Meta-Crafting: Improved Detection of Out-of-Distributed Texts via Crafting Metadata Space (Student Abstract)

Ryan Koo[1], **Yekyung Kim**[2], **Dongyeop Kang**[1], **Jaehyung Kim**[3]
[1]University of Minnesota
[2]Hyundai Motors Company
[3]KAIST
koo00017@umn.edu, dongyeop@umn.edu, jaehyungkim@kaist.ac.kr

Abstract

Detecting out-of-distribution (OOD) samples is crucial for robust NLP models. Recent works observe two OOD types: background shifts (style change) and semantic shifts (content change), but existing detection methods vary in effectiveness for each type. To this end, we propose *Meta-Crafting*, a unified OOD detection method by constructing a new discriminative feature space utilizing 7 model-driven metadata chosen empirically that well detects both types of shifts. Our experimental results demonstrate state-of-the-art robustness to both shifts and significantly improved detection on stress datasets.

Introduction

Detecting out-of-distribution (OOD) examples from in-distribution (ID, *i.e.*, training) ones is extensively explored as an important problem for reliable usage of deep neural networks (DNNs) in the real world. The general approach for *OOD detection* is finding a useful score to measure each sample's OODness, using model-driven *metadata* obtained from the trained classifiers.[1] However, Arora, Huang, and He 2021 recently showed that the effectiveness of such single-score methods largely varies depending on the type of distribution shifts. To this end, we propose a unified OOD detection method that *jointly* leverages information within model-driven *metadata* to better discriminate between ID and OOD samples capturing different types of information relevant to distinct areas of OOD examples.

Meta-Crafting: OOD Detection with Metadata

From the different effectiveness of Maximum Softmax Probability (*MSP*) and Perplexity (*PPL*) for detecting each type of OOD examples (Arora, Huang, and He 2021), we conjecture that each metadata extracts different kinds of information relevant to OOD detection. Motivated by this, we hypothesize that further improvements could be made by jointly leveraging various other metadata since each metadata captures specific characteristics of OOD samples that others do not; hence, it results in an effective unified method from their complementary effect.

[1]We use this term to refer to the measurements from the model (e.g., confidence or uncertainty) that provide additional information on examples.

Background vs. Semantic Shifts We introduce the problem space by defining the different shifts a set of OOD examples can take. Following (Ren et al. 2019), we characterize input x to be decomposed into 2 independent components: (1) *background* is characterized by population-level statistics, and (2) *semantic* is described by features that have a strong correlation with the class label. Intuitively, background or semantic shifts are based on the difference in 1) tasks or 2) class labels associated with each dataset respectively.

Selected Metadata for OOD Detection We choose seven metadata based upon (Kim et al. 2023) by selecting the most informative ones via regression model coefficients with a linear classifier. Specifically, we consider the 1) *Average confidence* and 2) *Variability* of the samples' confidence across training epochs, which observe that they effectively discriminate samples with small perturbations and unusual patterns. Next, we consider 3) *BALD* and 4) *MC-BALD* to measure model uncertainty, which estimates the mutual information between model predictions and parameters. Samples with high mutual information can reveal their true label, which we find effectively discriminates samples with small shifts. In addition, we consider 5) *Sentence density* using kNN distance on the sentence encoder's feature embeddings, which provides sentence-level uniqueness as an additional metadata. Finally, we incorporate 6) *MSP* and 7) *PPL* as metadata. In total, we utilize 7 different metadata to construct our new effective feature space for OOD detection.

Detecting OOD Samples using Metadata While the individual metadata is effective for capturing different aspects of information, combining these multi-dimensional features to calculate a single OODness is non-trivial. To this end, we utilize Mahalanobis Distance (MD) which provides an effective way to combine multi-dimensional features. We first generate each metadata for both in- and out-of-distribution samples independently; all of these metadata are then concatenated to create a new multi-dimensional feature space. With this feature z_x, we obtain the OODness s_x of each sample x by calculating MD to the closest cluster's center (Lee et al. 2018):

$$s_x = \min_{1 \le m \le M} (z_x - \mu_m)^T \sum_m^{-1} (z_x - \mu_m) \qquad (1)$$

ID	OOD	Shift	*AUROC*		
			PPL	*MSP*	***Meta* (ours)**
IMDB	c-IMDB	Semantic	53.5	63.7	**73.7**
MNLI	HANS	Background	**98.3**	55.0	92.4
	Negation	Background	44.5	60.5	**84.4**
	Len. Mismatch	Background	19.6	51.6	**82.3**
	Spell. Error	Background	43.9	57.7	**85.0**
	Word Overlap	Background	42.4	61.7	**84.8**
	Antonym	Semantic	4.5	55.3	**78.6**
	Num. Reason.	Semantic	27.5	75.8	**91.3**

Table 1: OOD detection performance on challenge datasets. The highest performing method is in bold, where a higher AUROC score indicates stronger performance. We can see that our method (*Meta*) generally exhibits an improvement as a result of considering various metadata on top of *PPL*/*MSP* holistically.

where μ and $\sum$ indicate the sample mean and covariance of the features $z_{\tilde{x}}$ of ID training samples $\tilde{x}$. Here, features are partitioned into M (the number of classes) clusters where distances are calculated for each test input x, and minimization happens over M clusters.

Experiments

Setup We demonstrate the effectiveness of our approach through a variety of ID and OOD pairs considered in Arora, Huang, and He 2021 and mark their results as a baseline for each dataset. We conduct each experiment by finetuning[2] RoBERTa (Liu et al. 2019) on the training split of each dataset as the in-distributed set with 3 different seeds at 5 epochs. Additionally, we finetune GPT-2 on each of the training splits for 1 epoch in order to generate the token perplexity (*PPL*) scores. For each ID/OOD sample, we compute the AUROC to compare the detection performance.

Datasets We benchmark the performance of our method on the Stress Test dataset of MNLI (Williams, Nangia, and Bowman 2018; Naik et al. 2018), c-IMDB (Kaushik, Hovy, and Lipton 2020), and HANS (McCoy, Pavlick, and Linzen 2019), to address three general failure cases: spurious semantic features, small shifts, and repetition. As they are constructed by adding small perturbations or unnatural patterns, hence difficult to discriminate from the original sentence, *MSP* and *PPL* are quite limited under these datasets (Table 1). For example, *Word Overlap* is generated by adding small shifts to the original MNLI dataset, *e.g.* large overlaps between the premise and hypothesis: "The country's history has been turbulent" $\rightarrow$ "The country's history has been turbulent and **true is true**".

Results Table 1 shows the results of our method's performance on the challenge datasets. We find that the consideration of multiple metadata is consistently able to better detect unnatural patterns contained by the challenge datasets and hence better generalizes compared to the single-score methods. On average, our results over all the datasets see an improvement by 101% against *PPL* and 39% against *MSP*. This result coincides with our motivation, where the consideration of additional, more sophisticated metadata incorporates other useful aspects of training data and thus better detects OOD samples.

[2]We use the trained model and parameters at https://github.com/uditarora/ood-text-emnlp

Conclusion

In this paper, we propose *Meta-Crafting*, a unified method to detect different types of a shift in OOD simultaneously by combining metadata. We demonstrated the effectiveness of our method in various types of challenge datasets, where our method shows an improvement over most challenge data that previous works generally fail to detect, such as human-generated counterfactual data and rule-based data. Our results illustrate the benefits of considering several features in a unified manner rather than a single feature at a time but also highlight areas of improvement where the current considered metadata fails.

References

Arora, U.; Huang, W.; and He, H. 2021. Types of Out-of-Distribution Texts and How to Detect Them. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kaushik, D.; Hovy, E.; and Lipton, Z. C. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *ICLR*.

Kim, J.; Kim, Y.; de Langis, K.; Shin, J.; and Kang, D. 2023. infoVerse: A Universal Framework for Dataset Characterization with Multidimensional Meta-information. arXiv:2305.19344.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *NeurIPS*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692.

McCoy, R. T.; Pavlick, E.; and Linzen, T. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Naik, A.; Ravichander, A.; Sadeh, N.; Rose, C.; and Neubig, G. 2018. Stress Test Evaluation for Natural Language Inference. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*.

Ren, J.; Liu, P. J.; Fertig, E.; Snoek, J.; Poplin, R.; Depristo, M.; Dillon, J.; and Lakshminarayanan, B. 2019. Likelihood ratios for out-of-distribution detection. In *NeurIPS*.

Williams, A.; Nangia, N.; and Bowman, S. R. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *NAACL*.