© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717189

Bridging Deep Learning and Ensemble Methods: Residual Networks as Implicit Boosting Models

KSHITIJ KATARIYA[1], SHALINI MAURYA[2], ANVESHA TRIPATHI[3]
[1,2,3]*Department of Computer Science (Data Science), Greater Noida Institue of Technology*

Abstract- ***The advent of deep learning has revolutionized predictive modeling across unstructured data domains, largely driven by the capacity of deep neural networks to learn hierarchical, high-dimensional feature representations. Among the most critical architectural innovations of the past decade, Residual Networks (ResNets) have been instrumental in enabling the training of ultra-deep models by mitigating the vanishing gradient problem through identity shortcut connections. Concurrently, ensemble learning methods—particularly Gradient Boosting Machines (GBMs)—have dominated tabular data tasks by iteratively combining weak learners to minimize an arbitrary differentiable loss function. This paper investigates the profound theoretical and empirical intersections between these two seemingly disparate paradigms. We posit and mathematically demonstrate that Residual Networks can be conceptually interpreted as implicit boosting models. By unrolling the recursive structural equations of ResNets, we illustrate that individual residual blocks function analogously to additive weak learners in a boosting ensemble, where each subsequent layer is jointly optimized to fit the residual error of the preceding layers' representations. This research formalizes the fundamental equivalence between additive modeling in gradient boosting and the residual mapping in deep neural networks. Furthermore, we explore the empirical implications of this equivalence, conducting theoretical analyses of lesion studies, stochastic depth optimization, and gradient flow stability. Our findings provide a unified framework that enhances the interpretability of deep representation learning, demystifies the robustness of skip-connections, and paves the way for novel hybrid architectures that leverage the strengths of both global backpropagation and ensemble robustness.***

Index Terms - Boosting, Deep Learning, Dynamical Systems, Ensemble Methods, Residual Networks

I. INTRODUCTION

The landscape of modern machine learning is broadly partitioned into two highly successful, yet historically isolated, paradigms: deep representation learning and ensemble learning. Deep learning architectures, primarily Convolutional Neural Networks (CNNs) and Transformers, excel in domains characterized by dense, unstructured data such as computer vision and natural language processing. The success of these models is predicated on the hierarchical composition of non-linear transformations, which iteratively extract increasingly abstract features from raw inputs. However, as the quest for higher accuracy drove architectures to unprecedented depths, researchers encountered severe optimization degradation, primarily the vanishing and exploding gradient phenomena. The introduction of Residual Networks (ResNets) by He et al. provided an elegant resolution by introducing identity shortcut connections, allowing gradients to flow unimpeded and enabling the stable training of networks spanning hundreds or thousands of layers.

Parallel to the evolution of neural architectures, ensemble learning—specifically Gradient Boosting—has established itself as the foundational methodology for predictive modeling on structured, tabular datasets. Boosting fundamentally relies on the principle of constructing a strong learner by sequentially adding weak learners (typically shallow decision trees). Each subsequent learner is specifically trained to correct the pseudo-residuals of the aggregated ensemble up to that stage. While deep learning is canonically viewed as a hierarchical feature extractor, boosting is fundamentally an additive model grounded in functional gradient descent.

The primary problem this paper addresses is the theoretical divergence and lack of structural synthesis between these two dominant methodologies. Traditionally, a deep neural network is perceived as a singular, cohesive estimator. A breakdown in any intermediate layer of a standard feed-forward

network catastrophically destroys the output. However, empirical observations—such as the resilience of ResNets to the arbitrary dropping of layers during inference (stochastic depth)—directly challenge this monolithic view. ResNets exhibit graceful degradation, a structural property that is highly characteristic of ensemble models like Random Forests or Boosting Machines, rather than sequential pipelines.

The motivation for explicitly connecting ResNets with Boosting lies in the pursuit of both mathematical interpretability and the next generation of architectural optimization. By framing residual connections mathematically as an additive boosting process, we can demystify the "black box" nature of deep network optimization. This equivalence suggests that deeper layers in a ResNet do not necessarily learn entirely new, higher-order semantic representations, but rather perform iterative, fine-grained refinement of features already learned by earlier layers.

The core contributions of this expanded research are:

1. We formalize the exact mathematical mapping between the iterative error correction mechanism in Gradient Boosting and the identity mapping formulation in Residual Networks.
2. We provide rigorous theoretical intuition, grounded in continuous-time dynamical systems and Ordinary Differential Equations (ODEs), detailing how individual residual blocks behave as weak learners within a functional gradient descent framework.
3. We present an extended discussion on the empirical implications of this formulation, comparing the optimization landscapes, gradient flow dynamics, and generalization capabilities of both paradigms under lesion constraints.

II. RESEARCH ELABORATION / METHODOLOGY

To establish a rigorous equivalence between Residual Networks and Boosting models, it is necessary to deconstruct the mathematical formulations and optimization trajectories governing both architectures.

A. Architecture of Residual Networks and the Unrolled View

In a standard, purely sequential feed-forward neural network (such as VGG or AlexNet), the output of the l-th layer is represented as a direct non-linear transformation of the previous layer's output:

$$x_{l+1} = H(x_l, W_l)$$

where x_l is the input representation to the layer, W_l represents the learnable parameters (weights and biases), and H denotes the composite transformation (e.g., a sequence of convolution, Batch Normalization, and ReLU activation).

Residual Networks introduce a paradigm shift by adding an identity mapping (the shortcut connection). Instead of hoping each few stacked layers directly fit a desired underlying mapping, ResNets explicitly force these layers to fit a residual mapping. The output of a standard residual block is defined as:

$$x_{l+1} = x_l + F(x_l, W_l)$$

To understand the ensemble nature of this architecture, we must analyze the unrolled network. Through recursive expansion of the above equation, the state of the network at any deeper layer L can be expressed as the summation of the initial input x_0 and the outputs of all preceding residual functions:

$$x_L = x_0 + \Sigma F(x_i, W_i) \text{ (from } i=0 \text{ to } L-1)$$

This unrolled equation is highly revealing. It demonstrates that x_L is fundamentally an additive model. The network does not compute a single deeply composed and fragile function, but rather an accumulation of numerous shallow residual functions. Furthermore, for any two layers L and l (where $L > l$), the relationship is mathematically defined as:

$$x_L = x_l + \Sigma F(x_i, W_i) \text{ (from } i=l \text{ to } L-1)$$

This indicates that the feature representation at layer L is merely the representation at layer l plus an additive refinement term.

B. Mechanics of Functional Gradient Boosting

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717189

Gradient Boosting is an ensemble machine learning technique that produces a prediction model by optimizing an arbitrary differentiable loss function L(y, F(x)). The objective is to find a function F_hat(x) that minimizes the expected value of this loss over the joint distribution of inputs and targets.
Boosting approaches this optimization problem by performing gradient descent not in parameter space, but in function space. The model is built in a stage-wise, additive fashion:

$F_m(x) = F_{m-1}(x) + \alpha_m * h_m(x)$

where $F_{m-1}(x)$ is the aggregated ensemble model at stage m-1, $h_m(x)$ is the newly added weak learner (a parameterized function, usually a regression tree), and α_m is the learning rate or shrinkage parameter.
Crucially, the weak learner $h_m(x)$ is trained to approximate the negative gradient of the loss function with respect to the previous ensemble's predictions. These are the pseudo-residuals:

$r_{i, m} = - [\partial L(y_i, F(x_i)) / \partial F(x_i)]$ at $F(x) = F_{m-1}(x)$

The weak learner is then fitted to the dataset $\{(x_i, r_{i, m})\}$.

C. The Conceptual Mapping: ResNets as Implicit Boosting
By juxtaposing the unrolled ResNet equation with the boosting additive model, a profound structural and functional alignment emerges:

1. Additive Structural Modeling: Both formalisms generate their final, high-capacity output via summation rather than pure composition. In boosting, $F_M(x) = F_0(x) + \Sigma \alpha_m * h_m(x)$. In a ResNet, $x_L = x_0 + \Sigma F(x_i, W_i)$.
2. The Role of Weak Learners: In boosting, $h_m(x)$ acts as a base weak learner with restricted capacity. In a ResNet, each residual block $F(x_i, W_i)$ acts identically as a weak learner that perturbs the identity mapping. Because modern neural network weights are typically initialized near zero (or with specific variance scaling like Kaiming initialization), F initially outputs values infinitesimally close to zero. This makes the residual block an exceptionally 'weak' learner at the start of training, which gradually increases in capacity as it refines the feature space.
3. Iterative Error Correction via Backpropagation: While boosting trains $h_m(x)$ to correct the explicitly computed pseudo-residuals of $F_{m-1}(x)$, the global backpropagation process in a ResNet achieves the same end implicitly. During gradient descent, the updates to W_i are directed to correct the representation errors accumulated up to layer i in order to minimize the global loss at layer L. The residual block computes a step in the feature space direction that minimizes the loss, effectively learning the 'residual' representation required.

D. Theoretical Intuition via Ordinary Differential Equations (ODEs)
The equivalence can be further formalized and generalized through the lens of continuous-time dynamical systems. A discrete residual block can be mathematically viewed as an Euler discretization step of an Ordinary Differential Equation (ODE) governing the evolution of the hidden state over 'time' (depth).

Let the continuous depth parameter be t. The dynamics of the hidden state x(t) are governed by:
$dx(t)/dt = F(x(t), t, \theta)$

Discretizing this continuous formulation with a standard Euler method and a step size of $\Delta t = 1$ yields the exact ResNet formulation:

$x(t+1) = x(t) + F(x(t), t, \theta)$

Simultaneously, Gradient Boosting can be formulated as the Euler integration of gradient flows in an infinite-dimensional function space. Therefore, both ResNets and Gradient Boosting are simply discrete numerical approximations of an underlying, continuous functional refinement process.

III. RESULTS / FINDINGS

While this paper focuses heavily on synthesizing theoretical frameworks, the validity of the 'ResNet-as-Boosting' hypothesis is strongly corroborated by observable structural behaviors and empirical studies in literature.

A. Robustness to Lesion Studies and Layer Ablation

The most compelling evidence for the ensemble nature of ResNets comes from lesion studies. In a traditional composed network (e.g., AlexNet), features are strictly hierarchical; removing a single intermediate convolutional layer breaks the representational chain, leading to a catastrophic drop in accuracy (reducing performance to random chance).

Conversely, pioneering work by Veit et al. demonstrated that dropping a residual block from a fully trained ResNet at inference time results in only a graceful, linear degradation in performance. If a ResNet-110 has a test error of 6%, removing one block might only increase the error to 6.5%. This behavior is the hallmark of an ensemble. If you remove a few trees from a Random Forest or an XGBoost model containing 500 trees, the overall predictive variance increases slightly, but the model does not collapse. The residual blocks, acting as implicit weak learners, lack the catastrophic co-dependence seen in traditional deep learning.

B. Optimization Stability and Gradient Flow

In gradient boosting, the additive nature strictly prevents the vanishing gradient problem in function space; errors are directly computed against the target and fitted sequentially. In ResNets, the gradient of the loss E with respect to the input of an intermediate layer x_l is derived by the chain rule over the unrolled addition:

$$\partial E/\partial x_l = \partial E/\partial x_L * \partial x_L/\partial x_l = \partial E/\partial x_L * [\ 1 + \partial/\partial x_l \Sigma F(x_i, W_i)\ ]$$

The presence of the additive constant 1 inside the gradient formulation is revolutionary. It acts as an information highway, ensuring that gradients $\partial E/\partial x_L$ are directly propagated back to arbitrarily shallow layers without passing through complex, multiplicative non-linearities that cause exponential decay. This is mechanically analogous to how boosting algorithms maintain strong learning signals across thousands of sequential trees by summing their independent contributions.

C. Sub-linear Performance Scaling and Effective Depth

Empirical results across computer vision benchmarks (such as ImageNet) show that the performance of ResNets scales sub-linearly with architectural depth. Adding layers from ResNet-18 to ResNet-50 yields significant accuracy gains, but the jump from ResNet-152 to ResNet-1001 provides only marginal, highly localized improvements. This perfectly mirrors the diminishing returns universally observed in Gradient Boosting Machines when adding thousands of weak trees. The earliest trees (or shallowest layers) capture the vast majority of the data's variance and foundational structures, while later trees (or deeper layers) make increasingly minute, asymptotic refinements to the decision boundary.

IV. DISCUSSION

A. Interpretability and the Myth of Deep Hierarchies

By mathematically acknowledging that ResNets are ensembles of exponentially many shallow networks (a ResNet with L blocks contains 2^L possible paths from input to output due to the binary choice of passing through or bypassing F), we must critically reassess the standard 'deep hierarchical feature' narrative. Deeper layers in modern architectures are not necessarily finding 'shapes constructed from edges, constructed from pixels.' Instead, they are performing an iterative, unrolled refinement of a feature space established quite early in the network. This ensemble perspective provides a clear, statistically sound justification for why ultra-deep networks generalize remarkably well without violently overfitting: ensemble averaging naturally acts as a strong variance-reduction regularizer.

B. Limitations and Divergences in Optimization

While the additive structural analogy is robust, it is critical to note that ResNets are not perfectly symmetrical to classical Gradient Boosting in their optimization trajectory. In classical GBMs, weak learners are trained greedily and strictly sequentially; a tree h_m is fixed and immutable once the next tree h_{m+1} begins training. In ResNets, all 'weak learners' (residual blocks) are trained simultaneously via global backpropagation. This simultaneous, parallel optimization allows layers to co-adapt. A residual block in a ResNet does not purely fit the pseudo-residual of the past in isolation; it

dynamically adjusts based on the transformations that subsequent blocks are concurrently learning to apply. Thus, ResNets represent an implicit, jointly-optimized form of boosting, rather than a strict sequential, greedy one.

C. Real-World Implications: From CV to Large Language Models (LLMs)

The additive ensemble perspective theoretically validates explicit regularization techniques like Stochastic Depth, where entire residual blocks are randomly bypassed during training. By treating the network as a vast ensemble, dropping layers acts as a form of architectural dropout, drastically reducing training time and implicitly training an ensemble of sub-networks, much like feature and row subsampling in XGBoost.

Furthermore, this mathematical framework extends far beyond Convolutional Neural Networks. Modern Natural Language Processing (NLP) is exclusively dominated by Transformer architectures, which rely intrinsically on residual streams. The multi-head self-attention and feed-forward sub-layers in Transformers are mathematically structured exactly as $x = x + \text{SubLayer}(x)$. Therefore, the unprecedented emergent capabilities of Large Language Models (LLMs) are fundamentally anchored in the very same ensemble-like, iterative functional error-correction mechanisms discussed throughout this paper.

V. CONCLUSION

This comprehensive research paper has elaborated on the profound theoretical and structural intersections between two of the most dominant algorithms in artificial intelligence: Residual Networks and Gradient Boosting Machines. By rigorously analyzing the mathematical framework of residual mappings, unrolled feature paths, and additive functional gradient descent, we demonstrated that ResNets operate as implicit, jointly-optimized ensembles of shallow networks. The individual residual blocks function fundamentally as weak learners, incrementally refining the feature representation by correcting the representation residuals of previous layers.

This unified mathematical perspective resolves long-standing mysteries behind the resilience of ResNets to layer ablation, explains their sub-linear performance scaling, and provides clarity on their robust gradient flow. Recognizing deep networks not as fragile deep chains, but as additive ensembles, paves the way for a new era of architectural innovation. Future work must focus on designing explicit hybrid models—such as dynamic Gradient Boosted Neural Networks—that intelligently construct network depth at runtime based on the complexity of the input instance, analogous to early stopping in boosting. Bridging these paradigms promises to combine the unmatched feature-extraction capabilities of deep learning with the unparalleled tabular robustness and interpretability of ensemble methods.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition,' CVPR, pp. 770-778, 2016.

[2] J. H. Friedman, 'Greedy Function Approximation: A Gradient Boosting Machine,' The Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.

[3] A. Veit, M. J. Wilber, and S. Belongie, 'Residual Networks Behave Like Ensembles of Relatively Shallow Networks,' NeurIPS, vol. 29, pp. 1046-1054, 2016.

[4] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, 'Deep Networks with Stochastic Depth,' ECCV, pp. 646-661, Springer, 2016.

[5] T. Chen and C. Guestrin, 'XGBoost: A Scalable Tree Boosting System,' KDD, pp. 785-794, 2016.

[6] W. E, 'A Proposal on Machine Learning via Dynamical Systems,' Communications in Mathematics and Statistics, vol. 5, no. 1, pp. 1-11, 2017.

[7] Y. Freund and R. E. Schapire, 'A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,' Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119-139, 1997.

[8] A. Vaswani et al., 'Attention is All You Need,' NeurIPS, vol. 30, pp. 5998-6008, 2017.

[9] H. Wang, S. Yang, D. Liu, J. Wang, W. Dong, and F. Wu, 'Connecting Deep Networks with Ensemble Learning,' IEEE TNNLS, vol. 31, no. 12, pp. 5312-5326, 2020.

[10] N. B. Erichko, K. G. Belyaev, and A. V. Ivanova, 'Gradient Boosting and Neural Networks: A Theoretical Unification,' IEEE Access, vol. 9, pp. 125432-125445, 2021.

[11] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, 'Neural Ordinary Differential Equations,' NeurIPS, vol. 31, 2018.