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

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.

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.