Analyzing Internal Sensitivity in Neural Network Representations via Jacobian Norms and Perturbation Analysis

Aninda Kumar Sharma(✉)

School of Computer and Mathematical Sciences, University of Adelaide, Adelaide, SA 5005, Australia
anindakumar.sharma@student.adelaide.edu.au

Abstract. Deep neural networks are widely used in computational and machine learning applications. However, deep neural networks with similar predictive power may have vastly different internal representation responses to perturbations. This is important in the computational analysis of neural networks because predictive power and input-output smoothness are not sufficient to describe the internal organization and output response in deep neural networks. This paper proposes a geometric concept called opacity to describe this intrinsic sensitivity in deep neural networks. Opacity is measured using the Jacobian norms in individual layers and the internal perturbation radius. This is a quantitative method to describe the influence of internal representations in deep neural networks. The paper also shows that opacity is not reducible to global input-output sensitivity. Architectures with comparable input sensitivity can differ substantially in internal perturbation radius. This suggests that sensitivity in deep neural networks is an intrinsic property. Because opacity is computed from standard forward passes without adversarial perturbations or auxiliary probes, it enables direct cross-architecture comparison under matched accuracy conditions. This paper also evaluates multilayer perceptrons, convolutional networks, and vision transformers using the MNIST and CIFAR-10 datasets. Our results show that architectures with comparable accuracy can differ by up to an order of magnitude in internal perturbation radius, highlighting the role of architectural inductive bias in shaping representational geometry. The findings suggest that opacity provides a quantitative and computationally grounded insight in analysing internal bottlenecks and stability properties in deep neural networks.

Keywords: Neural network · Internal sensitivity · Representation learning · Jacobian norms · Perturbation analysis

1 Introduction

Deep neural and transformer-based models are now widely used across computational machine learning applications, including natural language processing,

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16767, pp. 79–95, 2027.
https://doi.org/10.1007/978-3-032-30491-9_6

visual recognition, and cyber-security tasks [1–3]. This broad use of neural models motivates the need to study not only their predictive performance but also the internal structure of their learned representations. Computational analysis of neural networks typically emphasize predictive performance, input–output smoothness, or robustness under external perturbations. However, these views are limited as these do not show how changes relate to representations that are in between the final representation and the corresponding input. When models with similar predictive accuracy exhibit different internal stability regimes, a quantitative measure of hidden-state sensitivity can help identify architectural differences that are not visible from output performance alone. Internal layerwise sensitivity analysis is therefore a natural extension of computational approaches that relate representational structure to measurable differences in perturbation behaviour across neural architectures.

The Predictive accuracy is often the primary basis for comparing neural networks, but it does not directly characterize the internal representational structure of a model. Networks with similar predictive accuracy may exhibit substantially different internal stability regimes, suggesting that performance alone is insufficient for understanding how learned representations behave. Although accuracy and input-output smoothness have been widely studied, the geometry of hidden representations remains comparatively under-characterized. In particular, it is still unclear whether the internal sensitivity of learned representations is primarily governed by Lipschitz behavior, training dynamics, architectural structure, or their interaction.

This paper provides a structural concept we refer to as opacity, which refers to how difficult it is to induce changes in output from perturbations in the hidden representations or latent variables. More generally speaking, a representation is said to be opaque if perturbations of large size must be applied in hidden space to induce a given change in output size. The concept of opacity differs from traditional smoothness in that opacity is defined with respect to hidden (internal, intermediate) layers of the network rather than input (external) layers of the network. Thus, it captures more of a geometric property of the representation that is learned, rather than the learned function as a whole. In essence, from the perspective of measuring uncertainty, opacity reflects how stable a learned representation is at an internal level and serves as a complement to the input-output level measures of robustness that currently exist.

Currently, most analyses of the stability of neural networks emphasize global sensitivity to input perturbations (robustness) or Lipschitz bounds. The previous methods that were applied only consider the network from a single mapping point of view and therefore do not have the ability to distinguish between neural network architectures which exhibit similar input behaviour but are divergent internally. Hence, input based metrics cannot solely characterise the representational geometry of a neural network, and therefore additional structural metrics are required to quantify the organisation of information in a neural network across its depth.

To formalize the concept of opacity as dimensionality of neuron representation (in Sect. 3), we combine Jacobian norms and minimum internal perturbation radii to develop a layerwise sensitivity-radius functional for quantifying the opacity of each of a network's layers. We find that opacity is not contingent upon the predictive accuracy of output, nor is it solely based on the smoothness between an input and output of a network.

In Sect. 4, we develop a method for estimating the opacity of a network using layerwise perturbation search and randomized directional Jacobian projections. Furthermore, our estimators for opacity are architecture agnostic, and therefore can be applied uniformly across multilayer perceptrons, convolutional networks and transformers, without modification.

In Sect. 5, we provide empirical results on MNIST and CIFAR-10, where we find that at least some neural networks that exhibit similar predictive accuracy exhibit different internal perturbation radii, with transformers having significantly larger stability radii than multilayer perceptron baselines, and CNNs exhibiting intermediate stability radii.

All of the above empirical results demonstrate that the analysis of opacity is an under-characterised geometrical structure of a neural network's representation that is orthogonal to a neural network's predictive accuracy and input Lipschitz behaviour. Thus, our analysis of opacity as a dimension will add value to the analysis of neural networks.

2 Background Study

The stability and structure of neural representations have been studied from several theoretical and empirical perspectives, though typically through input-output analyses rather than internal geometric characterizations. A large body of work on adversarial robustness demonstrates that small input perturbations can result in large output changes in deep networks [4,5]. Robust optimization methods formalize this phenomenon through minmax training targets [6], at the same time as certification strategies offer provable input-space guarantees below Lipschitz or randomized smoothing assumptions [7–9]. These works ordinarily recognition on worldwide input sensitivity, treating networks as enter-output mappings without at once studying the internal geometry of intermediate representations.

Related theoretical efforts connect robustness and generalization to norm-based or spectral complexity measures. Margin-based totally and spectral norm bounds try to manipulate generalization through international Lipschitz constants or layerwise operator norms [10–12]. Empirical research has examined the correlations between sensitivity metrics and generalization performance [13,14]. While those techniques quantify elements of useful smoothness, they do no longer isolate internal representational balance as a separate structural amount. In unique, global norm manage does now not always suggest uniform balance across intermediate layers.

The geometry of discovered representations has additionally been investigated via representational similarity and subspace analyses. Canonical correlationprimarily based strategies [15,16] evaluate internal representations throughout networks and training ranges. Visualization of loss landscapes and flatness houses has supplied perception into parameter-area geometry [17]. More recently, phenomena consisting of neural fall apart characterize emergent geometric shape within the terminal layers of deep classifiers [18]. These research monitor structural regularities in illustration area, but they commonly describe statistical alignment or clustering consequences rather than perturbation balance with appreciate to output alternate.

From a dynamical systems perspective, analyses of sign propagation and dynamical isometry take a look at how structure impacts gradient float and conditioning [19,20]. The neural tangent kernel framework characterizes training dynamics in overparameterized regimes [21]. Information-theoretic viewpoints interpret deep learning through compression and mutual-information trade-offs [22,23]. While these views address stability and structure in different forms, none directly formalize a layerwise minimum perturbation radius that links hidden-space perturbations to controlled output variation.

While existing robustness and generalization frameworks primarily characterize input-space sensitivity, output stability, or norm-based smoothness, the present work focuses on perturbation stability inside intermediate representation spaces. In contrast to prior robustness, generalization, and representational-similarity frameworks, opacity is designed to quantify perturbation stability directly within hidden representation spaces. Its novelty lies in treating the minimum hidden-space perturbation required for controlled output change as a layerwise geometric property, rather than as an input-output robustness measure, a norm-based complexity bound, or a statistical similarity score.

Collectively, earlier work establishes that neural networks may be analyzed by entering robustness, norm-based complexity, representational similarity, dynamical conditioning, or information compression. However, those frameworks do not provide a unified geometric useful that (i) operates explicitly in hidden space, (ii) quantifies the minimal internal perturbation required for output change, and (iii) permits pass-structure comparison impartial of predictive accuracy. The opacity practical delivered in this work addresses this hole by formalizing inner perturbation stability as a structural property of discovered representations as opposed to completely of enteroutput mappings.

Despite sizable progress in knowledge robustness, generalization, and representation geometry, present frameworks usually quantify global sensitivity, alignment, or data waft. They do not formalize the minimal internal perturbation required to set off the controlled output variant as structural assets of a community. In particular, prior work does no longer isolate a layerwise hidden-state radius that hyperlinks inner representational stability to useful behavior in a principled manner. Our formulation of opacity addresses this gap by defining a geometric functional over intermediate representations, thereby separating architectural structure from training-dependent sensitivity. This perspective

complements current robustness and generalization theories at the same time as introducing a brand new axis along which neural models may be analyzed.

3 Opacity as an Internal Geometric Quantity

This section formalizes opacity as a geometric property of internal representations. We define layerwise opacity radii, relate them to Jacobian norms under mild regularity conditions, and show that architectural structure alone can induce arbitrarily small global opacity—even when input–output sensitivity remains fixed. This section presents the main theoretical formulation and proof sketches; detailed proofs and technical derivations are provided in Appendix A.

3.1 Layerwise Opacity Radius

Let $f : \mathbb{R}^d \to \mathbb{R}^k$ be a neural network composed of L intermediate representations:

$$h_0 = x, \qquad h_{\ell+1} = g_\ell(h_\ell), \qquad f(x) = h_L. \tag{1}$$

For any layer $\ell \in \{0, \dots, L-1\}$, define the *tail mapping*

$$F_\ell(h) := g_{L-1} \circ \cdots \circ g_\ell(h), \qquad f(x) = F_\ell(h_\ell(x)). \tag{2}$$

Fix an internal layer ℓ and an output tolerance $\eta > 0$. We define the layerwise opacity radius at x as

$$r_\ell(x;\eta) := \inf \Big\{ r > 0 \ : \ \sup_{\|\varepsilon\|_2 \le r} \|F_\ell(h_\ell(x) + \varepsilon) - F_\ell(h_\ell(x))\|_2 \ \ge \ \eta \Big\}. \tag{3}$$

This is the smallest internal perturbation magnitude that can induce an output change of size at least η.

Define the *pointwise opacity*

$$R(x;\eta) = \min_{\ell=0,\dots,L-1} r_\ell(x;\eta), \tag{4}$$

and the *global opacity*

$$R(f;\eta) = \mathbb{E}_{x\sim\mathcal{D}}\big[R(x;\eta)\big]. \tag{5}$$

Opacity thus measures the smallest internal bottleneck through which a significant output change can be induced.

3.2 First-Order Characterization

Let

$$J_\ell(x) = \left\| \frac{\partial F_\ell}{\partial h}\big(h_\ell(x)\big) \right\|_{2\to 2}. \tag{6}$$

denote the operator norm of the Jacobian from representation h_ℓ to the output.

Lemma 1 (First-Order Opacity Approximation). *Fix $\ell \in \{0, \dots, L-1\}$ and let F_ℓ denote the tail mapping so that $f(x) = F_\ell(h_\ell(x))$. Assume F_ℓ is differentiable at $h_\ell(x)$ and its Jacobian is locally Lipschitz in a neighborhood of $h_\ell(x)$. Then for sufficiently small η,*

$$r_\ell(x;\eta) = \frac{\eta}{J_\ell(x)} + O(\eta^2), \tag{7}$$

where $J_\ell(x) = \left\| \frac{\partial F_\ell}{\partial h}(h_\ell(x)) \right\|_{2\to 2}$.

Interpretation. Opacity at a layer is inversely proportional, to first order, to the sensitivity of the output with respect to that representation. Layers with large Jacobian norm admit small opacity radii.

Remark (Smoothness Requirement). If the Jacobian $\partial f / \partial h_\ell$ is locally Lipschitz, the $O(\eta^2)$ remainder is controlled by the second-order variation of the network in that layer. No global linearity assumption is required.

3.3 Architectural Decoupling

We now show that opacity can be driven to zero independently of input–output sensitivity.

Theorem 1 (Architectural Opacity Decoupling). *There exists a piecewise-linear (ReLU) network architecture and a data distribution supported inside a single linear region such that:*

1. *the input–output Lipschitz constant is fixed, and*
2. *the global opacity $R(f;\eta)$ can be made arbitrarily small.*

Proof sketch. Inside a fixed ReLU activation region, the network reduces to a linear map:

$$f(x) = W_L W_{L-1} \cdots W_1 x. \tag{8}$$

Choose a decomposition

$$W_1 = \frac{1}{\alpha} W, \qquad W_L = \alpha I, \tag{9}$$

with all intermediate layers identity maps. Then $f(x) = Wx$, so the input–output mapping remains unchanged. However, the internal Jacobian from representation h_1 to the output satisfies

$$J_1(x) = \left\| \frac{\partial f}{\partial h_1}(x) \right\|_{2\to 2} = \|W_L\| = \alpha \|I\|. \tag{10}$$

As $\alpha \to \infty$, the internal sensitivity $J_1(x)$ becomes unbounded, and by Lemma 1,

$$r_1(x;\eta) = \frac{\eta}{J_1(x)} + O(\eta^2) \to 0. \tag{11}$$

Thus $R(f;\eta) \to 0$ while the global Lipschitz constant of f remains fixed. □

Interpretation. Opacity is not determined by the overall input–output mapping. It depends critically on internal factorization. Architecture alone can create arbitrarily thin internal bottlenecks.

3.4 Attention and Representation Diffusion

Transformer layers redistribute information across tokens via attention weights. Let a single attention block produce

$$z = A(h)Vh, \tag{12}$$

where $A(h)$ is the attention matrix and V a value projection. The Jacobian with respect to a token representation h_i decomposes as

$$\frac{\partial z}{\partial h_i} = V^\top A_{\cdot i} + \frac{\partial A}{\partial h_i} Vh. \tag{13}$$

Thus opacity depends not only on value projections but also on how attention weights vary with the representation. Architectures that distribute attention mass broadly across tokens may increase layerwise radii by diffusing local perturbations, whereas sharply concentrated attention may reduce them. Opacity therefore reflects structural information flow patterns intrinsic to the architecture.

3.5 Summary of Theoretical Implications

The preceding subsections have shown that opacity admits a local geometric characterization through layerwise perturbation radii and is inversely related, to first order, to layerwise Jacobian norms. They also show that opacity can be made arbitrarily small under architectural reparameterization without changing input-output sensitivity. Finally, the analysis of attention mechanisms shows that structured diffusion effects can modulate opacity through both value projections and attention gradients.

These results position opacity as a structural internal property of neural representations rather than a mere consequence of global Lipschitz behavior. Detailed proofs and additional technical derivations corresponding to these results are provided in Appendix A.

4 Estimation of Opacity

Section 3 introduced opacity as a layerwise sensitivity–radius functional. We now describe practical estimators for this quantity that are architecture-agnostic and scalable to modern neural networks.

Our objective is to estimate both the input sensitivity proxy and the minimal internal perturbation radius.

Both procedures rely only on automatic differentiation and forward evaluation, and apply uniformly to multilayer perceptrons, convolutional networks, and transformer architectures.

4.1 Directional Jacobian Sensitivity Estimation

Exact computation of Jacobian operator norms is computationally expensive. Instead, we approximate layerwise sensitivities using randomized directional projections.

Given a model f and input x, we approximate

$$\|J_{h_\ell \to f}(x)\|_{2\to 2}$$

by sampling random unit vectors u in output space and computing

$$\|\nabla_{h_\ell} \langle f(x), u \rangle\|_2 .$$

For sufficiently many sampled directions, the maximum observed norm provides a stable proxy for the operator norm. This procedure requires only backpropagation and scales linearly with the number of sampled directions.

The same estimator is used both for input sensitivity (when $\ell = 0$) and for layerwise sensitivity terms entering opacity.

4.2 Layerwise Radius Estimation

Opacity depends on the minimal perturbation in hidden space required to induce a fixed output change. For each layer ℓ, we estimate

$$r_\ell(x) = \inf \{\|\delta\|_2 : \|f_\ell(h_\ell(x) + \delta) - f(x)\|_2 \geq \eta\}, \tag{14}$$

where f_ℓ denotes the remainder of the network from layer ℓ onward.

We approximate this value through directional search combined with binary refinement. For each hidden layer, random unit directions v are sampled in hidden space. The perturbation magnitude is then expanded along each direction until the output change exceeds the threshold η, after which binary search is used to refine the boundary.

The minimal radius across sampled directions provides an upper bound on the true minimal perturbation radius.

This procedure requires only forward passes and scales linearly in the number of directions and refinement steps.

4.3 Opacity Estimator

For each input x, opacity is computed as the minimum layerwise radius:

$$R(x) = \min_\ell r_\ell(x).$$

In practice, we evaluate this quantity on a random subset of test data and report the mean input sensitivity L_{avg}, the maximum input sensitivity $L_{\max}$, and the mean opacity $\widehat{R}$.

The estimator is architecture-agnostic and does not rely on model-specific assumptions.

4.4 Matched-Sensitivity Separation Protocol

To empirically validate the structural decoupling established in Sect. 3.3, we compare networks under matched sensitivity conditions.

Specifically, we identify pairs of models with comparable predictive accuracy, within 1%, and comparable input sensitivity L_{avg}, within a fixed tolerance.

Within such matched pairs, differences in opacity cannot be attributed to predictive performance or global smoothness. Observed discrepancies, therefore, provide empirical evidence that opacity captures structural information not encoded in traditional metrics.

4.5 Computational Complexity

Let N denote the number of evaluation samples, L the number of layers, S the number of hidden directions, and T the number of binary refinement steps. Let K denote the number of sampled output directions for Jacobian estimation.

The overall computational cost scales as

$$\mathcal{O}\big(NL(K + S\ \mathrm{T})\big).$$

In practice, moderate values of S and T yield stable estimates, making the procedure feasible for medium-scale architectures.

5 Empirical Evidence for Architecture-Dependent Opacity

Sections 3–4 defined *opacity* as an internal property of neural representations, formalized through bottleneck geometry and operationalized via the empirical estimator $\hat{R}$. In this section, we evaluate whether opacity, so defined, exhibits systematic structure across architectures and whether such structure can be attributed primarily to architectural design rather than to training noise, regularization, or predictive performance. All experiments follow exactly the measurement pipeline introduced in Sect. 4 and are interpreted strictly through the theoretical lens developed in Sect. 3.

Table 1. Architecture-Level Comparison of Test Accuracy, Average Input Sensitivity (L_{avg}), and Estimated Bottleneck Radius ($\hat{R}$).

Architecture	Dataset	Test Acc.	L_{avg}	$\hat{R}$
MLP	MNIST	0.98	7.93	0.46
CNN	MNIST	0.99	3.21	2.43
CNN	CIFAR-10	0.69	9.05	1.70
TinyViT	CIFAR-10	0.66	2.09	4.10

5.1 Architecture-Level Separation in Opacity

Table 1 above reports test accuracy, average input sensitivity L_{avg}, and estimated bottleneck radius $\hat{R}$ for the MLP, CNN, and Transformer (TinyViT) models on MNIST and CIFAR-10.

Figure 1 below visualizes the same results by embedding models in the $(L_{\text{avg}}, \hat{R})$ plane. The results Indicate That Internal Representational Geometry Varies Across Architectures and Is Not Determined Solely by Predictive Performance or Input-Level Sensitivity.

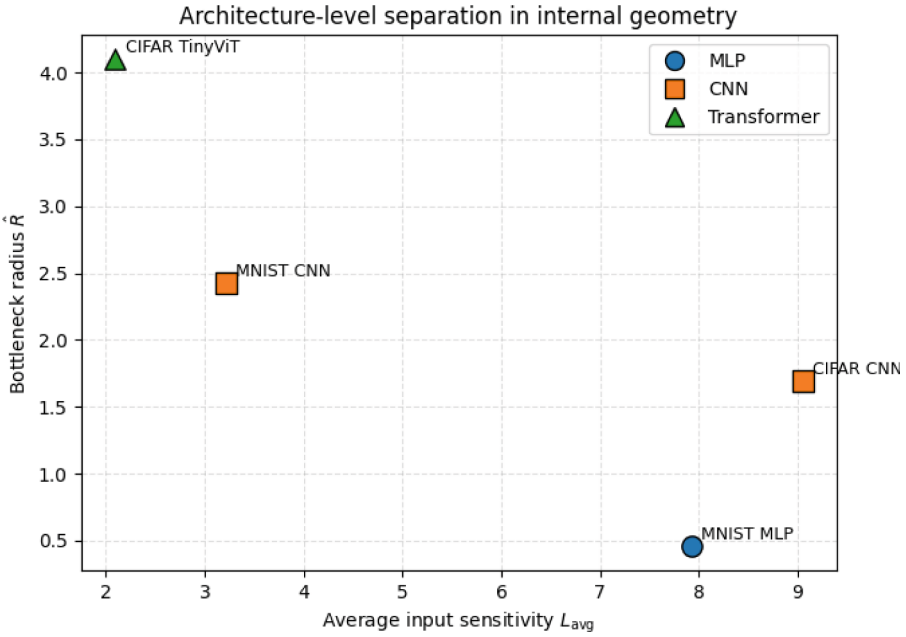

Fig. 1. Architecture-level Separation In Opacity.

Specifically, two observations are central. First, **opacity is not determined by predictive accuracy**. For example, the MNIST MLP and MNIST CNN both achieve near-saturated accuracy, yet their opacity estimates differ sharply (MLP: $\hat{R} \approx 0.46$; CNN: $\hat{R} \approx 2.43$). This directly reflects the bottleneck formulation of Sect. 3: even when end-task performance is matched, the minimum-radius layer governing opacity may occur at different depths and scales depending on architectural constraints.

Second, **opacity is not reducible to input-level sensitivity**. As shown in Fig. 1, models with low L_{avg} do not necessarily exhibit low opacity. In particular, the TinyViT model on CIFAR-10 occupies a regime characterized by simultaneously low input sensitivity ($L_{\text{avg}} \approx 2.09$) and large bottleneck radius

($\hat{R} \approx 4.10$), whereas the CIFAR-10 CNN exhibits substantially higher input sensitivity but a smaller $\hat{R}$. This separation demonstrates that opacity captures an internal property not controlled by input smoothness alone.

Taken together, Table 1 and Fig. 1 provide direct empirical support for the title claim: **opacity varies systematically with architecture** and cannot be inferred from accuracy or input sensitivity.

5.2 Layerwise Sensitivity Profiles and the Structure of Opacity

Section 3 characterizes opacity through a *min-over-layers* bottleneck mechanism. If opacity is genuinely architecture-dependent, this dependence should manifest in the layerwise internal sensitivity profiles J_ℓ.

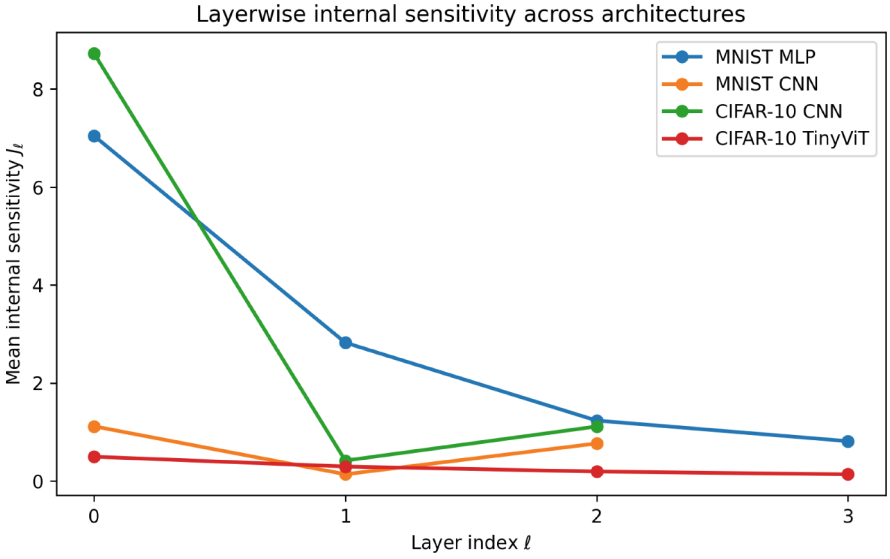

Fig. 2. Layerwise Sensitivity Profiles and the Structure of Opacity.

The MNIST MLP exhibits a high initial sensitivity followed by gradual decay, consistent with diffuse early representations and a deep bottleneck that yields a small $\hat{R}$. Above, Fig. 2 plots J_ℓ in depth for the evaluated models, calculated as described in Sect. 4. CNNs display non-monotonic profiles, characterized by sharp early compression and partial re-expansion, reflecting localized architectural bottlenecks and corresponding to larger opacity on MNIST. In contrast, the TinyViT model shows uniformly low and smoothly decreasing sensitivity across layers, consistent with distributed representation mixing and a comparatively large bottleneck radius.

These patterns provide *mechanistic grounding* for the scalar opacity estimates reported in Table 1. Differences in $\hat{R}$ arise not from noise, but from coherent depth-wise sensitivity structure induced by architecture.

5.3 Limited Tunability of Opacity Within an Architecture (Table 2)

A natural counter-hypothesis is that opacity can be freely adjusted through training-time interventions such as dropout or weight decay. To test this, Table 2 reports matched-performance configurations within a fixed architecture (MNIST MLP), selected to have similar test accuracy and closely matched L_{avg}, while varying regularization strength. Configurations are selected to have similar test accuracy ($\pm 1\%$) and comparable input sensitivity (L_{avg} within $\pm 7\%$). The estimated bottleneck radius ($\hat{R}$) varies only modestly, indicating limited tunability of opacity via regularization.

Table 2. Matched-Performance Configurations Within MNIST MLP.

Dropout	Weight Decay	Test Acc.	$\hat{R}$
0.0	10^{-3}	0.9749	0.44
0.2	10^{-4}	0.9781	0.48
0.0	10^{-3}	0.9749	0.44
0.2	10^{-3}	0.9694	0.48
0.0	10^{-4}	0.9737	0.46
0.0	10^{-3}	0.9749	0.44

Despite these variations, the resulting opacity estimates remain confined to a narrow range. Once predictive performance and input sensitivity are controlled for, changes in $\hat{R}$ are modest relative to the cross-architecture differences observed in Table 1. This indicates that opacity, as formalized in Sect. 3, is **not a fragile training artifact**, but is instead strongly constrained by architectural structure.

6 Conclusion

We redefined opacity as a formally defined geometric functional over internal neural representations and demonstrated that it constitutes a structural property distinct from classical input-based smoothness measures. Our theoretical evaluation and empirical results provide evidence that opacity captures an internal property of representation geometry that cannot be characterized solely in terms of predictive accuracy or global sensitivity. Specifically, two architectures with equivalent predictive performance may differ substantially with respect to their opacity properties; therefore, the internal geometric organization of a network is likely determined primarily by its architectural structure. The findings outlined in the paper indicate that, from a computational science perspective, a study of internal representation behavior can also be conducted using other measurable quantities in addition to external performance. Therefore, when comparing

architectures, predictive accuracy is not the only factor of concern; how internal perturbations are handled and transmitted between layers is also important.

This perspective reframes opacity as a feature of representational bottlenecks and perturbation stability rather than simply as a high-level concept that is synonymous with interpretability or robustness. By identifying a small perturbation radius for hidden states that is shown to be independent of standard Lipschitz-type metrics, we establish an architectural dimension of neural networks that has generally remained unexplored in the literature on generalization and robustness. Our findings suggest that the internal geometry of neural networks may influence the way in which perturbations move through the layers of the network and potentially, in turn, shape the nature of the uncertainty associated with prediction accuracy, regardless of calibration and probabilistic outputs.

One avenue for future research is to develop a deeper analytical understanding of opacity through the lens of compositional nonlinear operators. By investigating this avenue, it may be possible to obtain deeper insights related to the nature of opacity in extremely deep networks and/or networks that have been heavily over-parameterized. Investigating the extension of our model to probabilistic or Bayesian architectures may also provide insight into potential formal relationships between the underlying geometric constraints of neural networks and epistemically based uncertainties.

Our findings suggest that the internal geometry of neural networks may influence how perturbations propagate through network layers and may, in turn, be related to aspects of predictive uncertainty that are not fully captured by calibration or probabilistic output measures alone. Opacity does not replace existing robustness or uncertainty notions; instead, it complements them by exposing internal geometric regimes that are invisible to conventional smoothness metrics. More broadly, this view supports the computational study of neural networks as layered systems whose internal behaviour can be analysed quantitatively. We suggest that viewing opaqueness as a geometric attribute of architecture provides a new perspective on neural representations and offers an avenue for studying structural invariants in modern learning systems.

A Proofs and Technical Details for Section 3

A.1 Preliminaries and Notation

Let

$$f : \mathbb{R}^d \to \mathbb{R}^k$$

be a feedforward neural network with intermediate representations

$$h_0 = x, \qquad h_{\ell+1} = g_\ell(h_\ell), \qquad f(x) = h_L.$$

For any layer ℓ, denote the tail mapping

$$F_\ell(h) := g_{L-1} \circ \cdots \circ g_\ell(h),$$

so that

$$f(x) = F_\ell(h_\ell(x)).$$

Let

$$J_\ell(x) = \left\| \frac{\partial f}{\partial h_\ell}(x) \right\|_{2\to 2} = \left\| \frac{\partial F_\ell}{\partial h_\ell}(h_\ell(x)) \right\|_{2\to 2}.$$

All norms are Euclidean unless stated otherwise.

A.2 First-Order Approximation of the Opacity Radius (Proof of Lemma 3.1)

Recall the definition of the layerwise opacity radius:

$$r_\ell(x;\eta) = \inf \left\{ r > 0 : \sup_{\|\varepsilon\|_2 \le r} \|F_\ell(h_\ell(x)+\varepsilon) - F_\ell(h_\ell(x))\|_2 \ge \eta \right\}. \tag{15}$$

Assumptions. Assume:

1. F_ℓ is differentiable at $h_\ell(x)$,
2. its Jacobian is locally Lipschitz in a neighborhood of $h_\ell(x)$.

Step 1: Taylor Expansion. For sufficiently small perturbations ε,

$$F_\ell(h_\ell(x)+\varepsilon) = F_\ell(h_\ell(x)) + J_\ell(x)\varepsilon + R(\varepsilon),$$

where

$$\|R(\varepsilon)\|_2 \le C\|\varepsilon\|_2^2$$

for some local constant $C > 0$.

Step 2: Factorization Trick. Let the overall linear mapping be

$$f(x) = Wx.$$

Construct a depth-3 network (within a fixed linear region) by choosing

$$W_1 = \frac{1}{\alpha} W, \qquad W_2 = I, \qquad W_3 = \alpha I,$$

with $\alpha > 0$. Then

$$W_3 W_2 W_1 = (\alpha I)\, I \left(\frac{1}{\alpha} W\right) = W,$$

so the input–output function is unchanged on this linear region.

Step 3: Internal Jacobian at Layer 1. The Jacobian from representation h_1 to the output is

$$J_1 = W_3 W_2 = \alpha I,$$

and hence

$$\|J_1(x)\| = \alpha \|I\| \xrightarrow[\alpha \to \infty]{} \infty.$$

Therefore, by the first-order radius relation,

$$r_1(x;\eta) \approx \frac{\eta}{\|J_1(x)\|} = \frac{\eta}{\alpha \|I\|} \xrightarrow[\alpha \to \infty]{} 0.$$

Thus opacity vanishes while the input–output mapping remains fixed. □

A.3 Opacity Under Piecewise Linear Activations

Even outside a single linear region, the first-order analysis applies locally.

For almost every input (except on measure-zero activation boundaries), ReLU networks are differentiable. Thus Lemma 3.1 applies almost everywhere under the data distribution.

Therefore global opacity

$$R(f;\eta) = \mathbb{E}_{x \sim \mathcal{D}}[R(x;\eta)]$$

inherits the decoupling phenomenon except possibly on activation boundaries.

A.4 Attention Layer Derivative Details

Consider a single attention head:

$$z = A(h)Vh,$$

where

$$A(h) = \text{softmax}\left(\frac{Q(h)K(h)^\top}{\sqrt{d}}\right).$$

For a token representation h_i,

$$\frac{\partial z}{\partial h_i} = V^\top A_{\cdot i} + \left(\frac{\partial A}{\partial h_i}\right) Vh.$$

The first term corresponds to value projection sensitivity. The second term captures sensitivity induced through attention weight changes.

Thus the Jacobian norm reflects both:

1. mixing strength (via A),
2. sensitivity of attention scores.

This justifies the claim in Sect. 3.4 that attention can increase or decrease opacity depending on its structural diffusion properties.

References

1. Nadif, M.A., Hasan, M.D.S., Sharma, A.K.: BanglaTag: next-gen Bangladeshi number plate recognition hierarchical processing network. In: 2024 27th International Conference on Computer and Information Technology (ICCIT). IEEE (2024)
2. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
3. Sharma, A.K.: DistilBERT based malicious URL detection system. In: 2024 4th International Conference on Digital Society and Intelligent Systems (DSInS), pp. 307–312. IEEE (2024)
4. Szegedy, C., et al.: Intriguing Properties of Neural Networks. In: International Conference on Learning Representations (ICLR) (2014)
5. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (ICLR) (2015)
6. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (ICLR) (2018)
7. Hein, M., Andriushchenko, M.: Formal guarantees on the robustness of a classifier against adversarial manipulation. In: Advances in Neural Information Processing Systems. NeurIPS (2017)
8. Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., Usunier, N.: Parseval networks: improving robustness to adversarial examples. In: International Conference on Machine Learning (ICML) (2017)
9. Cohen, J.M., Rosenfeld, E., Kolter, Z.: Certified adversarial robustness via randomized smoothing. In: International Conference on Machine Learning (ICML) (2019)
10. Bartlett, P.L., Foster, D.J., Telgarsky, M.: Spectrally-normalized margin bounds for neural networks. In: Advances in Neural Information Processing Systems. NeurIPS (2017)
11. Neyshabur, B., Bhojanapalli, S., McAllester, D., Srebro, N.: Exploring generalization in deep learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
12. Arora, S., Ge, R., Neyshabur, B., Zhang, Y.: Stronger generalization bounds for deep nets via a compression approach. In: International Conference on Machine Learning (ICML) (2018)
13. Novak, R., Bahri, Y., Abolafia, D.A., Pennington, J., Sohl-Dickstein, J.: Sensitivity and generalization in neural networks: an empirical study. In: International Conference on Learning Representations (ICLR) (2018)
14. Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., Bengio, S.: Fantastic generalization measures and where to find them. In: International Conference on Learning Representations (ICLR) (2020)
15. Raghu, M., Gilmer, J., Yosinski, J., Sohl-Dickstein, J.: SVCCA: singular vector canonical correlation analysis for deep learning dynamics and interpretability. In: Advances in Neural Information Processing Systems. NeurIPS (2017)
16. Morcos, A.S., Raghu, M., Bengio, S.: Insights on representational similarity in neural networks with canonical correlation. In: Advances in Neural Information Processing Systems (NeurIPS) (2018)
17. Li, H., Xu, Z., Taylor, G., Studer, C., Goldstein, T.: Visualizing the loss landscape of neural nets. In: Advances in Neural Information Processing Systems (NeurIPS) (2018)

18. Papyan, V., Han, X., Donoho, D.L.: Prevalence of neural collapse during the terminal phase of deep learning training. Proc. Natl. Acad. Sci. **117**(40), 24652–24663 (2020)
19. Saxe, A.M., McClelland, J.L., Ganguli, S.: Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In: International Conference on Learning Representations (ICLR) (2014)
20. Pennington, J., Schoenholz, S.S., Ganguli, S.: Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In: Advances in Neural Information Processing Systems. NeurIPS (2017)
21. Jacot, A., Gabriel, F., Hongler, C.: Neural tangent kernel: convergence and generalization in neural networks. In: Advances in Neural Information Processing Systems (NeurIPS) (2018)
22. Tishby, N., Zaslavsky, N.: Deep learning and the information bottleneck principle. In: IEEE Information Theory Workshop. ITW (2015)
23. Achille, A., Soatto, S.: Information dropout: learning optimal representations through noisy computation. IEEE Trans. Pattern Anal. Mach. Intell. (2018)