Detection of Misclassified and Out of Distribution Samples

Paide Ashish
200101072
Computer Science and Engineering(B.tech)
a.paide@iitg.ac.in

Nikhil Jaishwal
236150102
MFSDS&AI
j.nikhil@iitg.ac.in

Abstract

Drawing inspiration from the groundbreaking paper titled *"A BASELINE FOR DETECTING MISCLASSIFIED AND OUT-OF-DISTRIBUTION EXAMPLES IN NEURAL NETWORKS"* our methodology delves into the key problems of identifying misclassified and out-of-distribution examples in neural networks. We access the observing performance of the different models on validating the detecting of the misclassified and the out-of-distribution examples. This report serves as a exploration into the area of anomaly detection in neural networks, offering valuable contributions to the field and paving the way for future advancements in this area of research.

1 Introduction

When the Neural Network classifiers employed in the real-world task, they often encounter challenges when the training and the testing distribution differ. These classifiers may fail by giving the high confidence predictions that are outrageously incorrect (Goodfellow et al., 2015; Amodei et al., 2016).

This occurs particularly when classifiers do not indicate the uncertainty can impede their adoption or lead to serious consequences. For instance, in medical domain, a model may consistently provide confident classifications even for difficult cases where human intervention is warranted. This lack of indication could result in erroneous diagnoses, hindering the progress of machine learning in medicine and raising concerns in AI safety (Amodei et al., 2016).

A common culprit behind this high-confidence misclassified is the softmax function, in which the probabilities change just by minor alterations in the input logits because of its exponential nature. Similarly out-of-distribution examples may also yield seemingly high confident prediction. Despite such high confidence values, anecdotal evidence and empirical studies suggest that SoftMax probabilities do not directly correspond to confidence (Nguyen & O'Connor, 2015; Yu et al., 2010; Provost et al., 1998; Nguyen et al., 2015).

To tackle this, this paper introduced a novel method that surpasses the baseline on select tasks by evaluating the quality of neural network input reconstruction to identify abnormal examples and provided methodologies for supplying out-of-distribution examples at test time, such as using images from distinct datasets or introducing realistic input distortions, with the aim of encouraging further exploration and advancement by the research community.

We adopted this methodology on various transformer models (DistilBERT , Bert , Xlnet) for the sentiment analysis task and validated the performance of the models in predicting the misclassified and the out of distribution examples.

In summary this model mentioned that Soft-Max classifier probabilities may not directly translate the confidence of the estimates . Simple statistics that are derived from the SoftMax distribution gives the effective means of detecting misclassification and out-of-distribution examples. [1]

2 Methodology

We are going to address the two key problems. The first one is error and success prediction: which can predict whether a trained model will make an error on the particular held-out test sample. Ans the second one is in-and-out-of distribution detection: which can predict whether the sample is in or out of distribution from the training data. Below we present the simple baseline for solving these problems. We evaluate our solution with two evaluation metrics.

Before addressing the two evaluation metrics , we first note that comparing detectors is not

[1] Code is available at https://github.com/ashish-paide/Detecting-Misclassified-and-OOD-data

straightforward by using the accuracy. The detector have to output the scores for the two classes for the both positive and negative classes. If the positive class is far more likely than the negative class, the model may always guess the positive class and gets high accuracy which can be misleading (Provost et al., 1998). We must then specify the score threshold so that some negative samples are classified correctly, but this depends on the trade-off between the false negatives(fn) and the false positives(fp),

To Tackle this issue we employ the Area Under the Receiver Operating Characteristic curve(AUROC) which is threshold independent performance evaluation. The ROC graph is the graph between the true positive rate (tpr = tp/(tp + fn)) and false positive rate (fpr = fp/(fp + tn)) and interpreted as the probability with the positive example has a greater detector score/value than the negative example.

The AUROC sometimes sidesteps the issue of the threshold selection and it is not ideal when data have the different base rates, Area Under the Precision-Recall curve(AUPR) adjust for these different base rates which is sometimes deemed more informative. For this reason AURP is the second evaluation metric, The PR curve plots the precision (tp/(tp+fp)) and recall (tp/(tp + fn)) against each other.

2.1 Prediction of misclassified samples:

Following the finetuning the model , our methodology as follows:

- We extract the logits and softmax probabilities from the testing dataset. These probabilities are then divided between correctly and incorrectly predicted samples.
- Next, we isolate the prediction probabilities associated with both correctly and incorrectly predicted samples.
- We construct a simple model by concatenating the prediction probabilities as inputs, with binary labels (0 for incorrectly predicted samples and 1 for correctly predicted samples).
- Subsequently, we validate the model's performance using two key metrics: area under the precision-recall curve (AUPR) and area under the receiver operating characteristic curve (AUROC).
- Additionally, we apply a similar process using Kullback-Leibler (KL) divergence, evaluating the effectiveness of the model in detecting misclassifications examples through this alternate metric.

2.2 Prediction of in and out distribution:

- We take the in and out of distribution testing samples. From this we extract the softmax probabilities of those samples respectively.
- Next, we isolate the prediction probabilities associated with both in and out distribution samples.
- We construct a simple model by concatenating the prediction probabilities as inputs, with binary labels (0 for in distribution samples and 1 for out of distribution samples).
- Subsequently, we validate the model's performance using two key metrics: area under the precision-recall curve (AUPR) and area under the receiver operating characteristic curve (AUROC).
- Additionally, we apply a similar process using Kullback-Leibler (KL) divergence, evaluating the effectiveness of the model in detecting out-of-distribution examples through this alternate metric.

3 Results and Analysis

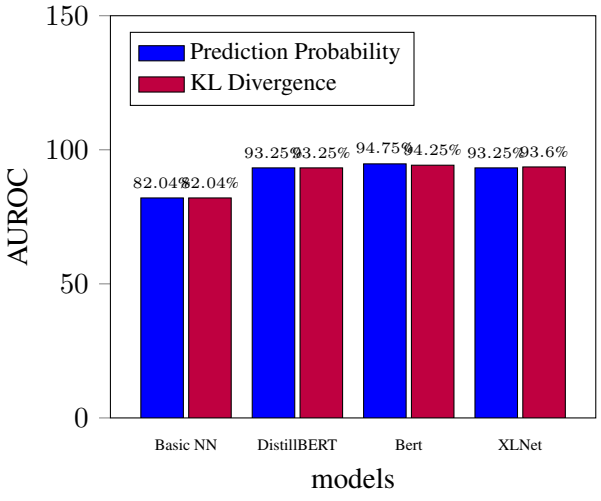

Figure 1: Success detection of Right/Wrong classification

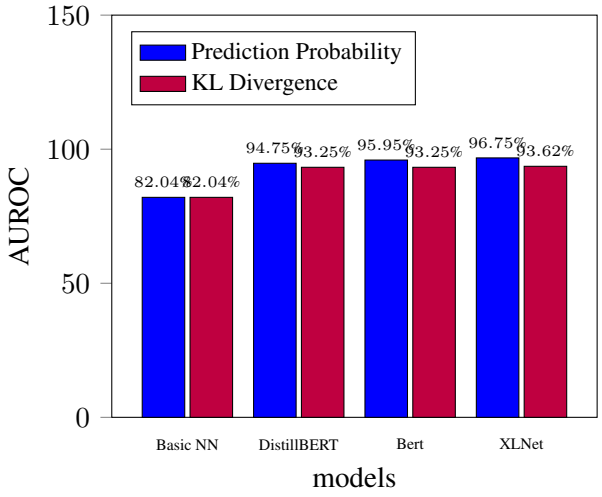

Figure 2: Error detection of Right/Wrong classification

- The high AUROC values obtained for both features, KL divergence and prediction probability, across all models indicate their effectiveness in predicting misclassified data and out-of-distribution samples. This underscores the significance of these features in identifying anomalous data instances across various models.

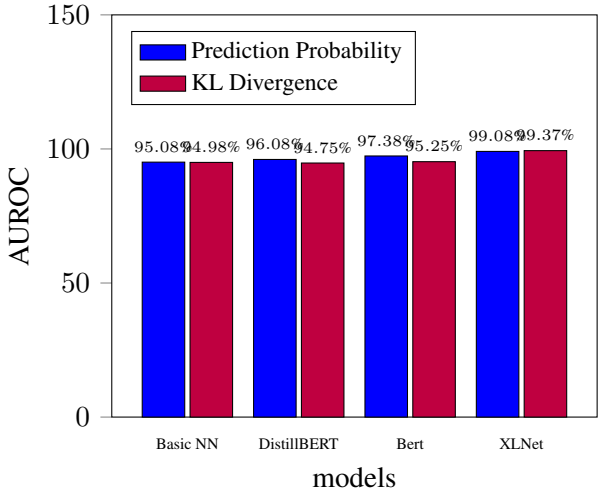

Figure 3: Abnormality detection

- The observation that high KL divergence and low prediction probability datapoints are more likely to be out of distribution suggests that these metrics effectively discriminate between in-distribution and out-of-distribution samples.

- DistilBERT and XLNet achieve exceptionally high AUPR scores for error detection (97.95% and 99.95% respectively), indicating strong performance in identifying errors.

- While BERT's AUPR score for error detection is not as high, it still performs well with an AUROC score of 95.95%.

- XLNet outperforms the other models in error detection when considering the in/out classification distinction, achieving a very high AUPR score of 99.35% and AUROC score of 99.37%.

- For both right/wrong classification and in/out distribution classification, AUROC scores are relatively high across all models, indicating good discrimination ability between classes.

- AUPR scores for prediction probability are also consistently high across models, indicating good precision in predicting probabilities.

4 Limitations

- Our approach assumes that datasets that are taken from the another source are purely out of distribution samples.

- Finetuning the models from the smaller datasets containing 20,000 samples may introduce limitation of not capturing the complexity and diversity of the underlying data distribution.

5 References

References

- Code is available at https://github.com/ashish-paide/Detecting-Misclassified-and-OOD-data

- A Baseline For Detecting Misclassified and Out of distribution Examples in Neural Networks.