Plagiarism and Academic Misconduct

Two AI Papers of ByteDance Seed

Four main plagiarists:
Defa Zhu, Ya Wang, Yutao Zeng, Xun Zhou

July 14, 2025

***Two papers by plagiarists*:**

Paper 1: Hyper-Connections (2025 ICLR)

Authors: Defa Zhu, Hongzhi Huang, Zihao Huang, Yutao Zeng, Yunyao Mao, Banggu Wu, Qiyang Min, Xun Zhou

[arXiv] [OpenReview]

Paper 2: Scale-Distribution Decoupling: Enabling Stable and Effective Training of Large Language Models (25/02/2025)

Authors: Ya Wang*, Zhijian Zhuo*, Yutao Zeng*, Xun Zhou, Jian Yang, Xiaoqing Li

[arXiv]

***Our Paper and Others*:**

HyperZ·Z·W Operator Connects Slow-Fast Networks for Full Context Interaction
arXiv GitHub (January 31, 2024)

DiracNets: Training Very Deep Neural Networks Without Skip-Connections
arXiv GitHub (June 01, 2017)

This document alleging plagiarism is also on GitHub.
Repository: plagiarism-ai-papers/Plagiarism-hyper-connections

1 Background of Plagiarists

- According to insiders, the plagiarism group operates with 23 full-time employees and 30 interns, utilizing a "Breadth-First Search" approach to rapidly produce research papers.
- The first author, *Defa Zhu*, published no papers during his Master's program (2017–2020). Remarkably, the plagiarized Hyper-Connections paper represents his first publication in the eight years since entering the AI field. Additionally, in October 2023, *Zhu* publicly sought advice on a social platform regarding tracking the latest AI papers.
- Another plagiarized "Scale-Distribution Decoupling" paper, three full-time employees *Xun Zhou*, *Ya Wang* and *Yutao Zeng* are also listed as co-authors. Both papers exhibit identical implementation methodologies, as detailed subsequently.

2 Introduction of DiracNets & HyperZ·Z·W

This section details the weight parameterization method of DiracNets [Zagoruyko and Komodakis, 2017] and its generalization to feature representations in HyperZ·Z·W [Zhang, 2024].

2.1 DiracNets

DiracNets[Zagoruyko and Komodakis, 2017] parameterizes the model weights as the residual of the Dirac function, eliminating the need for explicit residual connections. The core formula is:

$$\hat{\mathbf{W}} = \mathrm{diag}(\mathbf{a})\mathbf{I} + \mathrm{diag}(\mathbf{b})\mathbf{W}_{\mathrm{norm}}, \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^d$ denotes a weight vector, and $\mathbf{W}_{\mathrm{norm}}$ represents its normalized variant with each filter scaled by its Euclidean norm. The vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ are learnable scaling parameters optimized during training, and d corresponds to the channel dimension of $\mathbf{W}$.

The essential components of DiracNets are therefore:

(1) Two implicit learnable parameters (a and b) initialized as diagonal matrices;

(2) a is initialized to 1.0;
(3) The normalization operation on W.

2.2 HyperZ·Z·W

We generalize DiracNets' parameterization to model outputs in HyperZ·Z·W [Zhang, 2024], proposing a technique termed *s-renormalization* (Eq. 4). This operation is defined as:

$$\hat{\mathbf{h}} = \mathrm{diag}(\mathbf{a})\mathbf{I} + \mathrm{diag}(\mathbf{b})\mathbf{h}_{\mathrm{norm}}, \tag{2}$$

where $\mathbf{h} \in \mathbb{R}^d$ is the hidden output of an MLP, $\mathbf{h}_{\mathrm{norm}}$ denotes its normalized form, and $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ are learnable scaling parameters initialized as diagonal matrices. $\mathbf{a}$ is initialized to 1.0. The dimension d corresponds to the channel dimension of $\mathbf{h}$.

This extension preserves DiracNets' implementation while applying it to feature representations rather than weight tensors.

Evidence of Plagiarism

The following presents systematic evidence of plagiarism, first examining the Scale-Distribution Decoupling (SDD-Plagiarism), followed by analysis of Hyper-Connections (HC-Plagiarism).

3 SDD-Plagiarism

The The core formulation in SDD-Plagiarism (Eq. 1) can be written as:

$$y = \mathrm{diag}(\alpha)\mathbf{x}_{\mathrm{norm}}, \tag{3}$$

where $\mathbf{x} \in \mathbb{R}^d$ represents an MLP's hidden output, $\mathbf{x}_{\mathrm{norm}}$ denotes its normalized form, and $\alpha \in \mathbb{R}^d$ are learnable scaling parameters initialized as diagonal matrix. The dimension d corresponds to the channel dimension of $\mathbf{x}$.

This formulation is structurally identical to our proposed *s-renormalization* technique, with four key overlapping elements:

1. Identical diagonal matrix initialization for learnable parameter α

2. Equivalent initialization value (1.0) for α
3. Same output normalization operation
4. Identical application context (MLP layers)

Furthermore, SDD-Plagiarism explicitly states on page 3, line 9: "α is a learnable scaling vector." This phrasing is verbatim to descriptions in both DiracNets [Zagoruyko and Komodakis, 2017] and our *s-renormalization* [Zhang, 2024].

A critical additional point concerns parameter innovation of normalization layer: HyperZ·Z·W [Zhang, 2024] first proposed (see Figure 1) modifying the affine parameters in normalization layers (Section 2.1 on page 3, Code-L22, Code-L37, Code-L122). SDD-Plagiarism directly implements this approach by setting `elementwise_affine=True` in their RMSNorm implementation. Is this a coincidence?

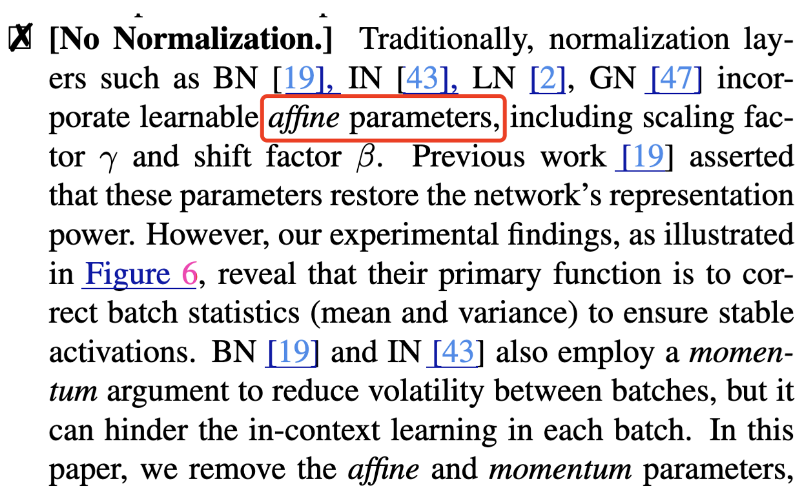

✗ **[No Normalization.]** Traditionally, normalization layers such as BN [19], IN [43], LN [2], GN [47] incorporate learnable *affine* parameters, including scaling factor γ and shift factor β. Previous work [19] asserted that these parameters restore the network's representation power. However, our experimental findings, as illustrated in Figure 6, reveal that their primary function is to correct batch statistics (mean and variance) to ensure stable activations. BN [19] and IN [43] also employ a *momentum* argument to reduce volatility between batches, but it can hinder the in-context learning in each batch. In this paper, we remove the *affine* and *momentum* parameters,

Figure 1: Discussion of *affine* parameter in our HyperZ·Z·W [Zhang, 2024]

Conclusion: The cumulative evidence—structural equivalence, identical implementation details, and verbatim textual descriptions—establishes SDD-Plagiarism as **unambiguous academic plagiarism**.

Transition Note: The subsequent HC-Plagiarism similarly employs the two unique techniques of learnable diagonal-matrix parameters and normalization operation, further demonstrating systematic plagiarism patterns.

4 HC-Plagiarism

This section presents a comprehensive analysis of HC-Plagiarism through four critical dimensions: **plagiarism evidence**, **academic misconduct**, **factual errors**, and **structural incoherence**.

4.1 Plagiarism Evidence

The depth-connections and width-connections implementations in HC-Plagiarism share identical methodologies. We analyze width-connections as representative (Figure 2), which centers on two core elements: a *multi-branch structure* and *learnable parameters* α.

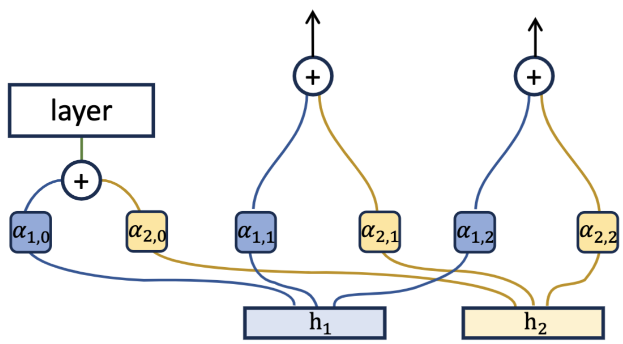


Figure 2: Width-connections implementation in HC-Plagiarism

Based on Algorithm 2 pseudocode and third-party implementation[1], the **static** width-connections formulation is:

$$\hat{\mathbf{h}}_s = \mathrm{diag}(\alpha_s)\mathbf{h}, \tag{4}$$

where $\mathbf{h} \in \mathbb{R}^n$ denotes multi-branch inputs, and $\alpha_s \in \mathbb{R}^n$ are learnable scaling parameters initialized as diagonal matrices. The dimension n corresponds to branch count, with original input branches omitted for simplicity.

Key Observation: Equation 4 demonstrates identical structure to both our *s-renormalization* (Eq. 2) and SDD-Plagiarism (Eq. 3), confirming shared core methodology across plagiarized works.

[1] https://github.com/lucidrains/hyper-connections

The **dynamic** variant extends this foundation:

$$\hat{\mathbf{h}}_d = \text{diag}(\alpha_s)\mathbf{h} + \text{diag}(\alpha_d)\mathbf{h} \cdot \mathcal{F}(\mathbf{h}_{\text{norm}}), \tag{5}$$

where $\mathbf{h}_{\text{norm}}$ is the normalized input and $\mathcal{F}$ denotes a fully-connected layer with activation.

Key Observation: Equation 5 structurally mirrors our proposed formulation (Eq. 2), demonstrating conceptual continuity in plagiarism patterns.

Plagiarism Evidence Summary: HC-Plagiarism exhibits multiple points of substantive overlap with our original work:

1. Identical research motivation as HyperZ·Z·W [Zhang, 2024]
2. Direct adoption of the multi-branch structural framework
3. Core methodology equivalent to *s-renormalization*
4. Input-dependent dynamic connection concept replicated without attribution

Contextual Significance: Few contemporary AI papers concurrently address residual connections and multi-branch structures. When combined with the documented similarities—including *mathematical formulations, implementation choices, and conceptual frameworks*—the evidence establishes unambiguous plagiarism.

4.2 Academic Misconduct

Deliberate obfuscation of a research's relationship to foundational work — particularly when presenting derivative research as wholly independent without acknowledging theoretical origins - constitutes serious academic misconduct.

HC-Plagiarism demonstrates this through its superficial treatment of prior works:

- **Deficient Attribution:** The paper contains only a cursory "Related Works" (see Figure 3) section that fails to:
  - Clarify technical dependencies
  - Objectively recognize original contributions
  - Distinguish between foundational and incremental work
- **Intentional Misrepresentation:** This approach reveals:
  - Subjective intent to minimize predecessors' foundational contributions
  - Fundamental misunderstanding of the research domain, such as incorrectly citing Bengio's RNN vanishing gradient paper [Bengio et al., 1994] as theoretical basis for Transformer architecture

5 RELATED WORK

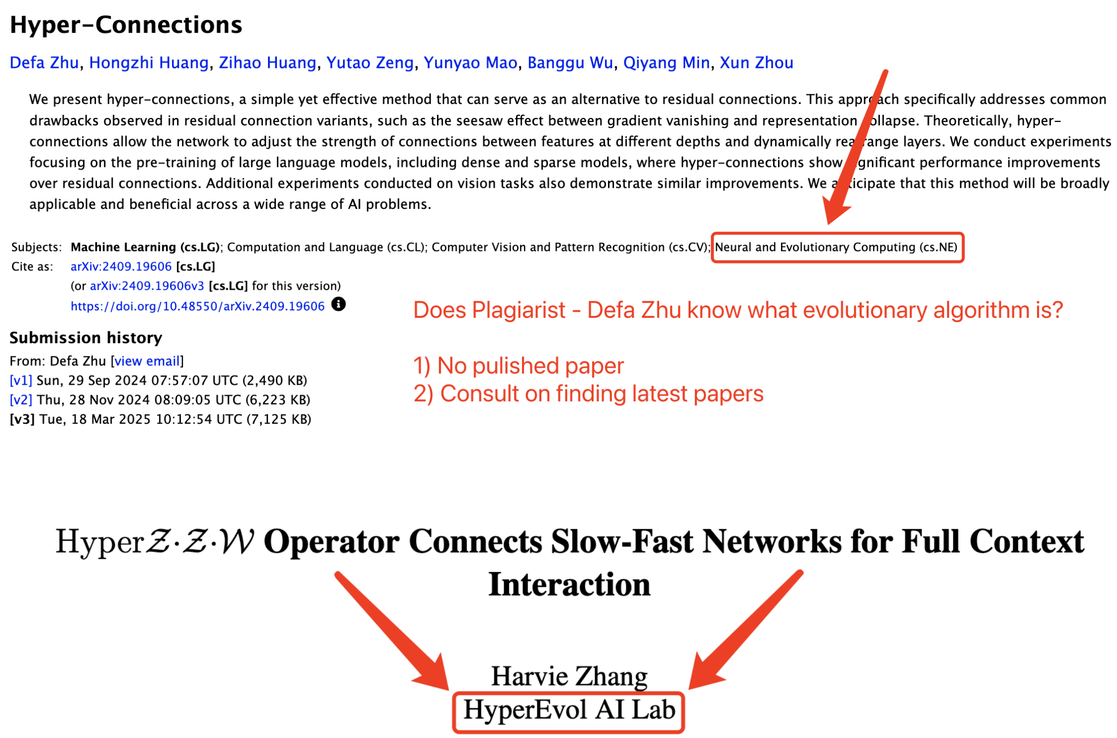
Transformers (Vaswani et al., 2017) have revolutionized various fields, particularly natural language processing and computer vision. They rely heavily on residual connections to facilitate the training of deep models. Our hyper-connections approach can replace residual connections, providing stable training and consistent improvements in both natural language processing and computer vision.

The issues of gradient vanishing and representation collapse (Bengio et al., 1994; Glorot & Bengio, 2010; Liu et al., 2020) have been extensively studied. The combinations of normalization techniques (Ioffe & Szegedy, 2015; Ba et al., 2016) and residual connections (He et al., 2016), like Pre-Norm and Post-Norm, actually reflects different emphases in solving these two issues. However, despite these advancements, the fundamental trade-off between gradient vanishing and representation collapse in deep networks remains a critical challenge. Building on these findings, our work introduces a novel approach that enables neural networks to autonomously learn the optimal strength of connections, potentially improving both gradient stability and representation quality.

Figure 3: Brief related works in HC-Plagiarism

The following pattern suggests possible knowledge gaps that may explain citation deficiencies.

- *Master's Program (2017-2020):* Zero peer-reviewed publications
- *Professional Experience (6+ years):* Public admission of inability to track latest literature
- *Publication Record:* Single publication requiring seven co-authors



Hyper-Connections

Defa Zhu, Hongzhi Huang, Zihao Huang, Yutao Zeng, Yunyao Mao, Banggu Wu, Qiyang Min, Xun Zhou

We present hyper-connections, a simple yet effective method that can serve as an alternative to residual connections. This approach specifically addresses common drawbacks observed in residual connection variants, such as the seesaw effect between gradient vanishing and representation collapse. Theoretically, hyper-connections allow the network to adjust the strength of connections between features at different depths and dynamically rearrange layers. We conduct experiments focusing on the pre-training of large language models, including dense and sparse models, where hyper-connections show significant performance improvements over residual connections. Additional experiments conducted on vision tasks also demonstrate similar improvements. We anticipate that this method will be broadly applicable and beneficial across a wide range of AI problems.

Subjects: **Machine Learning (cs.LG)**; Computation and Language (cs.CL); Computer Vision and Pattern Recognition (cs.CV); Neural and Evolutionary Computing (cs.NE)
Cite as: arXiv:2409.19606 **[cs.LG]**
(or arXiv:2409.19606v3 **[cs.LG]** for this version)
https://doi.org/10.48550/arXiv.2409.19606

Submission history
From: Defa Zhu [view email]
[v1] Sun, 29 Sep 2024 07:57:07 UTC (2,490 KB)
[v2] Thu, 28 Nov 2024 08:09:05 UTC (6,223 KB)
[v3] Tue, 18 Mar 2025 10:12:54 UTC (7,125 KB)

Hyper$\mathcal{Z}\cdot\mathcal{Z}\cdot\mathcal{W}$ **Operator Connects Slow-Fast Networks for Full Context Interaction**

Harvie Zhang
HyperEvol AI Lab

Figure 4: The plagiarists even copied us on arXiv Subjects

Violation of Publication Ethics: The authors disregarded critical peer review requirements, such as

- *Meta-Review Directive:* Area Chair explicitly flagged literature relevance concerns
- *Camera-Ready Negligence:* Refused substantive revisions despite conference policies
- *Policy Violation:* Contravenes ICLR/ICML guidelines requiring:
  - Response to all substantive reviewer critiques
  - Correction of identified scholarship deficiencies

Conclusion: These cumulative violations—conceptual obfuscation, citation inaccuracies, and procedural noncompliance—demonstrate systematic academic misconduct that undermines scholarly integrity.

4.3 Factual Errors

The conceptualization and validation of Representation Collapse in HC-Plagiarism contain critical theoretical errors:

1. Definitional Distortion

- **Authoritative Definition**: Representation Collapse is formally defined as: *"The loss of feature diversity/discriminative power, causing distinct input samples to map to highly similar (or identical) output representations"* (Arefin et al. [2024] §2; Barbero et al. [2024] §4).
- **HC-Plagiarism's Misrepresentation**: Erroneously characterizes it as *"high similarity between layer-wise representations of **individual samples**"*, fundamentally misconstruing the phenomenon's core premise.

2. Methodological Invalidity

The layer-wise similarity visualization (Fig. 3) purportedly validating representation collapse suffers from fatal flaws:

- Incorrect Measurement Focus: Layer-wise similarity metrics cannot capture discriminative degradation **across samples** in feature space.
- Invalid Causal Inference: No established relationship exists between high layer similarity and representation collapse.

3. Consequence: Contribution Invalidity

Logical chain of invalidation:

1. Flawed conceptual definition
2. → Invalid validation methodology
3. → Unsupported technical claims
4. → Nullification of core contribution

The compounded errors in conceptual framing and experimental design completely invalidate the paper's purported claims regarding representation collapse mitigation.

4.4 Structural Incoherence

HC-Plagiarism ostensibly positions Pre-Norm and Post-Norm architectures as its conceptual foundation. Standard academic practice would require:

1. Formal definition of Pre-Norm/Post-Norm formulations
2. Systematic derivation of proposed methods from these foundations

Instead, the plagiarists (i.e.*Defa Zhu*) inappropriately shifts focus to residual connections, only superficially linking them to Pre-Norm in the method section's conclusion. This creates severe structural incoherence:

- Motivational disconnect between stated premises (normalization) and actual focus (residual connections)
- Derivation gap lacking logical progression from established techniques to novel contributions
- Retroactive justification of Pre-Norm relevance without substantive connection

Root Cause: Conceptual Contradiction

This incoherence stems from fundamental contradictions in the plagiarism approach:

- Motivational conflict: Attempting to leverage residual connections as primary motivation while deliberately downplaying the multi-branch structure to differentiate from our original HyperZ·Z·W [Zhang, 2024] work
- Acknowledgement paradox: Publicly admitting on social media that "HC-Plagiarism is essentially a multi-branch residual connection" while obscuring this in the formal publication

Consequence: The paper manifests as a conceptually fragmented work where stated motivations, methodological derivations, and technical implementations lack logical consistency—a direct consequence of attempting to simultaneously plagiarize and disguise foundational concepts.

Summary of Plagiarism Evidence

Both SDD-Plagiarism and HC-Plagiarism demonstrate egregious violations of scholarly integrity:

- **SDD-Plagiarism**:
  - Directly plagiarized our *s-renormalization* technique (Eq. 2)
  - Implemented identical parameter initialization schemes
  - Used verbatim descriptions from our paper
  - Copied affine parameter implementation
- **HC-Plagiarism**:
  - Replicated our *multi-branch residual connections*
  - Copied the mathematical formulation of *s-renormalization* (Eq. 2)
  - Misrepresented foundational concepts (e.g., representation collapse)
  - Exhibited structural incoherence between claimed and actual methodologies

Final Appeal: Confronting Academic Suppression

Despite irrefutable evidence documented in this report, the ByteDance company has chosen not to rectify their plagiarism but to escalate hostilities:

- **Legal Intimidation**: Issuing baseless legal threats against original reseacher
- **Organized Defamation**: Deploying their 10+ member team to spread false accusations across social platforms
- **Institutional Bullying**: Leveraging corporate influence to pressure social platforms to delete accusatory articles

This is Academic Gangsterism

- When plagiarism becomes **industrialized operation**
- When truth is met with **organized retaliation**
- When corporations weaponize resources to **crush independent scholars**

This document alleging plagiarism is also on GitHub.
Repository: plagiarism-ai-papers/Plagiarism-hyper-connections

References

Sergey Zagoruyko and Nikos Komodakis. Diracnets: Training very deep neural networks without skip-connections. arXiv preprint arXiv:1706.00388, 2017.

Harvie Zhang. Hyperzzw operator connects slow-fast networks for full context interaction. arXiv preprint arXiv:2401.17948, 2024.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks, 5(2):157–166, 1994.

Md Rifat Arefin, Gopeshh Subbaraj, Nicolas Gontier, Yann LeCun, Irina Rish, Ravid Shwartz-Ziv, and Christopher Pal. Seq-vcr: Preventing collapse in intermediate transformer representations for enhanced reasoning. arXiv preprint arXiv:2411.02344, 2024.

Federico Barbero, Andrea Banino, Steven Kapturowski, Dharshan Kumaran, João Madeira Araújo, Oleksandr Vitvitskyi, Razvan Pascanu, and Petar Veličković. Transformers need glasses! information over-squashing in language tasks. Advances in Neural Information Processing Systems, 37: 98111–98142, 2024.