Revisiting Mahalanobis Distance for Transformer-Based Out-of-Domain Detection

Alexander Podolskiy[1], **Dmitry Lipin**[1], **Andrey Bout**[1],
Ekaterina Artemova[1, 2], **Irina Piontkovskaya**[1]

[1] Huawei Noah's Ark Lab, Moscow, Russia
[2] HSE University, Moscow, Russia
{podolskiy.alexander, dmitry.lipin, bout.andrey, artemova.ekaterina, piontkovskaya.irina}@huawei.com

Abstract

Real-life applications, heavily relying on machine learning, such as dialog systems, demand out-of-domain detection methods. Intent classification models should be equipped with a mechanism to distinguish seen intents from unseen ones so that the dialog agent is capable of rejecting the latter and avoiding undesired behavior. However, despite increasing attention paid to the task, the best practices for out-of-domain intent detection have not yet been fully established.

This paper conducts a thorough comparison of out-of-domain intent detection methods. We prioritize the methods, not requiring access to out-of-domain data during training, gathering of which is extremely time- and labor-consuming due to lexical and stylistic variation of user utterances. We evaluate multiple contextual encoders and methods, proven to be efficient, on three standard datasets for intent classification, expanded with out-of-domain utterances. Our main findings show that fine-tuning Transformer-based encoders on in-domain data leads to superior results. Mahalanobis distance, together with utterance representations, derived from Transformer-based encoders, outperforms other methods by a wide margin (1-5% in terms of AUROC) and establishes new state-of-the-art results for all datasets.

The broader analysis shows that the reason for success lies in the fact that the fine-tuned Transformer is capable of constructing homogeneous representations of in-domain utterances, revealing geometrical disparity to out of domain utterances. In turn, the Mahalanobis distance captures this disparity easily.

Introduction

The usability of dialog systems depends crucially on the capability of dialog agents to recognize user intents. Recently deep classifiers have been widely used to recognize user intents, leveraging efficient pre-training, and large amounts of labeled data. However, the scope of annotated corpora is inherently limited, leading to unsatisfactory results when presented with unseen intents. Rather than trying to match user utterances to a limited number of intent classes, dialog agents may be equipped with an auxiliary mechanism to distinguish between seen and unseen intents, i.e., to identify out-of-domain (OOD) utterances. The OOD detection mechanism must handle unseen intents to prevent the erroneous actions of dialog agents.

Multiple recent papers emphasize the increasing importance of OOD utterances detection caused by the spreading integration of classification models to real-life applications and dialog systems. Simultaneously, in the overwhelming majority of papers, the task is approached in an unsupervised way, see (Gangal et al. 2020; Larson et al. 2019; Zheng, Chen, and Huang 2020). To this end, the primary approach relies on a decision rule, which is defined to score each utterance. The scores are further used to reject OOD utterances or to subject to further processing in-domain (ID) ones. An intuitive yet efficient decision rule determines a threshold for `softmax` output probabilities, measuring the classifier's confidence. The less confident the classifier is, the higher are the chances to reject the utterance. Other decision rules rely on distance-based approaches to check whether an utterance falls out of ID space.

With Transformer-based contextual encoders becoming core to almost, if not all, Natural Language Processing (NLP) methods, undoubtedly, their performance for intent classification is well-studied. However, the performance of Transformers in the OOD detection task so far has been little explored. Hendrycks et al. (2020) provide evidence that Transformers generalize well to unseen domains in sentiment classification and sentence pair modeling tasks, suggesting that Transformers will perform better than previous models for the task of OOD utterance detection, too. This paper fills this gap in the evaluation of Transformers.

The key idea of this paper is to conduct a comprehensive comparison of the performance of different contextual encoders in multiple settings. We adopt three dialog datasets designed for the task of OOD intent detection along with the current best practices and state-of-the-art methods for the task. Although Transformers primarily outperform other contextual encoders by a wide margin, they serve as especially useful embedders for the distance-based methods of OOD detection. When fine-tuned on ID data, Transformers form dense clusters of ID utterances, which are easy to locate with Mahalanobis distance.

To summarize, the key contributions of the paper are as follows:

1. We evaluate multiple contextual encoders and best practices for OOD detection on three common datasets for intent classification, expanded with out-of-domain utterances;
2. We show that not only fine-tuning Transformers on ID data consistently improves OOD detection, but also that when combined with Mahalanobis distance, it established new state-of-the-art results;
3. We discover that the fine-tuned Transformer is capable of constructing homogeneous representations of ID utterances, revealing geometrical disparity to OOD ones, captured easily in turn by the Mahalanobis distance.

Related Work

Methods for OOD detection can be roughly grouped based on whether they have access to OOD data and whether they utilize ID labels.

Classification methods require access to OOD data for supervision. Larson et al. (2019) use two supervised settings: 1) binary classification, so that all ID classes are classified against OOD one, and 2) training an additional class for OOD inputs. (Kamath, Jia, and Liang 2020) train an additional model, a **calibrator**, which identifies inputs on which the classifier errs, and rejects those inputs, for which an error is likely. (Hendrycks, Mazeika, and Dietterich 2018) utilize OOD data for outlier detection by training models to increase entropy on OOD examples.

However, in real-life applications gathering and maintaining OOD data is complicated by the lexical and stylistic variation of user utterances (Schuster et al. 2019). For this reason, methods without OOD supervision gain more attention.

Outputs of the classifier, trained with the supervision of ID classes, can be exploited as a score for OOD inputs. Maximum softmax probability (Hendrycks and Gimpel 2017) is recognized as a strong baseline, when used with deep classifiers, improved further by introducing temperature scaling (Liang, Li, and Srikant 2018). KL-divergence captures the changes in prediction distributions learned for an ID class by the classifier and detects the arbitrary guesses made for OOD inputs (Yilmaz and Toraman 2020).

Generative methods use a natural ability of language models and other generative models to estimate the likelihood of the inputs (Nalisnick et al. 2018; Ren et al. 2019). (Zheng, Chen, and Huang 2020) utilize ID inputs and unlabeled data to generate pseudo-OOD utterances with a generative adversarial network, improving OOD detection on a dialog dataset.

Distance-based methods treat distance estimation as an OOD score: the further an input is from ID inputs, the higher are the chances that it is OOD (Mandelbaum and Weinshall 2017; Gu, Akoglu, and Rinaldo 2019; Lee et al. 2018).

Other research direction include **Bayesian estimation for uncertainty** derived from learned distributions over network weights (Malinin and Gales 2018; Blundell et al. 2015), processing of **lexical features** (Ghosal et al. 2018) and training prototypical networks to define class prototypes for each ID class (Tan et al. 2019).

Background

Let $\mathscr{D}_{ID} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a dataset, where x_i is an input utterance and $y_i \in \Upsilon$ is its class label. Than Υ is the set of seen, in-domain classes, and the total number of classes is $|\Upsilon| = N$. Assume that ID utterances are drawn from the distribution P_{ID} and that there exists an OOD distribution, P_{OOD}, which differs from P_{ID}. Finally, suppose that a scoring function maps an utterance x into a real number. The OOD detector then accepts the ID utterances and rejects the OOD utterances according to the decision rule in Eq. 1.

$$R(x) = \begin{cases} \textbf{reject}, & \text{if } d(x) \geq \theta \\ \textbf{accept}, & \text{otherwise} \end{cases} \tag{1}$$

where θ is a threshold, d can be either independent from y, otherwise model joint $d(x,y)$ or conditional $d(x|y)$ dependence. Ideally, we want $d(x) < d(\hat{x})$ for all $x \sim P_{IN}$, $\hat{x} \sim P_{OOD}$.

Methods

We adopt several methods that do not rely on access to OOD data and are shown to be effective for OOD detection in vision and natural language tasks. We exploit Maximum Softmax Probability (MSP) as a strong baseline (Hendrycks and Gimpel 2017), Likelihood ratio (Gangal et al. 2020) as the current state-of-the-art method for dialog data. We use Mahalanobis distance (De Maesschalck, Jouan-Rimbaud, and Massart 2000), an advanced distance-based method, computed in multiple ways. It is the most straightforward to compute the Mahalanobis distance to the closest ID classes, assuming that the ID labels are provided. If not, marginal Mahalanobis distance allows computing the distance to the ID data centroid.

Maximum Softmax Probability (MSP) requires a pretrained classifier f with a softmax output layer (Hendrycks and Gimpel 2017). Let $p_y(x)$ denote the probability, assigned by f, to the utterance x to belong to class y. The less classifier is confident with its prediction, the higher is the OOD score:

$$d(x) = 1 - \max_{y \in \Upsilon} p_y(x). \tag{2}$$

To prevent the classifier from becoming too confident in its prediction, (Liang, Li, and Srikant 2018) introduce softmax temperature scaling at the test time: $p_y(x) = e^{\frac{z_y}{\tau}} / \sum_{y \in \Upsilon} e^{\frac{z_y}{\tau}}$, where z_y denotes the logit for label y while τ denotes the softmax temperature.

Likelihood Ratio (LLR) exploits two language models (Gangal et al. 2020). One of them, $L(x)$, is trained on source data and aims to capture ID utterances' semantics. The second language model, L_{bg}, addressed as a background model, is trained on corrupted with some noise source data and aimed at learning the background statistics. The final score is computed as follows in Eg. 3.

$$d(x) = -\log \frac{L(x)}{L_{bg}(x)}, \tag{3}$$

Mahalanobis distance is a way to determine the closeness of an utterance to a set of utterances belonging to the

class c. Following (Lee et al. 2018), we define Mahalanobis distance, serving as OOD score, as:

$$d(x) = \min_{c \in \Upsilon} (\psi(x) - \mu_c)^T \Sigma^{-1} (\psi(x) - \mu_c), \quad (4)$$

where $\psi(x)$ is a vector representation of the utterance x, μ_c is the centroid for a class c and Σ is the co-variance matrix. The estimations of μ_c and Σ are defined by

$$\mu_c = \frac{1}{N_c} \sum_{x \in \mathscr{D}_{in}^c} \psi(x),$$

$$\Sigma = \frac{1}{N} \sum_{c \in \Upsilon} \sum_{x \in \mathscr{D}_{in}^c} (\psi(x) - \mu_c)(\psi(x) - \mu_c)^T,$$

where $\mathscr{D}_{IN}^c = \{x | (x,y) \in \mathscr{D}_{in}, y = c\}$, N is the total number of utterances, and N_c is the number of utterances belonging to class c.

Datasets

To the best of our knowledge, we are the first to evaluate OOD detection with three NLU datasets, consisting of both ID and OOD utterances.

CLINC150 is an intent classification dataset, modeling a real-life situation. Some utterances fall out of domains, covered by train data (Larson et al. 2019). The total number of ID classes in CLINC150 is equal to 150. The OOD utterances relate to actions not supported by existing ID intents.

ROSTD extends the English part of multilingual dialog dataset with OOD utterances (Schuster et al. 2019; Gangal et al. 2020). The hierarchical label structure of ROSTD allows us to experiment with both a larger number of classes (12) or "coarsened" classes (3). Following (Gangal et al. 2020), we experiment with both variants and refer to them as ROSTD and ROSTD-COARSE. The OOD part consists mainly of subjective, under-specified, or over-emotional utterances that do not fall into ID classes.

SNIPS has no explicit ID/OOD split. The total number of intents is 7. Following (Lin and Xu 2019) setup, we randomly split all labels into ID and OOD parts. The ID part covers about 75% of the whole dataset. We average the results of all splits.

Table 1 presents with dataset statistics.

Embeddings and Encoders

We evaluate three representation models, ranging from bag-of-words, static pre-trained word embeddings up to contextualized encoders.

	CLINC150	ROSTD	SNIPS
Number train IND	15K	30K	13K
Number val IND	3K	4K	0.7K
Number val OOD	0.1K	1.5K	–
Number test IND	4.5K	8.6K	0.7K
Number test OOD	1K	3K	–

Table 1: Dataset statistics

Bag-of-words. We use the bag-of-words model (Harris 1954), which shows stable performance due to its low variance.

Static word embeddings. We use GloVe (Pennington, Socher, and Manning 2014) as inputs to a convolutional neural network (CNN) and long short-term memory (LSTM), trained further with the supervision of ID data. The CNN architecture follows one used in (Zheng, Chen, and Huang 2020). We use GloVe vectors as inputs to language models needed for LLR. LSTM is used as an underlying model of LLR, trained on ID data with language modeling objective. We train the background model on the ID data with added uniform noise. We find that 0.5 noise probability performs the best.

Pre-trained Transformers. We utilize multiple BERT-based models (Devlin et al. 2019), which are pre-trained Transformers, trained with a self-supervised masked language modeling objective. Additionally to BERT-base and BERT-large, we use RoBERTa-base and RoBERTa-large models (Liu et al. 2019). We use distilled versions of both BERT and RoBERTa, DistillBERT and DistillRoBERTa (Sanh et al. 2019).

Each CNN, LSTM, and Transformer model is used as a classifier with the MSP method and as an embedder with Mahalanobis distance. We follow the standard fine-tuning procedure to fine-tune each model for three ID intent classification tasks. We tune hyper-parameters to maximize performance on the validation set for each of the ID intent classification tasks. We perform our experiments with PyTorch[1], PyTorch Lightning and Hugging Face Transformers library (Wolf et al. 2019).

Evaluation

The task of OOD detection is a binary classification task, where OOD utterances should be distinguished from ID utterances. In the unsupervised setting, a scoring function is used to assign an OOD score.

AUROC, the area under the Receiver Operating Characteristic, can be interpreted as the probability of randomly sampled ID utterance having a lower OOD score than randomly sampled OOD one.

AUPR_{OOD}, the area under Precision-Recall Curve, requires taking OOD as the positive class. It is more suitable for highly imbalanced data in comparison to AUROC.

FPR@X corresponds to False Positive Ratio with the decision threshold is set to $\theta = \sup\{\tilde{\theta} \in \mathbb{R} \mid \text{TPR}(\tilde{\theta}) \leq \text{X}\}$ where $\text{TPR} \in [0,1]$ is a True Positive Rate. This metric also requires selecting one class as positive. Different approaches are used, e.g. Gangal, Arora, Einolghozati, and Gupta (2020) treat the OOD class as positive one, while Zheng, Chen, and Huang (2020) choose ID class. We report both metrics: FPR@X_{ID} means that ID class is treated as positive, and $\text{FPR@X}_{\text{OOD}}$ means the same for OOD class.

Two metrics, AUROC and AUPR_{OOD} are threshold-independent. FPR@X requires picking a threshold.

[1]PyTorch version 1.4.0, PyTorch Lightning version 0.7.5, Hugging Face Transformers version 2.8.0

Dataset	Model	AUROC ↑	AUPR_{OOD} ↑	$\text{FPR@95}_{\text{OOD}}$ ↓	$\text{FPR@95}_{\text{ID}}$ ↓
CLINC150	BoW MSP	91.5 ± 0.0	66.7 ± 0.2	31.7 ± 0.4	43.9 ± 0.9
	LSTM MSP	90.9 ± 0.6	67.8 ± 2.1	31.2 ± 2.0	50.7 ± 3.0
	CNN MSP	94.1 ± 0.6	80.8 ± 2.1	26.4 ± 4.0	24.4 ± 2.8
	CNN Maha	95.2 ± 0.2	76.2 ± 1.4	16.4 ± 1.1	27.8 ± 1.6
	LLR	91.4 ± 0.3	73.1 ± 1.0	37.0 ± 1.5	39.9 ± 1.5
	$\text{BERT}_{\text{base}}$ Maha	97.3 ± 0.1	88.6 ± 1.0	10.9 ± 0.7	12.5 ± 1.1
	$\text{BERT}_{\text{base}}$ SNGP[1]	96.9 ± 1.0	88.0 ± 1.0	–	–
	RoBERTa MSP	97.1 ± 0.6	91.2 ± 1.3	11.6 ± 2.4	12.5 ± 2.2
	RoBERTa Maha	$\mathbf{98.4 \pm 0.1}$	$\mathbf{94.5 \pm 0.5}$	$\mathbf{6.8 \pm 0.8}$	$\mathbf{7.3 \pm 1.1}$
ROSTD	BoW MSP	94.2 ± 0.1	86.7 ± 0.1	30.5 ± 0.4	25.8 ± 0.2
	LSTM MSP	73.7 ± 8.3	60.6 ± 12.1	63.0 ± 6.0	57.4 ± 13.8
	CNN MSP	95.2 ± 1.2	88.2 ± 2.8	22.2 ± 6.3	32.5 ± 6.0
	CNN Maha	98.1 ± 0.2	93.3 ± 0.7	7.6 ± 1.5	7.8 ± 1.3
	LLR	97.7 ± 0.2	95.6 ± 0.3	12.3 ± 1.7	9.3 ± 1.0
	RoBERTa MSP	99.3 ± 0.2	98.2 ± 0.4	2.2 ± 0.8	1.8 ± 0.9
	RoBERTa Maha	$\mathbf{99.8 \pm 0.1}$	$\mathbf{99.5 \pm 0.3}$	$\mathbf{0.5 \pm 0.4}$	$\mathbf{1.0 \pm 0.5}$
ROSTD-coarse	BoW MSP	98.0 ± 0.1	96.0 ± 0.1	7.8 ± 0.7	6.6 ± 0.2
	LSTM MSP	86.3 ± 7.8	80.2 ± 10.6	52.7 ± 13.5	32.0 ± 15.3
	CNN MSP	97.0 ± 0.8	94.7 ± 1.3	19.8 ± 8.1	10.4 ± 2.7
	CNN Maha	99.0 ± 0.2	97.5 ± 0.4	4.5 ± 1.1	4.6 ± 0.8
	LLR	97.7 ± 0.2	95.5 ± 0.4	12.5 ± 1.4	9.1 ± 0.9
	RoBERTa MSP	99.2 ± 0.5	98.8 ± 0.5	0.6 ± 0.5	1.7 ± 0.9
	RoBERTa Maha	$\mathbf{99.8 \pm 0.1}$	$\mathbf{99.6 \pm 0.1}$	$\mathbf{0.2 \pm 0.1}$	$\mathbf{0.7 \pm 0.4}$
SNIPS 75	BoW MSP	92.4 ± 2.0	76.9 ± 6.8	30.7 ± 4.3	41.6 ± 6.3
	LSTM MSP	81.7 ± 10.9	59.6 ± 15.3	49.9 ± 24.7	59.0 ± 15.3
	CNN MSP	93.7 ± 2.3	78.7 ± 9.1	24.4 ± 9.1	20.2 ± 8.8
	CNN Maha	87.1 ± 9.4	75.4 ± 12.6	49.3 ± 33.6	37.8 ± 18.7
	LLR	83.5 ± 5.2	61.3 ± 12.9	65.1 ± 16.0	58.1 ± 9.0
	RoBERTa MSP	95.3 ± 2.8	85.7 ± 5.6	25.5 ± 22.3	18.2 ± 9.1
	RoBERTa Maha	$\mathbf{97.6 \pm 1.9}$	$\mathbf{92.9 \pm 5.4}$	$\mathbf{12.3 \pm 10.3}$	$\mathbf{11.2 \pm 10.5}$

Table 2: Comparison of OOD detection performance. Each result is an average of 10 runs. ↑ – greater is better, ↓ – lower is better

Out-of-Domain Detection

Transformers With Mahalanobis Distance are Better at OOD Detection Than Other Models

Table 2 presents with the results of experiments. On all datasets, RoBERTa equipped with the Mahalanobis distance outperforms baselines, and other methods, including RoBERTa with the MSP score. Advantages are even more evident for CLINC150, which is less lexically and syntactically diverse and challenging.

On the CLINC150 dataset, Mahalanbois distance combined with Transformer-based embeddings outperforms recently proposed BERT SNGP (Spectral-normalized Neural Gaussian) (Liu et al. 2020). In order to make a fair comparison, we show the performance of the $\text{BERT}_{\text{base}}$ Maha.

LSTM with MSP performs at the baseline level and is outperformed with CNN with MSP, followed by the previously established state-of-the-art method, LLR (Gangal et al. 2020). In turn, it does not cope well with CLINC150 and SNIPS and is slightly outperformed by CNN with Mahalanobis distance. LLR might be challenging to apply, as the background model can still learn semantics from the data, even though it is trained on the noisy inputs. There is a high variance in the background model training due to the extensive vocabulary size.

The Mahalanobis distance and its variants depend primarily on the embeddings learned by a model. All models equipped with MSP were fine-tuned using cross-entropy loss for intent classification. The Table 2 confirms that such fine-tuning does not always help the models generate informative embeddings. CNN with the Mahalanobis distance shows moderate performance on the ROSTD dataset and its coarse version. However, performance severely drops on more challenging datasets.

Semantically Close ID and OOD Classes are Often Confused

RoBERTa with the Mahalanobis distance score is affected by multiple factors, such as annotation issues and confusion

[1]Results are adopted from (Liu et al. 2020)

	RoBERTa			CNN		
	CLINC150	ROSTD	SNIPS	CLINC150	ROSTD	SNIPS
Pairwise similarity between centroids	0 ± 0.08	-0.05 ± 0.13	-0.16 ± 0.12	0.35 ± 0.11	0.24 ± 0.09	0.15 ± 0.02
Centroid length	19.75 ± 0.23	18.09 ± 0.36	18.57 ± 0.64	23.07 ± 1.38	19.12 ± 1.11	18.29 ± 1.08
Similarity between ID instances and centroids	0.96 ± 0.11	0.95 ± 0.08	0.98 ± 0.04	0.92 ± 0.08	0.96 ± 0.44	0.99 ± 0.05

Table 3: Descriptive statistics of embedding space. Both spaces are derived from fine-tuned models with ID supervision. We show statistics for only one of the SNIPS splits

Figure 1: t-SNE visualization of CLINC150 ID classes. Embeddings are derived from fine-tuned RoBERTa for ID classification. ID classes are easily separated

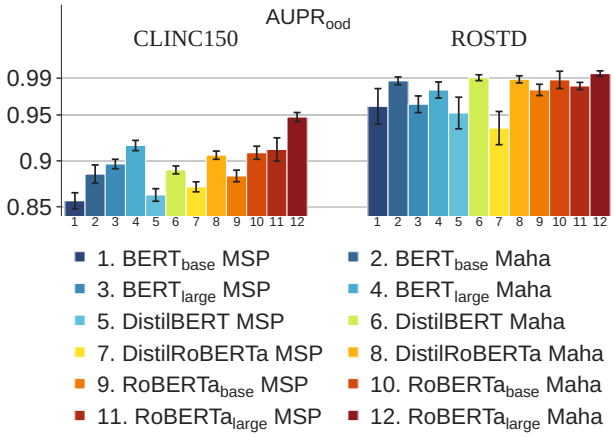


Figure 2: Comparison of models of different sizes on ROSTD and CLINC150. Maha stands for Mahalanobis distance

between semantically related utterances.

Mislabeled instances cause top errors made for CLINC150, e.g. *give me the weather forecast for today* is labeled as OOD but is related to the intent **weather**. Similarly, an OOD utterance *how old is Jennifer Anniston?* is incorrectly assigned with the intent **how old are you?**, used to question about the dialog agent's personality.

Other errors include confusion between semantically related utterances. For example, CLINC150 contains intent **text** that is related only to sending a text message. Utterances related to similar actions, such as *read my friend's text message*, are erroneously accepted.

Similar issues appear in SNIPS. The structure of ID-OOD splits explains the high variance of metrics. If two semantically related or often confused intents get into different sets, the resulting measures drop significantly. For example, it is challenging if the intent **SearchScreenEvent** is ID and **SearchCreativeWork** is OOD. The rest of the ID intents do not provide enough supervision to learn the former intent's exact semantics to more clearly separate from the latter.

On the ROSTD dataset, we observe the same errors caused by the semantic similarity between an OOD utterance and ID intents. Another source of errors is the lexical discrepancy between ID utterances in the train and test sets.

Is Bigger Model Better?

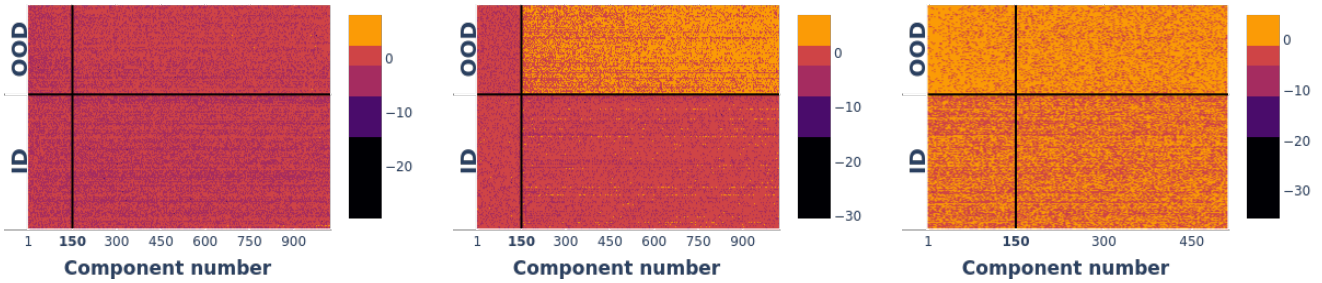
We utilize base, large, and distilled versions of BERT-based models. Fig. 2 compares the performance of models with different sizes on two datasets, CLINC150 and ROSTD. On a more diverse dataset, CLINC150, we see that larger models outperform smaller versions. On the ROSTD dataset, this difference is not so prominent but persists. By comparing distilled versions with their respective teachers, we note that the distillation does not affect the Mahalanobis distance, unlike the MSP score. Hence, Mahalanobis distance is more robust to distillation than MSP.

Diverse Pre-Training Data Improves OOD Detection

Fig. 2 shows that RoBERTa has better OOD detection capabilities than BERT. The core difference between the two models is that RoBERTa was pre-trained on a larger and more diverse dataset than BERT. Thus, we hypothesize that pre-training on larger amounts of data improves model robustness to OOD instances. Recent studies confirm this effect in computer vision (Hendrycks, Lee, and Mazeika 2019; Orhan 2019) and natural language processing (Hendrycks et al. 2020).

Features of Embeddings Space

Transformer models, fine-tuned on ID data, coupled with Mahalanobis distance, show excellent performance for OOD detection. A possible explanation could be the way the space of Transformer-based embeddings is settled. Further, we compare geometrical features of two different embedding spaces, derived from RoBERTa model and CNN for com-



(a) RoBERTa, no fine-tuning (b) RoBERTa, fine-tuned (c) CNN, trained

Figure 3: These heatmaps represent each utterance from the CLINC150 test set as the vector of Mahalanobis distance terms, computed according to Eq. 5 and sorted in the decreasing order of explained variance. Each row stands for an utterance. The horizontal solid line separates the OOD utterances (above the line) from the ID ones (below the line). The vertical solid line splits each heatmap into two parts: to the left are components numbered lower than 150, to the right are components numbered above 150. 150 is the number of classes in the CLINC150 dataset. Only fine-tuned RoBERTa-based vectors clearly distinguish ID and OOD utterances. The difference between ID and OOD is less evident in (c) and almost indistinguishable in (a). However, in (b), the values of the components, starting from the 150^{th} one (in yellow), are lower than those of ID ones (in red).

parison (see Table 3). The differences between the spaces are even sharper if the number of ID classes is high.

ID class centroids are mutually orthogonal. The pairwise cosine similarity between centroids approaches zero as its mean value and standard deviation are close to zero. This reveals that the centroids are mutually orthogonal, as all angles are close to $\frac{\pi}{2}$. This phenomenon is present for the space of Transformer-based embeddings and does not hold for CNN-based embeddings.

ID class centroids lay on a sphere. The length of Transformer-based centroids does not vary much, as the deviation from the sphere is less than 2% of its radius. On the other hand, CNN embeddings deviate more significantly.

ID classes form clusters around centroids. The deviation of ID instances from the centroids according to cosine similarity is small both for CNN and RoBERTa embeddings. ID data is well clustered, and the classes are well separated from each other, as depicted in Fig. 1.

ID data can be approximated by low-dimensional subspace in the embedding space, because ID embeddings are close to class centroids, and the number of classes is significantly lower than the dimension of the embeddings ($N \ll d$).

For further analysis, we consider several Mahalanobis distance variants. Following Kamoi and Kobayashi (2020), we introduce the equivalent Mahalanobis distance form, based on Principal Component Analysis of the class-wise centered ID data:

$$d(\psi(x)) = \min_c \sum_{i=1}^{d} \frac{y_i^2(\psi(x) - \mu_c)}{\lambda_i}, \tag{5}$$

where $y_i(\psi(x))$ is the i-th component of the PCA transform of $\psi(x)$, λ_i are explained variances of the corresponding principal components, μ_c are class centroids.

Kamoi and Kobayashi (2020) introduced two modifications of Eq. 5, namely, **marginal Mahalanobis distance**, which ignores class information and uses instead a single mean vector for all ID classes, (see Eq. 6) and **partial Mahalanobis distances**: it is the version of the equations (5) and (6) with the summation starting from N-th component. Eq. 7 corresponds to the **partial marginal** variant. Marginal Mahalanobis distance aims at using more compact data representation in the form of a single ID centroid, helping to reduce the amount of data needed for OOD detection. Partial variant utilizes the most important terms only.

$$d(\psi(x)) = \sum_{i=1}^{d} \frac{y_i^2(\psi(x) - \mu)}{\lambda_i} \tag{6}$$

$$d(\psi(x), N) = \sum_{i=N}^{d} \frac{y_i^2(\psi(x) - \mu)}{\lambda_i}, \tag{7}$$

where

$$\mu = \frac{1}{N} \sum_{x \in \mathcal{D}_{in}} \psi(x),$$

stands for the ID data centroid.

Mahalanobis distance can efficiently utilize low-dimensional nature of ID data. Following the properties of PCA (Murphy 2012), if the data is approximately N-dimensional, it is explained by the first N principal components. That means that for ID data, all the terms in the Eq. 6, 7 are little, while OOD data can be detected by important loadings of the terms $\frac{y_i^2}{\lambda_i}$ with $i > N$. To check this, we plot the terms of the Eq. 7 for ID and OOD data, Fig. 3. Fig. 3 shows that when decomposed with the Mahalanobis distance embeddings of fine-tuned RoBERTa fall into two parts. The last components of OOD embeddings have a higher variance when compared to the first ones. This phenomena is observed neither for ID embeddings nor for RoBERTa without fine-tuning nor for the trained CNN.

Comparison of Other Distances

We compare Mahalanobis distance variants to explore this matter: original, marginal Mahalanobis, and their partial versions. Additionally, we exploit Euclidean distance to complete our evaluation.

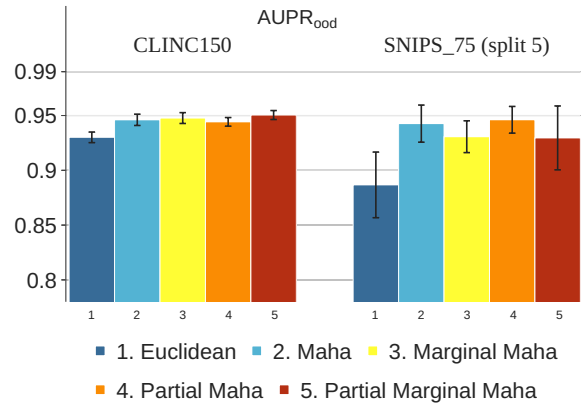


Figure 4: Comparison of different distances. Mahalanobis distance and its variants outperform Euclidean distance by a wide margin.

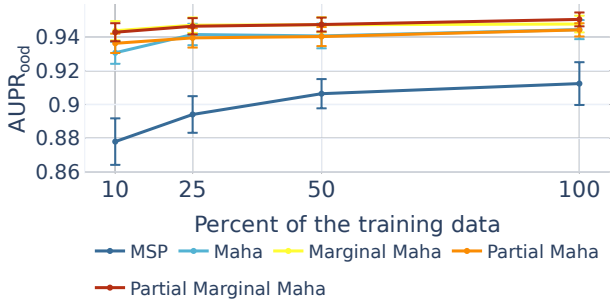


Figure 5: OX: fraction of train data used, CLINC150, OY: performance of OOD detection score. Mahalanobis distance and its variants need less data for OOD detection.

All Mahalanobis distance variants outperform Euclidean distance by far; see Fig. 4. Euclidean distance does not take the correlation between features into account. Although there is little difference between Mahalanobis distance variants, partial and marginal variants are more stable when varying training data size. Marginal Mahalanobis distance is less affected by the reduction of training data; see 5.

Conclusion

Out-of-Domain (OOD) detection task is becoming core to modern dialog systems. Successful detection and rejection of OOD utterances in real-life applications increase the dialog assistant's credibility and improves user experience. This paper compared multiple techniques for unsupervised OOD detection, applied to three commonly used NLU datasets, in particular, CLINC150, ROSTD, and SNIPS. We exploited different text representation models, ranging from the old-fashioned bag-of-word modes to the most recent pre-trained Transformers. We adopted best practices used in the vision domain and previously established state-of-the-art methods within the scope of unsupervised methods, namely, Maximum Softmax Probability, Likelihood Ratio, and Mahalanobis distance, along with its modifications.

With the help of Transformer-based models, equipped with Mahalanobis distance, we establish new state-of-the-art results. To that end, we show that fine-tuning with ID data's supervision plays a crucial role, allowing re-shaping, favorable for the task, of the embedding space. These results are supported in line with (Reimers and Gurevych 2019), confirming that fine-tuning Transformers improves the performance of the downstream unsupervised tasks. What is more, the proposed pipeline, i.e., fine-tuning a Transformer and using Mahalanobis distance, is robust to distillation. Supporting smaller models is essential for edge devices, where distilled models are usually deployed. Reduced in size, distilled versions of pre-trained Transformers models perform on par with the full-size models. Mahalanobis distance remains stable, even when used with a distilled model.

Still, there are some limitations to the Mahalanobis OOD score. In the first place, it depends on the geometrical features of the embedding space, which could be spoilt if, for example, the embedder is used simultaneously as a classification model and happens to overfit. Secondly, the greatest challenge is caused by semantically similar utterances, of which one is ID, and the other is OOD. For example, this can happen if the dialog assistant supports only one of two related actions. Future research directions should consider such cases and the trade-off between the accuracy of intents classification and OOD detection performance.

Acknowledgments

Ekaterina Artemova is partially supported by the framework of the HSE University Basic Research Program.

References

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight Uncertainty in Neural Network. In *Proceedings of the International Conference on Machine Learning*, 1613–1622.

De Maesschalck, R.; Jouan-Rimbaud, D.; and Massart, D. L. 2000. The Mahalanobis distance. *Chemometrics and intelligent laboratory systems* 50(1): 1–18.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Gangal, V.; Arora, A.; Einolghozati, A.; and Gupta, S. 2020. Likelihood Ratios and Generative Classifiers for Unsupervised Out-of-Domain Detection in Task Oriented Dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 7764–7771.

Ghosal, T.; Sonam, R.; Saha, S.; Ekbal, A.; and Bhattacharyya, P. 2018. Investigating domain features for scope detection and classification of scientific articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 7–12.

Gu, X.; Akoglu, L.; and Rinaldo, A. 2019. Statistical Analysis of Nearest Neighbor Methods for Anomaly Detection. In *Advances in Neural Information Processing Systems*.

Harris, Z. S. 1954. Distributional structure. *Word* 10: 146–162.

Hendrycks, D.; and Gimpel, K. 2017. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In *Proceedings of International Conference on Learning Representations*.

Hendrycks, D.; Lee, K.; and Mazeika, M. 2019. Using Pre-Training Can Improve Model Robustness and Uncertainty. In *Proceedings of the International Conference on Machine Learning*, 2712–2721.

Hendrycks, D.; Liu, X.; Wallace, E.; Dziedzic, A.; Krishnan, R.; and Song, D. X. 2020. Pretrained Transformers Improve Out-of-Distribution Robustness. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2744–2751.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. 2018. Deep Anomaly Detection with Outlier Exposure. In *Proceedings of International Conference on Learning Representations*.

Kamath, A.; Jia, R.; and Liang, P. 2020. Selective Question Answering under Domain Shift. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 5684–5696.

Kamoi, R.; and Kobayashi, K. 2020. Why is the Mahalanobis Distance Effective for Anomaly Detection? *arXiv preprint arXiv:2003.00402* .

Larson, S.; Mahendran, A.; Peper, J. J.; Clarke, C.; Lee, A.; Hill, P.; Kummerfeld, J. K.; Leach, K.; Laurenzano, M. A.; Tang, L.; et al. 2019. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 1311–1316.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Advances in Neural Information Processing Systems*.

Liang, S.; Li, Y.; and Srikant, R. 2018. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In *Proceedings of International Conference on Learning Representations*.

Lin, T.-E.; and Xu, H. 2019. Deep Unknown Intent Detection with Margin Loss. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 5491–5496.

Liu, J.; Lin, Z.; Padhy, S.; Tran, D.; Bedrax Weiss, T.; and Lakshminarayanan, B. 2020. Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness. In *Advances in Neural Information Processing Systems*, volume 33, 7498–7512.

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 Pre-training Approach. *arXiv preprint arXiv:1907.11692* .

Malinin, A.; and Gales, M. 2018. Predictive Uncertainty Estimation via Prior Networks. In *Advances in Neural Information Processing Systems*, 7047–7058.

Mandelbaum, A.; and Weinshall, D. 2017. Distance-based Confidence Score for Neural Network Classifiers. *arXiv preprint arXiv:1709.09844* .

Murphy, K. P. 2012. Machine Learning: A Probabilistic Perspective. *The MIT Press* .

Nalisnick, E.; Matsukawa, A.; Teh, Y. W.; Gorur, D.; and Lakshminarayanan, B. 2018. Do Deep Generative Models Know What They Don't Know? In *Proceedings of International Conference on Learning Representations*.

Orhan, A. E. 2019. Robustness Properties of Facebook's ResNeXt WSL models. *arXiv preprint arXiv:1907.07640* .

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1532–1543.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3973–3983.

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 *Advances in Neural Information Processing Systems*, volume 32, 66–77.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv preprint arXiv:1910.01108* .

Schuster, S.; Gupta, S.; Shah, R.; and Lewis, M. 2019. Cross-lingual Transfer Learning for Multilingual Task Oriented Dialog. In *Proceedings of the Conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3795–3805.

Tan, M.; Yu, Y.; Wang, H.; Wang, D.; Potdar, S.; Chang, S.; and Yu, M. 2019. Out-of-Domain Detection for Low-Resource Text Classification Tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3557–3563.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771* .

Yilmaz, E. H.; and Toraman, C. 2020. KLOOS: KL Divergence-based Out-of-Scope Intent Detection in Human-to-Machine Conversations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2105–2108.

Zheng, Y.; Chen, G.; and Huang, M. 2020. Out-of-domain detection for natural language understanding in dialog systems. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, volume 28, 1198–1209. IEEE.