INTERPRETATIONS ARE USEFUL: PENALIZING EXPLANATIONS TO ALIGN NEURAL NETWORKS WITH PRIOR KNOWLEDGE

By Laura Rieger, Chandan Singh, W.James Murdoch, Bin Yu

03/05/2020

Presenter: Zijie Pan

https://qdata.github.io/deep2Read/

Motivation

- Explanation methods must provide both insights and suggest corresponding actions for a objective to be effective

- Often, methods are able to provide insights but have no way to take actions.

  - Ex. Models learned spurious correlations to achieve high accuracy
    Explanation method uncover the relationships but is unable to alter the model.

Background

Explanation Methods:
- Gradient based
- Decomposition based

Contextual decomposition(CD):

For a given DNN f(x), outputs can be represented as a SoftMax on composition of logits functions

$$f(x) = \text{SoftMax}(g(x)) = \text{SoftMax}(g_L(g_{L-1}(...(g_2(g_1(x))))))$$

and CD algorithm decomposes the logits g(x) into a sum of two terms. β captures the self-contribution while γ captures the interaction contributions

$$\beta(x) + \gamma(x) = g(x)$$

Claim / Target Task

- Using CD as the explanation functions
- Augment the prediction loss and explanation loss to make model learn correct predictions as well as correct explanations.
- propose contextual decomposition explanation penalization (CDEP), which penalizes the CD scores of features that a user does not want the model to learn to be important

Proposed Solution

Objective function:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \underbrace{\mathcal{L}\left(f_\theta(X), y\right)}_{\text{Prediction error}} + \lambda \underbrace{\mathcal{L}_{\text{expl}}\left(\text{expl}_\theta(X), \text{expl}_X\right)}_{\text{Explanation error}}$$

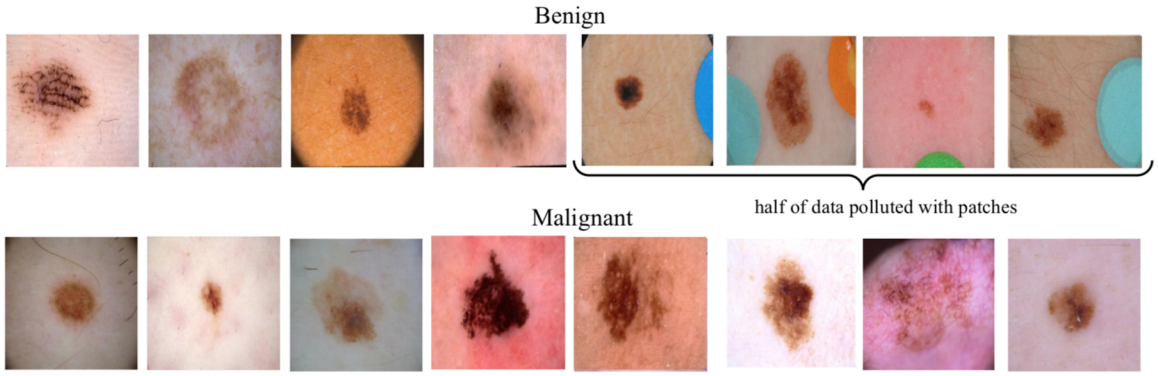
$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \underbrace{\sum_i \sum_c - y_{i,c} \log f_\theta(x_i)_c}_{\text{Classification error}} + \underbrace{\lambda \sum_i \sum_S ||\beta(x_{i,S}) - \text{expl}_{x_{i,S}}||_1}_{\text{Explanation error}}$$

Data Summary

- Ignore spurious feature (data bias):

  Cancer images from ISIC (International Skin Imaging Collaboration):

  Half benign images contains colorful patches, but none in the malignant images.



21,654 images (19,372 benign)

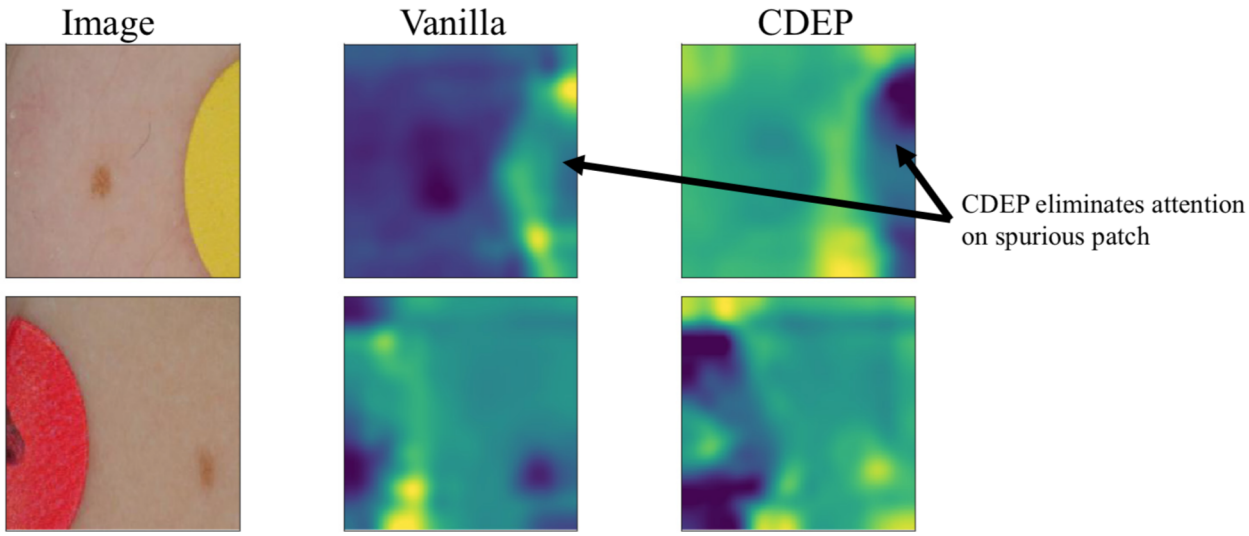
Model Architecture

- VGG16 architecture pre-trained on the ImageNet Classification task

- Freeze the weights of early layers so that only the fully connected layers are trained

Experimental Results

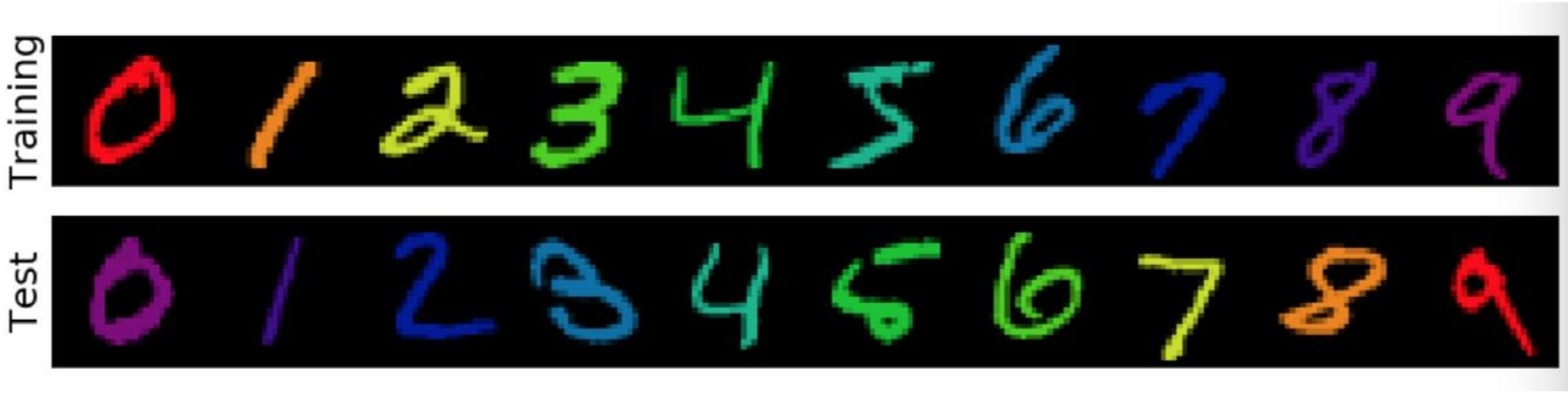
	AUC (no patches)	F1 (no patches)	AUC (all)	F1 (all)
Vanilla (excluded data)	0.86	0.59	0.92	0.59
Vanilla	0.85	0.56	0.92	0.56
With RRR	0.66	0.39	0.82	0.39
With CDEP	**0.88**	**0.61**	**0.93**	**0.61**



Experiment 2

Colored MNIST:

- Training: assign each class a distinct color
- Testing: each class will have a different color compared with training



	Unpenalized	CDEP	RRR	Expected Gradients
Test Accuracy	0.01 ± 0.2	25.5 ± 0.4	0.4 ± 0.2	0.4 ± 0.8

Experiment 3

Decoy MNIST:

DecoyMNIST adds a class-indicative gray patch to a random corner of the image

spurious features are not entangled with any other feature and are always at the same location

Table 3: Results on Grayscale Decoy set.

	Unpenalized	CDEP	RRR	Expected Gradients
Test accuracy	60.1 ± 5.1	97.2 ± 0.8	99.0 ± 1.0	97.8 ± 0.2
Run time/epoch (seconds)	4.7	17.1	11.2	821.0
Maximum GPU RAM usage (GB)	0.027	0.068	0.046	3.15

Experiment 4

SST dataset with spurious signals:

Injecting indicator words to each class at random positions.

Positive

pacino is the best **she**'s been in years and keener is marvelous

she showcases davies as a young woman of great charm , generosity and diplomacy

shows **she** 's back in form , with an astoundingly rich film .

proves once again that **she**'s the best brush in the business

Negative

green ruins every single scene **he**'s in, and the film, while it 's not completely wreaked, is seriously compromised by that

i'm sorry to say that this should seal the deal - arnold is not, nor will **he** be, back .

this is sandler running on empty , repeating what **he** 's already done way too often .

so howard appears to have had free rein to be as pretentious as **he** wanted

Results

	Unpenalized	CDEP
Random words	56.6 ± 5.8	$\mathbf{75.4 \pm 0.9}$
Biased (articles)	57.8 ± 0.8	$\mathbf{68.2 \pm 0.8}$
Biased (gender)	64.2 ± 3.1	$\mathbf{78.0 \pm 3.0}$

Experiments above only compared the training results of unpenalized model and CDEP model. It is better to show the results on unpolluted data to have concrete evidence (to show unpenalized model performance actually drops)

Conclusion and Future Work

- CDEP can penalize complex features and feature interactions.
- CDEP is more computationally efficient than previous work and does not rely on backpropagation, enabling its use with more complex neural networks.
- CDEP can be used to remove bias and improve predictive accuracy on a variety of toy and real data.