Interpreting Deep Classifiers by Visual Distillation of Dark Knowledge

Daniela Schacherer

Seminar: Explainable Machine Learning

Summer Term 2018

Content

1) Introduction
2) DarkSight
3) Experiments and Evaluation
4) Take Home

[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.

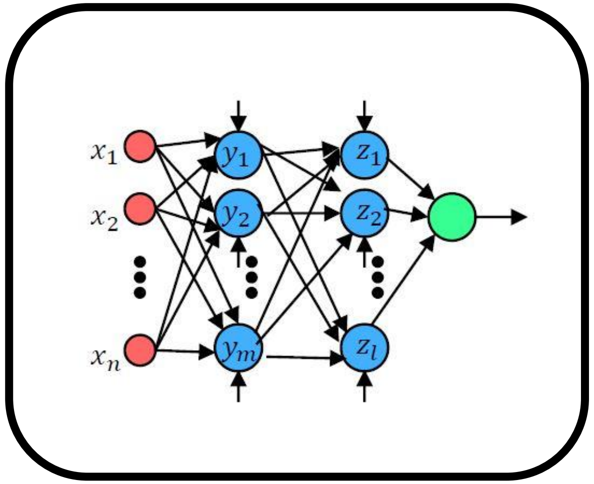
Introduction

Interpretability of Deep Classifiers

- Can we see what the network “sees”?

Black Box

Input



Decision

Related work

- Knowledge distillation/model compression
  - Training a simpler model to generalize in the same way as a more complex model
- Dimension reduction
  - Transformation of high-dimensional data into a lower-dimensional space
  - e.g. PCA, multidimensional scaling, t-SNE, …
- Interpreting deep networks
  - Compression of neural networks into more interpretable models like a decision tree
  - Interpretation via hidden activations or influence functions
- Visualizing deep networks
  - Feature visualization → visualize different layers learnt by the neural network (low-level features, mid-level features etc.)
  - Attribution methods → visualize how different parts of the input contribute to the final output (sensitivity maps)

t-SNE

- **t**-distributed **s**tochastic **n**eighbor **e**mbedding
- **Basic idea:** transform a list of high-dimensional vectors $x_i \dots x_n$ into a list of lower dimensional vectors $y_i \dots y_n$ (usually 2D) while **keeping the relative similarity of instances.**

- High dimensional space:

$$p_{i|j} = \frac{\exp\left(\left|x_i - x_j\right|^2 / 2\,\sigma_i^2\right)}{\sum_{k \neq i} \exp(|x_i - x_k|^2 / 2\,\sigma_i^2)} \rightarrow \quad p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

Gaussian distributed

- Lower-dimensional space:

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

t-Student distributed

t-SNE /2

- Minimize distance between the two similarity matrices using stochastic gradient descent

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

t-SNE /3

„We will observe that these (t-SNE) plots can be misleading because they contain well separated clusters even when, in fact, there are many points nearby the decision boundary"

Source: https://www.oreilly.com/learning/an-illustrated-introduction-to-the-t-sne-algorithm

DarkSight

“See what the network sees by performing dimension reduction and model compression jointly.”

Dark Knowledge

- Classifier that outputs probabilistic predictions

```
pred = [cat:0.92, dog:0.03, car:0.01, … ]
```

- Idea: full vector of class probability – not just the highest probability – contains implicit knowledge that the classifier has learned

```
pred1 = [cat:0.92, dog:0.06, car:0.01, …]
pred2 = [cat:0.92, dog:0.01, car:0.06, …]
```

Dark knowledge

- Dark knowledge can be extracted using model compression techniques

DarkSight

- **Intention:** visualize predictions of a black-box classifier in a lower-dimensional space
- **Given:**

Trained classifier „*Teacher*“
→ produces probability distribution $\mathrm{P}_{\mathrm{T}}(c|x)$

- prediction vector for x_i: $\pi_i = P_{\mathrm{T}}(c_i|x_i)$

Validation set
$D_V = \{(x_i, c_i)\}$

- **Task:** visually summarize predictions made by the teacher for D_V
- **Approach:** combine <u>dimension reduction</u> and <u>model compression</u>

DarkSight

a. **<u>Dimension reduction</u>:** represent each point x_i/prediction π_i in a lower-dimensional space (here: **2D**) as embedding y_i

b. **<u>Model Compression</u>:** train a simple and interpretable „*Student*“ classifier $P_S(\cdot\,|y;\theta)$ in the low-dimensional space, θ: classifier parameters

Aim: Student’s prediction vector should match the teacher’s prediction vector

$$\rightarrow P_S(\cdot\,|y_i;\theta) \approx \pi_i$$

→ optimize parameters of student classifier θ AND inputs of the student classifier $Y = \{y_i\}$ simultaneously

DarkSight – Objective

- We want to match the predictive distributions of teacher and student

$$L(Y, \theta) = \frac{1}{N} \sum_{i=1}^{N} D\big(P_{\mathrm{T}}(\cdot \mid x_i), P_S(\cdot \mid y_i; \theta)\big)$$

- Xu *et al.* empirically found that symmetric Kullback-Leibler divergence works best

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

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

DarkSight – „*Student*“ model

- „*Student*“ as Naive Bayes classifier:

$$P_S(c_i = k | y_i; \theta) = \frac{\boxed{P(y_i | c_i = k; \theta_c)}\,\boxed{P\left(c_i = k; \theta_p\right)}}{P(y_i | \theta)}$$

- Advantage:
  - models data from each class separately → embeddings are more likely to cluster well

- $\boxed{P(y_i | c_i = k; \theta_c)}$: non-centered Student‘s t-distribution $t_\nu(y_i | \mu_k, \Sigma_k)$
- Prior $\boxed{P\left(c_i = k; \theta_p\right)}$: Categorical distribution $\mathcal{C}at\left(c_i = k; \sigma\left(\theta_p\right)\right)$

DarkSight – Summary

- Assign low-dimensional representation to every data point x such that the simpler “student” classifier can mimic the complicated “teacher” model (and we get an output in 2D space).
- Representations y_i and interpretable classifier are trained end-to-end by stochastic gradient descent (SGD)

x

“Teacher”

π_T

SGD

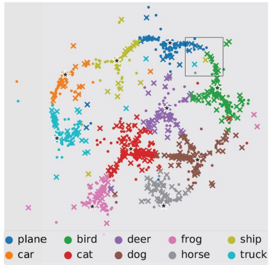
Train student to generalize the same way as the teacher

y, θ

“Student”

π_S

2d scatter plot



DarkSight – new confidence measure

- Byeffect: DarkSight allows for a new confidence measure:
  - **Intuition:** If full prediction vector is unusual compared to the others then we should not trust the prediction
  - But density estimation on full prediction vector space is expensive → better: density estimation on **embeddings**
  - **Formally:** Kernel density estimation $\hat{p}_{KDE}(y_i)$

  - Usually used: predictive entropy $H[P_T\,(c_i|x_i)] = \sum_k p(c_i = k|x_i) \log p(c_i = k|x_i)$ → but this does not take dark knowledge into account

π_1 `= [cat:0.95, dog:0.03, … ]`

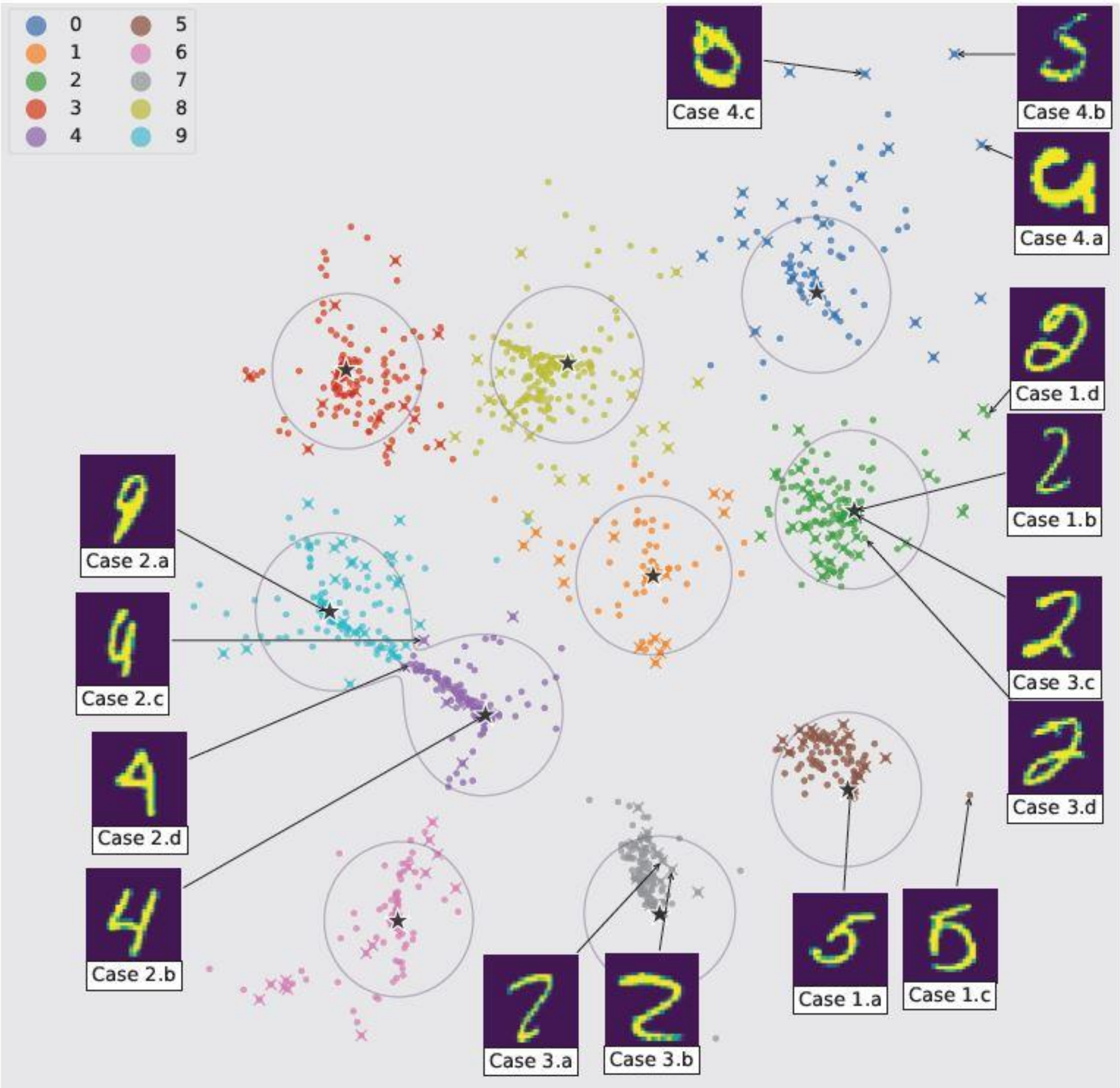
π_2 `= [cat:0.95, airplane:0.03, … ]`

Experiments and Evaluation

Design Principles

1. **Cluster Preservation:**
   - Points in the low-dimensional space are clustered by the predicted label
   - The prediction confidence of the classifier monotonically decreases from the cluster center to the outer borders of a cluster

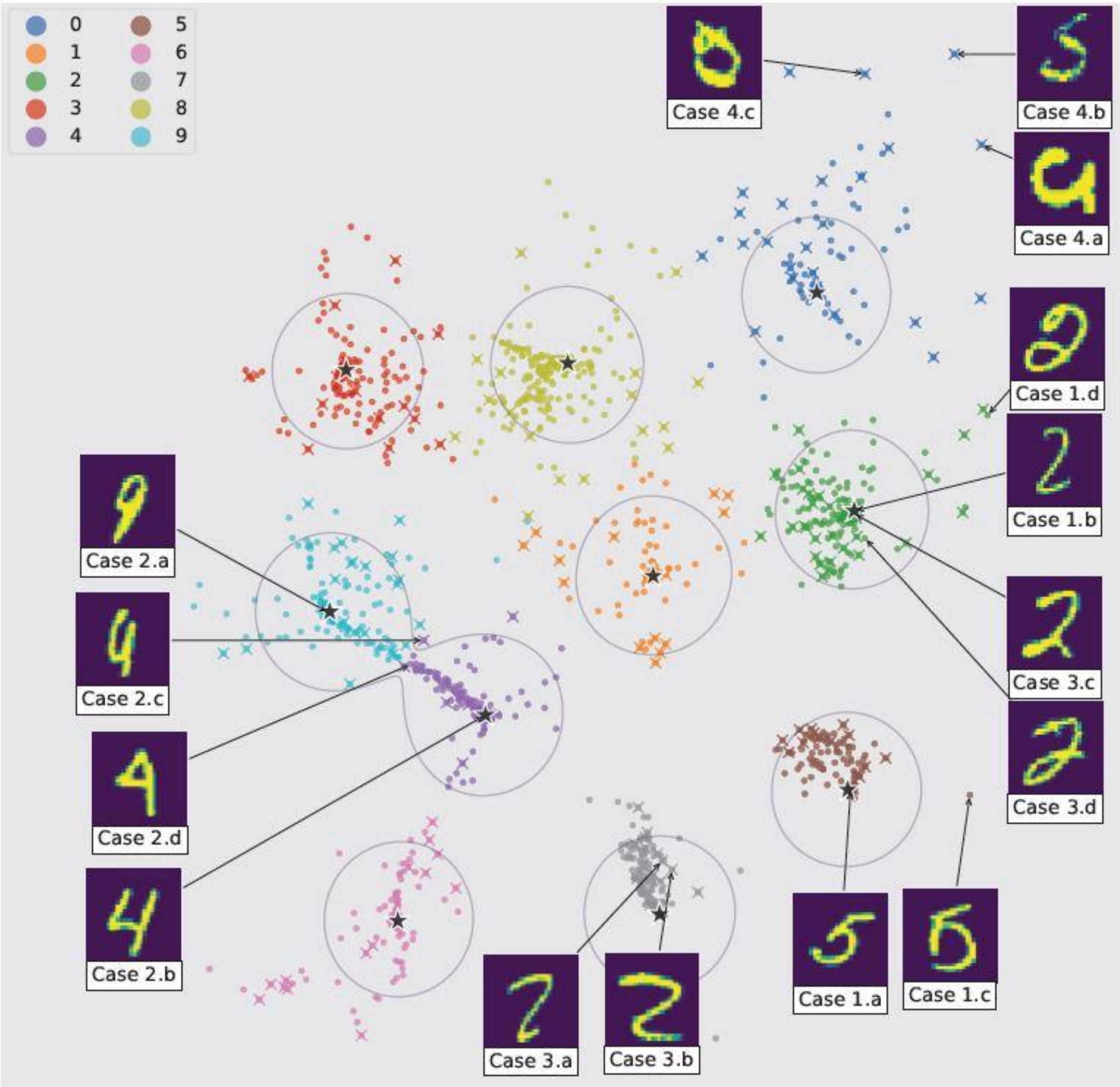
2. **Global Fidelity**
   - The relative position of clusters in the low-dimensional space is meaningful (nearby clusters get confused more likely)



Design Principles /2

3. **Outlier Identification**
   - Data points with a nontypical predicted probability vector are easy to find in the low-dimensional space

4. **Local Fidelity**
   - Points that are near to each other in the low-dimensional space have similar predicted probability vectors.



Experimental Setup

"Teacher" Classifier	Dataset	Test accuracy on dataset
LeNet	MNIST	98.23 %
VGG16	Cifar10	94.01 %
Wide-ResNet	Cifar100	79.23 %

- Comparison against *t-SNE*
  - *t-SNE prob:* uses the original predictive probability vectors
  - *t-SNE logit:* uses logits of predictive probability vectors = output of last layer before softmax
  - *t-SNE fc2:* uses final feature representations of the input = layer before logit

How does the model compression work?

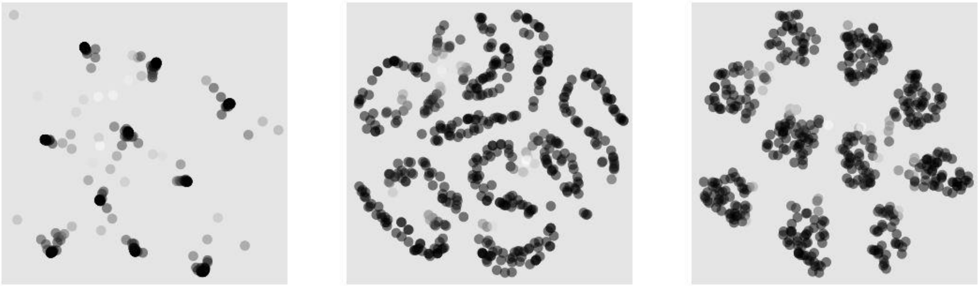
- How well can the student's model match the teacher's predictions? → Quality of model compression:

Table 2. Training results of DarkSight for different datasets. Note: Acc#ground is the accuracy towards true labels and Acc#teacher is the accuracy towards the predictions from the teacher.

DATASET	KL_{sym}	ACC#GROUND	ACC#TEACHER
MNIST	0.0703	98.2%	99.9%
CIFAR10	0.0246	94.0%	99.7%
CIFAR100	0.383	79.2%	99.9%

AccTeacher#Ground
98.23 %
94.01 %
79.23 %

Results on Cluster Preservation

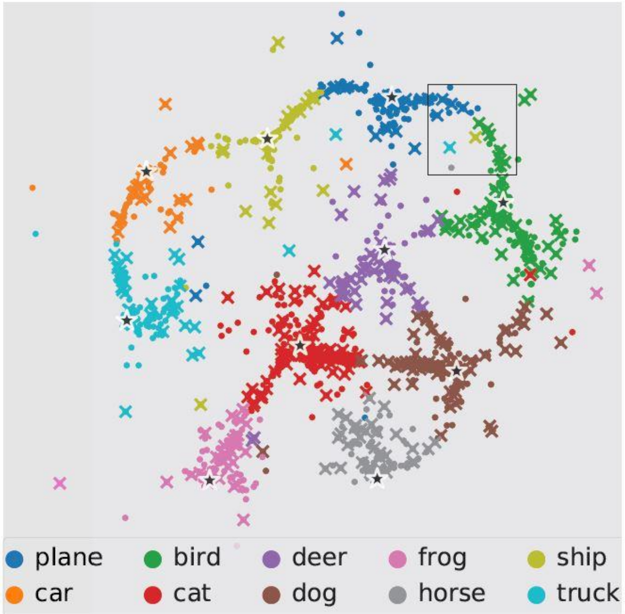


(a) DarkSight (b) t-SNE prob (c) t-SNE logit

Figure 1: Scatter plots generated by DarkSight/t-SNE for predictions of LeNet on MNIST. Points are coloured by predictive entropy. Dark points have large values. All plots show the same random subset (500 out of 10000 points).

- **Expected:**
  - points close to the cluster center have higher confidence than points at cluster edges
- **Observed:**
  - DarkSight
  - t-SNE spreads points with high predictive confidence all over the cluster

Results on Cluster Preservation /2



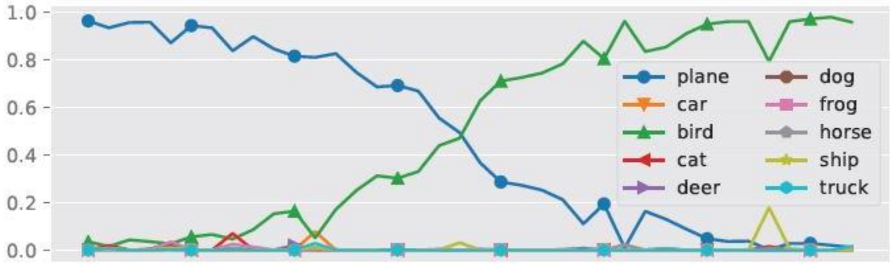
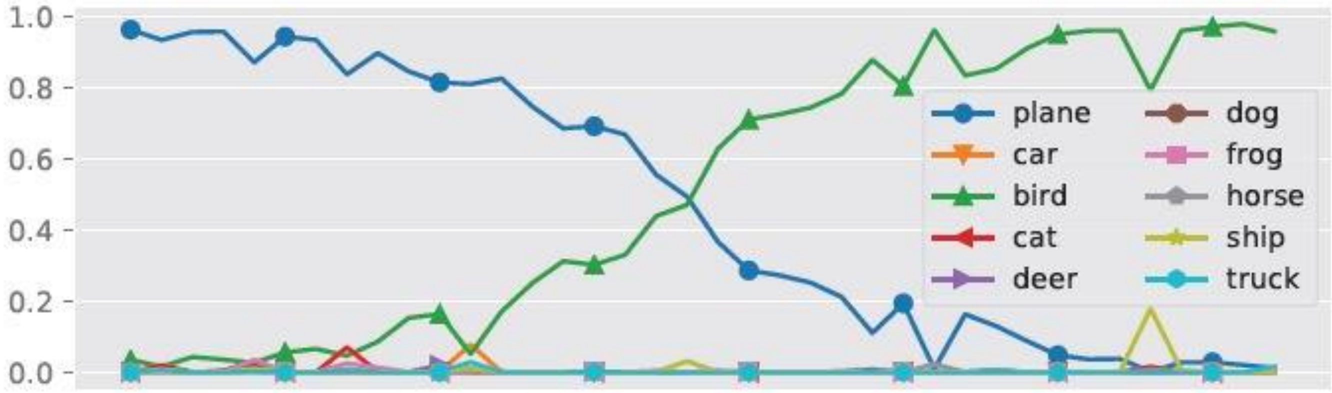


Figure 2: Top: Scatterplot by DarkSight for predictions of VGG16 on Cifar10. Bottom: Predictive probabilities of points in the black box of 2a)

- **Expected:**
  - Data from points between two clusters should look similar to both classes
- **Observed**:
  - Points in the box of Fig. 2a) form a transition from class blue to green → values of the two top probabilities in the prediction vector smoothly interchange with each other (see Fig. 2b).

Results on Cluster Preservation /2








Prediction: plane plane plane bird bird bird

Results on Cluster Preservation /3

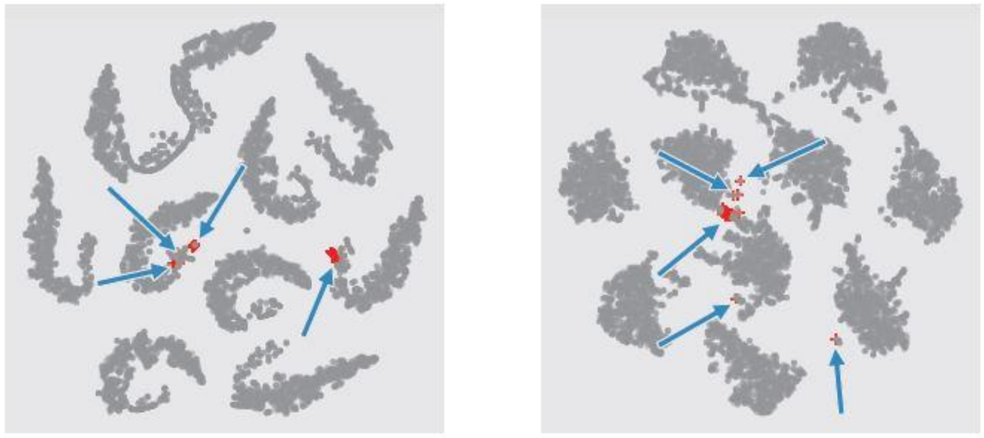


Figure 3: Scatterplot by t-SNE prob (left) and t-SNE logit (right) for predictions of VGC16 on Cifar10. Marked points correspond to the points in the box of upper image in Figure 2.

- **Expected:**
  - Data from points between two clusters should look similar to both classes
- **Observed**:
  - Points in the box of Fig. 2a) form a transition from class blue to green → values of the two top probabilities in the prediction vector smoothly interchange with each other (see Fig. 2b).
  - This can not be observed for the t-SNE visualization
  - DarkSight ✓ t-SNE ✗

Results on Global Fidelity

- **Expected:** global position of clusters in the low-dimensional space has a meaning
- **Observed** (based on the confusion matrix):
  - Both:
    - Classes which are close to each other are **often but not always** confused by the classifier
  - DarkSight shows global patterns:
    - Upper left: vehicles
    - Lower right: animals

  - DarkSight ✓✓ t-SNE ✓

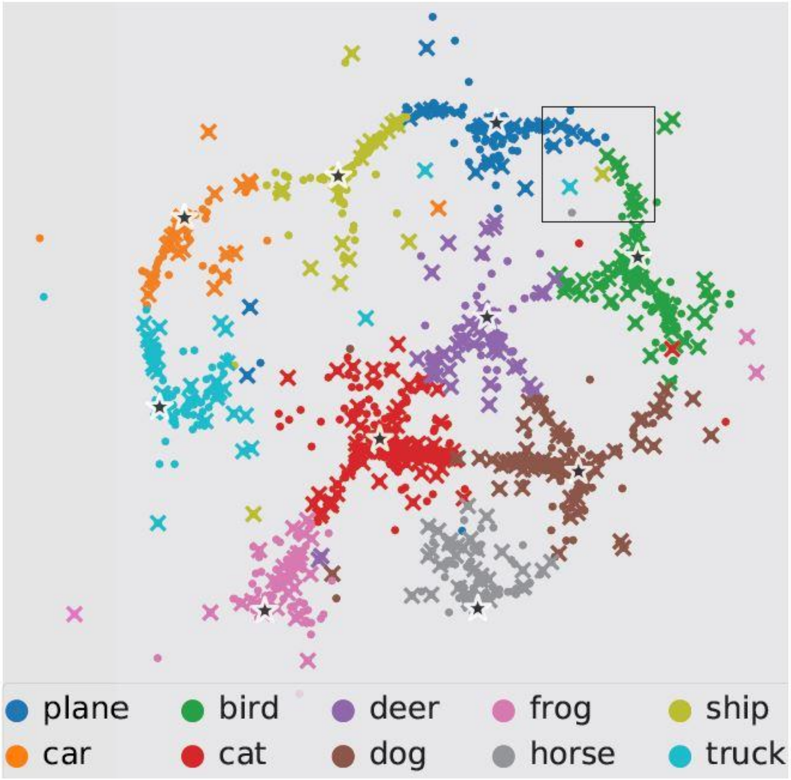


Figure 4: Scatterplot by DarkSight for predictions of VGC16 on Cifar10

Results on Outlier Identification

- **Expected:** Outliers in DarkSight visualizations correspond to points with less reliable predictions → DarkSight confidence is a good measure for reliability of predictions
- **Experiment:**
  - A confidence measure is effective if the classifier is more accurate on predictions with high confidence
  - First: run density estimation (KDE, GME…) on embeddings → confidence
  - Second: apply teacher classifier → when confidence $< \delta$ allow to reject that point without penalty on the performance

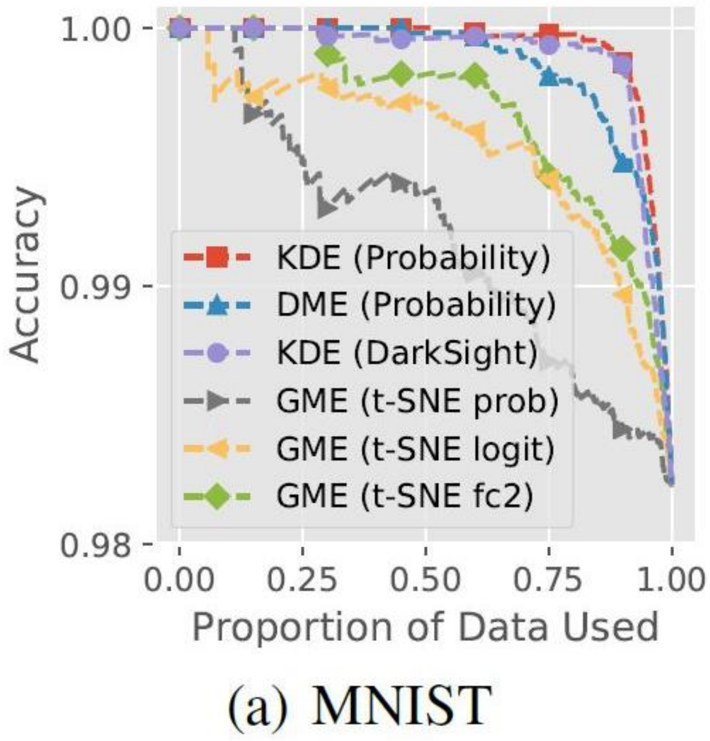


Figure 5: Data-Accuracy plot

Results on Outlier Identification /2

- **Observed:**
  - Density of DarkSight embeddings seems to be a more useful confidence measure than density of t-SNE embeddings

- DarkSight t-SNE
- “Outlier detection can be done by simply picking instances on the corner of the scatter plot or using a confidence measure based on density of DarkSight embedding”

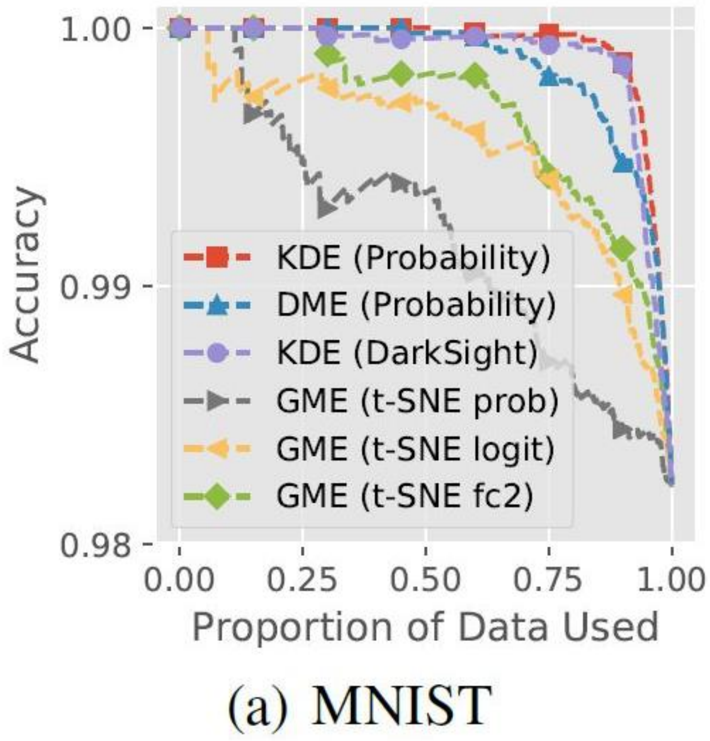


Figure 5: Data-Accuracy plot

Results on Outlier Identification /3

Outlier-car

Source Image Related

P(y) = 0.00025095485034398735

Outlier-car

Source Image Related

P(y) = 0.00003587365063140169

Outlier-dog

Source Image Related

P(y) = 0.00005148481432115659

Results on Local Fidelity

- **Expected:** predictive distributions of nearby points in the visualization are similar
- **Observed:**
  - t-SNE prob & DarkSight > t-SNE logit & t-SNE fc2
  - The performance of t-SNE seems to depend on the visualized quantities
  - t-SNE better for low k, Darksight for high k

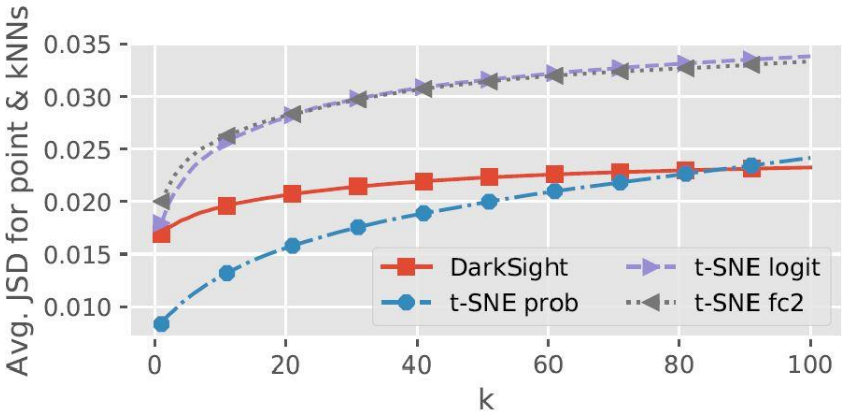


Figure 6: Local fidelity $M_k(Y)$ on MNIST as function of the number of neighbours k.

- DarkSight ✓ t-SNE ✓✓

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

- $p_i = P_S(\cdot\,|y_i)$
- JSD = Jensen-Shannon distance
- $NN_k(y_i)$: set of indices of k nearest neighbours of y_i in the 2D space

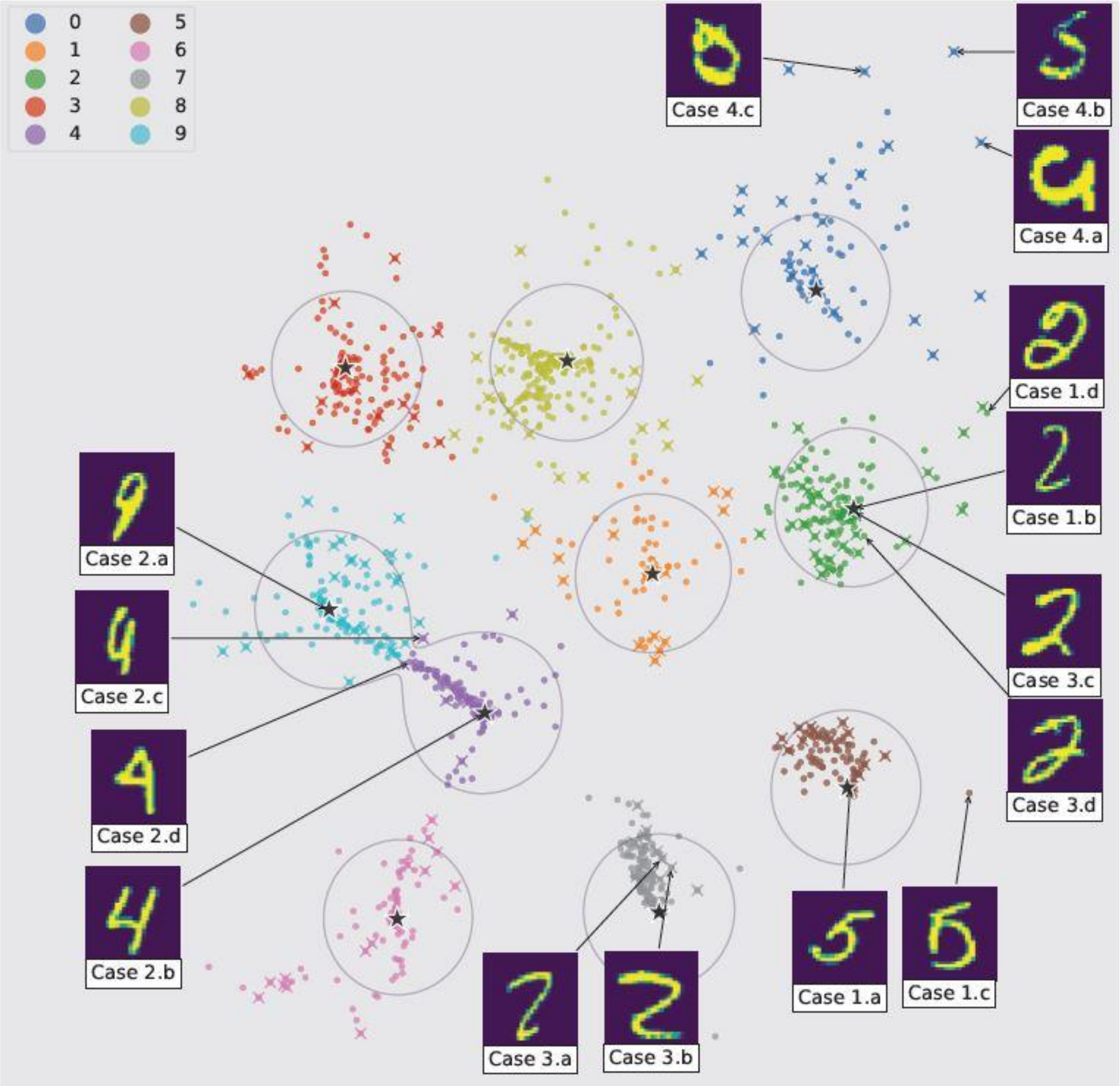
Summary

DarkSight: 6 x ✓

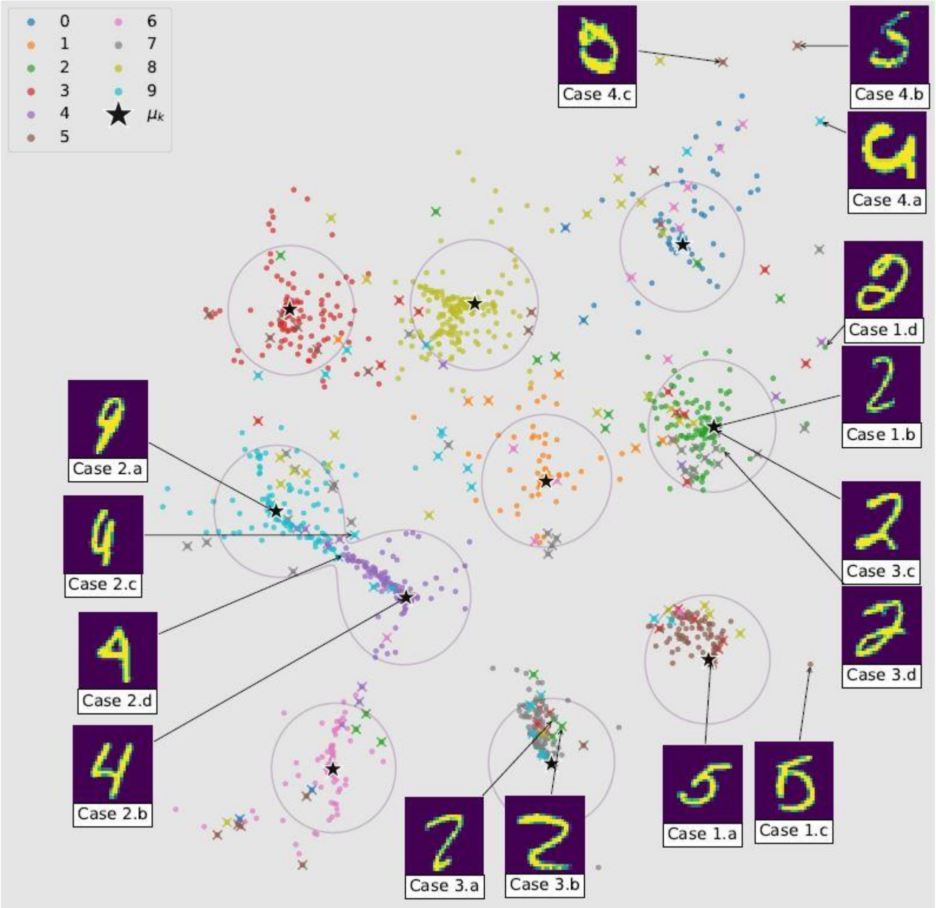
t-SNE: 3x ✓

Table 1. Comparisons between DrakSight and t-SNE. Properties are defined in Section 1.1. $\surd$: good, $\times$: bad and $\sim$: acceptable.

METHOD / PROPERTY	1	2	3	4	TIME
DARKSIGHT	$\surd$	$\surd$	$\surd$	$\sim$	$O(N)$
T-SNE PROB	$\sim$	$\times$	$\times$	$\surd$	$O(N^2)$ OR $O(N \log N)$
T-SNE LOGIT	$\times$	$\sim$	$\sim$	$\times$	
T-SNE FC2	$\times$	$\times$	$\times$	$\times$	



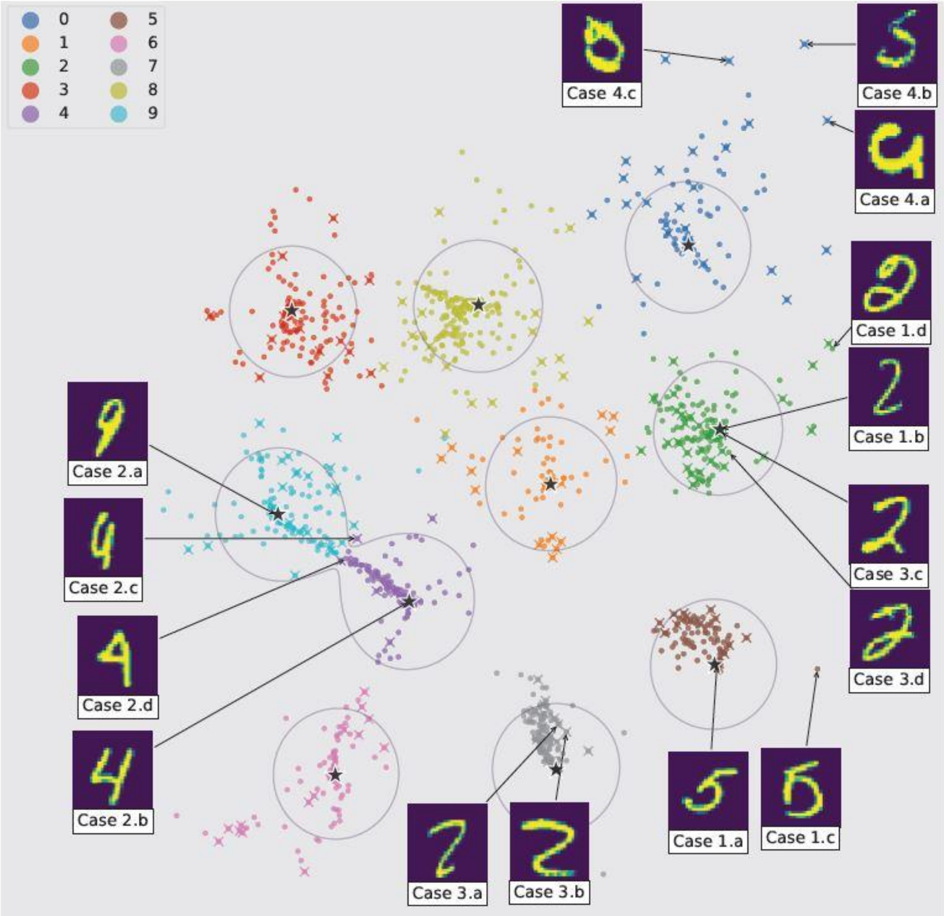
LeNet on MNIST – colored by predicted label



LeNet on MNIST – colored by true label

http://xuk.ai/darksight/demo/mnist.html

Limitations



LeNet on MNIST

π_{1c} = [5:0.52, 0:0.44, …]

π_{1d} = [9:0.12, 3:0.11, 8: 0.08 …]

Keep in mind that it is not possible

- to capture all the information of many dimensions in just 2 dimensions
- to visualize all multi-dimensional relations in 2 dimensions.

Take home

- DarkSight visualizes what the network sees by combining dimension reduction and model compression.
- Comparison against t-SNE proves that DarkSight provides additional useful information
- However, limitations have to be kept in mind: it is not possible to capture all the information of many dimensions in just 2 dimensions

References

- [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.
- Demo: http://xuk.ai/darksight/demo/mnist.html / http://xuk.ai/darksight/demo/cifar.html
- Homepage: http://xuk.ai/darksight/
- [van der Maaten and Hinton, 2008] van der Maaten, L. and Hinton, G. (2008). Visualizing high-dimensional data using t-sne.
- https://www.oreilly.com/learning/an-illustrated-introduction-to-the-t-sne-algorithm