Interpretability in Machine Learning

Tameem Adel

Machine Learning Group
University of Cambridge, UK

February 12, 2019

Growth of ML

ML algorithms optimized:

- Not only for task performance, e.g. accuracy.
- But also other criteria, e.g. safety, fairness, providing the right to explanation.
- There are often trade-offs among these goals.

However,

- Accuracy can be quantified.
- Not precisely the case for the other criteria.

What is interpretability?

- Interpret means to explain or to present in understandable terms.
- In the ML context: The ability to explain or to present in understandable terms to humans.
- What constitutes an explanation? What makes some explanations better than others? How are explanations generated? When are explanations sought?
- Automatic ways to generate and, to some extent, evaluate interpretability.

Taxonomy

Task-related:

- Global interpretability: A general understanding of how the system is working as a whole, and of the patterns present in the data.
- Local interpretability: Providing an explanation of a particular prediction or decision.

Method-related (what are the basic units of the explanation?):

- Raw features.
- Derived features that have some semantic meaning to the expert.
- Prototypes.

The nature of the data/tasks should match the type of the explanation.

Visualizing Deep Neural Network Decisions: Prediction Difference Analysis

Zintgraf, Cohen, Adel, Welling, ICLR 2017

Idea

- Visualize the response of a deep neural network to a specific input.

- For an individual classifier prediction, assign each feature *a relevance value* reflecting its contribution towards or against the predicted class.

Visualizing deep networks

- Looking under the hood: explaining why a decision was made.
- Can help to understand strengths and limitations of a model, help to improve it [wolves/huskies based on presence/absence of snow].

- Important for liability: why does the algorithm decide this patient has Alzheimer?
- Can lead to new insights and theories in poorly understood domains.

Approach

- Relevance of a feature x_i can be estimated by measuring how the prediction changes if the feature is *unknown*.
- The difference between $p(c|\mathbf{x})$ and $p(c|\mathbf{x}_{\setminus i})$, where $\mathbf{x}_{\setminus i}$ denotes the set of all input features except x_i.
- But how would a classifier recognize a feature as *unknown*?
  - Label the feature as unknown.
  - Retrain the classifier with the feature left out.
  - Marginalize the feature.

Marginalization of a feature

$$p(c|\mathbf{x}_{\setminus i}) = \sum_{x_i} p(x_i|\mathbf{x}_{\setminus i})p(c|\mathbf{x}_{\setminus i}, x_i) \tag{1}$$

Assume x_i is independent of $\mathbf{x}_{\setminus i}$

$$p(c|\mathbf{x}_{\setminus i}) \approx \sum_{x_i} p(x_i)p(c|\mathbf{x}_{\setminus i}, x_i) \tag{2}$$

Weight of evidence

UNIVERSITY OF CAMBRIDGE

Compare $p(c|\mathbf{x}_{\setminus i})$ to $p(c|\mathbf{x})$:

$$\text{odds}(c|\mathbf{x}) = \frac{p(c|\mathbf{x})}{(1-p(c|\mathbf{x}))}$$

$$\text{WE}_i(c|\mathbf{x}) = \log_2(\text{odds}(c|\mathbf{x})) - \log_2\left(\text{odds}(c|\mathbf{x}_{\setminus i})\right), \qquad (3)$$

- A large prediction difference $\rightarrow$ the feature contributed substantially to the classification.
- A small prediction difference $\rightarrow$ the feature was not important for the decision.
- A positive value WE_i $\rightarrow$ the feature has contributed evidence *for* the class of interest.
- A negative value WE_i $\rightarrow$ the feature displays evidence *against* the class.

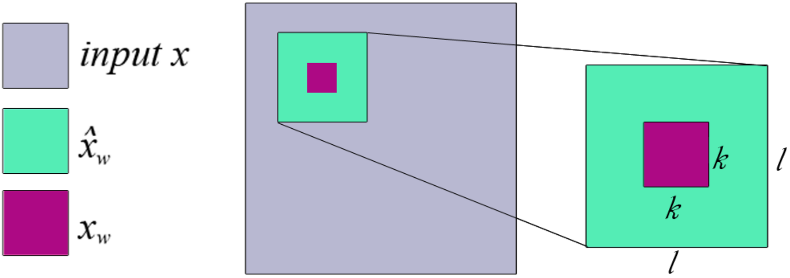
Conditional sampling

- A pixel depends most strongly on a small neighbourhood around it.
- The conditional of a pixel given its neighbourhood does not depend on the position of the pixel in the image.

$$p(x_i|\mathbf{x}_{\setminus i}) \approx p(x_i|\hat{\mathbf{x}}_{\setminus i}) \tag{4}$$

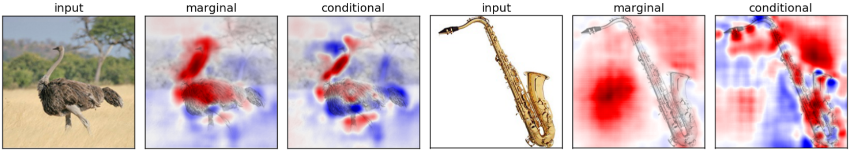
A neural network is relatively robust to the marginalization of just one feature.

- Remove several features at once
- Connected pixels.
- patches of size $k \times k$.



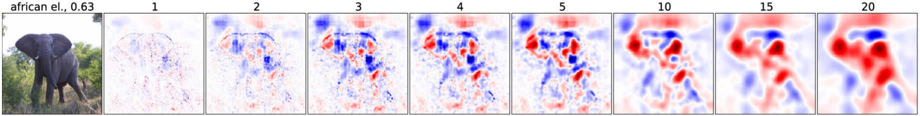
Experiments

Conditional sampling

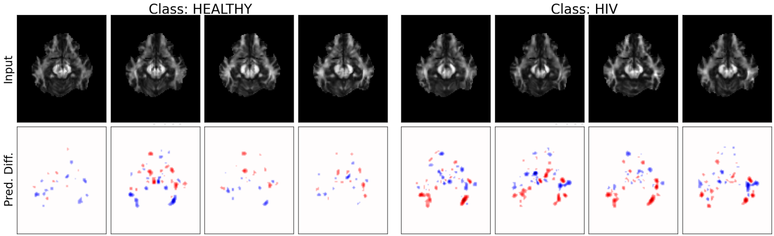


- Red: For.
- Blue: Against.

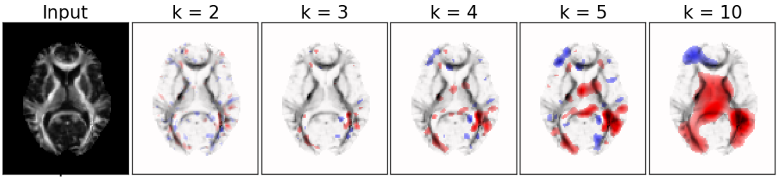
Multivariate analysis



MRI data



MRI data



Conclusions

- A method for visualizing deep neural networks by using a more powerful conditional, multivariate model.

- The visualization method shows which pixels of a specific input image are evidence for or against a node in the network.

Discovering Interpretable Representations for Both Deep Generative and Discriminative Models

Adel, Ghahramani, Weller, ICML 2018

Generative models and interpretability

- Generative models seek to infer the data-generating latent space.
- This implies capturing to some extent the salient characteristics of the data.
- Generative models can potentially provide disentangled (and perhaps interpretable?) data representations (Kingma et al., 2014; Chen et al., 2016; Desjardins et al., 2012; Higgins et al., 2017; Kulkarni et al., 2015) .

"What I cannot create, I do not understand.", Richard Feynman

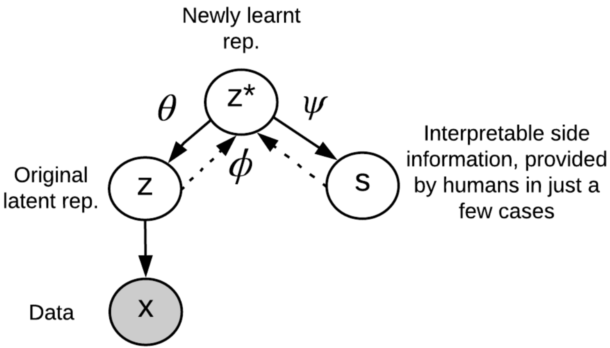
Contributions

UNIVERSITY OF CAMBRIDGE

We propose:

- An interpretability framework as a lens on an existing model using fully invertible transformations.
- An active learning methodology basing the acquisition function on mutual information with interpretable data attributes.
- A quantitative metric. We define interpretability as a *simple relationship to something we can understand*.
- A second interpretability framework jointly optimized for reconstruction and interpretability. This provides a novel analogy between data compression and regularization.
- Qualitative and quantitative state-of-the-art results on three datasets.

Interpretable Lens Variable Model (ILVM)



- Interactive 'human-in-the-loop' interpretability
- Choose the point with index $\mathbf{j}$ that maximizes :

$$\hat{\mathbf{j}} = \mathrm{argmax}_{\mathbf{j}}\, \mathbf{I}(\mathbf{s_j}, \psi) = \mathbf{H}(\mathbf{s_j}) - \mathbb{E}_{\mathbf{q}_\phi(\mathbf{z}^*|\mathbf{s})}[\mathbf{H}(\mathbf{s_j}|\mathbf{z_j^*})]$$
$$= -\int \mathbf{p}(\mathbf{s_j}) \log \mathbf{p}(\mathbf{s_j})\, ds$$
$$+ \mathbb{E}_{\mathbf{q}_\phi(\mathbf{z}^*|\mathbf{s})}\left[\int \mathbf{p}_\psi(\mathbf{s_j}|\mathbf{z}^*) \log \mathbf{p}_\psi(\mathbf{s_j}|\mathbf{z}^*)\, ds\right]. \quad (5)$$

- Choose the point possessing side information about which:
  - the model is most uncertain -maximized $\mathbf{H}(\mathbf{s_j})$-, but
  - in which the individual settings of the founding latent space $\mathbf{z}^*$ are confident -minimized $\mathbb{E}_{\mathbf{q}_\phi(\mathbf{z}^*|\mathbf{s})}[\mathbf{H}(\mathbf{s_j}|\mathbf{z_j^*})]$-

Interpretability Metric

- Interpretability refers to a *simple relationship to something we can understand*.

- A latent space is (more) interpretable if it manages to explain the relationship to salient attributes (more) simply.

Contributions

We propose:

- An interpretability framework as a lens on an existing model using fully invertible transformations.
- An active learning methodology basing the acquisition function on mutual information with interpretable data attributes.
- A quantitative metric. We define interpretability as a *simple relationship to something we can understand*.
- A second interpretability framework jointly optimized for reconstruction and interpretability. This provides a novel analogy between data compression and regularization.
- Qualitative and quantitative state-of-the-art results on three datasets.

Jointly Learnt Variable Model (JLVM)

- JLVM jointly optimizes for interpretability and reconstruction fidelity.
- It is based on the information bottleneck concept:
- Make $\mathbf{z}^*$ maximally expressive about the side information $\mathbf{s}$ while being maximally compressive about the data $\mathbf{x}$. :

$$\mathbf{IB}(\mathbf{z}^*, \mathbf{x}, \mathbf{s}) = \mathbf{I}(\mathbf{z}^*, \mathbf{s}) - \beta \mathbf{I}(\mathbf{z}^*, \mathbf{x}).$$

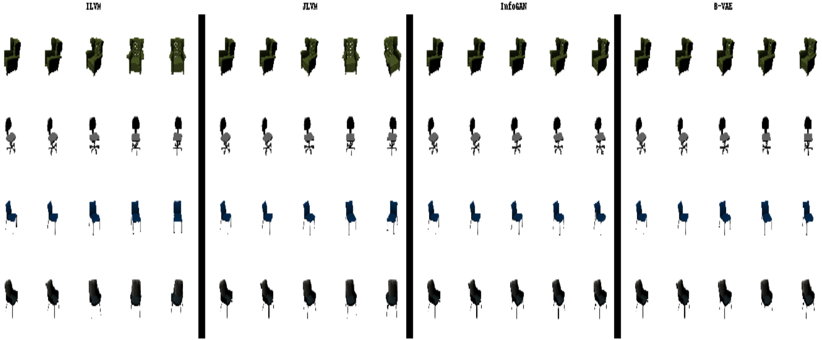
- We prove that being maximally compressive about the input for the sake of interpretability is analogous to further regularization.

Contributions

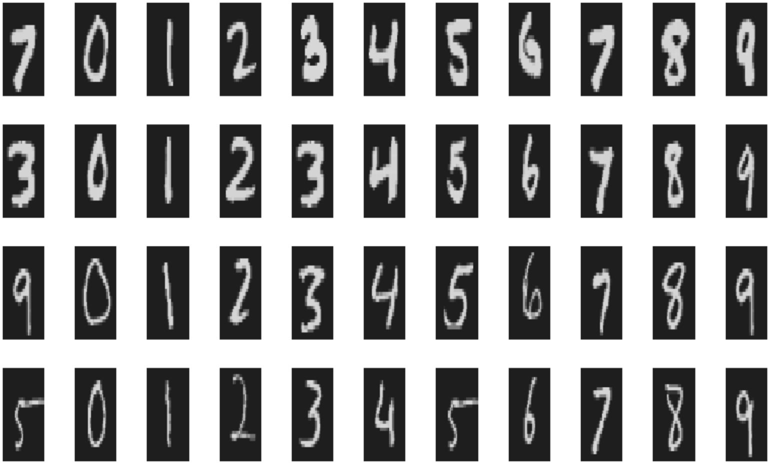
We propose:

- An interpretability framework as a lens on an existing model using fully invertible transformations.
- An active learning methodology basing the acquisition function on mutual information with interpretable data attributes.
- A quantitative metric. We define interpretability as a *simple relationship to something we can understand*.
- A second interpretability framework jointly optimized for reconstruction and interpretability. This provides a novel analogy between data compression and regularization.
- Qualitative and quantitative state-of-the-art results on three datasets.

Qualitative Evaluation

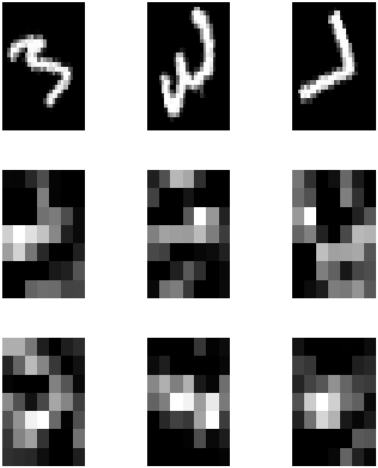


Qualitative Evaluation



Interpretable Lens Variable Model (ILVM)

Interpretable Lens on a Hidden Layer of a Neural Network



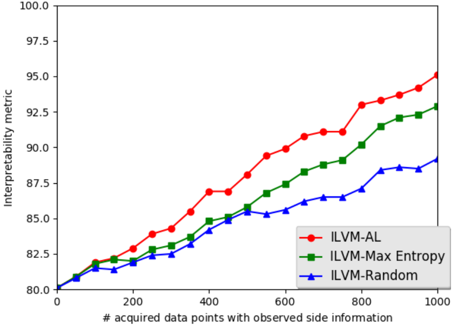
Quantitative Evaluation

Interpretability Metric

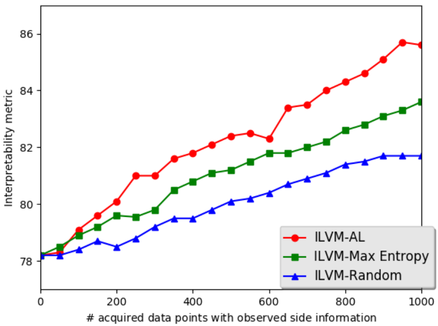
	MNIST	SVHN	Chairs
`ILVM`	$\mathbf{95.2 \pm 1.3}$ %	85.7 ± 0.9 %	87.4 ± 1.0 %
`JLVM`	89.8 ± 0.9 %	$\mathbf{90.1 \pm 1.1}$ %	$\mathbf{89.8 \pm 1.5}$ %
InfoGAN	83.3 ± 1.8 %	83.9 ± 1.3 %	85.2 ± 1.4 %

Quantitative Evaluation

Active Learning



(a) MNIST



(b) SVHN

Conclusion

- In `ILVM`, interpretability does not conflict with the original objective, be it reconstruction fidelity or classification accuracy.
- A strategy to bring human subjectivity into interpretability to yield interactive 'human-in-the-loop' interpretability.
- `JLVM` sheds light on a newly derived relationship between compression and regularization.
- The introduced frameworks achieve state-of-the-art results on three datasets.

The End