DL for Case Based Reasoning - Reproduction

Berend Jansen, Tom Lotze, Stan Lochtenberg, Cees Kaandorp

FACT in AI, Group 4

January 30, 2020

Outline

1. Introduction
   - Transparency in AI
   - Prototype based reasoning
2. Methodology & Experiments
   - Datasets
3. Results
   - Accuracy
   - Prototypes
4. Conclusion
5. Discussion

Transparency in black box models

- Decision making is inherently nontransparent in Neural Nets
- Explaining a model's reasoning can expose problems:

(a) Husky classified as wolf (b) Explanation

Figure 1: Raw data and explanation of a bad model's prediction [Ribeiro et al., 2016]

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.

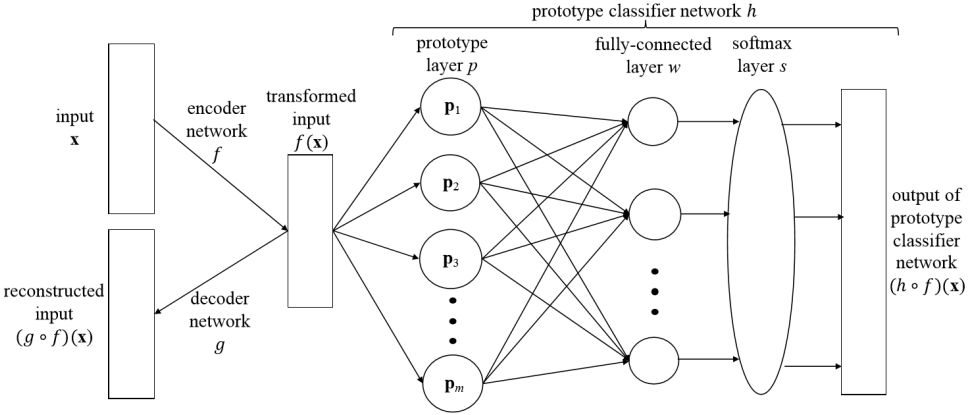

Figure 2: Prototype classification model [Li et al., 2018]

Model: Ensure proper functioning through 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 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
  - MNIST gray (natural) background
- Grid search for optimal $\boldsymbol{\lambda}$ in colored & gray MNIST

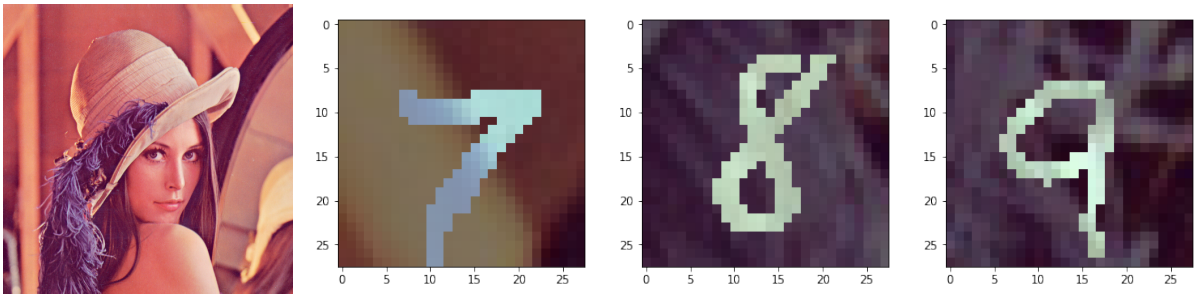

Figure 3: Lena & Examples of colored MNIST dataset

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
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

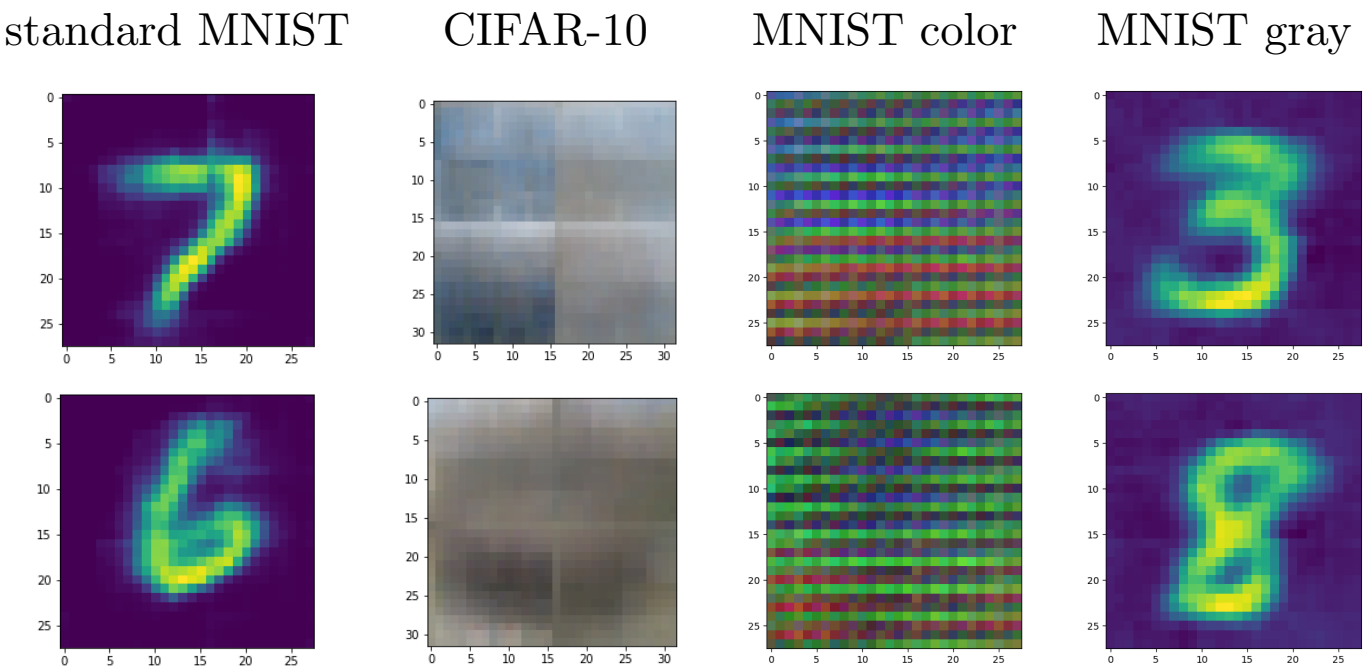

Figure 4: Prototype examples per dataset

Conclusion

- 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"
- Real backgrounds can be handles, but color is too complex
- Little intra-class variation required for meaningful prototypes

Discussion

- FACT original paper
  - 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

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*.

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.