

NLP Seminar - March 2021

Explainability in DL & NLP

bhaecker

Overview

- Introduction
- Datasets/models
- LIME Framework
- Faithful Interpretations
- Explanation sets
- Counterfactual examples ("what-if"-method)
- Discussion

Overview

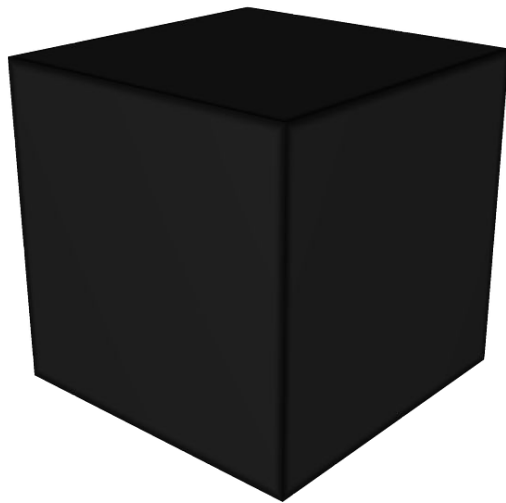
- Introduction
- Datasets/models
- LIME Framework
- Faithful Interpretations
- Explanation sets
- Counterfactual examples ("what-if"-method)
- Discussion



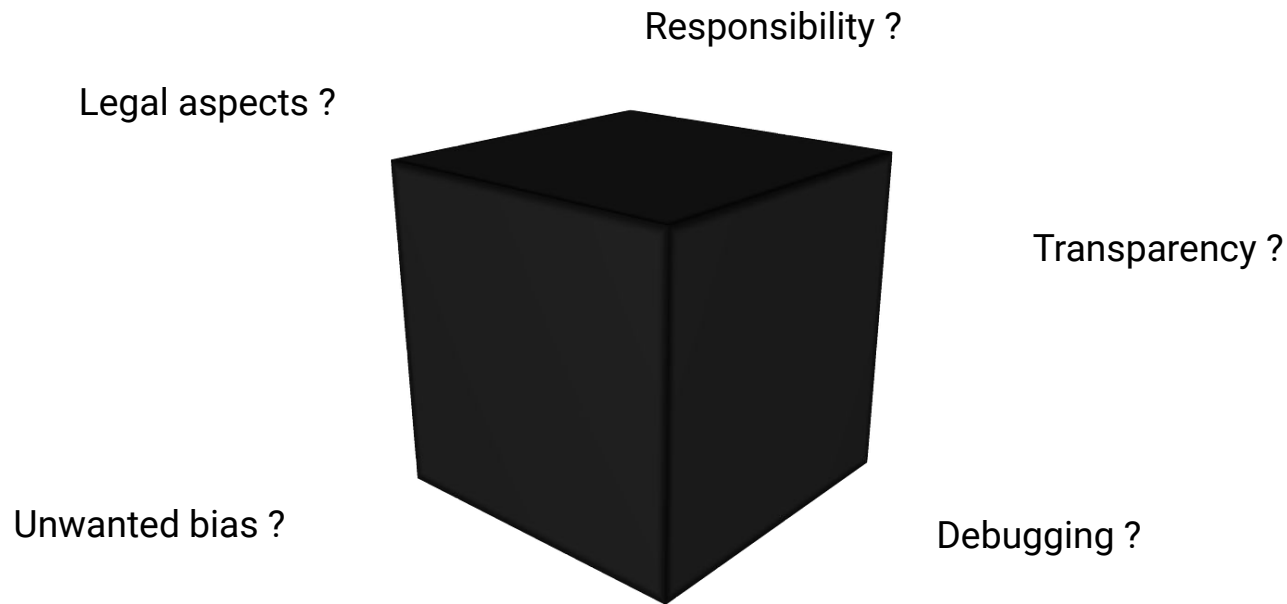
basically three methods on two data sets/tasks

Introduction

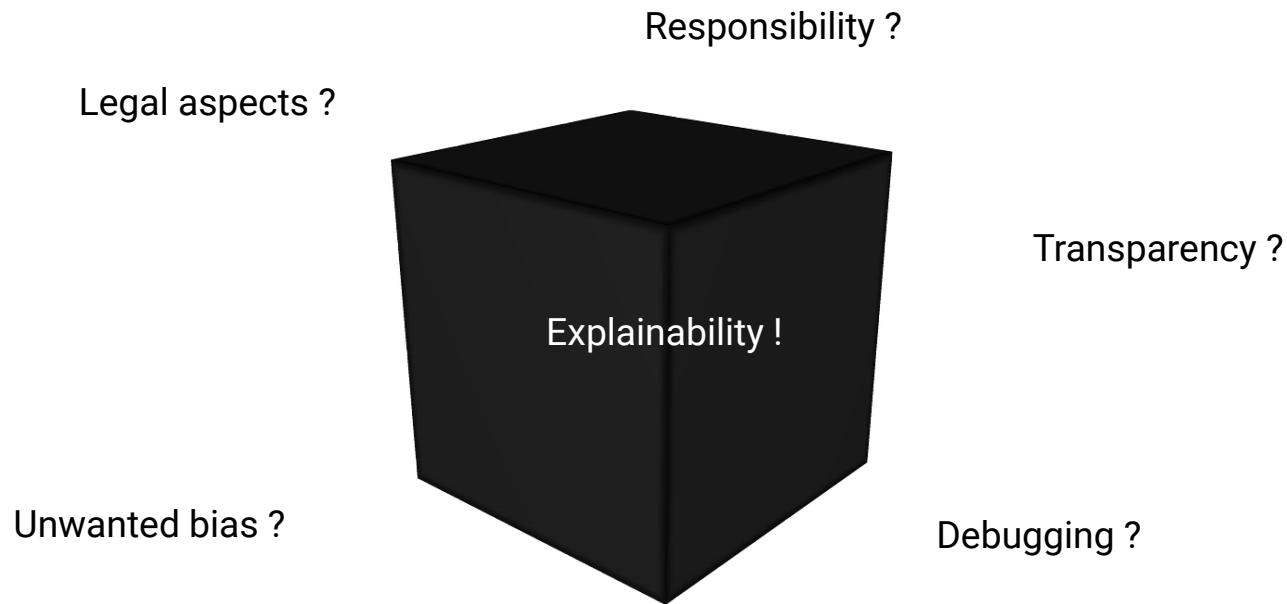
Introduction



Introduction



Introduction



Introduction

- **Interpretability is the degree to which a human can understand the cause of a decision**

(Tim Miller: “Explanation in Artificial Intelligence: Insights from the Social Sciences.”)

“You can ask a human, but, you know, what cognitive psychologists have discovered is that when you ask a human you’re not really getting at the decision process. They make a decision first, and then you ask, and then they generate an explanation and that may not be the true explanation.”

- Peter Norvig

(Muhamad Aurangzeb et al.: “Explainable AI in Healthcare”)

Introduction

Individual predictions (local)	Whole prediction process (global)
Directly from prediction process (self-explaining)	Post-processing required (post-hoc)
Convincing humans (Plausibility)	Reflects true reasoning process of model (Faithfulness)

- Model-specific vs. model-agnostic
- Examples and counter-examples
- Proxy models
- ...

(Alon Jacovi and Yoav Goldberg.: “Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness?”
Andreas Rauber.: “Security, Privacy and Explainability In ML - Explainability” TU Wien)

Introduction

Individual predictions (local)	Whole prediction process (global)
Directly from prediction process (self-explaining)	Post-processing required (post-hoc)
Convincing humans (Plausibility)	Reflects true reasoning process of model (Faithfulness)

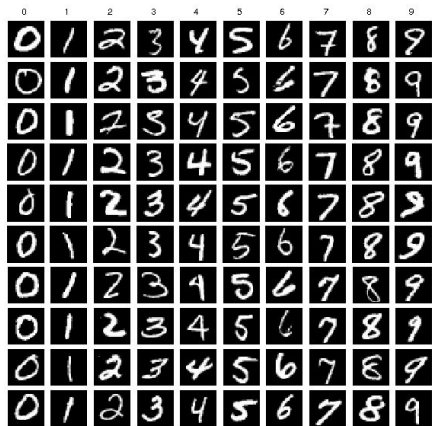
- Model-specific vs. model-agnostic
- Examples and counter-examples
- Proxy models
- ...

Introduction

Individual predictions (local)	Whole prediction process (global)
Directly from prediction process (self-explaining)	Post-processing required (post-hoc)
Convincing humans (Plausibility)	Reflects true reasoning process of model (Faithfulness)

- Model-specific vs. model-agnostic
- Examples and counter-examples
- Proxy models
- ...

MNIST data & IMDB Dataset of 50K Movie Reviews

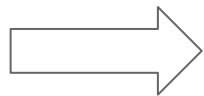
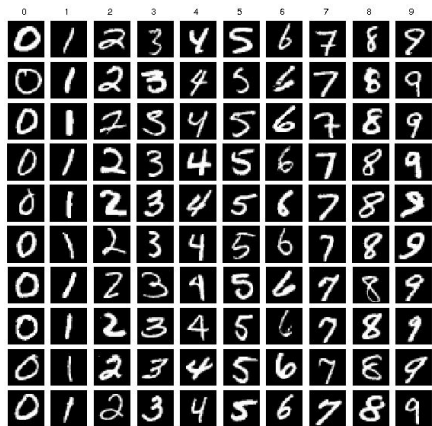


(LeCun et al.: “The MNIST Dataset Of Handwritten Digits”)



(Maas et al.: “Learning Word Vectors for Sentiment Analysis”)

MNIST data & IMDB Dataset of 50K Movie Reviews

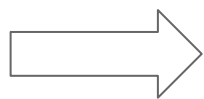
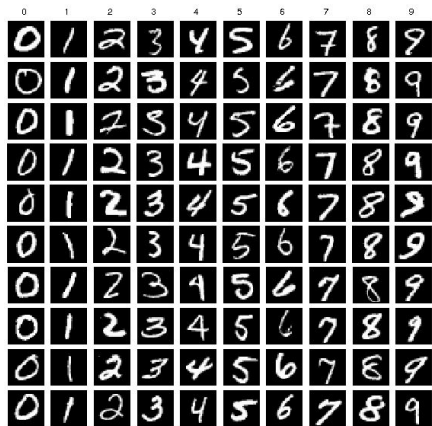


10 classes

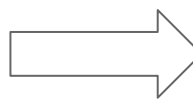


2 classes

MNIST data & IMDB Dataset of 50K Movie Reviews



Convolutional &
Dense layers



LSTM &
Dense layers

LIME

LIME

- Model agnostic
- Local explanation
- Explanation through visual or textual artifacts
- Usage of interpretable classifier

“Why Should I Trust You?” Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marco@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

Xiv:1602.04938v3 [cs.LG] 9 Aug 2016

ABSTRACT

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing *trust*, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one.

In this work, we propose LIME, a novel explanation technique that explains the predictions of *any* classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. We also propose a method to explain models by presenting representative individual predictions and their explanations in a non-redundant way, framing the task as a submodular optimization problem. We demonstrate the flexibility of these methods by explaining different models for text (e.g. random forests) and image classification (e.g. neural networks). We show the utility of explanations via novel experiments, both simulated and with human subjects, on various scenarios that require trust: deciding if one should trust a prediction, choosing between models, improving an untrustworthy classifier, and identifying why a classifier should not be trusted.

1. INTRODUCTION

Machine learning is at the core of many recent advances in science and technology. Unfortunately, the important role of humans is an oft-overlooked aspect in the field. Whether humans are directly using machine learning classifiers as tools,

how much the human understands a model’s behaviour, as opposed to seeing it as a black box.

Determining trust in individual predictions is an important problem when the model is used for decision making. When using machine learning for medical diagnosis [6] or terrorism detection, for example, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.

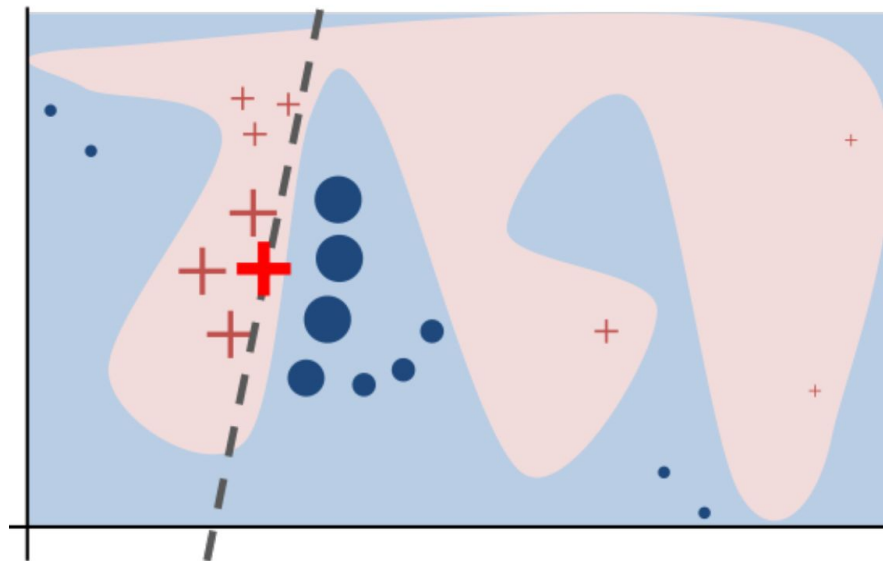
Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it “in the wild”. To make this decision, users need to be confident that the model will perform well on real-world data, according to the metrics of interest. Currently, models are evaluated using accuracy metrics on an available validation dataset. However, real-world data is often significantly different, and further, the evaluation metric may not be indicative of the product’s goal. Inspecting individual predictions and their explanations is a worthwhile solution, in addition to such metrics. In this case, it is important to aid users by suggesting which instances to inspect, especially for large datasets.

In this paper, we propose providing explanations for individual predictions as a solution to the “trusting a prediction” problem, and selecting multiple such predictions (and explanations) as a solution to the “trusting the model” problem. Our main contributions are summarized as follows.

- LIME, an algorithm that can explain the predictions of *any* classifier or regressor in a faithful way, by approximating it locally with an interpretable model.
- SP-LIME, a method that selects a set of representative instances with explanations to address the “trusting the model” problem, via submodular optimization.

LIME algorithm

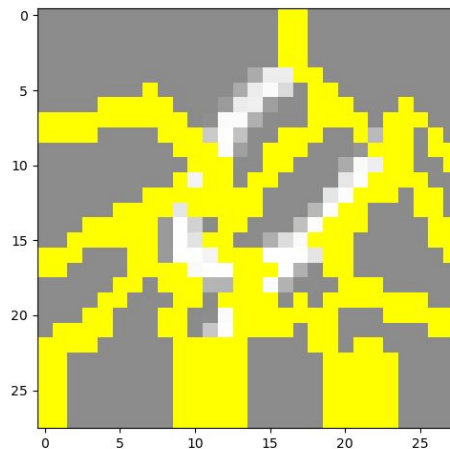
- Sampling around sample of interest
- Distances from new samples to original one
- Training of an interpretable classifier with weighted samples
- Extraction of coefficients from interpretable classifier
- Usage of coefficients for artifact creation



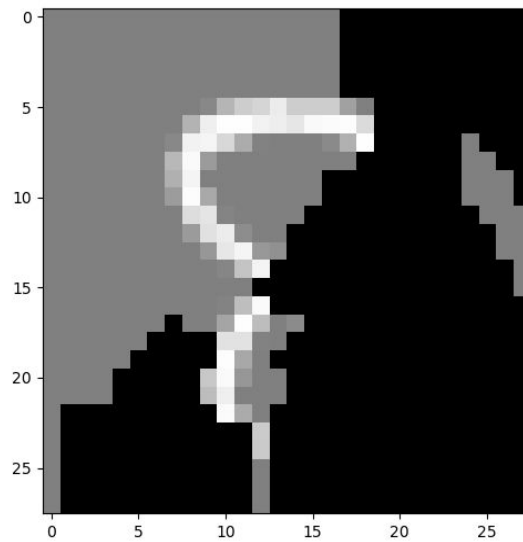
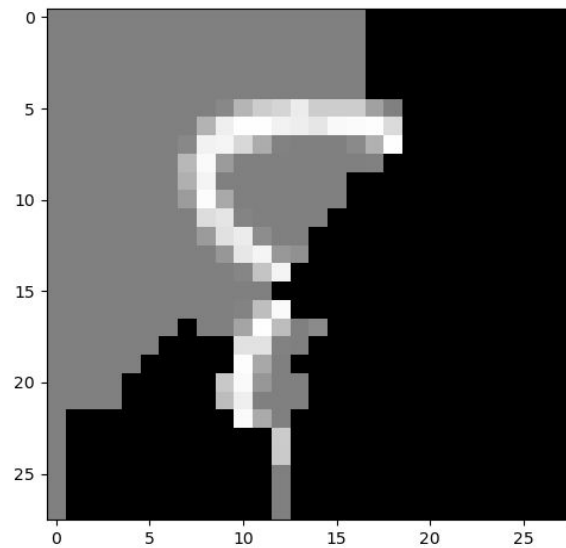
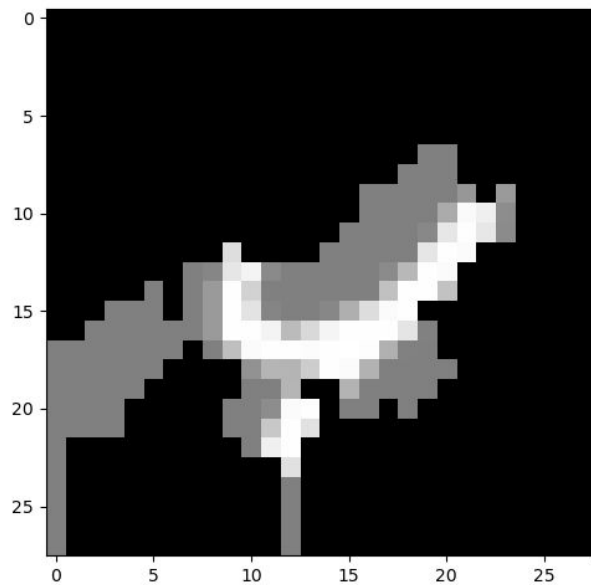
(Ribeiro et al.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier)

- Sampling: Switching attributes off, leaving out word, switching (super-)pixels off
- Distance calculation: Cosine or euclidean distance for text, pixels, etc.
- Interpretable Classifier: Linear classifiers like Linear Regression
- Order on attributes by sorting the coefficients of the interpretable classifier
- Artifact creation by choosing top-N attributes or colour coding of attributes with respect to coefficients

- ☐ this movie is amazing while being funny and entertaining it is also profoundly deep and eye opening
- ☐ this ____ is amazing while ____ funny and entertaining it is ____ profoundly ____ and eye ____
- ☐ ____ movie ____ amazing ____ being funny and entertaining it is also ____ deep and eye ____
- ☐ this ____ is amazing while being ____ and ____ entertaining it is also profoundly ____ and eye opening
- ☐ this ____ is amazing while being funny ____ entertaining ____ is also ____ deep and eye opening
- ☐ ...

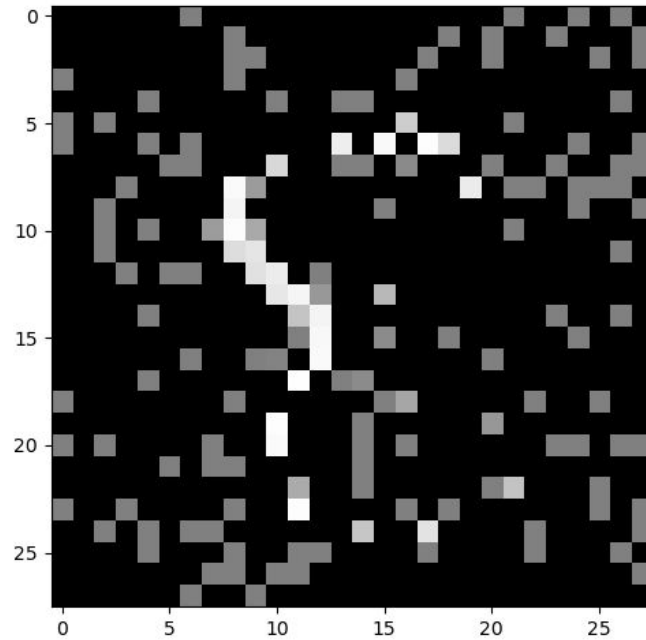


...or just use all pixels

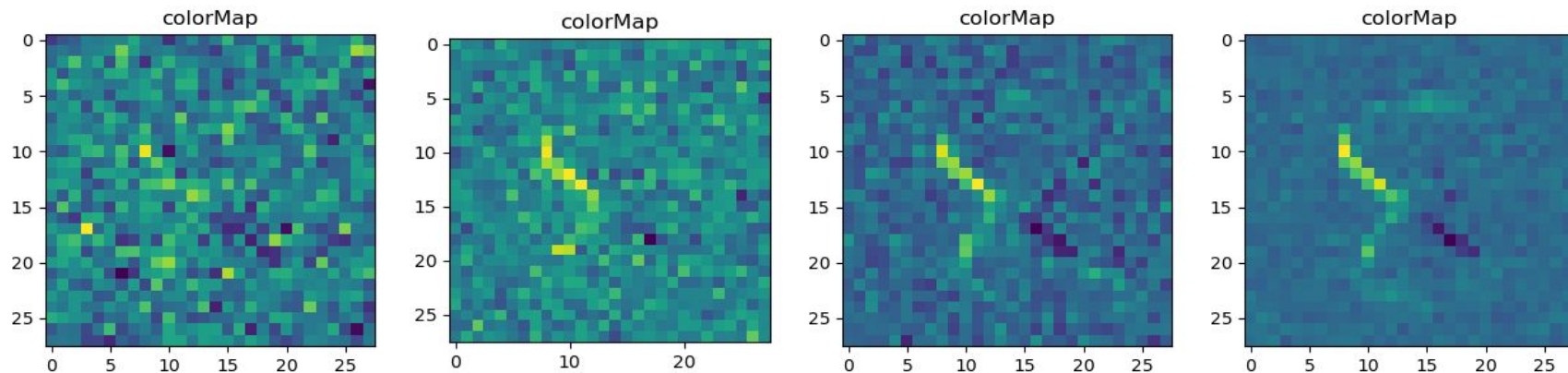


2, 3 and 4 superpixels active

LIME on MNIST - examples



most important 20% pixels active



Importance of individual pixels with 111, 222, 1111 and 2222 samples from left to right

Sentiment Classification

Example

negative

the king maker will be success in where the similar but superior the legend of set box office records the film directed by after screenplay by sean casey based on historical fact in has some amazingly beautiful visual elements but is by one of the pedestrian scripts and story development on film the event the picture relates is the arrival of the portuguese soldier of fortune de gary stretch whose vengeance for this father murderer drives him to captured and thrown into slavery and put on the in in the kingdom of where he is purchased by the beautiful maria cindy with the of her father john rhys davies man with name and past that are revealed as the story progresses there is plot to the king and and his new sidekick after some gratuitous cgi enhanced choreographed martial arts silliness are first rewarded by the king to become his only to be imprisoned together once queen reveals her plot to kill the king and son to allow her lover lord oliver to take over the rule of yet of course and escape and are condemned to fight each other to save the lives of their families wife and children and now firm love affair with maria with the expected consequences the acting with the exception of john rhys davies is so weak that the film occasionally seems as though it were meant to be camp the cast struggle with the poorly written dialog making us wish they had used their native with subtitles the musical score by ian sounds as though form old tv soap operas but if it is visual you re after there is plenty of that and that alone makes the movie worth watching it is film that has obvious high financial backing for all the special effects and masses of cast and sets and shows its good intentions it is just the that are missing

Sentiment Classification

Example

-----positive-----
vanilla sky was wonderfully thought out movie or rather
los was well thought out watched that movie late one night
excited about what was to come wasn't disappointed by the
end of the movie was couldn't get it off my mind the whole
idea of it just blew me away the ending was more of surprise
than could ever do the plot line was also something that
kept me interesting through and through the cast superb
it was an all around wonderful movie the kind of movie
you can watch again and again and always find something
new we've seen it four or five times and I'm always finding
something new it's a movie to keep you interested forever

-----positive-----

it should however their leader played by davis is delightfully memorable creation and watching how they handle the troops technology with their simple natural weapons provides nice contrast by the time we get to the third act though the pace picks up again as we between the battle against the troops and the rebel forces an attack against the empire all new half completed death star and luke final showdown with darth vader and the emperor the latter ties with the palace sequence as the highlight of the movie mark his acting chops once again as luke in these scenes and watching him as fully jedi knight makes for an unforgettable performance also as iconic as james earl jones voice as darth vader is he is only by the like emperor played with deliciously frightening evil by ian the tension between this trio the excitement of this climactic moment which is appropriately darkly lit and the star wars movies have always set standards for special effects and the technical work in return of the jedi can easily hold candle to its predecessors the space battle fights are as as always and the bike chase through the forest is of course given that this movie was made after new hope and the empire strikes back it probably shouldn't be so surprising that the special effects have reached an even greater level of excellence the acting is classic star wars fare harrison ford and carrie fisher all mature and into their roles and anthony daniels provides more hilarious moments as frank oz only appears in two scenes but he makes the most of it and ucs there also inhn

Sentiment Classification - StarWars example

positive

it should however their leader played by davis is delightfully memorable creation and watching how they handle the troops technology with their simple natural weapons provides nice contrast by the time we get to the third act though the pace picks up again as we between the battle against the troops and the rebel forces an attack against the empire all new half completed death star and luke final showdown with darth vader and the emperor the latter ties with the palace sequence as the highlight of the movie mark his acting chops once again as luke in these scenes and watching him as fully jedi knight makes for an unforgettable performance also as iconic as james earl jones voice as darth vader is he is only by the like emperor played with deliciously frightening evil by ian the tension between this trio the excitement of this climactic moment which is appropriately darkly lit and the star wars movies have always set standards for special effects and the technical work in return of the jedi can easily hold candle to its predecessors the space battle fights are as as always and the bike chase through the forest is of course given that this movie was made after new hope and the empire strikes back it probably shouldn't be so surprising that the special effects have reached an even greater level of excellence the acting is classic star wars fare harrison ford and carrie fisher all mature and into their roles and anthony daniels provides more hilarious moments as frank oz only appears in two scenes but he makes the most of it and yes there also john

williams music all told while return of the jedi little bit in the middle the first and third acts deliver in style making this rather satisfactory finale to one of the greatest ever in george lucas re released the classic star wars in restored and special which featured added in effects and or shots as well as some of the three return of the jedi appears to have caused the most with star wars fans perhaps it can be due to the out of place albeit funny if you're not so easily offended jedi rocks musical number in palace which although technically amazing does the flow of the film however did like the ending montage

Sentiment Classification - StarWars example

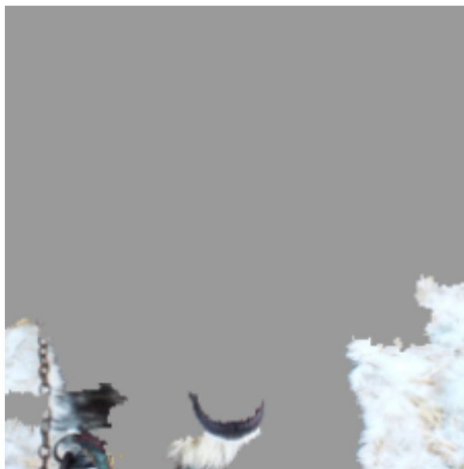
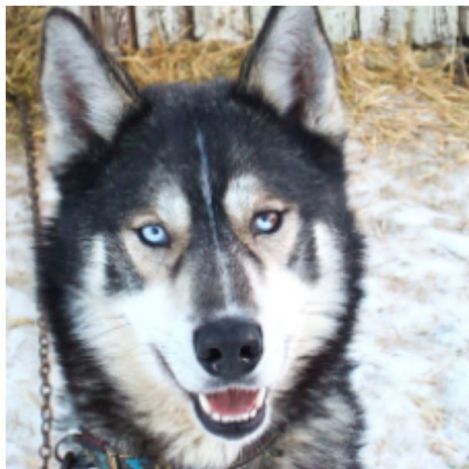
positive

it should however their leader played by davis is delightfully memorable creation and watching how they handle the troops technology with their simple natural weapons provides nice contrast by the time we get to the third act though the pace picks up again as we between the battle against the troops and the rebel forces an attack against the empire all new half completed death star and luke final showdown with darth vader and the emperor the latter ties with the palace sequence as the highlight of the movie mark his acting chops once again as luke in these scenes and watching him as fully jedi knight makes for an unforgettable performance also as iconic as james earl jones voice as darth vader is he is only by the like emperor played with deliciously frightening evil by ian the tension between this trio the excitement of this climactic moment which is appropriately darkly lit and the star wars movies have always set standards for special effects and the technical work in return of the jedi can easily hold candle to its predecessors the space battle fights are as as always and the bike chase through the forest is of course given that this movie was made after new hope and the empire strikes back it probably shouldn't be so surprising that the special effects have reached an even greater level of excellence the acting is classic star wars fare harrison ford and carrie fisher all mature and into their roles and anthony daniels provides more hilarious moments as frank oz only appears in two scenes but he makes the most of it and yes there also john

williams music all told while return of the jedi little bit in the middle the first and third acts deliver in style making this rather satisfactory finale to one of the greatest ever in george lucas re released the classic star wars in restored and special which featured added in effects and or shots as well as some of the three return of the jedi appears to have caused the most with star wars fans perhaps it can be due to the out of place albeit funny if you're not so easily offended jedi rocks musical number in palace which although technically amazing does the flow of the film however did like the ending montage

lucas re released the classic star wars in restored and special which featured added in effects and or shots as well as some of the three return of the jedi appears to have caused the most with star wars fans perhaps it can be due to the out of place albeit funny if you're not so easily offended jedi rocks musical number in palace which although technically amazing does the flow of the film however did like the ending montage scenes where we see victory occurring on the various planets of the galaxy this dvd version features yet more we get to see more montage finale scenes notably on where we hear what sounds like jar jar screaming free and in what is probably the most controversial change as the of in the closing scenes probably due to the intense and unfair disdain fans have for his somewhat shaky work in episode ii attack of the it seems inevitable that fans would put this edition down for that alone however if you're watching the star wars saga and about it chances are you may react little differently nonetheless it is an issue that fans have raised so it probably best to be warned beforehand as nice as it would be to have lucas release the original versions of these three classic films he nonetheless stands by what he said about these being the definitive of his classic trilogy and when viewing the star wars movies altogether as one complete saga as lucas intended it actually makes sense to keep them technically and consistent the original films will always be in our memories but these new are just as much fun if one can give them chance

Sentiment Classification - StarWars example



Husky classified as wolf and superpixels as explanation

(Ribeiro et al.: ""Why Should I Trust You?": Explaining the Predictions of Any Classifier")

Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)

Subject: Another request for Darwin Fish

Organization: University of New Mexico, Albuquerque

Lines: 11

NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.

This is the same question I have and I have not seen an answer on the

net. If anyone has a contact please post on the net or email me.

Text classified as atheism and
highlighted words as explanation

(Ribeiro <https://github.com/marcotcr/lime>)

Unwanted Bias detection with LIME

Faithful Interpretations & Explanation Sets

Faithful Interpretations

- Faithfulness: “[A] faithful interpretation is one that accurately represents the reasoning process behind the model’s prediction.”

Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?

Alon Jacovi

Bar Ilan University
alonjacovi@gmail.com

Yoav Goldberg

Bar Ilan University and Allen Institute for AI
yoav.goldberg@gmail.com

Abstract

With the growing popularity of deep-learning based NLP models, comes a need for interpretable systems. But what is interpretability, and what constitutes a high-quality interpretation? In this opinion piece we reflect on the current state of interpretability evaluation research. We call for more clearly differentiating between different desired criteria an interpretation should satisfy, and focus on the faithfulness criteria. We survey the literature with respect to faithfulness evaluation, and arrange the current approaches around three assumptions, providing an explicit form to how faithfulness is “defined” by the community. We provide concrete guidelines on how evaluation of interpretation methods should and should not be conducted. Finally, we claim that the current binary definition for faithfulness sets a potentially unrealistic bar for being considered faithful. We call for discarding the binary notion of faithfulness in favor of a more graded one, which we believe will be of greater practical utility.

One such pain is the challenge of defining—and evaluating—what constitutes a quality interpretation. Current approaches define interpretation in a rather ad-hoc manner, motivated by practical use-cases and applications. However, this view often fails to distinguish between distinct aspects of the interpretation’s quality, such as readability, plausibility and faithfulness (Herman, 2017).² We argue (§2, §5) such conflation is harmful, and that faithfulness should be defined and evaluated *explicitly*, and independently from plausibility.

Our main focus is the *evaluation of the faithfulness* of an explanation: **a faithful interpretation is one that accurately represents the reasoning process behind the model’s prediction**. We find this to be a pressing issue: in cases where an explanation is required to be faithful, imperfect or misleading evaluation can have disastrous effects.

While literature in this area may implicitly or explicitly evaluate faithfulness for specific explanation techniques, there is no consistent and formal definition of faithfulness. We uncover three assumptions that underlie all these attempts. By making the assumptions explicit and organizing the

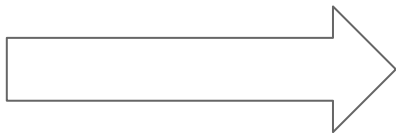
1 Introduction

- **Assumption 1 (The Model Assumption).**
Two models will make the same predictions if and only if they use the same reasoning process.
- **Assumption 2 (The Prediction Assumption).**
On similar inputs, the model makes similar decisions if and only if its reasoning is similar.
- **Assumption 3 (The Linearity Assumption).**
Certain parts of the input are more important to the model reasoning than others. Moreover, the contributions of different parts of the input are independent from each other.

- **Corollary 1.1.** An interpretation system is unfaithful if it results in different interpretations of models that make the same decisions
- **Corollary 2.** An interpretation system is unfaithful if it provides different interpretations for similar inputs and outputs

(Alon Jacovi and Yoav Goldberg.: “Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness?”)

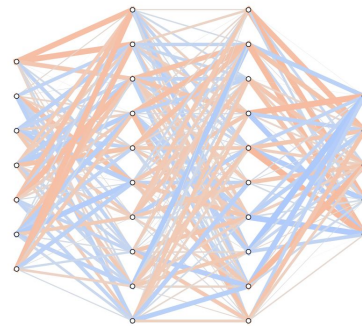
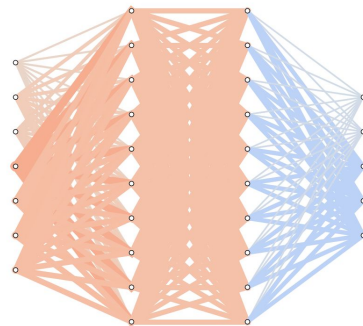
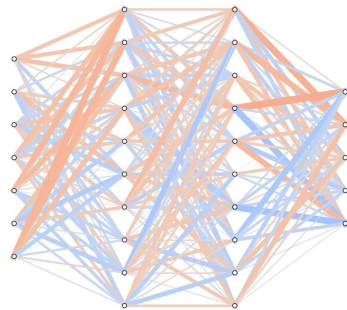
- **Assumption 1 (The Model Assumption).**
Two models will make the same predictions if and only if they use the same reasoning process.
- **Assumption 2 (The Prediction Assumption).**
On similar inputs, the model makes similar decisions if and only if its reasoning is similar.
- **Assumption 3 (The Linearity Assumption).**
Certain parts of the input are more important to the model reasoning than others. Moreover, the contributions of different parts of the input are independent from each other.
- **Corollary 1.1.** An interpretation system is unfaithful if it results in different interpretations of models that make the same decisions
- **Corollary 2.** An interpretation system is unfaithful if it provides different interpretations for similar inputs and outputs



Explanation Sets

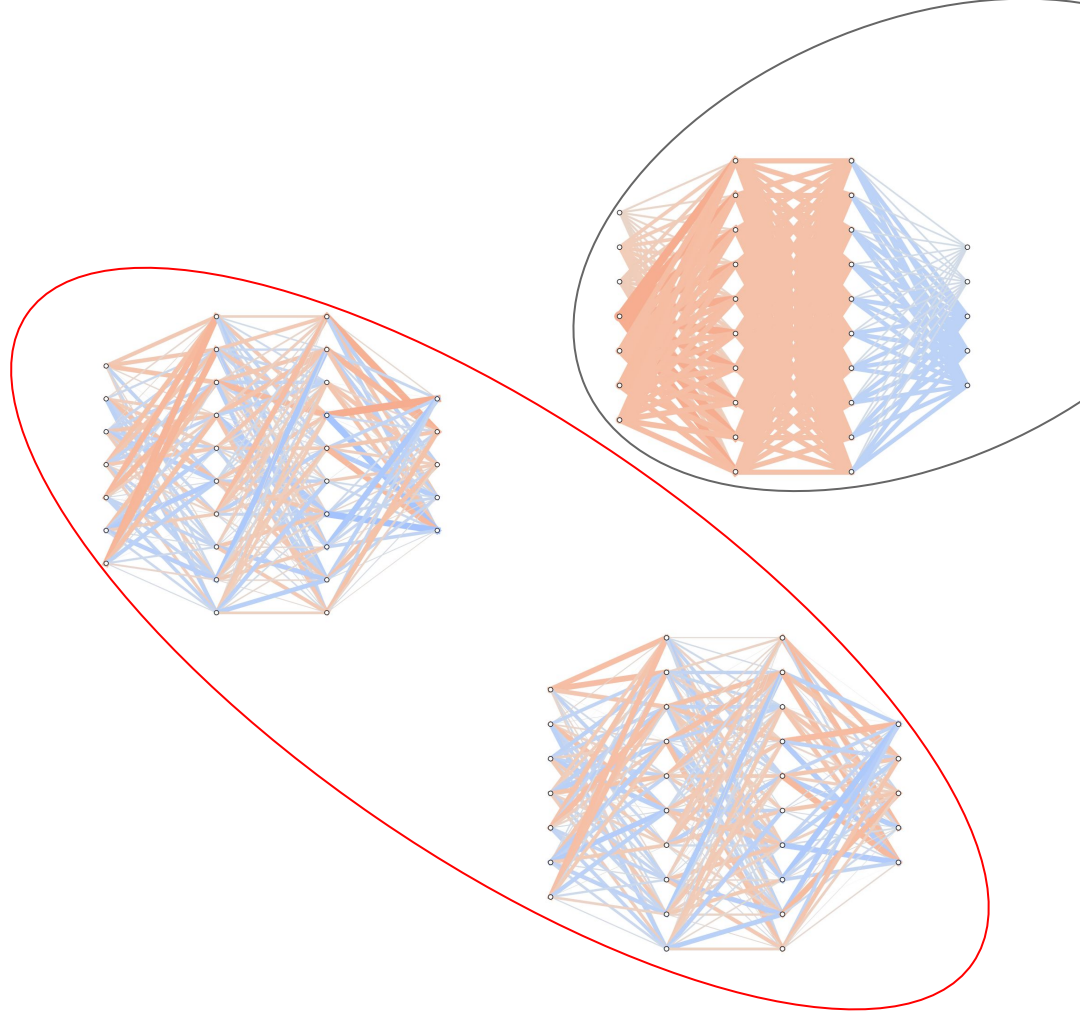
Explanation Sets

- Different samples lead to different activations of neurons
- Grouping of all training samples with respect to their neuron activation



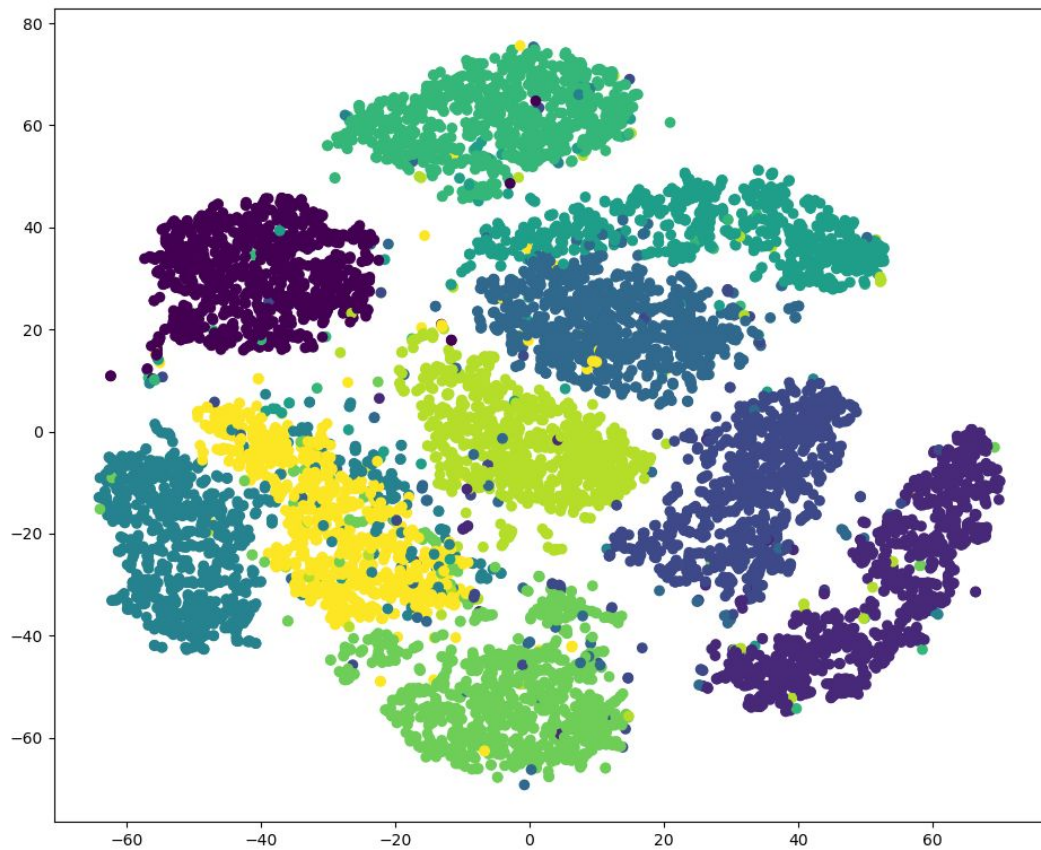
Explanation Sets

- Different samples lead to different activations of neurons
- Grouping of all training samples with respect to their neuron activation



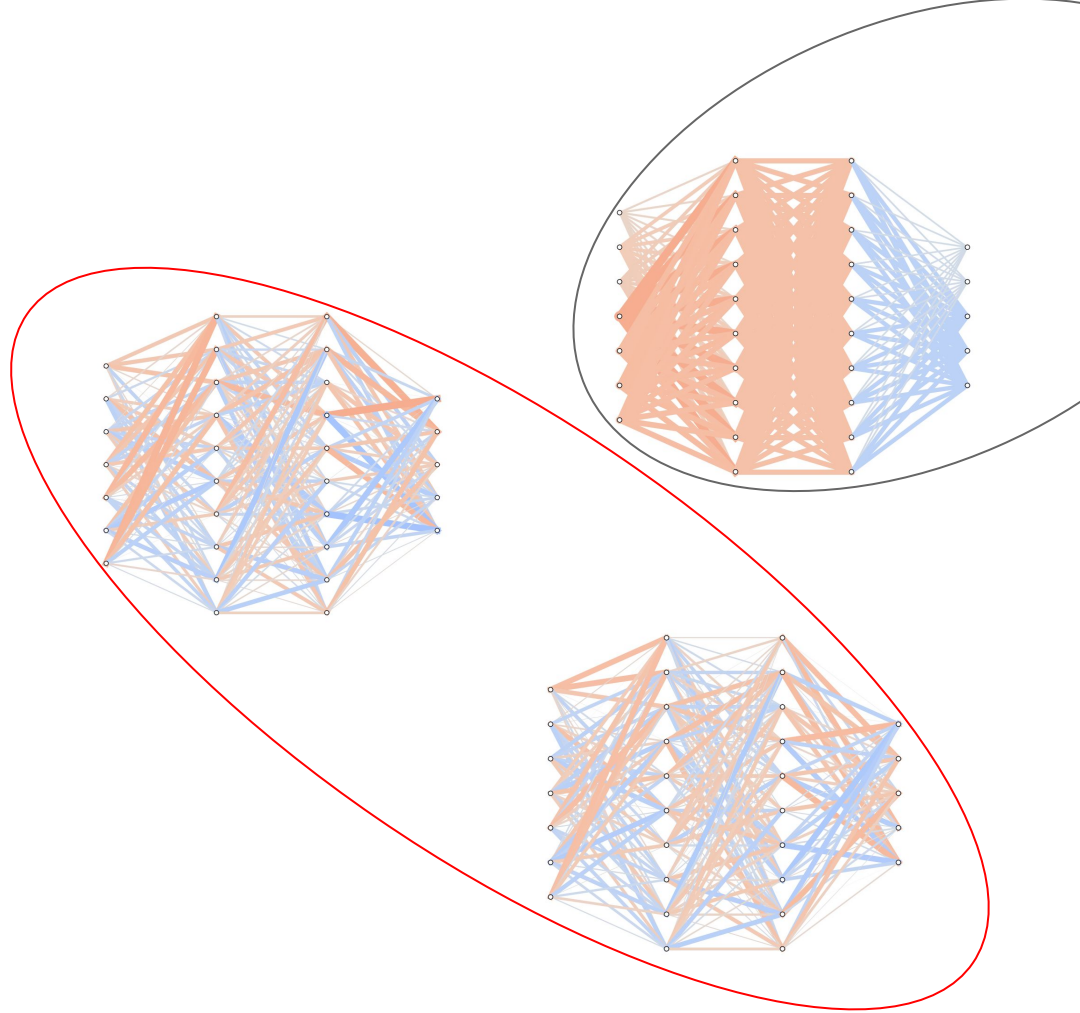
Explanation Sets

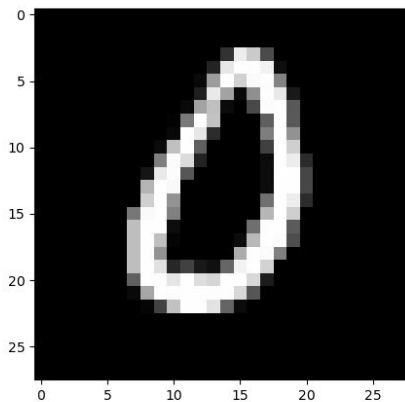
Activations clustered in 2 dimensional space



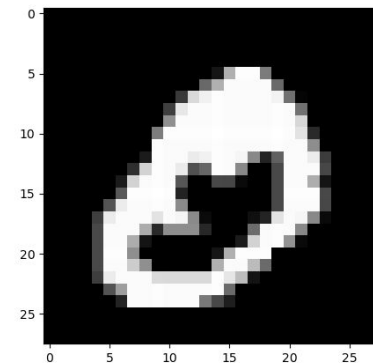
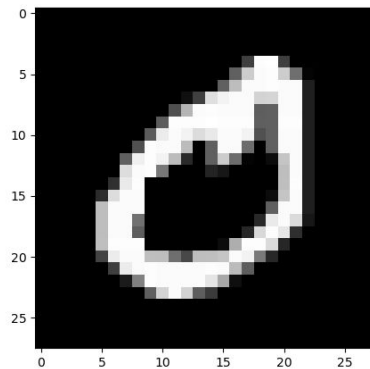
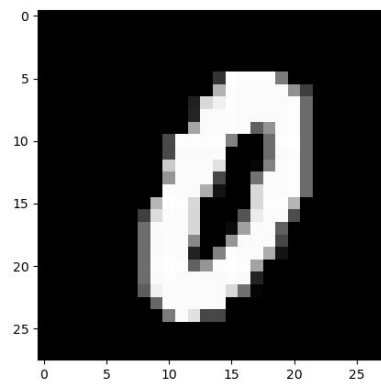
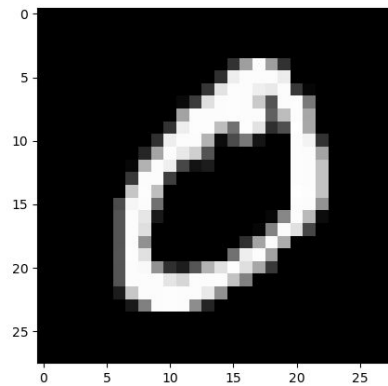
Explanation Sets

- Different samples lead to different activations of neurons
- We group all training samples with respect to their neuron activation
- A new sample gets grouped into one of the group by its own neuron activity
- Training samples from this group serve as “faithful” explanations

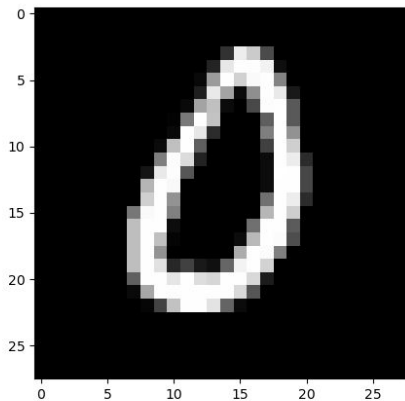




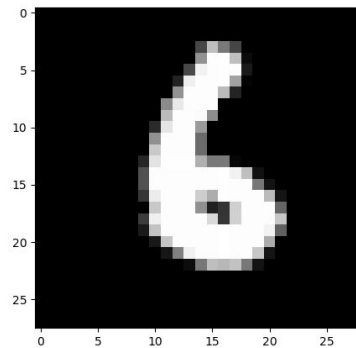
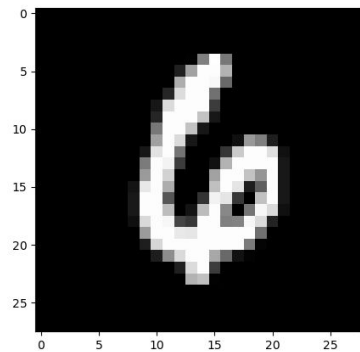
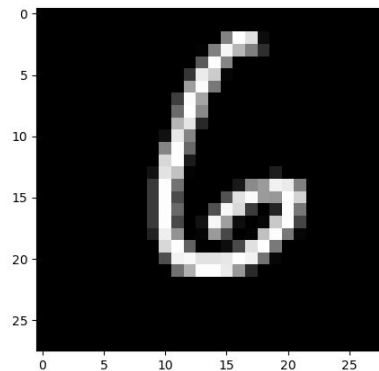
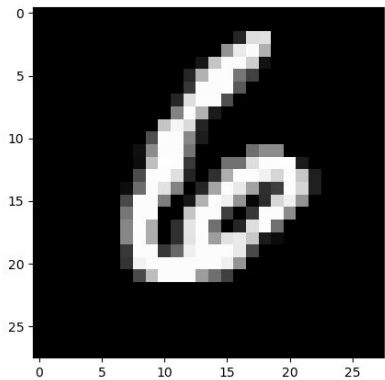
Sample to explain



Samples which evoke similar neuron activity



Sample to explain



Samples which evoke similar neuron activity

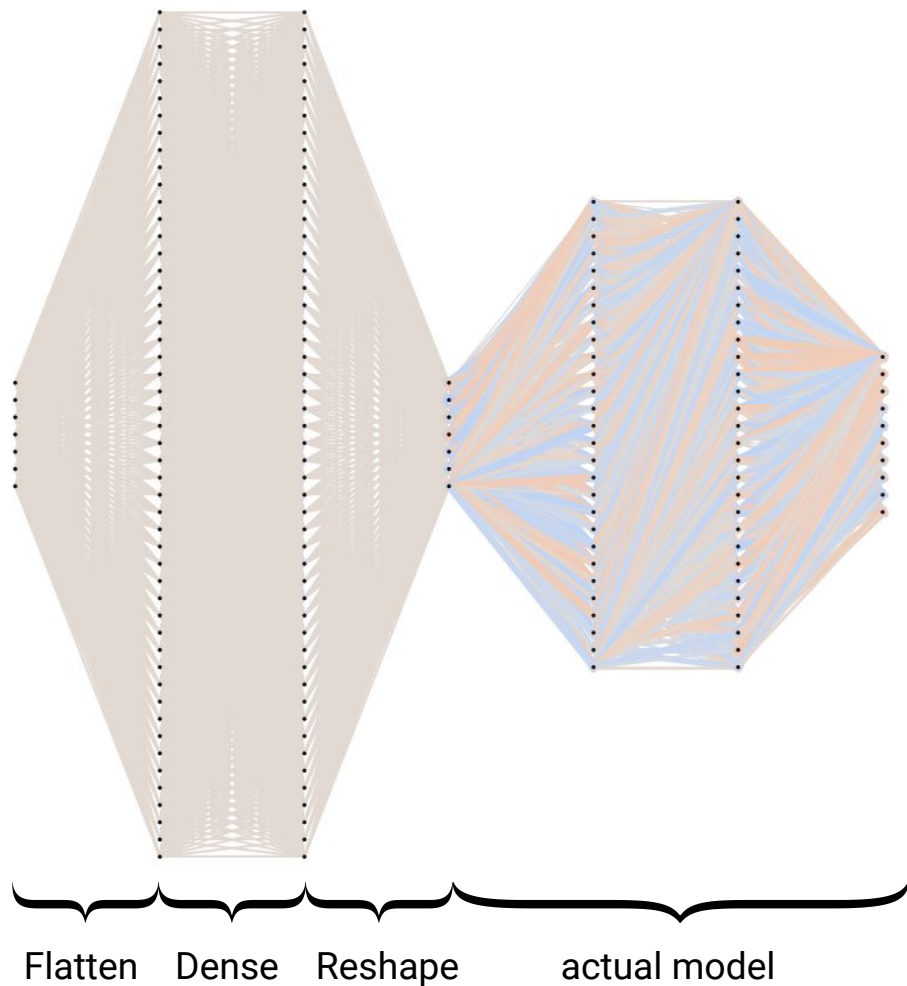
- ❖ **When the model "saw"**
 - Im guessing that the movie was based on book given the number of characters and subplots during thought that the movie creators perhaps the writer or director intended to create an epic movie [...]
- ❖ **the neuron activity was similar to when it "saw" the following training samples:**
 - there is no such thing as perfect murder lieutenant columbo knows that ken franklin who is the other half of the writing team of detective stories doesn't know that he kills his partner jim who had plans [...]
 - Ive been fan of columbo since childhood and still enjoy watching them there was break for many years that they weren't showing columbo stories at all but now he back back for one more question
 - this is my first awful rating ever on imdb and couldn't think of more deserving film to honor it with hoped for entertaining trash and found trash of the saddest kind found film which no one can possibly have cared bit about including its creator hell ride directed written by and starring larry friend of bishop [...]
 - the main reason for writing this review is found this of great play and worthy film horrible movie experience if can save someone from watching it will have done good thing this new version is [...]
 - I rented this when it came out on video in after it again my idea about it hasn't changed much was an adult then and Im still an adult now illogical elements mentioned by other reviewers didn't bother me this isn't a documentary it's a fantasy story where animals can talk [...]

Counterfactual Examples

Counterfactual Examples

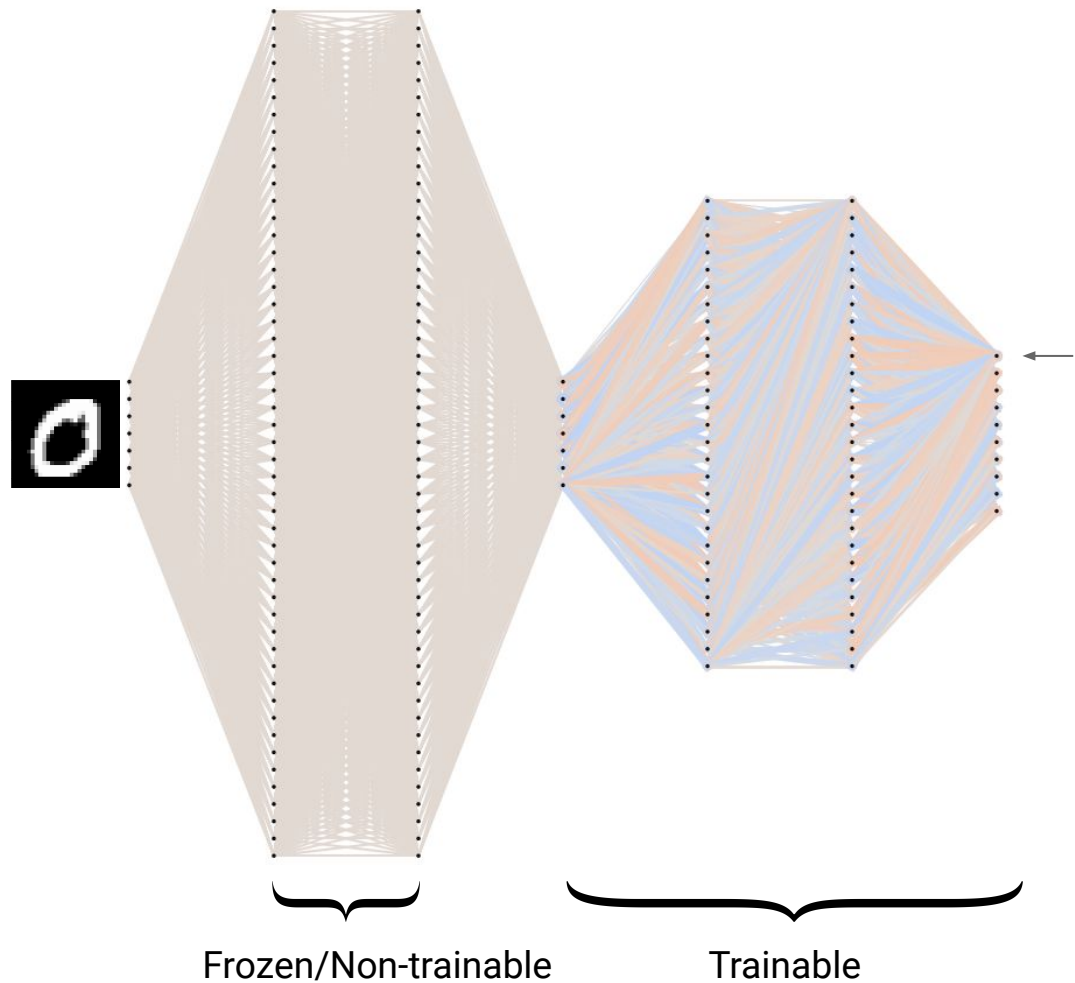
“What-if”-method:

- the model should tell us how a sample must look like in order to be classified in a specific class.
- Achieved by an encoder decoder structure in front of the model.
- Encoder Decoder is initialized as an identity function.



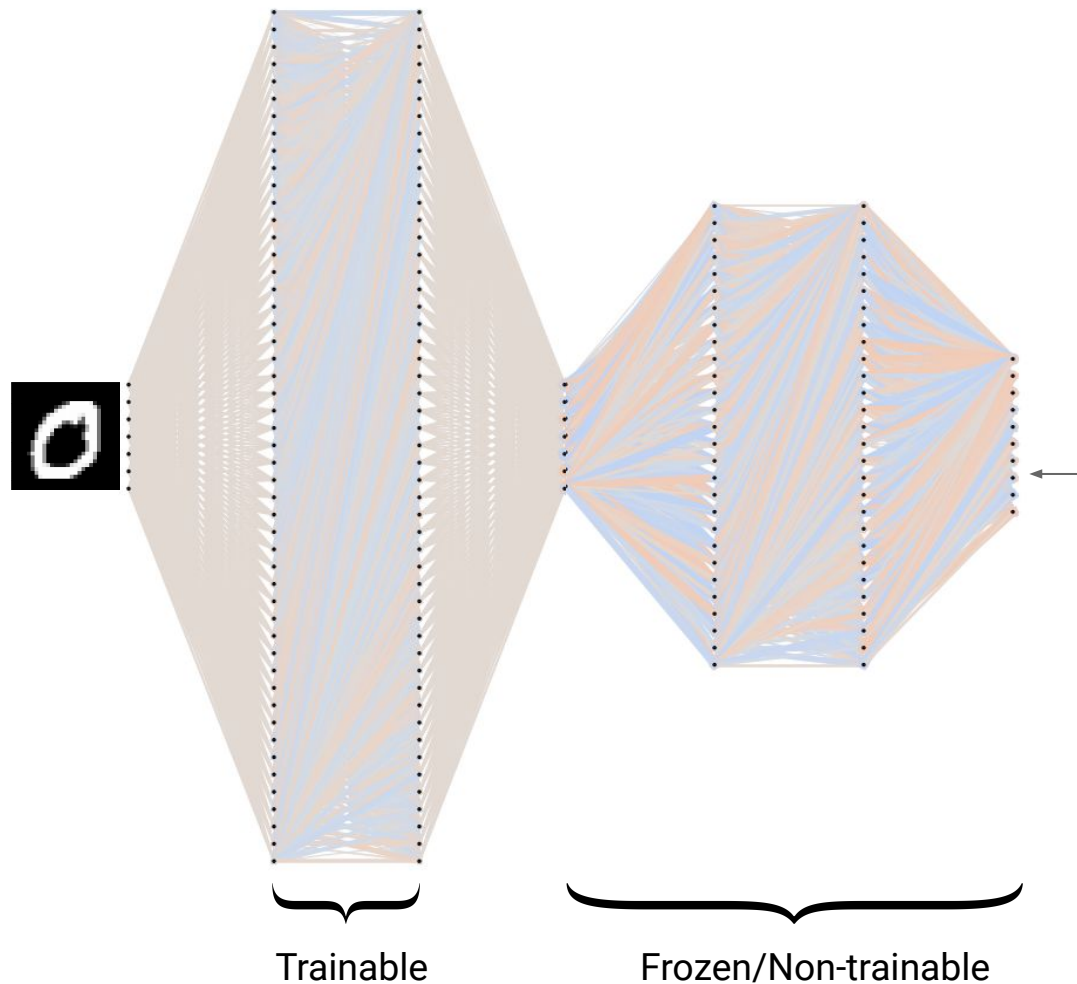
Counterfactual Examples

Normal training phase



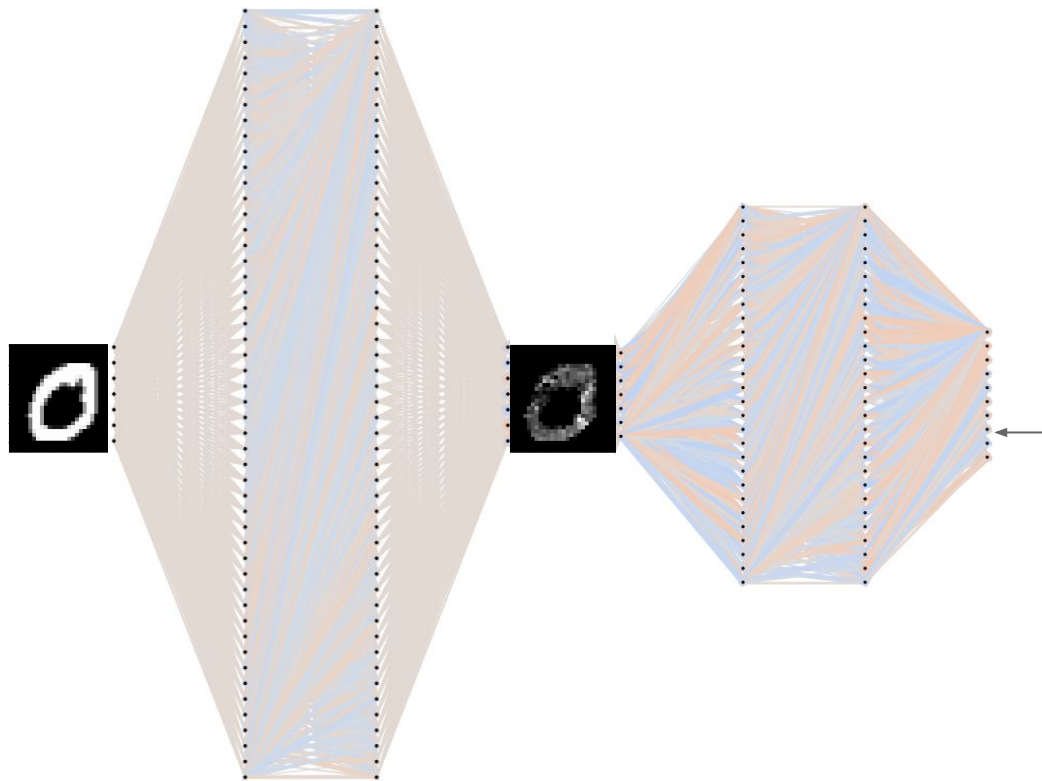
Counterfactual Examples

Sample creation phase



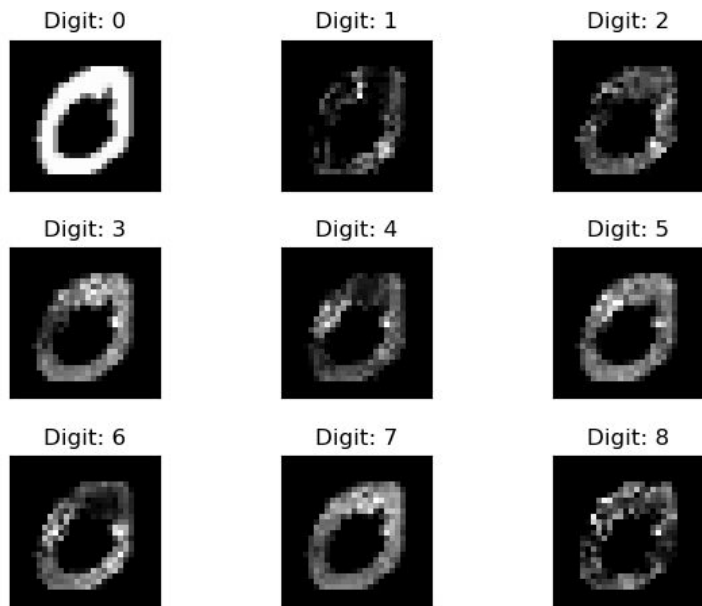
Counterfactual Examples

New sample extraction

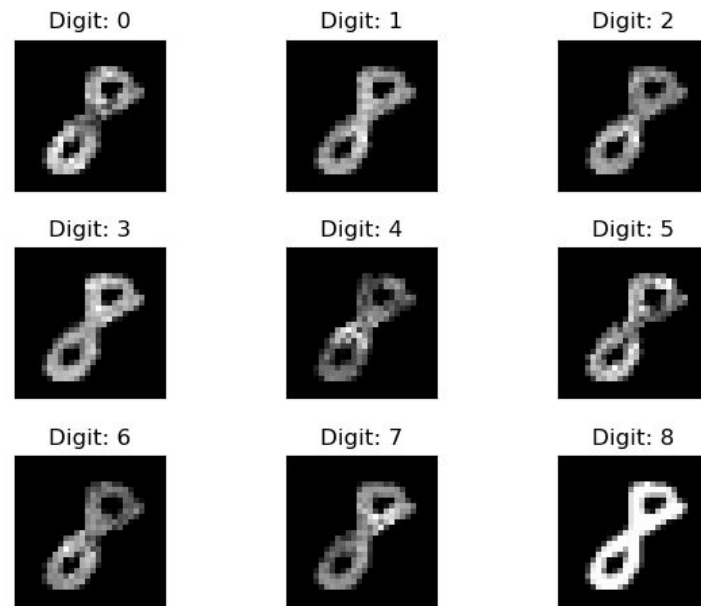


New sample belongs to class x with probability p

Counterfactual Examples

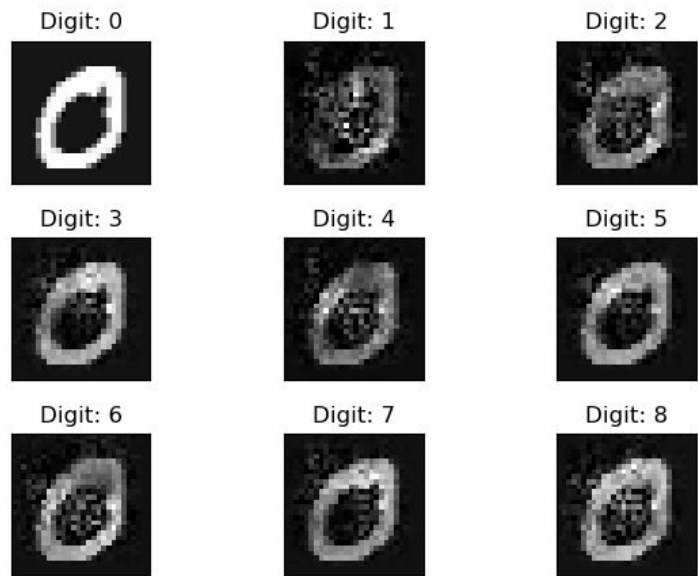


more than 75% probability

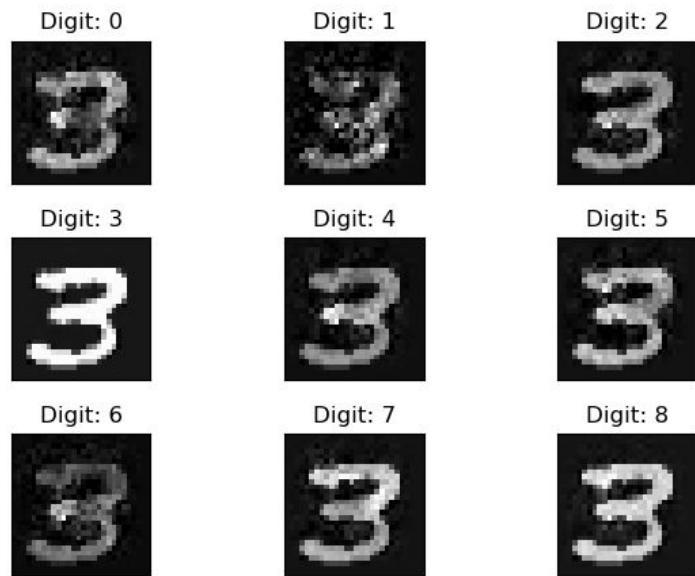


more than 85% probability

Counterfactual Examples



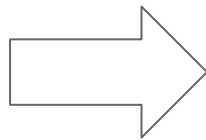
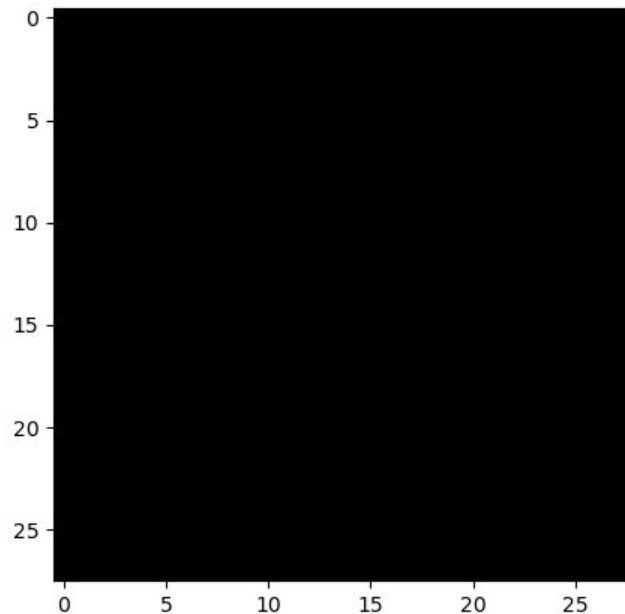
more than 85% probability



more than 90% probability

Examples - non black background

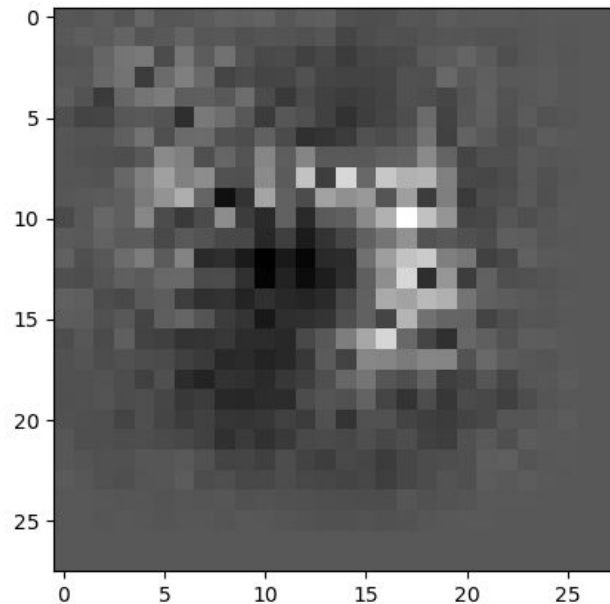
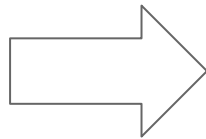
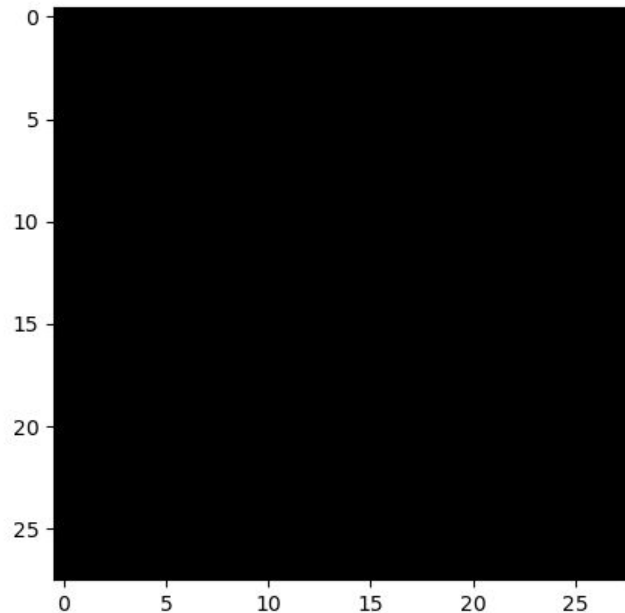
Counterfactual Examples



Class 7 with more than 75%

Examples - non black background - from nothing

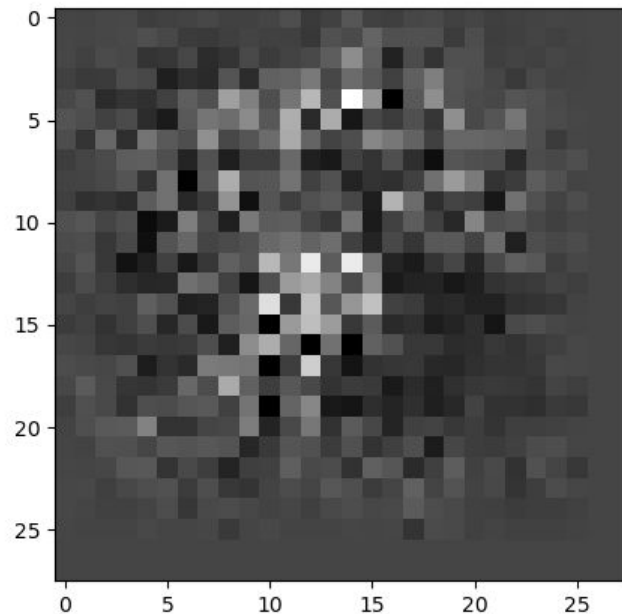
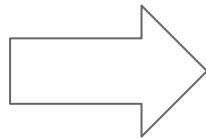
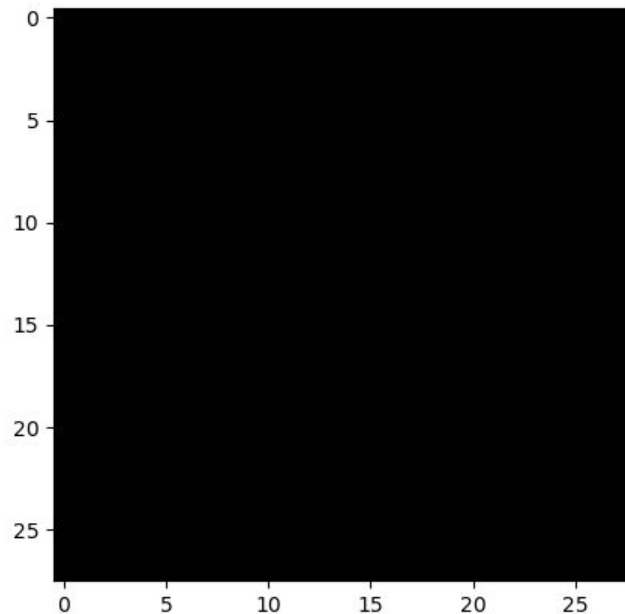
Counterfactual Examples



Class 7 with more than 75%

Examples - non black background - from nothing to something

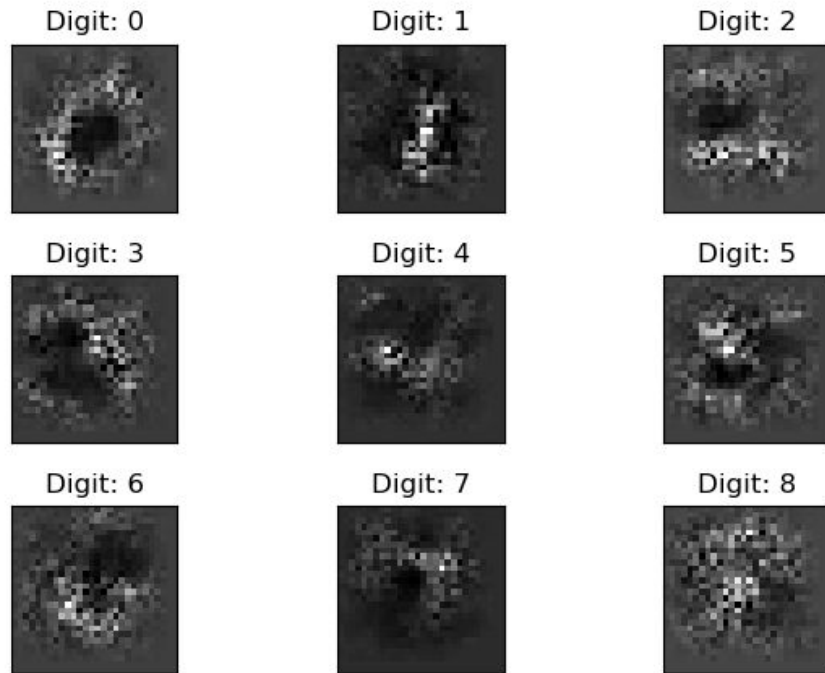
Counterfactual Examples



Class 8 with more than 75%

Examples - non black background - from nothing to something

Counterfactual Examples



more than 75% probability

Examples - non black background - from nothing to something

Counterfactual Examples

Possible Extensions:

- Only allow change in some parts of the image
- Permutation of (super-)pixels

Counterfactual Examples

Possible Extensions:

- Only allow change in some parts of the image
- Permutation of (super-)pixels
- Ideas ?



Counterfactual creation with encoder-decoder structure not possible for the sentiment classification task using GD, since :

- Embedding layer is not differentiable
- Inverse Embedding is hard to find

Thank you for your attention!

Questions/ideas/feedback ?

Pictures from papers or own if nothing specified