Why did the network make this prediction?

Ankur Taly (Google Inc.)

Joint work with Mukund Sundararajan and Qiqi Yan

Some Deep Learning Successes

Google weaves machine learning into new Google Photos features

Facebook Uses Deep Learning To Make Faster Translations

By Tyler Lee on 05/10/2017 04:09 PDT

Google AI Matches Diabetic Retinopathy Screens

Google's AlphaGo AI beats the world's best human Go player

Posted May 23, 2017 by Darrell Etherington (@etherington)

captricity deepomatic sentient Sighthound

terraloupe imagenii pilot.ai

DigitalGenius LEAPMIND Arya

neuro twentybn

FRACTAL INDUSTRIES MARIANA DEEPGRAM

True AI KOVID ditto Gridspace

CORE AI: VOICE INTERFACE

VIV Mobvoi 出门问问 MaluubA

OBEN

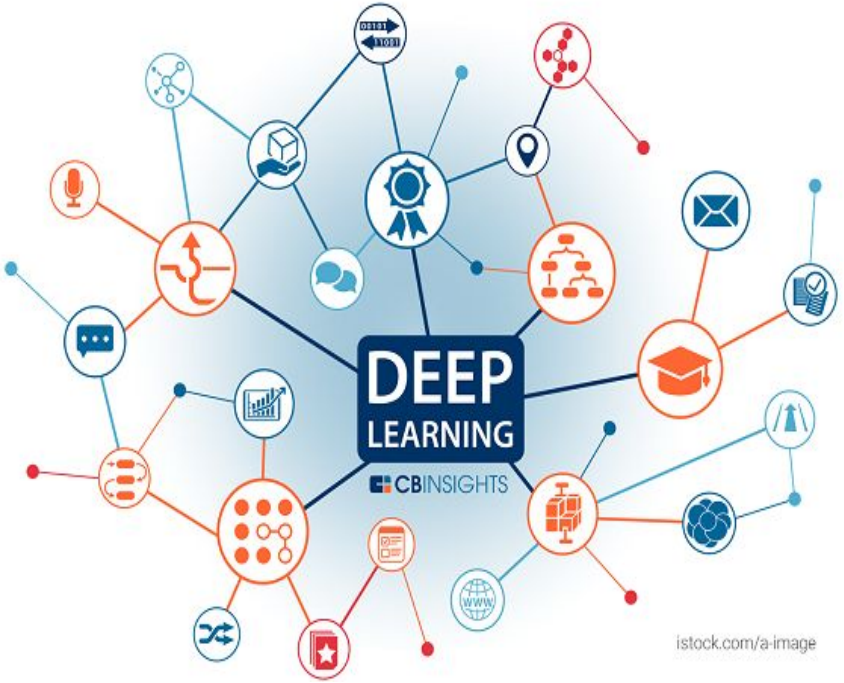
ROBOTICS & AUTO

netradyne AdasWorks

drive.ai Rokid 图灵机器人 TURING ROBOT

comma.ai

HEALTHCARE



E-COMMERCE

Reflektion CORTEXICA

VISENZE STAQU

ACQUIRED

(2014-2016YTD)

Etsy Blackbird a Orbeus IBM AlchemyAPI

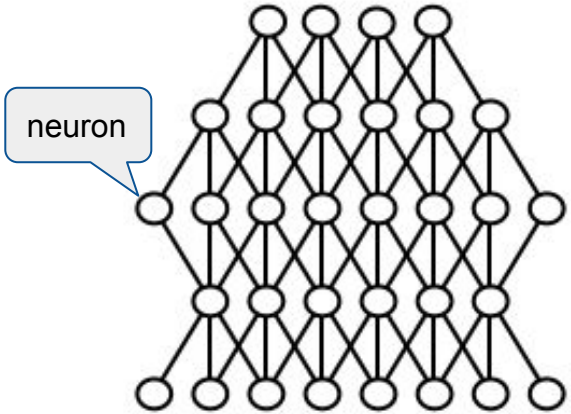
intel Movidius nervana SYSTEMS

Perceptio vocaliq turi

Source: https://www.cbsinsights.com

Deep Neural Networks

Output
(Image label, next word, next move, etc.)



Input
(Image, sentence, game position, etc.)

Flexible model for learning arbitrary non-linear, non-convex functions

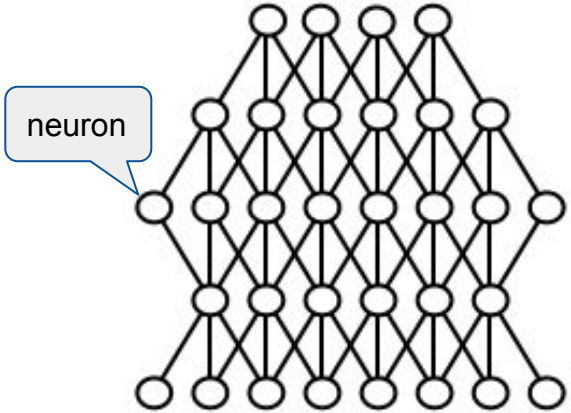
Transform input through a network of neurons

Each neuron applies a non-linear activation function (σ) to its inputs

$$n_3 = \sigma(w_1 . n_1 + w_2 . n_2 + b)$$

Deep Neural Networks

Output
(Image label, next word, next move, etc.)



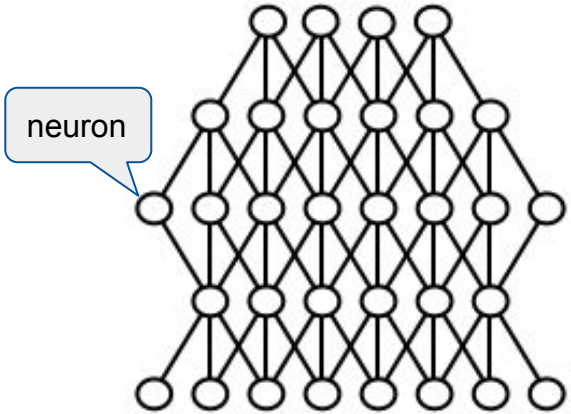
Input
(Image, sentence, game position, etc.)

Highly expressive model

- Many different architectures
  - Can encode linear models, decision trees, markov models, and their combinations
- Less feature engineering needed
  - Lower layers can automatically extract complex features from raw inputs and feed them above

Deep Neural Networks

Output
(Image label, next word, next move, etc.)



Input
(Image, sentence, game position, etc.)

Used to be notoriously hard to train

Rapid advances over the last few years

- **Stochastic gradient descent**
- Lots of training data
- Several “training” tricks
- Better hardware and software

We can now train networks with millions of parameters over billions of training examples!

Understanding Deep Neural Networks

We understand them enough to:

- Design architectures for complex learning tasks
  - For both supervised, unsupervised training datasets
- Train these architectures to favorable optima
- Help them generalize beyond training set (prevent overfitting)

But, a trained network still remains a black box to humans

Our long-term objective

Understanding the input-output behavior of Deep Networks

i.e., *we ask why did it make this prediction on this input?*

Benefits:

- Debug and understand models
- Build trust in the model
- Surface an explanation to the end-user
- Intellectual curiosity

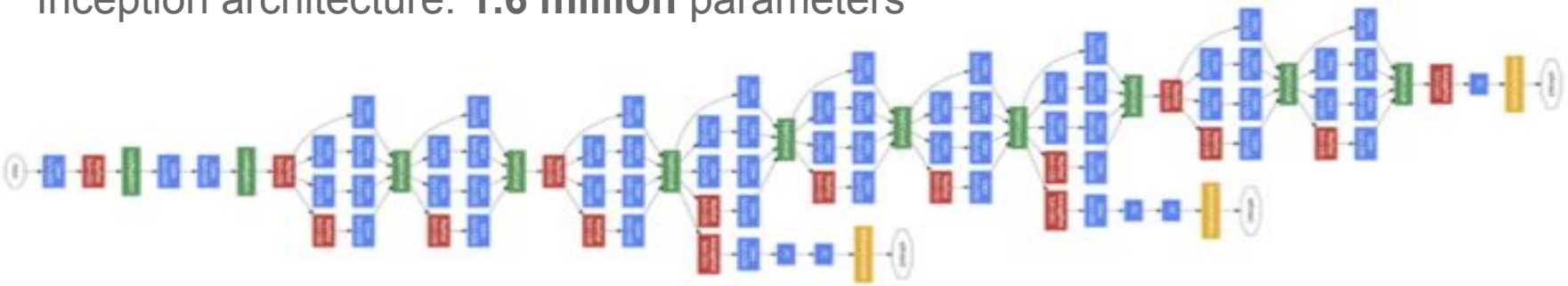
Why did the network label this image as “drilling platform”?

Retinal Fundus Image

Why does the network label this image with “mild” Diabetic Retinopathy?

Analytical Reasoning is very hard

Inception architecture: **1.6 million** parameters



- Modern architectures are way too complex for analytical reasoning
- The meaning of individual neurons is not human-intelligible
- **Faithfulness** vs. **Interpretability**

Our approach

- Explain the network’s behavior in terms of the input
- What “features” of the input were important for this prediction?

(Getting the right problem statement is the hard part here)

The Attribution Problem

Distribute the prediction score to each input feature in proportion to its contribution to the prediction with respect to a certain baseline input

- Input features could be pixels, words etc.
- Baseline input is one where the prediction is neutral, e.g., black image
- The amount assigned to each feature is its attribution
- Large attribution indicates feature importance

Outline

- Our attribution method: Integrated Gradients
- Applications of the method
- Justifying Integrated Gradients
- Discussion

Naive approach: Ablations

Ablate each input feature and measure the change in prediction

- Costly, especially for dense models with (224*224*3) pixel features
- Over or under attribution of interactive features
  - E.g., Query=”Facebook” AND Domain=”facebook.com” IMPLIES high click through rate
- Unrealistic inputs

Gradient-based Attribution

Attribute using gradient* of the output w.r.t each base input feature

Attribution for feature x_i is $x_i * \partial y/\partial x_i$

- Standard approach for understanding linear models
  - Same as feature weights
- First-order approximation for non-linear models

Inception on ImageNet

Drilling platform (0.986)	Crane (0.002)	Container ship (0.001)	Pier (0.001)	Dock (0.001)

Visualizing Attributions

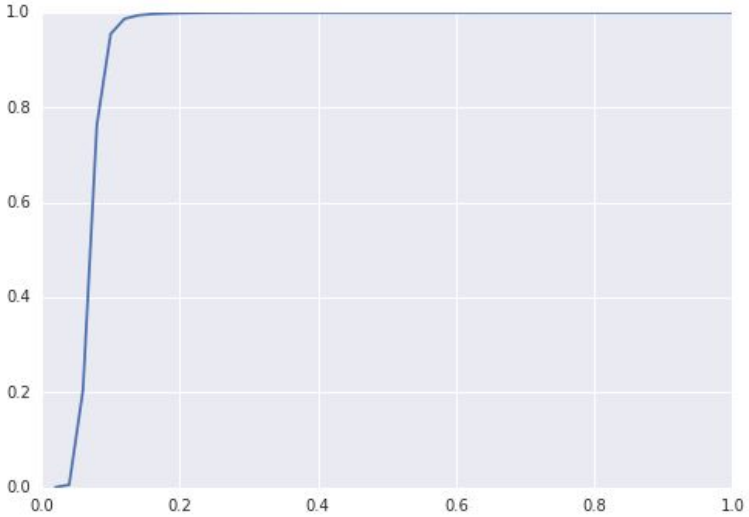
Visualization: Use (normalized) attribution as mask/window over image

Attribution using gradients

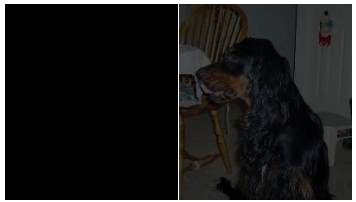
Attribution using gradients

Saturation

Prediction Score

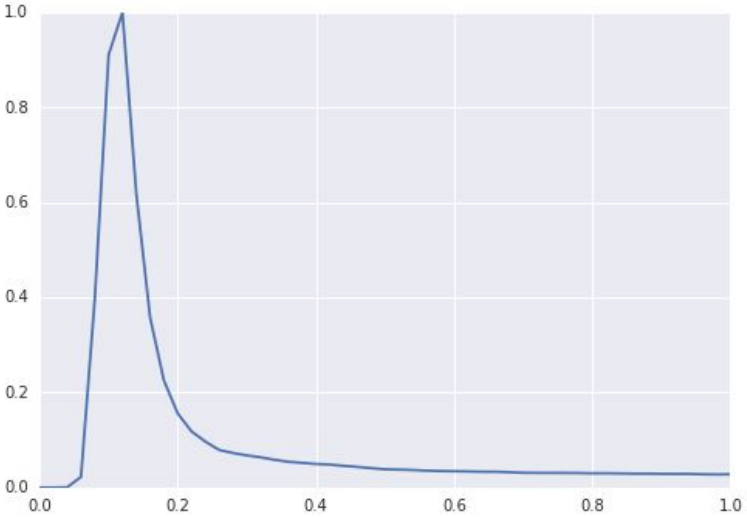


Intensity α

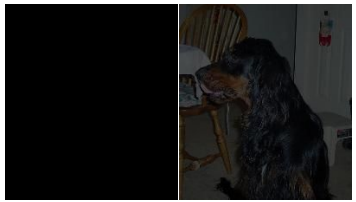


Saturation

Average pixel gradient (normalized)

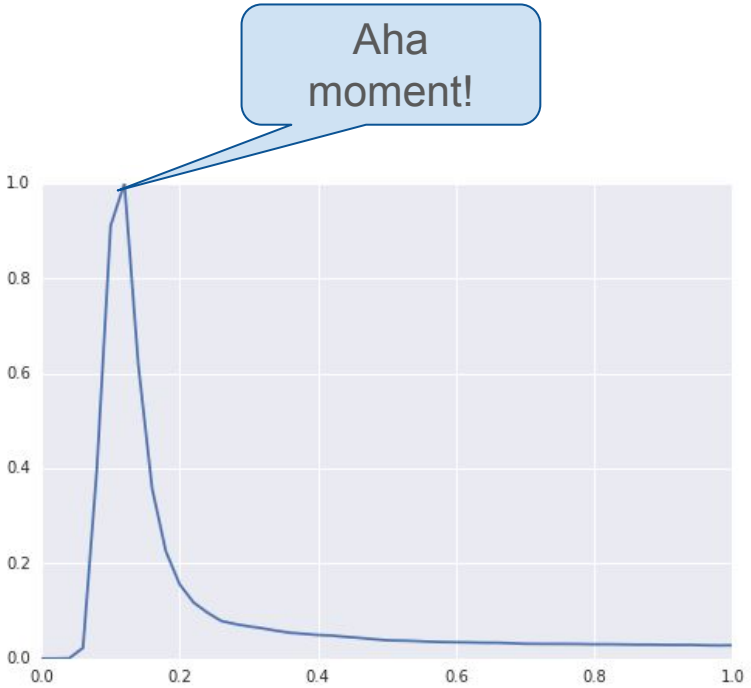


Intensity α

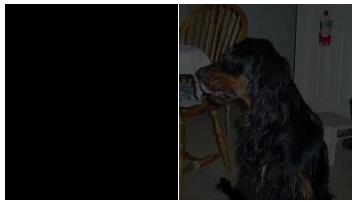


Saturation

Average pixel gradient (normalized)



Intensity α



Saturation Animated

- Compute gradient for images that range from black to actual image
- Use these gradients as attributions
- Gif the sequence of resulting visualizations
- Blue screen starts animation

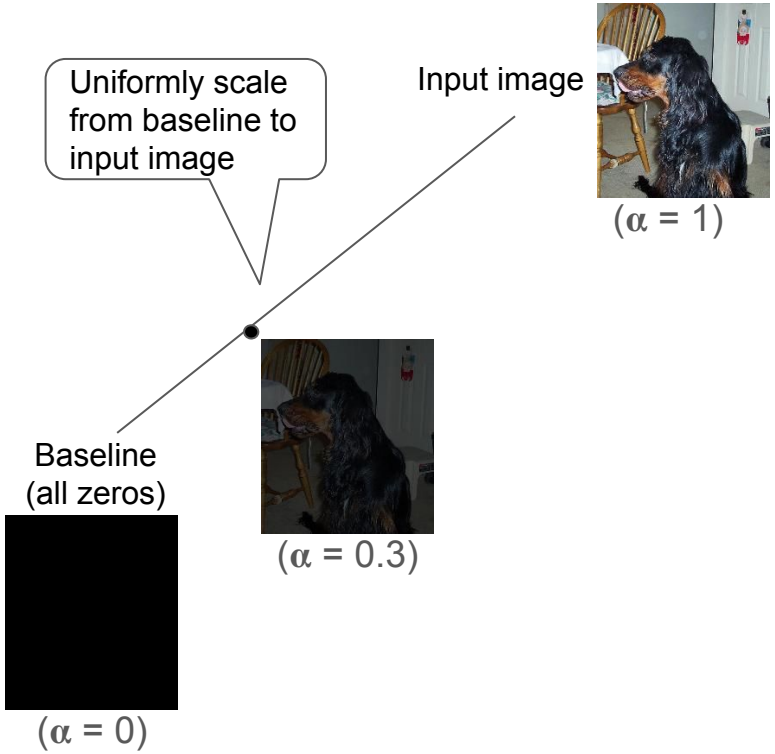
Saturation occurs...

- across images
  - Not just the two images we discussed
- across layers
  - Not just the output layer
- across networks
  - Not just Inception on ImageNet
  - Severity varies

(see this paper for details)

Integrated Gradients

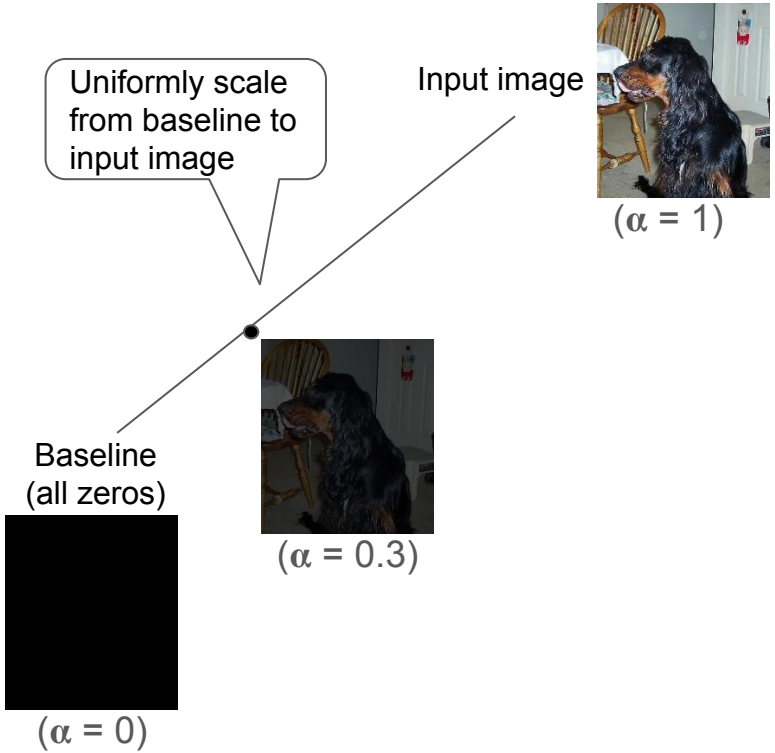
The method: Integrated gradients



Construct a sequence of images interpolating from a baseline (black) to the actual image

Average the gradients across these images

The method: Integrated gradients



Mathematically,

$$IG_i(image) = image_i * \int_{0}^{-1} \nabla F_i(\alpha * image)\, d\alpha$$

where:

- F is the prediction function for the label
- $image_i$ is the intensity of the i^{th} pixel
- $IG_i(image)$ is the integrated gradient w.r.t. the i^{th} pixel, i.e., attribution for i^{th} pixel

Integrated gradients for Inception

Original image (**Drilling platform**)

Gradient at image

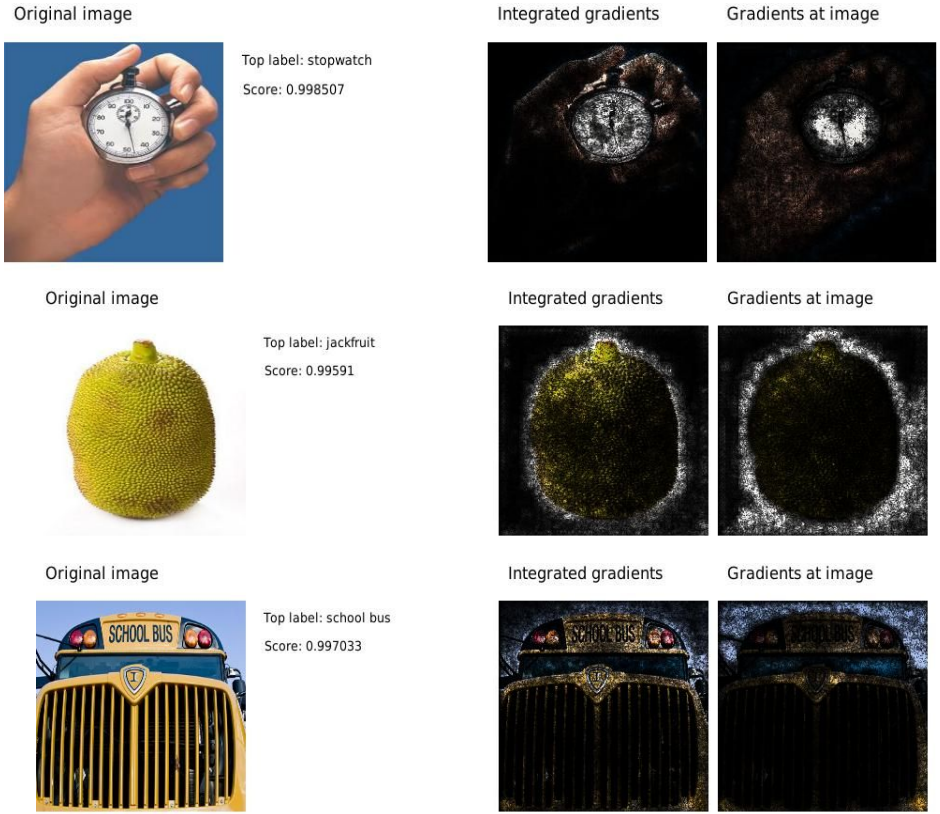
Integrated gradient

Integrated gradients for Inception

Original image (**Drilling platform**)

Gradient at image

Integrated gradient



Many more Inception+ImageNet examples here

Misconception

Human label: accordion
Network's top label: toaster

Misconception

Human label: accordion
Network's top label: toaster

Integrated gradient

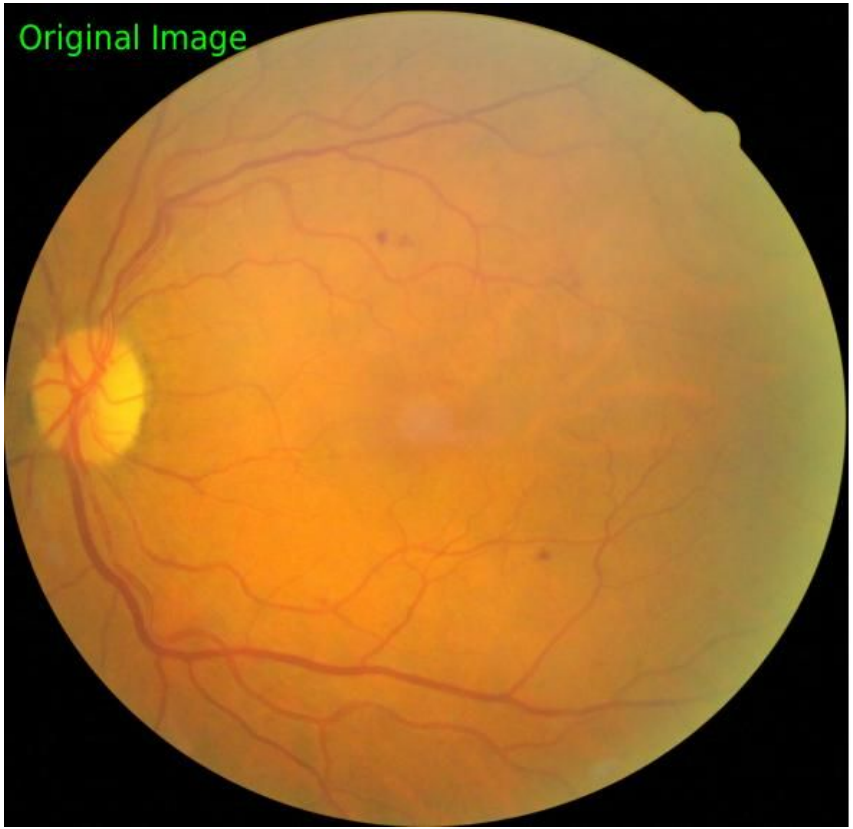
Diabetic Retinopathy

Diabetes complication that causes damage to blood vessels in the eye due to excess blood sugar.

An Inception-based network for predicting diabetic retinopathy grade from retinal fundus images achieves **0.97 AUC** [JAMA paper]

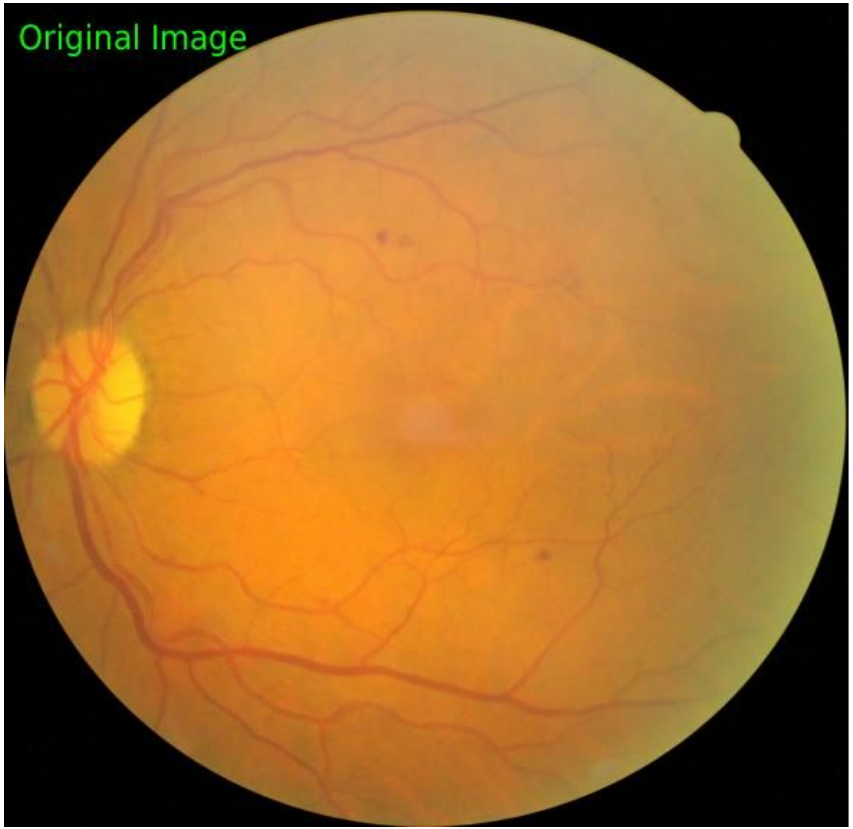
On what basis, does the network predict the DR grade?

Diabetic Retinopathy

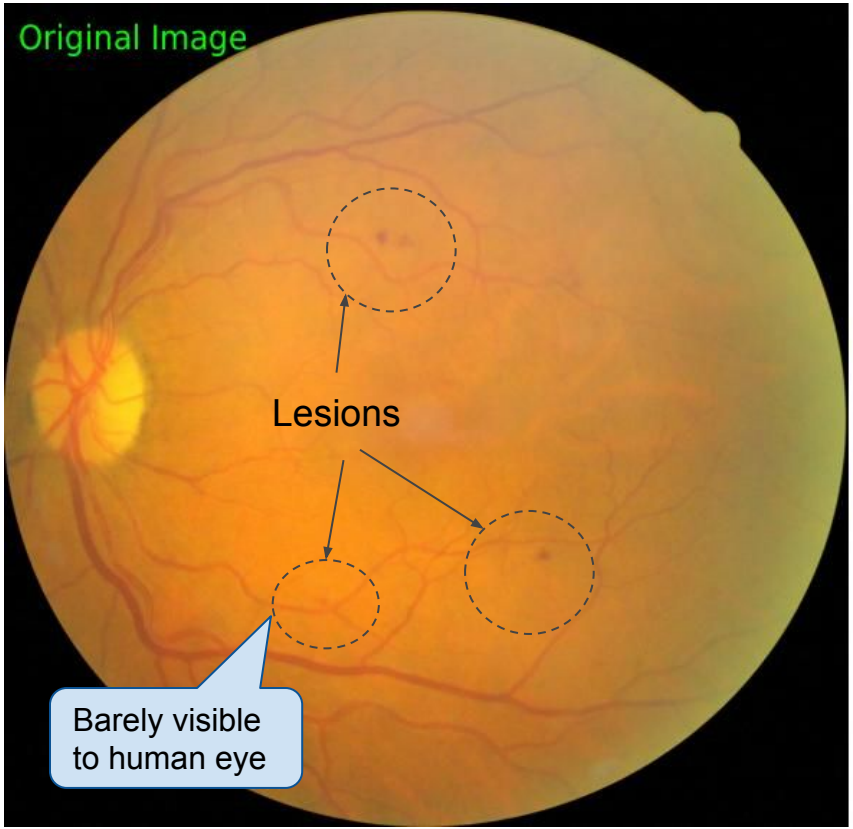


Predicted DR grade: Mild

Attributions using Integrated Gradients



Attributions using Integrated Gradients



Text Classification

- We have a data set of questions and answers
  - Answer types include numbers, strings, dates, and yes/no
- Can we predict the answer type from the question?
  - Answer: Yes using a simple feedforward network
- Can we tell which words were indicative of the answer type?
  - Enter attributions
- **Key issue**: What is the analog of the black image?
  - Answer: the zero embedding vector

Text Classification

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]

Red is positive attribution
Blue is negative attribution
Shades interplolate

Text Classification

Several sensible results, can almost harvest these as grammar rules

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]

Overfitting?

Negative signals too

Many Other Applications

- Drug Discovery
  - Why part of the molecule causes it to bind to this protein?
- Search Ranking
  - What makes one result rank higher than another?
- Language translation
  - Which input word does this output word correspond to?

Justifying Integrated Gradients

How do you evaluate an attribution method?

How do you evaluate an attribution method?

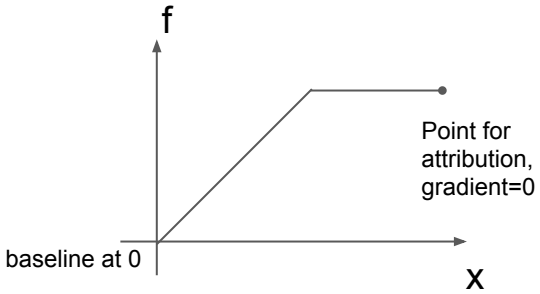
Our approach:

- Define a set of reasonable axioms for attribution methods
- Show that integrated gradients satisfies them

Sensitivity Axioms

Sensitivity: If starting from baseline, varying a variable changes the output, then the variable should receive some attribution.

- Pure gradients do not satisfy this when predictions saturate.



Insensitivity: A variable that has no effect on the output gets no attribution.

Functional Axioms

Implementation Invariance:
Two functionally equivalent networks have identical attributions for all inputs and baseline

Linearity:
If the function **F** is a linear combination of two functions $\mathbf{F_1}$,$\mathbf{F_2}$ then the attributions for **F** are a linear combination of the attributions for $\mathbf{F_1}$,$\mathbf{F_2}$

Symmetry:
If a function is symmetric across two input variables then the variables should receive identical attribution

An Accounting Axiom

Completeness: Sum(attributions) = F(input) - F(baseline)

Break down the predicted click through rate (pCTR) of an ad like:

- 55% of pCTR is because it’s at position 1
- 25% is due to its domain (a popular one)
- ...

Result

Theorem: Integrated Gradients is the unique method satisfying:

- Sensitivity, Insensitivity
- Implementation Invariance, Linearity, Symmetry
- Completeness

up to the errors from approximating integration.

Historical note:

- It's essentially the Aumann-Shapley method in cost sharing, which has a similar characterization. (Friedman 2004)

A note on the baseline

- The need for a baseline is central to any explanation method
  - In a sense, it is the counterfactual for causal reasoning
- The network must have a truly neutral prediction at the baseline input

Highlights of Integrated Gradients

- Easy to implement
  - Gradient calls on a bunch of scaled down inputs
  - No instrumentation of the network, no new training
- Widely applicable
- Backed by an axiomatic guarantee

References

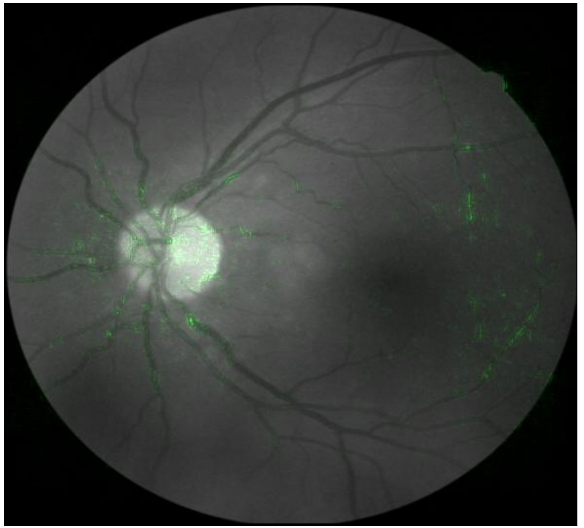
- Google Data Science Blog: Attributing a deep network's prediction to its input
- Paper [ICML 2017]: Axiomatic Attribution for Deep Networks

Discussion

Interpreting Attributions

Hypothetically, suppose we train a model to predict age from Retina images

How do we interpret the resulting attributions?



Interpreting Attributions

Attributions help when the causality is somewhat known (to the human)

- Confirm attributions to known* causal features is high
- Identify new features that contribute to the prediction

When the causality is unknown, attributions alone don’t offer much insight

- Cluster attributions across examples?
- Explore feature interactions?

Other Limitations

- Attributions do not give you a global understanding of the model
- Attributions do not deal with correlations
    - data understanding vs. model understanding

A different problem statement [Liang et al.]

- *“Explain a prediction in terms of the training data”*
- Paper: Understanding Black-box Predictions via Influence Functions [ICML 2017]

Questions?