Towards Understanding and Designing Transformers: From Sparse Rate Reduction to Generalized Hopfield Energy

Yunzhe Hu

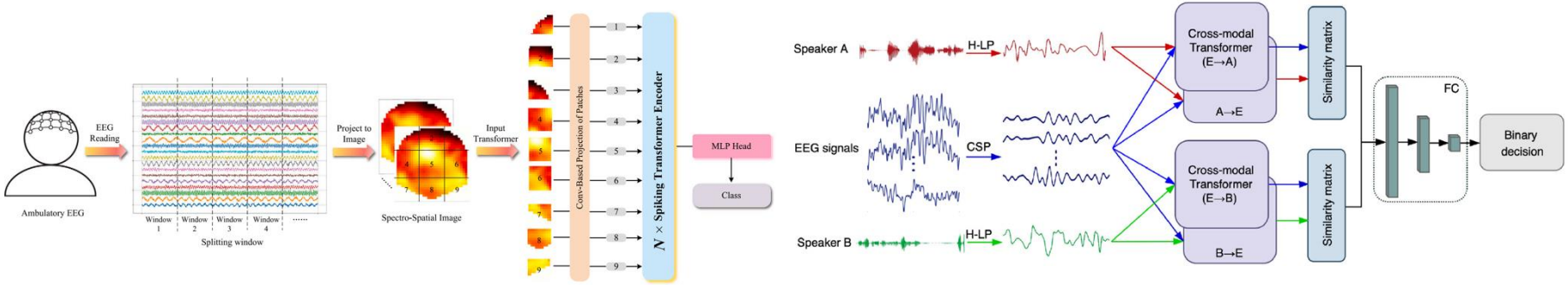
The University of Hong Kong

CRF Yearly Workshop, June 23 2025

Transformers for EEG-based Auditory Attention Decoding

- Traditional signal processing methods, while powerful, sometimes struggle with the non-linear, high-dimensional nature of neural data.

- Transformers have shown promise here due to their long-range modeling capability.



Lan et al., Low-power and lightweight spiking transformer for EEG-based auditory attention detection, Neural Network, 2025
Cai et al., Auditory Attention Detection via Cross-Modal Attention, Frontiers in Neuroscience, 2021

Motivation & Problem Statement

- Issues with current Transformer-based systems:
  - **Heuristic Design**: Building layers often engineered from bottom-up, lacking principled derivation
  - **Redundancy**: Representation uniformity, suggesting convergent functionality that one layer represents
  - **Post hoc Interpretability**: Exist techniques often mechanistic and phenomenological, hardly inform practices

- *Can we find or design a function prior that induces a model that is interpretable by construction?*
  - Our quest for top-down designed, intrinsically interpretable models

Today's Talk

Part I: Investigating Sparse Rate Reduction (SRR): An Information-Theoretic Perspective

Part II: Hyper-spherical Energy-Informed Design: A more Principled Energy-based Approach

Today's Talk

Part I: Investigating Sparse Rate Reduction (SRR): An Information-Theoretic Perspective

Part II: Hyper-spherical Energy-Informed Design: A more principled Energy-based Approach

Background

Traditional View: visualizing attention maps, inspecting weight matrices, mechanistic interpretability, etc. (don't know what to do about them)

New focus: Theoretical connections to mathematically describable concepts.

- Differential Optimization as Layers

$$z_{i+1} = \underset{z}{\text{argmin}} \ \frac{1}{2} z^T Q(z_i) z + q(z_i)^T z$$
$$\text{subject to} \ A(z_i) z = b(z_i)$$
$$G(z_i) z \leq h(z_i)$$

OptNet
[Amos et al. ICML 2017]

- Algorithm Unrolling into Layers

$$\min_Z E_{W_d}(X, Z) = \frac{1}{2} \|X - W_d Z\|_2^2 + \alpha \|Z\|_1$$

Algorithm 1 ISTA

function ISTA(X, Z, W_d, α, L)
  Require: $L >$ largest eigenvalue of $W_d^T W_d$.
  Initialize: $Z = 0$,
  repeat
    $Z = h_{(\alpha/L)}(Z - \frac{1}{L} W_d^T (W_d Z - X))$
  until change in Z below a threshold
end function

Learned ISTA
[Gregor et al. ICML 2010]

Sparse Rate Reduction: An Information-Theoretic Objective

Sparse Rate Reduction (SRR):

$$\max_{\boldsymbol{Z}\in\mathbb{R}^{d\times N}} R(\boldsymbol{Z}) - R^c(\boldsymbol{Z};\boldsymbol{U}) - \lambda\|\boldsymbol{Z}\|_0$$

$$R(\boldsymbol{Z}) \doteq \tfrac{1}{2}\log\det\left(\boldsymbol{I} + \tfrac{d}{N\epsilon^2}\boldsymbol{Z}^T\boldsymbol{Z}\right)$$

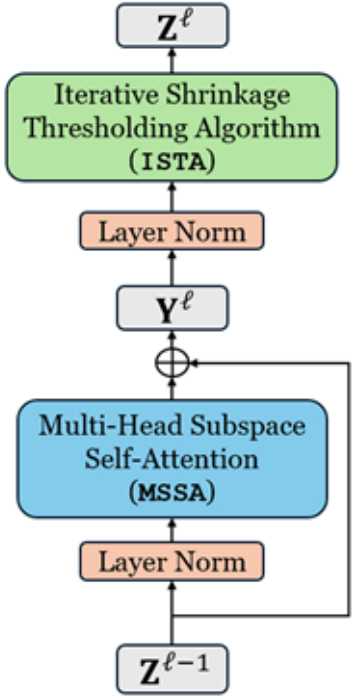
$$R^c(\boldsymbol{Z};\boldsymbol{U}) \doteq \textstyle\sum_{k=1}^{K} R(\boldsymbol{U}_k^T\boldsymbol{Z})$$

Alternating Minimization:

$$\boldsymbol{Y}^\ell = \underbrace{\boldsymbol{Z}^{\ell-1} - \alpha\nabla R^c(\boldsymbol{Z}^{\ell-1};\boldsymbol{U}^\ell)}_{\text{minimize } R^c(\boldsymbol{Z};\boldsymbol{U})} \approx \boldsymbol{Z}^{\ell-1} + \alpha\gamma^2\,\mathrm{MSSA}(\boldsymbol{Z}^{\ell-1};\boldsymbol{U}^\ell)$$

$$\mathrm{MSSA}(\boldsymbol{Z};\boldsymbol{U}) = \sum_{k=1}^{K}\boldsymbol{U}_k\boldsymbol{U}_k^T\boldsymbol{Z}\,\mathrm{softmax}((\boldsymbol{U}_k^T\boldsymbol{Z})^T(\boldsymbol{U}_k^T\boldsymbol{Z}))$$

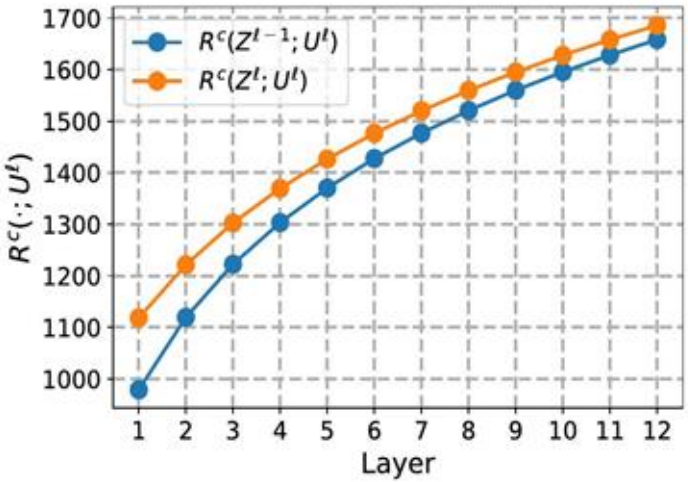
$$\boldsymbol{Z}^\ell = \underbrace{\mathrm{ReLU}\left(\boldsymbol{Y}^\ell + \beta(\boldsymbol{D}^\ell)^T(\boldsymbol{Y}^\ell - \boldsymbol{D}^\ell\boldsymbol{Y}^\ell) - \beta\lambda\boldsymbol{1}\right)}_{\text{minimize } \lambda\|\boldsymbol{Z}\|_0 - R(\boldsymbol{Z})}$$



White-box Transformer **CRATE**
[Yu et al. NeurIPS 2023]

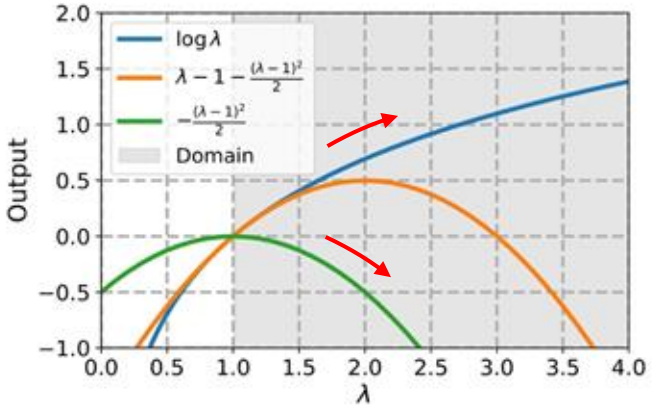
A Critical Pitfall: Is Self-Attention Playing its Role?

- **Empirically, No.** We can demonstrate the update actually diverges from its purpose.

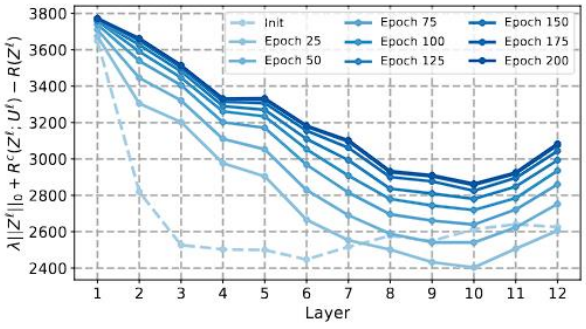


- **Theoretically, No.** We can also reveal this derivation artifacts from the eigenvalue perspective (see our paper).

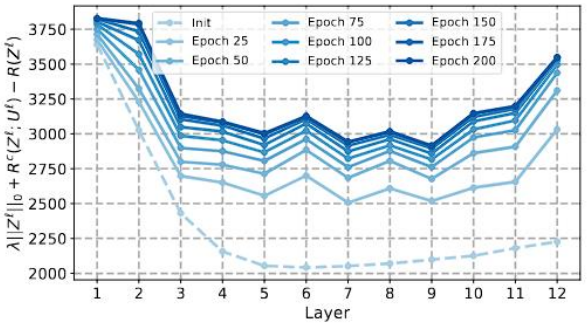
$$R^c(\boldsymbol{Z};\boldsymbol{U}) \geq \sum_{k=1}^{K} \left(\underbrace{\frac{\gamma}{2}\|\boldsymbol{U}_k^T\boldsymbol{Z}\|_F^2}_{\text{First-order term}} \underbrace{-\frac{\gamma^2}{4}\|(\boldsymbol{U}_k^T\boldsymbol{Z})^T\boldsymbol{U}_k^T\boldsymbol{Z}\|_F^2}_{\text{Second-order term}} \right)$$



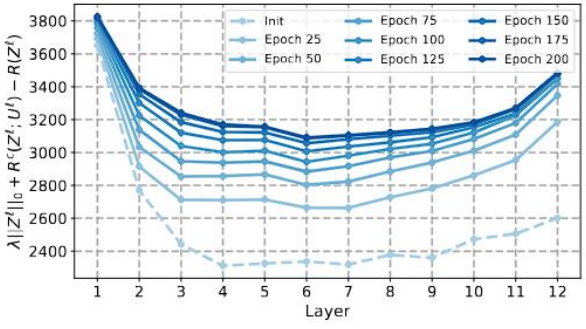
Is SRR Truly Optimized from a Macro View?



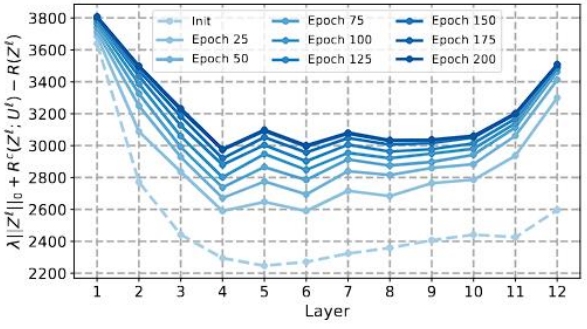
(a) Sparse rate reduction measure of CRATE-C (7)



(b) Sparse rate reduction measure of CRATE-N (8)



(c) Sparse rate reduction measure of CRATE-T (9)



(d) Sparse rate reduction measure of CRATE

Figure 2: Sparse rate reduction measure $\lambda\|\boldsymbol{Z}\|_0 + R^c(\boldsymbol{Z};\boldsymbol{U}) - R(\boldsymbol{Z})$ of CRATE and its variants evaluated at different layers and epochs on CIFAR-10.

- **Mixed Results.** The implementations *partially* align with the design purpose.
- Robust to add-ons like layer normalization.
- Gradient approximation challenges the exact implementation.

Whether Sparse Rate Reduction Benefits Generalization?

- SRR as a Complexity Measure and Proxy for Generalization Gap

$$\mu_{\text{SRR}}(\boldsymbol{w}; \boldsymbol{Z}) = \frac{1}{L}\sum_{\ell=1}^{L} \mu^{\ell}_{\text{SRR}}(\boldsymbol{w}^{\ell}; \boldsymbol{Z}^{\ell}) = \frac{1}{L}\sum_{\ell=1}^{L} \Big(\lambda\|\boldsymbol{Z}^{\ell}\|_0 + R^{c}(\boldsymbol{Z}^{\ell}; \boldsymbol{U}^{\ell}) - R(\boldsymbol{Z}^{\ell})\Big)$$

- **Positive Correlation and Stronger than Others (e.g., Path-Norm, Sharpness).**

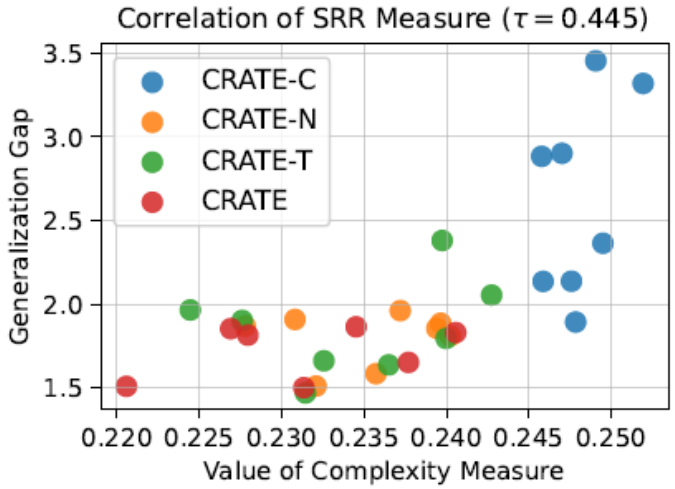


Table 1: Correlation of complexity measures with generalization gap (width $d = 384$).

Complexity measures	Batch size	Learning rate	Dropout	Model type	Overall τ	Ψ
ℓ_2-norm	0.200	-0.333	-0.333	-0.429	-0.363	-0.224
ℓ_2-norm-init	0.200	-0.200	-0.333	-0.286	-0.290	-0.158
# params	0.000	0.000	0.000	-0.572	-0.351	-0.143
1/margin	-0.067	0.467	0.467	0.238	0.415	0.276
sum-of-spec	0.200	-0.333	-0.467	-0.381	-0.290	-0.245
prod-of-spec	0.200	-0.333	-0.467	-0.476	-0.338	-0.269
sum-of-spec/margin	0.333	-0.333	-0.467	-0.048	-0.230	-0.129
prod-of-spec/margin	0.333	-0.333	-0.467	-0.143	-0.260	-0.152
fro/spec	-0.200	0.333	0.467	-0.476	0.019	0.031
spec-init-main	0.333	-0.333	-0.467	-0.190	-0.273	-0.164
spec-orig-main	0.200	-0.333	-0.467	-0.095	-0.252	-0.174
sum-of-fro	0.200	-0.333	-0.333	-0.381	-0.325	-0.212
prod-of-fro	0.200	-0.333	-0.333	-0.429	-0.372	-0.224
sum-of-fro/margin	0.333	-0.200	-0.467	-0.048	-0.217	-0.095
prod-of-fro/margin	0.333	-0.200	-0.467	-0.143	-0.247	-0.119
fro-distance	0.200	-0.200	-0.333	-0.286	-0.290	-0.155
spec-distance	0.200	-0.200	-0.333	-0.286	-0.290	-0.155
param-norm	0.200	-0.333	-0.333	-0.429	-0.363	-0.224
path-norm	0.333	-0.600	-0.467	-0.286	-0.191	-0.255
pac-bayes-init	0.200	0.200	-0.600	0.238	0.015	-0.009
pac-bayes-orig	-0.200	0.333	0.467	0.381	0.333	0.245
$1/\sigma$ pac-bayes-flatness	-0.267	0.333	0.333	0.455	0.333	0.213
SRR	-0.067	0.467	0.333	0.714	**0.445**	**0.362**

Whether Sparse Rate Reduction Benefits Generalization?

- SRR as a Regularization Technique for Training

$$\min_{\boldsymbol{w}} \mathcal{L}_{\text{ce}}(\boldsymbol{w}) + \lambda \cdot \frac{1}{L} \sum_{\ell=1}^{L} \mu^{\ell}_{\text{SRR}}(\boldsymbol{w}^{\ell}; \boldsymbol{Z}^{\ell}_{\text{StopGrad}})$$

- **Consistently Improves Performance.**

Table 2: Top-1 accuracy for CRATE and its variants trained with or without SRR regularization on CIFAR-10/100 from scratch (width $d = 384$).

Models	CIFAR-10		CIFAR-100	
	cross-entropy	+ SRR regularization (L=12)	cross-entropy	+ SRR regularization (L=12)
CRATE-C	76.87	**77.61**	43.40	**44.53**
CRATE-N	81.52	**81.91**	55.11	**55.62**
CRATE-T	85.49	**85.52**	60.59	**60.69**
CRATE	86.67	**86.79**	62.40	**62.52**

Table 5: Top-1 accuracy for CRATE and its variants trained with efficient implementations of SRR regularization on CIFAR-10 from scratch (width $d = 384$).

Training methods	CIFAR-10			
	CRATE-C	CRATE-N	CRATE-T	CRATE
cross-entropy	76.87	81.52	85.49	86.67
+ Layer 2 reg	77.75	**82.41**	**85.84**	**87.03**
+ Layer 4 reg	**77.95**	81.57	85.46	87.03
+ Layer 6 reg	77.48	80.83	85.22	87.02
+ Layer 8 reg	77.04	81.29	85.12	86.64
+ Layer 10 reg	77.44	81.19	85.68	86.67
+ Layer 12 reg	77.61	81.91	85.52	86.79
+ Random layer reg	75.19	79.66	84.27	85.36

Today's Talk

Part I: Investigating Sparse Rate Reduction (SRR): An Information-Theoretic Perspective

Part II: Hyper-spherical Energy-Informed Design: A more Principled Energy-based Approach

Conceptualization: Token Uniformity and Alignment

SRR provides an information-theoretic lens; Can we design Transformers by framing forward dynamics as minimizing a general "energy" function?

We propose a dual objective:

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

Interpretation:

- **Mode seeking**: tokens align with semantic basis on the **high-dim** hyperspheres
- **Mass covering**: tokens distribute uniformly to avoid collapse on the **low-dim** hyperspheres

Theoretical Justification: Maximum Likelihood on Hyperspheres

Formalize token dynamics as joint Maximum Likelihood Estimation (MLE) on Hyperspheres:

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

Under some mild assumptions:

- Conditional i.i.d. $p_\phi(\boldsymbol{z}^1, \ldots, \boldsymbol{z}^H|\boldsymbol{x}) = \Pi_{h=1}^{H} p_\phi(\boldsymbol{z}^h|\boldsymbol{x})$
- Uniform prior on hyperspheres $p(\boldsymbol{z})$

$$\begin{aligned}\max_{\boldsymbol{x}} \quad & \mathbb{E}_{(\boldsymbol{z}^1,\ldots,\boldsymbol{z}^H)\sim p(\boldsymbol{z}^1,\ldots,\boldsymbol{z}^H)} \left[\log p(\boldsymbol{x}, \boldsymbol{z}^1, \ldots, \boldsymbol{z}^H; \theta, \phi)\right] \\ &= -\sum_{h=1}^{H} \mathrm{D}_{\mathrm{KL}}(p(\boldsymbol{z}) \parallel p_\phi(\boldsymbol{z}^h|\boldsymbol{x})) - H \times \mathcal{H}(p(\boldsymbol{z})) + \log p_\theta(\boldsymbol{x})\end{aligned}$$

Dual Energy: Quantifying with Extended Hopfield Energy Functions

We quantify the uniformity and alignment terms with extended Hopfield energy:

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

- Repulsive Energy (Self-Attention):

$$E_{\text{ATTN}}^h = \beta^{-1} \sum_{i=1}^{N} \log \left(\sum_{j=1}^{N} \exp \left(\beta (\boldsymbol{z}_i^h)^T (\boldsymbol{z}_j^h) \right) \right)$$

- Attractive Energy (Feed-Forward)

$$E_{\text{FF}} = -\frac{1}{2} \sum_{i=1}^{N} \sum_{m=1}^{M} \left(\text{ReLU} \left(\boldsymbol{d}_m^T \boldsymbol{x}_i \right) \right)^2$$

Hyper-SET: Transformers Emerging from Energy Minimization

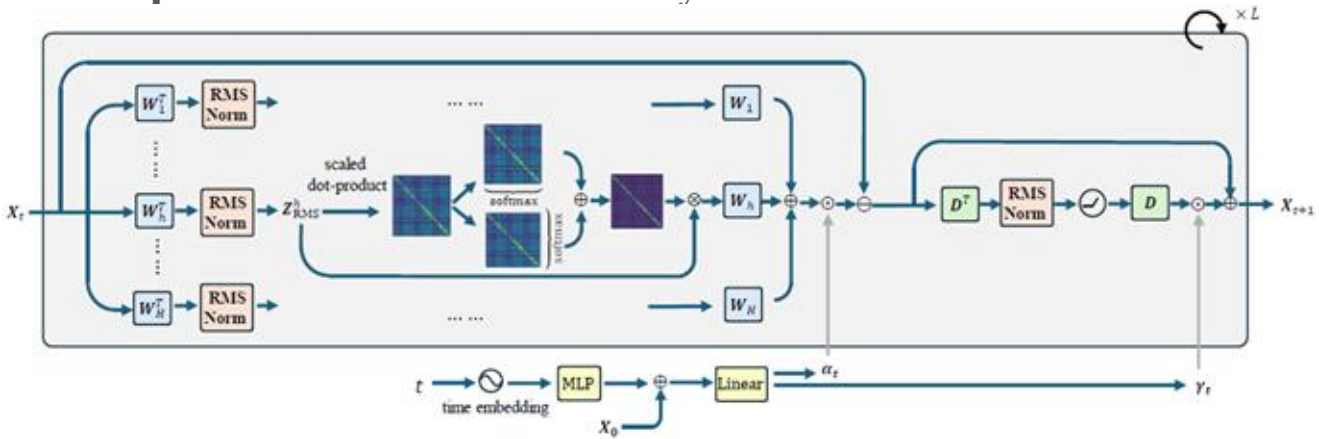
Iterative minimization of the dual energy induces core Transformer components:

$$\dot{X} = -\nabla_X E_{\text{ATTN}} = -\sum_{h=1}^{H} W_h W_h^T X \left(\underbrace{\text{softmax}}_{\text{column-wise}} \left(\beta (W_h^T X)^T (W_h^T X)\right) + \underbrace{\text{softmax}}_{\text{row-wise}} \left(\beta (W_h^T X)^T (W_h^T X)\right) \right)$$

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

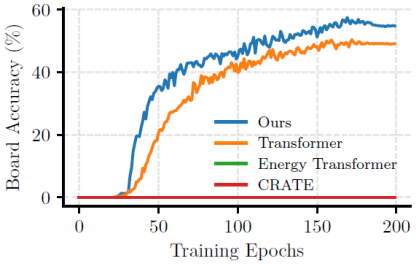
A recurrent-depth model:

- Shared parameters across layers, leading to **compactness** and **scalability**
- **Adaptive step sizes** learned for flexibility

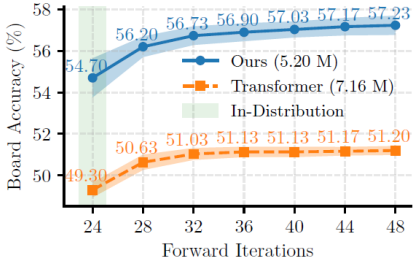


Hyper-SET: Interpretable by Design & Empirically Competitive

Superior on solving Sudoku and image classification



(a) Sudoku training dynamics



(b) Sudoku test-time extrapolation

Table 1: Top-1 accuracy on image classification. Models are under the same hidden dimension and 1 layer with 12 iterations. Scaling up our model to match Transformer's parameters gives better performance. Parameters are measured on CIFAR-10.

Models	Config (# Params)	Dataset		
		CIFAR-10	CIFAR-100	IM-100
Transformer	`Small` (1.79 M)	89.90	61.89	69.44
CRATE-T [46]	`Small` (0.32 M)	86.77	60.52	61.12
CRATE [102]	`Small` (0.47 M)	86.97	61.13	63.52
Energy Transformer [41]	`Small` (0.91 M)	75.98	52.50	40.40
Ours	`Small` (0.96 M)	88.89	62.48	67.64
Ours	`Small` Scale-up (1.61 M)	**90.11**	**63.41**	**70.16**

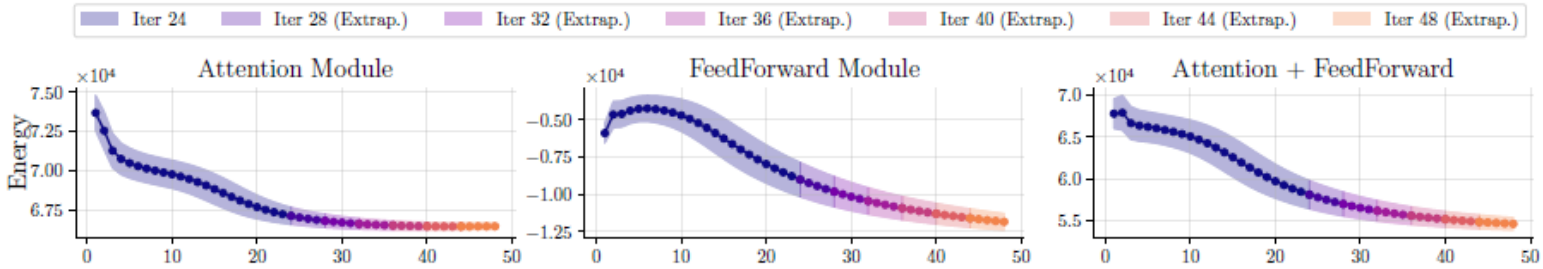
Competitive on masked image modeling

Table 2: Comparison of masked image modeling performance on ImageNet-100 (5k). Our model lags behind Transformer when given the same iterations, but matches its performance if scaling up them and the width of the feedforward module (larger M). Our model is also more parameter-efficient.

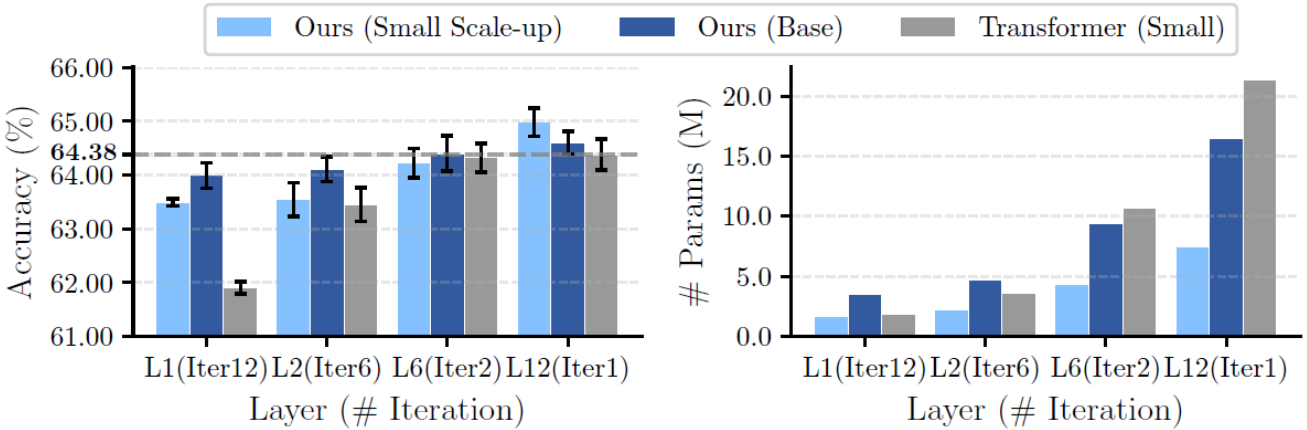
Models	Layer / Iteration / FF Ratio M	PSNR (↑)	SSIM (↑)	Multi-Scale SSIM (↑)	LPIPS (↓)	FID (↓)
Transformer	1 / 12 / $4d$ (8.85 M)	15.953	0.417	**0.599**	**0.327**	**43.428**
Ours	1 / 12 / d (3.94 M)	15.713	0.411	0.576	0.358	59.841
Ours	1 / 24 / $8d$ (8.07 M)	**15.955**	**0.417**	0.596	0.332	45.174

Hyper-SET: Interpretable by Design & Empirically Competitive

Intrinsic Interpretability: Dynamics are directly describable via energy functions



Compactness: Parameter-efficient when scaling up



Hyper-SET: Alternative Design with Generalized Energy

A more generalized expression:

$$E_{\text{ATTN}} = \sum_{h=1}^{H} E_{\text{ATTN}}^{h} = \sum_{h=1}^{H} \sum_{i=1}^{N} f\left(\sum_{j=1}^{N} K(\boldsymbol{W}_h^T \boldsymbol{x}_i, \boldsymbol{W}_h^T \boldsymbol{x}_j) \right)$$

Generality: Linear attention induced by an alternative energy function:

Table 18: Alternative designs on attention energy E_{ATTN} and the induced operators.

Operator	$f(x)$	$K(\boldsymbol{x}, \boldsymbol{y})$	E_{ATTN}	$-\nabla_{\boldsymbol{X}} E_{\text{ATTN}}$
Bi-Softmax (Ours)	$\beta^{-1}\log(x)$	$\exp(\beta \boldsymbol{x}^T \boldsymbol{y})$	(3)	(7)
Sigmoid-Softmax	$\frac{\beta^{-1}}{2}x$	$\sigma(\beta \boldsymbol{x}^T \boldsymbol{y})$	$\frac{1}{2}\sum_{h=1}^{H}\sum_{i=1}^{N}\sum_{j=1}^{N}\sigma\left(\beta(\boldsymbol{W}_h^T\boldsymbol{x}_i)^T\boldsymbol{W}_h^T\boldsymbol{x}_j\right)\beta^{-1}$	$\sum_{h=1}^{H}\boldsymbol{W}_h\boldsymbol{W}_h^T\boldsymbol{X}\sigma(1-\sigma)\left(\beta(\boldsymbol{W}_h^T\boldsymbol{X})^T\boldsymbol{W}_h^T\boldsymbol{X}\right)$
Linear Attention	$\frac{\beta^{-1}}{2}(x)$	$\frac{1}{2}\left(\beta\Phi(\boldsymbol{x})^T\Phi(\boldsymbol{y})\right)^2$	$\frac{1}{4}\sum_{h=1}^{H}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\beta\Phi(\boldsymbol{W}_h^T\boldsymbol{x}_i)^T\Phi(\boldsymbol{W}_h^T\boldsymbol{x}_j)\right)^2\beta^{-1}$	$\sum_{h=1}^{H}\boldsymbol{W}_h\Phi'(\boldsymbol{W}_h^T\boldsymbol{X})\odot\left(\beta\Phi(\boldsymbol{W}_h^T\boldsymbol{X})\Phi(\boldsymbol{W}_h^T\boldsymbol{X})^T\Phi(\boldsymbol{W}_h^T\boldsymbol{X})\right)$

Table 3: Alternative designs on key components measured by top-1 accuracy on CIFAR-10 and CIFAR-100.

Components	Alternative Designs	Dataset	
		CIFAR-10	CIFAR-100
E_{ATTN}	Bi-Softmax Attention (Ours)	**90.11**	**63.41**
	Sigmoid Attention	85.93	59.72
	Linear Attention	84.88	56.97
E_{FF}	ReLU FF (Ours)	**90.11**	**63.41**
	Softmax FF	88.20	62.44
	Gated FF	84.99	59.29
Step Size	Learning Step Size (Ours)	**90.11**	**63.41**
	$\alpha_t = \gamma_t = 0.5$	25.81	57.92
	$\alpha_t = \gamma_t = 0.1$	81.45	58.29

Future Application to EEG-Based Auditory Attention Decoding

In summary, Hyper-SET provides

- a powerful top-down design principle, deriving (novel) Transformer components directly from the minimization of generalized Hopfield energy functions;
- intrinsically interpretable, compact, and high-performing models.

EEG signals are long time-series with noisy input:

- Develop scalable Transformers with **linear complexity** w.r.t. input length
- Endow Transformers with explicit **denoising** property

Thank You !

- **Y.Hu**, D.Zou, D.Xu, An In-depth Investigation of Sparse Rate Reduction in Transformer-like Models, NeurIPS 2024.
- **Y.Hu**, D.Zou, D.Xu, Hyper-SET: Designing Transformers via Hyperspherical Energy Minimization, arxiv 2025.