Learning Discriminative Representations to Interpret Image Recognition Models

Thèse de Doctorat

Felipe Torres Figueroa

École Centrale de Marseille

Laboratoire d'Informatique et de Systèmes (LIS)

Marseille, September 23rd 2024

Table of Contents

Table of Contents

Motivation

Low Stakes

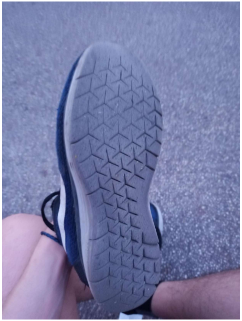
My go to exercise is running, **but...**

Motivation

Low Stakes

My go to exercise is running, **but...**

I think my running shoes
are getting *worn*

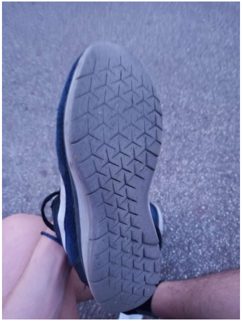


Motivation

Low Stakes

My go to exercise is running, **but...**

I think my running shoes are getting *worn*



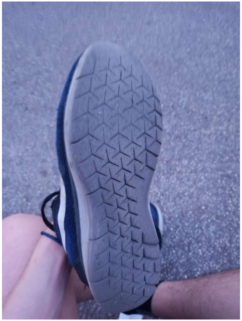
I want a replacement, *but* I know about machines, not shoes!

Motivation

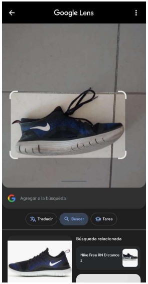
Low Stakes

My go to exercise is running, **but...**

I think my running shoes are getting *worn*



I want a replacement, *but* I know about machines, not shoes!



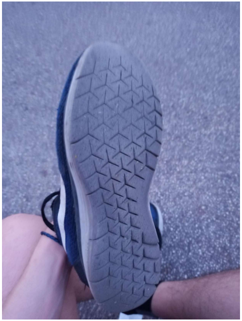
My phone can *identify* my current shoes

Motivation

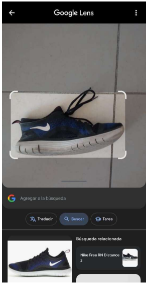
Low Stakes

My go to exercise is running, **but...**

I think my running shoes are getting *worn*



I want a replacement, *but* I know about machines, not shoes!



My phone can *identify* my current shoes

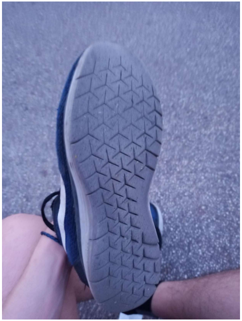
Nike Free RN Distance 2

Motivation

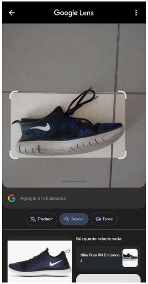
Low Stakes

My go to exercise is running, **but...**

I think my running shoes are getting *worn*



I want a replacement, *but* I know about machines, not shoes!



My phone can *identify* my current shoes

Nike Free RN Distance 2

How **could my phone identify that model?**

Motivation

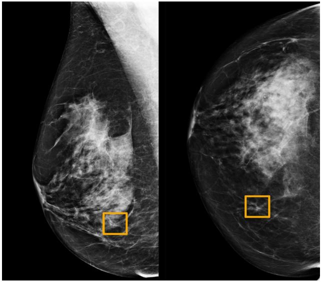
Raising the stakes

Now let's consider riskier situations:

Motivation

Raising the stakes

Now let's consider riskier situations:



Motivation

Raising the stakes

Now let's consider riskier situations:

Motivation

Raising the stakes

Now let's consider riskier situations:

Motivation

Straight to the point

- How do we **know how** a system works?

Motivation

Straight to the point

- How do we **know how** a system works?
- How do we **know how** safe a system is?

Motivation

Straight to the point

- How do we **know how** a system works?
- How do we **know how** safe a system is?
- If a system fails, **who** is accountable?

We must **understand** the behaviour of these models.

Step by step

Computation, Computer Vision and AI $\rightarrow$ Explainable AI $\rightarrow$ Thesis objectives

Computation, Computer Vision and AI

Computation

Alan Turing forefather of current computer science.

Better known as *Computer Science*.

Computation, Computer Vision and AI

Computation

Alan Turing forefather of current computer science.

Better known as *Computer Science*.

Study of:

- Algorithms.
- Data structures.
- Design of hardware and software.

Computation, Computer Vision and AI

Computer Vision

Replication of human vision capabilities.

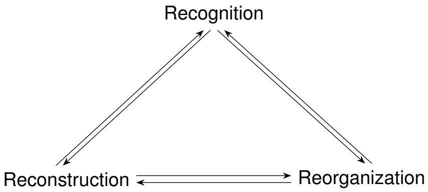
Computation, Computer Vision and AI

Computer Vision

Replication of human vision capabilities.

Three fundamental tasks[1]:

- Recognition.
- Reconstruction.
- Reorganization.



Computation, Computer Vision and AI

Artificial Intelligence

Systems capable of performing tasks requiring human intelligence [2].

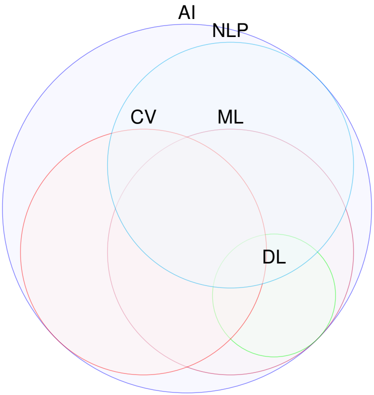
Computation, Computer Vision and AI

Artificial Intelligence

Systems capable of performing tasks requiring human intelligence [2].

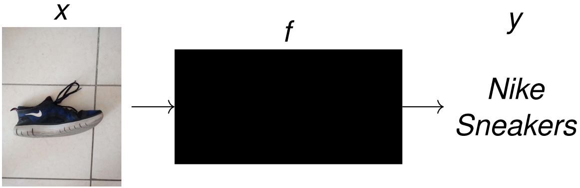
Subfields:

- Machine Learning *(ML)* & Deep Learning (*DL*).
- Computer Vision(*CV*).
- Natural Language Processing (*NLP*).
- Robotics.



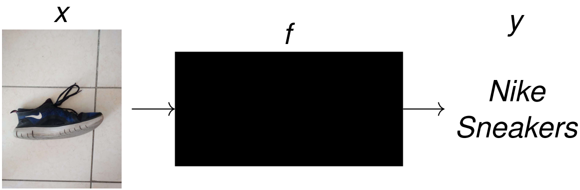
Explainable AI

We are interested in understanding models,
behaving like a black box model:



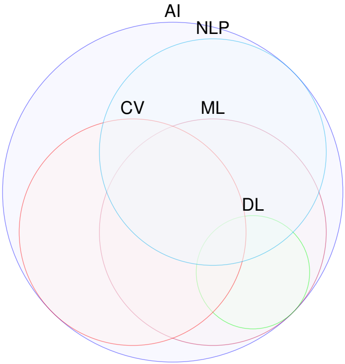
Explainable AI

We are interested in understanding models,
behaving like a black box model:

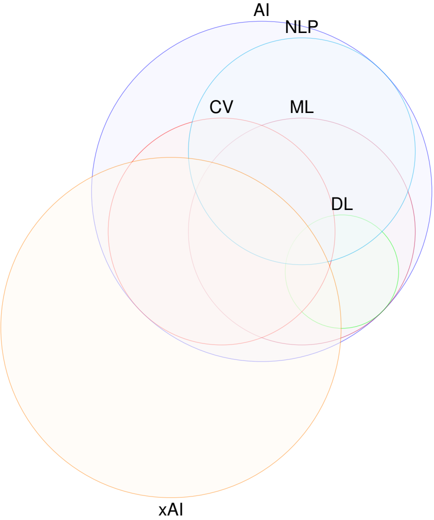


We want to *know why* $f(x) \to y$

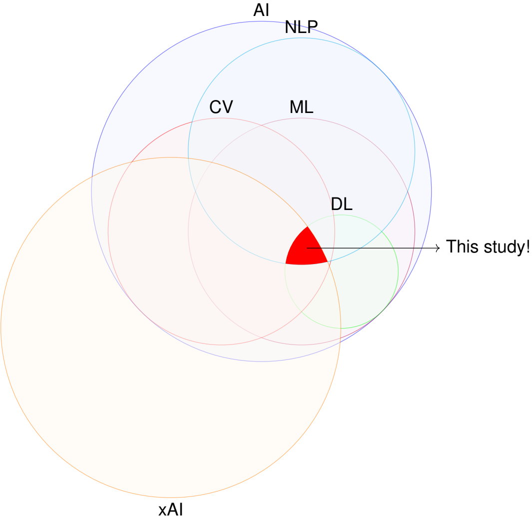
Fitting it all together



Fitting it all together



Fitting it all together



Thesis Objectives

Improvement of recognition and interpratable properties of model predictions.

Thesis Objectives

Improvement of recognition and interpratable properties of model predictions.

In particular:

- Development of low cost/complexity explainability approaches.
- Establishment of a fixed evaluation protocol.
- Differenciation of human based and machine explanations.

Table of Contents

Background

To familiarize with this work, we split it into three points:

Background

To familiarize with this work, we split it into three points:

Preliminaries

- Approaching Vision.
- David Marr's approach.
- CV currently.
- Desiredata of Interpretability Study.

Image Recognition Models

- Traditional Models.
- Convolutional Neural Networks (CNN).
- The Current Landscape.

Interpretability

- Transparency.
- Post-Hoc Interpretability.
  - Class Activation Methods.
- Evaluating Interpretability.

Preliminaries

Approaching Vision

Preliminaries

David Marr's approach

Addressing vision on three levels:

- Algorithmic.
- Implementation.
- Computational.
  - Three fundamental tasks. [1]

Preliminaries

David Marr’s approach

Addressing vision on three levels:

- Algorithmic.
- Implementation.
- Computational.
  - Three fundamental tasks. [1]

Computer Vision focuses on the last level.

Preliminaries

CV Currently

Preliminaries

Desiredata of Interpretability Study

We ask questions regarding *black box models*.

- How does it work?
- How safe is it?
- Who is accountable for it?
- Who benefits from it?
- Who uses it?.

Accepted AI systems within society must answer this.

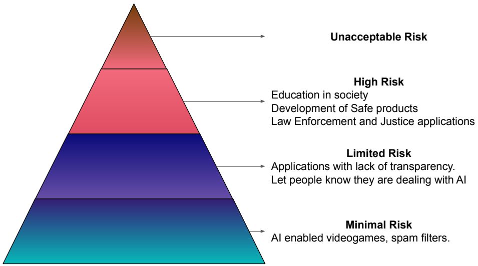
Preliminaries

Desiredata of Interpretability Study

We ask questions regarding *black box models*.

- How does it work?
- How safe is it?
- Who is accountable for it?
- Who benefits from it?
- Who uses it?.

Accepted AI systems within society must answer this.



Regulation planned with the European AI act[3]

Image Recognition Models

Classic Models

A two step process:

Image Recognition Models

Classic Models

A two step process:

Feature Extraction:
Extracting meaningful features from high noise/complexity $H \times W \times 3$ images.

Examples such as:

Image Recognition Models

Classic Models

A two step process:

Feature Extraction:
Extracting meaningful features from high noise/complexity $H \times W \times 3$ images.

Examples such as:

Classifier Training:
Relying on *Machine Learning* methodologies.

Image Recognition Models

Convolutional Neural Networks

Based on the **convolution operation**.
A representation $f \star g$ is computed for a feature map f and a kernel g.

First approach with *Neocognitron*[4]

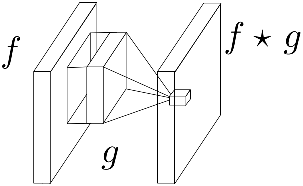
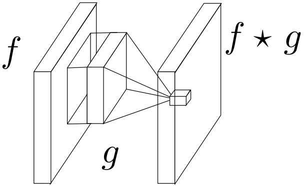


Image Recognition Models

Convolutional Neural Networks

Based on the **convolution operation**.
A representation $f \star g$ is computed for a feature map *f* and a kernel *g*.

First approach with *Neocognitron*[4]



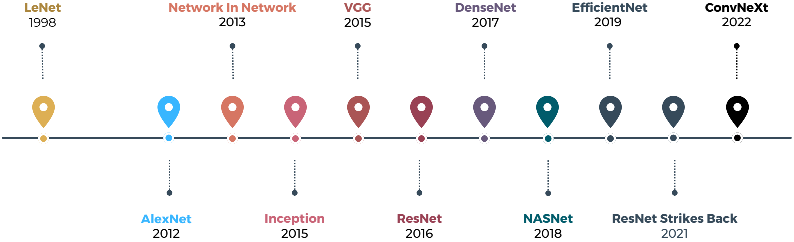


Image Recognition Models

Self Attention Architectures

Updates a representation, using the relevance of each element relative to others in an embedding.

An input is projected to three spaces (QKV), the projection weights (W_Q, W_K, W_V) control the relevance of each element.

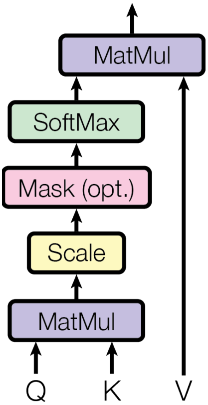
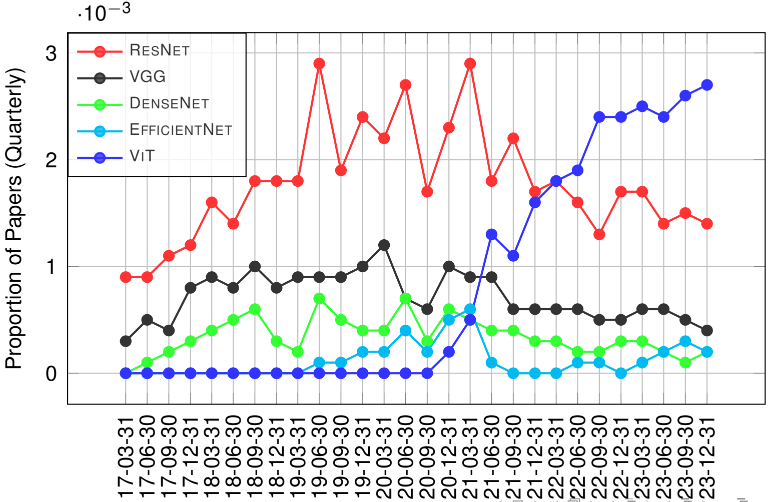


Image Recognition Models

The Current Landscape

Transformers had a strong impact on image recognition.



Interpretability

Transparency

Interpretability

Post-Hoc Interpretability

Interpretability

Class Activation Methods

Interpretability

Evaluating Interpretability

Table of Contents

Table of Contents

Table of Contents

Table of Contents

References I

- J. Malik, P. Arbeláez, J. Carreira, K. Fragkiadaki, R. Girshick, G. Gkioxari, S. Gupta, B. Hariharan, A. Kar, and S. Tulsiani, "The three r's of computer vision: Recognition, reconstruction and reorganization," *Pattern Recognition Letters*, vol. 72, pp. 4–14, 2016.
- J. McCarthy *et al.*, "What is artificial intelligence," 2007.
- T. Madiega, "Artificial intelligence act," *European Parliament: European Parliamentary Research Service*, 2021.
- K. Fukushima, "Cognitron: A self-organizing multilayered neural network," *Biological cybernetics*, vol. 20, no. 3-4, pp. 121–136, 1975.