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 2024

Table of Contents

Table of Contents

- Introduction

1. Background
2. Opti-CAM: Optimizing saliency maps for interpretability
3. CA-Stream: Attention-based pooling for interpretable image recognition
4. A learning paradigm for interpretable gradients

- References

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!

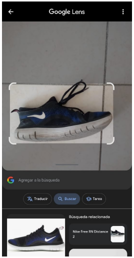
still, I know my phone can *identify* my current shoes

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!



still, I know my phone can *identify* my current shoes

and obtain a new pair of the shoes I like

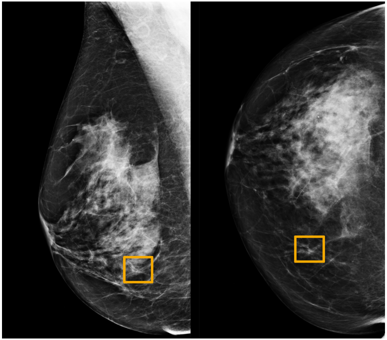
The *Nike Free RN Distance 2*

Motivation: Raising the stakes

Now let’s consider riskier situations:

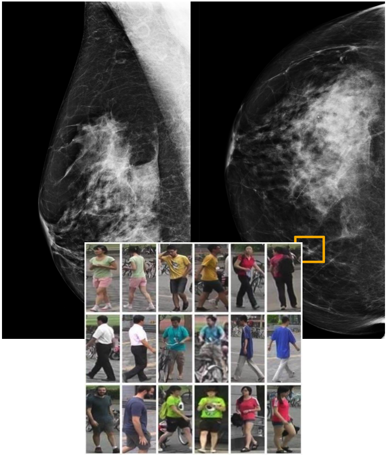
Motivation: Raising the stakes

Now let’s consider riskier situations:



Motivation: Raising the stakes

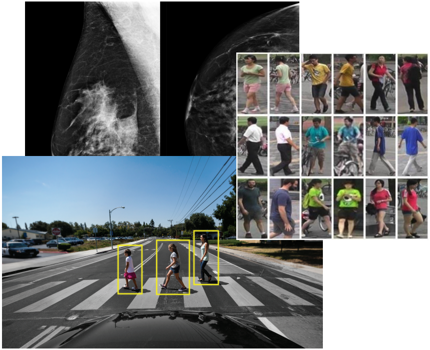
Now let’s consider riskier situations:



Motivation: Raising the stakes

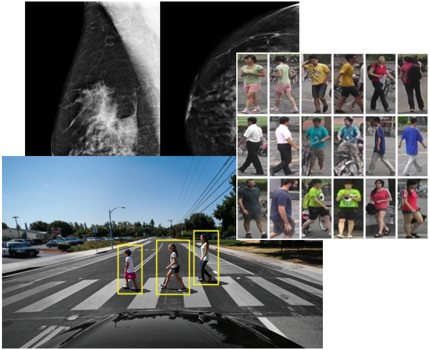
Now let’s consider riskier situations:

Motivation: Straight to the point



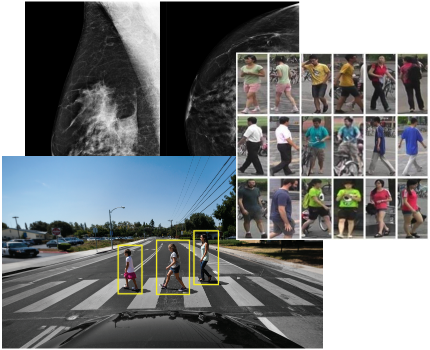
- How do we know **how** safe a system is?

Motivation: Straight to the point



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

Motivation: Straight to the point



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

Let’s slow for a bit
and go step by step:

Computation, Computer Vision and AI

[1]

Objectives

[1]

Table of Contents

Introduction

1 Background

2 Opti-CAM: Optimizing saliency maps for interpretability

3 CA-Stream: Attention-based pooling for interpretable image recognition

4 A learning paradigm for interpretable gradients

References

This is a text in second frame. For the sake of showing an example.

- Text visible on slide 1

Table of Contents

Introduction

1 Background

2 Opti-CAM: Optimizing saliency maps for interpretability

3 CA-Stream: Attention-based pooling for interpretable image recognition

4 A learning paradigm for interpretable gradients

References

Table of Contents

- Introduction

1 Background

2 Opti-CAM: Optimizing saliency maps for interpretability

3 CA-Stream: Attention-based pooling for interpretable image recognition

4 A learning paradigm for interpretable gradients

- References

Table of Contents

- Introduction
- 1 Background
- 2 Opti-CAM: Optimizing saliency maps for interpretability
- 3 CA-Stream: Attention-based pooling for interpretable image recognition
- 4 A learning paradigm for interpretable gradients
- References

Table of Contents

- Introduction
- 1 Background
- 2 Opti-CAM: Optimizing saliency maps for interpretability
- 3 CA-Stream: Attention-based pooling for interpretable image recognition
- 4 A learning paradigm for interpretable gradients
- References

References I

Z. C. Lipton, “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.” *Queue*, vol. 16, no. 3, pp. 31–57, 2018.