Abstract

In recent years, deep neural networks have revolutionized many application domains of machine learning and are key components of many critical decision or predictive processes. Therefore, it is crucial that domain specialists can understand and analyze actions and predictions, even of the most complex neural network architectures. Despite these arguments neural networks are often treated as black boxes. In the attempt to alleviate this shortcoming many analysis methods were proposed, yet the lack of reference implementations often makes a systematic comparison between the methods a major effort. The presented library \textit{iNNvestigate} addresses this by providing a common interface and out-of-the-box implementation for many analysis methods, including the reference implementation for PatternNet and PatternAttribution as well as for LRP-methods. To demonstrate the versatility of \textit{iNNvestigate}, we provide an analysis of image classifications for variety of state-of-the-art neural network architectures.
Keywords: Artificial neural networks, deep learning, analyzing classifiers, explaining classifiers, computer vision

1. Introduction

In recent years deep neural networks have revolutionized many domains, e.g., image recognition, speech recognition, speech synthesis, and knowledge discovery (Krizhevsky et al., 2012; LeCun et al., 2012; Schmidhuber, 2015; LeCun et al., 2015; Van Den Oord et al., 2016). Due to their ability to naturally learn from structured data and exhibit superior performance, they are increasingly used in practical applications and critical decision processes, such as novel knowledge discovery techniques, autonomous driving or medical image analysis. To fully leverage their potential it is essential that users can comprehend and analyze these processes. E.g., in neural architecture (Zoph et al., 2018) or chemical compound searches (Montavon et al., 2013; Schütt et al., 2017) it would be extremely useful to know which properties help a neural network to choose appropriate candidates. Furthermore for some applications understanding the decision process might be a legal requirement.

Despite these arguments neural networks are often treated as black boxes, because their complex internal workings and the basis for their predictions are not fully understood. In the attempt to alleviate this shortcoming several methods were proposed, e.g., Saliency Map (Baehrens et al., 2010; Simonyan et al., 2013), SmoothGrad (Smilkov et al., 2017), IntegratedGradients (Sundararajan et al., 2017), Deconvnet (Zeiler and Fergus, 2014), GuidedBackprop (Springenberg et al., 2015), PatternNet and PatternAttribution (Kindermans et al., 2018), LRP (Bach et al., 2015; Lapuschkin et al., 2016a, b, 2019; Montavon et al., 2018), and DeepTaylor (Montavon et al., 2017). Theoretically it is not clear which method solves the stated problems best, therefore an empirical comparison is required (Samek et al., 2017, Kindermans et al., 2017). In order to evaluate these methods, we present iNNvestigate which provides a common interface to a variety of analysis methods.

In particular, iNNvestigate contributes:

- A common interface for a growing number of analysis methods that is applicable to a broad class of neural networks. With this instantiating a method is as uncomplicated as passing a trained neural network to it and allows for easy qualitative comparisons of methods. For quantitative evaluations of (image) classification task we further provide an implementation of the method “perturbation analysis” (Samek et al., 2017).

- Support of all methods listed above—this includes the first reference implementation for PatternNet and PatternAttribution and an extended implementation for LRP—and an open source repository for further contributions.

- A clean and modular implementation, casting each analysis in terms of layer-wise forward and backward computations. This limits code redundancy, takes advantage of automatic differentiation, and eases future integration of new methods.

iNNvestigate is available at repository: https://github.com/albermax/innvestigate

It can be simply installed as Python package and contains documentation for code and applications. To demonstrate the versatility of iNNvestigate we provide examples for the analysis of image classifications for a variety of state-of-the-art neural networks.
**Terminology:** The different methods pose different assumptions to tasks and are designed for different objectives, yet they are related to “explaining” or “interpreting” neural networks, see Montavon et al. (2018). We actively refrain from using this terminology in order to prevent misunderstandings between the design choices of the algorithms and the implicit assumption these terms bring along. Therefore we will solely use the neutral term analyzing and leave any interpretation to the user.

2. Library

**Interface:** The main feature is a common interface to several analysis methods. The workflow is as simple as passing a Keras neural network model to instantiate an analyzer object for a desired algorithm. Then, if needed, the analyzer will be fitted to the data and eventually be used to analyze the model’s predictions. The corresponding Python code is:

```python
import innvestigate
model = create_a_keras_model()
analyzer = innvestigate.create_analyzer('analyzer_name', model)
analyzer.fit(X_train) # if needed
analysis = analyzer.analyze(X_test)
```

**Implemented methods:** At publication time the following algorithms are supported: Gradient Saliency Map, SmoothGrad, IntegratedGradients, Deconvnet, GuidedBackprop, PatternNet and PatternAttribution, DeepTaylor, and LRP including LRP-Z, -Epsilon, -AlphaBeta. In contrast, current related work is limited to gradient- and perturbation-based methods (Kotikalapudi and contributors, 2017; Ancona et al., 2018) or focuses on a single algorithm (E.g., Lundberg and Lee, 2017; Ribeiro et al., 2016). For an overview see Alber (2019). We intend to extend this selection and invite the community to contribute implementations as new methods emerge.

**Documentation:** The library’s documentation contains several introductory scripts and example applications. We demonstrate how the analyses can be applied to the following state-of-the-art models: VGG16 (Simonyan and Zisserman, 2014), InceptionV3 (Szegedy et al., 2016), ResNet50 (He et al., 2016), InceptionResNetV2 (Szegedy et al., 2017), DenseNet (Huang et al., 2017), NASNet (Zoph et al., 2018). Figure 1 shows the result for a subset.

![Figure 1: Result of methods applied to various neural networks (blank, if a method does not support a network’s architecture).](image-url)
2.1. Details

Modular implementation: All of the methods have in common that they perform a back-propagation from the model outputs to the inputs. The core of iNNvestigate is a set of base classes and functions that is designed to allow for rapid and easy development of such algorithms. The developer only needs to implement specific changes to the base algorithm and the library will take care of the complex and error-prone handling of the propagation along the graph structure. Further details can be found in the repositories documentation.

Another advantage of the modular design is that one can extend any analyzer with a given set of wrappers. One application of this is the smoothing of the analysis results by adding Gaussian noise to the copies of the input and averaging the outcome. E.g., SmoothGrad is realized in this way by combining a smoothing wrapper with a gradient analyzer.

Training: PatternNet and PatternAttribution (Kindermans et al., 2018) are two novel approaches that condition their analysis on the data distribution. This is done by identifying the signal and noise direction for each neuron of a neural network. Our software scales favorably, e.g., one can train required patterns for the methods on large data sets like Imagenet (Deng et al., 2009) in less than an hour using one GPU. We present the first reference implementation of these methods.

Quantitative evaluation: Often analysis methods for neural networks are compared by qualitative (visual) inspection of the result. This is can lead to subjective evaluations and one approach to create a more objective and quantitative comparison of analysis algorithms is the method “perturbation analysis” (Samek et al., 2017, also known as “PixelFlipping”). The intuition behind this method is that perturbing regions which are recognized as important for the classification task by the analyzing method, will impact the classification most. This allows to assess which analysis method best identifies regions that matter for a specific task and neural network. iNNvestigate contains an implementation of this method.

Installation & license: iNNvestigate is published as open-source software under the BSD-2-license and can be downloaded from: https://github.com/albermax/innvestigate. It is built as a Python 2 or 3 application on top of the popular and established Keras (Chollet et al., 2015) framework. This allows to use the library on various platforms and devices like CPUs and GPUs. At the time of publication only the TensorFlow (Abadi et al., 2016) Keras-backend is supported. The library can be simply installed as Python package.

3. Conclusion

We have presented iNNvestigate, a library that makes it easier to analyze neural networks’ predictions and to compare different analysis methods. This is done by providing a common interface and implementations for many analysis methods as well as making tools for training and comparing methods available. In particular it contains reference implementations for many methods (PatternNet, PatternAttribution, LRP) and example application for a large number of state-of-the-art applications. We expect that this library will support the field of analyzing machine learning and facilitate research using neural networks in domains such as drug design or medical image analysis.
Acknowledgments

Correspondence to MA, SL, KRM, WS and PJK. This work was supported by the Federal Ministry of Education and Research (BMBF) for the Berlin Big Data Center BBDC (01IS14013A). Additional support was provided by the BK21 program funded by Korean National Research Foundation grant (No. 2012-005741) and the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (no. 2017-0-00451, No. 2017-0-01779).
References


