Understanding and Characterizing the Geometry of Representations Across Layers, Models, and Modalities

Quentin Bouniot

Scientific Context

Foundation Models (FMs) such as large vision models (e.g. DinoV2 (Oquab et al. 2024)), language models (e.g., GPT (Brown et al. 2020)), and multimodal architectures (e.g. CLIP (Radford et al. 2021)) have achieved remarkable performance across a wide range of tasks. However, their increasing scale and complexity have made them largely opaque, raising critical concerns regarding interpretability, trustworthiness, and reliability. These concerns are now amplified by regulatory frameworks such as the EU AI Act. Understanding how these models process information internally, beyond simply evaluating predictions and generations, has become a central challenge in modern AI research.

Recent work in mechanistic interpretability has begun to reveal intriguing structural properties within FMs. Studies have shown that representations evolve through depth in structured ways: emergent linearity or non-linearity patterns appear across layers (Bouniot et al. 2025) (Sun et al. 2025), clusters of functionally similar layers can be identified (Chen et al. 2025), and progressively abstract concepts form along the network's depth (Kim et al. 2025). Other lines of research have uncovered internal sub-networks that behave as causal circuits (Elhage et al. 2021). These findings suggest that the internal structure of FMs is far from arbitrary and carries meaningful, exploitable information.

Yet, despite these promising results, the field remains fragmented. Existing analyses are often ad hoc, tied to specific architectures or tasks, and lack a unified theoretical grounding. There is currently no systematic framework for characterizing and comparing the internal representations of FMs across their layers, across different pretrained models, or across data modalities (text, images, graphs, time series). Furthermore, the relationship between measurable internal properties (e.g., linearity, intrinsic dimension, effective rank) and downstream behaviors (e.g., transferability, robustness to distribution shifts, interpretability) remains poorly understood.

Motivation

This PhD project is motivated by the need for a principled, systematic, and theoretically grounded approach to analyzing the internal structure of Foundation Models. Bridging this gap is essential not only for advancing our fundamental understanding of deep learning, but also for enabling practical downstream applications: identifying optimal layers for fine-tuning or alignment, guiding the design of more interpretable or robust models, and informing efficient adaptation strategies for new domains and modalities.

Objectives

The objective of this thesis is to develop a comprehensive framework for the mechanistic analysis of the internal representations of Foundation Models. The research will be organized around three complementary

axes:

- Defining theoretically grounded measures of internal structure. The student will develop and formalize tools to quantify both functional properties (e.g., degree of linearity, information compression) and geometric properties (e.g., intrinsic dimension, rank, manifold curvature) of representations at each layer. These measures should be principled, comparable across layers and models, and grounded in mathematical frameworks such as optimal transport, information geometry, or representation topology.
- Characterizing how representations and concepts evolve through depth. Building on these tools, the student will design methods for layer-wise concept extraction and representation analysis, aiming to understand how the representation space transforms across the depth of a model. This includes studying how abstract concepts emerge, how redundancy or specialization develops across layers, and whether universal patterns exist across architectures and modalities.
- Linking internal structure to downstream behaviors. The student will investigate both correlational and causal relationships between the identified internal properties and key downstream phenomena, including transferability, sensitivity to distribution shifts, alignment quality, and interpretability. This axis aims to turn mechanistic observations into actionable insights—for instance, predicting which layers are most suitable for intervention, adaptation, or explanation extraction.

The analysis will be conducted across multiple data modalities (vision, language, and potentially graphs or time series) and over a diverse panel of models, with the ambition of identifying both modality-specific and universal structural patterns.

References

Bouniot, Quentin, Ievgen Redko, Anton Mallasto, Charlotte Laclau, Oliver Struckmeier, Karol Arndt, Markus Heinonen, Ville Kyrki, and Samuel Kaski. 2025. "From Alexnet to Transformers: Measuring the Non-Linearity of Deep Neural Networks with Affine Optimal Transport." In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." *NeurIPS*.

Chen, Haoran, Junyan Lin, Xinghao Chen, Yue Fan, Jianfeng Dong, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. 2025. "Multimodal Language Models See Better When They Look Shallower." In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 6688–6706.

Elhage, Nelson, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, et al. 2021. "A Mathematical Framework for Transformer Circuits." *Transformer Circuits Thread*.

Kim, Jinyeong, Junhyeok Kim, Yumin Shim, Joohyeok Kim, Sunyoung Jung, and Seong Jae Hwang. 2025. "Interpreting Vision Transformers via Residual Replacement Model." *NeurIPS*.

Oquab, Maxime, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, et al. 2024. "DINOv2: Learning Robust Visual Features Without Supervision." *TMLR*.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. 2021. "Learning Transferable Visual Models from Natural Language Supervision." In *ICML*.

Sun, Wenfang, Xinyuan Song, Pengxiang Li, Lu Yin, Yefeng Zheng, and Shiwei Liu. 2025. "The Curse of Depth in Large Language Models." *NeurIPS*.