Ruprecht-Karls-University Heidelberg

Faculty of Mathematics and Computer Science

Seminar: Explainable Machine Learning

Interpreting Deep Classifiers
by Visual Distillation of Dark Knowledge

Author:
Daniela Schacherer
Lecturer:
Dr. Ullrich Koethe
Date
July 9, 2018

Contents

1 Introduction

1.1 Motivation and Related Work

Deep neural networks are currently the method of choice in machine learning and research areas that apply machine learning techniques like, for instance, image classification. Neural Networks achieve very convincing results, however, it is often hard to retrace how the decision came about. In order to get an insight into the neural network "black box" classifiers, various methods have been developed. One way to interpret neural networks is by means of compressing neural networks into simpler models e.g. a decision tree [W. Craven and W. Shavlik, 1999, Frosst and Hinton, 2017] which is referred to as knowledge distillation or model compression [Hinton et al., 2015, Bastani et al., 2017]. Other approaches for which [Olah et al., 2017] provides an overview try to visualize different layers that the network has learned (feature visualization) or depict by sensitivity maps how different parts of the input contribute to the output (attribution methods).

Xu *et al.* have recently developed a new method called DarkSight which can provide a visual interpretation of black-box classifier predictions [Xu et al., 2018]. Closely related to the work of Xu *et al.* is the t-distributed stochastic neighbour embedding (t-SNE)
[van der Maaten and Hinton, 2008] when applied to the features from the second to last layer in a deep classifier. In the following t-SNE is explained in more detail.

1.2 t-SNE

t-SNE is a technique for dimensionality reduction which is very widely used for the visualization of high-dimensional datasets, such as high-dimensional prediction vectors obtained by a complex neural network or another so-called black-box classifier. t-SNE was first introduced by van der Maaten and Hinton in 2008 [van der Maaten and Hinton, 2008]. The basic idea is to transform high-dimensional vectors $x_1, ..., x_n$ into low-dimensional vectors $y_1, ...y_n$ while keeping the relative similarity of all instances. The dimensionality of y_i is usually desired to be 2 such that $Y = \{y_i\} \in \mathcal{R}^2$ can easily be visualized by a scatter plot.

As mentioned before, t-SNE wants to keep the similarity between data points, i.e. if two vectors x_i, x_j are close together, the lower-dimensional vectors y_i, y_j should be close as well. Therefore, van der Maaten and Hinton define conditional probabilities between two data points considering a Gaussian distribution with given variance σ_i around each point x_i :

$$p_{j|i} = \frac{exp(\frac{-|x_i - x_j|^2}{2\sigma_i^2})}{\sum_{k \neq i} exp(\frac{-|x_i - x_k|^2}{2\sigma_i^2})} \tag{1}$$

The similarity is then defined as a symmetrized version of the conditional probability:

$$P_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{2}$$

with n being the total amount of data points. Doing this for every pair of data points results in a matrix P that represents similarities. It is also possible to compute a similar matrix Q for the low-dimensional counterparts $y_i, ..., y_n$, however, the Gaussian distribution is replaced by a t-student distribution. This is done in order to account for the fact that the distribution of distances is so different between a high-dimensional space and a low-dimensional space [van der Maaten and Hinton, 2008]. The similarity q_{ij} is thus defined as follows:

$$q_{ij} = \frac{exp(-|y_i - y_j|^2)}{\sum_{k \neq i} exp(-|y_i - y_k|^2)} \tag{3}$$

The distance between the two similarity matrices is minimized using stochastic gradient descent on the Kullback-Leibler divergence between P and Q:

$$KL(P||Q) = \sum_{i,j} p_{ij} log(\frac{p_{ij}}{q_{ij}}) \tag{4}$$

The final mapping is obtained when the gradient descent algorithm converges
[van der Maaten and Hinton, 2008]. Figure 1 shows a resulting t-SNE plot for the MNIST dataset.

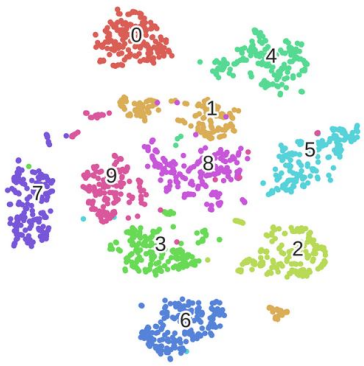


Figure 1: t-SNE visualization for the MNIST dataset.
Source: [Xu et al., 2018]

2 DarkSight

2.1 Principle

DarkSight is another recently published approach to visually summarize high-dimensional predictions of a black-box classifier in a lower-dimensional space (usually 2D). For this purpose, DarkSight combines model compression techniques with dimension reduction [Xu et al., 2018].

DarkSight is inspired by the concept of dark knowledge which refers to the idea that the full vector of predicted class probabilities - not just the highest probability - contains implicit knowledge that the classifier has learned. Therefore, consider the two class probability

vectors presented in Figure 2: It obviously makes sense that an image with associated probabilities `[cat:0.92, dog: 0.06, ...]` is somehow different from an image with probabilities `[cat:0.92, car:0.06, ...]`, since a car and a dog have a quite different appearance. Dark knowledge like this can be extracted using model compression techniques [Hinton et al., 2015].

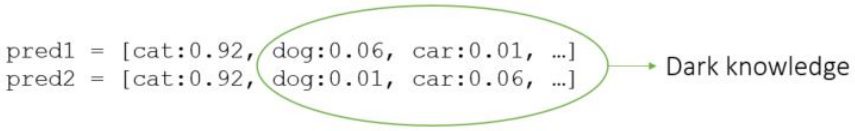


Figure 2: Dark Knowledge in vectors of predicted class probabilities.
Source: [Xu et al., 2018]

The DarkSight method expects to be given a trained classifier, called the *teacher*, as well as a validation dataset D_v. The *teacher* outputs a probability distribution $P_T(c|x)$ with $\pi_i = P_T(c_i|x_i)$ being the class probability vector for the data point x_i. The goal now is to represent every data point x_i by a low-dimensional embedding y_i i.e. to reduce the dimensionality. To do so, a simple and interpretable *student* classifier $P_S(c|y, \theta)$ is trained in the low-dimensional space such that the student's predictions match the teacher's predictions, i.e. $P_S(c_i|y_i, \theta) \approx \pi_i \forall i$. The training objective $L(Y, \theta)$ of the *student* is defined as the symmetric Kullback-Leibler divergence between the predictive distributions of *teacher* and *student*:

$$L(Y, \theta) = \frac{1}{N} \sum_{i=1}^{N} KL_{sym}(P_T(c_i|x_i), P_S(c_i|y_i; \theta)) \tag{5}$$

The symmetric KL divergence is defined from the unsymmetric version as:

$$KL_{sym}(P, Q) = \frac{1}{2}(KL(P, Q) + KL(Q, P)) \tag{6}$$

with the KL divergence given as:

$$KL(P, Q) = -\sum_{k=1}^{K} P(k) log(\frac{Q(k)}{P(k)}) \tag{7}$$

As the training objective does only depend on x_i via the prediction vectors π_i, the embeddings y_i can also be viewed as a lower-dimensional representation of π_i. The representations and the *student* classifier are trained end-to-end using stochastic gradient descent (SGD). This means that the *student* classifier's parameters θ and the inputs $Y = \{y_i\}$ to the *student* classifier are optimized simultaneously [Xu et al., 2018]. Figure 3 summarizes the DarkSight approach visually.

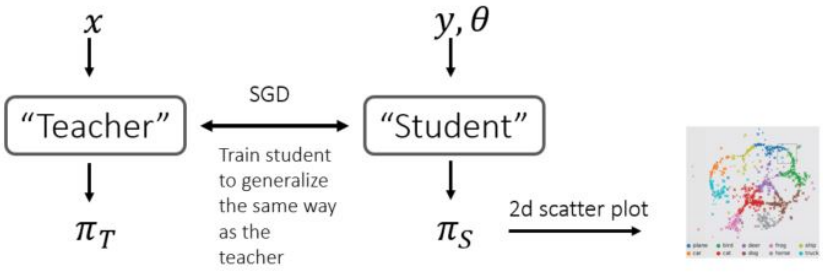


Figure 3: Schema of the DarkSight approach.

2.2 Choice of the Student Model

There exist different possibilities of how to define the *student* classifier. Xu *et al.* decided to use a very simple model namely the Naive Bayes classifier:

$$P_S(c_i = k)|y_i; \theta) = \frac{P(y_i|c_i = k; \theta_c)P(c_i = k; \theta_p)}{P(y_i|\theta)} \tag{8}$$

$P(y_i|c_i = k; \theta_c)$ can either be modelled by a Gaussian or a Student' t-distribution. Xu and his colleges decided for the latter due to the same reason as mentioned in section 1.2.
The prior $P(c_i = k; \theta_p)$ over the classes is described by a categorical distribution [Xu et al., 2018].

2.3 Confidence Measure

As a by-effect DarkSight allows for the definition of a new confidence measure. Intuitively, a full prediction vector which is unusual compared to all others, should not be trusted. This is measured quantitatively by density estimation in the space of the prediction vectors, however, this is computationally quite expensive. Xu *et al.* thus propose to perform kernel density estimation on the DarkSight embeddings $Y = \{y_i\}$. This yields an estimate $\hat{p}_{KDE}(y_i)$ that can be used as confidence measure for the *teacher's* prediction x_i. Commonly the predictive entropy

$$H(P_T(c_i|x_i) = \sum_k = p(c_i = k|x_i)logp(c_i = k|x_i) \tag{9}$$

is used to describe how reliable a prediction is, i.e. how possible it is that the *teacher* might have failed on classifying this data point. However, the predictive entropy does not take Dark Knowledge into account as it is invariant towards relabeling of classes. For instance, the prediction vectors `[cat:0.92, dog: 0.06, ...]` and `[cat:0.92, airplane: 0.06, ...]` have the same predictive entropy, while the corresponding $\hat{p}_{KDE}(y_i)$ will differ. The latter is what we expect, as we should have more trust in the first prediction vector, because a cat is more similar to a dog than to an airplane [Xu et al., 2018].

3 Design Principles

Xu *et al.* specified four properties, that low-dimensional embeddings like DarkSight and t-SNE should satisfy in order to provide a reliable basis for the interpretation of a classifier's results [Xu et al., 2018]

- **Cluster Preservation**: Points in the low-dimensional space are clustered according to the predicted class label. Additionally, the prediction confidence is highest in the cluster centers and monotonically decreases towards the cluster borders.
- **Global Fidelity**: The relative position of clusters towards each other has a meaning. From this follows that nearby clusters get confused more likely by the classifier than clusters which are located far away from each other.
- **Outlier Identification**: Data points with unusual predicted probability vectors are easily identifiable in the low-dimensional space as they are outliers.
- **Local Fidelity**: Points which are close to each other in the low-dimensional space are assigned similar probability vectors.

4 Experiments and Evaluation of Results

4.1 Experimental Setup

Xu *et al.* applied DarkSight to the predictions of three different *teacher* classifiers each trained on one dataset. Table 1 provides an overview, as well as the test accuracies of the *teacher* classifiers on the respective dataset.

Table 1: Overview of *Teacher* classifier, the datasets they were trained on and accuracy achieved.

Classifier	Dataset	Accuracy on Dataset (%)
LeNet	MNIST	98.23
VGG16	Cifar10	94.01
LeNet	Cifar100	79.23

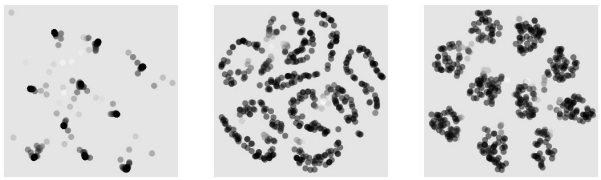
After verifying that the model compression works well [Xu et al., 2018], DarkSight was evaluated with respect to the four design principles (see section 3) and compared to the respective results of t-SNE. Thereby Xu and his colleges considered three different approaches, each of which applies t-SNE to different inputs:

- *t-SNE prob* uses the original predictive probability vectors.
- *t-SNE logit* uses logits of the predictive probablity vectors, i.e. the output of the last layer before applying the softmax.

- *t-SNE fc2* uses the final feature representations of the input, i.e. the output of the layer before logit.

4.2 Cluster Preservation

For a good cluster preservation we expect points close to the cluster center to have a higher confidence than points at the cluster border. Figure 4 shows scatter plots generated by DarkSight/t-SNE for predictions of LeNet on MNIST. The points are coloured by their confidence, which was here measured as the predictive entropy (see formula 9). Although the DarkSight clusters are quite small, one can observe that points with higher confidence are mostly located near the cluster center, whereas this can not be observed for the t-SNE visualizations.



(a) DarkSight (b) t-SNE prob (c) t-SNE logit

Figure 1. Scatter plots with points colored by predictive entropy. Dark points have large values. In all three plots, the same random subset (500 out of 10000) points is shown.

Figure 4: Scatterplots generated from DarkSight/t-SNE embeddings for predictions of LeNet on MNIST. Points are coloured by predictive entropy, with darker points having large values.

Source: [Xu et al., 2018]

From the concept of cluster preservation also follows that data from points between two clusters should be similar to both classes. Figure 5 shows the DarkSight plot of VGG16 predictions on the Cifar10 dataset. One can see that several clusters are directly adjacent along a curve. The points in the black box of Figure 5 form a transition of the bird class to the airplane class. In Figure 6 c) the predictive probabilities of the points within the black box are visualized and it can be seen that the values of the two top probabilities (bird and airplane) smoothly interchange with each other along the curve. Corresponding points for the t-SNE prob and t-SNE logit visualization are marked in Figure 6 a) and b) and show that such transition between two classes are hardly visible in t-SNE plots [Xu et al., 2018]. However, for a fair comparison one should have also taken adjacent points from a t-SNE plot and examine the values of the probability vectors in the same way as it was done for DarkSight.

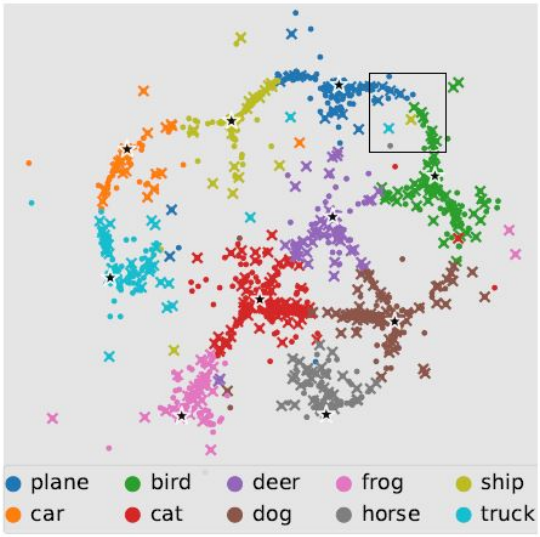


Figure 5: DarkSight visualization of VGG16 on Cifar10
Source: [Xu et al., 2018]

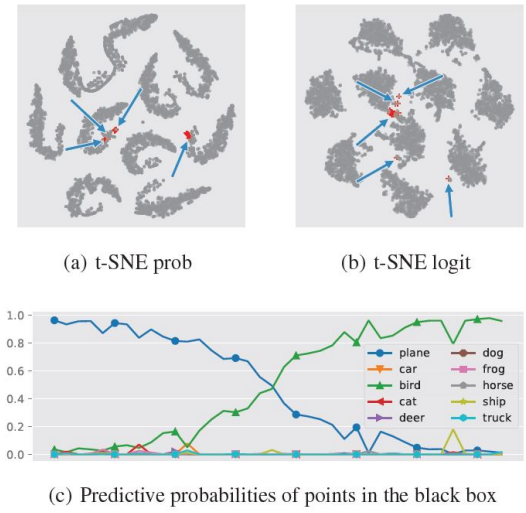


Figure 6: Interpretation of points in the black box of Figure 5
Source: [Xu et al., 2018]

Some images along the transition from plane to bird obtained from the online demo at `http://xuk.ai/darksight/demo/cifar.html` are shown in Figure 7 highlighting that at least in this case the decisions of the network are quite understandable for a human.

Figure 7: Real images from the transition within the black box of Figure 5 obtained from `http://xuk.ai/darksight/demo/cifar.html`.

4.3 Global Fidelity

If the principle of global fidelity is met, the global position of clusters in the low-dimensional space, i.e. in the DarkSight visualization, should be meaningful. By observing the confusion matrix Xu *et al.* could prove that nearby classes are often - but not always - confused by the classifier. This observation holds for DarkSight as well as t-SNE plots.
Referring to Figure 5 the DarkSight visualization provides a nicely interpretable global pattern: whereas all animal classes are located at the lower right, the remaining classes of vehicles can be found in the upper left corner. The two groups are only connected by the transition between birds and airplanes, which makes sense as these are semantically similar [Xu et al., 2018].

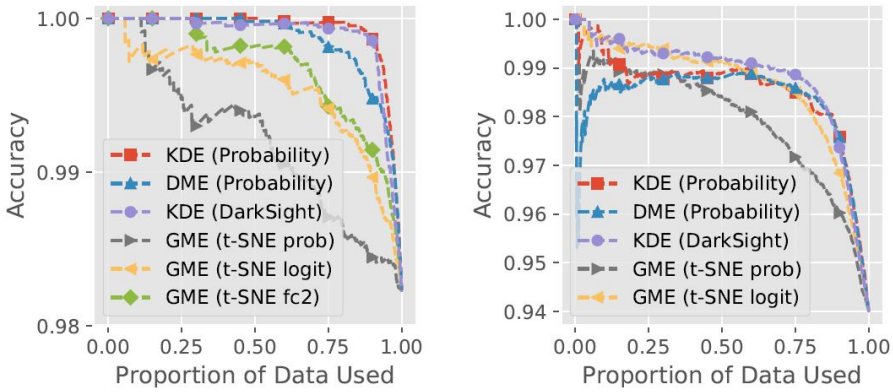
4.4 Outlier Identification

Furthermore, Xu *et al.* examined whether kernel density estimation on DarkSight embeddings (see section 2.3) is a suitable measure for the reliability of predictions and can thus be used as a confidence measure.

In general, a confidence measure is effective if the classifier is more accurate on predictions with high confidence and less accurate on predictions with low confidence. Thus, Xu and his colleges conducted the following experiment: first, they ran density estimation on DarkSight embeddings and t-SNE embeddings as well as - for comparison - in the original space of prediction vectors. For the latter, KDE and Dirichlet mixture estimation (DME) was employed. Then, the classifier was applied again on the validation dataset and performance was measured. However, this time the classifier was allowed to reject a point when confidence was below a predefined threshold δ, that is making a (possibly wrong) prediction without paying a penalty in accuracy. Figure 8 shows the prediction accuracy plotted against the proportion of data points the classifier was forced to make a prediction on (i.e. points with confidence $\leq \delta$). When the classifier is forced to make a prediction for all data points (proportion of data used is 100%) a decrease in accuracy due to low confidence points - more precisely points with confidence below δ - is expected. When not forced to predict for the total amount of data, the classifier should reject all the low-confidence points and thus not expose a decrease in accuracy. To summarize, the curves should be close to the upper right corner of the plot.

This expectation is met by KDE on DarkSight embeddings for MNIST as well as Cifar10. KDE outperforms all other confidence measures based on low-dimensional embeddings and is comparable to DME on the original probability space. This supports the statement of the authors that density estimation on DarkSight embeddings can serve as a reliable confidence measure. In the application this means that "outlier detection can be done by simply picking instances on the corner of the scatter plot" [Xu et al., 2018], as these relate to points with low confidence and thus the classifier may have failed on. Two random outlier examples together with their prediction vectors are presented in Figure 9 and 10 proving that outliers may indeed be instances which are quite hard to classify.

For the Cifar10 dataset KDE produces even better results than DME, which led the authors to the conclusion that DarkSight is possibly able to capture information which can not be captured by DME [Xu et al., 2018].



(a) MNIST (b) Cifar10

Figure 8: Accuracy-data plot for MNIST and Cifar10. KDE = Kernel Density Estimation, GME = Gaussian Mixture Estimation, DME = Dirichlet Mixture Estimation.
Source: [Xu et al., 2018]

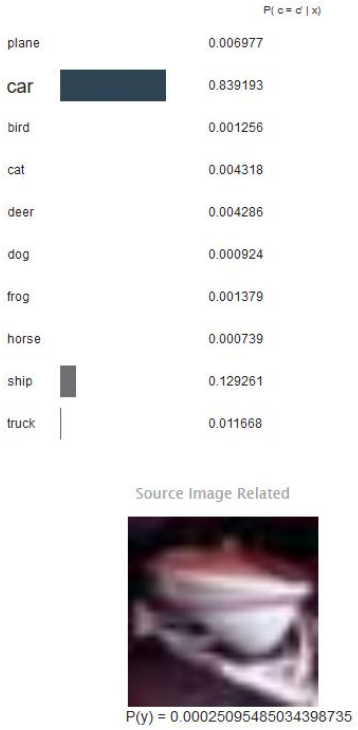


Figure 9: Real image outlier of the car cluster. Obtained from `http://xuk.ai/darksight/demo/cifar.html`.

Figure 10: Real image outlier of the dog cluster. Obtained from `http://xuk.ai/darksight/demo/cifar.html`.

4.5 Local Fidelity

In order to evaluate the local fidelity performance, Xu and colleges defined the following metric:

$$M_k(Y) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{k}\sum_{j\in NN_k(y_i)} JSD(p_i, p_j) \tag{10}$$

where p_i denotes the student classifier's prediction, JSD is the Jensen-Shannon distance, and $NN_k(y_i)$ is the set of k nearest neighbours to y_i in the visualization.

Figure 11 shows the results of $M_k(Y)$ on MNIST as function of the number of neighbours k. Since t-SNE is specifically optimized for local fidelity it naturally performs best in this task. Yet it is interesting that only t-SNE prob performs that good, whereas t-SNE logit and t-SNE fc2 produce rather bad results. This indicates that the t-SNE visualization depends on the quantities which are visualized. Whereas t-SNE logit and t-SNE fc2 process data from earlier layers of the network (i.e. layers with low-level features), t-SNE prob processes data from the last layer (with high-level features, containing more discriminative information). DarkSight can compete with t-SNE prob for higher k, however not for smaller k. The authors propose that this might be the case because the objective of DarkSight does globally rather than locally match the *teacher*'s with the *student*'s probability distribution [Xu et al., 2018].

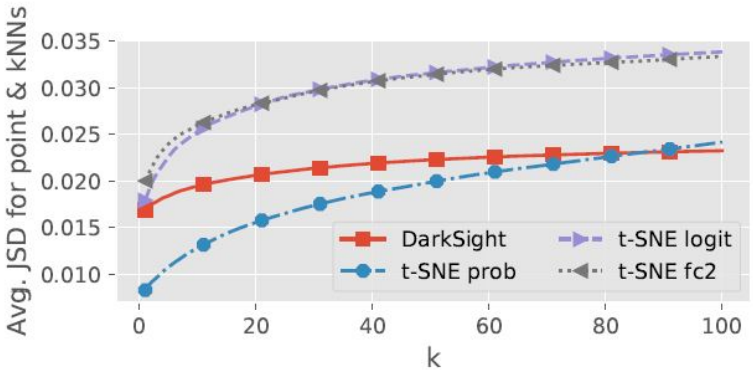


Figure 11: Local fidelity performance of t-SNE prob, t-SNE logit, t-SNE fc2 and DarkSight visualized as plot of average Jenson-Shannon distance (JSD) vs. the number of neighbours k.

Source: [Xu et al., 2018]

4.6 Case Study

In Figure 12 one can see a DarkSight visualization for LeNet on MNIST on which the design principles are satisfied and can be illustrated very well. Cluster preservation is visible as points from the cluster center (e.g. case 1.a. and 1.b) look very typical for the class whereas points at the border (e.g. case 1.c and 1.d) look more unusual and thus have lower confidence. Also, points at the transition of two classes look similar to both of them (compare Cases 2.a to 2.d). Global Fidelity is insofar fulfilled as nearby classes might get confused more easily as e.g. 3

and 8 or 4 and 9. Outliers (e.g. case 4.a, 4.b, 4.c) can be spotted as points far away from all clusters [Xu et al., 2018]. However, one has to keep in mind that it is not possible to capture all information of many dimensions in just two dimensions and especially it is hard to visualize all multi-dimensional relations in two dimensions. For example, $\pi_{1c} = [9 : 0.52, 0 : 0.44, ...]$ is expected to be located between cluster 5 and cluster 0, instead it can be found at the right border of the figure.

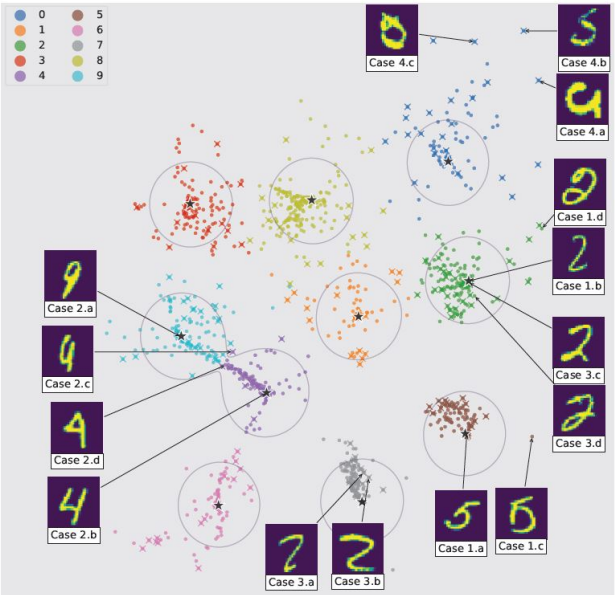


Figure 12: LeNet on MNIST. Data are coloured according to the predicted label.
Source: [Xu et al., 2018]

5 Discussion

With DarkSight, Xu *et al.* proposed a theoretically well comprehensible approach that converts predictions of a complex classifier into an easily interpretable visualization. They achieved this using a combination of model compression and dimension reduction. For evaluation of DarkSight they mainly focused on four properties, namely cluster preservation, global fidelity, outlier identification and local fidelity. Results in each category were compared with the outcome of the widely used t-SNE method.

Cluster preservation and Global Fidelity were mainly evaluated in a qualitative way with focus on specific examples. Therefore, one can not assume that similarly good results are produced for every classifier and dataset. Also, as mentioned in section 4.2, comparison with t-SNE could have been improved in some places. Apart from that, DarkSight offers an interesting and convincing way of presenting black-box classifier results in an understandable way from which one can draw further conclusions. Especially Figure 5 is very nicely interpretable. Therefore, when faced with such an issue I would personally try DarkSight as well as t-SNE and derive hypotheses from the combination of both.

6 References

[Bastani et al., 2017] Bastani, O., Kim, C., and Bastani, H. (2017). Interpretability via model extraction. *CoRR*, abs/1706.09773.

[Frosst and Hinton, 2017] Frosst, N. and Hinton, G. E. (2017). Distilling a neural network into a soft decision tree. *CoRR*, abs/1711.09784.

[Hinton et al., 2015] Hinton, G. E., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

[Olah et al., 2017] Olah, C., Mordvintsev, A., and Schubert, L. (2017). Feature visualization. *Distill*, 2(11).

[van der Maaten and Hinton, 2008] van der Maaten, L. and Hinton, G. (2008). Visualizing high-dimensional data using t-sne.

[W. Craven and W. Shavlik, 1999] W. Craven, M. and W. Shavlik, J. (1999). Extracting tree-structured representations of trained networks. 8.

[Xu et al., 2018] Xu, K., Park, D. H., Yi, C., and Sutton, C. A. (2018). Interpreting deep classifier by visual distillation of dark knowledge. *CoRR*, abs/1803.04042.