Utilizing Representation Learning for Robust Text Classification Under Datasetshift

Max Lübbering[1], Michael Gebauer[2], Rajkumar Ramamurthy[1], Maren Pielka[1], Christian Bauckhage[1] and Rafet Sifa[1]

[1]*Fraunhofer IAIS, Sankt Augustin, Germany*
[2]*TU Berlin, Berlin, Germany*

Abstract

Within One-vs-Rest (OVR) classification, a classifier differentiates a single class of interest (COI) from the rest, i.e. any other class. By extending the scope of the rest class to corruptions (dataset shift), aspects of outlier detection gain relevancy. In this work, we show that *adversarially trained autoencoders* (ATA) representative of autoencoder-based outlier detection methods, yield tremendous robustness improvements over traditional neural network methods such as multi-layer perceptrons (MLP) and common ensemble methods, while maintaining a competitive classification performance. In contrast, our results also reveal that deep learning methods solely optimized for classification, tend to fail completely when exposed to dataset shift.

Keywords

Dataset shift, Representation Learning, Outlier detection, One-vs-rest classification

1. Introduction

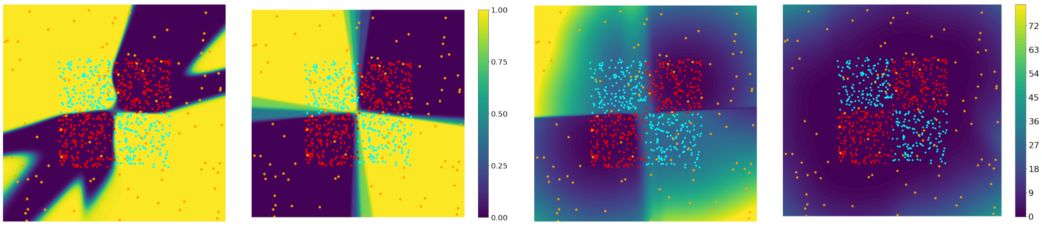
In recent years, deep neural networks (DNNs) have constantly achieved new SOTA results [1]. Despite these tremendous breakthroughs, they often fall behind expectations in reality[2].

Firstly, previous research exposed major blunders of DNNs providing wrong predictions with high confidence when exposed to dataset shift and adversarial examples[3, 4, 5, 6]. These robustness deficiencies can be visualized through the conceptual example in Fig. 1, in which the MLP has successfully learned to distinguish the XOR squares. However, when exposed to the uniform noise samples, the model wrongly classifies the noise with high confidence to belong to one of the two classes. Partially, this can be attributed to the softmax function as a fast-growing exponential function approximating a smooth indicator function, rendering predictions instable[3], and to the training process which is only separation oriented and focused on empirical risk minimization[7].

Secondly, models are often trained and evaluated in artificial environments, raising concerns on the transferability of the reported performances when applied in practice[2].

In this research paper, we investigate the issue of dataset shift robustness deficiencies for DNNs within the one vs rest (OVR) classification setting and empirically show that significant improvements can be achieved when incorporating reliable outlier detection techniques. By

LWDA'21: Lernen, Wissen, Daten, Analysen September 01–03, 2021, Munich, Germany
max.luebbering@iais.fraunhofer.de (M. Lübbering)



(a) MLP (b) MIMO (c) ATA (d) OCA

Figure 1: Class probabilities of MLP / MIMO and reconstruction errors of ATA / OCA visualized as contours on the noisy non-linear XOR squares dataset.

definition[8], outliers are samples that are generated from a completely different distribution than the inliers. Analogously, OVR classification aims to filter a single class of interest (COI) from the rest, i.e. all the remaining classes (RC). By incorporating out-of-distribution data within RC, i.e. classes unrelated to the training domain, methods from outlier detection gain relevancy. To evaluate the approaches w.r.t. classification and robustness performance, we depict a specific task for each concern: 1) For the classification task T_c, we evaluate the model on the classes it was trained on. 2) For the dataset shift task T_d, the model is evaluated on the inlier class of T_c and rest samples, i.e. outliers, derived from an unrelated dataset, similar to the evaluation approach in [3].

As previously shown by [9, 10, 11, 12, 13], autoencoder-based representation learning has been proven successful in detecting outliers. In this work, we utilize the two outlier detection methods *adversarially trained autoencoders* (ATA) [10] and *one class autoencoders* (OCA) to compare their classification and robustness performance to MLPs and the recently published ensemble method MIMO[14]. In contrast to the semi-supervised OCA, ATA not only minimizes the reconstruction error of COI samples but also actively maximizes the reconstruction error of RC samples, thereby making the reconstruction error a richer outlierness feature, especially when COI samples are correlated with RC samples.

Our contributions are summarized as follows: We show that traditional DNNs such as MLP and ensemble methods are highly unreliable when exposed to dataset shift. As a viable solution, we propose autoencoder-based outlier detection methods to OVR classification resulting in accurate classifiers that are highly robust to dataset shift. Furthermore, our results indicate that robustness can slightly harm classification performance, which is in line with previous research results [15, 16].

2. Adversarially Trained Autoencoders

Adversarially trained autoencoders (ATA) have been proven to be highly effective for outlier detection by incorporating a priori outlier information into the training process[10, 11]. While semi-supervised methods such as OCA, minimize the reconstruction loss $L_{\text{MSE}}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n}\sum_i^n (x_i - \hat{x}_i)^2$ for samples $\mathbf{x} \in$ inliers only, ATA additionally maximizes the L_{MSE} for outliers. Given sample $\mathbf{x}$, its reconstruction $\hat{\mathbf{x}}$ and target t, the so called adversarial loss function computes to

$$L_{adv}(\mathbf{x}, \hat{\mathbf{x}}, t) = \begin{cases} 0, & L_{\text{MSE}} \in \text{[l, u]} \quad \wedge \quad t \in \text{outliers} \\ L_{\text{MSE}}(\mathbf{x}, \hat{\mathbf{x}}), & L_{\text{MSE}}(\mathbf{x}, \hat{\mathbf{x}}) \geq u \quad \vee \quad t \in \text{inliers} \\ -\alpha L_{\text{MSE}}(\mathbf{x}, \hat{\mathbf{x}}), & \text{otherwise}, \end{cases} \tag{1}$$

where outlier weighting factor α determines the outlier maximization intensity. This loss function captures the reconstruction error of outliers within the bounds l, u, by maximizing / minimizing the reconstruction loss accordingly, as unlimited maximization could lead to exploding gradients. Inlier samples are generally minimized. The network architecture is given by

$$f(\mathbf{x}) = e_{\text{MSE}}(\mathbf{x}, d(e(\mathbf{x}))), \tag{2}$$

where $e_{\text{MSE}}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n} \sum_i^n (x_i - \hat{x}_i)^2$ is the reconstruction error, which takes a sample $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$ as input. The reconstruction is provided by the autoencoder defined by the nesting of the decoder d and encoder e.

Since ATA does not build upon unstable output functions like softmax[3] and learns a concrete representation of the COI, it is by design more robust to corruptions compared to other deep learning models like MLPs or ensemble methods like MIMO[14]. This makes ATA not only a compelling method for outlier detection but also a robust method for OVR.

3. Experiments and Results

To compare ATA to the three baselines MLP, MIMO and OCA on an algorithmic rather than model level, we perform nested cross validation (CV) [17]. For each algorithm, we select the best models by the *area under the precision recall curve* (AUPR) and report the score aside with *area under the receiver operating characteristics* (AUROC) and F1 score. AUPR and F1 score are calculated w.r.t. COI. Since AUROC and AUPR are threshold-independent, they yield a more comprehensive evaluation compared to e.g., F1 score. Unlike AUROC, AUPR takes the base rate of the positive class into account and thus is more applicable to settings with high class imbalance[3]. AUROC can be interpreted as the probability of ranking a random positive sample higher than a random negative sample[18].

For a fair comparison, ATA and the baselines are defined to have a comparable parameter complexity. MIMO, MLP, as well as decoder and encoder (but in reverse) each possess three hidden layers of size 50, 25 and 12 with sigmoid activations. Due to its five parallel input layers, MIMO has the highest number of trainable parameters. All approaches have a binary output, which alleviates aforementioned softmax stability. As part of the nested CV, all algorithms were hyperparameter-tuned w.r.t. *learning rate* and *weight decay*. Additionally, ATA was optimized for *outlier weighting factor* and *bin range*.

	ATIS		Reuters		Newsgroups	
	Rest	COI	Rest	COI	Rest	COI
Train/ S_c	a_c	flight	r_c	acq, earn	n_c	sci.space
S_{d1}	t_d	flight	t_d	acq, earn	t_d	sci.space
S_{d2}	r_d	flight	a_d	acq, earn	a_d	sci.space
S_{d3}	n_d	flight	n_d	acq, earn	r_d	sci.space

Table 1
Class assignment within splits S_c, S_{d1}, S_{d2} and S_{d3} for each dataset: Splits S_c and S_d are representative of tasks T_c and T_d, respectively. Mapping of rest classes specified in Tabl. 2

Dataset	Abbr.	Rest labels
Reuters	r_c	crude, interest, money-fx, money-supply, ship, retail, wpi, cpi, jobs, cotton, ipi, reserves, gnp, tin, carcass, housing, nat-gas, pet-chem, oilseed, rubber, orange, lumber, livestock, heat, wpi
	r_d	trade, grain, ship, gold, interest, money-fx, money-supply, jobs, sugar, tin, ipi, cpi, cocoa, coffee, cotton, coppper, alum, rubber, yen, nat-gas,reserves
ATIS	a_c	airfare, ground_service, airline
	a_d	abbreviation, restriction, airport, quantity, meal, city, flight_no, ground_fare, flight_time, flight, distance, aircraft, capacity
News groups	n_c	sci.crypt, sci.med, talk.politics.guns, misc.forsale, rec.sport.baseball, talk.politics.misc, comp.os.ms-windows.misc, soc.religion.christian
	n_d	rec.sport.hockey, sci.crypt, sci.med, comp.sys.ibm.pc.hardware, talk.politics.mideast, comp.sys.mac.hardware, rec.autos, sci.electronics, talk.religion.misc, alt.atheism, rec.motorcycles, comp.windows.x, comp.graphics, sci.space, talk.politics.guns, misc.forsale, rec.sport.baseball, talk.politics.misc, comp.os.ms-windows.misc, soc.religion.christian
TREC[4]	t_d	HUM, NUM, LOC, ABBR

Table 2
Indicated by the matching indices, rest labels of each dataset for each of the splits S_c, S_{d1}, S_{d2} and S_{d3}, as specified in Tabl. 1.

3.1. Datasets

To evaluate dataset shift robustness, we consider the three textual datasets Reuters[1], ATIS[2] and Newsgroups[3]. As shown in Tabl. 1 and Tabl. 2, the COI always deals with a very narrow topic and rest samples originate from a diverse set of classes, as characteristic for OVR. Split S_c resembles the classification task T_c since it contains all training classes. Splits S_{d1}, S_{d2} and S_{d3}, representative of dataset shift task T_d, contain rest samples from a novel dataset, similar to [3].

3.2. Results

As summarized in Tabl.3, MIMO and MLP yield strong classification results on T_c, which however is tainted by significant performance degradation when exposed to dataset shift within T_d. Conversely, OCA provides strong robustness during dataset shift exposure, whereas fails completely on task T_c.

[1]http://www.daviddlewis.com/resources/testcollections/reuters21578/
[2]http://www.ai.sri.com/natural-language/projects/arpa-sls/atis.html
[3]http://qwone.com/ jason/20Newsgroups/

		ATIS			REUTERS			Newsgroups		
		AUROC	AUPR	F1 Score	AUROC	AUPR	F1 Score	AUROC	AUPR	F1 Score
S_c	ATA	95.0 ± 0.6	98.9 ± 0.2	92.0 ± 0.7	99.4 ± 0.1	99.8 ± 0.0	97.8 ± 0.6	94.8 ± 1.6	86.7 ± 2.1	74.7 ± 1.6
	OCA	71.9 ± 2.2	91.3 ± 1.2	64.3 ± 1.7	82.3 ± 4.7	94.3 ± 2.0	76.1 ± 3.5	67.7 ± 2.1	29.2 ± 2.8	30.0 ± 2.9
	MLP	**99.3** ± 0.1	**99.8** ± 0.0	**98.0** ± 0.4	**99.7** ± 0.1	**99.9** ± 0.1	**99.0** ± 0.3	**97.5** ± 0.4	**90.9** ± 0.6	79.2 ± 4.3
	MIMO	98.9 ± 0.2	**99.8** ± 0.0	97.8 ± 0.1	99.6 ± 0.1	**99.9** ± 0.0	98.7 ± 0.1	97.3 ± 0.5	89.6 ± 1.2	**82.3** ± 1.7
	BASE	50.0	81.9	31.0	50.0	77.3	30.4	50.0	11.4	9.3
S_{d1}	ATA	**98.7** ± 0.1	**96.8** ± 0.5	**90.5** ± 0.4	97.6 ± 1.3	96.4 ± 1.6	74.2 ± 12.4	97.2 ± 0.6	55.3 ± 1.8	**70.3** ± 1.7
	OCA	97.8 ± 0.2	96.5 ± 0.3	66.1 ± 1.6	**98.6** ± 0.2	**98.3** ± 0.4	77.5 ± 3.1	**99.3** ± 0.5	**98.9** ± 0.6	41.9 ± 3.8
	MLP	73.1 ± 8.2	37.4 ± 7.7	46.3 ± 3.3	90.1 ± 1.8	68.0 ± 6.3	67.0 ± 1.8	91.3 ± 2.9	30.7 ± 5.7	40.9 ± 8.9
	MIMO	81.4 ± 2.0	41.0 ± 3.4	44.8 ± 1.3	93.1 ± 2.2	83.2 ± 5.0	59.0 ± 2.9	84.6 ± 1.7	26.5 ± 4.5	28.1 ± 2.3
	BASE	50.0	20.9	14.7	50.0	29.0	18.4	50.0	6.1	5.4
S_{d2}	ATA	**99.1** ± 0.2	**98.8** ± 0.3	**92.5** ± 0.6	**98.5** ± 0.9	97.1 ± 1.5	75.6 ± 13.2	90.2 ± 2.0	38.8 ± 4.0	26.9 ± 3.2
	OCA	94.2 ± 0.5	94.3 ± 0.5	66.1 ± 1.6	**98.5** ± 0.3	**97.9** ± 0.5	**77.5** ± 3.1	**99.4** ± 0.5	**99.1** ± 0.6	**41.9** ± 3.8
	MLP	82.0 ± 10.0	68.1 ± 16.3	64.5 ± 5.5	84.2 ± 5.6	49.2 ± 16.0	47.5 ± 4.4	81.4 ± 11.1	20.3 ± 18.8	18.8 ± 5.7
	MIMO	95.9 ± 1.1	92.6 ± 2.6	75.1 ± 3.1	83.8 ± 7.8	62.5 ± 16.8	39.8 ± 2.0	60.0 ± 3.5	9.7 ± 5.9	8.1 ± 0.2
	BASE	50.0	34.6	20.5	50.0	21.6	15.1	50.0	4.2	3.9
S_{d3}	ATA	**93.7** ± 2.0	82.4 ± 4.8	50.6 ± 17.5	**94.3** ± 2.6	84.1 ± 1.8	21.2 ± 10.5	96.2 ± 1.0	78.7 ± 2.8	**71.8** ± 2.1
	OCA	92.2 ± 0.6	**83.1** ± 1.0	**66.1** ± 1.6	87.2 ± 8.5	74.8 ± 18.9	**67.7** ± 21.4	**98.8** ± 0.7	**98.1** ± 0.8	41.9 ± 3.8
	MLP	69.9 ± 8.1	12.7 ± 12.9	9.2 ± 1.3	97.1 ± 1.1	67.6 ± 13.3	27.6 ± 4.4	95.8 ± 2.1	74.0 ± 8.8	69.8 ± 9.6
	MIMO	92.7 ± 2.1	57.8 ± 8.2	11.9 ± 1.8	95.0 ± 4.1	**85.9** ± 10.2	14.7 ± 2.1	88.8 ± 2.1	57.9 ± 11.8	56.5 ± 4.6
	BASE	50.0	4.1	3.8	50.0	6.2	5.5	50.0	11.5	9.3

Table 3
Performance of ATA and baselines on splits S_c, S_{d1}, S_{d2} and S_{d3}: Across the two subtasks of OVR, ATA yields robust results, while MLP/ MIMO and OCA show a significant performance degradation on the dataset shift and classification task, respectively. Metrics and confidence reported in %. Failures highlighted in red for AUROC $< 90\%$. AUPR and F1 score failures due to base rate dependency not considered. BASE resembles a random classifier predicting COI with probability $p \sim U[0,1]$ for reference.

In contrast, ATA represents an effective trade-off between classification performance and robustness to dataset shift. The results are always at least close to the best performing model on each task and never show complete model failures. On the contrary, MLP and MIMO each fail in terms of AUROC on T_d in 5/9 cases. Analogously, OCA fails in all cases on task T_c.

The results are well-aligned with the visualization in Fig. 1, in which MLP, MIMO and ATA are capable of separating the inlier class from the rest class, however, in a fundamentally different fashion. ATA learns a hull around the COI samples and therefore is able to reject the rest class including any corruptions. This is also reflected in the experiments, in which ATA not only provides a strong performance on the classification task but also higher robustness on the dataset shift task. The contours of OCA in Fig. 1, reveal a major overlap of COI and rest class, as rest samples get minimized implicitly when minimizing COI. This inherent problem is also present in the experiments, where OCA is robust to corruptions, but consistently fails on T_c.

4. Conclusion

We investigated model robustness on one-vs-rest classification by extending the scope of the rest class to strong corruptions (dataset shift). We find that conventional DNNs such as MLP and deep ensembles (MIMO), provide highly unstable predictions when exposed to dataset shift.

With ATA, as an outlier detection method based on autoencoders, we showed that tremendous robustness improvements can be achieved, while slightly compromising classification performance. Especially in safety-related and volatile environments with model robustness as a principal concern, ATA poses a worthwhile consideration.

References

[1] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature (2015).

[2] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, D. Mané, Concrete Problems in AI Safety, 2016.

[3] D. Hendrycks, K. Gimpel, A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks, Proc. of Int. Conf. on Learning Representations (2017).

[4] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, 2017.

[5] A. Nguyen, J. Yosinski, J. Clune, Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images, in: CVPR, 2015.

[6] I. Goodfellow, J. Shlens, C. Szegedy, Explaining and Harnessing Adversarial Examples, in: Int. Conf. on Learning Representations, 2015.

[7] J. Van Amersfoort, L. Smith, Y. W. Teh, Y. Gal, Uncertainty Estimation Using a Single Deep Deterministic Neural Network, in: Proc. of the 37th Int. Conf. on Machine Learning, 2020.

[8] C. C. Aggarwal, Outlier analysis, in: Data mining, Springer, 2015.

[9] R. Chalapathy, A. K. Menon, S. Chawla, Anomaly Detection using One-Class Neural Networks, CoRR (2018). URL: http://arxiv.org/abs/1802.06360.

[10] M. Lübbering, R. Ramamurthy, M. Gebauer, T. Bell, R. Sifa, C. Bauckhage, From Imbalanced Classification to Supervised Outlier Detection Problems: Adversarially Trained Auto Encoders, in: Artificial Neural Networks and Machine Learning – ICANN 2020, 2020.

[11] M. Lübbering, M. Gebauer, R. Ramamurthy, R. Sifa, C. Bauckhage, Supervised autoencoder variants for end to end anomaly detection, in: ICPR Int. Workshops and Challenges, 2021.

[12] S. Hawkins, H. He, G. Williams, R. Baxter, Outlier Detection using Replicator Neural Networks, in: Proc. of Int. Conf. on Data Warehousing and Knowledge Discovery, 2002.

[13] M. Lübbering, M. Pielka, K. Das, M. Gebauer, R. Ramamurthy, C. Bauckhage, R. Sifa, Toxicity detection in online comments with limited data: A comparative analysis, in: ESANN, 2021. In press.

[14] H. Marton, et al., Training Independent Subnetworks for Robust Prediction, in: International Conference on Learning Representations, 2021.

[15] R. G. Lopes, D. Yin, B. Poole, J. Gilmer, E. D. Cubuk, Improving robustness without sacrificing accuracy with patch gaussian augmentation, arXiv preprint (2019).

[16] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, B. Lakshminarayanan, Augmix: A simple data processing method to improve robustness and uncertainty, arXiv preprint arXiv:1912.02781 (2019).

[17] S. Raschka, Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning, arXiv preprint arXiv:1811.12808 (2018).

[18] T. Fawcett, An Introduction to ROC Analysis, Pattern recognition letters (2006).