Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions (Supplementary Material)

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

A Supplementary Material

A.1 Possible Techniques for Improving the Performance of Our Model

In our demonstrations we work with a basic network that demonstrates the main idea. If one wants to apply our model to more complex datasets, there are many techniques that can be adopted. We list a few of them below:

- Optimize the structure of the network. For instance, add more layers or using more powerful architectures such as ResNet (He et al. 2015), DenseNet (Huang, Liu, and Weinberger 2016), etc. to improve the accuracy.
- Add stronger regularizers to the training objective to make the prototype more realistic, such as using Generative Adversarial Networks (GAN) (Goodfellow et al. 2014). GAN has been used for improved visualization in posthoc analysis such as AM (Nguyen et al. 2016).

A.2 Weight Matrices

We show the learned weight matrices of Case Studies 2 and 3 in this section. The most negative weight connections are shaded for each prototype.

References

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2672–2680.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep residual learning for image recognition. *CoRR* abs/1512.03385.

Huang, G.; Liu, Z.; and Weinberger, K. Q. 2016. Densely connected convolutional networks. *CoRR* abs/1608.06993.

Nguyen, A. M.; Dosovitskiy, A.; Yosinski, J.; Brox, T.; and Clune, J. 2016. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *CoRR* abs/1605.09304.

*Contributed equally

	$-75°$	$-60°$	$-45°$	$-30°$	$-15°$	$0°$	$15°$	$30°$	$45°$	$60°$	$75°$
	1.33	0.45	0.01	0.70	-0.33	-2.11	0.12	0.61	-0.35	1.07	1.49
	1.69	-0.48	-2.93	0.07	1.39	0.63	0.86	1.92	0.86	1.09	0.50
	1.03	0.35	0.81	-0.21	0.54	-2.97	0.27	0.67	1.46	0.23	0.42
	1.36	1.15	1.62	0.65	0.39	0.76	0.33	-3.77	-0.13	0.78	1.05
	-0.88	0.46	1.80	1.46	1.55	1.93	1.20	0.80	2.25	-0.83	-3.01
	0.26	0.66	0.93	1.16	1.21	1.41	0.99	1.63	-0.44	-1.33	-1.09
	0.58	1.30	0.21	0.46	-3.93	0.68	0.76	0.68	0.86	0.57	1.42
	-2.09	-1.89	0.52	2.01	1.22	1.21	1.59	1.75	1.06	1.53	0.90
	0.17	1.14	0.83	1.29	1.39	0.99	1.51	0.57	-2.46	-1.26	1.19
	1.10	1.11	0.13	0.59	0.74	0.31	-4.29	0.05	0.05	1.17	0.90
	-0.02	0.26	0.68	-3.42	0.43	1.55	1.51	0.09	1.51	1.66	0.99

Table 4: Transposed weight matrix learned from Case Study 2 with both R_1 and R_2.

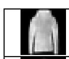
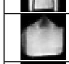
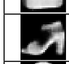
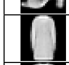
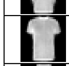
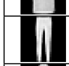
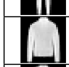
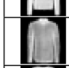
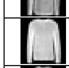
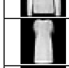
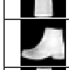
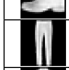
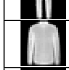
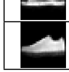
	T-shirt/top	Trouser	Pullover	Dress	Coat	Sandal	Shirt	Sneaker	Bag	Ankle boot
	-0.72	3.02	-11.00	1.72	-4.62	3.26	9.26	0.32	1.88	0.82
	0.18	4.20	5.24	-1.24	0.08	6.86	1.03	-2.11	-26.31	1.91
	3.16	0.30	2.78	3.69	0.78	-16.90	-3.51	3.12	5.05	-1.28
	4.50	6.88	3.39	-14.51	-2.76	5.59	-1.29	-1.44	-0.62	4.38
	-16.56	6.87	0.32	2.13	6.76	3.02	-5.43	4.58	-0.09	-0.17
	-0.92	-15.36	0.42	1.07	-1.11	-1.21	3.92	0.66	3.63	1.89
	-0.04	0.25	1.97	5.97	-14.13	0.91	0.90	2.36	1.15	3.38
	5.57	0.23	5.69	5.41	-2.03	-0.31	-12.96	2.48	2.76	3.23
	1.27	1.42	-11.92	1.09	7.31	3.07	-3.23	4.47	6.11	-1.23
	2.31	5.92	0.16	-17.12	3.65	5.38	-0.03	-0.53	1.21	-0.04
	-0.32	2.01	-0.97	-1.96	3.47	0.84	3.09	-5.15	-2.68	-24.82
	-0.97	-14.13	0.06	1.68	-1.30	-0.26	0.74	2.57	4.44	2.73
	1.73	1.06	0.40	-0.26	0.14	3.03	-5.41	2.63	-0.86	2.59
	0.15	0.38	2.89	1.21	2.41	-13.93	-1.93	1.10	2.04	0.58
	1.62	0.95	0.12	-2.27	0.28	1.45	4.55	-25.07	-4.31	-1.38

Table 5: Tranposed weight matrix learned in Case Study 3.