View Reviews

Paper ID
3299

Paper Title
Embarrassingly Simple Relevance-driven Training for Improved Network Calibration

Reviewer #1

Questions

1. Summary. In 5-7 sentences, describe the key ideas, experimental or theoretical results, and their significance.
The paper introduces RelCal, a new train-time calibration method. It aims to prioritize crucial features during training based on the understanding that not all dimensions contribute significantly to a class. By using relevance scores, it identifies and retains the most important features for each class while pruning irrelevant ones. The proposed method enhances both calibration performance and classification accuracy across diverse datasets.

2. Strengths. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Explain clearly why these aspects of the paper are valuable. Short bullet lists do NOT suffice.
- The proposed method has the potential to synergize with other train-time calibration techniques, enhancing the overall calibration capabilities of the networks.
- The experimental findings validate that the proposed method not only enhances calibration performance but also boosts accuracy.
- The proposed method demonstrates improved performance across various datasets.

3. Weaknesses. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Clearly explain why these are weak aspects of the paper, e.g., why a specific prior work has already demonstrated the key contributions, or why the experiments are insufficient to validate the claims, etc. Short bullet lists do NOT suffice. Be specific!
- The paper's motivation is unclear to me. The authors suggest that the miscalibration issue stems from learned features in the networks (L53-56) and emphasize the effectiveness of using only crucial features to tackle the problem (L62-64). However, the authors seem to prioritize regularization on the network weights over the features.
- The technical contribution of feature pruning appears lacking in novelty, resembling methods employed in existing OOD detection approaches [9, 48]. Specifically, the method for determining the pruning ratio closely resembles ASH [9], while the relevance score bears similarity to the contribution matrix in [48]. Additionally, the effectiveness of feature pruning for mitigating the overconfidence problem has already been

demonstrated in previous OOD methods [1, 47]. Given the close relationship between OOD detection and calibration, these similarities suggest that the novelty of the proposed method may be marginal. Although the proposed method differs from previous OOD methods in being a train-time technique, it remains unclear why train-time sparsification is more effective than test-time methods.
- There is a lack of quantitative comparisons with recent state-of-the-art methods, such as CALS[A] and ACLS[40], which significantly outperform the proposed method in terms of calibration performance on large-scale datasets like Tiny-ImageNet and ImageNet.
- To enhance the method's applicability and generalizability, it would be helpful if the authors include results for semantic segmentation, as discussed in previous methods.

[A] Class Adaptive Network Calibration, CVPR 2023

4. Paper rating (pre-rebuttal).
Weak Reject

5. Justification of rating. What are the most important factors in your rating?
The paper's motivation lacks clarity as it suggests that the miscalibration issue arises from learned features in the networks, yet emphasizes regularization on network weights rather than features. Furthermore, the technical contribution of feature pruning appears to lack novelty, resembling methods found in existing out-of-distribution (OOD) detection approaches.

6. Are there any serious ethical/privacy/transparency/fairness concerns? If yes, please also discuss below in Question 9.
No

7. Is the contribution of a new dataset a main claim for this paper? Have the authors indicated so in the submission form?
No dataset contribution claim

Reviewer #2

Questions

1. Summary. In 5-7 sentences, describe the key ideas, experimental or theoretical results, and their significance.
The manuscript addresses the miscalibration issue in DNNs via a train-time method. Specifically, a new method, ReCal, is proposed to enforce the model to selectively retains a subset of features from the penultimate layer based on their relevance scores. The relevance scores are defined as the absolute value of the multiplication between feature embedding and weight of the last clarifier layer. Extensive experiments have demonstrate the effectiveness of the proposed method on different benchmarks.

2. Strengths. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Explain clearly why these aspects of

the paper are valuable. Short bullet lists do NOT suffice.

1. The finding that the feature selection or pruning can enhance the calibration of the model is an interesting and valuable contribution. Motivated by it, a simple yet effective solution is implemented.

2. The paper is very well presented, with clear motivation, formulation and extensive experiments.

3. Weaknesses. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Clearly explain why these are weak aspects of the paper, e.g., why a specific prior work has already demonstrated the key contributions, or why the experiments are insufficient to validate the claims, etc. Short bullet lists do NOT suffice. Be specific!

1. What is the intuitive explanation that ReCal works much better with FL loss, compared to NLL and LS? I am also wondering whether it is a good idea to combine the ReCal with LS or FL. As mentioned in the paper, ReCal is not orthogonal to LS or FL but has similar effect of implicitly regularizing the network weight W.

2. The proposed method is encouraging sparse feature representation in a heuristic way. Therefore, I think it is necessary to compare with some sparsity method, like including L1 regularization on weight W, or other related works.

3. Is the proposed method computationally feasible for larger setting, like the vocabulary size of the target categories are very large, or the semantic segmentation task where a feature embedding is calculated for every pixel.

4. Paper rating (pre-rebuttal).

Borderline

5. Justification of rating. What are the most important factors in your rating?

The paper is well presented and proposes an interesting method to investigate the relationship between feature relevance and model calibration. Major concern is the justification of combining FL with the proposed method, and lack of comparison with other feature sparsity method. Please refer to the Weakness section for details.

6. Are there any serious ethical/privacy/transparency/fairness concerns? If yes, please also discuss below in Question 9.

No

7. Is the contribution of a new dataset a main claim for this paper? Have the authors indicated so in the submission form?

No dataset contribution claim

Reviewer #3

Questions

1. Summary. In 5-7 sentences, describe the key ideas, experimental or theoretical results, and their significance.

This paper introduces a calibration technique for DDNs classifiers to improve their uncertainty estimation. Unlike previous approaches addressing the problem (by different techniques such as temperature-scaling, inference optimization, label manipulation etc.), the authors propose RelCal, a masked training strategy that applies regularization by restricting features' most redundant dimensions of the second-to-last layer. They do so by calculating 'relevance score' to each feature. Experiments evaluating calibration metrics such as confidence and ECE demonstrate that the proposed method consistently reduces uncertainty and produces better calibrated distributions.

2. Strengths. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Explain clearly why these aspects of the paper are valuable. Short bullet lists do NOT suffice.

1. The paper is well-written, the motivation is clear as well the proposed approach is simply and easily understood.
2. The overall results show that RelCal does reduce ECE, without negatively affecting accuracy. The authors perform extensive experiments, ablations and define decent baselines.

3. Weaknesses. Consider the significance of key ideas, experimental or theoretical validation, writing quality, data contribution. Clearly explain why these are weak aspects of the paper, e.g., why a specific prior work has already demonstrated the key contributions, or why the experiments are insufficient to validate the claims, etc. Short bullet lists do NOT suffice. Be specific!

1. Model calibration is a well-recognized and significant issue. This paper focuses solely on classification tasks, though calibration remains critically important for other tasks that are currently drawing more attention within the community in 2024.
2. Technically, the method presented is straightforward and bears a close resemblance to prior works. I am not convinced that it will inspire further research.
3. The method presented resemble with techniques designed to mitigate shortcut learning. It would be valuable to hear the authors' perspective on this, particularly because the proposed method could potentially enhance shortcut learning.
4. Most experiments and ablation studies were conducted on:
a. CIFAR, which has a very low resolution and does not adequately represent the more complex image data typically encountered today.
b. CNN-based models (specifically those related to ResNet), which are no longer considered state-of-the-art.
5. Including top-k metrics such as Accuracy@k and Recall@k could be beneficial.

4. Paper rating (pre-rebuttal).

Weak Reject

5. Justification of rating. What are the most important factors in your rating?

Given the concerns outlined above, I am uncertain whether this work meets the ECCV standards in terms of both significance and technical innovation.

6. Are there any serious ethical/privacy/transparency/fairness concerns? If yes, please also discuss below in Question 9.

No

7. Is the contribution of a new dataset a main claim for this paper? Have the authors indicated so in the submission form?

No dataset contribution claim