Explainable AI

An introduction

Introduction

Explainable AI

- large research area
- much recent attention in the machine learning community
- long history in AI research
- much domain-specific work

Black box nature of many AI systems leads to lack of transparency

Makes it difficult to explain their decisions

Explainable AI (XAI) promotes AI algorithms that can show their internal process and explain how they make their decisions.

XAI aims to extract insights about how predictions have been performed or how the AI model operates for a set of inputs.

Deep learning has outperformed traditional ML

Reliance on AI decisions:

Product recommendations
Friend suggestions
News recommendation
Autonomous Vehicles in transportation
Financial decisions
Medical recommendations
.....

Regulations

GDPR
FDA on medical decisions/predictions
algorithmic accountability act 2019
many others

Trust and Adoption

Users need to understand why AI makes specific recommendations

If there is a lack of trust, then lower adoption etc.

Ethical Responsibility

Accountability for algorithmic decisions

Detecting and mitigating bias

Debugging & Improvement

Understanding failures (various types)

Model and algorithm enhancement

Detecting adversarial attacks.

Informing feature engineering and future data collection.

Many well known cases

COMPAS Recidivism Algorithm (2016): criminal risk assessment tool was biased against Black defendants

Amazon's Recruiting Tool (2018): discriminated against women. Trained on resumes mostly from men; learned a bias

A**pple Card Credit Limits** (2019): algorithm granted significantly lower credit limits to women compared to men with identical financial profiles

UK A-Level Algorithm (2020) system disproportionately downgraded students from disadvantaged backgrounds while favouring those from prestigious schools

XAI - evaluation

Evaluation of explanations ?

- **CCorrectness**: How accurately the explanation represents the model's actual decision process
- **CComprehensibility**: How understandable the explanation is to the target audience
- **EEfficiency**: Computational and cognitive resources required to generate and process explanations

Evaluation of explanations ?

UUser-Centered Methods

sSimulated task experiments: *do explanations improve user performance on specific tasks?*

Effect on Trust : *Assessing if explanations appropriately increase or decrease user trust based on model capabilities*

Humans prefer simple explanations - causal structures etc., difficulty in capturing edge cases

Computational Evaluation Methods

Perturbation based changes

Identify top k features

Perturb the features (alter, delete, replace with random)

Plot prediction versus number of features perturbed

Usually the bigger the change following perturbation, the better the feature

Computational Evaluation Methods

Example-based explanation

Generation of an example to explain the prediction

Prototypes: Representative examples that illustrate typical cases

Counterfactuals: Examples showing how inputs could be minimally changed to get different outcomes

Influential instances: Training examples that has most influence

Boundary examples: Cases near the decision boundary that demonstrate model limitations

Computational Evaluation Methods

Example-based explanation - Evaluation metrics

PProximity (how close examples are to the original input)

DDiversity (variety of examples provided)

PPlausibility (whether examples seem realistic to users)

Saliency

Saliency methods highlight the input features or regions that most influence a model's prediction

GGradient-based methods: Calculate sensitivity of output with respect to input features

PPerturbation-based methods: Observe prediction changes when features are modified

AApplications:

IImage classification: Highlighting regions that influenced the classification
NNLP: Identifying influential words or phrases in text classification

XAI - approaches

XAI (Explainable AI)

Typically in AI systems we use data to give a recommendation, classification, prediction etc.

In XAI, we give the recommendation and an explanation and typically try to allow feedback.

XAI models:

Pre-modelling explainability

Interpretable models

Post model explainability

Pre-modelling explainability:

Data selection
Preparation transparency
Feature engineering (and documentation): why certain variables were selected
Design constraints documentation: Outlining constraints and considerations around
Success metrics definition: how the algorithm's performance will be measured beyond just technical accuracy

Explanation

Meaning behind a decision

Can be correct but complex (conjunction of many features)

Non-linear models – more difficult

Accuracy vs Explainability

Often as accuracy increases, explainability suffers

- linear models are relatively easy to explain
- NN and non-linear models - harder to explain

Usually a trade-off between the performance and explainability.

Much previous work has concentrated on improving performance and has largely ignored transparency

XAI – attempts to enable better model interpretability while maintaining performance

Accuracy vs Explainability

Some models are intrinsically explainable

Linear Regression

The effect of each feature is the weight of the feature times the feature value.

$$y = \beta_0 + \beta_1(x_1) + \ldots + \beta_p(x_p)$$

Accuracy vs Explainability

Intrinsic Models - Decision Trees

Tree based models split the data multiple times according to certain cutoff values in the features.

Decomposing the decision path into one component per feature. All the edges are connected by ‘AND’.

Can measure the importance of the feature by considering the information gain

Accuracy vs Explainability

Neural networks, connectionist, non-linear models

Much more difficult to generate explanations

Different approaches depending on what information you have access to

Neural networks:

https://playground.tensorflow.org/

Allows you to build NNS, change architecture, learning rules, data sets etc

Similarly, in reasoning systems, can generate explanations relatively easily.

Oftentimes, simple explanation concepts can be helpful - consider a complex MAS with learning - can be hard to explain dynamics; however, analysis of equilibria can give a reasonable explanation of likely outcomes

Basic learning approaches may give better understanding/explanations

If some function is learnable from a simple model, then use the simple model. Tends to lead to better explainability.

e.g. Could use a complex NN to learn a function; may be hard to derive an explanation.

However, we could perhaps learn the same function with a simple decision tree.

As we move to more complex models which are less interpretable, other approaches adopted

- feature importance
- dependence plots
- sensitivity analysis

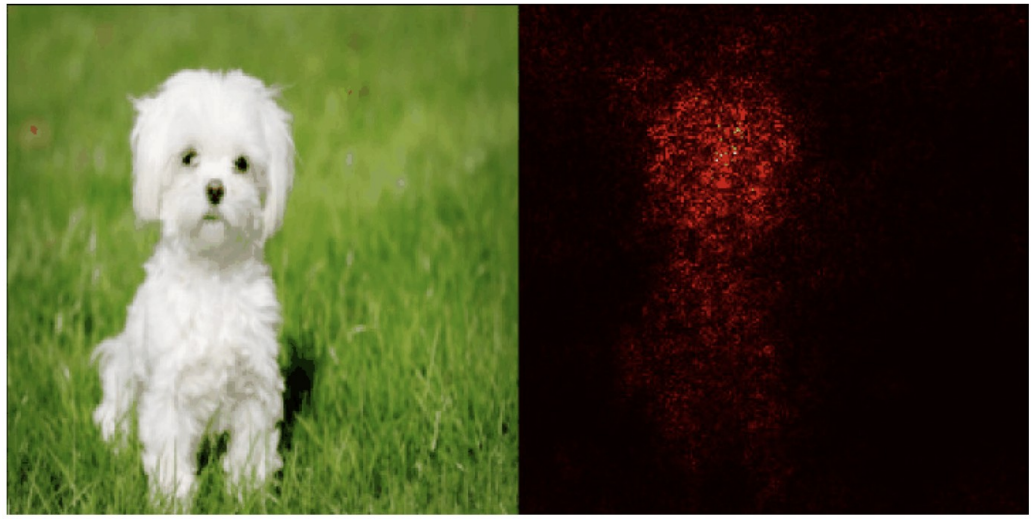
Explanations for Neural Networks

Can be difficult to generate

Neural networks can be extremely sensitive to perturbations.

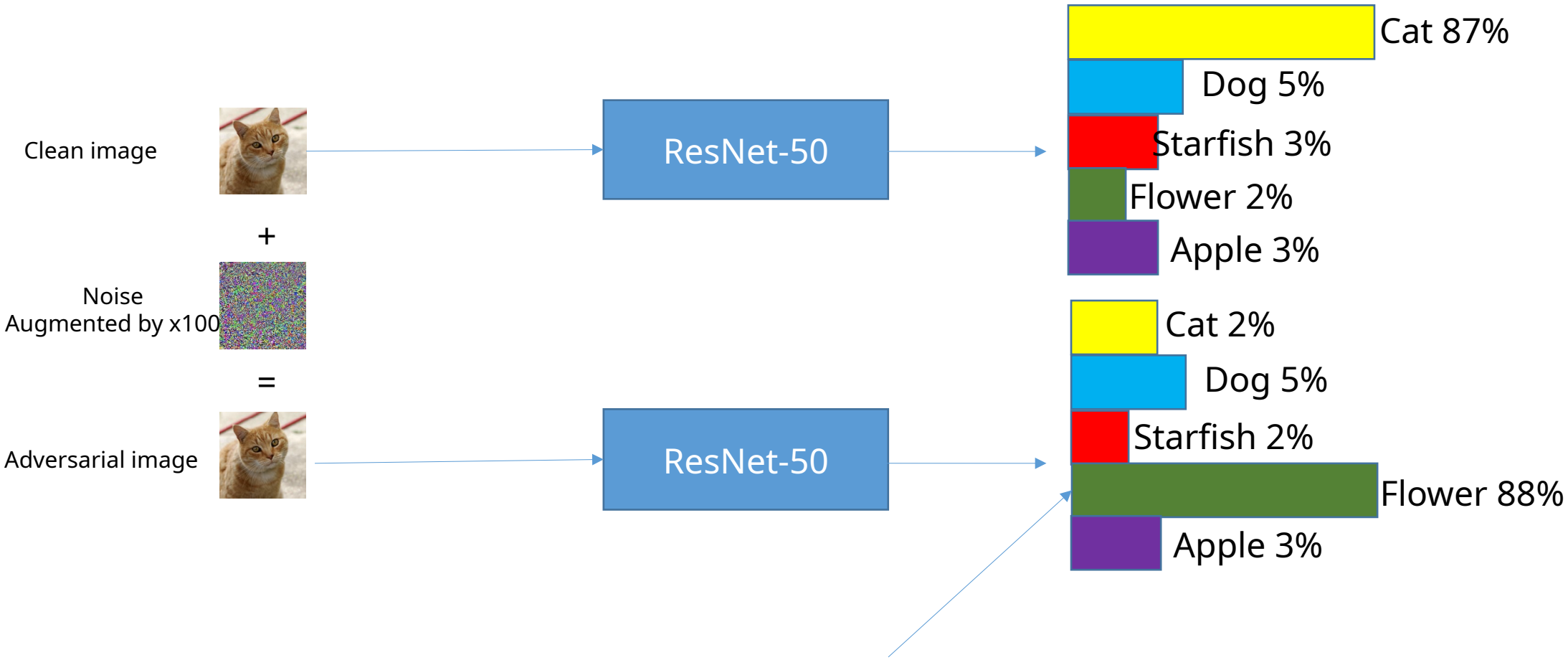
NNs susceptibility to adversarial attacks

Saliency - examples



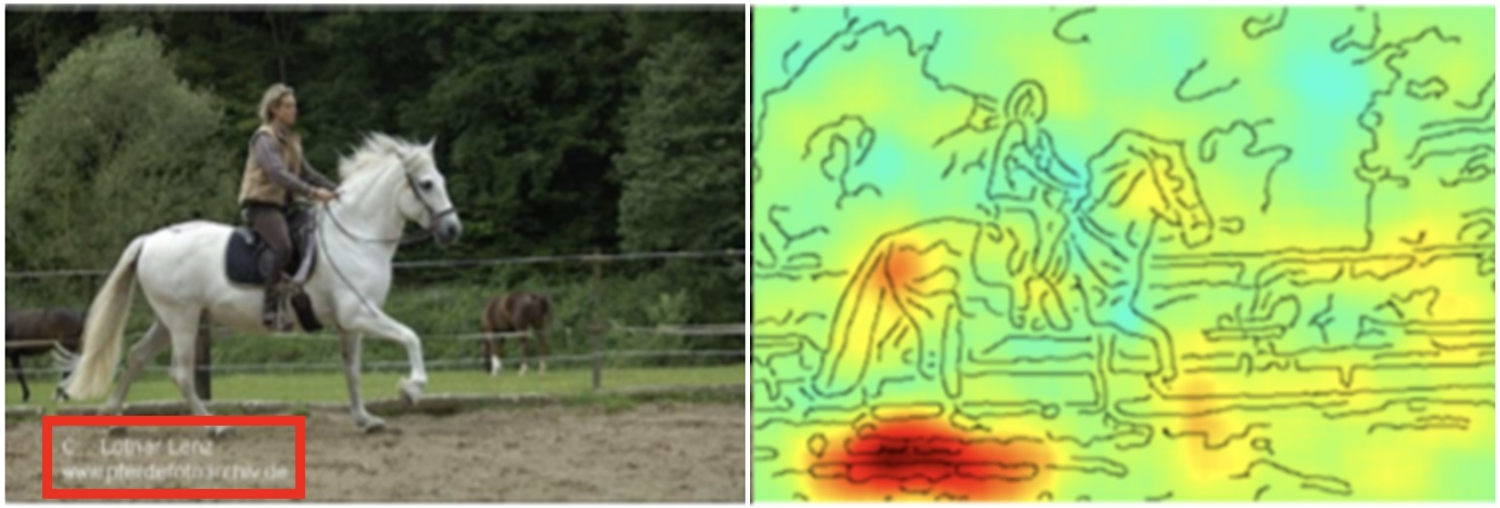
https://medium.datadriveninvestor.com/visualizing-neural-networks-using-saliency-maps-in-pytorch-289d8e244ab4
https://arxiv.org/pdf/1312.6034.pdf

Adversarial attack (an example)



Example of explanation showing the failing of a NN

NN learn to pay attentions to the captions



This slide is from: GCPR 2017 Tutorial — W. Samek & K.-R. Müller

ef: The image comes from Hungyi Lee's ML Lectures.

Explanations for Neural Networks

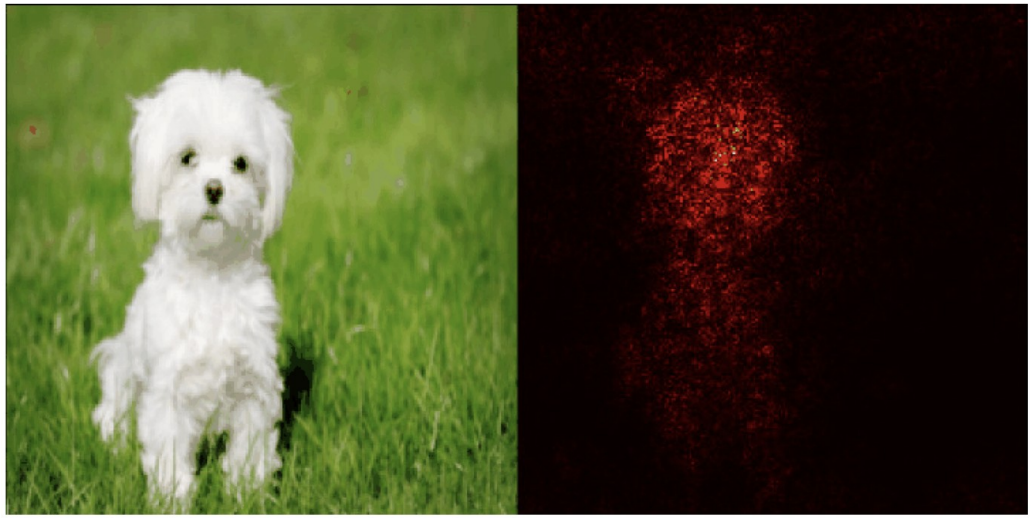
The predictions from NN must be aligned with humans and make sense to humans.

Approaches:

- Simplifying neural network
- Visualising
- Highlight aspects

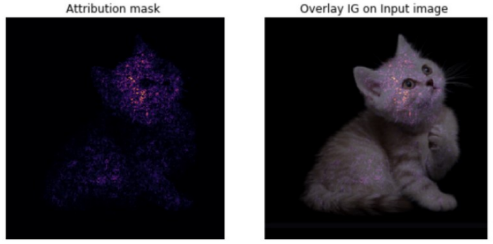
The motivation of gradient based explainable approach

- Large gradients for given inputs reflect the importance in activations.
- The area with large gradients reflect the interested area by models.



Integrated Gradients (2017)

- Integrated Gradients
  - This work will first create multiple versions of the input image
  - Typically interpolate from blank image to actual image
  - The created multiple versions of the input will be fed into the models for computing gradients
  - The computed gradients will be added together to form the final gradient map.
  - Gives insight into which pixels lead to decision



Integrated gradients

Shapley Values

Stems from work in cooperative game theory

Each pixel is viewed as a "player" in a cooperative "game" where the "payout" is the model's prediction. Each pixel's average marginal contribution across all possible combinations of pixels.

Knowledge Graphs in Explanation

Knowledge graph - collection of facts or relations
Typically stored as SPO triples (subject, predicate, object)

Many large databases available

- DBpedia - 4.6 million entities,
- YAGO3 - 17M
- Freebase - 40M
- .....

Some manually curated - by experts, or by collaborative events

Others automated - scraped from collections (prone to noise)

Coverage is an issue - many attribute values missing

Can be used in learning to:

i) Augment input features with semantic information from knowledge graph

ii) In NNs, can augment intermediate features with knowledge

Can be used to support explanations - incorporate new knowledge

Can also generate explanations for different audiences

Summary

Receiving much attention in recent years

Much studied for a long time

Easier to generate explanations in simpler learning models

Issues regarding evaluation of explanations

Much work in neural networks in recent years