ICLR

Hyper-SET: Designing Transformers via Hyperspherical Energy Minimization

Yunzhe Hu, Difan Zou, and Dong Xu

The University of Hong Kong

SCHOOL OF
COMPUTING &
DATA SCIENCE
The University of Hong Kong

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 believe 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 symmetric attention to skip connections—emerges naturally from a single, principled objective:

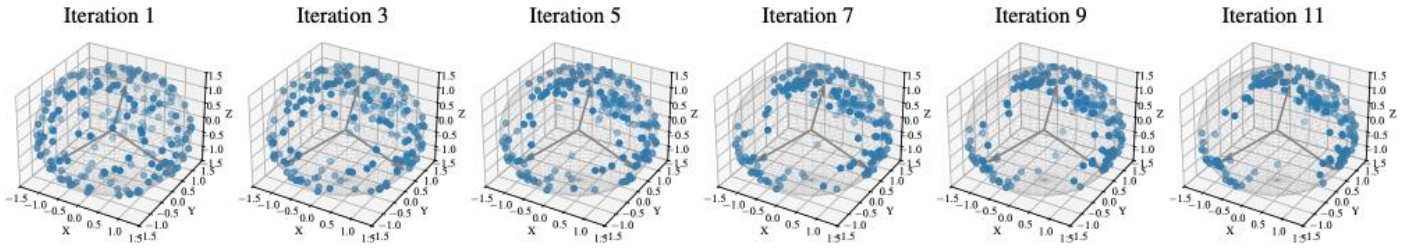
<u>maximum likelihood estimation on the hypersphere.</u>

Conceptualization

We conjecture that token dynamics should satisfy two complementary properties:

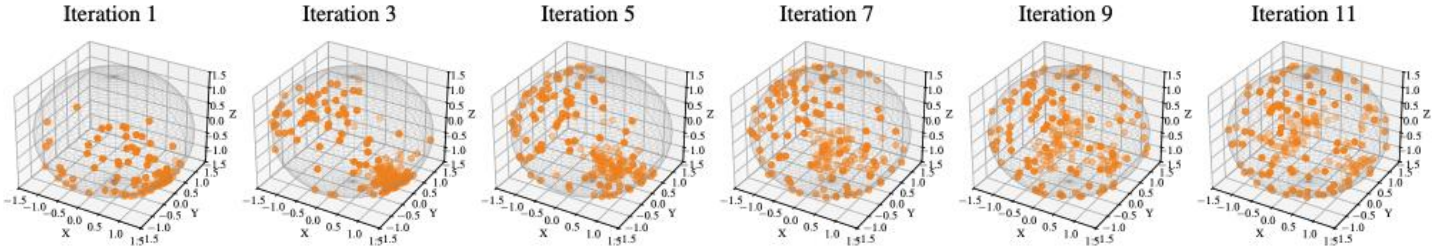
- **Semantic Alignment** aligns tokens with learned semantic directions to compress uninformative redundancy.

High-dim. Space



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

Low-dim. Subspace



Theoretical Justification

1. Joint maximum likelihood on the hypersphere (under mild assumptions):

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

2. Two complementary objectives:

Isotropic Gaussian supported on hypersphere

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

3. Quantified into optimizable, modified Hopfield energy functions:

$$\min_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N \in \boldsymbol{X}} E(\boldsymbol{X}; \boldsymbol{W}, \boldsymbol{D}) = E_{\mathrm{ATTN}} + E_{\mathrm{FF}},$$

$$\text{subject to} \quad \|\boldsymbol{W}_h^\top \boldsymbol{x}_i\|_2 = \sqrt{p}, \quad \|\boldsymbol{D}^\top \boldsymbol{x}_i\|_2 = \sqrt{M}, \quad i = 1, \ldots, N.$$

Hyper-SET Architecture

Symmetric self-attention:

$$\dot{\boldsymbol{X}} = -\nabla_{\boldsymbol{X}} E_{\text{ATTN}}$$

ODE discretization

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \alpha_t \sum_{h=1}^{H} \left(\boldsymbol{W}_h \boldsymbol{Z}^h_{\text{RMS},t} \underbrace{\text{softmax}}_{\text{column-wise}} \left([\boldsymbol{Q}\boldsymbol{K}]_{\text{RMS,t}} \right) + \boldsymbol{W}_h \boldsymbol{Z}^h_{\text{RMS},t} \underbrace{\text{softmax}}_{\text{row-wise}} \left([\boldsymbol{Q}\boldsymbol{K}]_{\text{RMS,t}} \right) \right)$$

Symmetric feedforward:

$$\dot{\boldsymbol{X}} = -\nabla_{\boldsymbol{X}} E_{\text{FF}} = \boldsymbol{D}\,\text{ReLU}\left(\boldsymbol{D}^\top \boldsymbol{X}\right)$$

ODE discretization

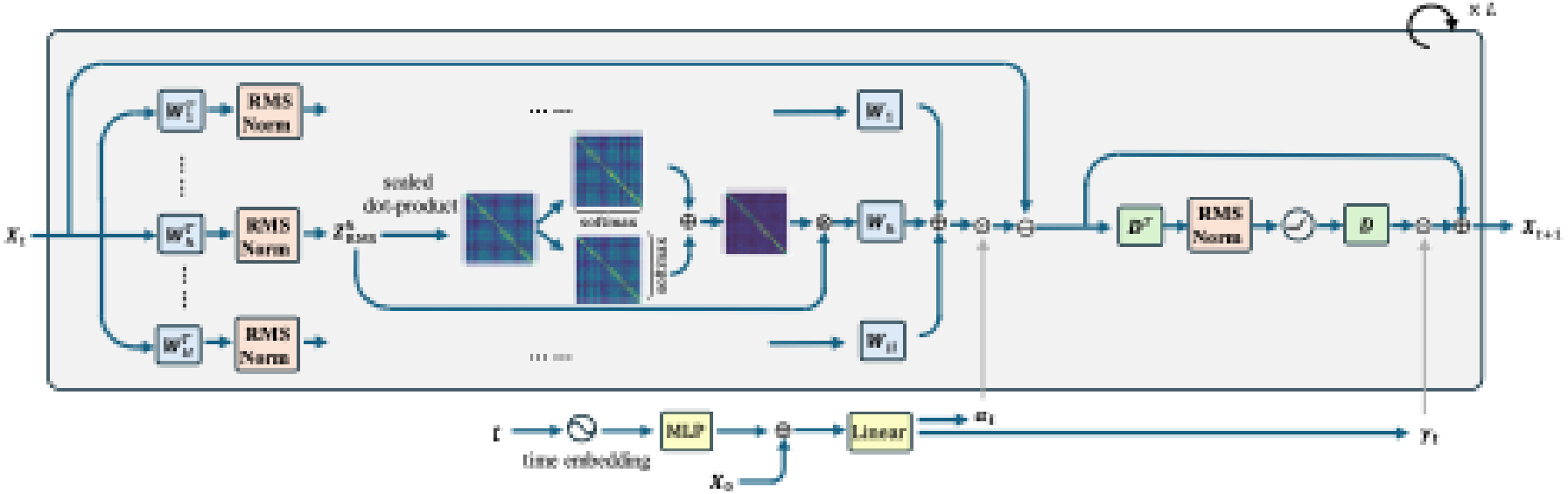
$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t + \gamma_t \boldsymbol{D}\,\text{ReLU}\left(\text{RMSNorm}\left(\boldsymbol{D}^\top \boldsymbol{X}_t\right)\right)$$

where **RMSNorm** emerges as spherical constraints

$$\boldsymbol{Z}^h_{\text{RMS}} = \text{RMSNorm}(\boldsymbol{Z}^h) = \text{RMSNorm}(\boldsymbol{W}_h^\top \boldsymbol{X})$$

$$[\boldsymbol{Q}\boldsymbol{K}]_{\text{RMS,t}} = \beta(\boldsymbol{Z}^h_{\text{RMS},t})^\top (\boldsymbol{Z}^h_{\text{RMS},t})$$

Hyper-SET Architecture



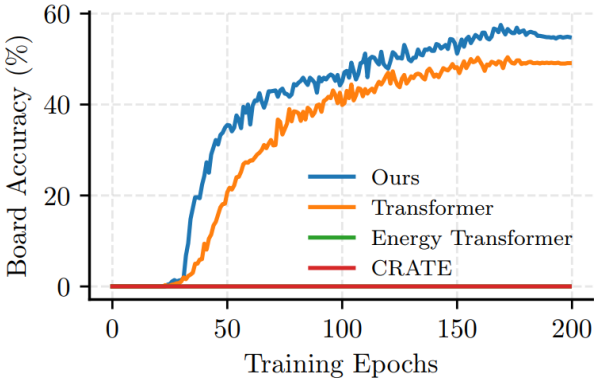
Hyper-SET layer with clear, built-in functional interpretation for representation learning

- forward pass as energy descent on the hypersphere
- recurrent-depth where shared weights indicate unique landscape

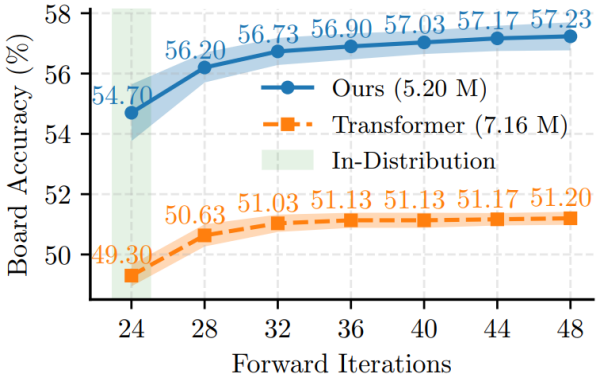
Practical Competitiveness, Validation & Extensibility

Sudoku Reasoning

- Train / Test



(a) Sudoku training dynamics.



(b) Sudoku test-time extrapolation.

Image Classification

- ImageNet-1K

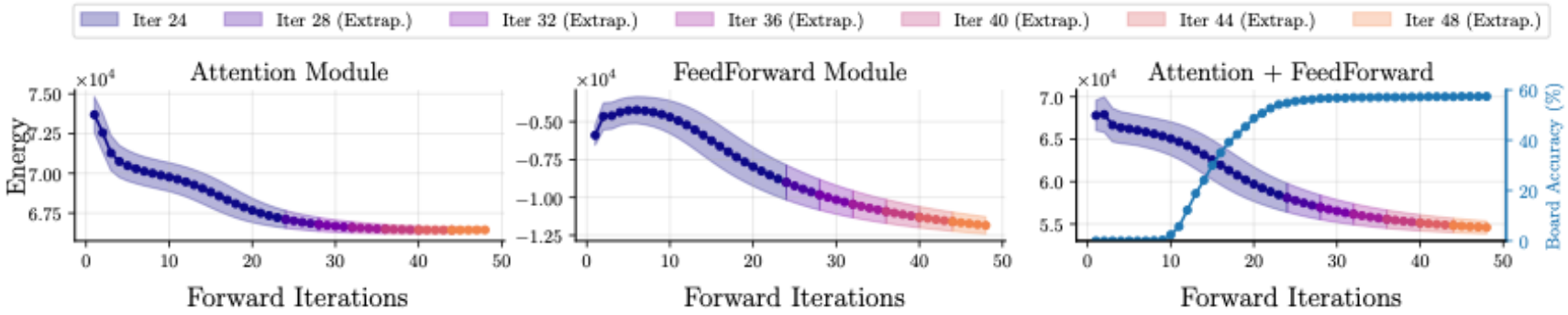
Table 1: Top-1 accuracy (%) for image classification with single-layer recurrent-depth models. Parameters are measured on ImageNet-1K. All models are trained from scratch on the listed datasets.

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

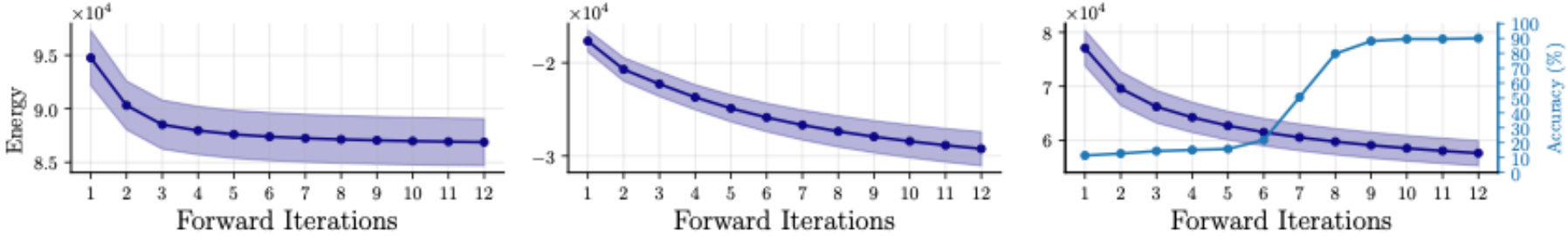
Practical Competitiveness, Validation & Extensibility

Energy Descent *positively* correlates with increase in accuracy

- Sudoku



- Image Classification



Practical Competitiveness, Validation & Extensibility

Extensibility: we are NOT re-interpreting existing Transformers!! but rather providing a more general design framework.

By defining various energy functions under our formalism, we can design novel components beyond what's widely known, like **linear attention** and **gated feedforward**.

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	$\frac{\beta^{-1}}{2}x$	$\frac{1}{2}\left(\beta\Phi(\boldsymbol{x})^\top\Phi(\boldsymbol{y})\right)^2$	$\frac{1}{4}\sum_{h=1}^{H}\sum_{i,j=1}^{N}\left(\beta\Phi(\boldsymbol{W}_h^\top \boldsymbol{x}_i)^\top\Phi(\boldsymbol{W}_h^\top \boldsymbol{x}_j)\right)^2\beta^{-1}$	$\sum_{h=1}^{H}\boldsymbol{W}_h\Phi'(\boldsymbol{W}_h^\top \boldsymbol{X})\odot\left(\beta\Phi(\boldsymbol{W}_h^\top \boldsymbol{X})\Phi(\boldsymbol{W}_h^\top \boldsymbol{X})^\top\Phi(\boldsymbol{W}_h^\top \boldsymbol{X})\right)$

Operator	$g(x)$	$h(x)$	E_{FF}	$-\nabla_{\boldsymbol{X}} E_{\text{FF}}$
ReLU FF (Default)	x	$\frac{1}{2}\text{ReLU}^2(x)$	Eq. 5	Eq. 10
Softmax FF	$\log(x)$	$\exp(x)$	$-\sum_{i=1}^{N}\log\left(\sum_{m=1}^{M}\exp(\boldsymbol{d}_m^\top \boldsymbol{x}_i)\right)$	$\boldsymbol{D}\underbrace{\text{softmax}}_{\text{column-wise}}(\boldsymbol{D}^\top \boldsymbol{X})$
Gated FF	$\frac{1}{2}x^2$	$\Phi(x)$	$-\frac{1}{2}\sum_{i=1}^{N}\left(\sum_{m=1}^{M}\Phi\left(\boldsymbol{d}_m^\top \boldsymbol{x}_i\right)\right)^2$	$\boldsymbol{D}\underbrace{\Phi(\boldsymbol{D}^\top \boldsymbol{X})}_{\text{column sum}}\odot\Phi'(\boldsymbol{D}^\top \boldsymbol{X})$

Summary & What's next

Hyper-SET

- frames representation learning as joint maximum likelihood estimation on hyperspheres
- bridges the gap between energy-based learning and practical Transformer design
- provides a general design principle that unlocks novel variants in core Transformer block

Looking ahead:

- How to extend it to autoregressive modeling?
- Can it connect to flow matching to discover new (layer-wise) training strategies?

Project Page

Full Paper

Project Page: hyper-set.github.io

Github Repo: https://github.com/huyunzhe/hyper-set