Energy Transformer

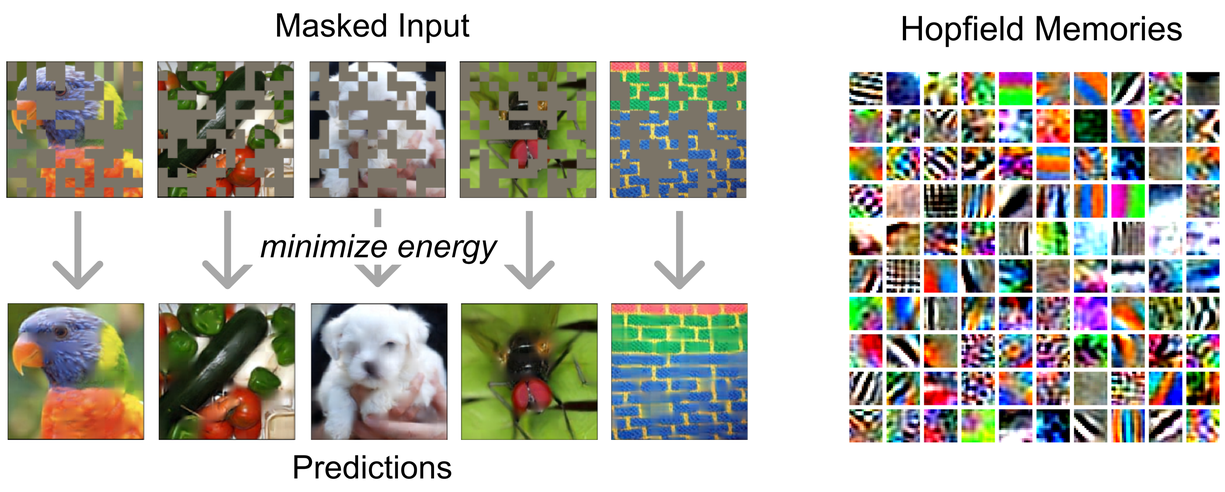
Benjamin Hoover* **Yuchen** Liang* **Bao** Pham* **Rameswar** Panda **Hendrik** Strobelt **Polo** Chau **Mohammed** J. Zaki **Dmitry** Krotov

*Authors contributed equally

Energy Transformer (E.T.) is a novel architecture that is an **Energy-Based Model**, an **Associative Memory**, and a **Transformer**. Specifically, E.T. looks like a recurrent Transformer block that defines an attractor system (*an O.D.E. with guaranteed fixed points convergence*) that performs error correction via energy descent. E.T. shows excellent performance on MASKed image in-painting, graph anomaly detection, and graph classification.

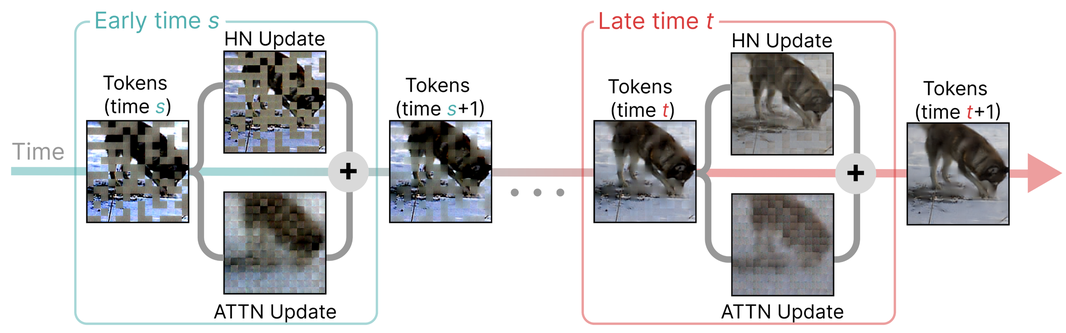
E.T. is Energy-Based Associative Memory

Minimizing energy (inference) **performs error correction on corrupted data.**



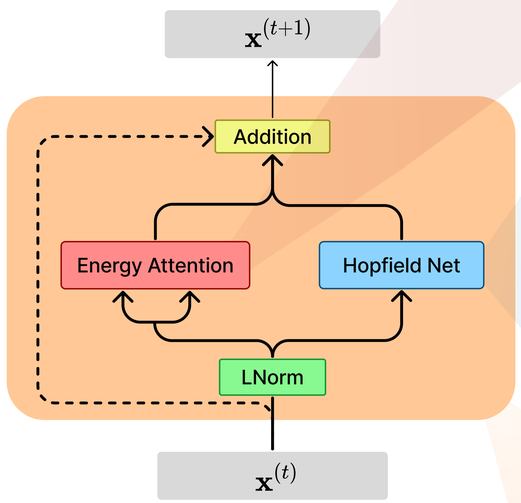
E.T. is Interpretable by Design

Visualize dynamic states (tokens and updates) **at any time**



E.T. is a Transformer

✅ **Energy Transformer Block**



Energy Attention makes Queries align with Keys in the latent space (and vice versa!)

$$E^{\text{ATT}} = -\frac{1}{\beta}\sum_h \sum_C \log\Big(\sum_{B\neq C} \exp\Big(\beta \sum_\alpha K_{\alpha hB} Q_{\alpha hC}\Big)\Big)$$

Update: $$-\frac{\partial E^{\text{ATT}}}{\partial g_{iA}} = \sum_{C\neq A}\sum_\alpha W^Q_{\alpha i}\, K_{\alpha C}\, \underset{C}{\text{softmax}}\Big(\beta \sum_\gamma K_{\gamma C}\, Q_{\gamma A}\Big) + W^K_{\alpha i}\, Q_{\alpha C}\, \underset{A}{\text{softmax}}\Big(\beta \sum_\gamma K_{\gamma A}\, Q_{\gamma C}\Big)$$

Hopfield Network makes tokens look like memories

$$E^{\text{HN}} = -\sum_{B=1}^{N}\sum_{\mu=1}^{K} G\Big(\sum_{j=1}^{D} \xi_{\mu j}\, g_{jB}\Big)$$

Update: $$-\frac{\partial E^{\text{HN}}}{\partial g_{iA}} = \sum_{\mu=1}^{K} \xi_{\mu i} G'\Big(\sum_{j=1}^{D} \xi_{\mu j} g_{jA}\Big)$$

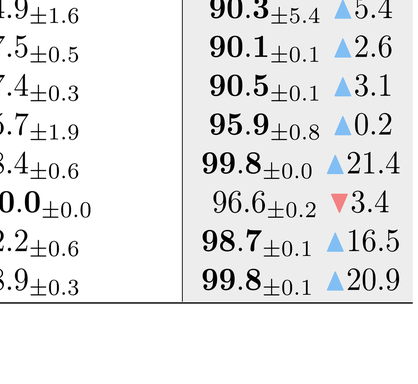
Energy Transformer minimizes the sum of these energies

$$E = E^{\text{ATT}} + E^{\text{HN}}$$

Continuous update: $$\tau \frac{dx_{iA}}{dt} = -\frac{\partial E}{\partial g_{iA}}$$

Discrete update: (stepsize α) $$x^{(t+1)}_{iA} = x^{(t)}_{iA} - \alpha\Big(\frac{\partial E^{\text{ATT}}}{\partial g_{iA}} + \frac{\partial E^{\text{HN}}}{\partial g_{iA}}\Big)$$

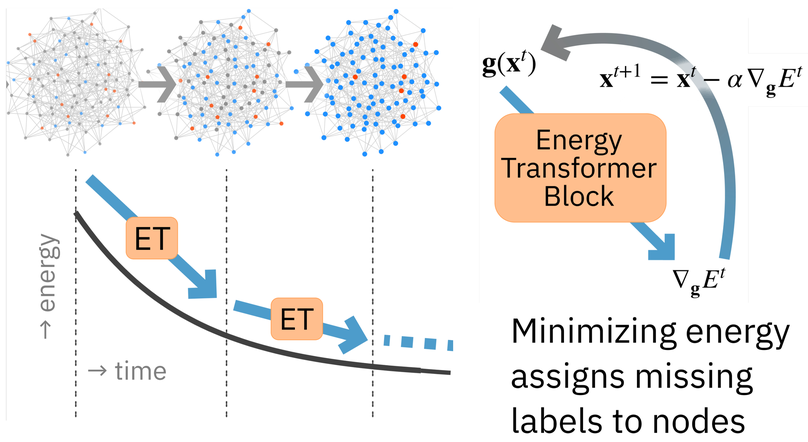
Standard Transformer Block



E.T. sets SOTA on Graphs

Node anomaly detection

	Datasets	Split	Top Baseline Score	ET (Ours)
Macro-F1	Yelp	1%	$62.1_{\pm 1.3}$	$\mathbf{63.0}_{\pm 0.6}$ ▲0.9
		40%	$71.0_{\pm 0.9}$	$\mathbf{71.5}_{\pm 0.1}$ ▲0.5
	Amazon	1%	$\mathbf{90.9}_{\pm \mathbf{0.7}}$	$89.3_{\pm 0.7}$ ▼0.7
		40%	$92.2_{\pm 0.4}$	$\mathbf{92.8}_{\pm \mathbf{0.3}}$ ▲0.6
	T-Finance	1%	$84.8_{\pm 0.0}$	$\mathbf{85.1}_{\pm 1.0}$ ▲0.3
		40%	$86.8_{\pm 0.0}$	$\mathbf{88.2}_{\pm 1.0}$ ▲1.4
	T-Social	1%	$75.9_{\pm 0.0}$	$\mathbf{79.1}_{\pm \mathbf{0.7}}$ ▲3.2
		40%	$\mathbf{83.9}_{\pm 0.0}$	$83.5_{\pm 0.4}$ ▼0.4
AUC	Yelp	1%	$\mathbf{75.4}_{\pm \mathbf{0.9}}$	$73.2_{\pm 0.8}$ ▼2.2
		40%	$84.0_{\pm 0.9}$	$\mathbf{84.9}_{\pm \mathbf{0.3}}$ ▲0.9
	Amazon	1%	$90.4_{\pm 2.0}$	$\mathbf{91.9}_{\pm 1.0}$ ▲1.5
		40%	$\mathbf{98.0}_{\pm \mathbf{0.4}}$	$97.3_{\pm 0.4}$ ▼0.7
	T-Finance	1%	$91.1_{\pm 0.0}$	$\mathbf{92.8}_{\pm 1.1}$ ▲1.7
		40%	$94.3_{\pm 0.0}$	$\mathbf{95.0}_{\pm \mathbf{3.0}}$ ▲0.7
	T-Social	1%	$88.0_{\pm 0.0}$	$\mathbf{91.9}_{\pm \mathbf{0.6}}$ ▲3.9
		40%	$\mathbf{95.2}_{\pm 0.0}$	$93.9_{\pm 0.2}$ ▼1.3



Minimizing energy assigns missing labels to nodes

Graph Classification

Datasets	Top Baseline Score	ET (Ours)
PROTEINS	$84.9_{\pm 1.6}$	$\mathbf{90.3}_{\pm 5.4}$ ▲5.4
NCI1	$87.5_{\pm 0.5}$	$\mathbf{90.1}_{\pm 0.1}$ ▲2.6
NCI109	$87.4_{\pm 0.3}$	$\mathbf{90.5}_{\pm 0.1}$ ▲3.1
DD	$95.7_{\pm 1.9}$	$\mathbf{95.9}_{\pm 0.8}$ ▲0.2
ENZYMES	$78.4_{\pm 0.6}$	$\mathbf{99.8}_{\pm 0.0}$ ▲21.4
MUTAG	$\mathbf{100.0}_{\pm 0.0}$	$96.6_{\pm 0.2}$ ▼3.4
MUTAGENICITY	$82.2_{\pm 0.6}$	$\mathbf{98.7}_{\pm 0.1}$ ▲16.5
FRANKENSTEIN	$78.9_{\pm 0.3}$	$\mathbf{99.8}_{\pm 0.1}$ ▲20.9

IBM Research · Rensselaer · Georgia Tech