (i.e., earlier top layer freezing) can also improve raw calibration, and this effect is consistent across the four datasets.

A.3. Comparison Methods

We compared several methods in the experiments of the main text: (stronger) weight decay, label smoothing, self-distillation, mixup, focal loss, Maximum Mean Calibration Error (MMCE), Brier and mixup in training (MIT). Among these methods, the former five ones are not original designed for calibration but has been found beneficial to it in recent studies. MMCE is a specifically designed loss function term for calibration. MIT is proposed to mitigate the harm of the original mixup for calibration. Brier loss is the simple mean squared error between the predicted confidences and the ground-truth one-hot labels, which was considered as an important baseline as it can be decomposed into calibration and refinement (DeGroot & Fienberg, 1983). For all these methods, we adopt the same optimizer, learning rate and the number of epochs in training as discussed at the beginning of Appendix. Most of these methods involve hyperparameters that need to be determined before training. Here is a brief introduction to these methods:



(a) SVHN (b) CIFAR-10 (c) CIFAR-100 (d) Tiny-ImageNet

(e) SVHN (f) CIFAR-10 (g) CIFAR-100 (h) Tiny-ImageNet

Figure 12. Raw ECE of PLP-trained ResNet-18 with varying γ on four datasets. The blue solid lines represents the results of models trained using the PLP strategy, while the red dashed line represents the results without employing the PLP strategy.

- Label smoothing is widely used to reduce overfitting of DNNs (Szegedy et al., 2016). The mechanism of LS is simple: when training with CE loss, the one-hot label vector $\boldsymbol{y}$ is replaced with *soft* label vector $\widetilde{\boldsymbol{y}}$, whose elements can be formally denoted as $\widetilde{y}_i = (1-\epsilon)y_i + \epsilon/K, \forall i \in \{1, ..., K\}$, where $\epsilon > 0$ is a strength coefficient.

- Mixup takes the convex combinations between pairs of examples and their labels: $\widetilde{x}_i = \lambda x_i + (1-\lambda)x_j, \widetilde{y}_i = \lambda y_i + (1-\lambda)y_j$, where λ is sampled from Beta(α, α) with $\alpha > 0$. A larger α will result in a higher degree of mixing strength, thus making the mixed labels smoother.

- Focal loss is originally proposed to address the class imbalance problem in object detection.It is formally defined as: $\mathcal{L}_f = -(1-f_y^\theta)^\gamma \log f_y^\theta$, where γ is a predefined coefficient. (Mukhoti et al., 2020) found that the models learned by focal loss produce output probabilities which are already very well calibrated.

- MMCE is a continuous and differentiable proxy for calibration error and is normally used as a regularizer alongside the commonly used cross-entropy loss, where a weighting factor β could be used to balance the contribution of MMCE (Kumar et al., 2018).

- Brier loss is the simple the mean squared error between the predicted confidences and the ground-truth one-hot labels, which was considered as an important baseline as it can be decomposed into calibration and refinement (DeGroot & Fienberg, 1983).

- MIT is a derived version of mixup for improving model calibration. The hyperparameter γ plys the same role with that of the original mixup.

In Figure 8, Table 1 and 2, the hyperparameters for the comparison methods are chosen based on commonly recommended values in the literature: a decay ratio of 5×10^{-4} for weight decay, ϵ as 0.1 for label smoothing, α as 1 for mixup and MIT, γ as 3 for focal loss and β as 0.5 for MMCE.

A.4. Future Work

We think the following problems are desired to be explored: (1) Conducting experiments to analyze calibratability on larger datasets and a variety of network architectures would be a valuable next step. This could involve examining the impact of the pre-training on the calibratability of downstream tasks, especially within the prevalent pre-training/fine-tuning paradigm in modern machine learning. (2) Investigating the performance of model calibratability under distribution shift is crucial for real-world applications. Also, developing post-hoc or test-time calibration methods that are robust to distribution shifts could lead to models that provide reliable predictions in diverse and changing environments. (3) Theoretical analysis of calibratability could offer insights into the fundamental principles to produce well-calibrated models.