ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Adversarial Robustness by Combining Prototype Models with Lipschitz Training

Adia Khalid and Benjamin Paaßen

Bielefeld University - Faculty of Technology
Inspiration 1, 33619 Bielefeld - Germany

Abstract. Beyond accuracy, adversarial robustness and interpretability are crucial elements of trustworthy machine learning systems. Prototype-based models such as generalized learning vector quantization (GLVQ) have favourable robustness and interpretability properties but do not achieve state-of-the-art accuracy in many practical domains, such as image classification. Applying prototype-based models in the embedding space of a deep neural network boosts their accuracy but removes their interpretability and robustness. We partially resolve this dilemma: We prove that robustness guarantees of shallow classifiers translate to robustness guarantees of deep classifiers when imposing Lipschitz continuity, we provide a training scheme to achieve Lipschitz continuity, and we empirically validate our approach on three image classification data sets against fast gradient sign attacks.

1 Introduction

With the increasing deployment of machine learning systems and their integration into critical systems, security has become a serious concern [1]. Among the various threats, adversarial attacks are particularly concerning, as they involve manipulating input data with small, often imperceptible perturbations to mislead a model's predictions [2]. However, adversarial examples not only expose security issues of machine learning models, but also their lack of interpretability, in the sense that models systematically perceive input different from humans [3]. Ideally, models become robust to adversarial examples in a sense that enhances interpretability, as well: to change a model's output, inputs need to change in a way that would make a human change their judgment, accordingly [4].

Among the most promising defense paradigms against adversarial attack is the use of Lipschitz continuity [5, 6]. Lipschitz continuity means that the change in output of a model is upper-bounded by the change in input times a constant, the so-called Lipschitz-constant. Accordingly, unexpectedly large changes in the output, including adversarial attacks, are prevented. Popular methods to achieve Lipschitz continuity are regularization methods, such as spectral norm regularization[6, 7]. However, prior Lipschitz defenses do not address the underlying interpretability concerns, limiting trust and complicating regulatory compliance, debugging, bias auditing and use of the models in decision-making processes [8, 9].

Prototype-based models like Generalized Learning Vector Quantization (GLVQ) offer inherently interpretable decision-making mechanisms [10, 11] as

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

well as certificates of adversarial robustness due to maximum-margin properties [12]. Classifications are performed by assigning data points to a closest prototype, such that any decision can be explained by showing the prototypes. However, LVQ models are, by default, shallow models with limited performance on real-world image data sets.

In this paper, we combine a GLVQ classifier with a deep embedding network to enable high accuracy on image classification tasks but we retain GLVQ's favourable robustness properties by imposing Lipschitz continuity on the deep embedding network. Specifically, our contributions are 1) a proof that a robustness guarantees on a shallow classifier translate to robustness guarantees on a deep classifier when imposing Lipschitz continuity, 2) A simple and efficient training scheme to achieve Lipschitz continuity, 3) an experimental evaluation of adversarial robustness on three image classification data sets.

2 Method

Our goal is to leverage the interpretability and robustness advantages of prototype-based models when combined with deep embedding networks for pre-processing. Assume a deep neural network $\phi : \mathcal{X} \to \Phi$, which maps input data from some input space $\mathcal{X}$ to an n-dimensional embedding space Φ. Further, assume that the input space $\mathcal{X}$ is equipped with some metric d and the embedding space Φ with some metric D. Finally, assume that we have trained a prototype-based classifier $f : \Phi \to \{1, \ldots, K\}$ which maps data in the embedding space to one of K labels based on the label of the closest prototype.

The core challenge we try to address is that a deep neural network ϕ may yield an embedding space that enables us to improve classification accuracy but it can break adversarial robustness. Points that look close in the input space may be suddenly far apart in the embedding space. Hence, even if the classifier f is adversarially robust in the embedding space, this does not help us to be adversarially robust in the input space (see Figure 1, top). Our core argument is now quite simple: Lipschitz-continuity exactly limits how far apart two points can be in the embedding space given their closeness in the input space (see Figure 1, bottom). Hence, a robust classifier in the embedding space in combination with a deep network that conforms to the Lipschitz property yields a robust classifier in the input space.

To be more precise, we define a classifier $f : \Phi \to \{1, \ldots, K\}$ as **ϵ-robust** around a point $x \in \Phi$ with respect to metric D if for all points $y \in \Phi$ where $D(x, y) \leq \epsilon$, we have $f(x) = f(y)$. For the purpose of our argument, we will assume that this property holds point-wise for all training data points for a GLVQ classifier in the embedding space and we point to extensive prior work for proofs of this property [12, 13, 14].

We further recall the definition of Lipschitz continuity: The function $\phi : \mathcal{X} \to \Phi$ is called (locally) **L-Lipschitz** around some point $x \in \mathcal{X}$ with respect to the metrics d and D if for all points $y \in \mathcal{X}$ it holds: $D(\phi(x), \phi(y)) \leq L \cdot d(x, y)$.

Theorem: If $\phi : \mathcal{X} \to \Phi$ is L-Lipschitz with respect to the metrics d and D

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

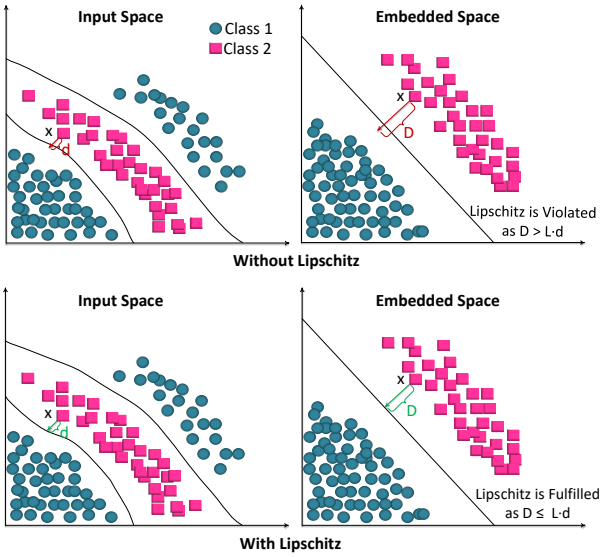


Fig. 1: Top Left: An adversarial attack of size d from point x. Top right: In embedding space, the attack is enlarged to size D, violating Lipschitz continuity. Bottom: With Lipschitz training, the classification boundary shifts such that the attack fails because D is now upper-bounded by $L \cdot d$.

around some point $x \in \mathcal{X}$ and $f : \Phi \to \{1, \ldots, K\}$ is ϵ-robust around $\phi(x)$ with respect to D, then $\phi \circ f$ (defined as $f(\phi(x))$) is $\frac{\epsilon}{L}$-robust around x with respect to d.

Proof: Consider any $y \in \mathcal{X}$ with $d(x, y) \leq \frac{\epsilon}{L}$. Then, $D(\phi(x), \phi(y)) \leq L \cdot d(x, y) \leq \epsilon$ because ϕ is L-Lipschitz around x with respect to d and D. Hence, $f(\phi(x)) = f(\phi(y))$ because f is ϵ-robust around $\phi(x)$ w.r.t. D.

While this theorem is, admittedly, somewhat trivial, it is highly useful for our purpose. The theorem implies that we can leverage the robustness and interpretability properties of GLVQ models in combination with deep networks if the deep networks are sufficiently Lipschitz continuous. Hence, we devise a training method that ensures this Lipschitz property.

Our core strategy is quite simple: For each training data point x, we sample one neighbor y by adding Gaussian random noise to x. Then, we punish violations of the Lipschitz property by imposing the loss

$$\mathcal{L}_{\text{Lipschitz}} := \text{ReLU}\Big(D(\phi(x), \phi(y)) - L \cdot d(x, y)\Big). \tag{1}$$

In our practical experiments, we use the squared Euclidean distance for both D and d. We, then, simply add this loss (with a weight) to the regular loss of the classifier training. Note that the neighbors y should be sampled anew for each minibatch of the training process. Because neighbors are sampled as Gaussian noise, this scheme is substantially more efficient than adversarially

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

robust training as proposed by [15] because y does not need to be an adversarial example. Also note that only local Lipschitz conditions around the training data are enforced (which is sufficient for robustness), retaining more model flexibility compared to global Lipschitz constraints.

3 Experiments

Data sets: We consider three image classification data sets, namely MNIST, CIFAR-10 (both via pytorch), and Imagenette-160[1].

Convolutional Backbone: We consider image classification with a convolutional neural network for feature extraction. For MNIST, we apply two CNN layers with 3x3 kernels, each followed by a ReLU nonlinearity and 2x2 max pooling, and conclude with a fully connected layer mapping to 16 feature dimensions. For CIFAR-10 and Imagenette-160, we apply three CNN layers with 3x3 kernels, each followed by batch normalization, ReLU, and 2x2 max pooling, and conclude with a 0.5-dropout layer and a fully connected layer mapping to 64 feature dimensions.

Models: We compare four different combinations of classification layers and training schemes: 1) CNN + fully connected layer and crossentropy loss for classification, 2) CNN + GLVQ classifier with standard GLVQ loss, 3) CNN + fully connected layer and Lipschitz training, and 4) CNN + GLVQ classifier with Lipschitz training. Our prediction is that only the final variant (our proposed method) will achieve substantial adversarial robustness by combining a robust classifier in the embedding space (GLVQ) with a Lipschitz-continuous CNN. For Lipschitz training, we use $L = 1$ and weigh the Lipschitz loss by a factor of 10. Across models, we train for 50 epochs. The batch sizes are 32 for MNIST, 128 for CIFAR-10, and 64 for Imagenette-160. For GLVQ, we used one prototype per class. For the Lipschitz training, we generated noise with standard deviation of 0.1 for MNIST and 0.03 for CIFAR-10 and Imagenette. The source code for the proposed model is freely available for research purposes at `https://gitlab.ub.uni-bielefeld.de/publications-ag-kml/adversarial_robustness_lipschitz_glvq`.

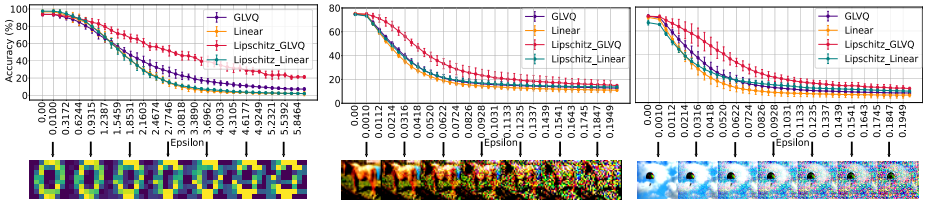
Evaluation measure: We evaluate the adversarial robustness by measuring the accuracy after adversarial attacks with the Fast Gradient Sign Method (FGSM) [9] across a wide range of attack strengths (in line with the robustness curve proposal of [16]), namely $\epsilon \in [0.001, 0.2]$ for CIFAR-10 and Imagenette-160 and $\epsilon \in [0.001, 6.0]$ for MNIST. To give some sense of interpretability, we also show example images and their attacked versions for various ϵ.

Results: Figures 2a, 2b, and 2c present the results on MNIST, CIFAR-10, and Imagenette-160, respectively. We observe a consistent pattern: For $\epsilon = 0$ (no adversarial attack), all models exhibit high accuracy with ca. 98% for MNIST and ca. 70% for CIFAR-10 and Imagenette-160. However, performance swiftly degrades with increasing ϵ and only Lipschitz-GLVQ consistently outperforms

[1] `https://github.com/fastai/imagenette`

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

the other approaches, often by more than a standard deviation. The second-most robust approach is GLVQ without Lipschitz training, which is likely due to the fact that our relatively small CNN is somewhat Lipschitz continuous even without Lipschitz training. The most convincing results are obtained for MNIST where GLVQ + Lipschitz only breaks down when the images are very strongly altered to the point where a human observer may not even recognize them anymore. In the other datasets, even high ϵ would likely not lead a human observer to change their predicted class (refer to the example images), illustrating that the current level of robustness is still not satisfactory for interpretability – albeit substantially improved.



(a) MNIST (b) CIFAR-10 (c) Imagenette-160

Fig. 2: Adversarial robustness curves for all models and data sets. Error bars indicate standard deviation. Below each curve, an example image is shown with FGSM attacks against the GLVQ+Lipschitz model.

4 Conclusion

We developed a novel approach to achieve adversarial robustness when combining prototype-based models – which are robust and inherently interpretable, but shallow – with deep neural networks. In particular, we proved that enforcing Lipschitz continuity on a deep embedding network translates robustness guarantees from the embedding space to the input space and we provided a training scheme to enforce Lipschitz continuity. Empirically, we verified on three image classification data sets (MNIST, CIFAR-10, Imagenette-160) that a robust GLVQ classifier combined with Lipschitz training on the underlying CNN embedding network outperforms CNN+GLVQ without Lipschitz training, CNN+Lipschitz training without GLVQ as well as a plain CNN.

Our paper suggests plenty of opportunities for further work: while we obtained some level of robustness, this is still not satisfactory for interpretability: changes that humans would be able to ignore still flip a classifier's output. Further robust training approaches and stricter Lipschitz constants may thus be needed. Investigations of alternative metrics and alternative embedding network or classifier architectures may be informative. User studies are needed to reveal how much interpretability is actually preserved when using a deep embedding network (even with Lipschitz training). We believe that a rich research field of interpretable models in combination with deep learning is yet to be discovered.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

References

[1] Shuai Zhou, Chi Liu, Dayong Ye, Tianqing Zhu, Wanlei Zhou, and Philip S. Yu. Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity. *ACM Computing Surveys*, 55(8), 2022.

[2] Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International journal of automation and computing*, 17(2):151–178, 2020.

[3] Jan Philip Göpfert, André Artelt, Heiko Wersing, and Barbara Hammer. Adversarial attacks hidden in plain sight. In *Proceedings of the IDA*, pages 235–247, 2020.

[4] Jon Vadillo, Roberto Santana, and Jose A Lozano. Adversarial attacks in explainable machine learning: A survey of threats against models and humans. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15(1):e1567, 2025.

[5] Monty-Maximilian Zühlke and Daniel Kudenko. Adversarial robustness of neural networks from the perspective of lipschitz calculus: A survey. *ACM Computing Surveys*, 57(6):1–41, 2025.

[6] Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941*, 2017.

[7] Hiskias Dingeto and Juntae Kim. Comparative study of adversarial defenses: Adversarial training and regularization in vision transformers and cnns. *Electronics*, 13(13), 2024.

[8] Biao Xu and Guanci Yang. Interpretability research of deep learning: A literature survey. *Information Fusion*, 115:102721, 2025.

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv:1412.6572*, 2014.

[10] Atsushi Sato and Keiji Yamada. Generalized learning vector quantization. *Advances in neural information processing systems*, 8, 1995.

[11] Fredrick Mumali and Joanna Kałkowska. Generalized matrix learning vector quantization computational method for intelligent decision making: A systematic literature review. *Archives of Computational Methods in Engineering*, pages 1–23, 2025.

[12] Barbara Hammer, Marc Strickert, and Thomas Villmann. On the generalization ability of grlvq networks. *Neural Processing Letters*, 21(2):109–120, 2005.

[13] Sascha Saralajew, Lars Holdijk, and Thomas Villmann. Fast adversarial robustness certification of nearest prototype classifiers for arbitrary seminorms. In *Proceedings of the NeurIPS*, 2020.

[14] Václav Voráček and Matthias Hein. Provably adversarially robust nearest prototype classifiers. In *Proceedings of the ICML*, 2022.

[15] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv:1706.06083*, 2017.

[16] Christina Göpfert, Jan Philip Göpfert, and Barbara Hammer. Adversarial robustness curves. In *Workshops of the ECML 2019*, pages 172–179, 2020.