Explainable AI paper

Presenter: YC, Choi

2019.02.18

Contents

Methods for Interpreting and Understanding Deep Neural Networks

Learning deep features for discriminative localization

Evaludating the visualization of what a Deep Neural Network has learned

Explaining Recurrent Neural Network Predictions in Sentiment Analysis

Contents

Methods for Interpreting and Understanding Deep Neural Networks

Learning deep features for discriminative localization

Evaludating the visualization of what a Deep Neural Network has learned

Explaining Recurrent Neural Network Predictions in Sentiment Analysis

Introduction

- Techiques for interpreting and understanding what the model has learned have therefore become a key ingredient of a robust validation procedure
- This paper gives an overview of techniques for interpreting complex machine learning models, with a focus on deep neural networks.
- First, it can be usefule to clarify the meaning we associate to definitions in this paper.

Introduction

- Definition 1. **Interpretation** is the mapping of an abstract concept (e.g. a predicted class) into a domain that the human can make sense of.

  About estimated model f

- Definition 2. **Explanation** is the collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g. classification or regression)

  About givan a input x and estimated model f

- Example of explanation is a heatmap highlighting which pixels of the input image most strongly support the classification decision.

Explaining DNN Decision

- We ask for a given data point x and pre-trained deep neural network f
- A common approach is to view the data point x as a collection of features $(x_i)_{i=1}^{d}$ and to assign to each of these, a score R_i determining how relevant the feature x_i is for explaining $f(x)$

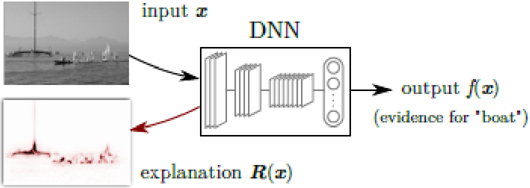

input $\boldsymbol{x}$
DNN
output $f(\boldsymbol{x})$
(evidence for "boat")
explanation $\boldsymbol{R}(\boldsymbol{x})$

Figure 3: Explanation of the DNN prediction "boat" for an image $\boldsymbol{x}$ given as input.

Sensitivity Analysis

- It is based on the models locally evaluated gradient or some other local measure of variation.
- $R_i(x) = (\frac{\partial f}{\partial x_i})^2$
- The most relevant input features are those to which the output is most sensitive.
- The technique is easy to implement for a deep neural network, since the gradient can be computed using backpropagation.

Sensitivity Analysis

- It is important to note that sensitivity analysis does not produce an explanation of the function value $f(x)$ itself, but rather a variation of it.
- $\sum_{i=1}^{d} R_i(x) = \|\nabla f(x)\|^2$

Simple Taylor Decomposition

- The Taylor decomposition is a method that explains the model's decision by decomposing the function value $f(x)$ as a sum of relevance scores.
- Consider some root point $\tilde{x}$ for which $f(\tilde{x}) = 0$
- This expansion lets us rewrite the function as:

$$f(x) = \sum_{i=1}^{d} R_i(x) + O(xx^T)$$

where $R_i(x) = \frac{\partial f}{\partial x_i}|_{x=\tilde{x}}(x_i - \tilde{x}_i)$

Simple Taylor Decomposition

- Consider a special class of functions :
  piecewise linear, $f(tx) = tf(x)$ for $t \geq 0$ (deep ReLU networks)
- Let f be a functio satisfying above condition, then root point $\tilde{x} = \lim_{\epsilon \to 0} \epsilon x$ and higher-order terms are zero
- $f(x) = \sum_{i=1}^{d} R_i(x)$ where $R_i(x) = \dfrac{\partial f}{\partial x_i} x_i$

LRP(Layerwise Relevance Propagation)

- Decomposong the prediction of a DNN is to make explicit use of its feed-forward graph structure.
- $w_{jk}^{(l)}$ be the weight from j-the node in l-th layer to k-th node in l+1-th layer.
- $x_j^{(l)}$ be the activated value of j-th node in l-th layer.
- $R_j^{(l)} = \sum_k \frac{x_j^{(l)} w_{jk}^{(l)}}{\sum_{j'} x_{j'}^{(l)} w_{j'k}^{(l)} + \epsilon} R_k^{(l+1)}$ is the score of j-the node in l-th layer where $\epsilon > 0$ is a stablization term.
- $R^{(L+1)} = f(x)$ where L is the number of hidden layer.

LRP(Layerwise Relevance Propagation)

- $\alpha\beta$-rule

$$R_j^{(l)} = \sum_k (\alpha \frac{x_j w_{jk}^+}{\sum_j x_j w_{jk}^+} - \beta \frac{x_j w_{jk}^-}{\sum_j x_j w_{jk}^-}) R_k^{(l+1)}$$

where $()^+$and $()^-$ denote the positive and negitive parts respectively, and where the parameter αandβ are chosen subject to the constraints $\alpha - \beta = 1$ and $\beta \geq 0$

LRP and Deep Taylor Decomposition

- LRP-$\alpha_1\beta_0$ and Deep Taylor decomposition are same in some sense.
- Consier deep ReLU networks
- $a_k = \max(0, \sum_j a_j w_{jk} + b_k)$ with $b_k \leq 0$
- Then, we can rewrite $R_k = a_k c_k$
- $R_k = \max(0, \sum_j a_j w^{'}_{jk} + b^{'}_k)$
- Using Taylor expansion, we can get a first-order them:

$$R_k = \sum_j \frac{\partial R_k}{\partial a_j}|_{(\tilde{a}_{j_j}}(a_j - \tilde{a}_j)$$

Handling Special Layers

- l_p-pooling layers(including sum pooling and max-pooling)
- LRP authors use a winner-take-all redistrubition policy for max pooling layer.
- Montavon et al. recommend to apply for l_p-pooling layers the following propagation rule:

$$R_j^{(l)} = \frac{x_j}{\sum_j x_j} R_k^{(l+1)}$$

Explanation Continuity

- Explanation continuity can be quantified by looking for the strongest variation of the explanation $R(x)$ in the input domain

$$\max_{x \neq x'} = \frac{\|R(x) - R(x')\|}{\|x - x'\|}$$

- When $f(x)$ is a deep ReLU network, both sensitivity and Taylor decomposition have sharp discontinuity.
- On the other hand, deep Taylor LRP produces continuous explanations.

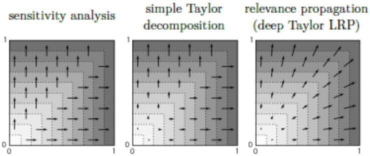


Figure 6: Explaining $\max(x_1, x_2)$. Function values are represented as a contour plot, with dark regions corresponding to high values. Relevance scores are represented as a vector field, where horizontal and vertical components are the relevance of respective input variables.

Contents

Methods for Interpreting and Understanding Deep Neural Networks

Learning deep features for discriminative localization

Evaludating the visualization of what a Deep Neural Network has learned

Explaining Recurrent Neural Network Predictions in Sentiment Analysis

CAM (Class Activation Mapping)

- The authors suggest a Class Activation Mapping which is a explanation method.
- Use a network architecture similar to GooLeNet.
- Before the final output layer, the author perform global average pooling on the convolutional feature map.

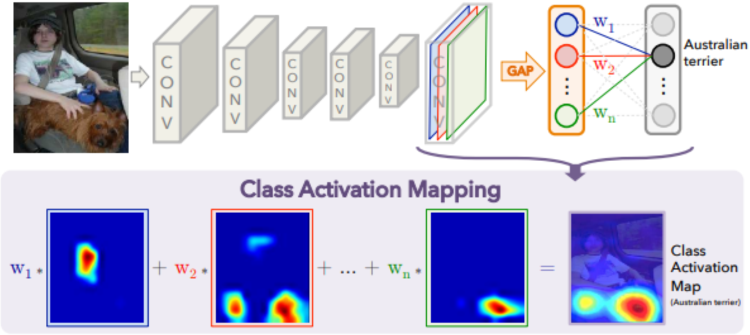
CAM (Class Activation Mapping)

- Let $f_k(x, y)$ represent the activation of unit k in the last convolutional layer at spatial location (x,y).
- For unit k, the result of performing global average pooling $F^k = \sum_{x,y} f_k(x, y)$
- So, for a given class c, the input to the softmax, $S_c = \sum_k w_k^c F_k$

$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x, y) = \sum_{x,y} \sum_k w_k^c f_k(x, y) = \sum_k M_c(x, y)$$

where $M_c(x, y) = \sum_k w_k^c f_k(x, y)$

CAM (Class Activation Mapping)



CAM (Class Activation Mapping)

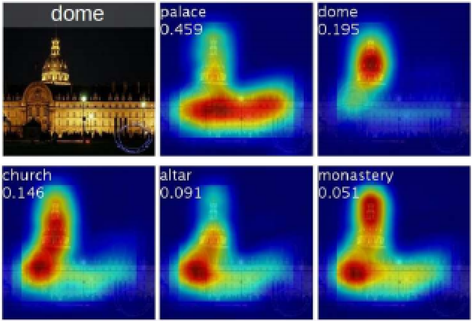


Figure 4. Examples of the CAMs generated from the top 5 predicted categories for the given image with ground-truth as dome. The predicted class and its score are shown above each class activation map. We observe that the highlighted regions vary across predicted classes e.g., *dome* activates the upper round part while *palace* activates the lower flat part of the compound.

Contents

Methods for Interpreting and Understanding Deep Neural Networks

Learning deep features for discriminative localization

Evaludating the visualization of what a Deep Neural Network has learned

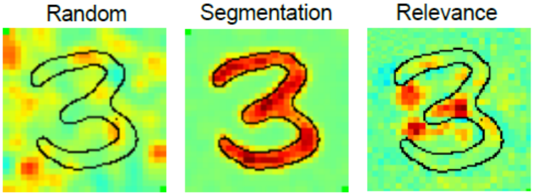
Explaining Recurrent Neural Network Predictions in Sentiment Analysis

Introduction

- It is difficult to intuitively and quantitatively understand the result of DNN inference.
- Note that this aspect differs from feature selection, where the question is: which features are on average salient for the ensemble of training data.
- It is difficult to quantitatively evaluate the quality of a heatmap.
- An automated objective and quantitative measure for assessing heatmap quanlity becomes necessary.

Heatmap

- A heatmap $h = \{h_p\}$ assigns each pixel p a value $h_p = \mathcal{H}(x, f, p)$ where $f\colon \mathbb{R}^d \to \mathbb{R}^+$ the scoring function.
- Since h the same dimensionality as x, it can be visualized as an image.

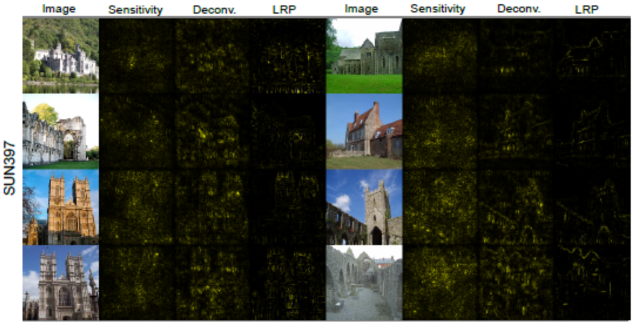


Evaluating Heatmaps

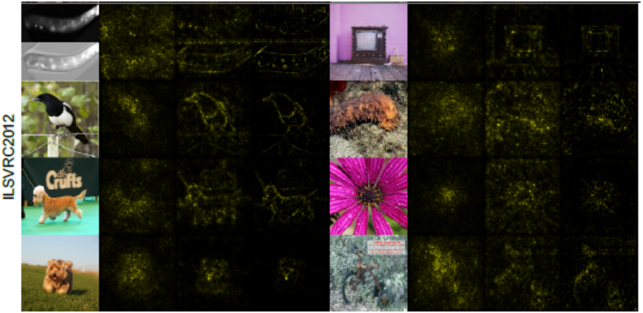
- Heatmap quality does not only depend on the algorithms used to compute a heatmap, but also on the performance of the classifier.
- If the training data does not contain images of the defits '3', then the classifier can not know that the absence of strokes in the above figure.
- Note that there is no guarantee that human and classifier explanations match.

Experimental Result

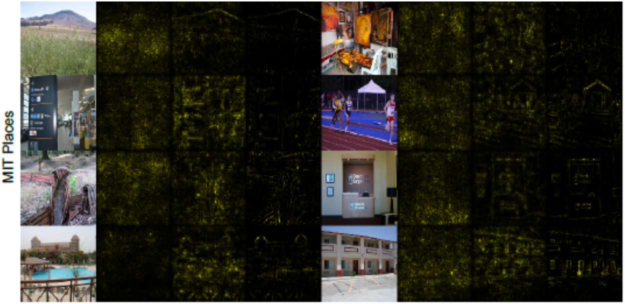
- Use 3 benchmark datasets (SUN397, ILSVRC2012, MIT Places)



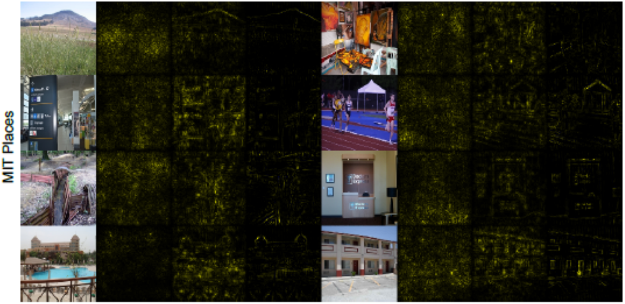
Experimental Result



Experimental Result



Experimental Result



Pertubation measure

- Define a heatmap as an ordered set of locations in the image, where these locations might lie on a predefined grid.

  $$\mathcal{O} = (r_1, r_2, \ldots, r_d)$$

  where each r_p, $(p = 1, \ldots, d)$ is for example a two-dimensional vector encoding the horizontal and vertical position on a grid of pixels. (can be a singple pixel or a local neighborhood)
- A heatmap function $h_p = \mathcal{H}(x, f, r_p)$ indicates how important the given location r_p of the image is for representing the image class.

MoRF and AOPC

- Region perturbation process that follows the ordered sequence of locations.

$$\begin{aligned} x_{MoRF}^{(0)} &= x && (1) \\ x_{MoRF}^{(k)} &= g(x_{MoRF}^{(k-1)}, r_k), \forall 1 \leq k \leq d. && (2) \end{aligned}$$

- The process of removing information in order of importance
- The quantity of interest in this case is the area over the MoRF perturbation curve(AOPC)

$$\text{AOPC(L)} = \frac{1}{L} < \sum_{k=1}^{L} f(x_{MoRF}^{(0)}) - f(x_{MoRF}^{(k)}) >_{p(x)}$$

where $< . >_{p(x)}$ denotes the average over all images in the dataset.

Experimental Result

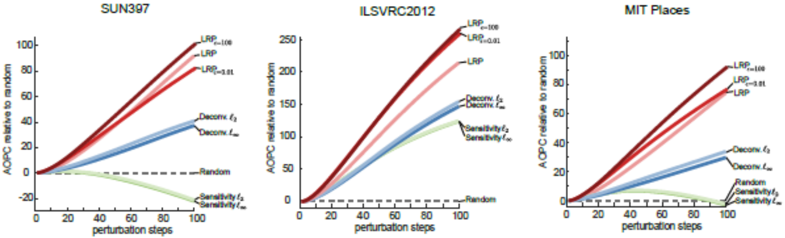


Fig. 4. Comparison of the three heatmapping methods relative to the random baseline. The LRP algorithms have largest AOPC values, i.e., best explain the classifier's decision, for all three data sets.

Contents

Methods for Interpreting and Understanding Deep Neural Networks

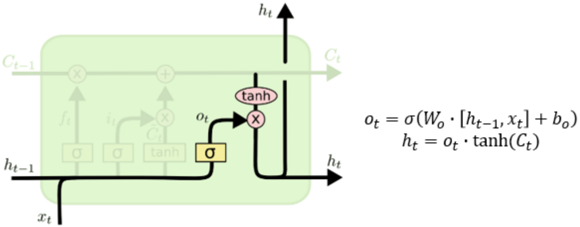
Learning deep features for discriminative localization

Evaludating the visualization of what a Deep Neural Network has learned

Explaining Recurrent Neural Network Predictions in Sentiment Analysis

Multiplicative Interactions

- Let z_j be an upper layer neuron, whose value in the forward pass is computed as the multiplication of the two lower layer neuron values z_g and z_s (i.e. $z_j = z_g z_s$)
- In LSTMs GRUs, there is always one of two lower-layer neurons that constitutes a gate.



$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$
$$h_t = o_t \cdot \tanh(C_t)$$

- In LSTM, $z_s = o_t, \; z_g = tanh(C_t)$

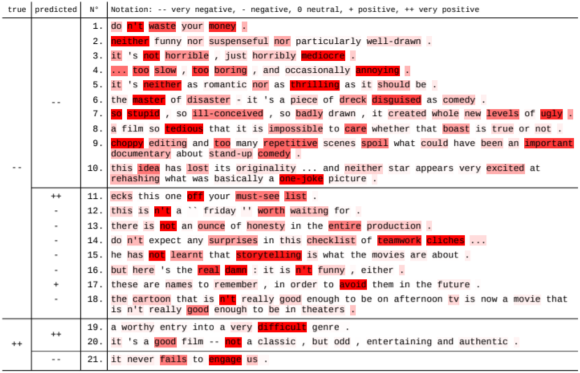
Multiplicative Interactions

- We set $R_g = 0$ and $R_s = R_j$
- That is(i think), for input x_t, the relevance score $R_t \approx \frac{o_t}{\sum_{t'} o_{t'}}$
- The intuition behind this reallocation rule is that the gate neuron decides already in the forward pass how much of the information contained in the source neuron should be retained to make the overall classification decision.

Experimental Result

- 2210 tokenized sentences of the Stanford Sentiment Treebank
- five-class (Very negative, negative, neutral, positive, very positive)
- The trained model achives 46.3% accuracy and for binary classification, 82.9%

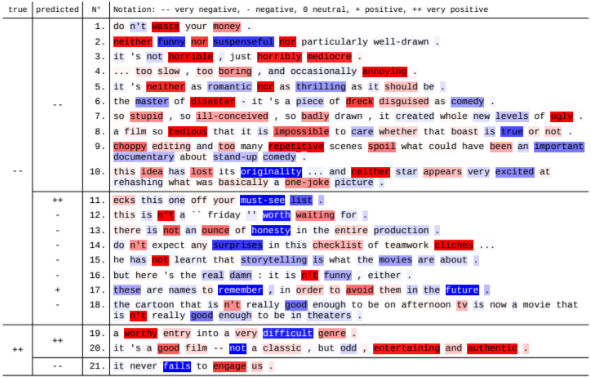
Multiplicative Interactions



true	predicted	N°	Notation: -- very negative, - negative, 0 neutral, + positive, ++ very positive
--	--	1.	do n't waste your money .
		2.	neither funny nor suspenseful nor particularly well-drawn .
		3.	it 's not horrible , just horribly mediocre .
		4.	... too slow , too boring , and occasionally annoying .
		5.	it 's neither as romantic nor as thrilling as it should be .
		6.	the master of disaster - it 's a piece of dreck disguised as comedy .
		7.	so stupid , so ill-conceived , so badly drawn , it created whole new levels of ugly .
		8.	a film so tedious that it is impossible to care whether that boast is true or not .
		9.	choppy editing and too many repetitive scenes spoil what could have been an important documentary about stand-up comedy .
		10.	this idea has lost its originality ... and neither star appears very excited at rehashing what was basically a one-joke picture .
	++	11.	ecks this one off your must-see list .
	-	12.	this is n't a `` friday '' worth waiting for .
	-	13.	there is not an ounce of honesty in the entire production .
	-	14.	do n't expect any surprises in this checklist of teamwork cliches ...
	-	15.	he has not learnt that storytelling is what the movies are about .
	-	16.	but here 's the real damn : it is n't funny , either .
	+	17.	these are names to remember , in order to avoid them in the future .
	-	18.	the cartoon that is n't really good enough to be on afternoon tv is now a movie that is n't really good enough to be in theaters .
++	++	19.	a worthy entry into a very difficult genre .
		20.	it 's a good film -- not a classic , but odd , entertaining and authentic .
	--	21.	it never fails to engage us .

Figure 1: SA heatmaps of exemplary test sentences, using as target class the *true* sentence class. All relevances are positive and mapped to red, the color intensity is normalized to the maximum relevance per sentence. The true sentence class, and the classifier's predicted class, are indicated on the left.

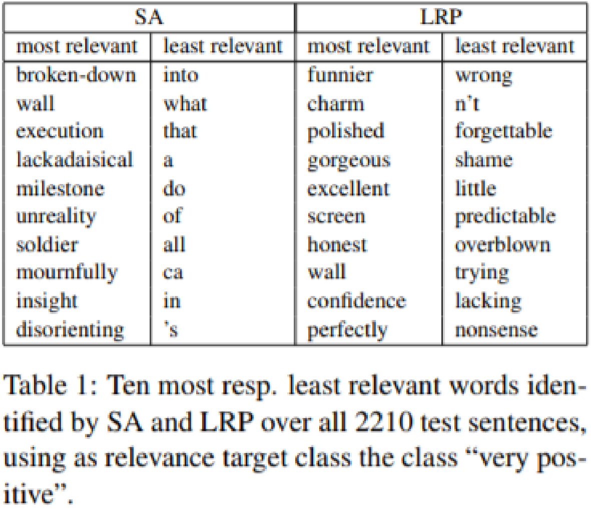
Multiplicative Interactions



true	predicted	N°	Notation: -- very negative, - negative, 0 neutral, + positive, ++ very positive
--	--	1.	do n't waste your money .
		2.	neither funny nor suspenseful nor particularly well-drawn .
		3.	it 's not horrible , just horribly mediocre .
		4.	... too slow , too boring , and occasionally annoying .
		5.	it 's neither as romantic nor as thrilling as it should be .
		6.	the master of disaster - it 's a piece of dreck disguised as comedy .
		7.	so stupid , so ill-conceived , so badly drawn , it created whole new levels of ugly .
		8.	a film so tedious that it is impossible to care whether that boast is true or not .
		9.	choppy editing and too many repetitive scenes spoil what could have been an important documentary about stand-up comedy .
		10.	this idea has lost its originality ... and neither star appears very excited at rehashing what was basically a one-joke picture .
	++	11.	ecks this one off your must-see list .
	-	12.	this is n't a `` friday '' worth waiting for .
	-	13.	there is not an ounce of honesty in the entire production .
	-	14.	do n't expect any surprises in this checklist of teamwork cliches ...
	-	15.	he has not learnt that storytelling is what the movies are about .
	-	16.	but here 's the real damn : it is n't funny , either .
	+	17.	these are names to remember , in order to avoid them in the future .
	-	18.	the cartoon that is n't really good enough to be on afternoon tv is now a movie that is n't really good enough to be in theaters .
++	++	19.	a worthy entry into a very difficult genre .
		20.	it 's a good film -- not a classic , but odd , entertaining and authentic .
	--	21.	it never fails to engage us .

Figure 2: LRP heatmaps of exemplary test sentences, using as target class the *true* sentence class. Positive relevance is mapped to red, negative to blue, and the color intensity is normalized to the maximum absolute relevance per sentence. The true sentence class, and the classifier's predicted class, are indicated on the left.

Multiplicative Interactions



SA		LRP	
most relevant	least relevant	most relevant	least relevant
broken-down	into	funnier	wrong
wall	what	charm	n't
execution	that	polished	forgettable
lackadaisical	a	gorgeous	shame
milestone	do	excellent	little
unreality	of	screen	predictable
soldier	all	honest	overblown
mournfully	ca	wall	trying
insight	in	confidence	lacking
disorienting	's	perfectly	nonsense

Table 1: Ten most resp. least relevant words identified by SA and LRP over all 2210 test sentences, using as relevance target class the class "very positive".