Hyper-SET: Designing Transformers via Hyperspherical Energy Minimization

Yunzhe Hu[1], Difan Zou[1,2], and Dong Xu[1]

[1]The University of Hong Kong, [2]Shenzhen Loop Area Institute

SCHOOL OF COMPUTING & DATA SCIENCE
The University of Hong Kong

深圳河套學院
Shenzhen Loop Area Institute

ICLR
International Conference On Learning Representations

Project Page w/ Code

Full Paper

Problem

Transformers are great, but…

- **Engineered from the bottom up.** Their architecture remains largely heuristics-driven—key components are arranged by trial and error.
- **Mysteriously redundant.** Evidence that representations are similar in the middle layers of LLMs suggests a convergent layer functionality.
- **Mostly interpreted post hoc.** Current tools to interpret their inner workings (e.g., SAEs, circuit analysis) are mostly with hindsight—hard to break the performance ceiling.

What we do

A *top-down* approach by asking:

Can we find or design a function prior that induces a model interpretable by construction?

Our response: We think the answer is **YES**, at least for a family of Transformers.

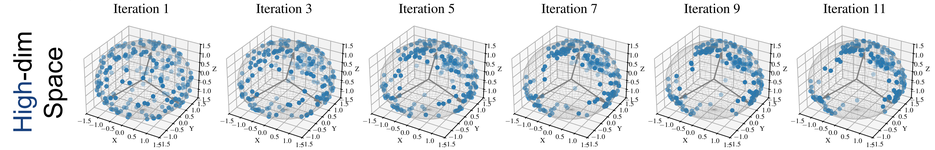
Main result: We introduce *Hyper-SET*, an **intrinsically interpretable** Transformer where every core component—from self-attention to skip connections—emerges naturally from a single, principled objective:

<u>maximum likelihood estimation on the hypersphere</u>

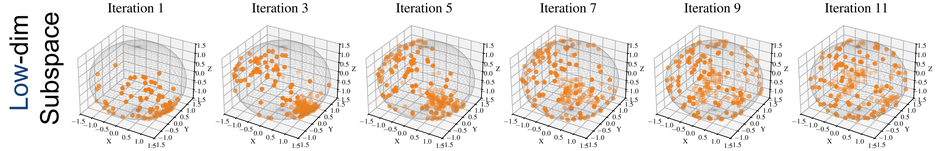
Conceptualization

We begin by conceptualizing what effective representations should look like. We posit that token dynamics should satisfy two complementary properties:

- **Semantic Alignment** aligns tokens with learned semantic directions to compress uninformative redundancy.



- **Distributional Uniformity** prevents representation collapse and ensures tokens spread out as isotropic Gaussian on the sphere to preserve volume.



Objective-driven Architecture

(Joint MLE)

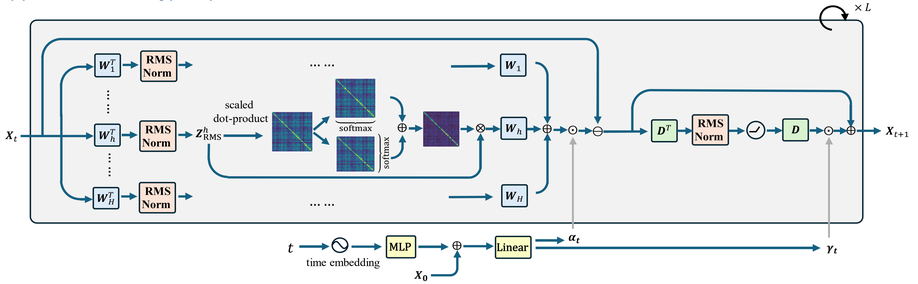
$$\max_{\boldsymbol{x}} \quad \mathbb{E}_{(\boldsymbol{z}^1,\ldots,\boldsymbol{z}^H)\sim p(\boldsymbol{z}^1,\ldots,\boldsymbol{z}^H)}\left[\log p(\boldsymbol{x},\boldsymbol{z}^1,\ldots,\boldsymbol{z}^H;\theta,\phi)\right]$$

$$\min_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\in\boldsymbol{X}} E(\boldsymbol{X};\boldsymbol{W},\boldsymbol{D}) = E_{\text{ATTN}} + E_{\text{FF}},$$
$$\text{subject to} \quad \|\boldsymbol{W}_h^\top \boldsymbol{x}_i\|_2 = \sqrt{p}, \quad \|\boldsymbol{D}^\top \boldsymbol{x}_i\|_2 = \sqrt{M}, \quad i = 1,\ldots,N.$$

Isotropic Gaussian supported on the hypersphere

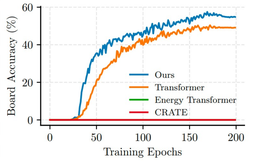
$$\min_{\boldsymbol{x}} \sum_{h=1}^{H} \underbrace{D_{\text{KL}}\left(p(\boldsymbol{z}) \| p_\phi(\boldsymbol{z}^h \mid \boldsymbol{x})\right)}_{\text{uniformity}} \underbrace{- \log p_\theta(\boldsymbol{x})}_{\text{alignment}}$$

Quantified into optimizable, modified Hopfield energy functions

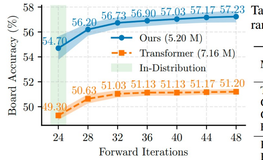


Empirical Results

a) Sudoku Reasoning



(a) Sudoku training dynamics.



(b) Sudoku test-time extrapolation.

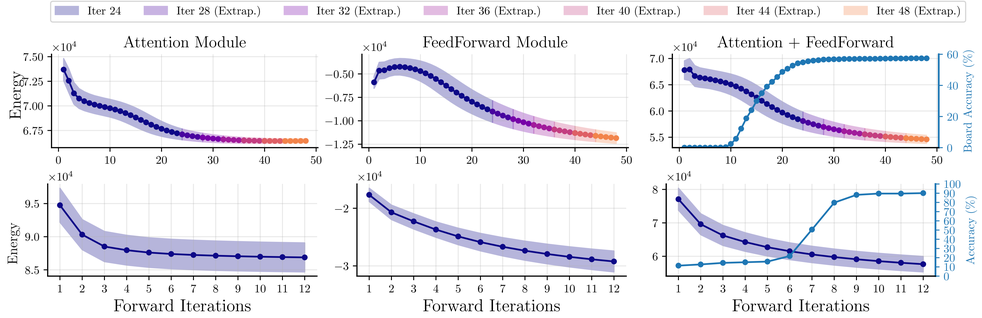
b) Image Classification

Table 1: Top-1 accuracy (%) for image classification with single-layer recurrent-depth models. Parameters are measured on ImageNet-1K. All models are trained from scratch on the listed datasets.

Model	Width d	# Params (M)	Dataset			
			CIFAR-10	CIFAR-100	IN-100	IN-1K
Transformer	384	2.38	89.90	61.89	69.44	**66.94**
CRATE-T (Hu et al., 2024c)	896	3.04	87.54	60.23	68.16	57.89
CRATE (Yu et al., 2023)	768	3.00	84.81	58.22	68.52	57.00
Energy Transformer (Hoover et al., 2024)	512	2.39	76.39	50.60	36.68	34.24
HYPER-SET (Ours)	512	2.39	**90.11**	63.41	**70.16**	62.76
HYPER-SET (Ours)	640	3.40	89.96	**64.60**	69.31	66.21

c) Energy Evolution

Energy Descent *positively* correlates with increase in accuracy



Extensibility

We are **NOT** re-interpreting existing Transformers! but instead providing a general design framework. Designing energy functions unlocks novel variants (e.g., **linear attention**).

Operator	$f(x)$	$K(\boldsymbol{x},\boldsymbol{y})$	E_{ATTN}	$-\nabla_{\boldsymbol{X}} E_{\text{ATTN}}$
Bi-Softmax (Default)	$\beta^{-1}\log(x)$	$\exp(\beta \boldsymbol{x}^\top \boldsymbol{y})$	Eq. 3	Eq. 7
Sigmoid Attention	$\frac{\beta^{-1}}{2}x$	$\sigma(\beta \boldsymbol{x}^\top \boldsymbol{y})$	$\frac{1}{2}\sum_{h=1}^{H}\sum_{i,j=1}^{N}\sigma\left(\beta(\boldsymbol{W}_h^\top \boldsymbol{x}_i)^\top \boldsymbol{W}_h^\top \boldsymbol{x}_j\right)\beta^{-1}$	$\sum_{h=1}^{H}\boldsymbol{W}_h\boldsymbol{W}_h^\top \boldsymbol{X}\sigma(1-\sigma)\left(\beta(\boldsymbol{W}_h^\top \boldsymbol{X})^\top \boldsymbol{W}_h^\top \boldsymbol{X}\right)$
Linear Attention	$\frac{\beta^{-1}}{2}x$	$\frac{1}{2}\left(\beta\Phi(\boldsymbol{x})^\top\Phi(\boldsymbol{y})\right)^2$	$\frac{1}{4}\sum_{h=1}^{H}\sum_{i,j=1}^{N}\left(\beta\Phi(\boldsymbol{W}_h^\top \boldsymbol{x}_i)^\top\Phi(\boldsymbol{W}_h^\top \boldsymbol{x}_j)\right)^2\beta^{-1}$	$\sum_{h=1}^{H}\boldsymbol{W}_h\Phi'(\boldsymbol{W}_h^\top \boldsymbol{X})\odot\left(\beta\Phi(\boldsymbol{W}_h^\top \boldsymbol{X})\Phi(\boldsymbol{W}_h^\top \boldsymbol{X})^\top\Phi(\boldsymbol{W}_h^\top \boldsymbol{X})\right)$

Summary

- Frames representation learning as <u>joint maximum likelihood estimation</u> on hyperspheres
- Bridges the gap between <u>energy-based learning</u> and practical Transformer design
- Provides a <u>general design principle</u> that unlocks novel variants in core Transformer block