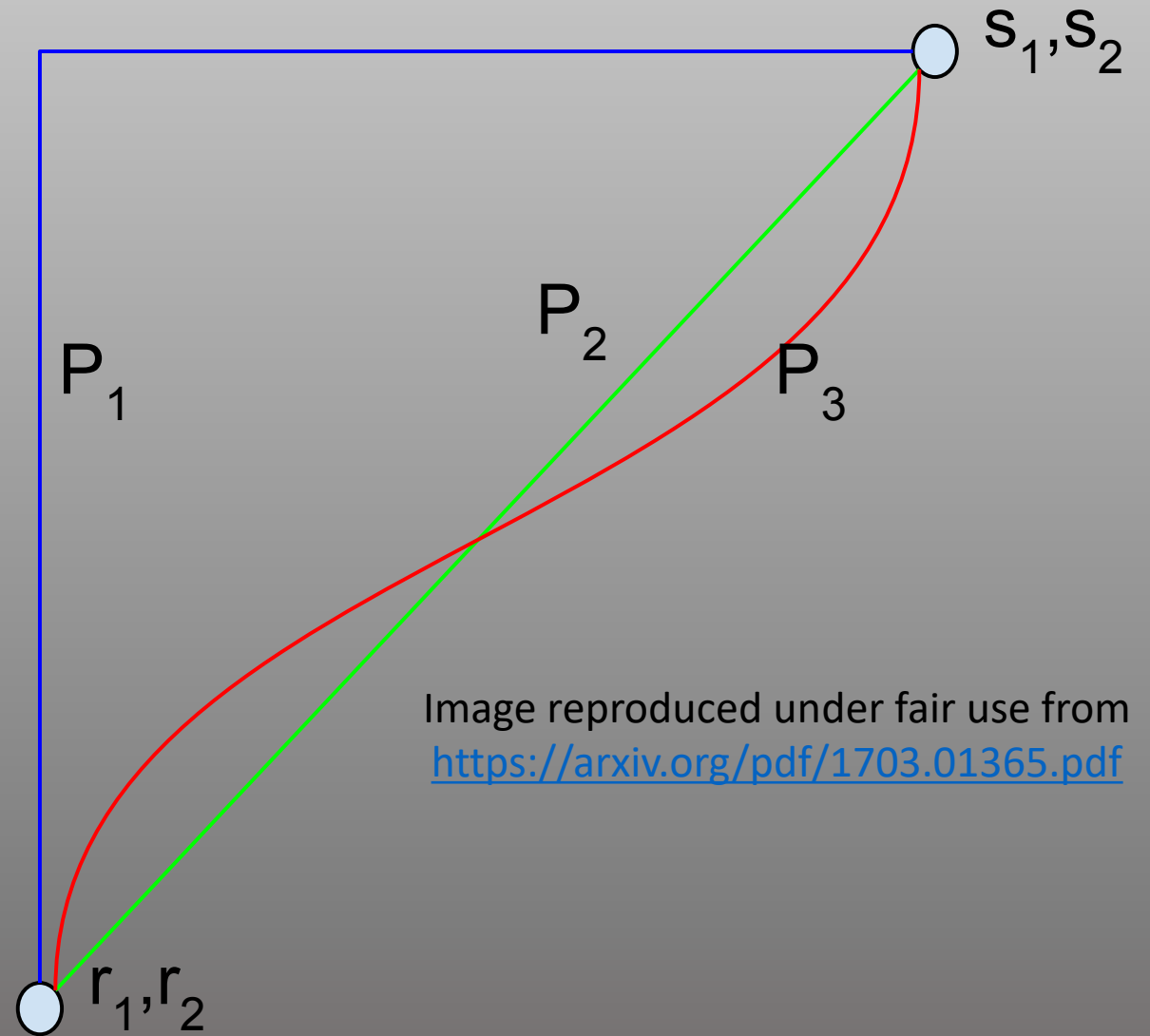


# Axiomatic Attribution for Deep Networks

Paper Authors: Mukund Sundararajan, Ankur Taly,  
Qiqi Yan



# Attributions of a DNN to its input features

- *Given a DNN  $F : \mathbb{R}^n \rightarrow [0,1]$*
- *and an input  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ ,*
- *an attribution for input  $x$  relative to a baseline input  $x^0$*
- *vector  $A_F(x, x^0) = (a_1, \dots, a_n) \in \mathbb{R}^n$*
- *Here,  $a_i$  is the contribution of  $x_i$  to the prediction  $F(x)$ .*

Baehrens, David, Schroeter, Timon, Harmeling, Stefan, Kawanabe, Motoaki, Hansen, Katja, and Muller, Klaus-Robert.

How to explain individual classification decisions.

*Journal of Machine Learning Research*, pp. 1803– 1831, 2010.

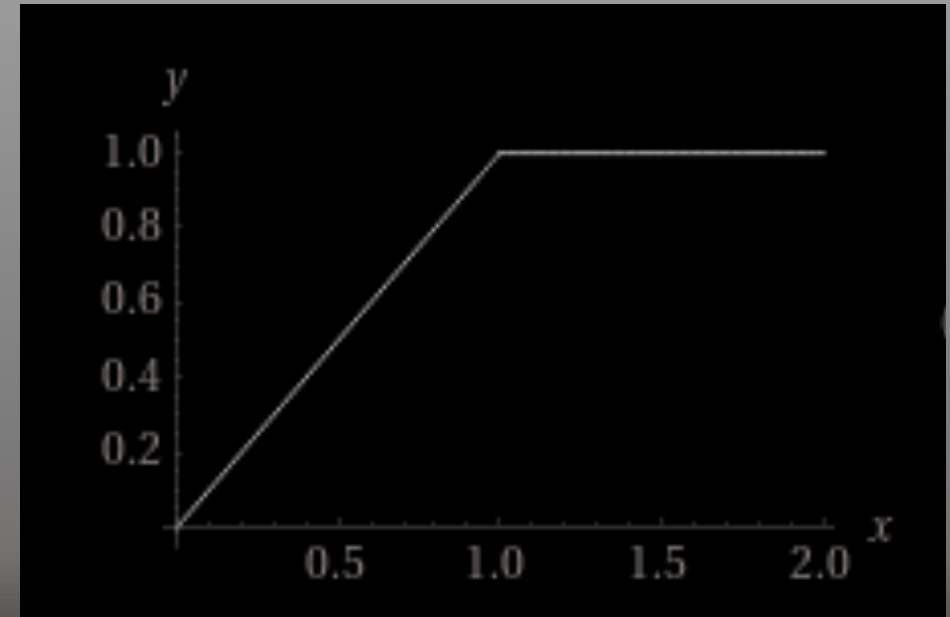
# Why choose a baseline?

- *Humans often perform attribution by exploiting*
  - *counterfactual intuition.*
- *Blame a feature → absence of feature is a baseline*
- *Here, absence of feature described using a single baseline input.*
- *DNNs: natural baseline*
  - *An input where the DNN is “neutral”.*
- *E.g. object recognition networks*
  - *Black image*

# Axiom 1: Sensitivity

- For an input and a baseline differing in 1 feature F
- with divergent predictions t,
- the feature F must have a non-zero attribution.

- $f(x) = 1 - \text{ReLU}(1-x)$
- Baseline:  $x = 0$ 
  - f is 0
- Input:  $x = 2$ 
  - f is 1
- **Gradient methods** assign 0 attribution to x
  - as the function is flat at  $x=2$



# Axiom 1: Sensitivity

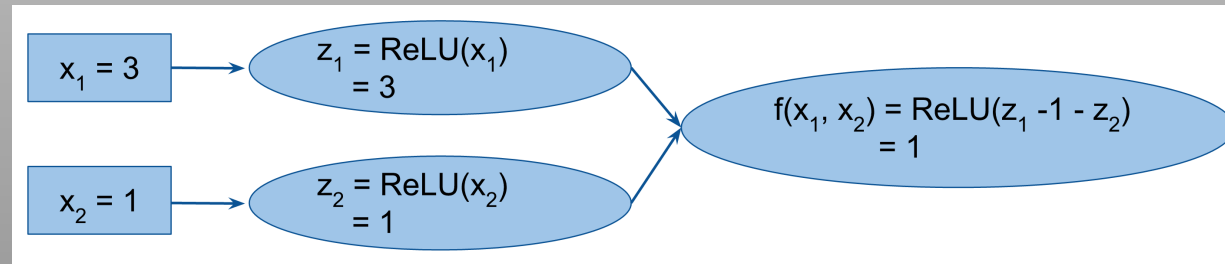


Image reproduced under  
fair use from  
[https://arxiv.org/pdf/1703.  
01365.pdf](https://arxiv.org/pdf/1703.01365.pdf)

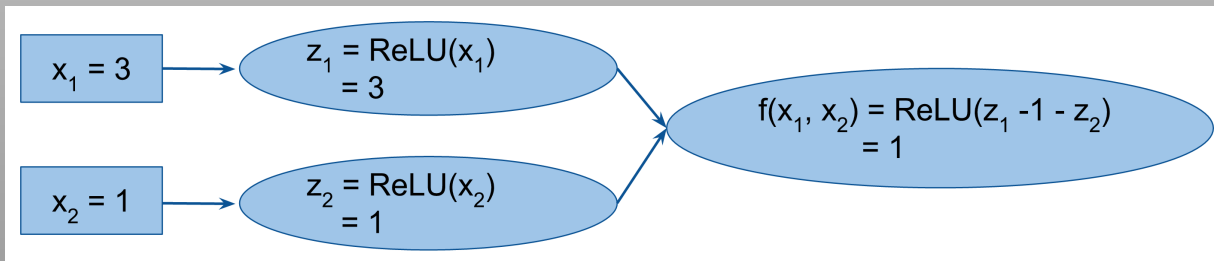
Network  $f(x_1, x_2)$

- For  $x_1 > 1$ , output decreases linearly as  $x_2$  increases from 0 to  $x_1 - 1$ .
- Yet, for all inputs, both of these assign 0 attribution for  $x_2$ 
  - Deconvolutional networks
  - Guided back-propagation
  - back-propagated signal at  $\text{ReLU}(x_2)$  is less than 0
    - and is therefore not back-propagated through the ReLU operation

# Axiom 2: Implementation Invariance

- Chain rule does not hold for discrete gradients
- DepLIFT and LRP use discrete gradients to tackle sensitivity.

# DeepLIFT and LRP break Axiom 2



Network  $f(x_1, x_2)$

Attributions at  $x_1 = 3, x_2 = 1$

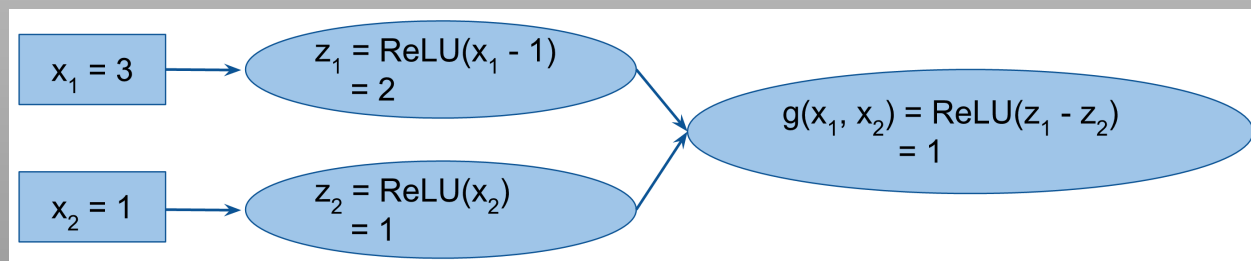
**Integrated gradients**  $x_1 = 1.5, x_2 = -0.5$

DeepLift  $x_1 = 1.5, x_2 = -0.5$

LRP  $x_1 = 1.5, x_2 = -0.5$

$$h(x_1, x_2) = \text{ReLU}(x_1) - 1 - \text{ReLU}(x_2)$$

$$k(x_1, x_2) = \text{ReLU}(x_1 - 1) - \text{ReLU}(x_2)$$



Network  $g(x_1, x_2)$

Attributions at  $x_1 = 3, x_2 = 1$

**Integrated gradients**  $x_1 = 1.5, x_2 = -0.5$

DeepLift  $x_1 = 2, x_2 = -1$

LRP  $x_1 = 2, x_2 = -1$

Different only when  $x_1 < 1$ , but then  $f=g=0$ .

# Integrated Gradients

$$\text{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha$$

Input                      Baseline                      Model output w.r.t. feature i

- Merge the two
  - Implementation Invariance of Gradients
  - Sensitivity of LRP or DeepLift.
- In practice, 20 to 300 discrete samples approximate the integral
  - within 5%.



# Fundamental Theorem of Calculus

- Let  $f$  be a real-valued function on a  $[a, b]$
- Let  $F$  be an antiderivative of  $f$  in  $(a, b)$  i.e.

$$F'(x) = f(x)$$

- If  $f$  is Riemann integrable on  $[a, b]$  then

$$\int_a^b f(x)dx = F(b) - F(a)$$


# Integrated Gradients - Completeness

$$\text{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha$$

↑                      ↑                      ↑  
Input                      Baseline                      Model output w.r.t. feature i

- If  $F : \mathbb{R}^n \rightarrow \mathbb{R}$  is differentiable almost everywhere,

$$\sum_{i=1}^n \text{IntegratedGrads}_i(x) = F(x) - F(x')$$

- Completeness  Sensitivity
- Gradients  Implementation-Invariance

# General Path Methods

- Let  $\gamma = (\gamma_1, \dots, \gamma_n) : [0, 1] \rightarrow \mathbb{R}^n$  be a smooth function
  - specifying a path in  $\mathbb{R}^n$
  - from baseline  $x^0$  to input  $x$ ,
  - i.e.,  $\gamma(0) = x^0$  and  $\gamma(1) = x$ .

$$\text{PathIntegratedGrads}_i^\gamma(x) ::= \int_{\alpha=0}^1 \frac{\partial F(\gamma(\alpha))}{\partial \gamma_i(\alpha)} \frac{\partial \gamma_i(\alpha)}{\partial \alpha} d\alpha$$

- All path methods satisfy Sensitivity and Implementation Invariance
- Integrated Gradients (IG) is a path method for the straight-line path
  - $\gamma(\alpha) = x^0 + \alpha \times (x - x^0)$
  - for  $\alpha \in [0, 1]$ .

# Additional Axioms satisfied by Path Methods

- Axiom Dummy:
  - If a DNN does not depend on a variable  $X$ ,
  - then the attribution to the variable  $X$  is zero.
- Axiom Linearity
  - If we linearly compose 2 DNNs  $f_1$  and  $f_2$  to form a third DNN  $a f_1 + b f_2$
  - Then the attributions for the new DNN
    - should be the weighted sum of the attributions for  $f_1$  and  $f_2$
    - with weights  $a$  and  $b$  respectively.
- Path methods are the only attribution methods satisfying

- Implementation Invariance
- Dummy,
- Linearity,
- Completeness.

Friedman, Eric J. Paths and consistency in additive cost sharing. *International Journal of Game Theory*, 32(4): 501–518, 2004.

Aumann, R. J. and Shapley, L. S. *Values of Non-Atomic Games*. Princeton University Press, Princeton, NJ, 1974.

# Symmetry Preserving Path Methods = IG

- *Integrated gradients (IG) is the unique general path method that respects symmetry-preserving.*

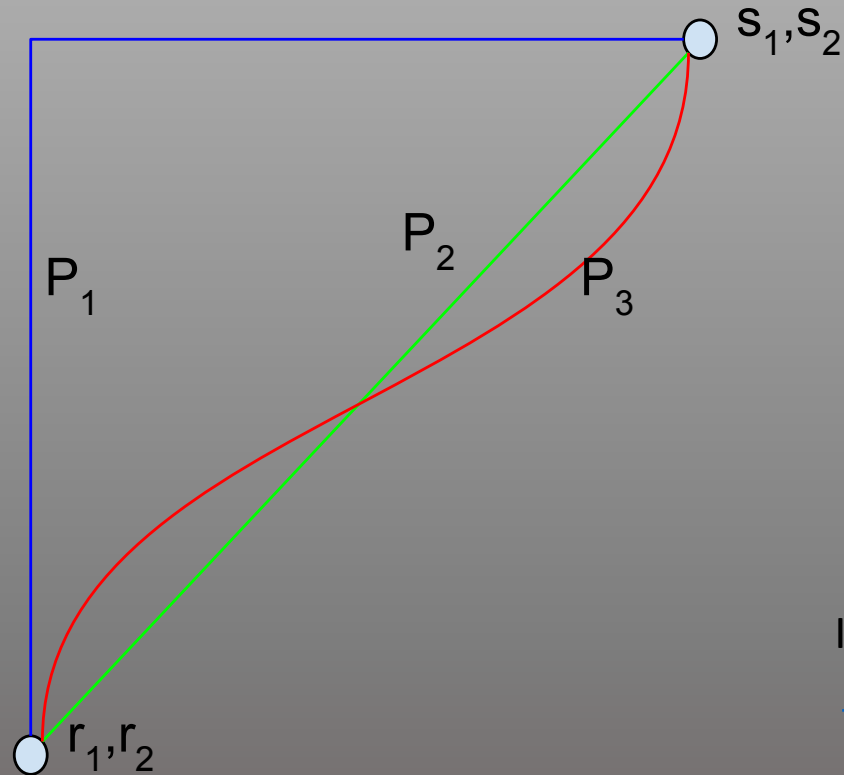




Image reproduced under fair use from  
<https://arxiv.org/pdf/1703.01365.pdf>

# Symmetry Preserving Path Methods = IG

- Consider a non-straightline path  $\gamma : [0,1] \rightarrow \mathbb{R}^n$  from baseline to input.
- WLOG, there exists  $t_0 \in [0,1]$  such that for two dimensions  $i,j$ ,  $\gamma_i(t_0) > \gamma_j(t_0)$ .
  - Otherwise, it is a straight line!
- Let  $(t_1, t_2)$  be the maximum real open interval containing  $t_0$ 
  - such that  $\gamma_i(t) > \gamma_j(t)$  for all  $t$  in  $(t_1, t_2)$ , and
- Then let  $a = \gamma_i(t_1) = \gamma_j(t_1)$ , and  $b = \gamma_i(t_2) = \gamma_j(t_2)$ .
- Define function  $f : x \in [0,1]^n \rightarrow \mathbb{R}$  as
  - 0 if  $\min(x_i, x_j) \leq a$ ,
  - $(b - a)^2$  if  $\max(x_i, x_j) \geq b$ ,
  - $(x_i - a)(x_j - a)$  otherwise.
- Note that  $f$  is symmetric w.r.t.  $x_i$  and  $x_j$

# Symmetry Preserving Path Methods = IG

- Consider a non-straightline path  $\gamma : [0,1] \rightarrow \mathbb{R}^n$  from baseline to input.
  - Let  $(t_1, t_2)$  be the maximum real open interval containing  $t_0$  such that  $\gamma_i(t) > \gamma_j(t)$  for all  $t$  in  $(t_1, t_2)$ , and let  $a = \gamma_i(t_1) = \gamma_j(t_1)$ , and  $b = \gamma_i(t_2) = \gamma_j(t_2)$ .
  - Define function  $f : x \in [0,1]^n \rightarrow \mathbb{R}$  as
    - 0 if  $\min(x_i, x_j) \leq a$ ,
    - $(b - a)^2$  if  $\max(x_i, x_j) \geq b$ ,
    - $(x_i - a)(x_j - a)$  otherwise.
  - Compute attributions of  $f$  at  $x = 1, \dots, 1$  with baseline  $x^0 = 0, \dots, 0$ .
  - Recall function  $f : x \in [0,1]^n \rightarrow \mathbb{R}$  as
    - 0 if  $\min(x_i, x_j) \leq a$ ,
    - $(b - a)^2$  if  $\max(x_i, x_j) \geq b$ ,
    - $(x_i - a)(x_j - a)$  otherwise.
- 
- 
- the function is a constant
  - the attribution of  $f$  is zero to all variables
  - the integrand of attribution of  $f$  is
    - $\gamma_j(t) - a$  to  $x_i$ , and
    - $\gamma_i(t) - a$  to  $x_j$
    - *one is larger than the other by our design.*
- Integrating, it follows that  $x_j$  gets a larger attribution than  $x_i$ , contradiction

# Experimental Results - I

- GoogLeNet
- ImageNet
- Black image as baseline
- Diabetic retinopathy

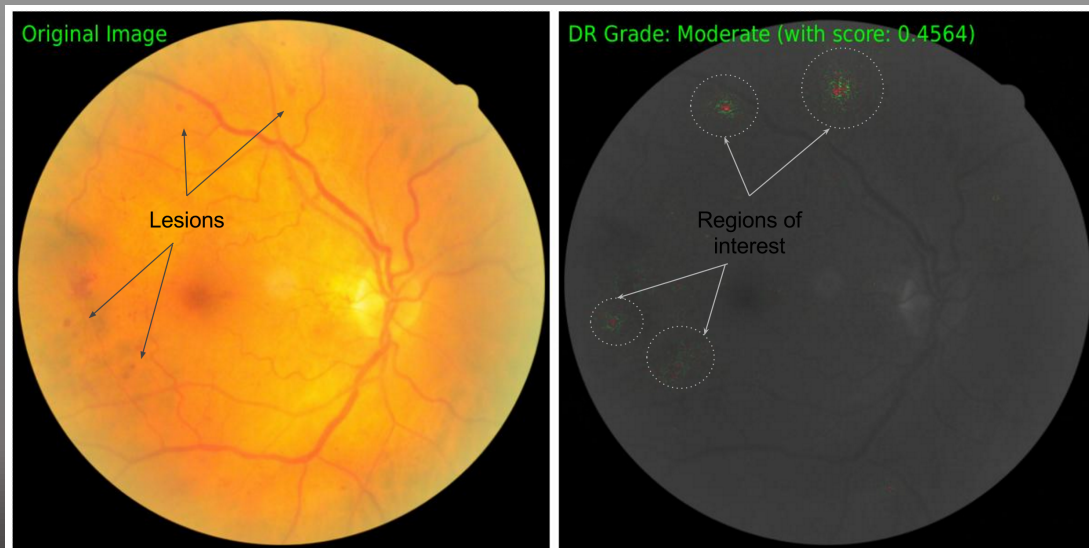


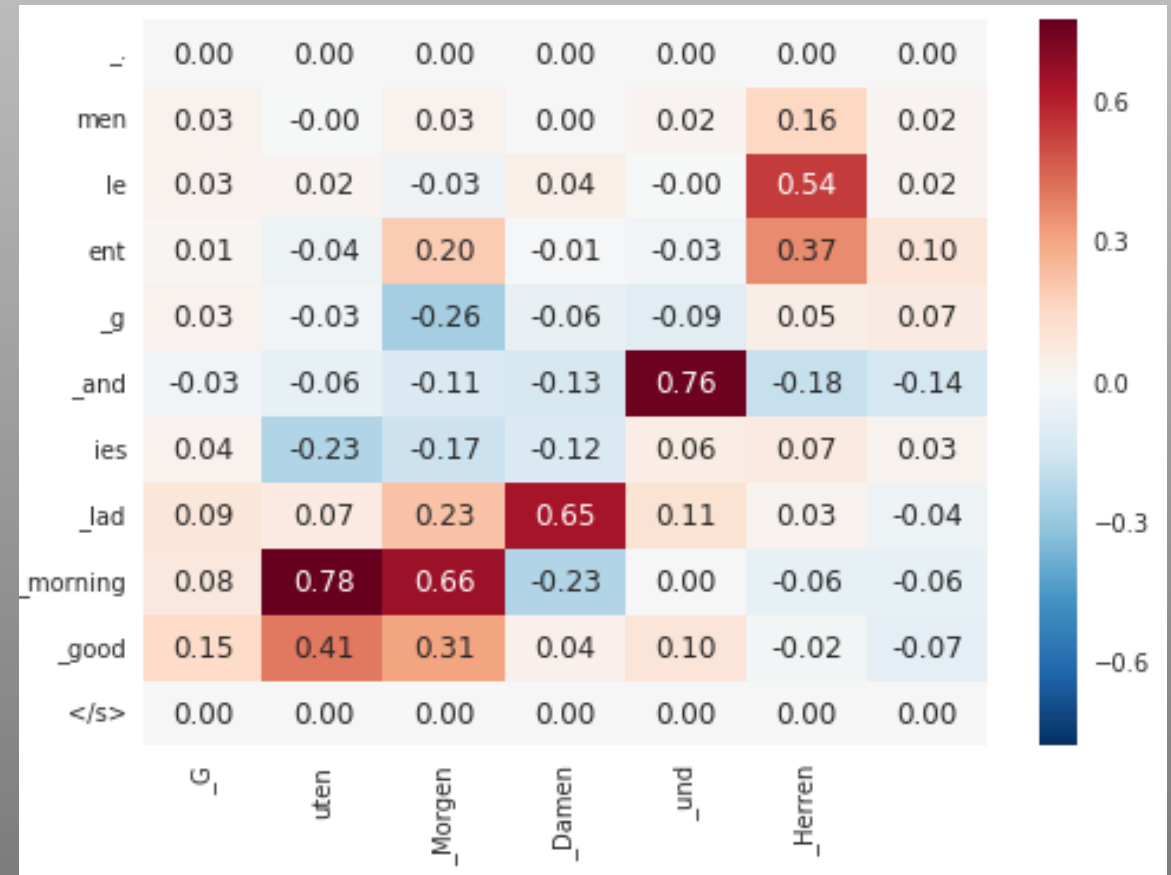
Image reproduced under fair use from  
<https://arxiv.org/pdf/1703.01365.pdf>



# Experimental Results II

- Model for question classification
  - text categorization architecture
  - WikiTableQuestions dataset
- IG to attribute the question terms
- Goal: Identify trigger phrases for answer type.
- Baseline = 0 embedding vector.

how many townships have a population above 50 ? [prediction: NUMERIC]  
 what is the difference in population between fora and masilo [prediction: NUMERIC]  
 how many athletes are not ranked ? [prediction: NUMERIC]  
 what is the total number of points scored ? [prediction: NUMERIC]  
 which film was before the audacity of democracy ? [prediction: STRING]  
 which year did she work on the most films ? [prediction: DATETIME]  
 what year was the last school established ? [prediction: DATETIME]  
 when did ed sheeran get his first number one of the year ? [prediction: DATETIME]  
 did charles oakley play more minutes than robert parish ? [prediction: YESNO]



IG Attributions for Language Translation

Image reproduced under fair use from  
<https://arxiv.org/pdf/1703.01365.pdf>

# Conclusions

- Primary contribution
  - a new method called integrated gradients
  - Attribute a DNN prediction to its inputs
  - Implemented using 10-1000 or so calls to the gradient operator
  - Applied to a variety of deep networks.
- Secondary contribution
  - axiomatic framework
  - cost-sharing from economics.
  - Axiomatic; hence, evaluation not strongly influence by
    - data artifacts,
    - network's artifacts
    - artifacts of the method.
    - The axiomatic approach rules out artifacts of the last type.