Interpretability in Machine Learning

Tameem Adel

Machine Learning Group
University of Cambridge, UK

February 22, 2018

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

- 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.

- 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

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.

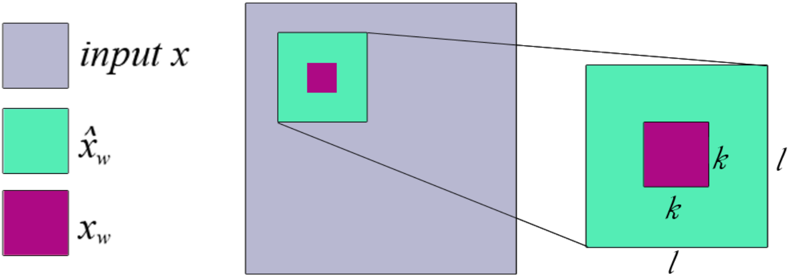
Conditional sampling

- 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.

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

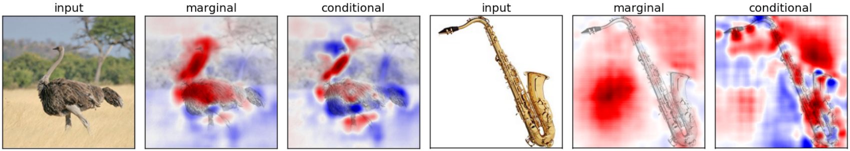
A neural network is relatively robust to the marginalization of just one feature.

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



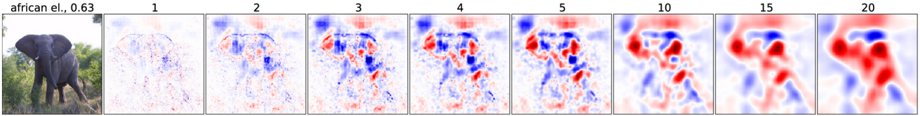
UNIVERSITY OF CAMBRIDGE

Conditional sampling

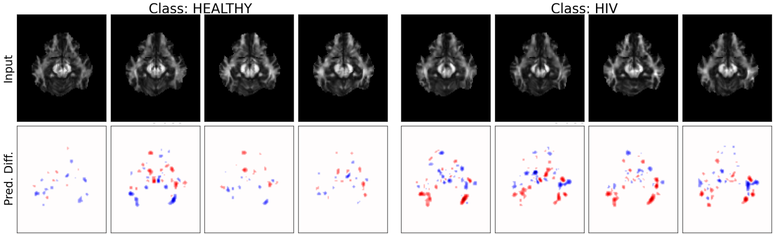


- Red: For.
- Blue: Against.

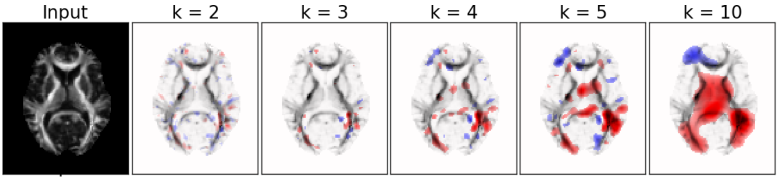
Multivariate analysis



MRI data



MRI data



- 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.

InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

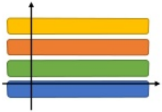
Chen, Duan, Houthooft, Schulman, Sutskever, Abbeel, NIPS 2016

Motivation

How can we achieve
unsupervised learning of disentangled representation?

In general, learned representation is entangled,
i.e. encoded in a data space in a complicated manner

When a representation is **disentangled**, it would be
more interpretable and easier to apply to tasks



Generative Adversarial Nets(GANs)

Generative model trained by competition between two neural nets:

✓Generator $x = G(z),\ z \sim p_z(Z)$
$p_z(Z)$: an arbitrary noise distribution

✓Discriminator $D(x) \in [0,1]$:
probability that x is sampled from the data dist. $p_{\mathrm{data}}(X)$ rather than generated by the generator $G(z)$

Optimization problem to solve:

$\min_G \max_D V_{\mathrm{GAN}}(G,D)$, where

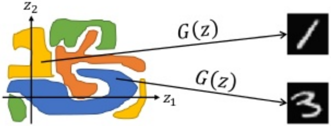
$$V_{\mathrm{GAN}}(G,D) \equiv E_{x\sim p_{\mathrm{data}}(X)}[\ln D(x)] + E_{z\sim p_z(Z)}\left[\ln\left(1 - D\left(G(z)\right)\right)\right]$$

Problems with GANs

From the perspective of representation learning:

✓No restrictions on how $G(z)$ uses z

- z can be used in a highly entangled way
- Each dimension of z does not represent any salient feature of the training data



Proposed Resolution: InfoGAN -Maximizing Mutual Information -

Observation in conventional GANs:
a generated date x does not have much information on the noise z from which x is generated
because of heavily entangled use of z

Proposed resolution = InfoGAN:
the generator $G(z, c)$ trained so that
it maximize the mutual information $I(C|X)$ between the latent code C and the generated data X

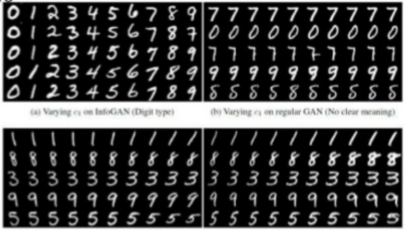
$$\min_{G} \max_{D} \{V_{\text{GAN}}(G, D) - \lambda I(C|X = G(Z, C))\}$$

Experiment
– Disentangled Representation –

- InfoGAN on MNIST dataset
- Latent codes
  - ✓ c_1: 10-class categorical code
  - ✓ c_2, c_3: continuous code

✓ c_1 can be used as a classifier with 5% error rate.

✓ c_2 and c_3 captured the rotation and width, respectively



(a) Varying c_1 on InfoGAN (Digit type) (b) Varying c_1 on regular GAN (No clear meaning)

(c) Varying c_2 from -2 to 2 on InfoGAN (Rotation) (d) Varying c_3 from -2 to 2 on InfoGAN (Width)

Figure 2 in the original paper

Experiment
– Disentangled Representation –

Dataset: P. Paysan, *et al.*, AVSS, 2009, pp. 296–301.

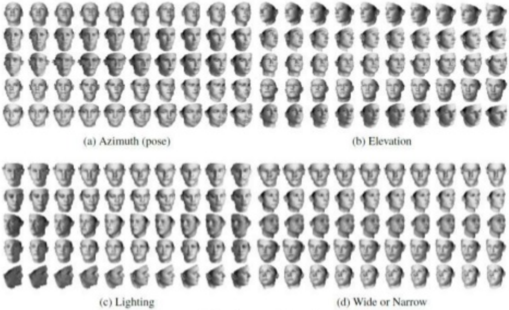


Figure 3 in the original paper

Experiment
– Disentangled Representation –

Dataset: M. Aubry, *et al.*, CVPR, 2014, pp. 3762–3769.

(a) Rotation (b) Width

Figure 4 in the original paper

InfoGAN learned salient features without supervision

Experiment
– Disentangled Representation –

Dataset: Street View House Number

(a) Continuous variation: Lighting

(b) Discrete variation: Plate Context

Figure 5 in the original paper

Experiment
– Disentangled Representation –

Dataset: CelebA

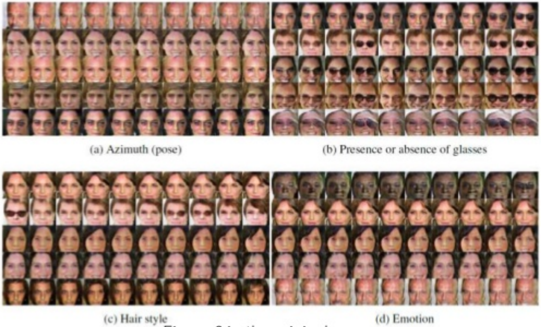


Figure 6 in the original paper

Future Prospect and Conclusion

✓Mutual information maximization can be applied to other methods, e.g. VAE

✓Learning hierarchical latent representation

✓Improving semi-supervised learning

✓High-dimentional data discovery

Goal
Unsupervised learning of disentangled representations

Approach
GANs + Maximizing Mutual Information between generated images and input codes

Benefit
Interpretable representation obtained without supervision and substantial additional costs

The End