Explaining Machine Learning Models

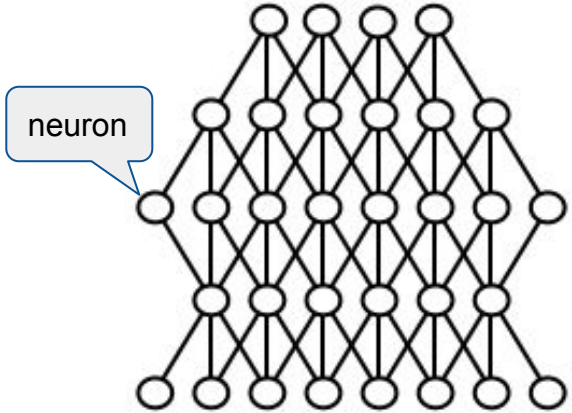
Ankur Taly, Fiddler Labs
ankur@fiddler.ai

Joint work with Mukund Sundararajan[1], Qiqi Yan[1], Kedar Dhamdhere[1], and Pramod Mudrakarta[2] and colleagues at Fiddler labs
[1]Google, [2]U Chicago

Machine Learning Models are Everywhere!

Output

(Label, sentence, next word, next move, etc.)



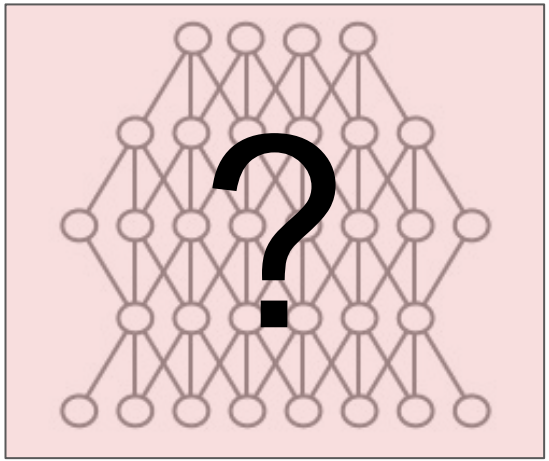
Input

(Image, sentence, game position, etc.)

Problem: Machine Learning Model is a Black Box

Output

(Label, sentence, next word, next move, etc.)



Input

(Image, sentence, game position, etc.)

Why did the model label this image as “fireboat”?

Top label: **“fireboat”**

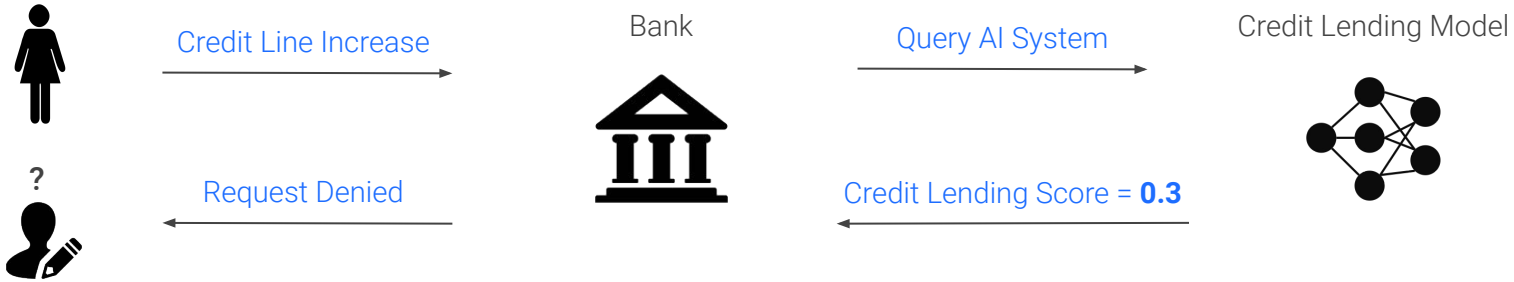
Why did the model label this image as “clog”?

Top label: **“clog”**

Beautiful design and execution, both the space itself and the food. Many interesting options beyond the excellent traditional Italian meatball. Great for the Financial District.

Why did the network predict **positive sentiment** for this review?

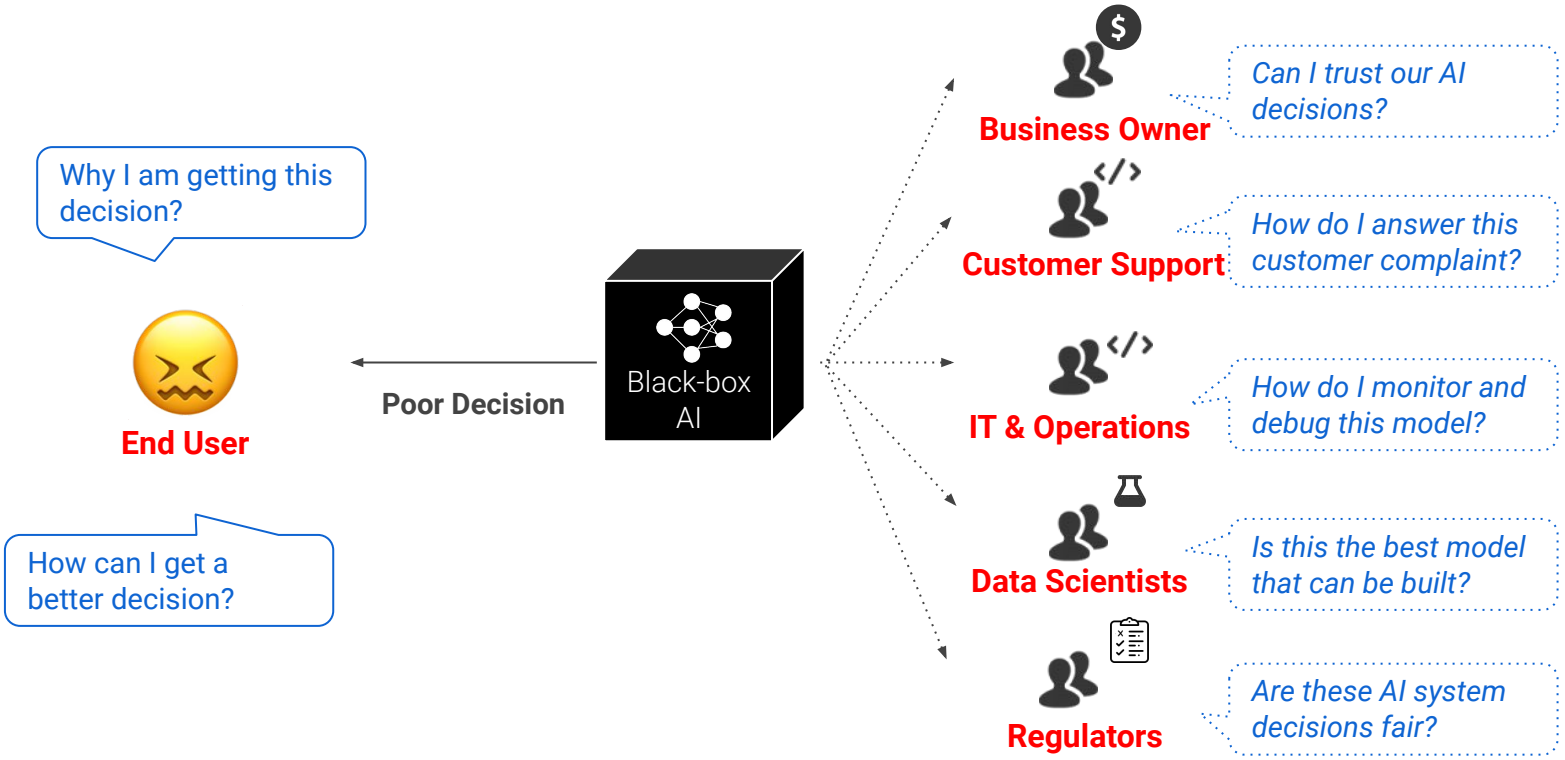
Example: Credit Lending in a black-box ML world



Why? Why not? How?

Fair lending laws [ECOA, FCRA] require credit decisions to be explainable

Black-Box Models create Confusion & Doubt



So, how do you explain a model?

The Attribution Problem

Attribute a model's prediction on an input to features of the input

Examples:

- Attribute a lending model's prediction to its features
- Attribute an object recognition network's prediction to its pixels
- Attribute a text sentiment network's prediction to individual words

A **reductive formulation** of "why this prediction" but surprisingly useful :-)

Applications of Attributions

- Debugging network predictions
- Generating an explanation for the end-user
- Analyzing network robustness
- Assessing prediction confidence

Agenda

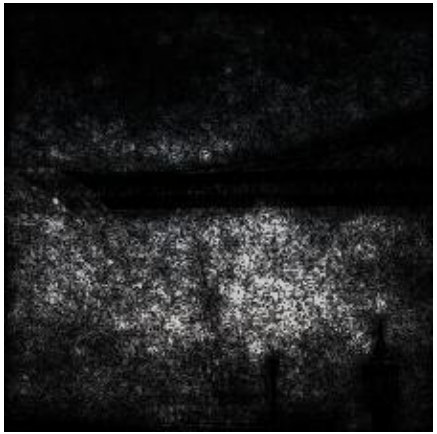
- Two Attribution Methods
  - Integrated Gradients
  - Shapley Values
- Applications of attributions
- Discussion

Naive approaches

- **Ablations**: Drop each feature and note the change in prediction
  - Computationally expensive, Unrealistic inputs, Misleading when features interact

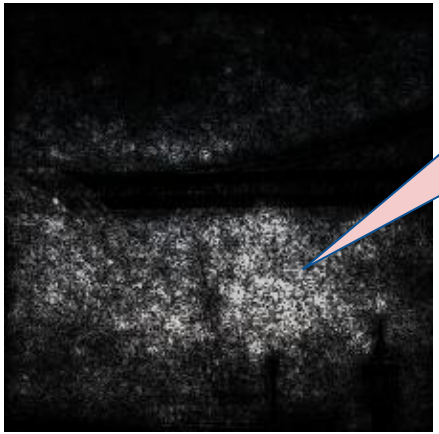
Naive approaches

- **Ablations**: Drop each feature and note the change in prediction
  - Computationally expensive, Unrealistic inputs, Misleading when features interact
- **Feature*Gradient**: Attribution for feature x_i is $x_i * \partial y/\partial x_i$

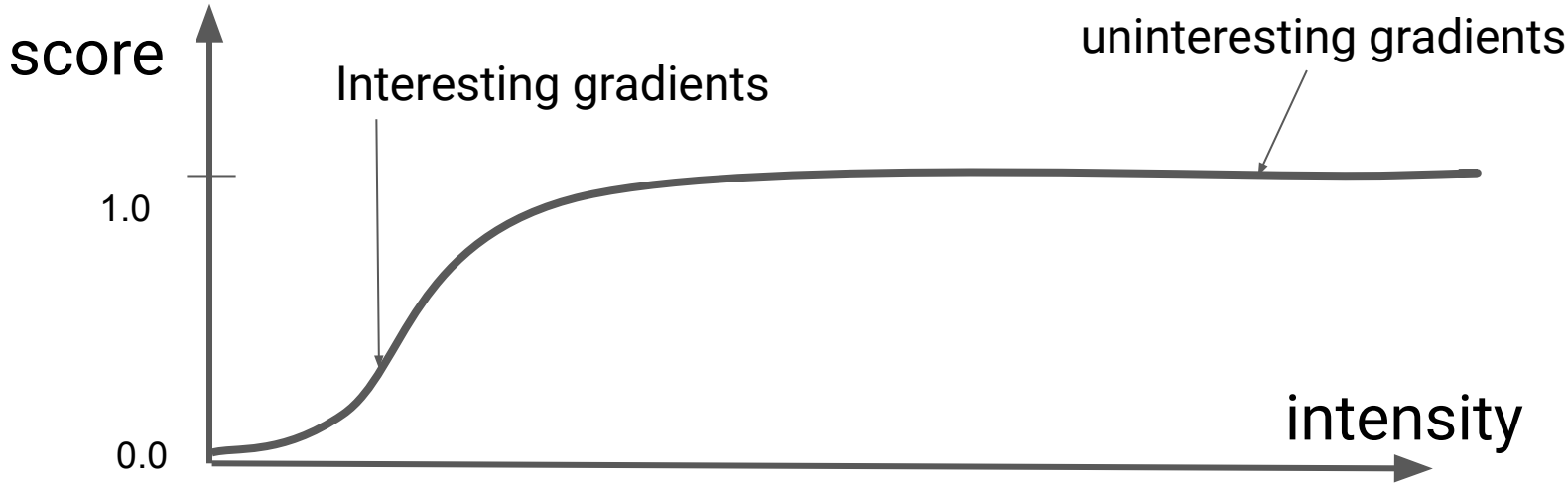


Naive approaches

- **Ablations**: Drop each feature and note the change in prediction
  - Computationally expensive, Unrealistic inputs, Misleading when features interact
- **Feature*Gradient**: Attribution for feature x_i is $x_i * \partial y/\partial x_i$



Gradients in the vicinity of the input seem like noise



Baseline

… scaled inputs ...

Input

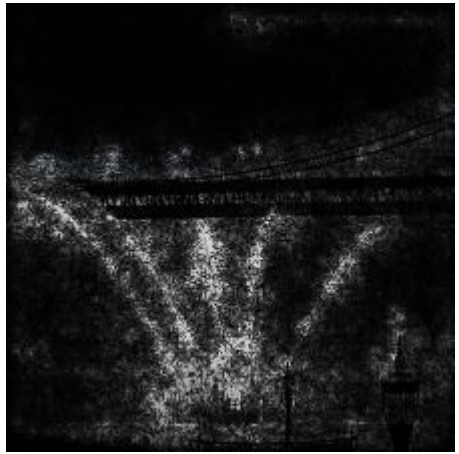
… gradients of scaled inputs ….

Our Method: **Integrated Gradients [ICML 2017]**

$$IG(input, base) ::= (input - base) * \int_{0}^{1} \nabla F(\alpha * input + (1-\alpha) * base)\, d\alpha$$

Original image

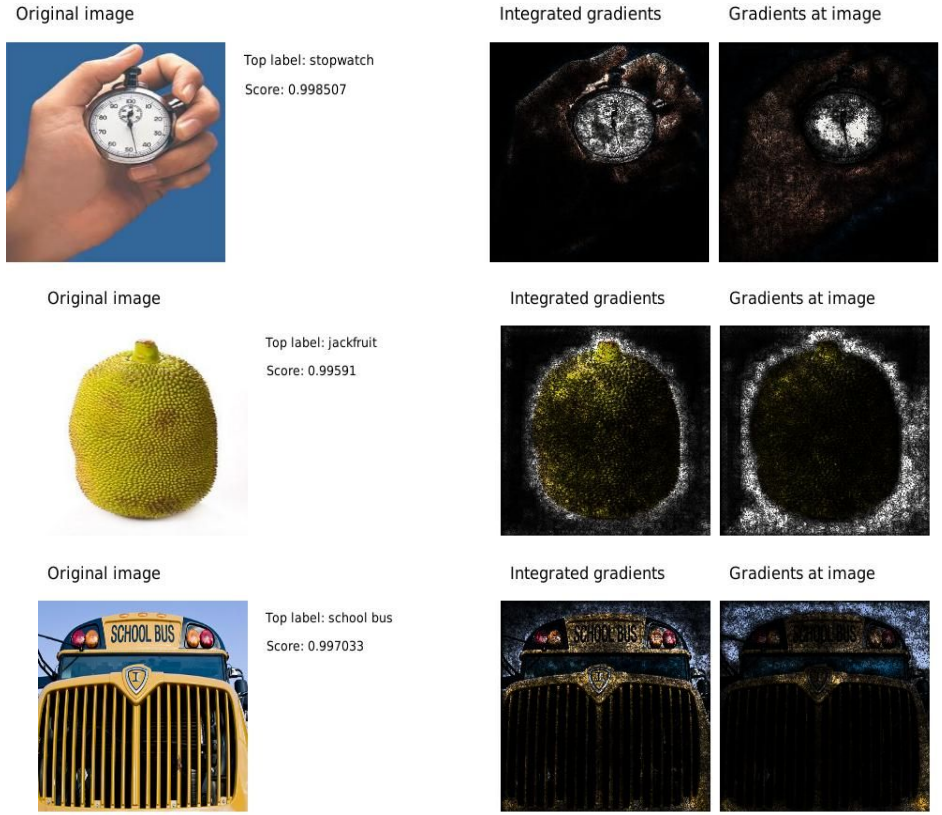
Integrated Gradients



What is a baseline?

- **A baseline is an informationless input that results in a neutral prediction**
  - E.g., Black image for image models
  - E.g., Empty text or zero embedding vector for text models
- Integrated Gradients explains the diff **F(input) - F(baseline)**

- Baselines (or Norms) are essential to all explanations [Kahneman-Miller 86]
  - E.g., A man suffers from indigestion. Doctor blames it to a stomach ulcer. Wife blames it on eating turnips. Both are correct relative to their baselines
  - The baseline may also be an important analysis knob



Many more Inception+ImageNet examples here

Evaluating an Attribution Method

- Ablate top attributed features and examine the change in prediction
  - Issue: May introduce artifacts in the input (e.g., the square below)

- Compare attributions to (human provided) groundtruth on “feature importance”
  - Issue 1: Attributions may appear incorrect because the network reasons differently
  - Issue 2 : **Confirmation bias**

Evaluating an Attribution Method

- Ablate top attributed features and examine the change in prediction
  - Issue: May introduce artifacts in the input (e.g., the square below)

- Compare attributions to (human provided) groundtruth on “feature importance”
  - Issue 1: Attributions may appear incorrect because the network reasons differently
  - Issue 2 : **Confirmation bias**

The mandate for attributions is to be faithful to the network’s reasoning

Our Approach: Axiomatic Justification

- List **desirable criteria (axioms)** for an attribution method
- Establish a uniqueness result: X is the **only** method that satisfies these criteria

Axioms

- **Insensitivity:** A variable that has no effect on the output gets no attribution
- **Sensitivity**: If baseline and input differ in a single variable, and have different outputs, then that variable should receive some attribution
- **Linearity preservation**: Attributions(α*F1 + ß*F2) = α*Attributions(F1) + ß*Attributions(F2)
- **Implementation invariance**: Two networks that compute identical functions for all inputs get identical attributions
- **Completeness**: Sum(attributions) = F(input) - F(baseline)
- **Symmetry**: Symmetric variables with identical values get equal attributions

Result

Theorem [ICML 2017]: Integrated Gradients is the **unique** path-integral method satisfying: Sensitivity, Insensitivity, Linearity preservation, Implementation invariance, Completeness, and Symmetry

Historical note:

- Integrated Gradients is the **Aumann-Shapley method** from cooperative game theory, which has a similar characterization; see [Friedman 2004]

Debugging network behavior

Why is this image labeled as “clog”?

Original image

“Clog”

Why is this image labeled as “clog”?

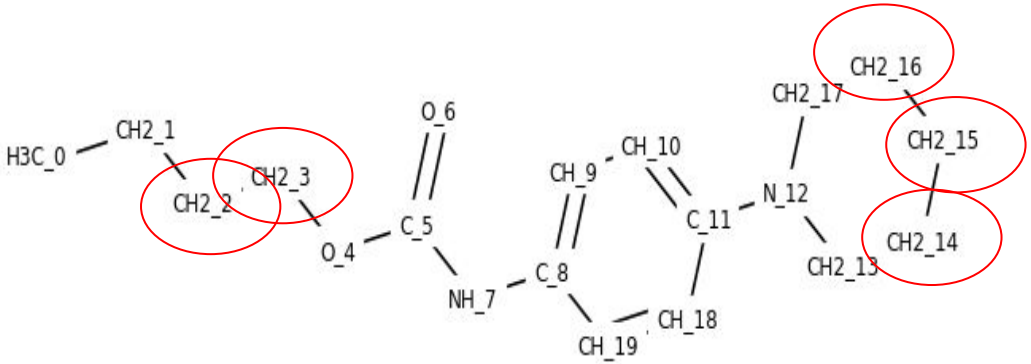
Original image

Integrated Gradients
(for label “clog”)

“Clog”

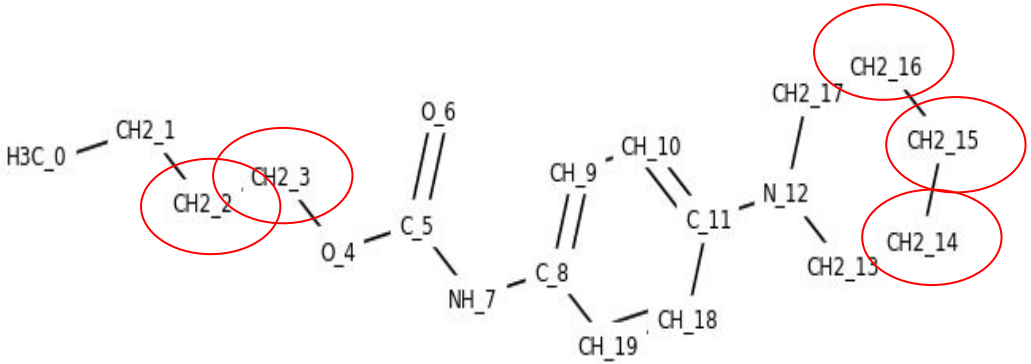
Detecting an architecture bug

- Deep network [Kearns, 2016] predicts if a molecule binds to certain DNA site
- **Finding**: Some atoms had identical attributions despite different connectivity



Detecting an architecture bug

- Deep network [Kearns, 2016] predicts if a molecule binds to certain DNA site
- **Finding**: S nt connectivity

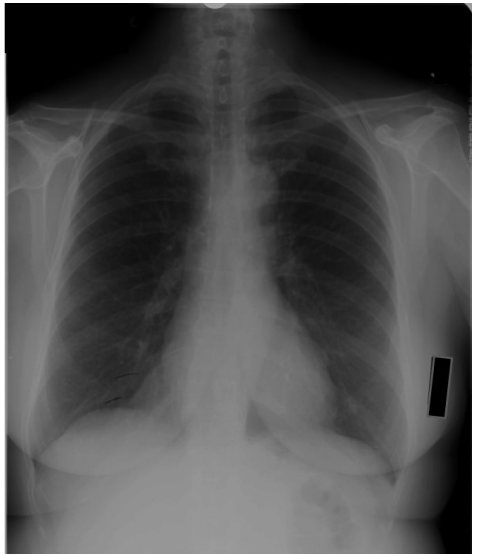


- **Bug**: The architecture had a bug due to which the convolved bond features did not affect the prediction!

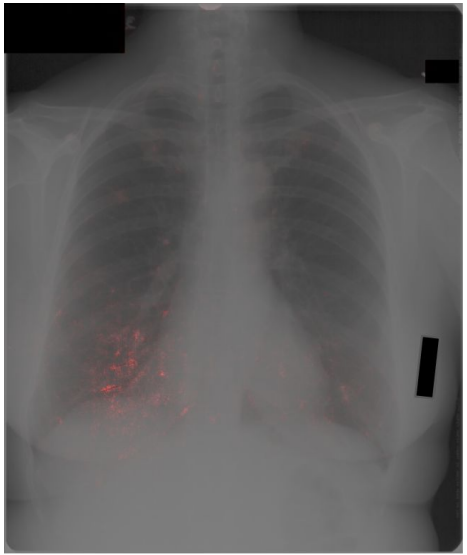
Detecting a data issue

- Deep network predicts various diseases from chest x-rays

Original image

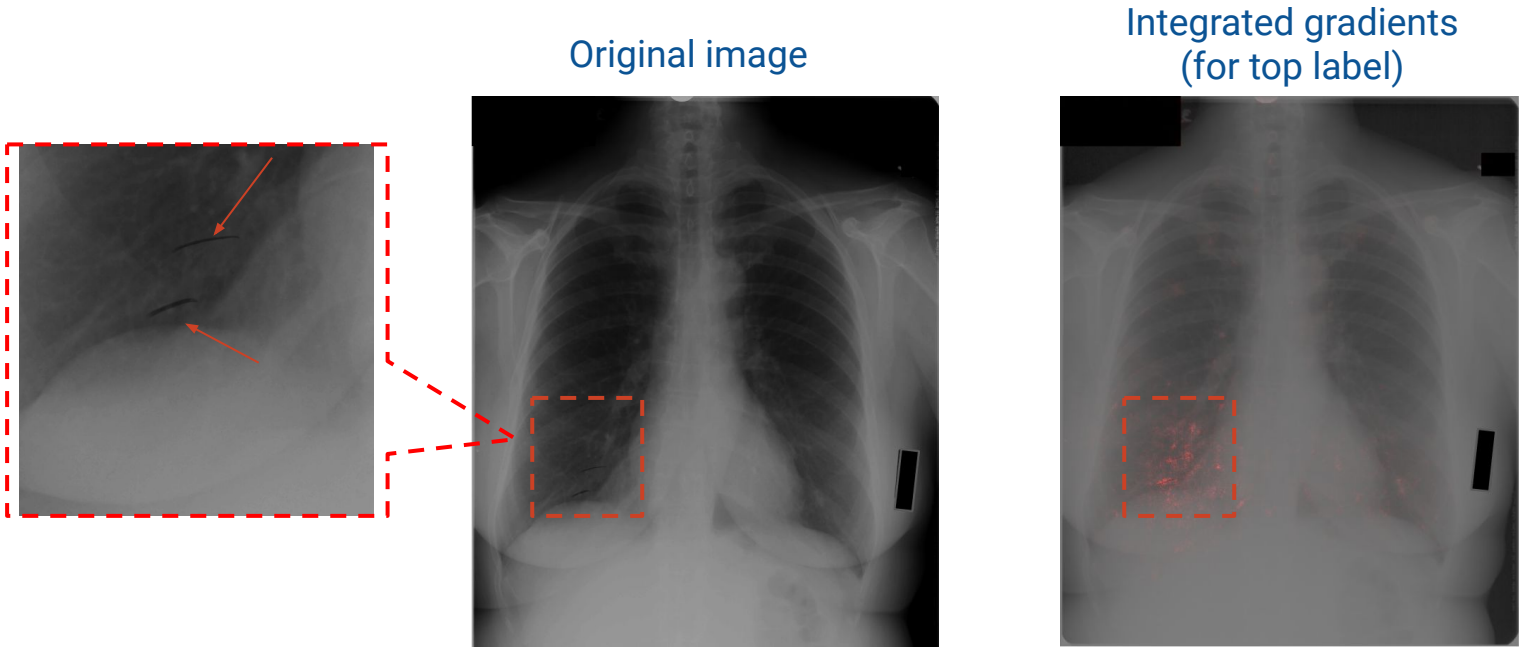


Integrated gradients
(for top label)



Detecting a data issue

- Deep network predicts various diseases from chest x-rays
- **Finding**: Attributions fell on radiologist's markings (rather than the pathology)



Generating explanations for end users

Retinal Fundus Image

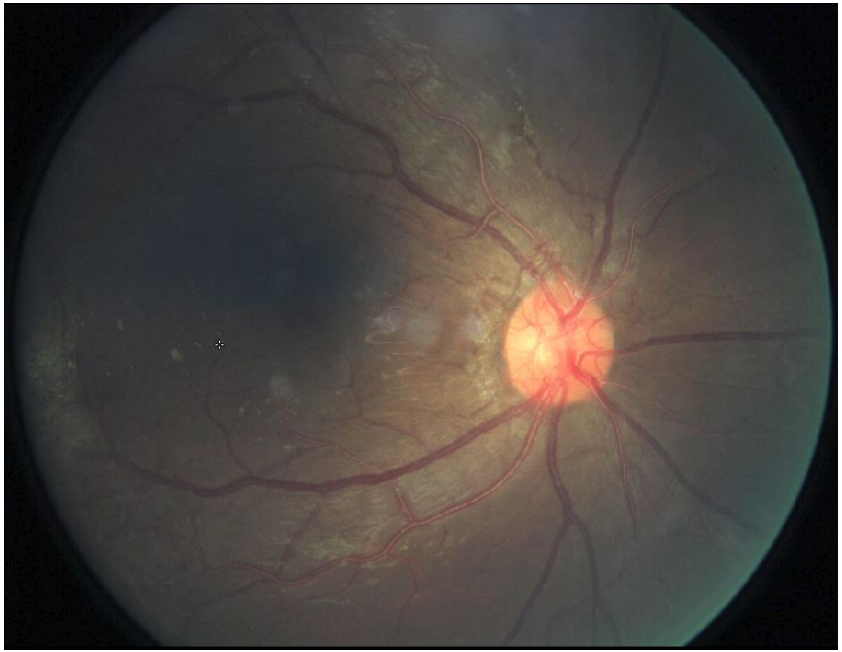
⇨ Prediction: **“proliferative”** DR[1]

- Proliferative implies vision-threatening

Can we provide an explanation to the doctor with supporting evidence for **“proliferative”** DR?

[1]**Diabetic Retinopathy (DR)** is a diabetes complication that affects the eye. Deep networks can predict DR grade from retinal fundus images with high accuracy (AUC ~0.97) [JAMA, 2016].

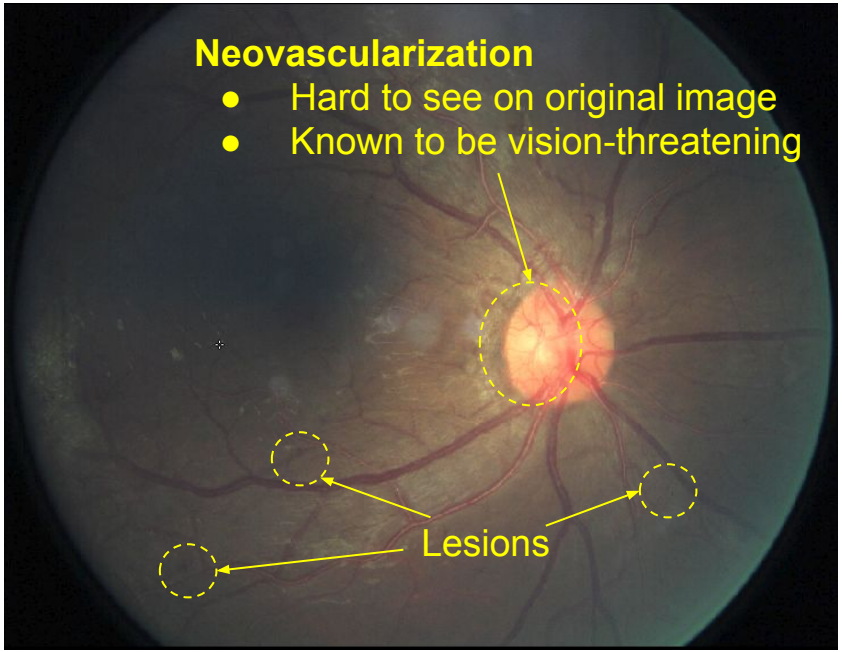
Retinal Fundus Image



Integrated Gradients for label: “proliferative”

Visualization: Overlay heatmap on green channel

Retinal Fundus Image



Integrated Gradients for label: “proliferative”

Visualization: Overlay heatmap on green channel

Efficacy of Explanations

Explanations help when:

- Model is right, and explanation convinces the doctor
- Model is wrong, and explanation reveals the flaw in the model's reasoning

But, Explanations can also hurt when:

- Model is right, but explanation is unintelligible
- Model is wrong, but the explanation convinces the doctor

Be careful about long-term effects too!

Humans and Automation: Use, Misuse, Disuse, Abuse - Parsuraman and Riley, 1997

Assisted-read study

9 doctors grade 2000 images under three different conditions

A. Image only
B. Image + Model's prediction scores
C. Image + Model's prediction scores + Explanation (Integrated Gradients)

Some findings:

- Seeing prediction scores (B) significantly increases accuracy vs. image only (A)
- Showing explanations (C) only provides slight additional improvement
  - Masks help more when model certainty is low
- Both B and C increase doctor ↔ model agreement

Paper: Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy --- Journal of Ophthalmology [2018]

Analyzing Model Robustness

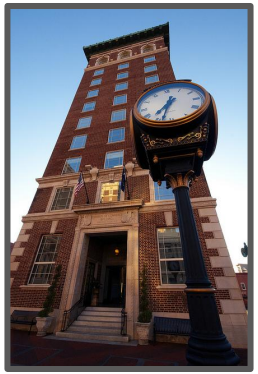
Tabular QA

Rank	Nation	Gold	Silver	Bronze	Total
1	India	102	58	37	197
2	Nepal	32	10	24	65
3	Sri Lanka	16	42	62	120
4	Pakistan	10	36	30	76
5	Bangladesh	2	10	35	47
6	Bhutan	1	6	7	14
7	Maldives	0	0	4	4

Q: How many medals did India win?
A: 197

Neural Programmer (2017) model
33.5% accuracy on WikiTableQuestions

Visual QA



Q: How symmetrical are the white bricks on either side of the building?
A: very

Kazemi and Elqursh (2017) model.
61.1% on VQA 1.0 dataset
(state of the art = 66.7%)

Reading Comprehension

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager

Q: Name of the quarterback who was 38 in Super Bowl XXXIII?
A: John Elway

Yu et al (2018) model.
84.6 F-1 score on SQuAD (state of the art)

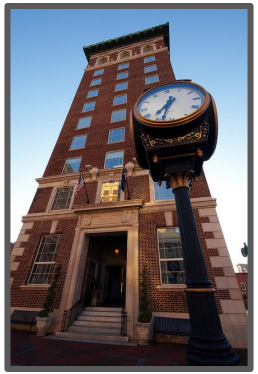
Tabular QA

Rank	Nation	Gold	Silver	Bronze	Total
1	India	102	58	37	197
2	Nepal	32	10	24	65
3	Sri Lanka	16	42	62	120
4	Pakistan	10	36	30	76
5	Bangladesh	2	10	35	47
6	Bhutan	1	6	7	14
7	Maldives	0	0	4	4

Q: How many medals did India win?
A: 197

Neural Programmer (2017) model
33.5% accuracy on WikiTableQuestions

Visual QA



Q: How symmetrical are the white bricks on either side of the building?
A: very

Kazemi and Elqursh (2017) model.
61.1% on VQA 1.0 dataset
(state of the art = 66.7%)

Reading Comprehension

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager

Q: Name of the quarterback who was 38 in Super Bowl XXXIII?
A: John Elway

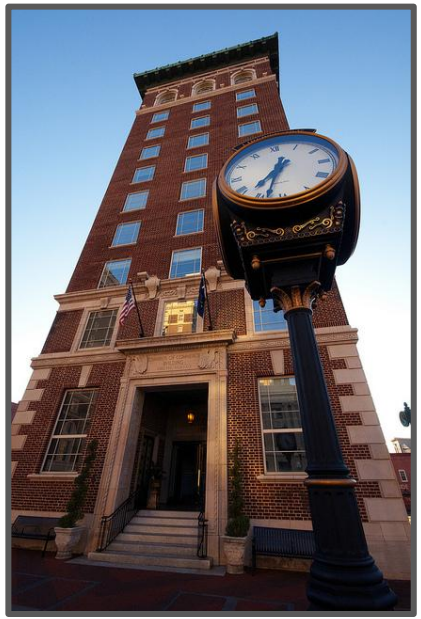
Yu et al (2018) model.
84.6 F-1 score on SQuAD (state of the art)

Robustness question: Do these network read the question carefully? :-)

Visual QA

Kazemi and Elqursh (2017) model.

Accuracy: **61.1%** (state of the art: 66.7%)

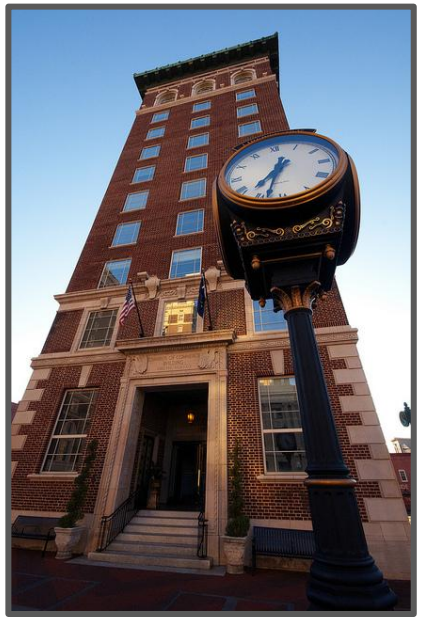


Q: How symmetrical are the white bricks on either side of the building?
A: very

Visual QA

Kazemi and Elqursh (2017) model.

Accuracy: **61.1%** (state of the art: 66.7%)



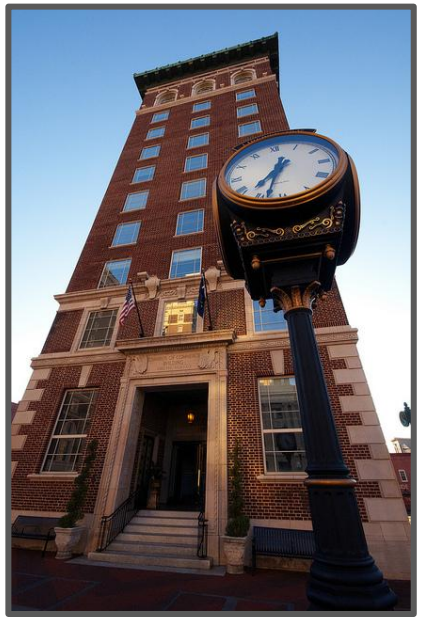
Q: How symmetrical are the white bricks on either side of the building?
A: very

Q: How asymmetrical are the white bricks on either side of the building?
A: very

Visual QA

Kazemi and Elqursh (2017) model.

Accuracy: **61.1%** (state of the art: 66.7%)



Q: How symmetrical are the white bricks on either side of the building?
A: very

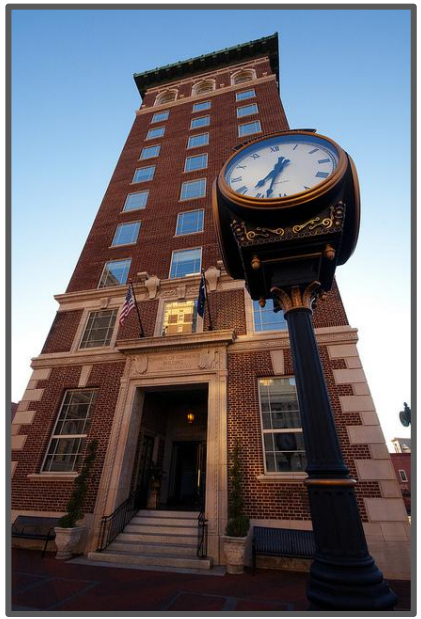
Q: How asymmetrical are the white bricks on either side of the building?
A: very

Q: How big are the white bricks on either side of the building?
A: very

Visual QA

Kazemi and Elqursh (2017) model.

Accuracy: **61.1%** (state of the art: 66.7%)



Q: How symmetrical are the white bricks on either side of the building?
A: very

Q: How asymmetrical are the white bricks on either side of the building?
A: very

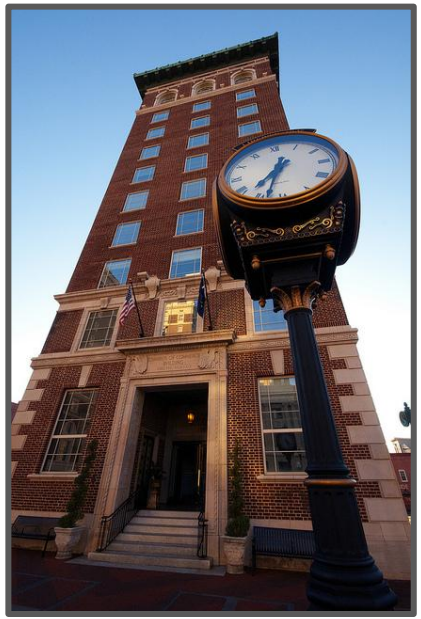
Q: How big are the white bricks on either side of the building?
A: very

Q: How fast are the bricks speaking on either side of the building?
A: very

Visual QA

Kazemi and Elqursh (2017) model.

Accuracy: **61.1%** (state of the art: 66.7%)



Q: How symmetrical are the white bricks on either side of the building?
A: very

Q: How asymmetrical are the white bricks on either side of the building?
A: very

Q: How big are the white bricks on either side of the building?
A: very

Q: How fast are the bricks speaking on either side of the building?
A: very

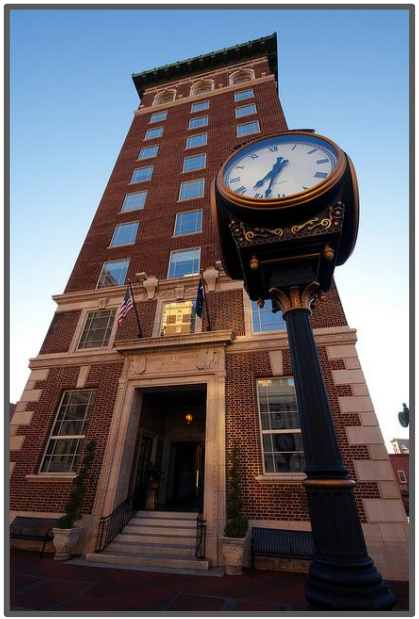
Test/dev accuracy does not show us the entire picture. Need to look inside!

Analysis procedure

- Attribute the answer (or answer selection logic) to question words
  - **Baseline**: Empty question, but full context (image, text, paragraph)
    - By design, attribution will **not** fall on the context

- Visualize attributions per example
- Aggregate attributions across examples

Visual QA attributions



Q: How symmetrical are the white bricks on either side of the building?
A: very

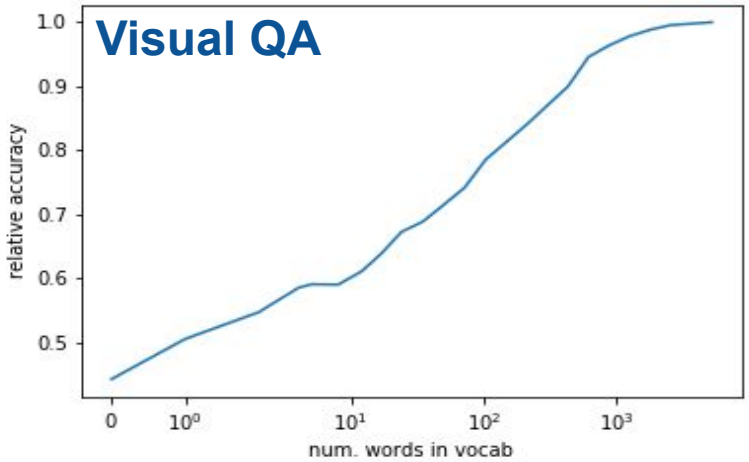
How symmetrical are the white bricks on either side of the building?

red: high attribution
blue: negative attribution
gray: near-zero attribution

Over-stability

During inference, drop all words from the dataset except ones which are frequently top attributions

- E.g. How many ~~red~~ ~~buses~~ ~~are~~ ~~in~~ ~~the~~ ~~picture?~~

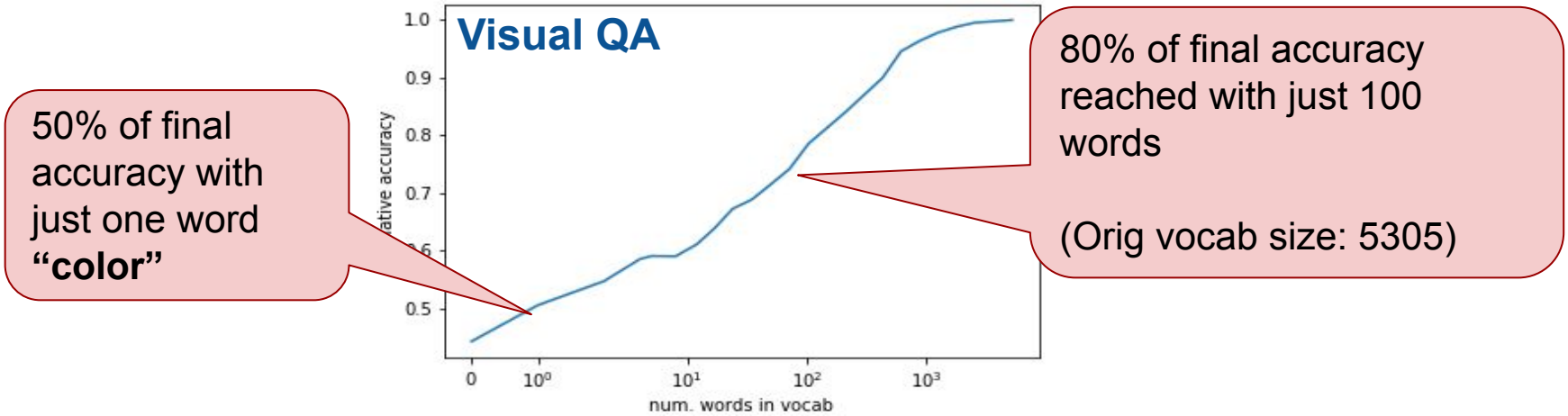


Top tokens: color, many, what, is, how, there, …

Over-stability

During inference, drop all words from the dataset except ones which are frequently top attributions

- E.g. How many ~~red buses are in the picture?~~



Top tokens: color, many, what, is, how, there, …

Attack: Subject ablation

Replace the subject of a question with a low-attribution noun from the vocabulary

- This **ought to change** the answer but often does not!

Low-attribution nouns

'tweet',
'childhood',
'copyrights',
'mornings',
'disorder',
'importance',
'topless',
'critter',
'jumper',
'fits'

What is the **man** doing? → What is the **tweet** doing?
How many **children** are there? → How many **tweet** are there?

VQA model's response remains the same 75.6% of the time on questions that it originally answered correctly

Many other attacks!

- Visual QA
  - Prefix concatenation attack (accuracy drop: **61.1% to 19%**)
  - Stop word deletion attack (accuracy drop: **61.1% to 52%**)
- Tabular QA
  - Prefix concatenation attack (accuracy drop: **33.5% to 11.4%**)
  - Stop word deletion attack (accuracy drop: **33.5% to 28.5%**)
  - Table row reordering attack (accuracy drop: **33.5 to 23%**)
- Paragraph QA
  - Improved paragraph concatenation attacks of Jia and Liang from [EMNLP 2017]

Paper: Did the model understand the question? **[ACL 2018]**

Summary

Integrated Gradients is a technique for attributing a deep network's (or any **differentiable model's**) prediction to its input features. It is very easy to apply, widely applicable and backed by an axiomatic theory.

References:

- Axiomatic Attribution for Deep Networks [ICML 2017]
- Did the model understand the question? [ACL 2018]
- Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy [Journal of Ophthalmology, 2018]
- Using Attribution to Decode Dataset Bias in Neural Network Models for Chemistry [PNAS, 2019]
- Exploring Principled Visualizations for Deep Network Attributions [EXSS Workshop, 2019]

But, what about non-differentiable models?

- Decision trees
- Boosted trees
- Random forests
- etc.

Shapley Value [Annals of Mathematical studies,1953]

Classic result in game theory on distributing gain in a **coalition game**

- **Coalition Games**
  - Players collaborating to generate some **gain** (think: revenue)
  - Set function v(S) determining the gain for any subset S of players
- **Shapley Values** are a fair way to attribute the total gain to the players based on their contributions
  - Concept: **Marginal contribution** of a player to a subset of other players (v(S U {i}) - v(S))
  - Shapley value for a player is a **specific weighted aggregation of its marginal** over all possible subsets of other players

Shapley Value Justification

Shapley values are unique under four simple axioms

- **Dummy:** If a player never contributes to the game then it must receive zero attribution
- **Efficiency:** Attributions must add to the total gain
- **Symmetry:** Symmetric players must receive equal attribution
- **Linearity:** Attribution for the (weighted) sum of two games must be the same as the (weighted) sum of the attributions for each of the games

Shapley Values for Explaining ML models

- We define a coalition game for each model input X
  - Players are the features in the input
  - Gain is the model prediction (output), i.e., gain = F(X)
- Feature attributions are the Shapley values of this game

Challenge: What does it mean for **only some players (features) to be present**?

- That is, how do we define F(x_1, <absent>, x_3, ..., <absent>) ?

Modeling Feature Absence

Key Idea: Take the expected prediction when the (absent) feature is sampled from a certain distribution.

Different approaches choose different distributions

- [SHAP, NIPS 2018] Use conditional distribution w.r.t. the present features
- [QII, S&P 2016] Use marginal distribution
- [Strumbelj et al., JMLR 2009] Use uniform distribution
- [Integrated Gradients, ICML 2017] Use a specific baseline point

Coming Soon: Paper investigating the best way(s) to handle feature absence

Computing Shapley Values

Exact Shapley value computation is exponential in the number of features

- Several approximations have been studied in the past
- Fiddler offers an efficient, distributed computation of approx. Shapley values

Our strategy

- For non-differentiable models with structured inputs, use Shapley values
- For differentiable models, use Integrated Gradients

Some limitations and caveats

Attributions are pretty shallow

Attributions do not explain:

- How the network combines the features to produce the answer?
- What training data influenced the prediction
- Why gradient descent converged
- etc.

An instance where attributions are useless:

- A network that predicts TRUE when there are **even number** of black pixels and FALSE otherwise

Attributions are useful when the network behavior entails that a strict subset of input features are important

Attributions are for human consumption

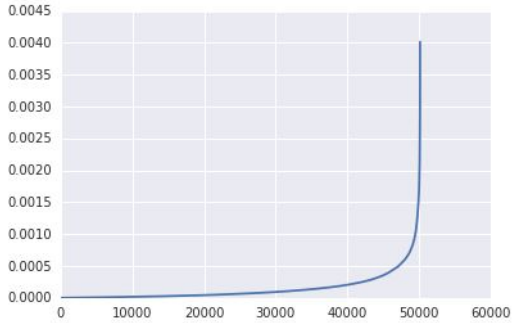
- **Humans** interpret attributions and generate insights
  - Doctor maps attributions for diabetic retinopathy to pathologies like microaneurysms, hemorrhages, etc.
- **Visualization** matters as much as the attribution technique

Attributions are for human consumption

- **Humans** interpret attributions and generate insights
  - Doctor maps attributions for diabetic retinopathy to pathologies like microaneurysms, hemorrhages, etc.
- **Visualization** matters as much as the attribution technique

Naive scaling of attributions from 0 to 255

Attributions have a **large range** and **long tail** across pixels



After clipping attributions at 99% to reduce range

Principles of Visualization for Image Attributions

Visualizations must satisfy:

- **Commensurateness**: Feature brightness must be proportional to attributions
- **Coverage**: Large fraction of the important features must be visible
- **Correspondence**: Allow correspondence between attributions and raw image
- **Coherence**: Visualizations must be optimized for human perception

Paper: Exploring Principled Visualizations for Deep Network Attributions [ExSS 2019]

Thank you!

Summary: **Integrated Gradients** is a technique for attributing a deep network's prediction to its input features. It is **very easy to apply**, **widely applicable** and backed by an **axiomatic theory.**

My email: ataly@google.com

References:

- Axiomatic Attribution for Deep Networks [ICML 2017]
- Did the model understand the question? [ACL 2018]
- Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy [Journal of Ophthalmology, 2018]
- Exploring Principled Visualizations for Deep Network Attributions [EXSS Workshop, 2019]
- Using Attribution to Decode Dataset Bias in Neural Network Models for Chemistry [preprint, 2019]