Universidade de São Paulo
Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto
Departamento de Computação e Matemática

Project Proposal: Comparing the Complexity Level of Neural Networks with their Inference Capability

Author: Lucas Miranda Mendonça Rezende
Supervisor: Ph.D. Luiz Otavio Murta Junior

August 3rd, 2025

Abstract

This research proposal investigates the relationship between neural network complexity measures and their inference capabilities. With the rapid advancement of Large Language Models (LLMs) and the increasing computational costs associated with scaling, understanding the complexity-performance relationship becomes crucial for efficient model development and deployment. This study aims to validate the hypothesis that model complexity and inference capability are meaningfully related, establish quantitative relationships between these measures, and explore their implications for model optimization and design. The research will analyze publicly available models using established complexity metrics and benchmark performance data, employing statistical methods to characterize the relationships discovered. Expected outcomes include theoretical insights into model behavior, practical guidelines for model selection and optimization, and foundations for future research in complexity-aware machine learning approaches.

Contents

1 Introduction

Since the creation of Transformers and the Large Language Models (LLMs) based on this technique (VASWANI et al., 2017), there has been a veritable gold rush within the machine learning world. The exponential growth in AI investments and research has led to increasingly sophisticated models with remarkable capabilities (BROWN et al., 2020; BUBECK et al., 2023).

However, over time, the need for more and more data to train increasingly larger models has become an impediment to the progress of these models' inference performance due to cost considerations (KAPLAN et al., 2020; HOFFMANN et al., 2022). The scaling laws for neural language models demonstrate that performance improvements require exponentially increasing computational resources and training data, making further advancement increasingly expensive and resource-intensive.

This creates the need for new advances that do not depend on adding more data but are instead improvements to the method itself, similar to what occurred with the creation of transformers (VASWANI et al., 2017). New methods to be created are generally based on observations of measures and relationships in the model weights.

One such measure was investigated in recent research that explores the relationship between the well-known measure of model complexity and inference capability (MURTA, 2025). Understanding this relationship could provide insights into model efficiency and performance optimization without requiring massive increases in training data or computational resources.

1.1 Objective

Work thesis: There exists a relationship between model complexity and its inference capability (MURTA, 2025).

The primary objectives of this work are to:

- Validate the hypothesis that model complexity and inference capability are related
- Measure and define the specific relationship that exists between these two measures, investigating whether complexity measures can serve as predictors of model performance across different architectures and tasks

2 Materials and Methods

2.1 Model Complexity

The complexity measure used in this study is based on the López-Ruiz-Mancini-Calbet (LMC) statistical complexity measure (LOPEZ-RUIZ; MANCINI; CALBET, 2010; MARTIN; PLASTINO; ROSSO, 2003), following recent research that demonstrates the relationship between complexity and inference performance (MURTA, 2025). For a neural network with weight parameters $\theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$, we define the complexity measure as follows:

Statistical Complexity of Model Weights: The LMC complexity $C_{LMC}(\theta)$ of a model's weight configuration is defined as:

$$C_{LMC}(\theta) = H(\theta) \times D(\theta) \tag{1}$$

where $H(\theta)$ is the normalized Shannon entropy and $D(\theta)$ is the disequilibrium measure. The normalized Shannon entropy is computed as:

$$H(\theta) = -\frac{1}{\log_2 n} \sum_{i=1}^{n} p_i \log_2 p_i \tag{2}$$

and the disequilibrium is defined as:

$$D(\theta) = \sum_{i=1}^{n} \left(p_i - \frac{1}{n} \right)^2 \tag{3}$$

where p_i represents the normalized probability distribution of weight bins obtained by discretizing the weight space, and n is the number of bins. This measure captures both the information content (entropy) and the degree of structure (disequilibrium) in the weight distribution.

The computation will be implemented using tools provided by frameworks like PyTorch (PASZKE et al., 2019). The chosen discretization approach and bin selection strategy will be explained in the final presentation.

2.2 Inference Capability

Inference capability will be assessed using standardized benchmarks and evaluation metrics commonly used in the machine learning community (LIANG et al., 2022; WANG et al., 2019). This requires careful investigation of comparison methodologies since benchmarks often become obsolete quickly due to rapid model improvements (RAJI et al., 2021).

2.3 Methodology

The research will follow the following step-by-step approach:

Step 1 - Model Selection

- Select publicly available open-source models with accessible weights. Models available through platforms like HuggingFace

Step 2 - Literature Review and Benchmark Data Collection

- Search for existing benchmark results for the selected models
- Conduct literature review on model evaluation methodologies
- Determine the most appropriate inference measures based on data availability

Step 3 - Complexity Measure Extraction

- Utilize tools such as PyTorch (PASZKE et al., 2019) to implement the calculation of the statistical complexity measure from model weights

Step 4 - Statistical Analysis and Relationship Modeling

- Structure results of the relationship between complexity and model performance
- Include measures of statistical significance (correlation analysis, regression modeling, confidence intervals...)

3 Planning

Steps 1 to 4 correspond to the steps shown in 2.3 Methodology.

Month	**Step 1**	**Step 2**	**Step 3**	**Step 4**
August	⋆	⋆		
September		⋆		
October		⋆	⋆	
November			⋆	⋆
December				⋆

Table 1: Project execution timeline.

4 Expected Results

The research aims to validate the existence of a meaningful relationship between algorithmic complexity of neural network weights and their inference performance (MURTA, 2025), as well as find the mathematical relation between those measures.

4.1 Primary Outcomes

- **Quantitative Relationship**: Mathematical characterization of the complexity-performance relationship with statistical significance measures

4.2 Expected Impact

- **Practical**: The results may enable new optimization approaches that consider both training error and model complexity, potentially leading to more efficient training methods (CHOLLET, 2019).
- **Theoretical**: Better understanding of how complexity relates to generalization in neural networks, contributing to the broader understanding of emergent capabilities in large language models (WEI et al., 2022).

References

BROWN, T. et al. Language models are few-shot learners. *Advances in neural information processing systems*, v. 33, p. 1877–1893, 2020.

BUBECK, S. et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

CHOLLET, F. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

HOFFMANN, J. et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

KAPLAN, J. et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

LIANG, P. et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

LOPEZ-RUIZ, R.; MANCINI, H. L.; CALBET, X. A statistical measure of complexity. *arXiv preprint arXiv:1009.1498v1*, 2010.

MARTIN, M.; PLASTINO, A.; ROSSO, O. Statistical complexity and disequilibrium. *Physics Letters A*, Elsevier, v. 311, n. 2-3, p. 126–132, 2003.

MURTA, L. O. Not published yet. 2025.

PASZKE, A. et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, v. 32, 2019.

RAJI, I. D. et al. Ai and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*, 2021.

VASWANI, A. et al. Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017.

WANG, A. et al. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*, 2019.

WEI, J. et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.