Medical diagnosis, Autonomous driving

Calibrating confidence in predictions by machine learning models

D3 Center

Associate Professor Hideaki Hayashi

Researchmap https://researchmap.jp/Hayashi_Hideaki/?lang=en

Abstract

In application fields requiring high reliability, such as medical diagnosis and autonomous driving, it is beneficial to understand the confidence level in predictions by machine learning models. In classification tasks, the posterior probability of the predicted class is commonly used as the confidence score. However, especially with recent deep learning models, there has been a problem with misalignment between this confidence score and the actual accuracy of predictions. This study proposes a method to calibrate the confidence score provided by machine learning models. The proposed approach incorporates a generative model into the final layer of a deep neural network classifier, allowing the model to compute class posterior probabilities while considering the input data distribution, thereby providing more accurate confidence estimates.

Background & Results

With the remarkable advancements in large language models and generative AI, the reliability of machine learning models has gained importance. Even when a model achieves high accuracy with plausible outputs, human decision-making often remains essential. In such cases, it would be useful if humans could refer to the model's confidence in its predictions. In classification tasks—a fundamental and important task in machine learning—the probability provided by a classifier along with a prediction serves as a valuable reference in decision-making. Classification is a task to predict which class an input belongs to, such as recognizing the object in an image. Given an input image, the model outputs the probability of each object class being present. Here, the highest class probability is called the confidence score and is regarded as an important metric of prediction reliability.

However, the confidence scores by machine learning models are often unreliable. By "unreliable," we mean that the confidence score does not align with the actual accuracy. For instance, consider a trained classifier applied to unseen data points to obtain predicted labels and their confidence scores. Ideally, the prediction accuracy for data points with a "70% confidence" would be 70%. However, particularly with recent deep learning models, confidence scores often do not match actual accuracy.

This study proposes a method that calibrates confidence during training by learning a generative model simultaneously in the final layer of a deep neural network classifier. Experiments demonstrated the effectiveness of the proposed method for confidence calibration in tasks involving general object and medical image classification. Notably, in semi-supervised classification tasks where labeled training data is limited, the proposed method significantly improved the confidence calibration metric, known as expected calibration error, compared to conventional methods.

Significance of the research and Future perspective

If the confidence score can accurately reflect the uncertainty of predictions by machine learning models, it becomes valuable in real-world decision-making. For example, in scenarios where medical data is used for disease diagnosis, the model provides class predictions (e.g., normal or abnormal) along with corresponding confidence scores. Since the final diagnostic decision is made by a physician, the model's confidence score assists the physician in deciding whether to rely on the predicted results. If the confidence score is below a certain threshold, the prediction can be disregarded, prompting a more thorough visual diagnosis instead.

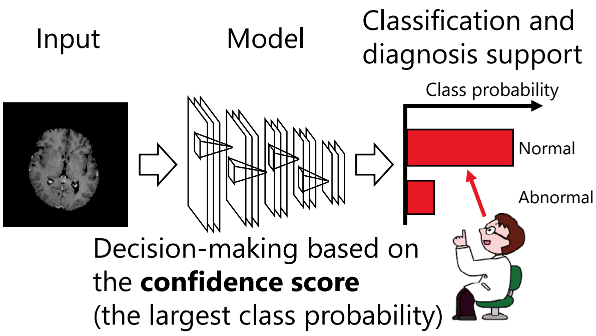

Figure 1

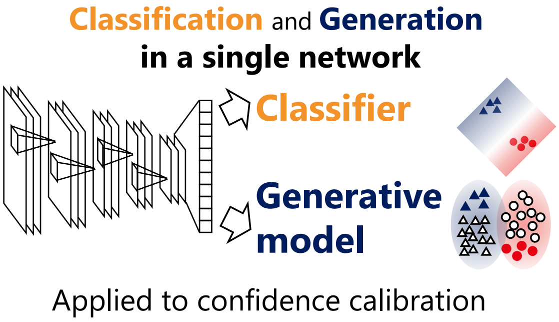

Figure 2

Patent	
Treatise	Hayashi, Hideaki. A hybrid of generative and discriminative models based on the Gaussian-coupled softmax layer. IEEE Transactions on Neural Networks and Learning Systems. 2024 (early access). doi: 10.1109/TNNLS.2024.3358113
URL	https://sites.google.com/view/hideakihayashi/home
Keyword	machine learning, pattern recognition, artificial intelligence