LioNets: Local Interpretation of Neural Networks through Penultimate Layer Decoding

Ioannis Mollas, Nikolaos Bassiliades, and Grigorios Tsoumakas

Aristotle University of Thessaloniki, Thessaloniki 54124, Greece
{iamollas,nbassili,greg}@csd.auth.gr

Abstract. Technological breakthroughs on smart homes, self-driving cars, health care and robotic assistants, in addition to reinforced law regulations, have critically influenced academic research on explainable machine learning. A sufficient number of researchers have implemented ways to explain indifferently any black box model for classification tasks. A drawback of building agnostic explanators is that the neighbourhood generation process is universal and consequently does not guarantee true adjacency between the generated neighbours and the instance. This paper explores a methodology on providing explanations for a neural network's decisions, in a local scope, through a process that actively takes into consideration the neural network's architecture on creating an instance's neighbourhood, that assures the adjacency among the generated neighbours and the instance.The outcome of performing experiments using this methodology reveals that there is a significant ability in capturing delicate feature importance changes.

Keywords: Explainable · Interpretable · Machine Learning · Neural Networks · Autoencoders.

1 Introduction

Explainable artificial intelligence is a fast-rising area of computer science. Most of the research in this area is currently focused on developing methodologies and libraries for interpreting machine learning models for two main reasons: a) increased use of black box machine learning models, such as deep neural networks, in safety-critical applications, such as self-driving cars, health care and robotic assistants, and b) radical law changes empowering ethics and human rights, which introduced the right of users to an explanation of machine learning models' decisions that concern them.

Local Explanators are methods aiming to explain individual predictions of a particular model. LIME [18] is a state-of-the-art methodology that first constructs a local neighbourhood around a given new unlabeled instance, by perturbing the instance's features, and then trains a simpler transparent decision model to extract the features' importance. Subsequent model agnostic methods like Anchors [19], X-SPELLS [12] and LORE [8] focused on generating better neighbourhoods.

This paper is concerned with generating better neighbourhoods too. However, it focuses on neural network models in particular, in contrast to the *model agnostic* local explanators mentioned in the previous paragraph that can work with any type of machine learning model. Our approach is inspired by the following observation: small changes at the input layer might lead to large changes at the penultimate layer of a (deep) neural network, based on which the final decision of the network is taken. We hypothesize that creating neighbourhoods at the penultimate layer of the neural network instead, could lead to better explanations.

To investigate this intuitive research hypothesis, we introduce our approach, dubbed LioNets (Local Interpretation Of Neural nETworkS through penultimate layer decoding). LioNets constructs a local neighbourhood at the penultimate layer of the neural network and records the network's decisions for this neighbourhood. However, in order to build a transparent local explanator, we need to have input representations at the original input space. To achieve this, LioNets trains a decoder that learns to reconstruct the input examples from their representations at the penultimate layer of the neural network. Taking together, the neural network model and the decoder resemble an autoencoder.

For the evaluation of LioNets, a set of experiments have been conducted, whose code is available at GitHub repository "LioNets"[1]. The results show that LioNets can lead to more precise explanations than LIME.

2 Background and Related Work

In order to be able to present LioNets architecture, this section will provide a sequence of definitions concerning the matter of explainable machine learning, autoencoders and knowledge distillation.

2.1 Explainable Machine Learning

Explainable artificial intelligence is a broad and fast-rising field in computer science. Recent works focus on ways to interpret machine learning models. Thus, this paper will focus on explainable machine learning. An accurate definition is the following:

> "An interpretable system is a system where a user cannot only see but also study and understand how inputs are mathematically mapped to outputs. This term is favoured over "explainable" in the ML context where it refers to the capability of understanding the work logic in ML algorithms" [1].

There are several dimensions that can define an interpretable system according to [9]. One interesting dimension is the *scope* of interpretability. There are two different scopes. An interpretable system can provide global or/and local

[1] https://github.com/iamollas/LioNets

explanations for its predictions. Global explanations can present the structure of the whole system, while local explanations are focused on particular instances.

In the same paper, they are also presenting the desired features of any interpretable system. Those are:

- **Interpretability**: Interpretability measures how much comprehensible is an explanation. In fact, there is not a formal metric because for every problem we measure different attributes.
- **Accuracy**: The accuracy, and probably other metrics, of the original model and the accuracy of the explanator.
- **Fidelity**: Fidelity describes the mimic ability of the explanator, namely the ability of the explanator on providing the same results as the model it explains for specific instances.

2.2 Autoencoders

Autoencoders is a growing area within deep learning [13]. An autoencoder is an unsupervised learning architecture and can be expressed as a function

$$f : X \to X. \tag{1}$$

Autoencoder networks are widely used for reducing the dimensionality of the input data. They initially encode the original data into some latent representation and subsequently reconstruct the original data by decoding this representation to the original dimensions. The most common varieties of autoencoders are the three following:

- **Vanilla**: A three-layered neural network with one hidden layer.
- **Multilayer**: A deeper neural network with more than one hidden or recurrent layers. For example Variational Autoencoders [11, 17].
- **Convolutional**: Used for image or textual data. In practice, the hidden layers are not fully connected, but convolutional layers.

2.3 Related Work

As already mentioned, LIME [18] is a state-of-the-art method for explaining predictions. It follows a simple pipeline. It generates a neighbourhood of a specific size for an instance by choosing randomly to put a zero value in one or more features of every neighbour. Then the cosine similarity of each neighbour with the original instance is measured and multiplied by one hundred. This constitutes the weight on which the simple linear model will depend on for its training. Thus, the most similar neighbours will have more impact on the training process of the linear model. A disadvantage of LIME is in sparse data. Due to the perturbation method that takes place on the original space, LIME can only generate 2^n different neighbours, where n the number of non-zero values. For example, in textual data, in a sentence of six words represented as a vector of four thousand features, where each feature corresponds to a word from the

vocabulary, the non-zero features are only six. Hence, only $2^6 = 64$ different neighbours can be generated. However, LIME will create a neighbourhood of five thousand instances by randomly sampling through the 64 unique neighbours.

X-SPELLS [12] is a forthcoming solution providing model agnostic local explanations to black boxes dealing with sentiment analysis problems. The core idea of this work is to generate neighbourhoods for instances, which they will contain semantically correct synthetic neighbours, using techniques similar to paraphrasing. By creating such neighbourhoods, using variational autoencoders [11, 17] to create new examples in the latent space, the goal is to present some of these neighbours to the user as the explanation. To accomplish this, they train a decision tree on the neighbourhood with labels assigned from the black box and subsequently they are extracting the exemplars.

Another set of methodologies in explaining decision systems, and specifically neural networks, are using Knowledge Distillation [10, 7]. Those methods are trying to explain globally the whole structure and the predictions of a deep neural network, by distilling its knowledge to a transparent system. This idea originates by the Dark Knowledge Distillation [20], which is trying to enhance the performance of a shallow network (the student) through the knowledge of a deeper and more complex network (the teacher).

3 LioNets

This section presents the full methodology and architecture of LioNets. LioNets consist of four fundamental sub-architectures, which are visible in Fig 1 at points 1, 2, 6 and 11. The main part of such system is the neural network, which will work as the predictor. A decoder based on the predictor is the second part. Finally, a neighbourhood generation process and a transparent predictor are the last two mechanisms. Hence, the following process should be executed.

3.1 Neural Network Predictor

For a given dataset, a neural network with a suitable fine-tuned architecture is being trained on this dataset. The output layer is by design in the same length as the number of classes of the classification problem. This process is similar to other supervised methodologies of building and training a neural network for classification tasks, which defines a function $f : X \rightarrow Y$.

3.2 Encoder and Decoder

When the training process of the neural network is over, a duplicate it is created. Then removing the last layer of this copy model and labelling every other layer as untrainable, the foundations for the autoencoder have been defined. Actually, these foundations would be the encoder, the first half of the autoencoder, thus only the decoder part is missing. By building successfully the decoder part and training it, the first two stages for Lionets' completion are achieved. Although,

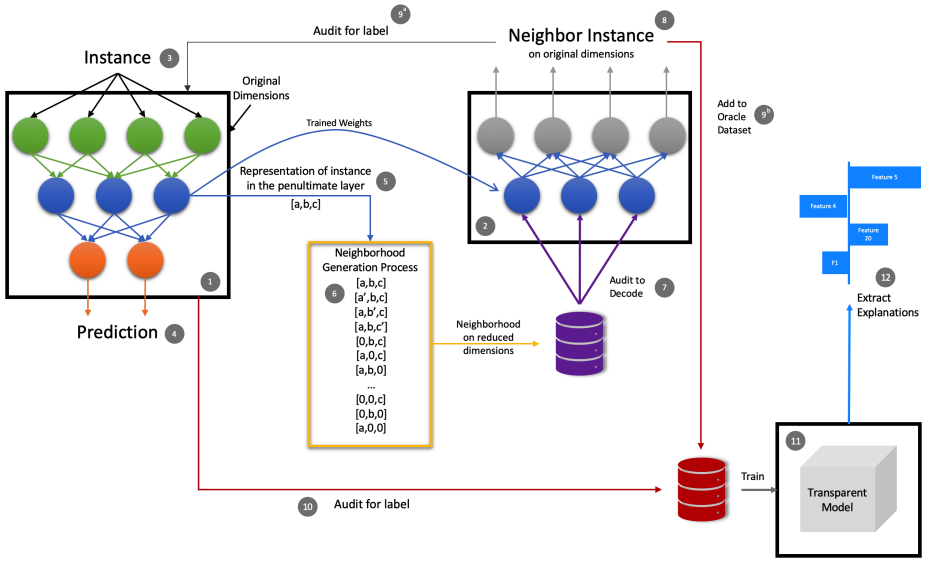


Fig. 1: LioNets' architecture. In this flow chart, the four fundamental mechanisms of LioNets are visible. In point 1 there is the predictor, while in point 2 the decoder. In point 3 there is the neighbourhood generation process and in point 4 the transparent model.

this is the most difficult stage to complete since it is not easy to successfully train autoencoders, especially when the first half of the autoencoder is untrainable. Another approach is to build the autoencoder firstly and afterwards to extract the layers in order to create the encoder, decoder and predictor networks.

Mathematically those neural networks can be expressed via these functions:

$$Encoder\colon X \to Z, \tag{2}$$

$$Decoder\colon Z \to X, \tag{3}$$

$$Autoencoder\colon X \to X, \tag{4}$$

$$Predictor\colon X \to Y. \tag{5}$$

By keeping the encoder part untrainable with stable weights, it guarantees that the generated neighbourhood is transforming from the reduced dimensions to the original dimensions with a decoder, which was trained with the original architecture of the neural network. That process will produce a more representative neighbourhood for the instance, without any semantic meaning to humans.

The academic community has extensively explored ways to create better neighbourhoods for an instance, but every methodology was focused on generating new instances in the level of the input. In this work, the generation processes take place to the latent representation of the encoded input.

3.3 Neighbourhood Generation Process

The neighbourhood generation process takes place after the training of the neural network, the encoder and the decoder. This process could be a genetic algorithm, like the proposed methods in LORE [8] or even another neural network, but simpler solutions are preferred. In LioNets for an instance, that is desirable to get explanations, after encoding it via the encoder neural network, extracting its new representation form from the penultimate level of the neural network, the neighbourhood generation process begins with input the instance with reduced dimensions. By making small changes in the reduced space it could affect more than one dimensions of the original space. Thus, the simple feature perturbation methods on low dimensions will lead to a complex generated neighbour, which most probably would have no semantic meaning for humans.

At that point in time, a specific number of neighbours is generated through a selected generation process and that set of neighbours is given to the decoder, in order to be reversed to the original dimensions.

3.4 Transparent Predictor

By the end of the neighbourhood generation stage, the neighbourhood dataset is almost complete. The only missing part is the neighbours' labels. Thus, the neural network is predicting each instance of the neighbourhood dataset assigning labels to every neighbour, in the form of probabilities. Afterwards, the final dataset with the neighbours and their labels are given as training data to any transparent regression model. The ultimate goal is to overfit that model to the training data.

4 Evaluation

The following section is presenting the setup for the experiments. The data preprocessing methods for two different datasets are described, alongside with the neural network models preparation and the neighbourhood generation process. Finally, there is a discussion about the results of the experiments.

4.1 Setup

Our experiments involve two textual binary classification datasets. The first one concerns the detection of hateful YouTube comments[2] [3] and contains 120 hate and 334 non-hate comments. The second dataset deals with the detection of spam SMS messages [2] and contains 747 spam and 4.827 ham (non-spam) messages. The pre-processing of these datasets consists of the following steps for each document:

– Lowercasing,

[2] https://intelligence.csd.auth.gr/research/hate-speech-detection

– Stemming and Lemmatisation through WordNet lemmatizer [14] and Snowball stemmer [15],
– Phrases transformations [Table 1],
– Removal of punctuation marks,
– Once again, Stemming and Lemmatisation.

“what’s”	to	“what is”
“don’t”	to	“do not”
“doesn’t”	to	“does not”
“that’s”	to	“that is”
“aren’t”	to	“are not”
“’s”	to	“ is”
“isn’t”	to	“is not”
“%”	to	“ percent”
“e-mail”	to	“email”
“i’m”	to	“i am”
“he’s”	to	“he is”
“she’s”	to	“she is”
“it’s”	to	“it is”
“’ve”	to	“ have”
“’re”	to	“ are”
“’d”	to	“ would”
“’ll”	to	“ will”

Table 1: Phrases and words transformations.

Then, for transforming the textual data to vectors a simple term frequency-inverse document frequency [21] (TF-IDF) vectorization technique is taking place.

Afterwards, the neural network predictor for these experiments consists of six layers [Fig 2a] and it has ‘binary_crossentropy’ as loss function. The encoder has five layers, which we extract from the predictor and the decoder has four layers as well [Fig 2b], which we train using ‘categorical_crossentropy’ loss function. The autoencoder is the combination of the encoder and the decoder.

In this set of experiments, a simple generation process via features perturbation methods is applied. Specifically, the creation of neighbours for an instance emerges by multiplying one feature value at a time with 0 and 2^z, $z \in \{-2, -1, 1, 2\}$. Concisely, the above process generates instances which are different in only one dimension in their latent representation.

As soon as the neighbourhood is acquired, every neighbour is transformed via the decoder to the original dimensions. Then, the transformed neighbourhood is given as input to the predictor to predict the class probabilities. Finally, combining the output of the predictor with the transformed neighbourhood a new oracle dataset has been created.

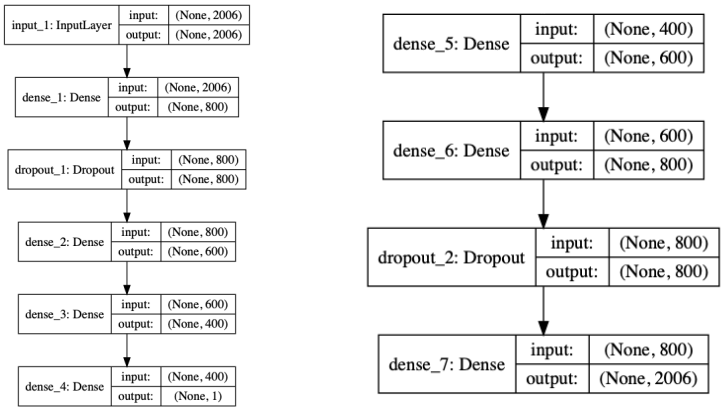


Fig. 2: The predictor's architecture(a) and the decoder's architecture(b).

```
input: neighbourhood
output: X, y
transformed_neighbourhood = decoder.predict(neighbourhood)
class_probabilities = predictor.predict(transformed_neighbourhood)
X = transformed_neighbourhood, y = class_probabilities
```

Algorithm 1.1: Oracle dataset synthesis

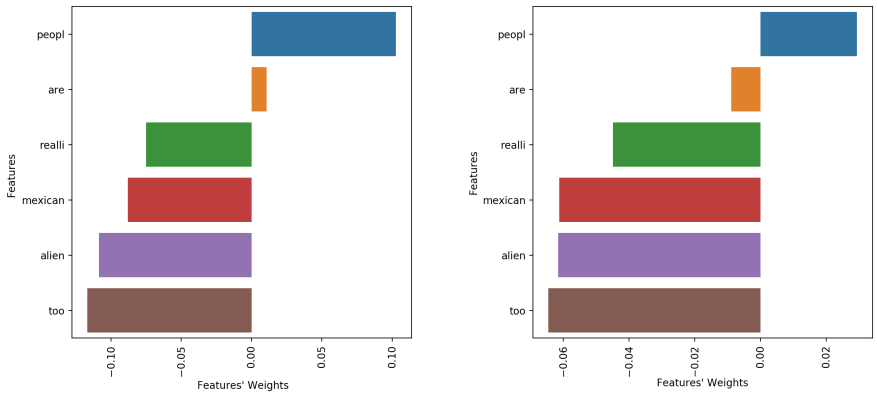
The last step is to train a transparent model with this oracle dataset. It might be useful to check the distribution of probabilities of this dataset and if needed to transform it to have a normal distribution. In these experiments, the transparent model chosen is a Ridge Regression model. By training this model, the coefficients of the features are extracted and transformed into explanations, presented as features' weights in the x-axis of the following figures.

```
input: X, y, instance, feature_names
transparent_model = Ridge().fit(X,y)
coef = transparent_model.coef_
plot_explanation(coef*instance, feature_names)
```

Algorithm 1.2: Explaining an instance

4.2 Results on the Hate Speech Dataset

We take the following YouTube comment from the hate speech dataset as an example: "aliens really, Mexicans are people too". The true class of this comment is *no hate*. According to the neural network, the probability of the *hate* class is approximately 0.00208.



(a) (b)

Fig. 3: Explanation plots of a hate speech instance via (a) LioNets and (b) LIME.

Fig. 3 visualizes the explanation of the neural network's decision via LioNets (3a) and LIME (3b). At first sight, they appear similar. Their main difference is that they assign the feature's "are" contribution to different classes. By removing this word from the instance the neural network predicts 0.00197, which is a lower probability. Thus, it is clear that the feature "are" it was indeed contributing to the "Hate Speech" class for this specific instance as LioNets explained.

Although to support LioNets explanations, the generated neighbourhoods' distances from the original instance computed and presented in Table 2. As it seems the neighbours generated by LIME on original space, in this example, when are encoded to the reduced space are further to the neighbours generated by LioNets in the encoded space. However, when the LioNets' generated neighbours are transformed back to the original space, are more distant to the original instance in comparison to LIME's neighbours, but that is the assumption that has been made through the beginning of these experiments. It is critical to mention at this point, that these distances measured with neighbours generated by changing only one feature at a time.

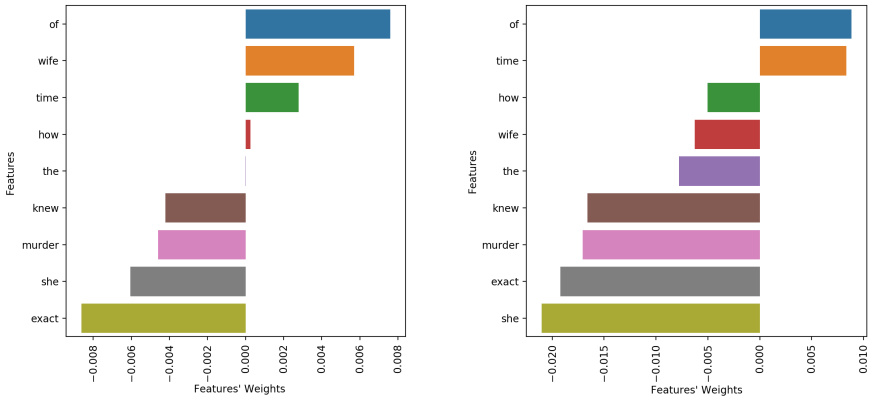
	Euclidean distance
LIME: Generated on Original Space	0.3961
LIME: Encoded	0.9444
LioNets: Generated on Encoded Space	0.2163
LioNets: Decoded to Original Space	0.7635

Table 2: Neighbourhood distances for instance of hate speech dataset.

4.3 Results on SMS spam dataset

The second example which is going to be explained belongs to the SMS spam dataset. The text of the preprocessed instance is the following: "Wife.how she knew the time of murder exactly". This instance has true class "ham". The classifier predicted truthfully 0.00014 probability to be "spam".

Fig. 4 presents two different explanations for the classifier's prediction. As before, Fig. 4a shows the explanation provided by LioNets and Fig. 4b shows LIME's explanation. The contribution of feature "wife" to the prediction is assigned to different classes in each explanation. To prove the stability and robustness of LioNets, this feature is removed and by auditing again the neural network the new prediction is lower with a probability of 0.000095. Thus, it is clear that feature "wife" was indeed contributing to the "spam" class as LioNets explained and captured.



(a) (b)

Fig. 4: Explanation plots of SMS spam instance using LioNets(a) and LIME(b).

Like before, the neighbourhoods' distances from the original instance are computed and presented in Table 3. As it seems the neighbours generated by LIME on original space, by projecting them into the encoded space, are more distant to the encoded instance, compared to the neighbours generated by LioNets directly in the encoded space.

5 Conclusion

In summary, the LioNets architecture provides valid explanations for the decisions of a neural network that are comparable to other state-of-the-art techniques, while at the same time it guarantees better adjacency between the generated neighbours of an instance because the generation of the neighbours is

	Euclidean distance
LIME: Generated on Original Space	0.3184
LIME: Encoded	0.8068
LioNets: Generated on Encoded Space	0.3459
LioNets: Decoded to Original Space	0.7875

Table 3: Neighbourhood distances for instance of SMS spam collection.

performed on the penultimate layer of the network. In addition, LioNets can create better, larger and more representative neighbourhoods, because the generation process takes place at the encoded space, where the instance has a dense representation. These are the main points of creating and using LioNets on decision systems like neural networks.

One main disadvantage of LioNets is that it is focused only on explaining neural networks, thus it is not a model agnostic method. Moreover, the overall process of building LioNets is harder than training neural network predictors, because they demand the training of a decoder, which is a difficult task.

Future work plans include testing the LioNets methodology on different variations of encoders and decoders and implementing more complex neighbourhood generation and neighbours selection processes. In addition, we would like to explore different transparent models for explaining the instances, such as rule-based models [5], decision tree models [16, 4] and models based on abstract argumentation [6]. Lastly, we plan to evaluate LioNets based on human subject experiments.

Acknowledgment

This paper is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825619. AI4EU Project[3].

References

1. Adadi, A., Berrada, M.: Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access **6**, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052, https://ieeexplore.ieee.org/document/8466590/
2. Almeida, T.A., Hidalgo, J.M.G., Yamakami, A.: Contributions to the study of sms spam filtering: new collection and results. In: Proceedings of the 11th ACM symposium on Document engineering. pp. 259–262. ACM (2011)
3. Anagnostou, A., Mollas, I., Tsoumakas, G.: Hatebusters: A Web Application for Actively Reporting YouTube Hate Speech. In: IJCAI (2017), http://www.inach.net/index.php

[3] https://www.ai4eu.eu

4. Breiman, L., Friedman, J.H., Olshen, R., Stone, C.: Classification and regression trees (belmont, ca: Wadsworth international group). Biometrics **40**(3), 17–23 (1984)
5. Clark, P., Niblett, T.: The cn2 induction algorithm. Machine Learning **3**(4), 261–283 (Mar 1989). https://doi.org/10.1023/A:1022641700528, https://doi.org/10.1023/A:1022641700528
6. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence **77**(2), 321 – 357 (1995). https://doi.org/https://doi.org/10.1016/0004-3702(94)00041-X, http://www.sciencedirect.com/science/article/pii/000437029400041X
7. Frosst, N., Hinton, G.: Distilling a Neural Network Into a Soft Decision Tree. arXiv preprint arXiv:1711.09784 (11 2017), http://arxiv.org/abs/1711.09784
8. Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., Giannotti, F.: Local Rule-Based Explanations of Black Box Decision Systems. arXiv preprint arXiv:1805.10820 (5 2018), http://arxiv.org/abs/1805.10820
9. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys **51**(5), 1–42 (8 2018). https://doi.org/10.1145/3236009, http://dl.acm.org/citation.cfm?doid=3271482.3236009
10. Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531 (3 2015), http://arxiv.org/abs/1503.02531
11. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
12. Lampridis, O.P.: Explaining Sentiment Prediction by Generating Exemplars in the LatentSpace. Undergraduate thesis, Aristotle University of Thessaloniki, School of Informatics
13. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., Alsaadi, F.E.: A survey of deep neural network architectures and their applications. Neurocomputing **234**, 11–26 (4 2017). https://doi.org/10.1016/j.neucom.2016.12.038, https://linkinghub.elsevier.com/retrieve/pii/S0925231216315533
14. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
15. Porter, M.F.: Snowball: A language for stemming algorithms. Published online (October 2001), http://snowball.tartarus.org/texts/introduction.html, accessed 11.03.2008, 15.00h
16. Quinlan, J.R.: C4. 5: programs for machine learning. Elsevier (2014)
17. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. arXiv preprint arXiv:1401.4082 (2014)
18. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16 (2 2016), https://arxiv.org/abs/1602.04938
19. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-Precision Model-Agnostic Explanations. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018), www.aaai.org
20. Sadowski, P., Collado, J., Whiteson, D., Baldi, P.: Deep Learning, Dark Knowledge, and Dark Matter (8 2015), http://proceedings.mlr.press/v42/sado14.html
21. Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. Journal of documentation **28**(1), 11–21 (1972)