# DL for Case Based Reasoning - Reproduction

Berend Jansen, Tom Lotze, Stan Lochtenberg, Cees Kaandorp

FACT in AI, Group 4

January 31, 2020

# Outline

# Transparency in black box models

- Decision making is inherently nontransparent in neural networks
- Visualizing model reasoning can give useful insights
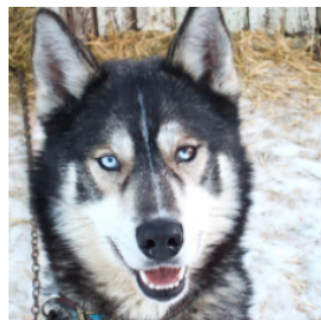


(a) Husky classified as wolf    (b) Explanation

Figure 1: Raw data and explanation of model prediction [4]

# Types of interpretability [2]

- Post hoc
  - LIME
  - SHAP
- Example-based
  - Counterfactual examples
  - Adversarial examples
  - Influential Instances
  - Prototypes

# Prototype based reasoning

▶ Classification based on similarity to prototypes in latent space

▶ Visualize prototypes to explain model

## Prototype

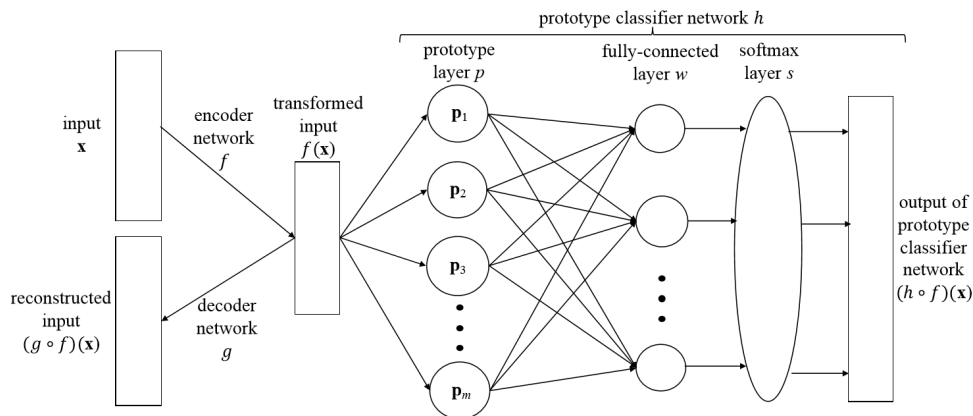Typical representation of the data, showing general characteristics.



Figure 2: Prototype classification model [1]

- 3 datasets:
  - MNIST handwritten digits
  - Car angles
  - Fashion MNIST

## Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions

Oscar Li[*1], Hao Liu[*3], Chaofan Chen[1], Cynthia Rudin[1,2]

[1]Department of Computer Science, Duke University, Durham, NC, USA 27708
[2]Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA 27708
[3]Kuang Yaming Honors School, Nanjing University, Nanjing, China, 210000
runliang.li@duke.edu, 141242059@smail.nju.edu.cn, {cfchen, cynthia}@cs.duke.edu

# Model: Cost function

$$L = \lambda_{class} * E + \lambda_{ae} * R + \lambda_1 * R1 + \lambda_2 * R2 \tag{1}$$

- $E$:  Cross Entropy loss
- $R$:  Reconstruction loss
- $R1$: Push prototypes to have meaningful decodings in pixel space
- $R2$: Cluster training examples around prototypes in latent space

# Methodology

- Conversion of code base from `TensorFlow` to `PyTorch`
  - Prototype model (`PM`) and Linear Model (`LM`)
- 4 different datasets (all 60,000 images)
  - Standard MNIST digits
  - CIFAR-10
  - MNIST colored (natural) background[1]
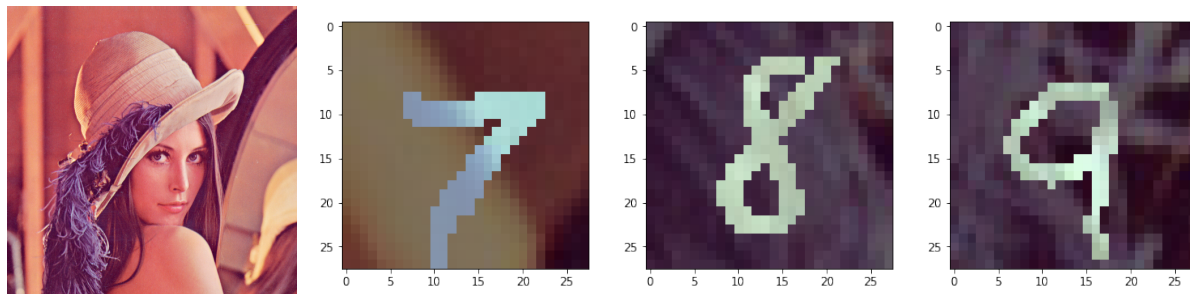  - MNIST gray (natural) background
- Grid search for optimal $\boldsymbol{\lambda}$ in colored & gray MNIST



Figure 3: Lena & Examples of colored MNIST dataset

---

[1]Inspired by `https://github.com/wouterbulten/deeplearning-resources`

# Results: Accuracy on different datasets

| Model on dataset | Accuracy | Nr. epochs |
|---|---|---|
| LM on MNIST color | 0.956 | 40 |
| LM on MNIST rgb2gray | 0.980 | 20 |
| Li et al. [1] on standard MNIST | 0.995 | 1500 |
| PM on standard MNIST | 0.991 | 20 |
| PM on MNIST color | 0.947 | 20 |
| PM on MNIST rgb2gray | 0.977 | 25 |
| PM on CIFAR-10 | 0.643 | 30 |

Table 1: Performance of the best model on different datasets. Accuracy is computed on hold-out test set. PM refers to the Prototype Model, LM refers to the Linear Model
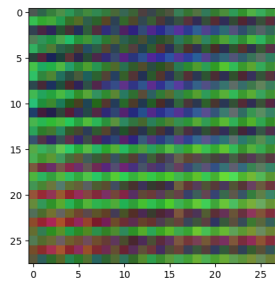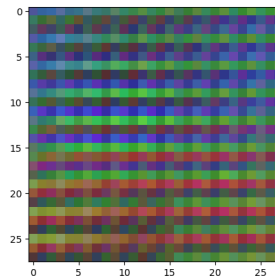
# Results: Prototype examples
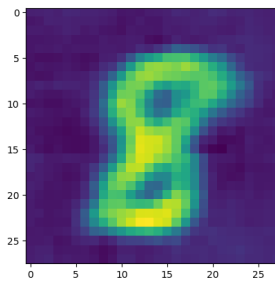
standard MNIST
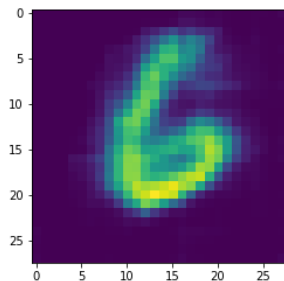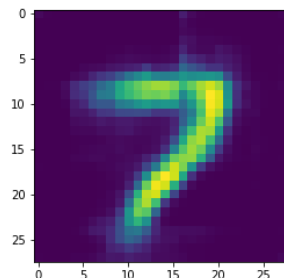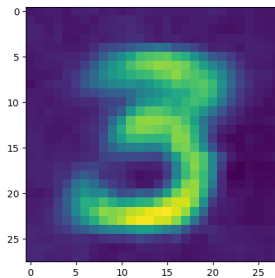
CIFAR-10
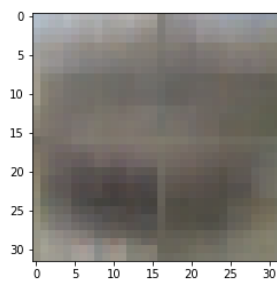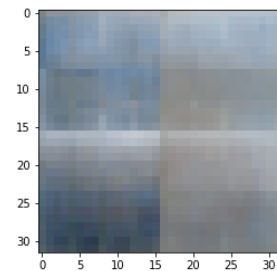
MNIST color

MNIST gray



Figure 4: Prototype examples per dataset

# Lambda grid search

- ▶ Effect on accuracy marginal
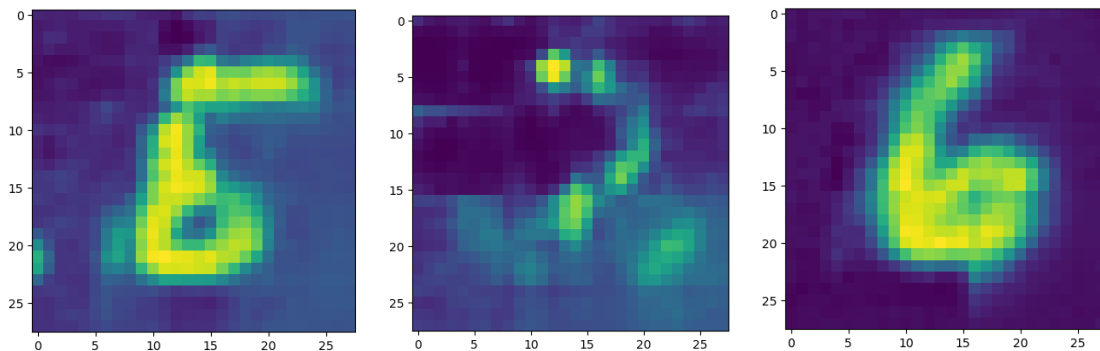- ▶ Effect on prototypes significant



Figure 5: Prototypes learned in grid search:
**Left**: $\lambda_{class} = 20$, $\lambda_{ae} = 10$, $\lambda_1 = 1$, $\lambda_2 = 10$;
**Middle**: $\lambda_{class} = 20$, $\lambda_{ae} = 10$, $\lambda_1 = 1$, $\lambda_2 = 20$;
**Right**: $\lambda_{class} = 20$, $\lambda_{ae} = 1$, $\lambda_1 = 10$, $\lambda_2 = 1$

# Conclusion

- More complicated backgrounds can still produce meaningful prototypes
- Colored images lead to problems
- Little to no intra-class variation required for meaningful prototypes
- Badge: Results Replicated
  - "The main results of the paper have been obtained in a subsequent study by a person or team other than the authors, using, in part, artifacts provided by the authors" [3]

# Discussion

- FACT Li et al. [1]
  - Discrepancies between code and original paper
  - Not for all datasets code available
- Limitations of our work
  - Color is (still) too complex for this architecture
  - Confidentiality: Prototypes contain information about training images.
    - Data removal requires retraining

# References

[1] Li, O., Liu, H., Chen, C., and Rudin, C. (2018). Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Thirty-Second AAAI Conference on Artificial Intelligence.*

[2] Molnar, C. (2019). *Interpretable Machine Learning.* https://christophm.github.io/interpretable-ml-book/.

[3] of Computing Machinery, A. (2018). *Artifact review and badging.*

[4] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.