Thomas Fel

Boston, USA
+33 7 86 85 64 43 • tfel@g.harvard.edu • thomasfel.me

Summary

Hi, my name is Thomas, and I am a Research Fellow at the Kempner Institute, Harvard, specializing in **Explainable AI for vision**. My primary research goal is to leverage Explainable AI to better understand intelligence. I approach this challenge through an interdisciplinary lens that merges computational science, mathematics, and neuroscience principles. My current work focuses on enhancing the interpretability of large-scale vision models, improving their alignment with human understanding, and making their outputs more actionable. Previously, I was a Ph.D. student with Thomas Serre at ANITI & Brown University. During my Ph.D., I was also a core team member of DEEL (ANITI), where I contributed to advancing AI interpretability in various projects and had the opportunity to intern at Google and GoPro. I am also the main author and maintainer of the Xplique library, which facilitates the interpretability of AI models, and I recently released another open-source project, LENS, to visualize the main concepts learned by large vision models. This tool is based on three major works completed during my thesis.

Education & Experiences

Harvard University — **Boston, USA**
Kempner Research Fellow — *2024–Present*

Google — **San Francisco, USA**
Student Researcher (*Internship*) — *2023–2024*

ANITI & Brown University — **Toulouse, France**
Ph.D. in Computer Science & Mathematics — *2021–2024*
Explainability for Large Vision Models.

GoPro — **Paris, France**
Machine Learning Researcher (*Internship*) — *2019–2020*

CNES National Center for Space Studies — **Toulouse, France**
Machine Learning Researcher (*Internship*) — *2019*

National Institute of Science — **Toulouse, France**
Master of Applied Mathematics — *2019–2021*
Machine Learning major.

Université Jean-Francois Champollion — **Castres**
Master of Computer Science — *2015–2019*
Computer Science and Information Systems for Health.

Publications

2024

Understanding Visual Feature Reliance through the Lens of Complexity:
Thomas Fel[*], L. Béthune[*], A. Lampinen, T. Serre, K. Hermann
Conference on Neural Information Processing Systems (***NeurIPS***) *2024*

This work introduces a metric based on $\mathcal{V}$-information to quantify feature complexity in deep learning models. Analyzing 10,000 features from an ImageNet-trained model, our study finds a spectrum from simple to complex features, with simpler ones emerging early in training. Interestingly, simpler features are often more important for decision-making and tend to bypass the visual hierarchy via residual connections. Finally, we found that important features follow a sedimentation process, becoming accessible early and building the foundation for the model's learning.

On the Foundations of Shortcut Learning:
K. Hermann, H. Mobahi, **T. Fel**, M.C. Mozer
International Conference on Learning Representations (***ICLR***) *2024*
This article investigates the interplay between predictivity and availability in deep learning models' feature utilization, highlighting how models tend to favor more available but less predictive features. The study reveals that linear models exhibit relatively unbiased behavior, while deeper architectures with ReLU or Tanh units tend to exhibit shortcut bias, shedding light on the impact of architectural choices on feature selection.

Saliency Strikes Back: How Filtering Out High Frequencies Improves Explanations:
S. Muzellec[*], **T. Fel**[*], V. Boutin, L. Andéol, R. VanRullen, T. Serre
International Conference on Machine Learning (***ICML***) *2024*
We empirically observe differences in the power spectra of attribution maps, notably higher-frequency content in gradient-based methods. By mitigating the high-frequency noise through low-pass filtering, we improve explainability scores and demonstrate a potential resurgence of interest in computationally efficient gradient-based explainability methods.

Influenciæ: A Library for Tracing the Influence Back to the Data-Points:
A. Picard, L. Hervier, **T. Fel**, D. Vigouroux
World Conference on Explainable Artificial Intelligence (***W-XAI***) *2024*
We introduce Influenciæ, a library that provides tools to trace the influence of training data points on a model's predictions. This aids in understanding model behavior and diagnosing issues related to data quality and model training.

2023

CRAFT: Concept Recursive Activation Factorization:
T. Fel[*], A. Picard[*], L. Béthune[*], T. Boissin[*], D. Vigouroux, J. Colin, R. Cadène, T. Serre
Conference on Computer Vision and Pattern Recognition (***CVPR***) *2023*
We proposed an automated method to extract concepts from neural networks that explains a specific class, tackling the "What" challenge in explainability. We go beyond understanding where the model looks, delving into what it sees at the precise location of interest. Taking it a step further, we enhance this method by (1) effectively estimating the importance of discovered concepts, and (2) generating localized heatmaps, revealing concept locations within each image.

A Holistic Approach to Unifying Automatic Concept Extraction and Concept Importance Estimation:
T. Fel[*], V. Boutin[*], M. Moayeri, R. Cadène, L. Béthune, M. Chalvidal, T. Serre
Conference on Neural Information Processing Systems (***NeurIPS***) *2023* (*Spotlight*)
In this article, we demonstrate that all concept extraction methods can be viewed as dictionary learning methods. We leverage this common framework to develop a comprehensive framework for comparing and improving concept extraction methods. Furthermore, we extensively investigate the estimation of concept importance and show that it is possible to determine optimal importance estimation formulas in certain cases. We also highlight the significance of local concept importance in addressing a crucial question in Explainable Artificial Intelligence (XAI): identifying data points classified based on similar reasons.

Unlocking Feature Visualization for Deeper Networks with Magnitude Constrained Optimiza-

tion:
T. Fel[*], T. Boissin[*], V. Boutin[*], A. Picard[*], P. Novello[*], J. Colin, D. Linsley, T. Rousseau, R. Cadène, L. Gardes, T. Serre
Conference on Neural Information Processing Systems (***NeurIPS***) *2023*
We address the challenges of feature visualization in deep neural networks by introducing magnitude-constrained optimization techniques. This allows for the generation of more interpretable visualizations for deeper layers, enhancing our understanding of complex models.

On the Explainable Properties of 1-Lipschitz Neural Networks: An Optimal Transport Perspective:
M. Serrurier, F. Mamalet, **T. Fel**, L. Béthune, T. Boissin
Conference on Neural Information Processing Systems (***NeurIPS***) *2023*
We argue that, when learning a 1-Lipschitz neural network with the dual loss of an optimal transportation problem, the gradient of the model is both the direction of the transportation plan and the direction to the closest adversarial attack. The proposed networks were already known to be certifiably robust, and we prove that they are also tailored for explainability.

Performance-Optimized Deep Neural Networks are Evolving into Worse Models of Inferotemporal Visual Cortex:
D. Linsley, I.F. Rodriguez, **T. Fel**, M. Arcaro, S. Sharma, M. Livingstone, T. Serre
Conference on Neural Information Processing Systems (***NeurIPS***) *2023*
Over the past decade, deep neural networks (DNNs) have excelled at object recognition but, surprisingly, their accuracy in predicting neural responses in the inferotemporal cortex (IT) has not improved despite their increasing accuracy on ImageNet. To address this, we introduce the neural harmonizer, which aligns DNN representations with human perception and breaks the trade-off between ImageNet accuracy and neural prediction accuracy, offering a more promising approach for modeling the visual cortex.

Diffusion Models as Artists: Are We Closing the Gap Between Humans and Machines?:
V. Boutin, **T. Fel**, L. Singhal, R. Mukherji, A. Nagaraj, J. Colin, T. Serre
International Conference on Machine Learning (***ICML***) *2023* (*Spotlight*)
We investigate AI's progress in creating lifelike drawings, using the 'diversity vs. recognizability' framework. One-shot diffusion models have narrowed the gap between humans and machines, but improving their guidance falls short of replicating human creativity. Our study reveals distinctions in visual strategies, emphasizing a remaining gap between human and machine drawing abilities.

COCKATIEL: Continuous Concept Ranked Attribution with Interpretable Elements for Explaining Neural Net Classifiers on NLP Tasks:
F. Jourdan, A. Picard, **T. Fel**, L. Risser, J.M. Loubes, N. Asher
Annual Meeting of the Association for Computational Linguistics (***ACL***) *2023*
This article focuses on the challenges of interpretability and explainability in Transformer architectures for NLP. We introduce a model-agnostic XAI technique that generates meaningful explanations from NLP models while preserving accuracy and avoiding retraining.

Confident Object Detection via Conformal Prediction and Conformal Risk Control: Application to Railway Signaling:
L. Andéol, **T. Fel**, F. De Grancey, L. Mossina
Symposium on Conformal and Probabilistic Prediction with Applications (***COPA***) *2023*
We showcase how the conformal prediction framework can create reliable railway signal detectors with trustworthy uncertainty estimates, using a unique dataset and state-of-the-art object detectors. Our study explores various conformal approaches and introduces a novel method based on conformal risk control.

2022.....

Harmonizing the Object Recognition Strategies of Deep Neural Networks with Humans:

T. Fel[*], I. Felipe[*], D. Linsley[*], T. Serre
Conference on Neural Information Processing Systems (***NeurIPS***) *2022*
Using an interesting property of explainability methods—the fact that they are themselves differentiable—we use the Click-Me dataset containing more than 300,000 human saliency map images to retrain a large number of models from the literature on ImageNet, and force them to 'look' at the same spots as humans. Furthermore, we highlight an interesting trade-off that seems to be emerging: the latest state-of-the-art models are no better (or even worse) than more modest models for predicting human saliency.

What I Cannot Predict, I Do Not Understand: A Human-Centered Evaluation Framework for Explainability Methods:
T. Fel[*], J. Colin[*], R. Cadène, T. Serre
Conference on Neural Information Processing Systems (***NeurIPS***) *2022*
We challenge current explainability metrics by asking whether better scoring implies better human understanding of the model. Our results show that the theoretical metrics used to score explainability methods poorly reflect the practical utility of individual attribution methods in real-world scenarios.

Don't Lie to Me! Robust and Efficient Explainability with Verified Perturbation Analysis:
T. Fel[*], M. Ducoffe[*], D. Vigouroux, R. Cadène, M. Capelle, C. Nicodème, T. Serre
Conference on Computer Vision and Pattern Recognition (***CVPR***) *2023*
To answer critical safety needs, we propose an explainability method with formal guarantees, showing that it obtains state-of-the-art results on the usual benchmarks and proposing a way to make it scalable.

Making Sense of Dependence: Efficient Black-box Explanations Using Dependence Measure:
P. Novello, **T. Fel**, D. Vigouroux
Conference on Neural Information Processing Systems (***NeurIPS***) *2022*
We improve the efficiency of black-box methods using the Hilbert-Schmidt Independence Criterion (HSIC). HSIC measures the dependence between regions of an input image and the output of a model based on kernel embeddings of distributions. Our experiments show that HSIC is up to 8 times faster than the previous best black-box attribution methods while being as faithful.

Xplique: A Deep Learning Explainability Toolbox:
T. Fel[*], L. Hervier[*], D. Vigouroux, A. Poche, J. Plakoo, R. Cadène, M. Chalvidal, J. Colin, T. Boissin, L. Béthune, A. Picard, C. Nicodème, L. Gardes, G. Flandin, T. Serre
Workshop on Explainable Artificial Intelligence for Computer Vision (***XAI4CV***), *CVPR 2022*
For more than a year, I worked alongside my thesis to implement more than fifty explainability papers. Xplique is the result of this work; it is a library that I develop and maintain. The library is composed of several modules: (1) the Attributions Methods module, (2) the Feature Visualization module, (3) the Concepts module, and (4) the Metrics module.

2021

Look at the Variance! Efficient Black-box Explanations with Sobol-based Sensitivity Analysis:
T. Fel[*], R. Cadène[*], M. Chalvidal, M. Cord, D. Vigouroux, T. Serre
Conference on Neural Information Processing Systems (***NeurIPS***) *2021*
We present a new attribution method obtaining state-of-the-art results while drastically reducing the computing time compared to current methods. Theoretically grounded in sensitivity analysis, we adapt Sobol' indices to explain Deep Neural Networks.

How Good is Your Explanation? Algorithmic Stability Measures to Assess the Quality of Explanations for Deep Neural Networks:
T. Fel[*], D. Vigouroux, R. Cadène, T. Serre
IEEE/CVF Winter Conference on Applications of Computer Vision (***WACV***) *2022*
We derive two novel metrics for Deep Learning Explainability methods grounded in Algorithmic Stability:

Representativity and Consistency. We then use these metrics to search for models with better explanations and find that 1-Lipschitz networks can be a good lead.

Open Source Projects

LENS: An open-source visualization tool. The purpose of this tool is to provide a quick overview of the main concepts (dictionary of features) employed by a large vision model for each of the 1,000 ImageNet classes, along with an associated importance score for these concepts.
https://serre-lab.github.io/Lens/

Xplique: An open-source Explainability toolbox implementing more than 50 papers of explainability, with proper documentation, tutorials, and notebooks.
https://github.com/deel-ai/xplique

CRAFT: An open-source repository to reproduce our work on concept-based explainability, in TensorFlow and PyTorch, with tutorials.
https://github.com/deel-ai/Craft

Harmonization: An open-source zoo of harmonized models trained, as well as notebooks and tutorials to evaluate other models, in TensorFlow and PyTorch.
https://github.com/serre-lab/Harmonization

Sobol: An open-source version of the Sobol attribution method, in TensorFlow and PyTorch.
https://github.com/fel-thomas/Sobol-Attribution-Method

Numkdoc: An open-source parser of MkDocs for NumPy style documentation, now used to build the documentation of Xplique and other open-source projects.

Invited Talks

- September 2024 – Invited talk at The Enigma Project, Stanford University
- July 2024 – Tutorial on Explainability at PFIA (French Association for Artificial Intelligence)
- June 2024 – Invited talk for Sinclair team at EDF
- June 2024 – AI for Vision at Davidson Paris, with Timothée Darcet
- May 2024 – Invited talk at the United Nations AI Conference in Geneva
- April 2024 – Invited talk at Centrale Paris
- March 2024 – Invited talk at CENTAI
- March 2024 – Invited talk at EPFL
- February 2024 – Invited talk at Centrale Marseille
- September 2023 – Invited talk at Max Planck Institute
- August 2023 – Invited talk at Princeton University
- May 2023 – Invited talk at Harvard University (Viégas & Wattenberg lab)
- March 2023 – Invited talk at Airbus Defense & Space
- February 2023 – Invited talk at École Normale Supérieure (ENS-Paris Saclay)
- November 2022 – Invited talk at Brown DSCOV
- February 2022 – Presentation of the Sobol paper at the Mathematical Institute of Toulouse

Professional Service

Reviewer

- NeurIPS 2023
- CVPR 2023 (Outstanding reviewer, top 3% of reviewers)
- CVPR 2023 XAI4CV Workshop
- NeurIPS 2022
- CVPR 2022 XAI4CV Workshop
- WACV 2022
- WACV 2021

Invited Talk Organizer

- Matthew Kowal (York University) at Brown University
- Devis Tuia (EPFL) at ANITI (DEEL)
- Timothée Darcet (Meta) at Brown University
- Mazda Moayeri (University of Maryland) at Brown University
- Filip Radenovic (Meta) at Brown University
- Arthur Douillard (DeepMind) at Brown University
- Peter Hase (University of North Carolina) at Brown University

Mentoring

PhD Student: Fenil Doshi (Harvard)
PhD Student: Cindy Luo (Harvard)
PhD Student: Chloe Su (Harvard)
Master Student: Julien Colin (ANITI), now Ph.D. student at ELLIS
Master Student: Pinyuan Feng (Brown University)

Languages

French: Native proficiency
English: Fluent
Spanish: Fluent
Italian: Basic proficiency

Honors & Awards

- Officially defended my thesis "Sparks of Explainability for Vision Models" (July 2024)
- Best thesis (both public & jury): "My thesis in 180 seconds" at SNCF (December 2022)
- Winner of "Nuit de l'info" for the realization of a web single-page application in one night (more than 1000 participants), 2017
- Winner of E-Health Eurocampus, UK, 2018
- Winner of "Trail du pastel", young category, a 27 km run in Toulouse, France, 2018
- Winner of regional Eloquence competition, 2017