Exploration of Design Choices in Domain Generalization of Zero/Few-Shot Techniques for Out-of-Domain (OOD) Detection in Task-oriented Dialogue Systems

Austin Nguyen[1], Caitlin Lau[1], and Aloysius Lim[1]

[1]Institute for Applied Computational Sciences (IACS), Harvard University
{*austinnguyen,caitlinlau,aloysius_lim*}@*g.harvard.edu*

Abstract

Task-oriented dialogue systems such as Alexa and Siri are often trained on a corpus of data to handle a restricted set of within-domain prompts. However, users engaging with these system can produce out-of-distribution (OOD) prompts. Current methods for OOD detection rely on fine-tuning pretrained models on labeled data. More recent research demonstrates that adaptive representation learning and density estimation — coined REDE — has great potential to discriminate OOD from ID [3]. Our contribution is a better understanding of the assumptions built into REDE approach. We outperform the author's approach by performing a comprehensive study of embedding representations and density estimators.

1 Introduction

Motivation. Task-oriented dialogue systems (i.e. Alexa Assistant, Siri, etc) are often restricted to within-domain prompts and answers. However, users engaging with a dialogue system may make requests that are beyond the scope of the data that is available. Existing methods for tasks rely on fine-tuning a pre-trained model and a corpus of annotated data. Jin, et al (2021)[3] propose a novel method called REDE to apply to zero/few-shot examples that quickly learns a high-performing detector comparable to the full-supervised setting. In this case, zero-shot refers to applying a model to unseen data or few-shot refers to fine-tuning training based on very small samples (i.e. low-resource setting). This was demonstrated on a specific dataset (DSTC9 Track 1) and resulted in the proposal of a new method called representation learning and density estimation (REDE).

Problem Statement. While this method is fast to train, it is not clear to what extent this model could be further improved by investigating the two components of REDE: (1) representation and (2) density estimation. Jin, et al. make particular assumptions around these two components, but we want to better understand how we can further improve this. There has been significant progress in exploring pre-trained S-BERT models. Given this progress, we hypothesize that improvements to current REDE approaches are achievable with newer and/or more relevant embedding representations being used as well as density estimators For example, how does this methodology that purports to perform well while only training on less than 5% of data generalizes well to other domains? What about REDE performing well? For example, how sensitive is REDE to hyperparameter tuning?

Contributions. Our contribution is the exploration of the design choices of embedding representationsm and density estimators that power REDE to more deeply understand how it can better generalize to other domains such as Tripadvisor Data.

2 Related Works

In research of task-oriented dialogue systems, there have been several highly influential papers focused on dialogue systems as optimization problems. Levin, et al. frames the problem space of dialog systems as an optimization problem in which a stochastic model (Markov Decisions and Reinforcement learning) helps to probabilistically estimate the interactions the dialog system will engage in with the user [5]. In our project, we will also use stochastic models via density estimation to analyze task-oriented problems.

Subsequent work by Rajpurkar, et al. work explores how to utilize the strength of crowd working to study the concept of Reading Comprehension (RC) or the ability to read, comprehend and answer questions given text. [7]. This paper compares the reading comprehension of over 100,000 questions and found humans were still able to have much higher reading comprehension than a simple logistic regression (86.8% vs 51%). Utilizing the very large dataset created in this paper, we learned about different types of questions you can ask in a task-orientated dialog system and the different responses can illicit.

Exploring the problem domain from a human-computer interaction (HCI) lens, researchers explore how to produce a natural conversational system in which a human-to-human online conversation data would be the most useful. [1]. The MultiWOZ team used online task workers on the “wizard” who asked questions and a “subject” that answered the questions in a simple manner [2]. These conversations were recorded in a large dataset which laid the foundation for data collection methods and HCI framework in task-based dialogue systems.

Research on task-oriented systems focused on an information retrieval (IR) lens [6]. Researchers re-purposed BERT as a passage re-ranker and developed a large dataset using human crowd workers to complete reading comprehension tasks. Passage ranking especially for task-oriented dialog systems is a critical feature because it increases the comprehension of the problem by the system and permit models to better prioritize key parts of the task.

3 Data

We obtain the data from the unstructured knowledge access in Track 1 of the DSTC9 dataset [4]. The data is sourced from the ”Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access” challenge. This challenging track takes turns between tasks that could be handled by the existing task-oriented conversational models with no extra knowledge and turns that require external knowledge resources to be answered by the dialogue system. The DSTC9 dataset is an augmented version of MultiWoz 2.1 [2] which includes newly introduced knowledge-seeking turns. To augment the data from DSTC9, we evaluated performance on real-world user queries obtained from a manually curated set of questions posed by real users on Tripadvisor (TA) forums, consistent with Jin, et al. [3]

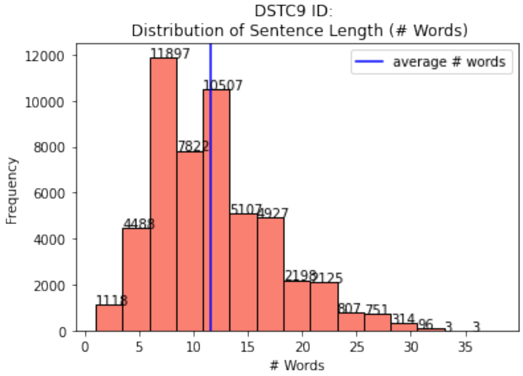

Figure 1: DSTC9: Distribution of Sentence Length for ID

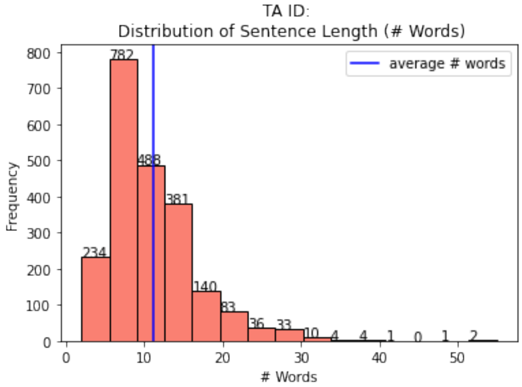

Figure 2: TA: Distribution of Sentence Length for ID

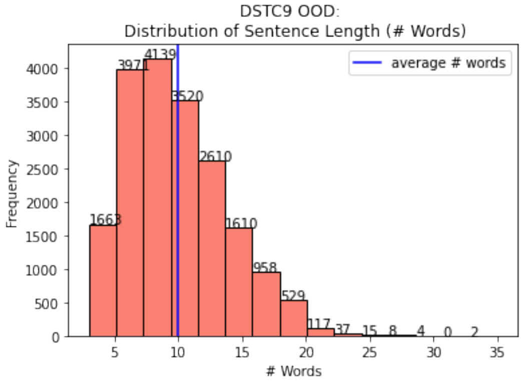

Figure 3: DSTC9: Distribution of Sentence Length for OOD

In the context of task-oriented dialogue systems, in-distribution (ID) examples are considered task-related, whereas out-of-distribution (OOD) examples are considered knowledge-seeking. In the DSTC9 dataset, we have 71346 observations in our dataset: 52163 are considered ID and 19183 are OOD. In addition to the 71k observations, we have 2816 observations sourced from Tripadvisor (TA) reviews that capture knowledge-seeking vs. task-oriented prompts related to the domain of travel. We treat this set of reviews as an external validation set.

We find that the distribution of sentence length is consistent for ID examples between DSTC9 and TA in which the average sentence length was 12

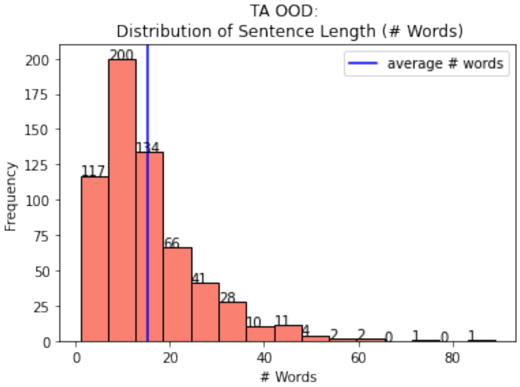

Figure 4: TA: Distribution of Sentence Length for OOD

words. This relationship holds for OOD examples between DSTC9 and TA in which the average sentence length was approximately 11 words. This confirms that distributions are at parity based on the metric of sentence length for ID and OOD examples between the source data (DSTC9) and the target data to which we want to generalize (TA).

Task-related (ID)	**Knowledge-seeking (OOD)**
I'd like to have some Chinese food.	Is a VISA card acceptable for this booking?
I also need a taxi booked	May I cancel my taxi booking later?
I'm looking for hotels in the center.	Does this restaurant have high chairs available for small children?
Yes please, make a reservation for 8 people at 13:00 on friday.	Do they allow BYOB at this location?
Could you try to book us there for 4 nights instead?	Do they have concierge service?

Table 1: Example of task-related (ID) vs. knowledge-seeking (OOD)

4 Methods

We applied term frequency-inverse document frequency (TF-IDF) to the data to produce a vectorized representation of each sentence. As a metric that is intended to reflect how important a word is to a document in a corpus, we use TF-IDF as a weighting factor in language modeling. TF-IDF value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general . We do this for the training, validation, and test data. We also remove stop words in English and specify that terms can be either unigrams or bigrams to preserve a small degree of locality.

With the TF-IDF, we leverage supervised learning approaches such as logistic regression to predict whether a given TF-IDF vectorized representation of a sentence is ID or OOD. The advantages of TF-IDF is that the metric is easy to compute . Furthermore, it is an intuitive heuristic for extracting the most descriptive terms in the document. The disadvantage of TF-IDF is that it is a BoW model that fails to capture word order, semantics, and co-occurrence. We employ ELI5 to help with the interpretability of the weights of our logistic regression applied to high-dimensional and highly sparse TF-IDF transformed sentences. This model serves as the baseline against which we explore more advanced techniques for OOD detection.

To address this limitation, we also explored another technique to produce word embeddings. Namely, we explore adaptive representation learning and density estimation approaches (REDE). In this approach, we learn a representation model via fine-tuning a pre-trained sentence encoder on task-oriented prompts that are considered in-distribution by using a masked language model (MLM). With this representation model, we learn a density estimator from the representations of this model. The density estimator returns a density score for a given prompt. If that prompt exceeds the threshold, it is considered in-distribution; otherwise, it will be considered out-of-distribution.

As the construction of BERT makes it unsuitable for semantic similarity search, we are using pre-trained word embeddings from Sentence-BERT (SBERT)[8]. It is a modification of a pre-trained BERT network that uses siamese and triplet network structures to drive semantically meaningful sentence embeddings which is more suitable to our context. We explored four different types of embedding representations: distllbert-base-nli-stsb-mean-tokens, All-mpnet-base-v2, all-MiniLM-L6-v2, multi-qa-mpnet-base-dot-v1 in order to explore new embedding representations that were not explored in the original research study. Our goal was to see if there was a significant difference in model robustness when using different types of embedding. Some key differences between these embedding models were that all-mpnet-base-v2 was typically shown to be the sentence transformer that provided the best quality on a majority of data sets overall. Furthermore, all-MiniLM-L6-v2 was shown to be 5 times faster than all the other embedding representation models. We want to see if trade off between the quality and time complexity was significant in our design space. Multi-qa-mpnet-base-dot-v1 has been trained on 215M (question, answer) pairs from diverse sources, which is more similar to the context of our data.

Motivated by recent work that demonstrated whitening operations in traditional ML can enhance the isotropy of sentence representations and reduce the dimensionality of sentence representations, we applied a whitening operation [10] to the embedding before similarly feeding it into a density scoring function. The steps of the transformation are detailed below. Note that steps 2 and 3 are only done for out-of-domain sentences.

```
Whitening- $k$ transformation
Input: Existing embeddings $\{x_i\}_{i=1}^{N}$
1: compute $\mu$ and $\Sigma$ of $\{x_i\}_{i=1}^{N}$
2: compute $U, \Lambda, U^T = \mathrm{SVD}(\Sigma)$
3: compute $W = \left(U\sqrt{\Lambda^{-1}}\right)[:, :k]$
4: for $i = 1, 2, \cdots, N$ do
5:     $\widetilde{x}_i = (x_i - \mu)\, W$
6: end for
Output: Transformed embeddings $\{\widetilde{x}_i\}_{i=1}^{N}$
```

We then feed the transformed word embedding into density estimation functions such as the Gaussian Mixture model (GMM) to derive the densities. With the densities, we use logistic regression to automatically set a threshold to differentiate between ID and OOD sentences.

4.1 Metrics

Since the task is OOD classification, the metrics of choice would be precision, recall, and F1-measure for the OOD class. We have focused on using weighted F1 score to take into account both precision and recall, and also allow greater contribution from class with more examples in the dataset.

5 Results & Discussion

5.1 TF-IDF approach

With TF-IDF as our baseline, we produced TF-IDF vectorized representations of the sentences and fit a logistic regression to the TF-IDF representations of the sentences on the training set. In the test set, we observed a 0.99 precision, 0.99 recall, and 1.00 accuracy.

Metric	DSTC9	TA
Precision (OOD)	0.99	0.95
Recall (OOD)	0.99	0.52
F1 (Weighted)	0.99	0.88
Accuracy	1.00	0.89

Table 2: Test metrics for DSTC9 vs. Tripadvisor reviews (TA)

However, we produced TF-IDF vectorized representations of TA reviews and obtain the following results: 0.95 for precision, 0.52 for recall, 0.88 for weighted F1 and 0.89 for accuracy. We observe significant degradation in the performance of our task of OOD detection. This suggests that the TF-IDF vectorized representations of sentences are limited in that this approach does not handle domain shift well.

5.2 TF-IDF: Interpretability

Top +	Feature	Top -	Feature
+16.023	station	-14.012	thank
+15.053	change	-13.947	looking
+14.692	wheelchair	-12.270	free parking
+14.030	bike	-11.596	cambridge
+13.561	cancel	-11.023	address

Table 3: Results from eli5: Top 5 weights related to OOD (left); bottom 5 weights related to OOD (right)

When we use eli5 to help with model explainability, we find that words that are associated with OOD (i.e. not related to a task) tend to be amenity-oriented words (i.e. ”wheelchair”, ”parking options”, ”onsite”) whereas words related to address/geography or calls to action (i.e. ”address”, ”looking”, etc) are ID (task-oriented). These results illustrate the trade-off of TF-IDF representations: while there is a sense degree explainability for unigram and bigram terms, semantic meaning is not captured in a BoW-oriented represenation of the words.

5.3 REDE approach

We fed our data through S-BERT to produce word embeddings and presented some examples of similar sentences below. The results illustrate that the embeddings are able to capture sentence similarities reasonably well.

Using these sentence embedding without whitening transformation, we derive the densities of the in-domain by running a BGM model — variational Bayesian estimation of a Gaussian mixture

Pair 1	Pair 2	Cosine Sim
No, that would be all. Thanks. Bye.	That was all I needed. Thanks. Bye.	0.8861
Thats everything I needed thanks for the help!	That was all I needed. Thanks. Bye.	0.9377
Thank you, that will be all. good bye.	Thank you, that's all I need today.	0.8763
Thank you, that's all I need today.	Thank you so much for all your help!	0.8490
Thank you so much for all your help!	That was all I needed. Thanks. Bye.	0.8389

Table 4: Results of cosine similarity as a sanity check for S-BERT

or Bayesian Gaussian Mixture. The BGM model returns a density estimate or an approximate posterior distribution over the parameters of a Gaussian mixture distribution. We use logistic regression to automatically set a threshold, where we observed the test results in table 4.

Metric	DSTC9	TA
Precision (OOD)	0.87	0.94
Recall (OOD)	0.76	0.55
F1 (weighted)	0.87	0.81
Accuracy	0.89	0.82

Table 5: REDE without whitening transformation: Test metrics for DSTC9 vs. TA reviews

Metric	DSTC9	TA
Precision (OOD)	0.99	0.74
Recall (OOD)	1.00	0.96
F1 (Weighted)	1.00	0.94
Accuracy	1.00	0.94

Table 6: REDE with whitening transformation: Test metrics for DSTC9 vs. TA reviews

We repeated the same steps, but used the whitening transformation applied to our word embedding. We observe a general improvement in the performance on TA reviews. For example, the REDE approach yields a higher F1 score (0.94) compared to TF-IDF (0.88). One limitation of this approach is that it loses interpretability when we adopted the pre-trained word embedding, which is unlike in our TF-IDF approach where we could highlight words that are associated with OOD.

5.4 Exploration of design space on REDE approach

Our results from experimentation in Table 7 with different pre-trained word embedding shows that performance based on weighted F1 is best on [multi-qa-mpnet-base-dot-v1]. One takeaway is that this embedding has been trained on QA-specific prompts and matches the domain of task-oriented dialogue systems, which is likely the reason for better performance. Another finding is that a smaller general purpose model [all-MiniLM-L6-v2] performs better than the bigger general purpose model [all-mpnet-base-v2].

We further explored using different covariance structures in our best-performing density estimator - Bayesian Gaussian Mixture (BGM) and Gaussian Mixture models (GMM). Having a more complex covariance structure results in more parameters which tends to fit better but not generalize well as shown in Table 8. This is evident in the better performance on the test data of the DCTC9 as covariance structure becomes more complex (i.e. from ”Spherical” to ”Full”) for both BGM and GMM. The same trend is however not observed in the TA dataset although the performance is the still the best for ”Full” covariance structure.

From our exploration of using different number of components in both BGM and GMM, we observed that performance is insensitive to our assumptions on the number of components as shown in Table 9. This is likely due to the whitening transformation which has pushed the ID features far away from the OOD features to the extent that the number of components do not matter. One advantage of BGM is that the number of components can be inferred from the data, so it will outperform GMM if the number of components specified in GMM is not appropriate. As the performance is insensitive to the number of components, BGM and GMM performs similarly well in our explorations.

6 Key Takeaways

TF-IDF is a reasonable baseline, but it is limited because it is a BoW model which does not take into account semantic meaning or order. Our results have shown that it also does not generalize well on the TA dataset. It however allows for better interpretability as demonstrated in how we have used eli5 to make sense of its results.

In the REDE approach, whitening used in preprocessing is found to have considerably improved

Embedding Representation	**n_components**	**Covariance Type**	**F1 (weighted)**	**F1 (ID / OOD)**
distllbert-base-nli-stsb-mean-tokens	5	Full	0.91	0.74 / 0.94
all-mpnet-base-v2	5	Full	0.92	0.78 / 0.95
all-MiniLM-L6-v2	3	Full	0.93	0.82 / 0.96
multi-qa-mpnet-base-dot-v1	4	Full	0.93	0.82 / 0.96

Table 7: Results: Exploration of pre-trained embedding representations

Density estimator:	BGM		GMM	
Test data / Covariance Structure	DCTC9	TA	DCTC9	A
Full	1.00	0.94	1.00	0.93
Tied	1.00	0.93	1.00	0.93
Diag	0.98	0.87	0.99	0.88
Spherical	0.97	0.91	0.98	0.92

Table 8: Results: Exploration of covariance structures in density estimators

Density estimator:	BGM		GMM	
Test data / no. of components	DCTC9	TA	DCTC9	A
1	1.00	0.93	1.00	0.93
4	1.00	0.94	1.00	0.94
8	1.00	0.93	0.99	0.89
12	0.99	0.93	0.99	0.93

Table 9: Results: Exploration of number of components in density estimators

performance. Our choice of GMM or BGM density estimators does not impact the performance. Performance is also insensitive to the number of components assumed in our density estimators. One hypothesis is that the whitening has separated the ID and OOD features far enough so that the performance is insensitive to these assumptions. We however found that a more complex covariance structure for GMM and BGM still results in the best performance for the TA dataset despite the risk of over-fitting. Another takeaway on using pre-trained word embedding is that using embedding trained on QA-specific prompts [multi-qa-mpnet-base-dot-v1] that match the domain of task-oriented dialogue systems outperforms representation using other pre-trained word embeddings that we have tried. Among those that we have tried, a smaller representation [all-MiniLM-L6-v2 w] outperformed MPNet [all-mpnet-base-v2] on this specific task of OOD detection on a new data set, demonstrating that performance using a word embedding from a smaller model can outperform one from a bigger model.

7 Future Work

Contrastive learning, specifically contrastive shifted instances (CSI), has been shown to be promising for OOD detection. Previous literature in OOD detection has focused on density-based, reconstruction-based, one-class classification, and self-supervised approaches to model representations that encode normality and define a detection score (Yang et al., 2021; Ruff et al., 2021) [12] [9]. Recent work develops a new approach by taking ideas from computer vision and audio processing — contrastive learning — to extract strong inductive bias from multiple views of a sample by letting them attract each other and repelling them from other samples (Tack et al., 2020) [11].

Tack et al. propose a simple method called contrasting shifted instances (CSI) specifically for OOD detection. This unique contribution to representation learning helps to learn a more discriminative representation for detecting OODs and designing a score function that utilizes this new representation form. Future work could explore contrastive learning to produce learned representations that better characterize OOD detection for task-oriented dialogue systems.

8 Impact Statement

One adverse impact will be to use our model under the wrong context. As the model has been trained on data using task-oriented dialogue systems for the travel industry, we would caution against making any domain generalization different task-oriented dialogue with applications for the medical domain and autonomous driving.

For example, in the medical domain, an improper use of this system would be to try and apply our system to a digital medical diagnosis system. If our model is trained on question-answer data related to context-specific tasks such as ordering food or hotel booking, application of this approach to medical data could have adverse outcomes, such as flagging OOD questions as ID and providing incorrect advice or providing an inappropriate task in response to a life-threatening task such as calling for an ambulance.

Another adverse impact could be using wrongly in the autonomous driving context. For example, if task-oriented dialogue systems flag certain user prompts as ID and dismisses the prompt in the case of autonomous driving, this could lead not only lead to a mismatch in user expectation with the autonomous vehicle but also can lead to unintended consequences such as the task-oriented dialogue system ignoring safety-preserving prompts by the user.

Either way, when there are human lives at risk in highly critical scenarios, our system should not be used without a thorough understanding of its limitations and assumptions. The key limitation to our work is that the embedding and data used for domain generalization must be context specific to the downstream application. If we use an embedding consistent with our domain, we achieve better results. Therefore, if one were to explore our approach for other task-oriented dialogue systems, we would recommend that machine learning practitioners use an embedding and data that is specific to that downstream task.

References

[1] Pawel Budzianowski et al. "MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling." In: *CoRR* abs/1810.00278 (2018). arXiv: 1810.00278. URL: http://arxiv.org/abs/1810.00278.

[2] Mihail Eric et al. "MultiWOZ 2.1: Multi-Domain Dialogue State Corrections and State Tracking Baselines." In: *arXiv preprint arXiv:1907.01669* (2019).

[3] Di Jin et al. *Towards Zero and Few-shot Knowledge-seeking Turn Detection in Task-orientated Dialogue Systems*. 2021. DOI: 10.48550/ARXIV.2109.08820. URL: https://arxiv.org/abs/2109.08820.

[4] Seokhwan Kim et al. "Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access." In: *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 1st virtual meeting: Association for Computational Linguistics, July 2020, pp. 278–289. URL: https://aclanthology.org/2020.sigdial-1.35.

[5] Esther Levin, Roberto Pieraccini, and Wieland Eckert. "A Stochastic Model of Human-Machine Interaction for Learning Dialog Strategies." In: *Speech and Audio Processing, IEEE Transactions on* 8 (Feb. 2000), pp. 11–23. DOI: 10.1109/89.817450.

[6] Rodrigo Frassetto Nogueira and Kyunghyun Cho. "Passage Re-ranking with BERT." In: *CoRR* abs/1901.04085 (2019). arXiv: 1901.04085. URL: http://arxiv.org/abs/1901.04085.

[7] Pranav Rajpurkar et al. "SQuAD: 100, 000+ Questions for Machine Comprehension of Text." In: *CoRR* abs/1606.05250 (2016). arXiv: 1606.05250. URL: http://arxiv.org/abs/1606.05250.

[8] Nils Reimers and Iryna Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. 2019. DOI: 10.48550/ARXIV.1908.10084. URL: https://arxiv.org/abs/1908.10084.

[9] Lukas Ruff et al. "A unifying review of deep and shallow anomaly detection." In: *Proceedings of the IEEE* 109.5 (2021), pp. 756–795.

[10] Jianlin Su et al. *Whitening Sentence Representations for Better Semantics and Faster Retrieval*. 2021. DOI: 10.48550/ARXIV.2103.15316. URL: https://arxiv.org/abs/2103.15316.

[11] Jihoon Tack et al. "CSI: Novelty detection via contrastive learning on distributionally shifted instances." In: *Advances in neural information processing systems* 33 (2020), pp. 11839–11852.

[12] Jingkang Yang et al. “Generalized out-of-distribution detection: A survey.” In: *arXiv preprint arXiv:2110.11334* (2021).