Example Based Methods for Interpretability

Apr 24, 2020

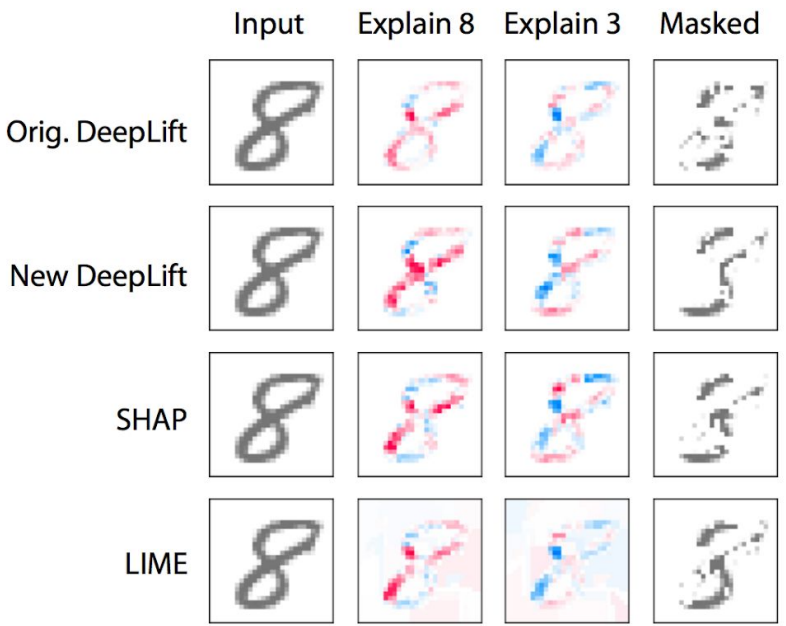
Dr. Wei Wei, Prof. James Landay

CS 335: Fair, Accountable, and Transparent (FAccT) Deep Learning

Stanford University

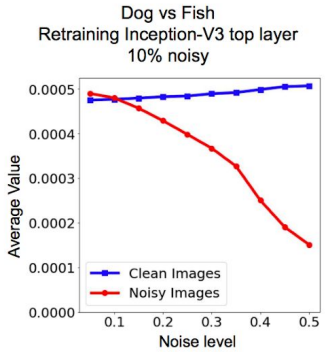
Recap

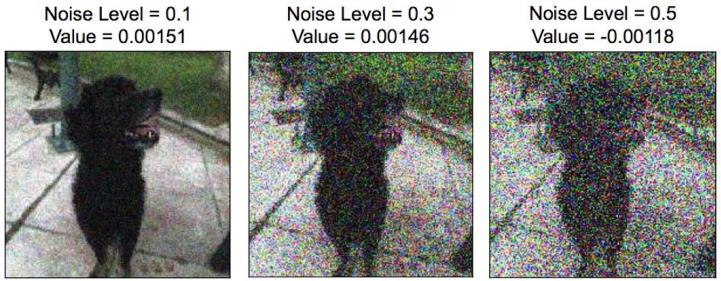
- Feature Interaction
  - SHAP
    - Inspired by game theory
    - Feature -> player
  - Kernel SHAP
    - Generalizes LIME
  - Deep SHAP
    - Generalizes LRP/DeepLift
  - Tree SHAP
    - Calculates SHAP efficiently
    - Embedded into tree models



Recap

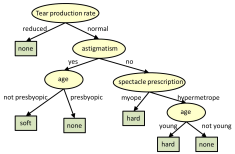
- Data Value Estimation
  - Assigns each data a Shapley Value
  - Assess the quality of data points



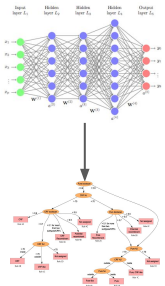


Recap

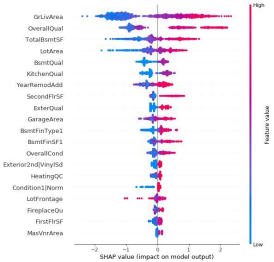
Lecture 3 Intrinsic Methods for Interpretability



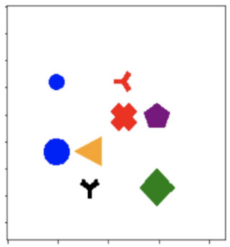
Post Hoc Methods for Interpretability



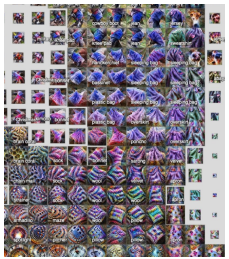
Lecture 4
Proxy Methods



Lecture 5
Feature Interaction



Lecture 6
Example Based
Methods



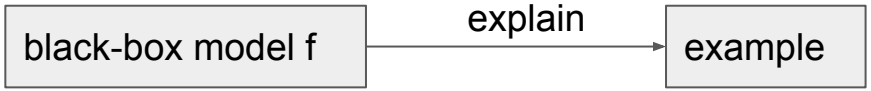
Lecture 7
Visualization Based
Methods

Outlines

- Example Based Methods
- Counterfactual Explanations
  - Targeted counterfactual examples
  - Untargeted counterfactual examples
- Contrastive Examples
- Concept Based Methods

Example Based Methods for Interpretability

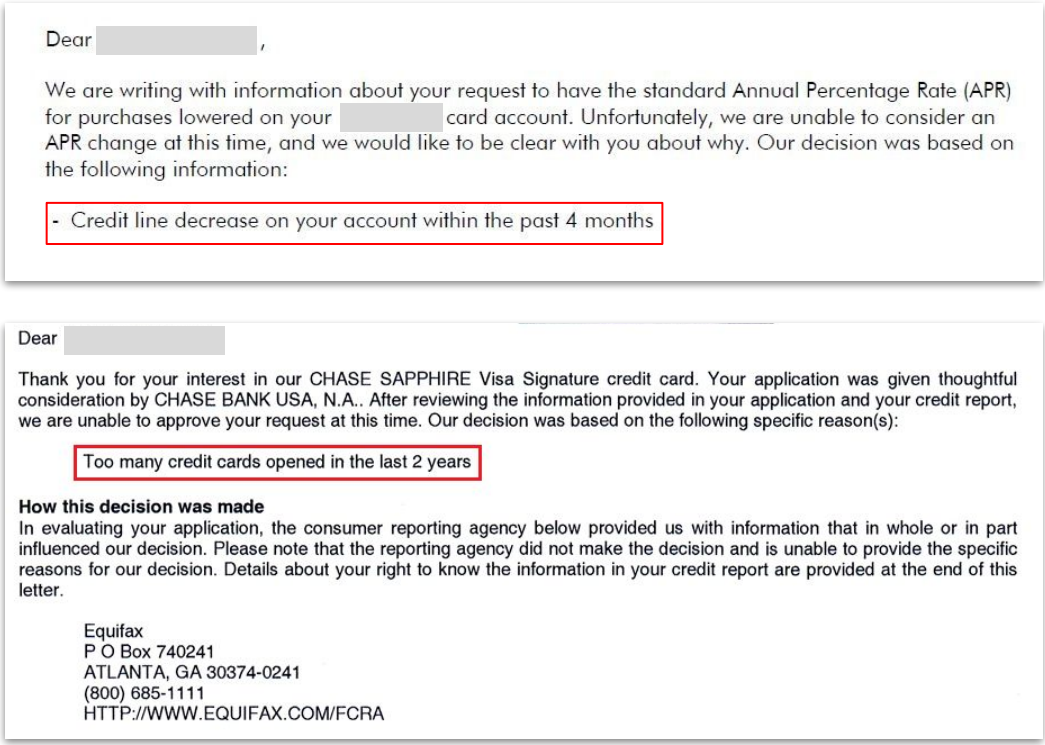
- Explain Model Behavior Using Examples
  - Counterfactual Examples
    - Explain models by asking "what-if" questions
  - Contrastive Examples
    - Explain models by generating examples that are "sufficiently present" and "necessarily absent"
  - Concept Based Methods
    - Explain models by discovering latent concepts no labeled in the data



Outlines

- Example Based Methods
- Counterfactual Explanations
  - Targeted counterfactual examples
  - Untargeted counterfactual examples
- Contrastive Examples
- Concept Based Methods

Counterfactual Explanations



Dear ,

We are writing with information about your request to have the standard Annual Percentage Rate (APR) for purchases lowered on your card account. Unfortunately, we are unable to consider an APR change at this time, and we would like to be clear with you about why. Our decision was based on the following information:

- Credit line decrease on your account within the past 4 months

Dear

Thank you for your interest in our CHASE SAPPHIRE Visa Signature credit card. Your application was given thoughtful consideration by CHASE BANK USA, N.A.. After reviewing the information provided in your application and your credit report, we are unable to approve your request at this time. Our decision was based on the following specific reason(s):

Too many credit cards opened in the last 2 years

How this decision was made

In evaluating your application, the consumer reporting agency below provided us with information that in whole or in part influenced our decision. Please note that the reporting agency did not make the decision and is unable to provide the specific reasons for our decision. Details about your right to know the information in your credit report are provided at the end of this letter.

Equifax
P O Box 740241
ATLANTA, GA 30374-0241
(800) 685-1111
HTTP://WWW.EQUIFAX.COM/FCRA

Counterfactual Explanations

- You were denied a loan because your annual income was £30,000.
- Counterfactual Example:
  - If your income had been £45,000, you would have been offered a loan.

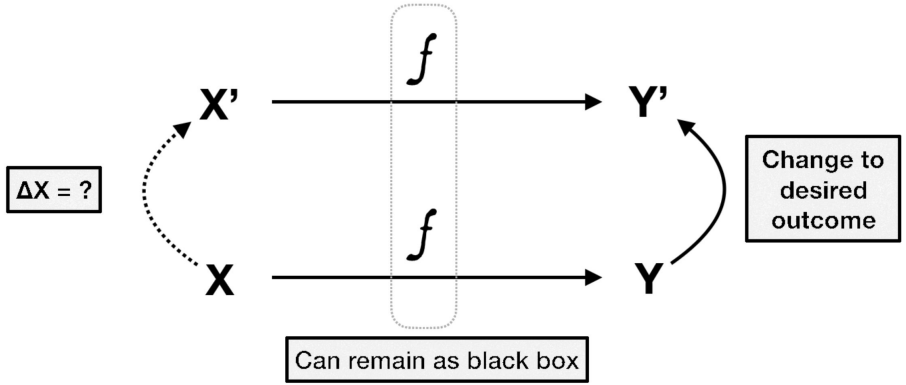
Grath et al, 2018

Formal Definitions

$$x' = arg\min_{x'} \lambda(\hat{f}(x') - y')^2 + d(x, x')$$

counterfactual example — desired outcome — distance function

Increase λ while $|\hat{f}(x') - y'| > \varepsilon$



Grath et al. 2018

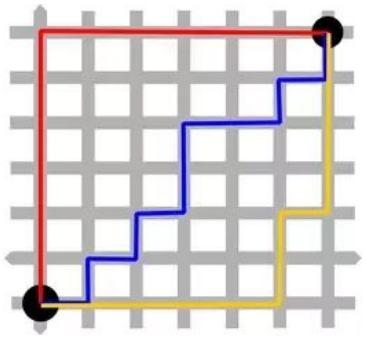
Minimum-Distance Counterfactual Examples

$$\underbrace{x'}_{\text{counterfactual example}} = arg\min_{x'} \lambda(\hat{f}(x') - \underbrace{y'}_{\text{desired outcome}})^2 + \underbrace{d(x, x')}_{\text{distance function}}$$

distance function

$$d(x, x') = \sum_{j=1}^{p} \frac{|x_j - x'_j|}{MAD_j}$$

$$MAD_j = median_{i \in \{1,\ldots,n\}}\left(\left|x_{i,j} - median_{l \in \{1,\ldots,n\}}(x_{l,j})\right|\right)$$

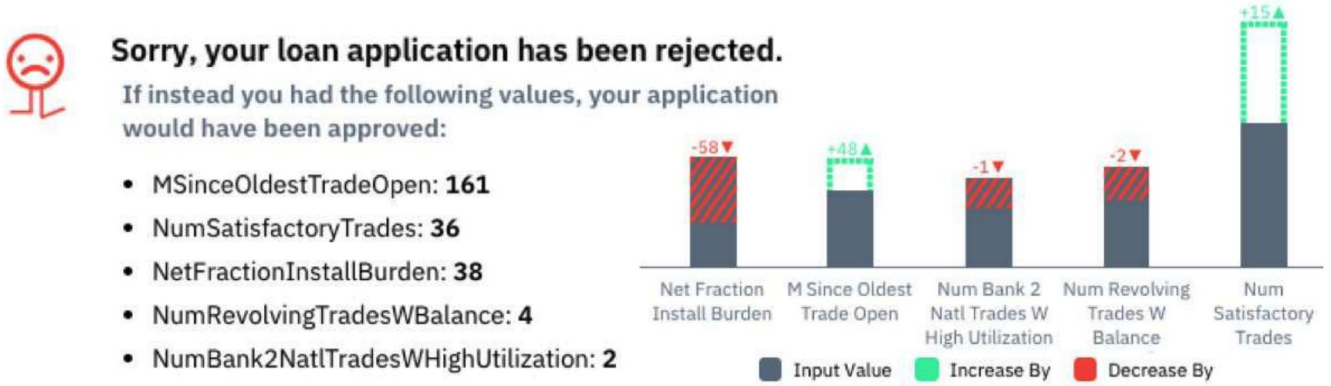


Manhattan distance

Manhattan distance weighted by inverse MAD is robust to outliers

Grath et al, 2018

Counterfactual Explanations for Loan Rejection



Grath et al, 2018

Counterfactual Explanations for Loan Acceptance

Congratulations, your loan application has been approved.

If instead you had the following values, your application would have been rejected:

- NetFractionRevolvingBurden: **55**
- NetFractionInstallBurden: **93**
- PercentTradesWBalance: **68**

Grath et al, 2018

Targeted and Untargeted Counterfactual Examples

Targeted Counterfactual Example

Generate a counterfactual example (from a sample in c) that classifiers will predict the targets class c'

Untargeted Counterfactual Example

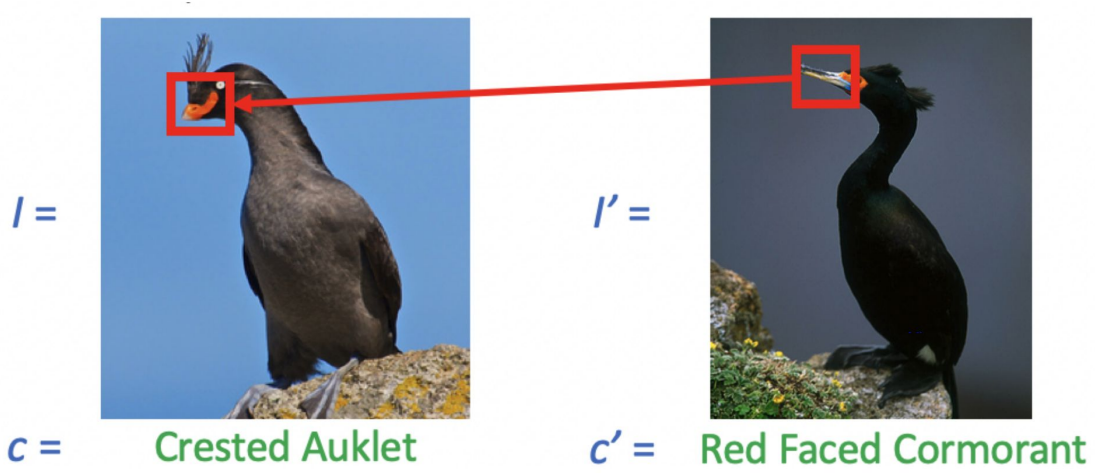
Generate a counterfactual example (from a sample in c) that classifiers have high changes of predicting the rest of the classes

Outlines

- Example Based Methods
- Counterfactual Explanations
  - Targeted counterfactual examples
  - Untargeted counterfactual examples
- Contrastive Examples
- Concept Based Methods

Counterfactual Vision Explanations

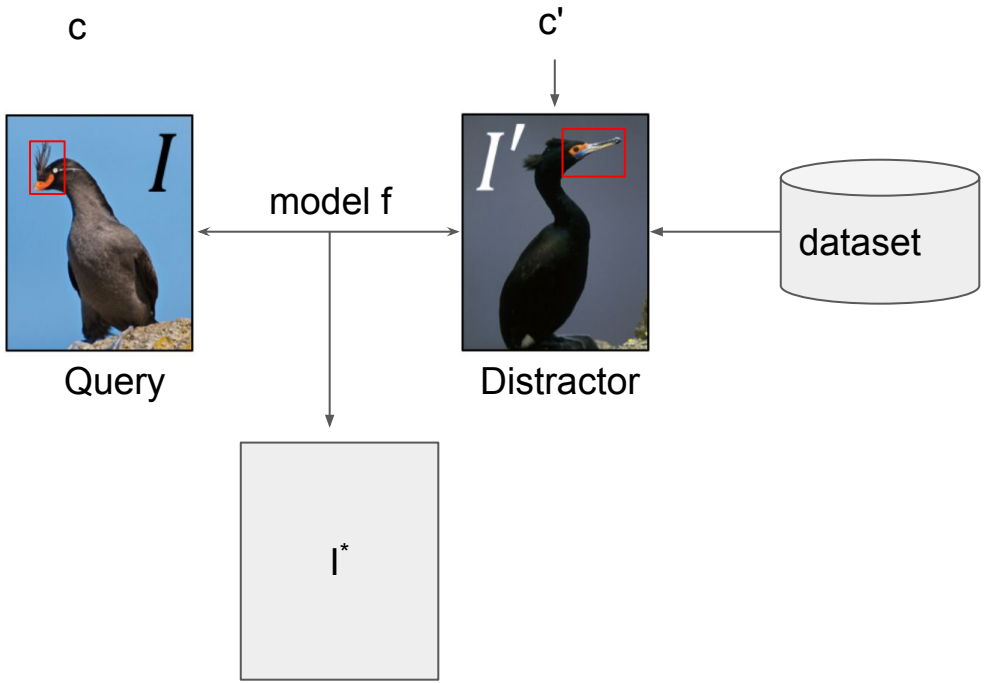
- Generate Targeted Counterfactual Image for Class c'



Goyal et al, 2019

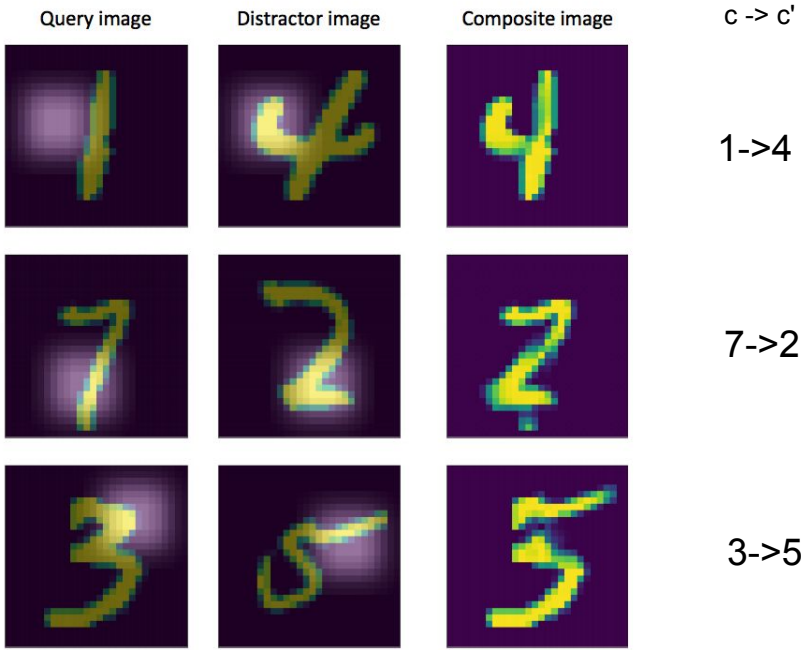
Counterfactual Vision Explanations

- Generate Targeted Counterfactual Image for Class c'



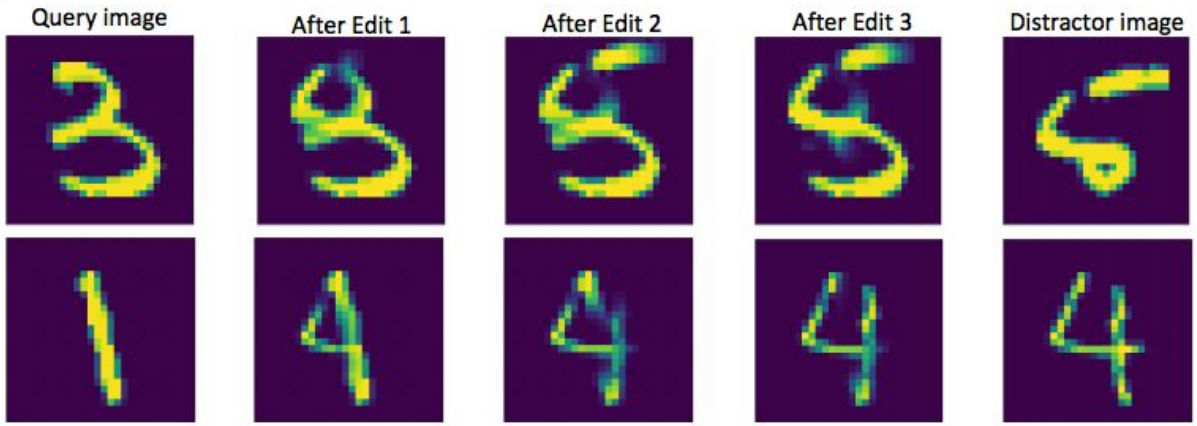
Goyal et al, 2019

MNIST Single Edits



Goyal et al, 2019

MNIST Multiple Edits



Goyal et al, 2019

Observations

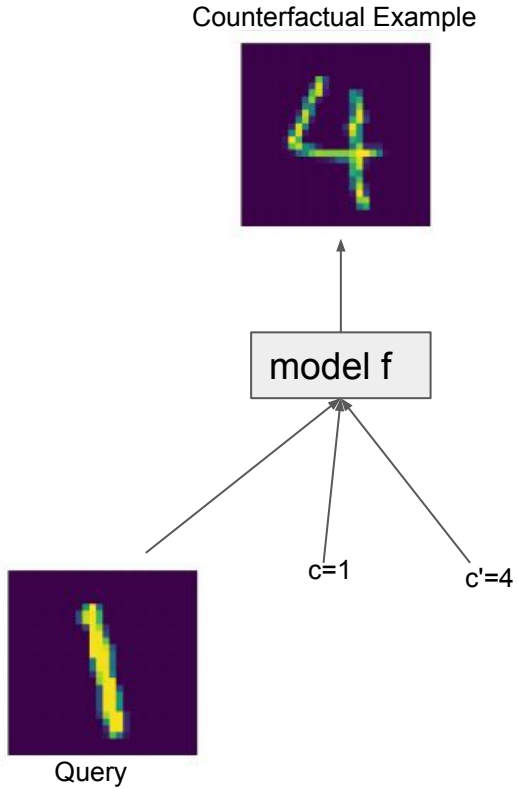
- Generate Examples for the Target Class c'
  - Based on an image of class c, I
  - With minimal changes

- Generating Counterfactual Example I*
  - Depends on c, I, c', and black-box model f
  - Explains how models make decisions
  - Different from investigating the relations between samples from c->c'

Observations

- Generate Examples for the Target Class c'
  - Based on an image of class c, I
  - With minimal changes

- Generating Counterfactual Example I*
  - Depends on c, I, c', and black-box model f
  - Explains how models make decisions
  - Different from investigating the relations between samples from c->c'

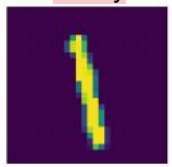


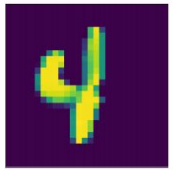
Learning Counterfactual Generators

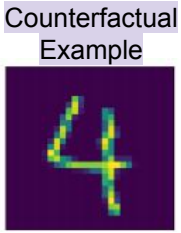
$$\underset{P,a}{\text{maximize}} \quad g_{c'}\underbrace{\left((\mathbb{1} - \mathbf{a}) \circ f(I) + \mathbf{a} \circ P f(I')\right)}_{\text{generated counterfactual example}}$$

Query

Distractor







Learning Counterfactual Generators

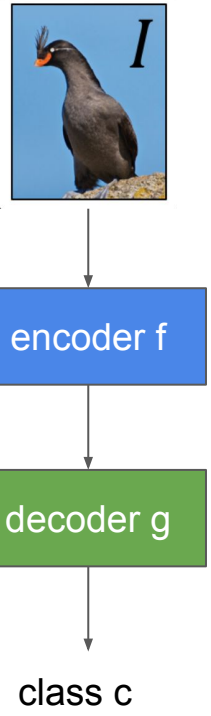
$$\underset{P,a}{\text{maximize}} \quad g_{c'}\big((\mathbb{1} - \mathbf{a}) \circ f(I) + \mathbf{a} \circ P f(I')\big)$$

Dec — $g_{c'}$

Enc Query — $f(I)$

Distractor — $P f(I')$

generated counterfactual example



Learning Counterfactual Generators

$$\underset{P,a}{\text{maximize}} \quad \underbrace{g_{c'}\left((\mathbb{1} - \mathbf{a}) \circ f(I) + \mathbf{a} \circ P f(I')\right)}_{\text{generated counterfactual example}}$$

Dec

Enc Query

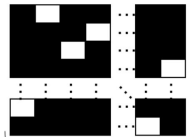
Distractor

$$\text{s.t.} \quad ||a||_1 = 1, \;\; a_i \geq 0 \;\; \forall i$$

$$||p_i||_1 = 1 \quad \forall i, \;\; P_{i,j} \geq 0 \;\; \forall i, j$$

Gating Vector

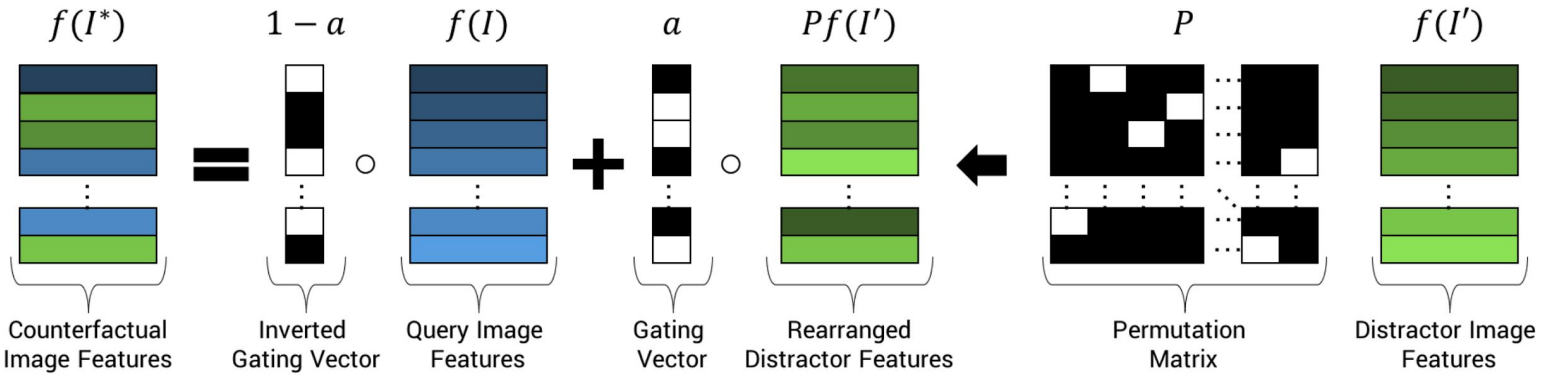
Permutation Matrix



Learning Counterfactual Generators

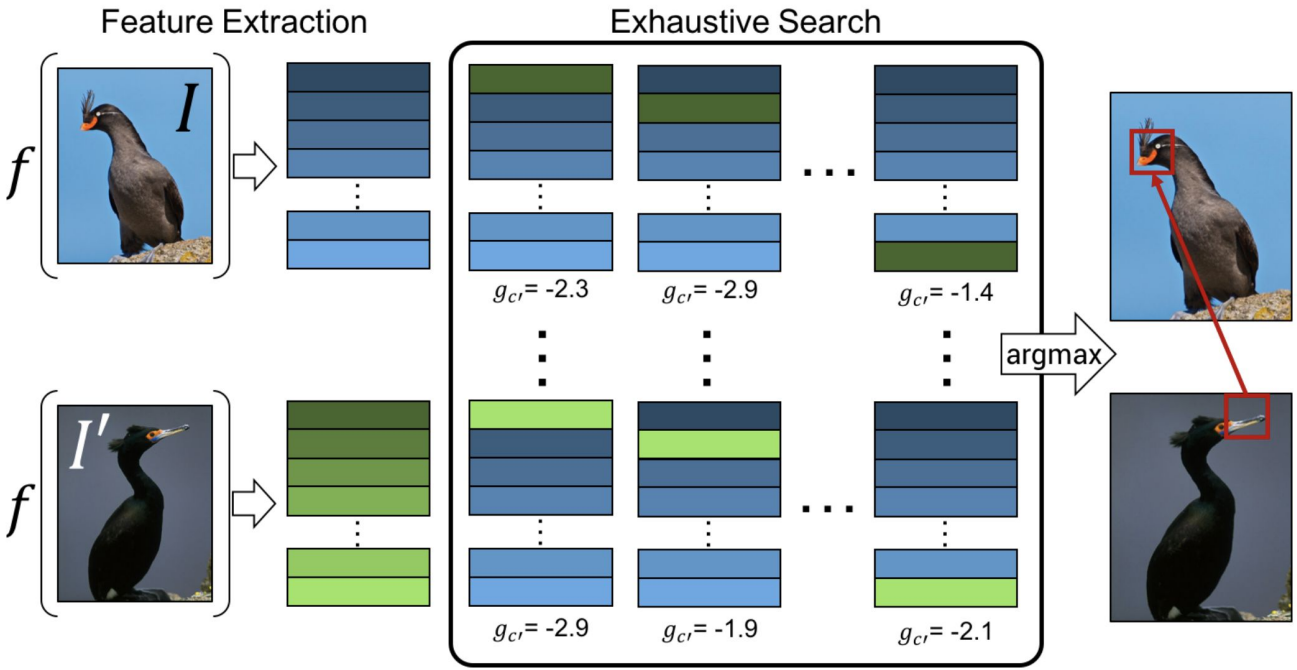
$$\underset{P,a}{\text{maximize}} \quad \underbrace{g_{c'}\big((\mathbb{1} - \mathbf{a}) \circ f(I) + \mathbf{a} \circ P f(I')\big)}_{\text{generated counterfactual example}}$$

Dec · Enc · Query · Distractor



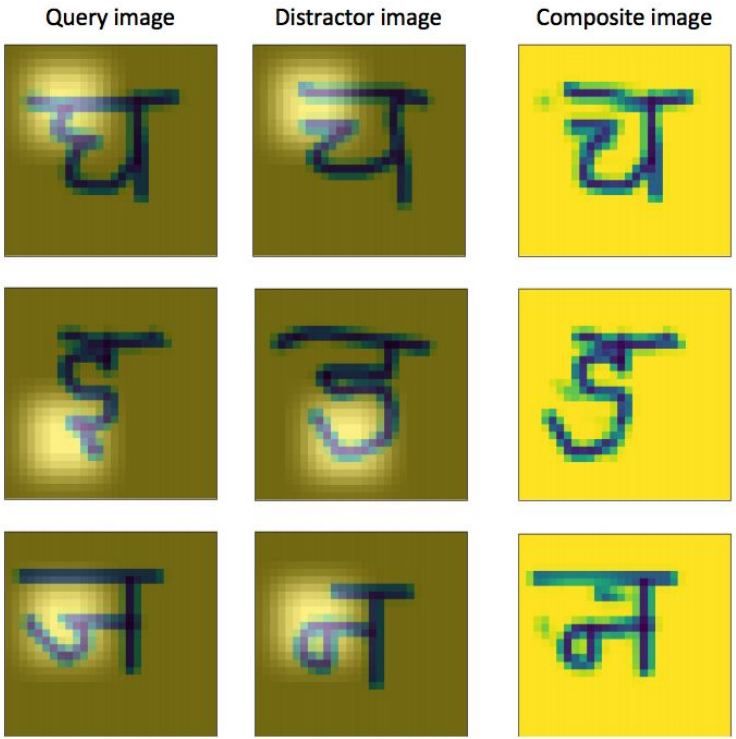
Goyal et al, 2019

Searching for A Distractor



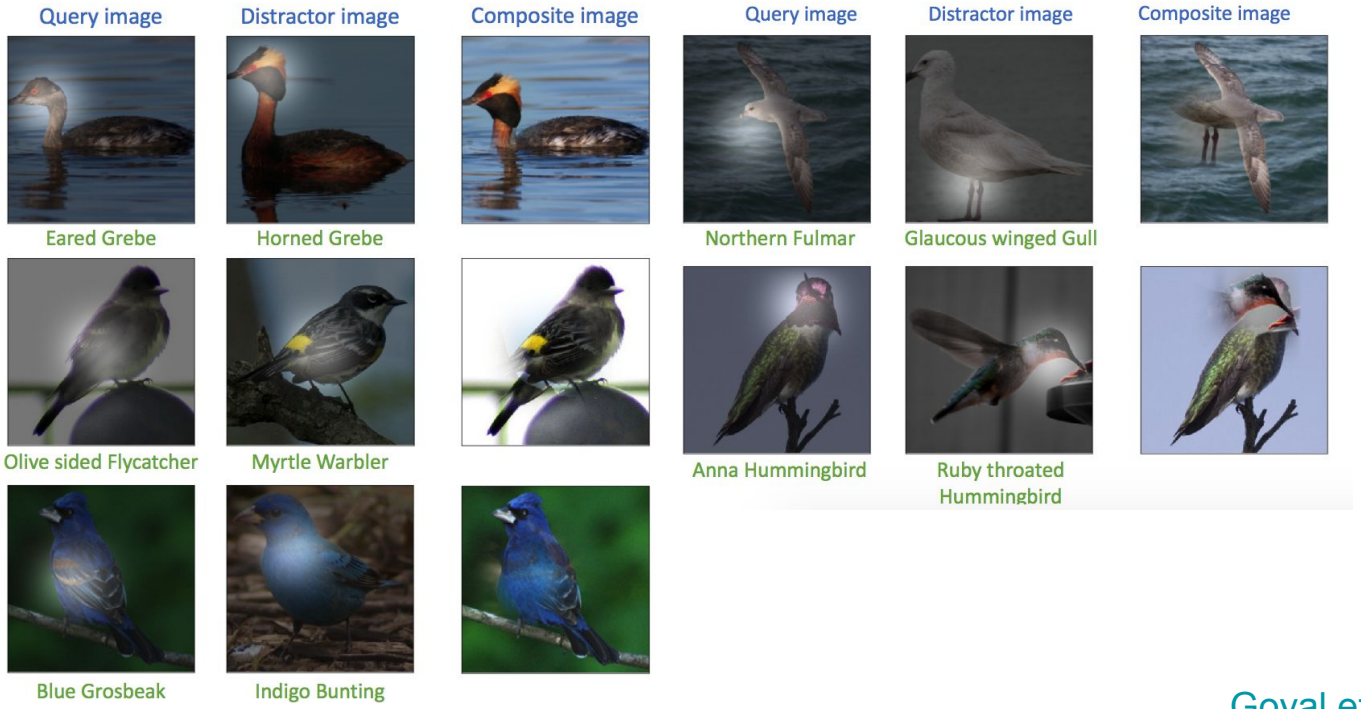
Goyal et al, 2019

Results on Omniglot



Goyal et al, 2019

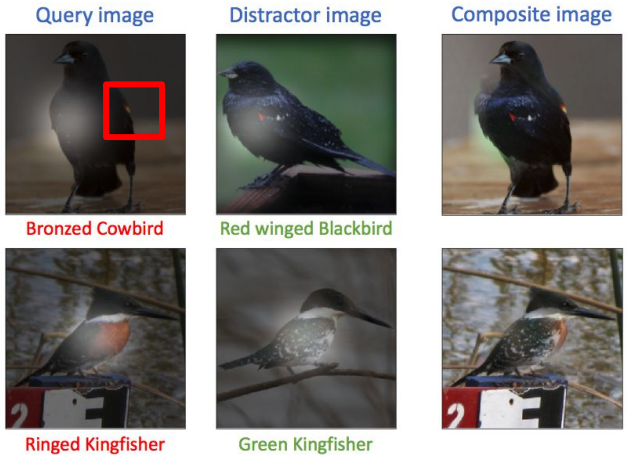
Results on Birds Dataset



Goyal et al, 2019

Results on Birds Dataset

- Results on Incorrect Predictions



Goyal et al, 2019

Outlines

- Example Based Methods
- Counterfactual Explanations
  - Targeted counterfactual examples
  - Untargeted counterfactual examples
- Contrastive Examples
- Concept Based Methods

Recall Targeted/Untargeted Counterfactual Examples

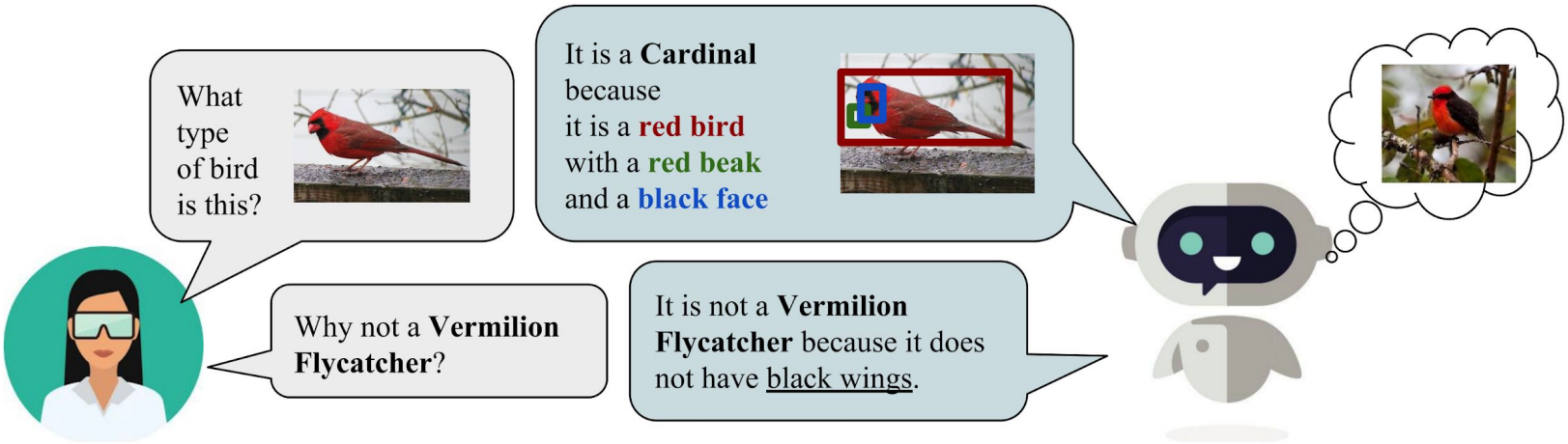
Targeted Counterfactual Example

Generate a counterfactual example (from a sample in c) that classifiers will predict the targets class c'

Untargeted Counterfactual Example

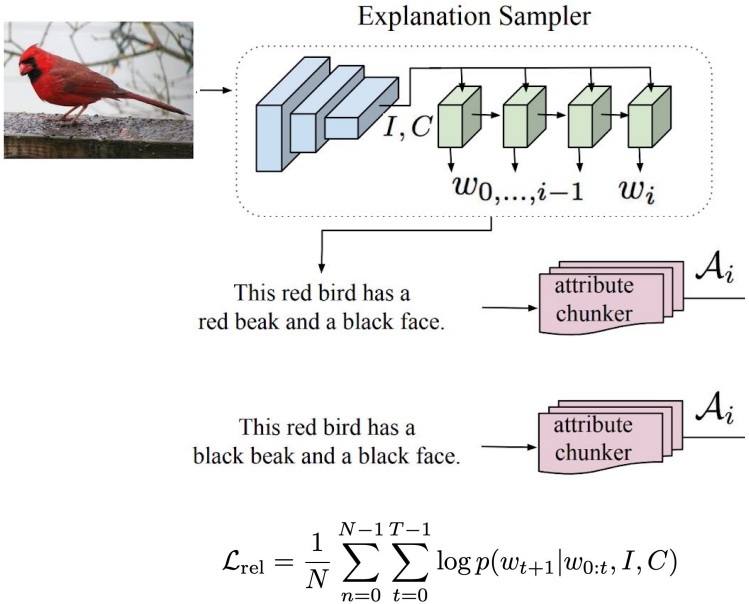
Generate a counterfactual example (from a sample in c) that classifiers have high changes of predicting the rest of the classes

Grounding Visual Explanations



Hendricks et al , 2018

Model Architecture

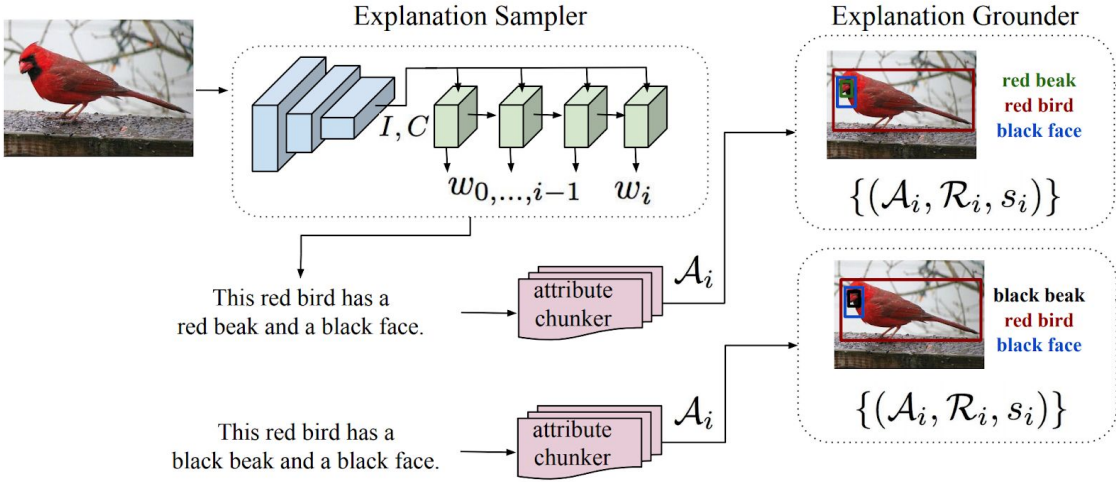


$$\mathcal{L}_{\text{rel}} = \frac{1}{N} \sum_{n=0}^{N-1} \sum_{t=0}^{T-1} \log p(w_{t+1} | w_{0:t}, I, C)$$

- Initially, each image has one ground truth sentence
- Generate ten negative explanation sentences
- Created negative sentences by flipping attributes corresponding to color, size and objects in attribute phrases
  - “yellow belly” -> “red head”
  - “yellow belly” -> “yellow beak”

Hendricks et al , 2018

Model Architecture



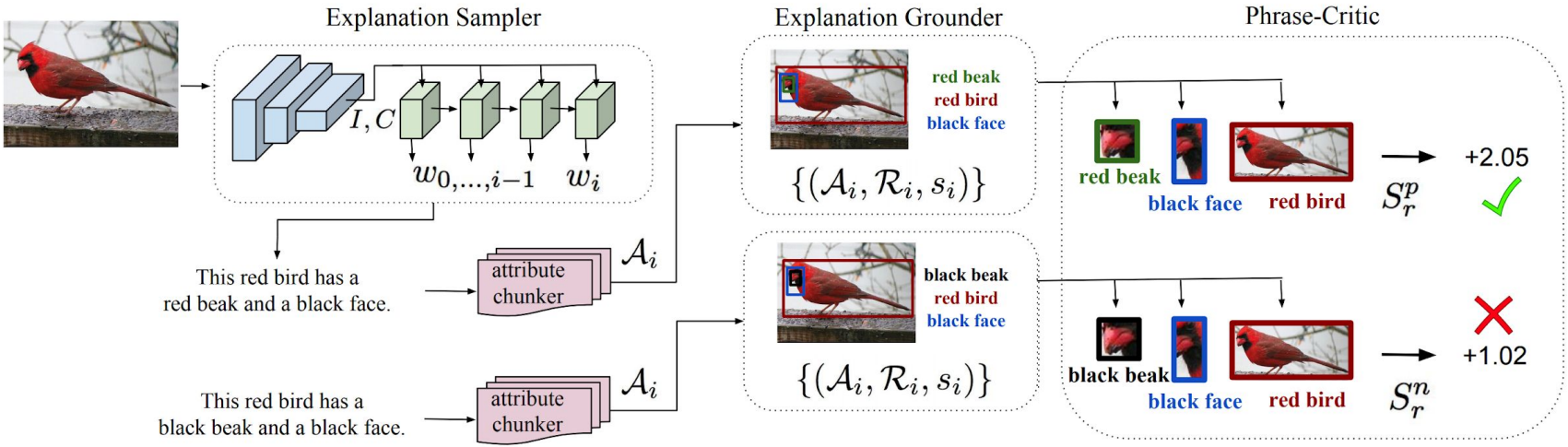
$$\mathcal{L}_{\text{rel}} = \frac{1}{N} \sum_{n=0}^{N-1} \sum_{t=0}^{T-1} \log p(w_{t+1} | w_{0:t}, I, C)$$

$$\mathcal{L}_{\text{rank}} = \max(0, \underbrace{f_{critic}(\{A_i'\}, I; \theta)}_{S_r^n} - \underbrace{f_{critic}(\{A_i^p\}, I; \theta)}_{S_r^p} + 1)$$

A_i = phrase
R_i = region
s_i = localization score

Hendricks et al , 2018

Model Architecture



$$\mathcal{L}_{\text{rel}} = \frac{1}{N} \sum_{n=0}^{N-1} \sum_{t=0}^{T-1} \log p(w_{t+1} | w_{0:t}, I, C)$$ Loss on Sentences

$$\mathcal{L}_{\text{rank}} = \max(0, \underbrace{f_{critic}(\{A_i^n\}, I; \theta)}_{S_r^n} - \underbrace{f_{critic}(\{A_i^p\}, I; \theta)}_{S_r^p} + 1)$$ Counterfactual Loss

$$\mathcal{L}_{\text{discr}} = \mathbb{E}_{\tilde{w} \sim p(w|I,C)} [R(\tilde{w})]$$

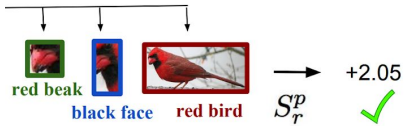
Loss on Rewards

Counterfactual Loss

$$\mathcal{L}_{\text{rank}} = \max(0, \underbrace{f_{critic}(\{A_i^n\}, I; \theta)}_{S_r^n} - \underbrace{f_{critic}(\{A_i^p\}, I; \theta)}_{S_r^p} + 1)$$

Mismatching phrase

Matching phrase



Hendricks et al , 2018

Counterfactual Examples


This bird is a **Crested Auklet** because this is a black bird with a small orange beak and it is not a **Red Faced Cormorant** because it does not have a long flat bill.


This bird is a **Parakeet Auklet** because this is a black bird with a white belly and small feet and it is not a **Horned Grebe** because it does not have red eyes.


This bird is a **Least Auklet** because this is a black and white spotted bird with a small beak and it is not a **Belted Kingfisher** because it does not have a long pointy bill.


This bird is a **White Pelican** because this is a large white bird with a long orange beak and it is not a **Laysan Albatross** because it does not have a curved bill.


This bird is a **Cardinal** because this is a red bird with a black face and it is not a **Scarlet Tanager** because it does not have a black wings.


This bird is a **Yellow Headed Blackbird** because this is a small black bird with a yellow breast and head and it is not a **Prothonotary Warbler** because it does not have a gray wing.

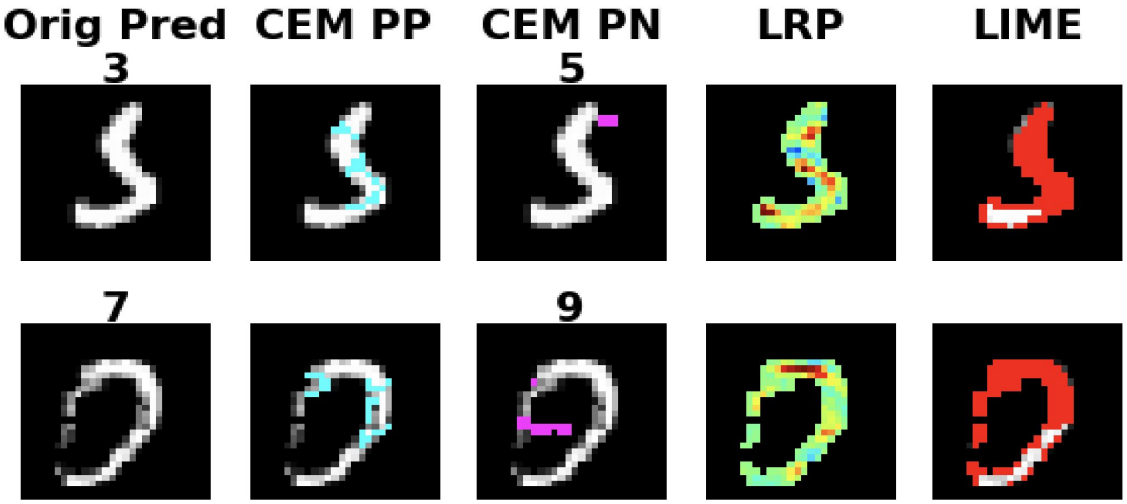
Hendricks et al , 2018

Outlines

- Example Based Methods
- Counterfactual Explanations
  - Targeted counterfactual examples
  - Untargeted counterfactual examples
- Contrastive Examples
- Concept Based Methods

Contrastive Examples

- Generate A Set of Examples to Explain Models
  - Pertinent Positive (PP) - critical features for the class
  - Pertinent Negative (PN) - what is missing in the model prediction



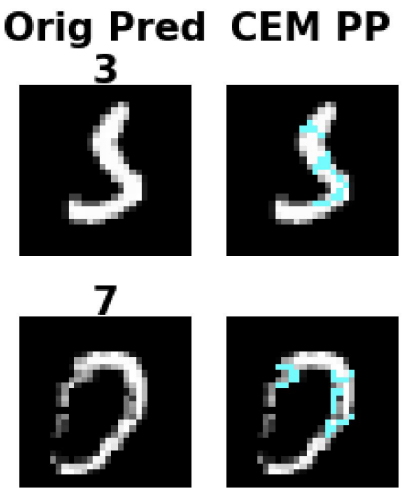
Contrastive Examples

Dhurandhar et al, 2018

Formal Definitions

$$\mathbf{x} = \boldsymbol{\delta}$$

generated example



- Pertinent Positive
  - Contains critical features

$$\min_{\boldsymbol{\delta} \in \mathcal{X} \cap \mathbf{x}_0} \; c \cdot \underbrace{\max\{\max_{i \neq t_0} [\text{Pred}(\boldsymbol{\delta})]_i - [\text{Pred}(\boldsymbol{\delta})]_{t_0}, -\kappa\}}_{\text{Max Margin Loss for Pertinent Positive}} + \underbrace{\gamma \|\boldsymbol{\delta} - \text{AE}(\boldsymbol{\delta})\|_2^2}_{\text{Reconstruction Loss}}$$

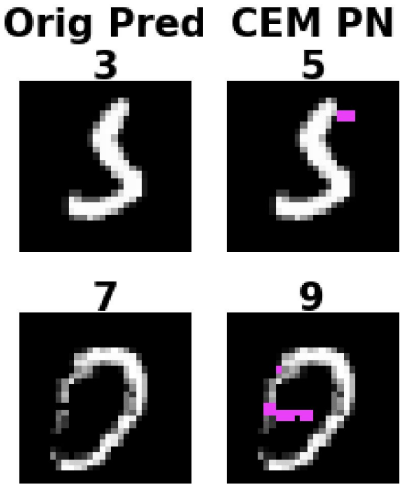
AE = AutoEncoder

Dhurandhar et al, 2018

Formal Definitions

$$\mathbf{x} = \mathbf{x}_0 + \boldsymbol{\delta}$$

generated example | original example | perturbations



- Pertinent Negative
  - Contains what is missing in the model prediction

$$\min_{\boldsymbol{\delta} \in \mathcal{X}/\mathbf{x}_0} \; c \cdot \max\{[\text{Pred}(\mathbf{x}_0 + \boldsymbol{\delta})]_{t_0} - \max_{i \neq t_0}[\text{Pred}(\mathbf{x}_0 + \boldsymbol{\delta})]_i, -\kappa\} + \gamma \|\mathbf{x}_0 + \boldsymbol{\delta} - \text{AE}(\mathbf{x}_0 + \boldsymbol{\delta})\|_2^2$$

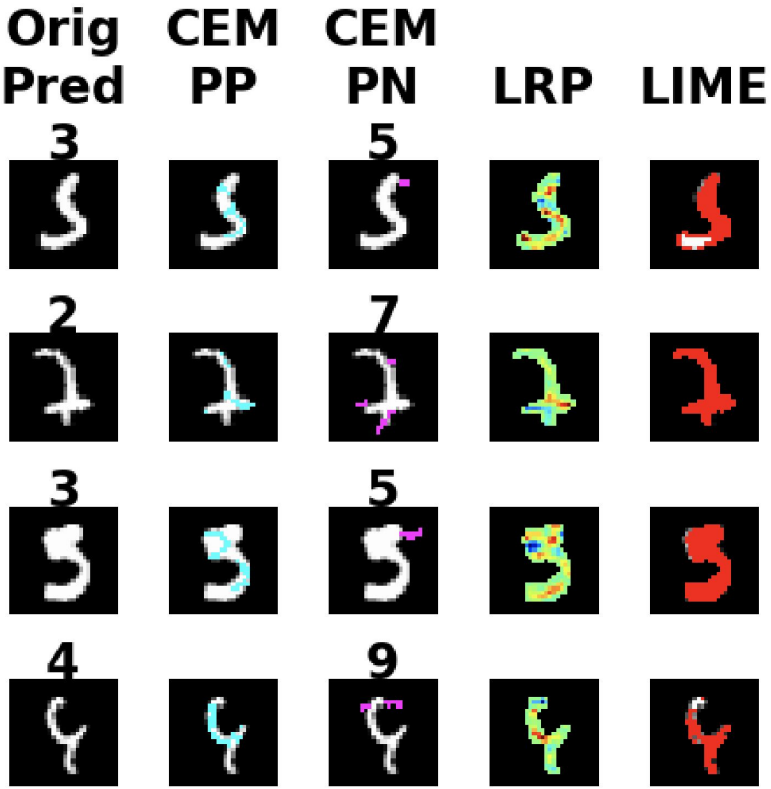
Max Margin Loss for Pertinent Negative

Reconstruction Loss

AE = AutoEncoder

Dhurandhar et al, 2018

Additional Examples



Dhurandhar et al, 2018

Outlines

- Example Based Methods
- Counterfactual Explanations
  - Targeted counterfactual examples
  - Untargeted counterfactual examples
- Contrastive Examples
- Concept Based Methods

Concept Based Methods for Interpretability

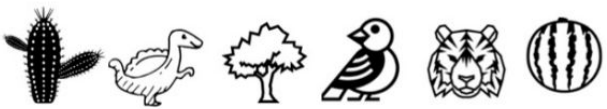
- Explain ML Models by Concepts
  - Concepts are higher level clusters that training samples belong to
  - Concepts are usually latent, not observed in the training data

- Testing With Concept Activation Vectors (TCAV)
  - Query model behavior by a concept
    - defined by a collect of user selected samples

Kim et al, 2018

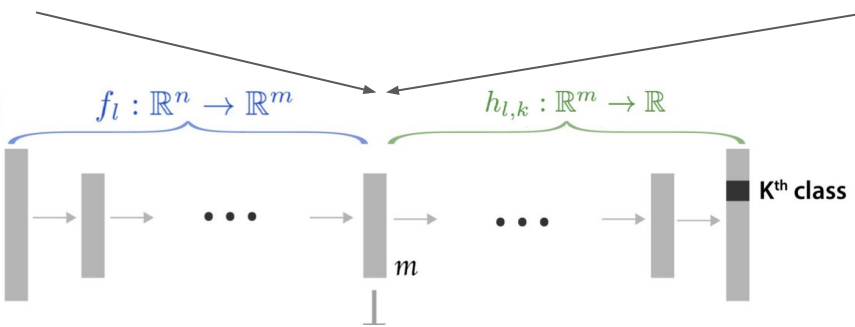
Concept Activation Vectors (CAV)

- 1) Select a collection of samples that represent a concept
- 2) Mix them with random samples
- 3) Feed into neural networks and generate activations



random samples

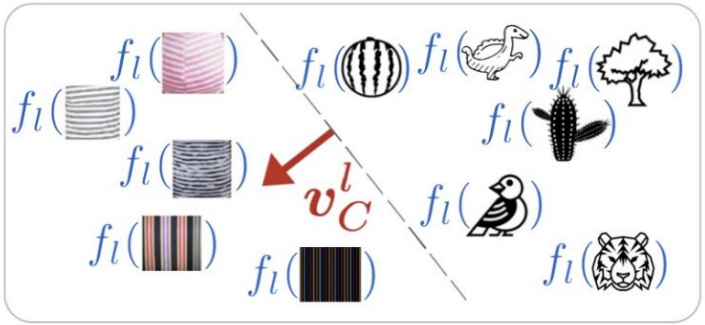
samples that represent a concept



Kim et al, 2018

Concept Activation Vectors (CAV)

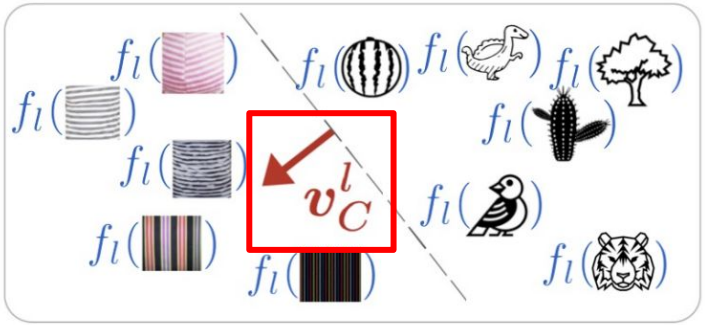
- 1) Select a collection of samples that represent a concept
- 2) Mix them with random samples
- 3) Feed into neural networks and generate activations
- 4) Generate Concept Activation Vectors (CAV)
  - Fit a linear model that separates the selected samples and the random samples
  - Concept Activation Vector is the unit vector that is orthogonal to the decision boundary



Kim et al, 2018

Concept Activation Vectors (CAV)

- 1) Select a collection of samples that represent a concept
- 2) Mix them with random samples
- 3) Feed into neural networks and generate activations
- 4) Generate Concept Activation Vectors (CAV)
  - Fit a linear model that separates the selected samples and the random samples
  - Concept Activation Vector is the unit vector that is orthogonal to the decision boundary



Kim et al, 2018

Testing With Concept Activation Vectors (TCAV)

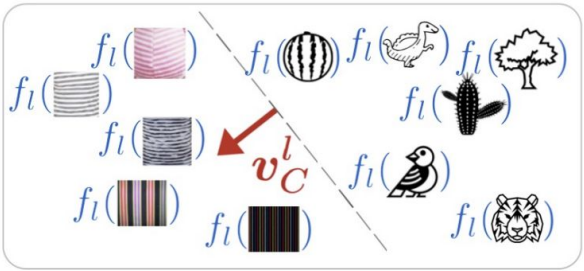
- Directional Derivative

TCAV (level l) activation vector (level l)

$$S_{C,k,l}(\boldsymbol{x}) \quad = \quad \lim_{\epsilon \to 0} \frac{h_{l,k}(f_l(\boldsymbol{x}) + \epsilon \boldsymbol{v}_C^l) - h_{l,k}(f_l(\boldsymbol{x}))}{\epsilon}$$

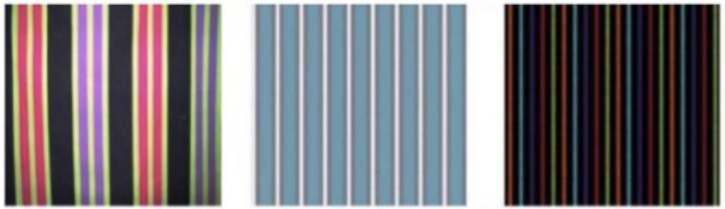
- TCAV Score

$$\text{TCAVQ}_{C,k,l} = \frac{|\{\boldsymbol{x} \in X_k : S_{C,k,l}(\boldsymbol{x}) > 0\}|}{|X_k|}$$



Querying Concepts In ImageNet

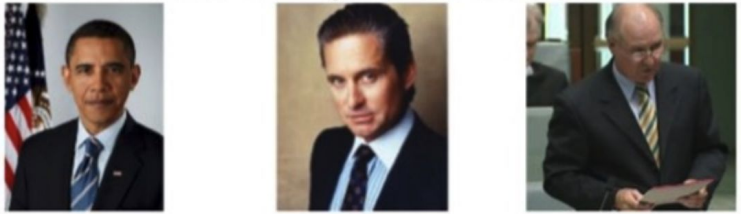
CEO concept: most similar striped images



CEO concept: least similar striped images

Model Women concept: most similar necktie images

Model Women concept: least similar necktie images



[Kim et al, 2018]

Querying Concepts In Deep Dream

Concept knitted texture

Concept corgis

Concept Siberian huskey

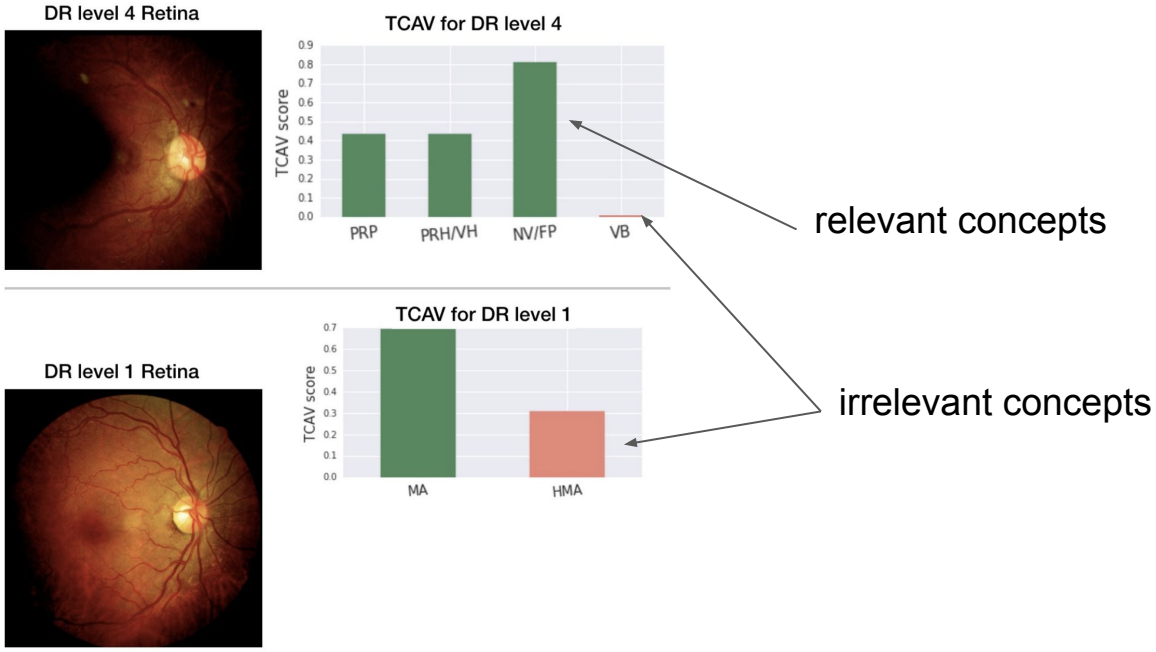
Kim et al, 2018

Case Study for Diagnosing Diabetic Retinopathy (DR)

- DR level uses a 5-point grading scale based on complex criteria
  - from level 0 (no DR) to 4 (proliferative)
- Doctors' diagnoses of DR depend on evaluating a set of diagnostic concepts
  - e.g. microaneurysms (MA), pan-retinal laser scars (PRP)
- Some concepts maybe more prominent at certain DR levels

Kim et al, 2018

Case Study for Diagnosing Diabetic Retinopathy (DR)



Kim et al, 2018

Summary

- We talked about three categories of example based methods for interpretability
- Counterfactual examples
  - Asking the models to answer "what-if" questions
  - Samples are generated to reflect models' decisions based outcomes
  - Outcomes can be targeted or untargeted
- Contrastive Examples
  - Generating a pair of examples
    - Explains the features that models use in making decisions
    - Explains the features that models do not use in making decisions
- Concept Based Methods
  - Cluster samples based on latent concepts

Required Reading

Molnar: Ch 6

Reading Assignments

- Wachter, Sandra, Brent Mittelstadt, and Chris Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harv. JL & Tech, 2017
- Yoon, Jinsung, James Jordon, and Mihaela van der Schaar. INVASE: Instance-wise variable selection using neural networks, ICLR 2018
- Yeh, Chih-Kuan, Been Kim, Sercan O. Arik, Chun-Liang Li, Pradeep Ravikumar, and Tomas Pfister. On Concept-Based Explanations in Deep Neural Networks, arXiv 2019
- Gurumoorthy, Karthik S., Amit Dhurandhar, Guillermo Cecchi, and Charu Aggarwal. Efficient Data Representation by Selecting Prototypes with Importance Weights, ICDM 2019
- Kim, Been, Cynthia Rudin, and Julie A. Shah. The bayesian case model: A generative approach for case-based reasoning and prototype classification. NeurIPS 2014

Next Lecture

Visualization Based Methods for Interpretability