Out of Scope Detection Using a Corpus of Examples

Simon Ellershaw
BSc MSc MRes
simje@live.co.uk

Abstract

Out-of-scope detection is the task of identifying inputs at inference time that have labels not seen at training time. We present a new distance based approach to this problem building on the simplex framework, in which an inputs final latent representation is approximated as the linear sum of corpus examples collected at training time. The euclidean distance between this approximation and the real latent state is defined as the corpus residual. We then classify a sample as out-of-scope by thresholding it's corpus residual in relation to the 95% percentile of the validation set's residual. This method achieves a 0.929 AUC on the CLINIC150 out-of-scope task outperforming the current distance based method of kNN. Further work is required to tune the method's hyperparameters, compare it to the method non-distance based approaches and evaluate the method on noisier real world data.

1 Introduction

Out-of-scope (OOS) or out-of-distribution detection refers to a group of methods which aim to identify inputs for which a model has not been trained on the samples underlying class. A trivial example of an OOS input is passing a picture of a fish to a dog vs cat classifier. The current generation of machine learning models would continue to classify such an input and not report the deviation in the input data. This becomes problematic in safety critical applications such as medical image classification for which diseases could potentially be misdiagnosed. For example, if a glaucoma image is passed to a diabetic retinopathy classifier the outcome is uncertain and most certainly erroneous.

Classifying OOS examples is difficult as the distribution of the OOS class is unknown at training time. Furthermore, to have utility an OOS method must generalise over a wide range of models and not effect their performance on the primary task. Such constraints make the naive approach of the addition of an 'other' category infeasible as exposing a model to every 'other' possible case is infeasible and would effect the ability of the model to learn the features of the actual task. A number of solutions to this problem have been put forward including gradient based methods [1], bayesian models [2], density based methods [3] and distance based models [4].

In this paper we propose a new method extending the recently proposed simplex method [5] to perform OOS detection. This is a novel type of distance based model in two regards. Firstly it proposes to measure the distances between latent states instead of inputs. Secondly, instead of using kNNs we use the simplex idea of corpus residuals.

2 Methodology

2.1 Problem Definition

We define the feature, $\mathcal{X} \subseteq \mathbb{R}^{d_\mathcal{X}}$, and label, $\mathcal{Y} \subseteq \mathbb{R}^{d_\mathcal{Y}}$, spaces to have dimensions $d_\mathcal{X}$ and $d_\mathcal{Y}$ respectively. Our task is to predict if an unseen sample input's, $\boldsymbol{x_i}$, label, y_i, is a member of the set of labels seen at training time $\mathcal{Y}_t$. If $y_i \notin \mathcal{Y}_t$ it is referred to as OOS.

This paper is restricted to considering models, $\boldsymbol{f} : \mathcal{X} \rightarrow \mathcal{Y}$, that can be decomposed as $\boldsymbol{f} = \boldsymbol{l} \circ \boldsymbol{g}$. Where $g : \mathcal{X} \rightarrow \mathcal{H}$ is a function that maps an input $\boldsymbol{x} \in \mathcal{X}$ to a latent vector $\boldsymbol{h} = \boldsymbol{g}(\boldsymbol{x}) \in \mathcal{H}$. $\boldsymbol{l}$ is then a linear function that maps $\boldsymbol{h}$ to the output $\boldsymbol{y}$ such that $\boldsymbol{y} = \boldsymbol{l}(\boldsymbol{h}) = \boldsymbol{A}\boldsymbol{h} \in \mathcal{Y}$. The space $\mathcal{H} \subseteq \mathbb{R}^{d_{\mathcal{H}}}$ will be referred to as the latent space and typically is of higher dimension than the label space, $d_{\mathcal{H}} > d_{\mathcal{Y}}$.

This assumption is not restrictive to the logit output of most modern machine learning classification methods including multi-layer perceptrons (MLPs) [6], convolutional neural networks (CNNs) [7] and transformers [8].

2.2 Simplex

The simplex approach [5] was first proposed as a method for post-hoc black box model explainability under the framework outlined in Section 2.1. It was also shown that the method was able to detect shifts in input data distributions. In this paper, we extend simplex to propose a novel method for OOS detection.

Simplex collects a corpus of N random training examples, $\mathcal{C}$. A given input's latent state, $\boldsymbol{h}$, is then approximated, $\hat{\boldsymbol{h}}$, as the weighted sum of latent states of the members of $\mathcal{C}$,

$$\hat{\boldsymbol{h}} = \sum_{i=1}^{N} w^i \boldsymbol{g}(\boldsymbol{x^i}) | \boldsymbol{x^i} \in \mathcal{C}. \tag{1}$$

The weights, $\boldsymbol{w}$, are found by minimising the euclidean distance between the actual and approximate hidden state called the corpus residuals, $r_{\mathcal{C}}(\boldsymbol{h})$,

$$r_{\mathcal{C}}(\boldsymbol{h}) = ||\boldsymbol{h} - \hat{\boldsymbol{h}}||_2 = \sqrt{\sum_{i=1}^{d_{\mathcal{H}}} (\boldsymbol{h_i} - \hat{\boldsymbol{h}_i})^2}, \tag{2}$$

by convex optimisation,

$$\boldsymbol{w} = \min_{\boldsymbol{w} \in \mathbb{R}^{\boldsymbol{d_N}}} r_{\mathcal{C}}(\boldsymbol{h}) \tag{3}$$

from which the K largest magnitude weights are kept.

The original motivation for this method was that given the constraint of the model's final layer, $\boldsymbol{l}$, being linear, if $\hat{\boldsymbol{h}} \approx \boldsymbol{h}$ the decomposition also approximates the model output $\hat{\boldsymbol{y}}$ when $\boldsymbol{l}$ is applied,

$$\boldsymbol{l}(\hat{\boldsymbol{h}}) = \boldsymbol{l}(\sum_{i=1}^{N} w^i \boldsymbol{g}(\boldsymbol{x^i})) = \sum_{i=1}^{N} w^i \boldsymbol{l}(\boldsymbol{g}(\boldsymbol{x^i})) = \sum_{i=1}^{N} w^i \boldsymbol{f}(\boldsymbol{x^i}) \approx \hat{\boldsymbol{y}}. \tag{4}$$

This allows weights, $\boldsymbol{w}$, to be used to explain the relative influence of training examples on the model's output. Note that without this linearity constraint this approximation does not hold. For example if the same decomposition is performed directly on the inputs, $\boldsymbol{x}$.

An additional experiment was also run to analyse the change in the magnitude of corpus residuals when the underlying data distribution between the training and test set changes. It was found that the magnitude of the residuals increases when the distribution shifts. However, a limitation of this experiment was that there was considerable overlap between the two distributions and so classification into classes according to their data origin was not always possible or optimal.

2.2.1 OOS Detection

This paper extends the simplex approach by analysing the change in corpus residuals when OOS data is inputted to a trained model. This has the advantage over previous studies of have clearly defined non-overlapping classes.

OOS detection is performed after the standard training [9] of a deep learning model, $\boldsymbol{f}$, on randomly selected training, X_t, and validation, X_v, datasets. The corpus residuals of each member of the validation set are calculated with respect to $N = 1000$ randomly selected samples from the training set. The original simplex fitting methodology [5] is followed with optimisation taking place over

10000 epochs with an initial regulation factor of 0.1 which is varied according to an exponential scheduler to a final value of 100 and $K = 10$

The discrete probability distribution of the validation's set corpus residuals is then found. A test sample is then classified as OOS if it's corpus residual is greater than that of 95% of the validation set's.

2.3 Experiments

To test the behaviour of the OOS extension to the simplex method, it has been applied to a transformer model trained on the natural language classification problem CLINIC150 [10]. This dataset contains 22,500 labelled sentences of 150 distinct classes and 1200 OOS examples. In the case of a transformer classifier the final latent state is taken as the state of the [CLS] token in the final hidden layer, as shown in Figure 1.

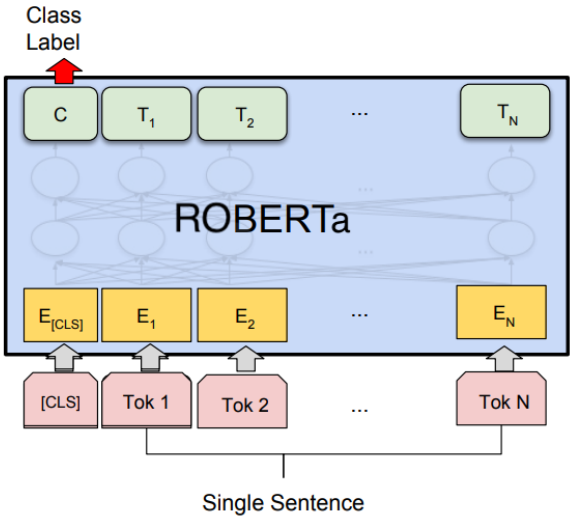

Figure 1: Visualisation of the Roberta transformer model [11]. Note token C, the final state of the [CLS] token, is taken as the latent state in all experiments in Section 2.3

The ability of the simplex method to detect OOS examples is compared to two baseline methods which also have 95% percentile thresholding applied. These are the corpus residuals resulting from approximation of the hidden latent state using weighted k nearest neighbours (kNN) [12] and the probability values of the most likely classes predicted by the model.

2.3.1 Transformer Training

A transformer model has been finetuned on the CLINIC150 dataset from the pretrained weights of the Roberta-Base model [13]. Input sentences are tokenised by the Roberta tokeniser and truncated to a maximum length of 512 tokens. Although due to being sentence inputs this limit is never reached. Tokenised inputs are then batched in groups of 32 samples and their lengths are dynamically padded to the longest sample their respective batch. It is important to note that only in-scope data is used in the training of the transformer.

The transformer's weights are then finetuned by backpropogation of the cross-entropy loss between the labels and model predictions for each batch for a total of 10 epochs. The training procedure uses an AdamW optimizer with an initial learning rate of 5×10^{-5} which decays with a linear schedule after a 10% warm-up period. Weight decay of 0.01 and gradient normal clipping of 1 are also used. At the end of training, the model weights at the epoch with the smallest validation loss are reloaded. This prevents over-fitting to the training set.

This results in a model which achieves the following macro classification metrics; precision=0.965, recall=0.967 and F1=0.965. Extensive hyperparameter tuning of the training procedure has not

been performed as this is not the objective of this paper and model performance is sufficient for the purposes of the OOS experiments.

3 Results

This section describes the results of experiments undertaken to evaluate the ability of simplex to identify OOS samples. This is done by comparing it's behaviour to baseline methods when applied to the finetuned transformer for inputs of the CLINIC150 OOD and test set. The test set is referred to as the in-scope dataset for the purposes of this results section.

The results presented in this section are fully reproducible using the code at `https://github.com/simonEllershaw/latent_variable_OOD`

3.1 Corpus residuals

A visualisation of the distribution of the corpus residuals for different datasets and methods is shown in Figure 2. As the maximum probability method has no concept of corpus residual it is not shown. When compared to the validation set it can be seen that the in-scope distributions are highly similar and the OOS distributions differs for both methods. Applying the two sample Kolmogorov-Smirnov test [14] to these comparisons confirms that these observations are statistically significant (p-values ≈ 1.0 and ≈ 0.0 respectively).

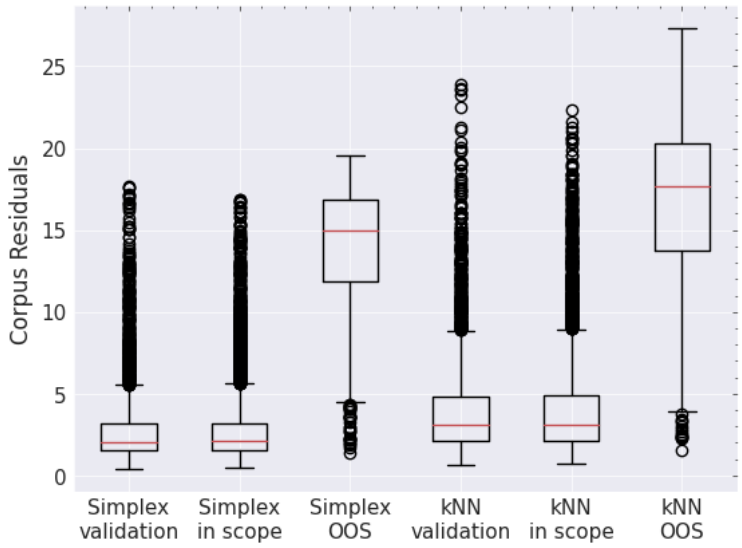

Figure 2: Plot of the distribution of corpus residuals for validation, in-scope and OOS datasets when latent states are estimated by simplex and distance weighted kNN methods.

This results proves the hypothesis that latent states of OOS have corpus residuals that are significantly different to those of in-scope examples. However, from this analysis it is not clear if simplex is preferable method.

3.2 95% Percentile Thresholding

Table 1: OOS detection metrics for simplex and baseline methods on the CLINIC150 dataset

Method	Precision	Recall	F1	AUC
Max Probability	0.803	0.826	0.814	0.886
kNN	0.816	0.855	0.835	0.902
Simplex	0.814	0.892	0.851	0.929

Classifying samples as OOS or in-scope based on 95% percentile thresholding has been performed for OOS and in-scope datasets. The performance of the simplex, kNN and maximum probability methods have been assessed by the precision, recall, F1 and AUC metrics and is shown in Table 1. It can be seen that simplex outperforms the two baselines, especially in terms of recall.

A visual explanation of this results can be seen in Figure 3. This plot visualises the distribution of probabilities that a sample drawn from the validation set has a latent state, $\boldsymbol{h_v} \in \boldsymbol{g(X_v)}$, with a corpus residual greater or equal to that of a given test sample's, $\boldsymbol{h_s}$, $P(r_{\mathcal{C}}(\boldsymbol{h_t}) \geq r_{\mathcal{C}}(\boldsymbol{h_v})$. For all methods the in-scope distributions are well behaved with means and interquartile ranges at approximately $0.5, 0.25, 0.75$ respectively.

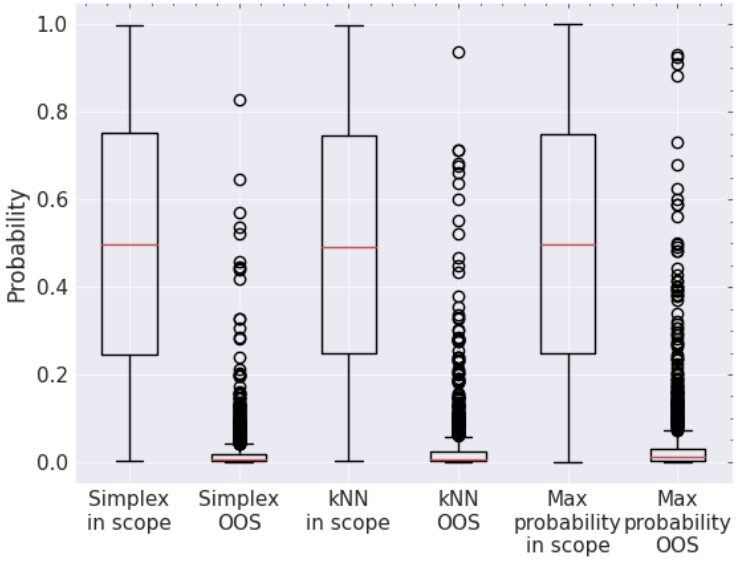

Figure 3: Plot of the distribution of probabilities that a sample drawn from the validation set has a latent state with a corpus residual greater or equal to that of a given test sample's

Furthermore, a distinct shift of the OOS samples towards zero for all methods can be seen. There is also separation between the majority of OOD and in-scope samples. However, the extent of the outlier tails with higher probabilities differs between methods, with the simplex methods having the shortest tail. This is an important observation as OOS with high probability will be misclassified as in-scope.

3.3 Model Mistake Classification

A disadvantage of the 95% percentile thresholding is that by definition 5% of in-scope samples will be classified as OOS. However, further exploration of the set of in-scope samples misclassified as OOS showed that the error-rate of their CLINIC150 label was 63.5% compared to the 3.5% error-rate in the dataset as a whole. The χ^2 test [14] of independence shows this difference to be statistically significant. This result suggests that corpus residuals are able to quantify not only shifts in the input data but also confidence in the model's output.

4 Discussion and Future Work

The results presented in Section 3 represent a proof of concept that corpus residuals resulting from the use of the simplex methodology can identify OOS inputs. However, further work is required to fully understand the behaviour of this method.

4.1 Simplex hyperparameter tuning

Although, simplex has few hyperparameters the tuning of these has not been explored. Firstly, the of the number of corpus examples, N, is a trade-off between computation time and potential

accuracy of the latent state approximation, $\hat{\boldsymbol{h}}$. How this parameter is efficiently and optimally set, it's data dependence and effect not only the accuracy of approximation but OOS detection is currently unknown.

Secondly, the number of weights kept in the approximation, K, requires further investigation. In the original work [5], it was shown that the accuracy of the approximation saturates at $K \approx 10$. Hence, the values choice in this paper. However, there is evidence that this number is data dependent and similarly to N it's effect on OOS detection is unknown.

Finally, the percentile threshold value of 95% has been chosen somewhat arbitrarily. Although it is inline with many other statistical tests which take this 95% threshold to show significance [14] it could potentially be tuned to optimise performance. The methodology to tune this hyper-parameter is non-trivial though as a base assumption and strength of the simplex approach is that it is not trained on OOS examples. Furthermore, a secondary in-scope validation set would be required for this operation as the threshold is set in reference to the original validation dataset.

4.2 Further Experiments

The simplex OOS method has been shown to outperform baseline kNN and maximum probability methods. However, these both also use the same percentile thresholding technique. Comparison to the current state of the art OOS methods using other approaches would benchmark the currently promising performance of the simplex method.

The results in Section 3.3 tentatively show that the corpus residual approach could be generalised to quantify a model's confidence in it's output. Currently maximum probability is commonly used if the confidence in a model is required to be known. However, in the OOS results presented in this paper we have shown simplex to outperform this approach. Reporting when a model is not confident in a prediction is essential in the safety critical application of machine learning models such as the detection of pedestrians crossing the road by autonomous cars to the prediction of prostate cancer by clinical risk models. However, due to a lack of simple yet accurate solution many applications are deployed without this behaviour. The simplex methodology could be a potential solution.

Finally, the CLINIC150 dataset represents a simple task for modern transformers. This is shown in the strong model performance on the classification task. Therefore, the set of experiments outlined in this paper also need to be repeated on a more challenging noisy real world task. Medical image classification tasks would be a strong candidate. For example, testing if the simplex methodology could flag images of glaucoma as OOS when applied to a CNN trained to classify cases of diabetic retinopathy from fundus images. If successful this would show the high utility of the simplex model and be potentially impactful in the future design of human in the loop automated medical imaging systems.

5 Conclusions

In conclusion, we have shown that there is a statistically significant difference between the distribution of the corpus residuals of OOS and in-scope data. Therefore, through the calculation of simplex corpus residuals at inference time, OOS inputs can be detected when thresholded with respect to the validation set's distribution of corpus residuals. This method requires no separate model or training on OOS data. The simplex method achieves an AUC score of 0.929 on the CLINIC150 dataset outperforming the currently used OOS distance method of kNNs. Further work is required to tune the methods hyperparameters, compare the simplex method to the state of the art non-distance based metrics and analyse it's performance on a noisier real world dataset. Successful results from the later work, in domains such as healthcare, would be impactful in the future design and monitoring safety critical systems.

References

[1] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[2] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[3] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018.

[4] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[5] Jonathan Crabbé, Zhaozhi Qian, Fergus Imrie, and Mihaela van der Schaar. Explaining latent representations with a corpus of examples. *Advances in Neural Information Processing Systems*, 34, 2021.

[6] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[7] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[9] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[10] Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China, November 2019. Association for Computational Linguistics.

[11] Sentiment classification using bert. `https://www.geeksforgeeks.org/sentiment-classification-using-bert/`. Accessed: 21-02-2021.

[12] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[14] Ifan Hughes and Thomas Hase. *Measurements and their uncertainties: a practical guide to modern error analysis*. OUP Oxford, 2010.