

Introducing Quality Guarantees for Autoencoders

Benedikt Böing¹, Rajarshi Roy², Emmanuel Müller¹ and Daniel Neider²

¹TU Dortmund, Germany

²Max Planck Institute for Software Systems, Germany

Abstract

Autoencoders are an essential concept in unsupervised learning. Currently, the quality of autoencoders is assessed either internally (e.g., by mean square error) or externally (e.g., by classification performance). Yet, there is no possibility to prove that autoencoders generalize beyond the finite training data, and hence they are not reliable for safety-critical applications requiring formal guarantees for unseen data. To address this issue, we propose the first framework to bound the worst-case error of an autoencoder within a safety-critical region of an infinite value domain, as well as the definition of unsupervised adversarial examples that cause such worst-case errors. Technically, our framework reduces the infinite search space for a uniform error bound to checking satisfiability of logical formulas in Linear Real Arithmetic. This allows us to leverage highly-optimized SMT solvers, a strategy that is very successful in the context of deductive software verification. We demonstrate our ability to find unsupervised adversarial examples as well as formal quality guarantees on a real-world dataset from the medical domain.

1. Introduction

Autoencoders are widely used for many unsupervised learning tasks such as cluster analysis [1], compression [2], anomaly detection [3], as well as a variety of preprocessing steps [4, 2, 5] in other machine learning pipelines. The general assumption is that data can be compressed into a lower dimensional latent space by an encoder function extracting the most relevant features of the data distribution. From this latent representation the decoder tries to reconstruct the original input. As the latent representation is an information bottleneck the autoencoder's input deviates from its output. Typically the autoencoder reconstructs better in dense regions (i.e., regions with many training examples) than in regions with few training examples [3] giving rise to its application in anomaly detection. Moreover even the small errors in dense regions are a desirable property as they allow it to be used, e.g., for denoising. At the same time it is necessary to control the error for all points in dense regions because otherwise the result - whether it is the latent representation or the reconstruction - is of little use. To this end current approaches to assess autoencoders either measure internally the mean square error (MSE) on the unsupervised training data or external performance on some supervised application such as classification performance [6, 7, 8].


However, a major shortcoming of these approaches is that they cannot provide a formal guarantee in terms of the maximum deviation between input and output of the autoencoder as it

OVERLAY 2021: 3rd Workshop on Artificial Intelligence and Formal Verification, Logic, Automata, and Synthesis, September 22, 2021, Padova, Italy

✉ benedikt.boeing@cs.tu-dortmund.de (B. Böing); rajarshi@mpi-sws.org (R. Roy); emmanuel.mueller@cs.tu-dortmund.de (E. Müller); neider@mpi-sws.org (D. Neider)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

is evaluated on test data only (i.e., with a finite number of inputs). This lack of formal quality guarantees for autoencoders leads to a very limited applicability of such unsupervised learning schemes for safety-critical applications. For instance, it is particularly important to consider the maximum deviation when working with data containing clusters. In such situations the autoencoder should not mix up the clusters because otherwise the autoencoder’s results are meaningless. If the maximum deviations for the respective clusters are small enough though, the autoencoder provably keeps the clusters separated.

To address this and other shortcomings of unsupervised learning with autoencoders, we provide the first methodology to bound an autoencoder’s worst-case error in a safety-critical region. Inspired by the vast literature on supervised adversarial attacks [9, 10, 11], we define the notion of *unsupervised adversarial examples* as inputs (not necessarily contained in the training data) on which the autoencoder’s error exceeds a user-defined threshold. Then we define the *worst-case error* of an autoencoder as the largest error that can possibly manifest.

Following a popular approach in the area of software verification, we reduce the problem of finding an unsupervised adversarial example to a satisfiability check of a formula in linear Real Arithmetic [12, 13]. This allows us to apply highly-optimized, off-the-shelf satisfiability modulo theory (SMT) solvers which can effectively reason about the infinite domains and, hence, can prove the existence or non-existence of unsupervised adversarial examples. Once we have found an unsupervised adversarial example, it serves as a lower bound for the worst-case error. Moreover, a binary search allows us to approximate the worst-case error arbitrarily well. We demonstrate the usefulness of our QUGA (QUality GUarantees for Autoencoders) approach by proving that a given autoencoder keeps two classes of a medical dataset separated. For more experiments and details please refer to the extended version of this paper [14].

2. QUGA: Problem Statement

In general, an autoencoder tries to reproduce its input; that is, it is trained to compute $f(x) = x$. However, it does so while propagating it through a latent space which typically has less dimensions than the input/output space. This latent space serves as an information bottleneck and therefore introduces errors to the identity function the autoencoder is supposed to learn. However, as most applications of autoencoders rely on a good approximation of the identity function, we are naturally interested in quantifying its error. More precisely, our goal is to give formal guarantees in terms of the maximum deviation from the identity function.

As a first step towards this goal, we define the notion of *adversarial examples of autoencoders*. Intuitively, such adversarial examples are inputs on which the “distance” between the input and the output of the autoencoder is larger than a (user-defined) threshold $\varepsilon > 0$.

Definition 1 (ε -adversarial examples). Let $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be an autoencoder, $dist: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}_+$ a distance function, and $\varepsilon > 0$. An ε -adversarial example is a point $x \in \mathbb{R}^n$ such that

$$dist(x, f(x)) > \varepsilon$$

(i.e., a point on which the input and output of f deviate more than ε).

Note that our definition of ε -adversarial examples is not restricted to inputs in the training or test sets but allows any input $x \in \mathbb{R}^n$. In the context of safety-critical systems, however, it

is not enough to identify individual ε -adversarial examples, but it is necessary to know the worst-case (i.e., maximum) error an autoencoder produces. As we cannot expect to find a global maximum of the unbounded error, we restrict the region for which we want to calculate the worst-case error.

Definition 2 (Worst-case error of autoencoders). Let $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be an autoencoder, $dist: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}_+$ a distance function, and $A \subseteq \mathbb{R}^n$ an (infinite) safety-critical region of inputs. Then, the worst-case error of f in A is defined as

$$wce(f, A) = \sup \{ dist(x, f(x)) \in \mathbb{R}_+ \mid x \in A \}$$

(i.e., the largest deviation of an input in the region A from its output).

Definition 2 serves as our novel *quality criterion* for autoencoders that reflects how good the identity function is learned in the specific region of interest. It overcomes the limitation of classical quality metrics that are defined on finite test data only. In total, this leads us to the main problem statement.

Problem 1 (QUGA: Quality Guarantees for Autoencoders). Given a region $A \subseteq \mathbb{R}^n$, an autoencoder $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ and a distance function $dist: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}_+$, compute $wce(f, A)$.

In general, computing the worst-case error is a very challenging and computationally intractable problem. Hence in the next section we introduce a restricted, solvable version of it.

3. Solution Framework

In this section we outline a framework for computing ε -adversarial examples and the worst-case error of autoencoders in a restricted and hence computationally tractable version of Problem 1. The following restrictions are designed in such a way that the solution framework remains applicable to a wide range of autoencoders used in practice:

1. We assume the the neurons of the autoencoder have linear or ReLU (Rectified Linear Unit) activation functions.
2. We assume the distance function to be the L_1 or L_∞ -norm.
3. We assume the safety-critical region A to be a finite union of convex compact polytopes.
4. We approximate the worst-case error up to a user-defined accuracy.

Given these assumptions we can encode the problem of finding an ε -adversarial example in a given region A as an instance of the SMT problem [15]. To this end we define formulas encoding the autoencoder (φ_{ae}), the region (φ_A) and the distance function ($\varphi_\varepsilon^{dist}$) and concatenate them to $\varphi_\varepsilon^{ae} = \varphi_{ae} \wedge \varphi_A \wedge \varphi_\varepsilon^{dist}$ such that the following theorem holds.

Theorem 1. Let f be an autoencoder, A a region, $dist$ a distance function, $\varepsilon > 0$, and φ_ε^{ae} as defined above. Then, the following two properties hold:

1. If A contains an ε -adversarial example, then φ_ε^{ae} is satisfiable.
2. If φ_ε^{ae} is satisfiable, then any solution of φ_ε^{ae} is an ε -adversarial example in A .

In order to approximate the worst-case error in a region we can thereafter exploit this property by performing a binary search over the values of ε and keeping track of lower and upper bounds on the worst-case error.

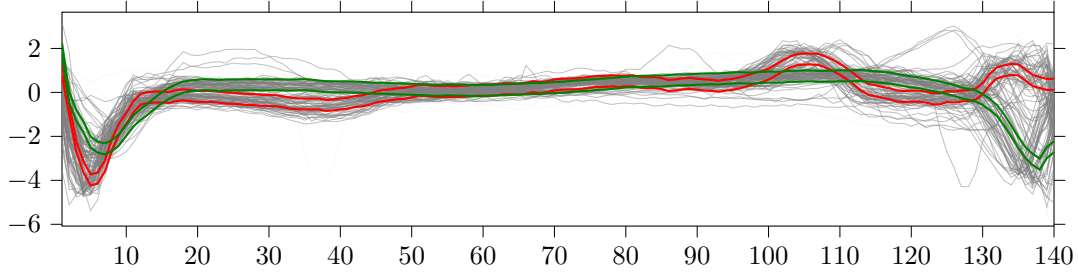


Figure 1: ECG5000 dataset with two safety-critical regions (red and green) obtained by extracting prototypes for two classes and adding a margin of 0.25.

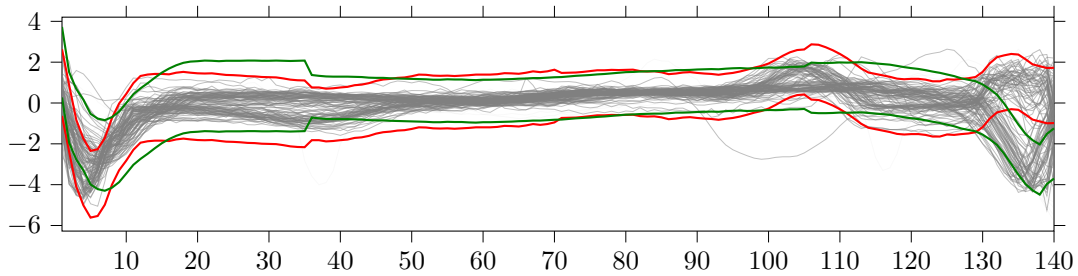


Figure 2: Image spaces of the autoencoder (red and green) into which points from the respective safety-critical regions in Figure 1 can theoretically be mapped by the autoencoder.

4. Empirical Evaluation

We demonstrate our QUGA framework on the ECG5000 dataset (Figure 1), by evaluating the unsupervised training based on two time series clusters. The goal of a traditional evaluation would be to show that all training objects are clearly separated in the latent space. In contrast, we care about all possible (infinitely many) objects in two safety-critical areas (green and red) that need to be distinguishable in the latent space. In Figure 2 we see the resulting corridors (green and red) into which points from the corresponding critical regions can be mapped. As the two corridors do not overlap on all timesteps (e.g., in time 137) the autoencoder keeps the two safety-critical classes apart allowing for the use of the latent space representations for follow up tasks such as, e.g., classification.

5. Conclusion

QUGA overcomes major shortcomings of unsupervised learning with autoencoders. We provide the first methodology to bound the error of an autoencoder in a safety-critical region. With our solution framework based on SMT solvers we propose to search for adversarial examples and the worst-case error in the infinite search space of a safety-critical region. Our QUGA approach formulates the autoencoder, the safety-critical region, and the error of the loss function with a logical conjunction of linear constraints. This allows us to prove separation of the input data in the latent space for further downstream applications.

References

- [1] S. E. Chazan, S. Gannot, J. Goldberger, Deep clustering based on A mixture of autoencoders, in: 29th IEEE International Workshop on Machine Learning for Signal Processing, 2019.
- [2] Q. Meng, D. R. Catchpoole, D. Skillicom, P. J. Kennedy, Relational autoencoder for feature extraction, in: 2017 International Joint Conference on Neural Networks, 2017.
- [3] M. Sakurada, T. Yairi, Anomaly detection using autoencoders with nonlinear dimensionality reduction, in: Proc. of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis, 2014.
- [4] L. Gondara, Medical image denoising using convolutional denoising autoencoders, in: IEEE International Conference on Data Mining Workshops, 2016.
- [5] L. Pasa, A. Sperduti, Pre-training of recurrent neural networks via linear autoencoders, in: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, 2014.
- [6] Q. V. Le, M. Ranzato, R. Monga, M. Devin, G. Corrado, K. Chen, J. Dean, A. Y. Ng, Building high-level features using large scale unsupervised learning, in: Proceedings of the 29th International Conference on Machine Learning, 2012.
- [7] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P. Manzagol, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, *J. Mach. Learn. Res.* (2010).
- [8] M. R. Min, D. A. Stanley, Z. Yuan, A. J. Bonner, Z. Zhang, A deep non-linear feature mapping for large-margin knn classification, in: ICDM 2009, The Ninth IEEE International Conference on Data Mining, 2009.
- [9] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: 3rd International Conference on Learning Representations, 2015.
- [10] N. N. Dalvi, P. M. Domingos, Mausam, S. K. Sanghai, D. Verma, Adversarial classification, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004.
- [11] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, R. Fergus, Intriguing properties of neural networks, in: 2nd International Conference on Learning Representations, 2014.
- [12] G. Katz, C. W. Barrett, D. L. Dill, K. Julian, M. J. Kochenderfer, Reluplex: An efficient SMT solver for verifying deep neural networks, in: Computer Aided Verification - 29th International Conference, 2017.
- [13] L. M. de Moura, N. Bjørner, Z3: an efficient SMT solver, in: Tools and Algorithms for the Construction and Analysis of Systems, 14th International Conference, 2008.
- [14] B. Böing, R. Roy, E. Müller, D. Neider, Quality guarantees for autoencoders via unsupervised adversarial attacks, in: Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD, 2020.
- [15] L. M. de Moura, N. Bjørner, Satisfiability modulo theories: introduction and applications, *Commun. ACM* (2011).