Understanding the Latent Spaces of OOD Detectors: A Study of Mahalanobis and IRW Approaches

Demoor Louise*
ENSAE
louise.demoor@ensae.fr

Maurel Benjamin*
ENSAE
benjamin.maurel@ensae.fr

Abstract

The objective of this project is to address concerns regarding the reliability of large neural networks in the field of Natural Language Processing (NLP), despite their impressive performance in recent years. The primary goal is to develop more resilient algorithms that can enhance the ability of NLP systems to resist data drifts and adversary attacks. While cutting-edge models can perform exceptionally well on input data that is similar to their training datasets, they can become ineffective in NLP scenarios due to the continuously changing nature of languages and distributional shifts. To tackle this challenge, the project proposes a methodology for measuring and identifying distributional shifts in different corpus/sentences by analyzing the latent representations of tokens. This analysis can be carried out using classical discrepancy measurement tools, which are tailored to the high-dimensional nature of transformers layers. This research is crucial for promoting the responsible application of promising NLP methods in critical systems, where robustness is a crucial consideration. In this project, the focus is on exploring the usefulness of incorporating information from all the layers to improve Out of Distribution detectors.

All our experiments and figures can be reproduced thanks to our code provided in our GitHub [1]

1 Introduction

The rise of large language models has revolutionized the field of Natural Language Processing (NLP) in recent years, enabling breakthroughs in tasks such as machine translation, sentiment analysis, and question answering. However, concerns about the ethical implications of these models have also grown in parallel with their success. Specifically, there are increasing worries about the potential for large language models to perpetuate biases and discrimination (Colombo et al., 2021, 2022b), leading to calls for greater attention to fairness in NLP. In addition, the threat of adversarial attacks, which aim to manipulate the output of machine learning models, has become more pronounced in recent times. This is especially concerning given the potential consequences of attacks (Picot et al., 2023a,b) on NLP models, such as the spread of misinformation or even the manipulation of democratic processes. To combat these challenges, researchers have developed various methods for detecting and mitigating adversarial attacks and Out-of-Distribution (OOD) samples in large language models (Darrin et al., 2023a,b; Gomes et al.).

These approaches aim to improve the robustness and fairness of these models and are becoming increasingly important in promoting the responsible and ethical use of NLP technology. This research paper aims to provide an overview of the latest developments in fairness, adversarial attack detection, and OOD detection for large language models, with the goal of identifying promising avenues for future research in this critical area (Colombo, 2021).

In this paper, we choose to work on Out-of-Distribution (OOD) detection which is essential because it is a critical challenge that large language models face. OOD samples are input data that are significantly different from the training data used to develop a model. These samples can cause NLP models make incorrect predictions, leading to unreliable or even harmful outputs. Overall, OOD detectors are crucial to developing more reliable and trustworthy NLP mod-

[1] https://github.com/BenJMaurel/NLP_project

els, ensuring their suitability for real-world applications and promoting the responsible adoption of NLP technology.

1.1 Problem Framing

In this section, we formalize the problem of out-of-distribution (OOD) detection in natural language processing (NLP). Let S_{train} denote a training dataset, consisting of n samples, where each sample is represented as a tuple (x_i, y_i), where x_i is an input sentence and y_i is its corresponding label.

The goal of OOD detection in NLP is to identify whether a new input sentence x_{new} is in-distribution regarding the training data (ID) or represents a novel or OOD sample. To accomplish this, we assume the availability of a separate validation dataset $S_{train} = (x_j, y_j)_{j=1,m}$ and a test dataset $S_{test} = (x_k, y_k)_{k=1,l}$, both drawn from different distributions $P_{train}(x, y)$ and $P_{test}(x, y)$, respectively.

Mathematically, we can represent the OOD detection problem as a binary classification task, where the input is a sentence x_i and the output is a label $y_i \in \{0, 1\}$, where $y_i = 0$ if x_i is an ID sample and $y_i = 1$ if x_i is an OOD sample.

In the context of OOD detection in NLP, two metrics are commonly used for evaluating the performance of a model: FPR and AUROC.

FPR (False Positive Rate at 95%) is a metric that measures the rate of false positives (FP) at a fixed true negative rate (TNR) of 95%. In other words, it measures the percentage of ID samples that are incorrectly classified as OOD samples. A lower FPR indicates better performance, as it means that the model is correctly identifying a higher proportion of ID samples.

AUROC (Area Under the Receiver Operating Characteristic Curve) is a metric that measures the overall performance of a binary classifier. It plots the true positive rate (TPR) against the false positive rate (FPR) at different classification thresholds and calculates the area under the resulting curve. In OOD detection, AUROC measures how well the model can distinguish between ID and OOD samples[2].

Since a high AUROC score does not necessarily mean that the classifier has a low FPR it is crucial to consider both the AUROC and FPR when designing an OOD detector.

In order to identify whether a given input belongs to the in-distribution or out-of-distribution (OOD) category we follow (Colombo et al., 2022a) and we rely on two models that incorporate the concept of data depth[3]. Specifically, the depth score of a given input is compared to the depth scores of samples in the training distribution to determine whether the input is in-distribution or OOD. The two models we employ are the Integrated Rank-Weighted Depth model (Ramsay et al., 2019) and the Mahalanobis-based score model (Mahalanobis, 1936). Both of them are measures of the distance between a point and a distribution.

Let X be a random variable and P_X the law of X. The IRW depth of $x \in \mathbb{R}^d$ w.r.t. to a probability distribution P_X is $D_{IRW}(x, P_X) = \int_{S^{d-1}} min\{F_u(\langle\, u, x\rangle), 1 - F_u(\langle\, u, x\rangle)\}\, du$ with $F_u(l) = P_X(\langle\, u, X\rangle \leq l)$ and S^{d-1} is the unit sphere.

The Mahalanobis-based score model uses the Mahalanobis distance that can be seen as a data depth function (Liu et al., 1999). The Mahalanobis depth is $D_M(x, P_X) = (1+(x-\mathbb{E}(X))^T\Sigma^{-1}(x-\mathbb{E}(X)))^{-1}$ where Σ^{-1} is the precision matrix of X.

2 Experiments Protocol

2.1 Context

Traditionally, OOD detectors are based on the output of the last layer of the neural network. However, recent research has shown that using all layers in the network can improve the performance of OOD detectors. The Avg-Avg (Chen et al., 2022) and TRUSTED (Colombo et al., 2022a) detectors are two examples of OOD detectors that use all layers of the network and achieve state-of-the-art performances. Both of them aggregate the information throughout the layers in the most simple way: they create a new embedding which is the mean of the embeddings over all the layers.

The goal of this project is to further investigate the advantages of using all layers of the network for OOD detection. The project aims to understand why taking into account all intermediate layers can be beneficial and in which cases.

[2]An AUROC score of 1 indicates perfect classification performance, while a score of 0.5 indicates random guessing.

[3]Data depth is a measure of how deep a data point is in a dataset, or how central it is relative to the other data points.

2.2 Dataset and model selection

When evaluating a method for detecting out-of-distribution data in natural language processing, it's essential to select an appropriate dataset. Given the lack of consensus on which benchmark to use for evaluating OOD detection methods in NLP, we choose to rely on a conventionally used benchmark (Chen et al., 2022).

We selected the SST-2 dataset as the training distribution and opted to evaluate the OOD detection performance on three different datasets, namely 20news, TREC, and WMT16.

Furthermore, we opted to work with the pre-trained encoder Roberta (Liu et al., 2019) in our study.

3 Results

3.1 Visualisation through Uniform Manifold Approximation and Projection

In this particular study, we are interested in analyzing the embedding of data both in distribution (i.e., data that is similar to the training data) and out of distribution (i.e., data that is dissimilar to the training data) across the layers of a neural network (here ROBERTA).

By visualizing the evolution of data embedding across the layers of the network using Uniform Manifold Approximation and Projection (UMAP), we can gain insight into how the network is processing and transforming the input data. Specifically, we are interested in whether the data becomes more separable (i.e., easier to distinguish between out and in distribution) as it passes through the layers of the network.

In order to make this visualisation, we used UMAP (McInnes et al., 2018), that achieved better performance for manifold visualisation than t-SNE that is more commonly used. UMAP is a dimensionality reduction technique that is based on the idea of preserving the local structure of the high-dimensional data in a low-dimensional space. UMAP has become increasingly popular in the machine learning community due to its ability to capture both global and local structure of the data, making it an effective tool for visualising complex datasets. In this section, we will briefly describe how UMAP works before presenting our results using this technique.

UMAP is a nonlinear dimensionality reduction technique that starts by constructing a weighted graph representing the high-dimensional data. The graph is constructed by connecting nearby points in the high-dimensional space with edges that are weighted according to a kernel function that measures the distance between the points.

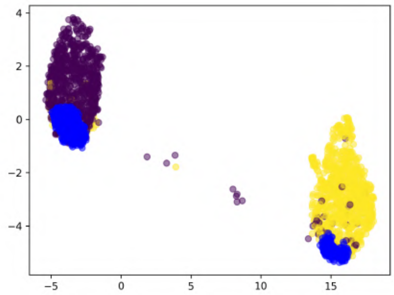

Figure 1: UMAP visualisation of the last layer with OOD dataset news20. Blue : OOD, Yellow: InD y = 1, Purple: InD y = 0

UMAP then optimises a low-dimensional embedding of the data that preserves both the global and local structure of the graph. This is achieved by minimising a cost function that balances the preservation of pairwise distances in the high-dimensional space with the preservation of the weighted graph structure in the low-dimensional space.

In practice, UMAP works by first randomly initialising a low-dimensional embedding of the data, and then iteratively refining it using stochastic gradient descent to minimise the cost function. The resulting embedding is a compressed representation of the original data that can be visualised in two or three dimensions.

As shown in Figure 1 and Appendix A, we observed that the distribution of test data becomes increasingly bimodal as we move through the layers of the network. This is a critical point in the analysis of OOD detector performance since all the models does not performing equaly when the probability distribution that we want to compare to are multimodals.

For example, using only the last layer with the IRW-based model would lead to poor results since the IRW distance is a poor estimator of the distance between a point and a distribution when the distribution is multimodal.

3.2 Metrics

We also computed the metrics (AUROC and FPR) evolution throughout the layers of the model and we compared them to the metrics computed on

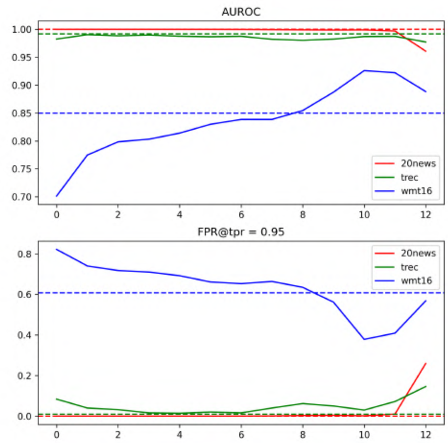

Figure 2: MAHALANOBIS: Evolution of the metrics throughout the layers. The dotted line represents the metric values calculated using the average of the embeddings

the mean of the embedding (as in (Colombo et al., 2022a) or in (Chen et al., 2022)) showned dashed in the figures 2 and 3. Our findings align with those of the authors: when facing challenging tasks, it is preferable to use the average of embeddings instead of relying solely on the last layer. Nevertheless, our investigation suggests that adopting the average approach may not always yield the best results. In fact, the optimal choice of layers varies depending on the model's type of underlying distance metric employed.

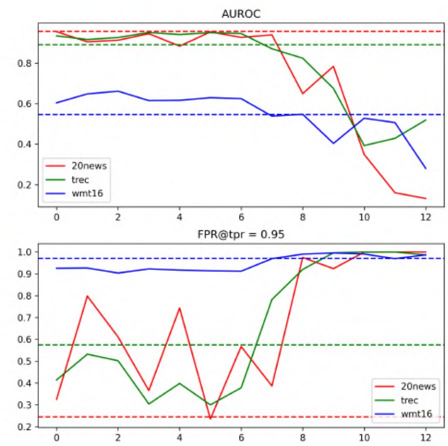

Figure 3: IRW: Evolution of the metrics throughout the layers. The dotted line represents the metric values calculated using the average of the embeddings

Figure 3 reveals a surprising finding: when we examine the false positive rate (FPR) for the 20news dataset, we observe poor performance across nearly all layers. However, the metric dashed, representing the metric computed on the mean of the embeddings across all layers stands exhibits strong performance. This show that averaging the embeddings other the whole layers help the model gain information other the initial distribution of the sentence.

The performance of the IRW model gradually declines as we move through the network layers, as indicated by the visualizations, particularly from layer 8 onwards. Meanwhile, the Mahalanobis model experiences a decrease in performance during the last layer, although it appears to be more stable overall. Interestingly, for difficult task (wm16 dataset) the Mahalanobis model shows an increase in AUROC and a decrease in FPR throughout the layers, indicating that the aggregated information in the early layers negatively impacts the model's performance. In summary, while the IRW model's performance can be improved by averaging the embeddings of the layers, it is most effective in the early layers, whereas the Mahalanobis model's performance benefits from averaging over the last layers.

To gain insight into the behavior of the IRW model around the eight layer, we can examine the false positive rates using visualization techniques. In this study, we will use the TREC dataset to illustrate our findings. As shown in figure 4, the false positive rates are plotted in red, revealing that the model's performance deteriorates significantly when the distribution becomes bimodal.

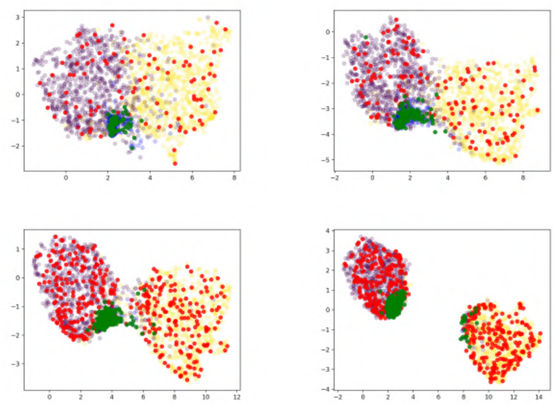

Figure 4: Embeddings of layer 6,7,8 and 9. Red: False Positive, Green: True Negative

The IRW distance is a powerful tool for measuring the similarity between data points, but it also has its limitations. One weakness of the IRW distance is that it relies on the halfspace depth (Tukey, 1975), which can lead to inaccurate metrics when there is no hyperplane to separate the OOD samples from the ID samples. In such cases, when the OOD samples are "in the middle" of the different modes, it is difficult to obtain reliable metrics. This limitation highlights the need for a robust choice of the layer that can avoid this kind of behaviour of the target distribution.

Based on the same idea, we can explain the high value of FPR for the dataset news20 (red) for layer 4 in Figure 5. The croissant shape induces much more miss classification because no hyperplane can separate OOD samples from ID samples.

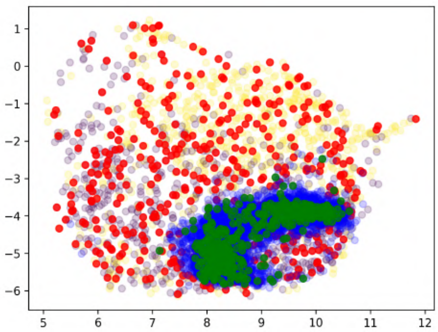

Figure 5: IRW: Embedding of layer 4 for dataset news20. Red: False Positive, Green: True Negative

4 Discussion/Conclusion

In conclusion, this study developed and evaluated two OOD detectors that incorporated all layers of a network, utilizing data depth models Mahalanobis and IRW. The research demonstrated the significance of visualizing the latent spaces of each layer to gain insight into the distribution characteristics that impact detector effectiveness. The findings revealed that it is disadvantageous for an IRW-type detector to consider layers where the distribution is bimodal and starts to resemble the output distribution. These insights offer valuable guidance for enhancing OOD detector performance in real-world applications.

.

References

Eduardo Dadalto Câmara Gomes, Pierre Colombo, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. A functional perspective on multi-layer out-of-distribution detection.

Prasanta Chandra Mahalanobis. 1936. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55.

John W. Tukey. 1975. Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, 2:523–531.

Regina Y. Liu, Regina Y. Parelius, and Kesar Singh. 1999. Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics*, 27(3):783–840.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction.

Kelly Ramsay, Stéphane Durocher, and Alexandre Leblanc. 2019. Integrated rank-weighted depth. *Journal of Multivariate Analysis*, 173(C):51–69.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Pierre Colombo. 2021. *Learning to represent and generate text using information measures*. Ph.D. thesis, (PhD thesis) Institut polytechnique de Paris.

Pierre Colombo, Chloe Clavel, and Pablo Piantanida. 2021. A novel estimator of mutual information for learning to disentangle textual representations. *ACL 2021*.

Pierre Colombo, Eduardo D. C. Gomes, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. 2022a. Beyond mahalanobis-based scores for textual ood detection.

Sishuo Chen, Xiaohan Bi, Rundong Gao, and Xu Sun. 2022. Holistic sentence embeddings for better out-of-distribution detection.

Pierre Colombo, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. 2022b. Learning disentangled textual representations via statistical measures of similarity. *ACL 2022*.

Marine Picot, Nathan Noiry, Pablo Piantanida, and Pierre Colombo. 2023a. Adversarial attack detection under realistic constraints.

Maxime Darrin, Pablo Piantanida, and Pierre Colombo. 2023a. Rainproof: An umbrella to shield text generators from out-of-distribution data. *arXiv preprint arXiv:2212.09171*.

Marine Picot, Guillaume Staerman, Federica Granese, Nathan Noiry, Francisco Messina, Pablo Piantanida, and Pierre Colombo. 2023b. A simple unsupervised data depth-based method to detect adversarial images.

Maxime Darrin, Guillaume Staerman, Eduardo Dadalto Câmara Gomes, Jackie CK Cheung, Pablo Piantanida, and Pierre Colombo. 2023b. Unsupervised layer-wise score aggregation for textual ood detection. *arXiv preprint arXiv:2302.09852*.

A Appendix A : 20news Dataset, UMAP Embedding of several layers

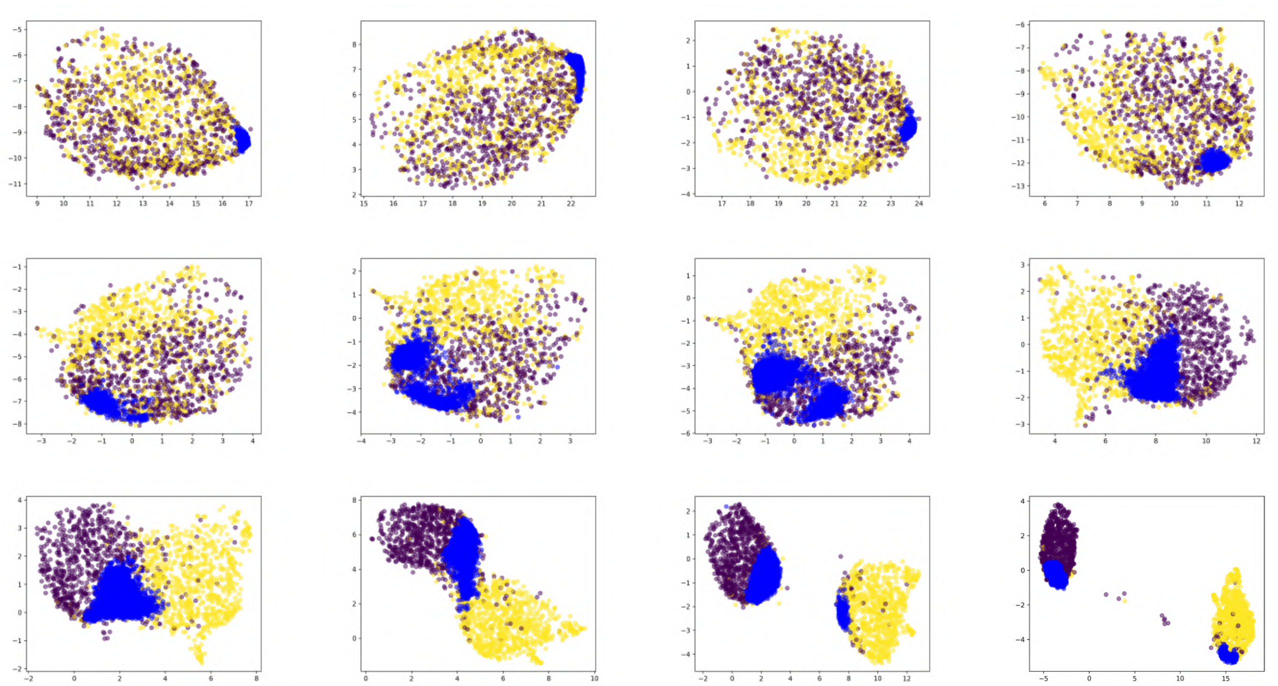

Figure 6: Embeddings of layer from 1 to 12. Blue: OOD, Yellow: In Distribution label = 1, Purple: In Distribution, label = 0

B Appendix B : TREC Dataset, UMAP Embedding of several layers

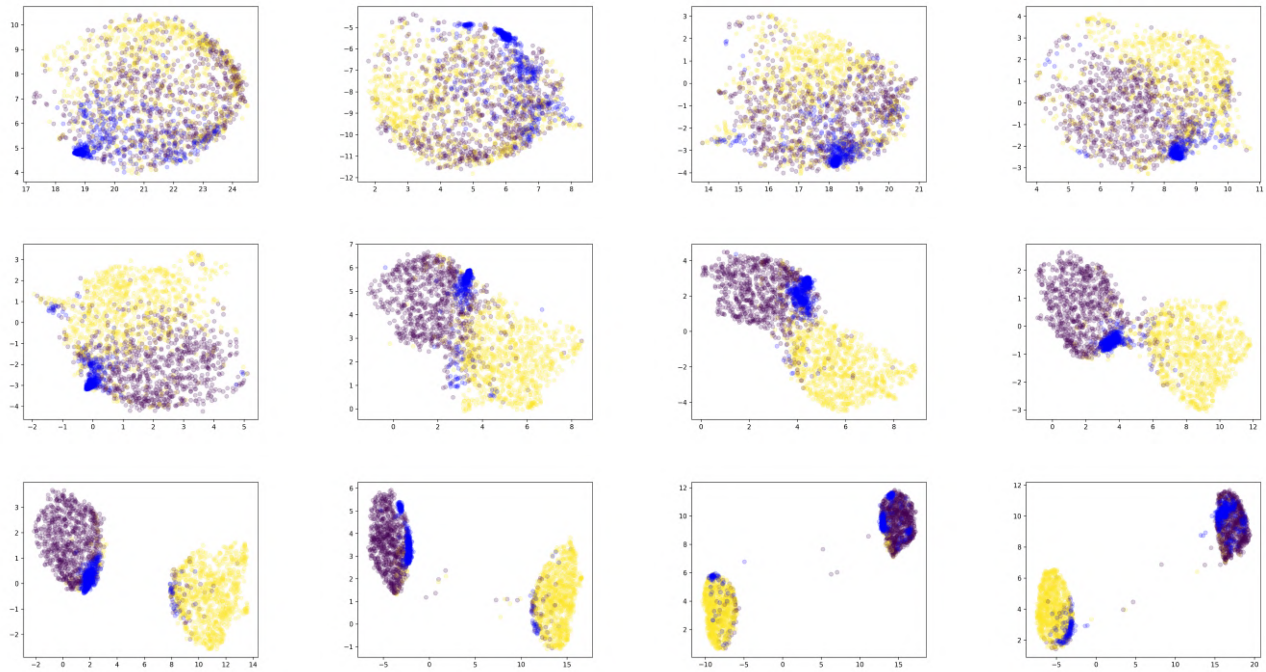

Figure 7: Embeddings of layer from 1 to 12. Blue: OOD, Yellow: In Distribution label = 1, Purple: In Distribution, label = 0

C Appendix C : WM16 Dataset, UMAP Embedding of several layers

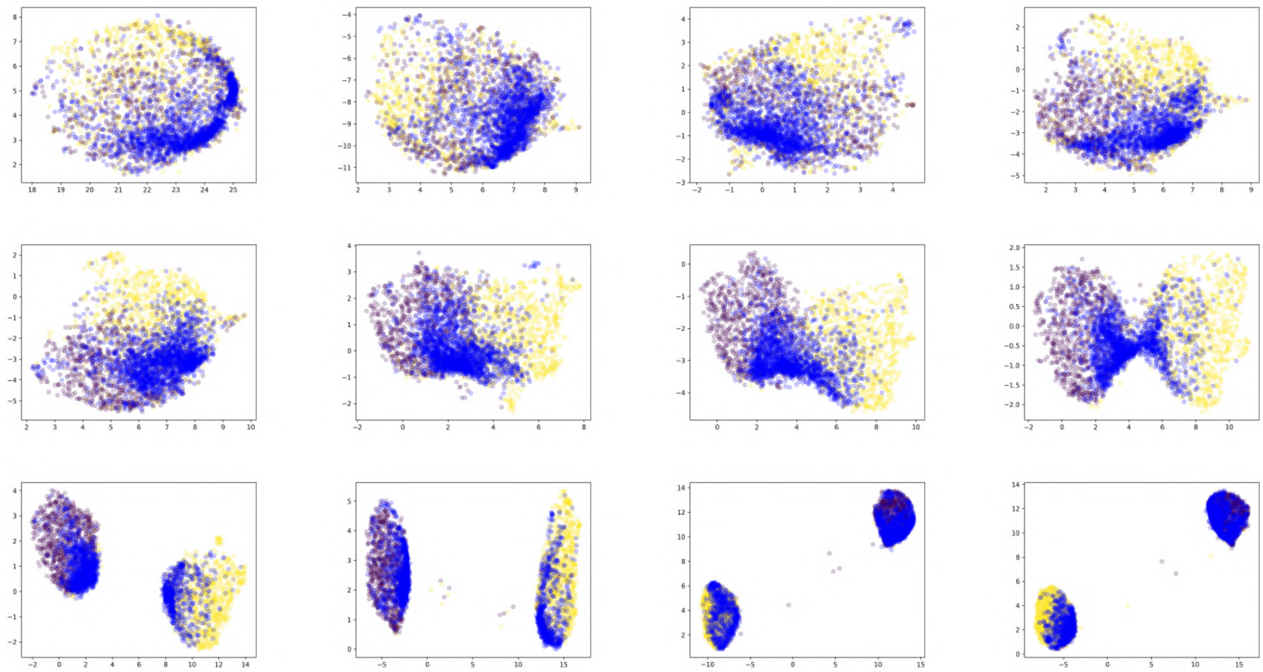

Figure 8: Embeddings of layer from 1 to 12. Blue: OOD, Yellow: In Distribution label = 1, Purple: In Distribution, label = 0