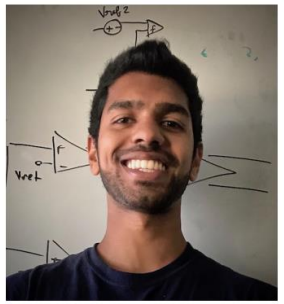
Beyond learning with labels

Laura **Ri**eger (lauri@dtu.dk)

Beyond learning with labels



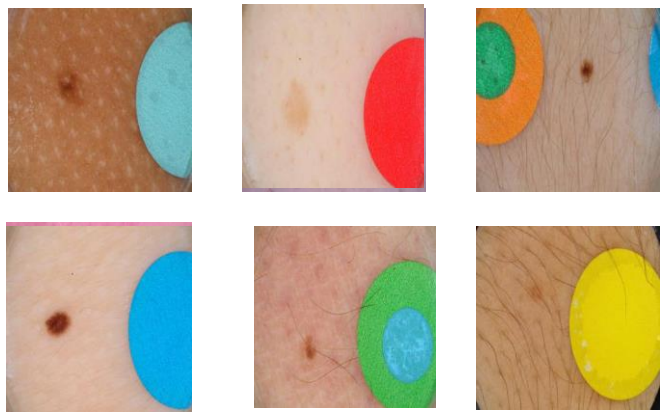
Rieger, Singh, Murdoch, Yu (2019).
Interpretations are useful: penalizing explanations to align neural networks with prior knowledge

In submission at ICLR

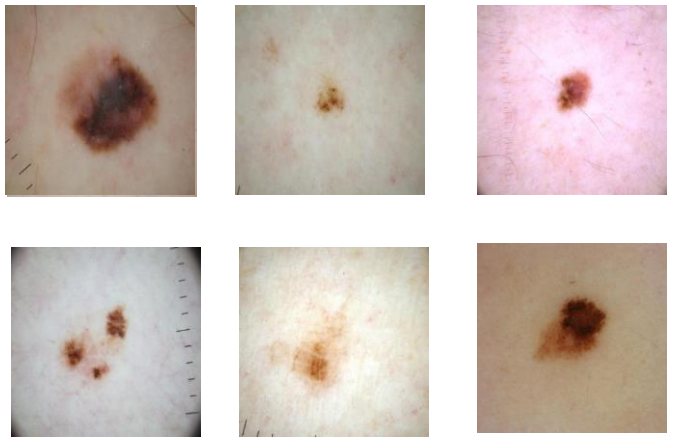
Are labels enough?

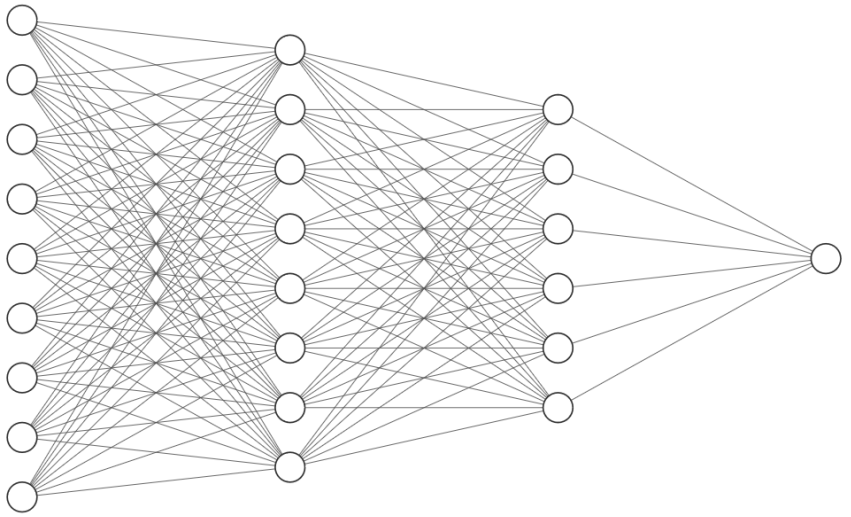
Learning from labels (step by step)

Benign



Cancerous

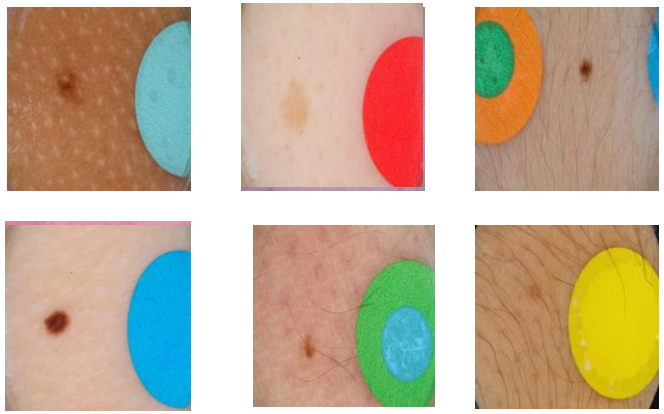




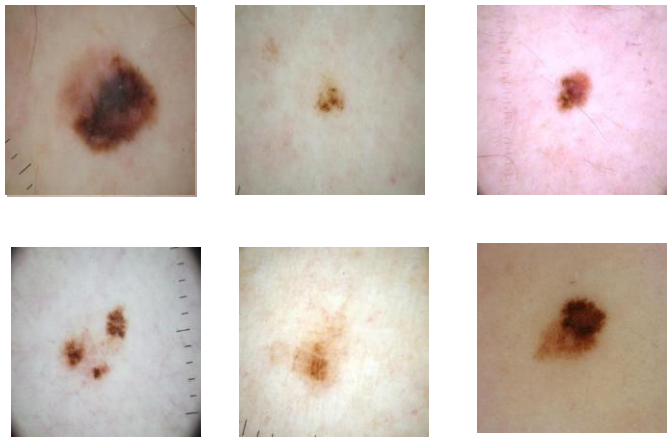
☺ **90% accurate**

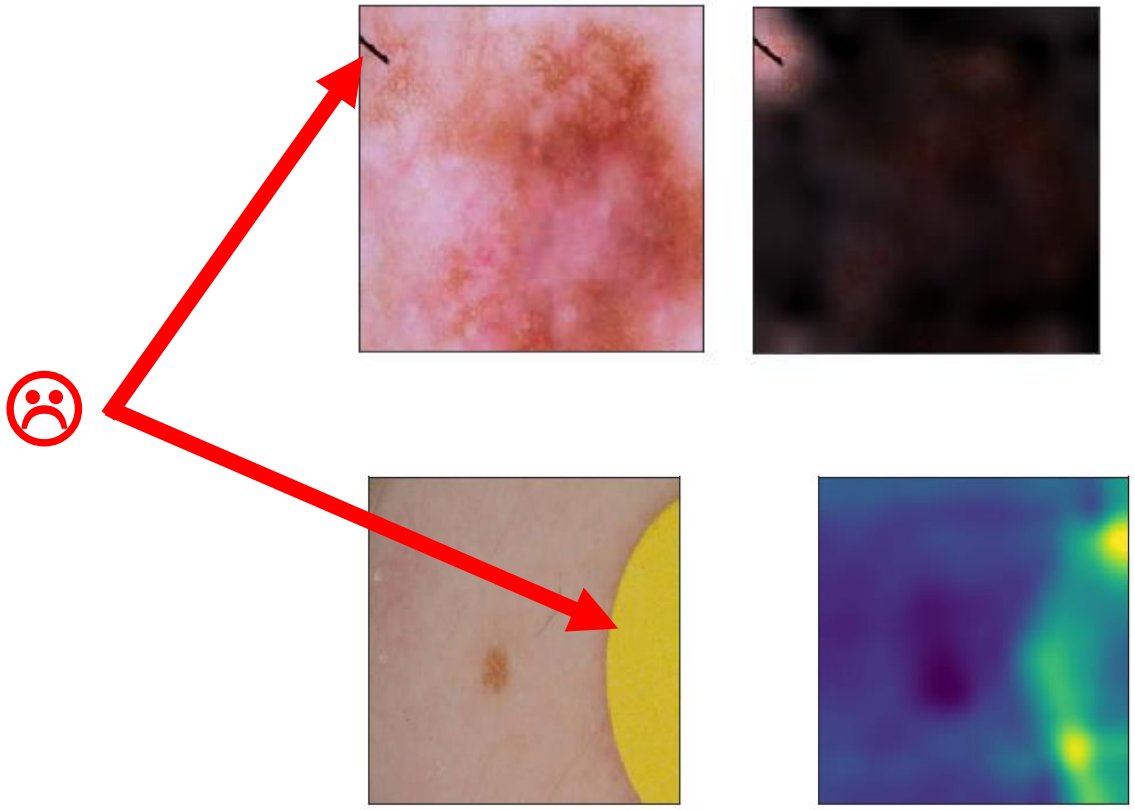
What did the network learn?

Benign



Cancerous





We know the bias (sometimes)

Gender is not important for job applications!

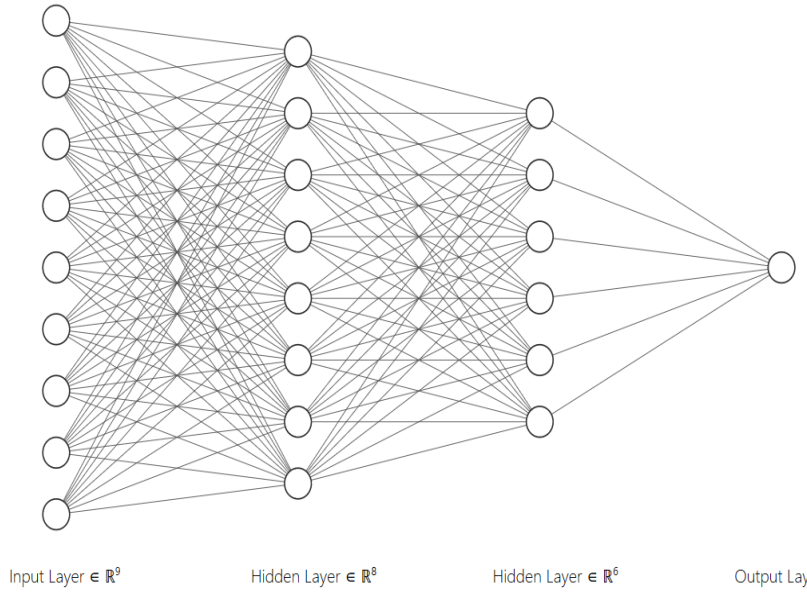
Race shouldn’t determine jail time!

Rulers aren’t cancerous!

Band aids don’t protect against cancer!

Regularize with prior knowledge

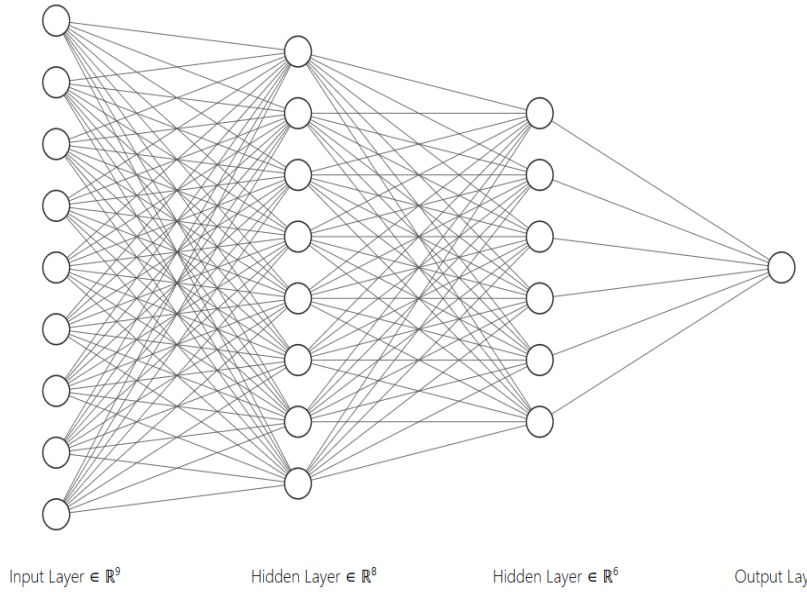
Augmenting the loss function



Prediction ← **True label**

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \mathcal{L}\left(f_\theta(X), y\right)$$

Augmenting the loss function



Prediction ← **True label**

Explanation ← **Prior knowledge**

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

Contextual **D**ecomposition **E**xplanation **P**enalty

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

Any differentiable explanation method works

We used Contextual Decomposition [1]

… skipping the math part here ….

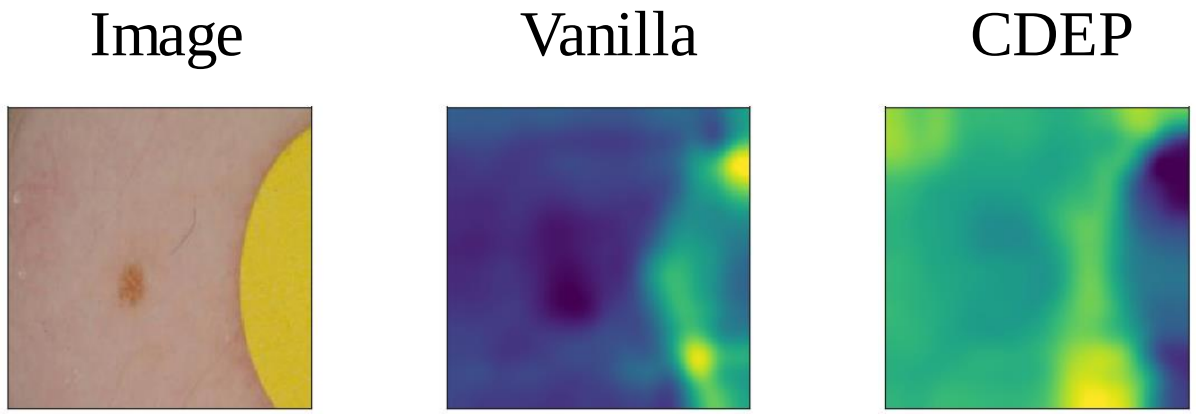
[1] Singh, Chandan, W. James Murdoch, and Bin Yu. "Hierarchical interpretations for neural network predictions."

Does it work?

Using CDEP improves accuracy

Test F1: 0.57 -> **0.62**

Saliency makes more sense



Text bias *57% -> **72%***

Biased MNIST *0% -> **31%***

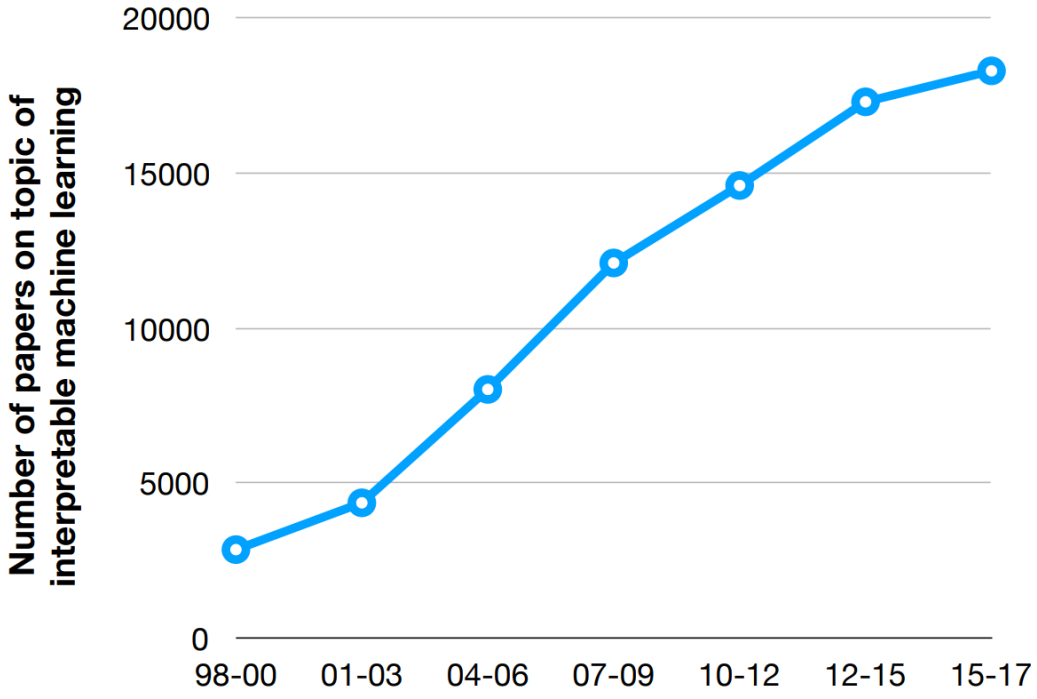
Conclusion

Conclusion

Interpretability is a growing field

Can be used for more than post-hoc analysis!

More research needed for more complicated priors



https://beenkim.github.io/slides/DLSS2018Vector_Been.pdf