记录阅读、代码等进度

第一篇 A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks 错误分类基准线

ABSTRACT

The **methodology** is to use the probabilities from the SoftMax Distributions.

Rationale: The correctly classified examples tend to have greater maximum probabilities than erroneously classified OOD examples.

The method is that the paper test the performance over many tasks in various fields

The **result** is a baseline to determine what is an OOD

1 INTRODUCTION

Gaussians Noise: named after Carl Gauss, is a term from signal processing theory denoting a kind of signal noise that has a probability density function equal to that of the normal distribution (which is also known as the Gaussian distribution).

$$p_G(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$$

Contribution 1: This new method evaluates the **quality of a neural network's input reconstruction** to determine if an example is abnormal.

Contribution 2: Another contribution of this work is the designation of **standard tasks and evaluation** metrics for assessing the automatic detection of errors and out-of-distribution examples.

In summary, while softmax classifier probabilities are not directly useful as confidence estimates, estimating model confidence is not as bleak as previously believed. This paper **creates a strong baseline** for detecting errors and out-of-distribution examples which we hope future research surpasses.

2 PROBLEM FORMULATION AND EVALUATION

Problem 1: The first is **error and success prediction**. Can we predict whether a trained classifier will **make an error** on a particular held-out test example' can we predict if it will **correctly classify said examples**?

Problem 2: The second is **in- and out-of-distribution detection**. Can we predict **whether a test example is from a different distribution** from the training data; can we predict **if it is from within the same distribution**?

To evaluate the solution, this paper uses two evaluation metrics.

Trade-off: The score threshold depends upon the trade-off between false negatives (fn) and false positive (fp).

Model employed: The Area Under the Receiver Operating Characteristics curve (AUROC) metric, which is a threshold-independent performance evaluation.

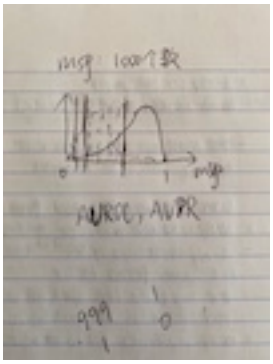
AUROC (Area Under the Receiver Operating Characteristics Curve): Is a measurement of the performance of classification model. It is threshold independent.

AUPR (Area Under the Precision-Recall Curve): Sometimes deemed more informative. It is similar idea but with two other rate which is recall and precision.

3 SOFTMAX PREDICTION PROBABILITY AS A BASELINE

Maximum softmax probability is simply the probability that the predicted class has.

For example, the output of the neural networks is [10, 5, 2], corresponding to the class 0, 1, 2 accordingly. Then we use the softmax function to convert it into an array of possibility [0.99, 0.009, 0.001].



Then we can see that the msp for the correctly classified and incorrectly classified have very **different msp distribution**.

In these two groups we also get the area under the PR and ROC curves. Basically these areas are used to determine if the model is doing great.

第二篇 Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks
增强错误分类在神经网络中的可信度

ABSTRACT

It **proposes ODIN**, which does not require any change to a pre-trained neural network.

ODIN is based on the observation that using temperature scaling and adding small perturbations to the input can separate the softmax score distributions between in- and out-of-distribution images, allowing for more effective detection.

Advantage: It is compatible with diverse network architectures and datasets. It consistently outperforms the baseline approach (from the first paper) by a large margin.

1 INTRODUCTION

Based on Hendrrycks & Gimpel's work that a well-trained neural network **tends to assign higher softmax scores** to in-distribution examples than out-of-distribution examples.

Contribution: We make the gap between in- and out-of-distribution examples further enlarged by **adding controlled perturbations**. By three bullet points, the contributions are listed:

1. It proposes a simple and effective method, ODIN (**O**ut-of-**DI**stribution detector for **N**eural networks)

2. Tests the ODIN on state-of-art network architectures under a diver set of in- and out-distribution dataset pairs.

3. Empirical analyze how parameter settings affect the performance.

2 PROBLEM STATEMENT

Assume **Px** and **Qx** denote two distinct data distributions defined on the image space **X**.

f is trained on a dataset drawn from the distribution Px.

Temperature Scaling: I think it is a denominator under the predicated output value from the neural network given input x. It's claimed to be able to **distill** the knowledge in neural networks and **calibrate** the prediction confidence in classification tasks.

Input Preprocessing: In addition to temperature scaling, this paper preprocess the input by adding small perturbations:

[SUPER COMPLEX EQUATION]

The rationale behind the preprocessing is based on Goodfellow's work, where small perturbations are added to decrease the softmax score for the true label and force the neural network to make a wrong prediction.

However, in this paper, they aim to increase the softmax score of any given input, without the need for a class label at all. As a result, **the perturbation can have stronger effect on the in-distribution images than that on out-of-distribution images.** (The perturbations can be easily computed by back-propagating the gradient of the cross-entropy loss w.r.t the input).

3 ODIN:OUT-OF-DISTRIBUTION DETECTOR

Out-of-distribution Detector: For each image **x**, the research first calculate the preprocessed image **x ~.** Next, the researcher feed the preprocessed image into the neural network, calculate its calibrated softmax score and compare the score to the threshold. If it's larger than the threshold, it's classified as in-distribution.

4 EXPERIMENTS

[TOO COMPLEX]

5 DISCUSSIONS

[TOO COMPLEX]

6 RELATED WORKS AND FUTURE DIRECTIONS

Other OOD Detection Methods:

Density estimation: It uses probabilistic models to estimate the in-distribution density and declares a test example to be out-of-distribution if it locates in the low-density areas.

Clustering method: It is based on the statistical distance, and declares an example to be out-of-distribution if it locates far from its neighborhood.

(Known to be unreliable in high-dimensional space such as image space)

- Generative adversarial networks

- Convolutional network

- Transfer representation-learning

All these works require enlarging or modifying ht eneural networks.

In the future, this paper suggests:

1. ON in-distribution images, modern neural networks tend to produce outputs with larger variance across class labels, and

2. Neural networks have larger norm of gradient of log-softmax scores when applied on in-distribution images.

7 CONCLUSIONS