NEURAL INFORMATION PROCESSING SYSTEMS

Deep Tree Tensor Networks

Chang Nie

Nanjing University of Science and Technology

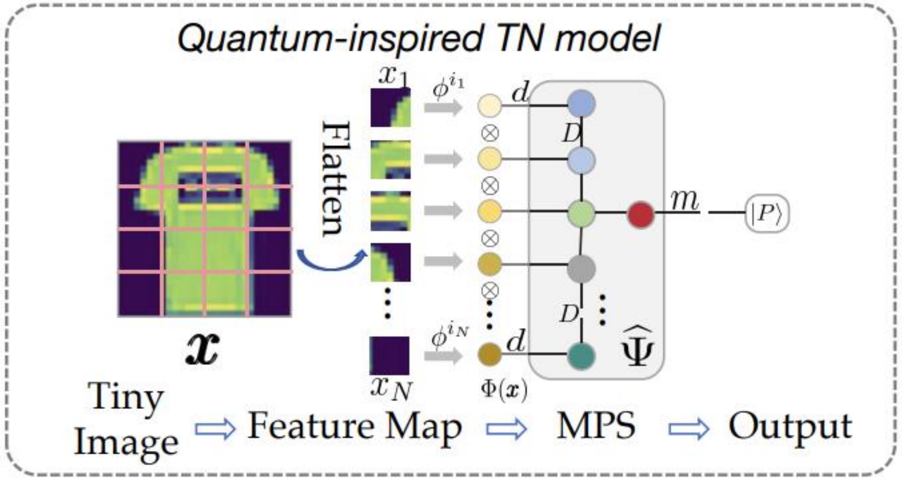
changnie@njust.edu.cn

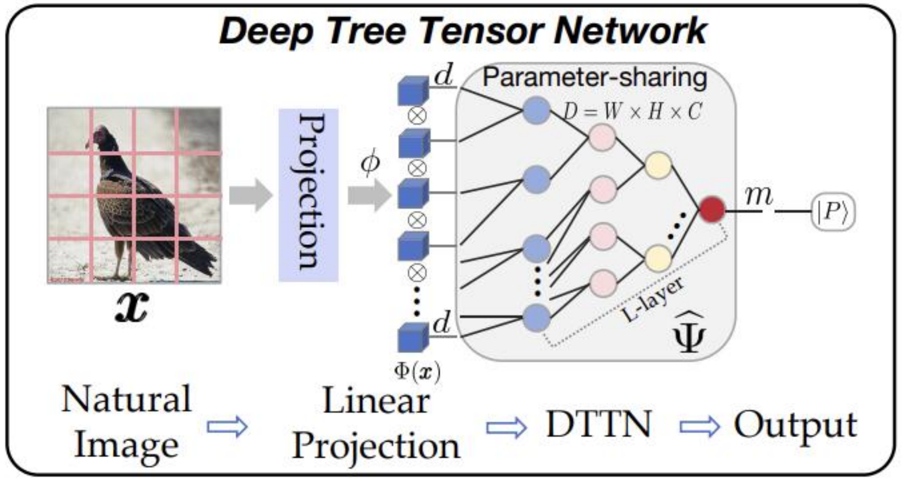
Code

Wechat

The Problem: Scaling Quantum-Inspired Models

- **The Promise of Tensor Networks (TNs)**
  - Originated in quantum physics to combat the "curse of dimensionality".
  - Offer a powerful, interpretable framework for modeling high-order interactions.
- **But, they face two key challenges in deep learning:**
  1. **Scalability Challenge:** Traditional TNs like MPS are limited to small-scale inputs and low dimensions, making them impractical for benchmarks like ImageNet.
  2. **Expressivity Challenge:** When used only for parameter compression, they lose their core strength: modeling **exponential-order feature interactions**.
- **The Gap:** A significant gap exists between the theoretical power of TNs and their practical application on large-scale vision tasks.





Our Approach: The Deep Tree Tensor Network (DTTN)

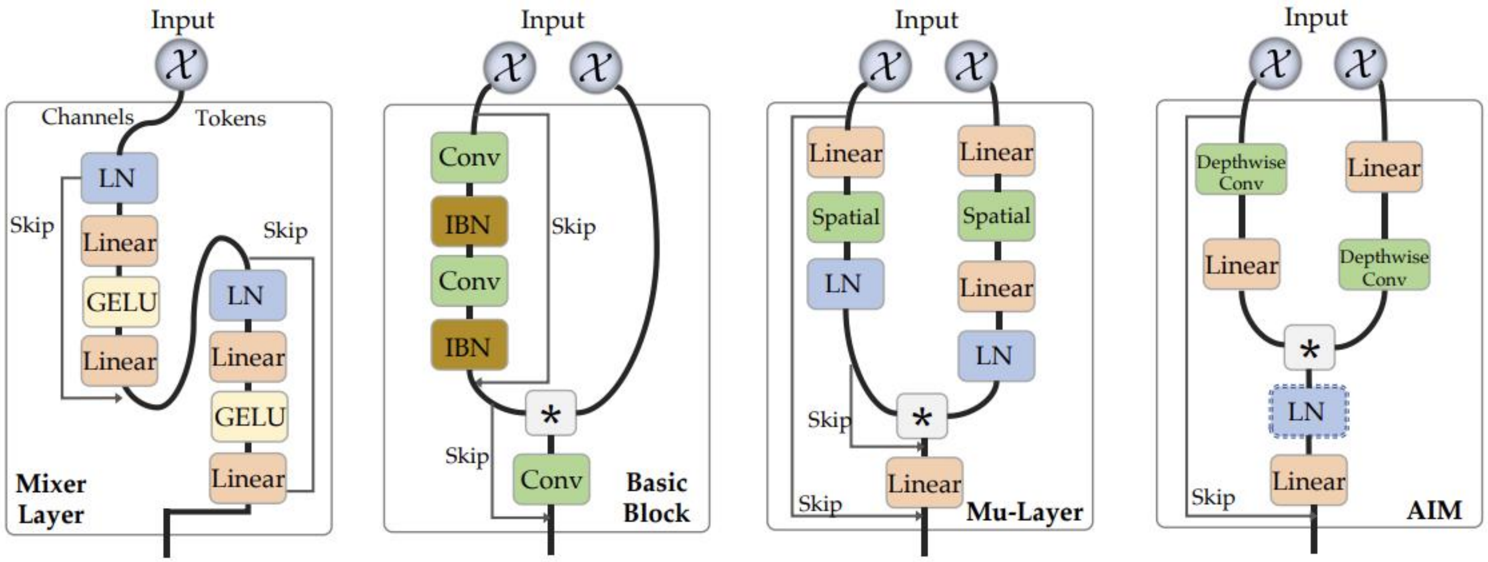
- **Goal:** To bridge the gap by designing a new architecture that is:
  - **Scalable:** Natively handles large-scale, high-resolution images.
  - **Expressive:** Explicitly models high-order multiplicative interactions.
  - **Principled:** Grounded in the theory of Tensor Networks.
- **Our Solution: A Novel, Purely Multilinear Architecture**
  - We abandon non-linear activations entirely.
  - We build the network from a simple, efficient interaction module.
  - The entire network unfolds into a classic **Tree Tensor Network (TTN)**.

The Core Idea: Antisymmetric Interaction Module (AIM)

Design: An antisymmetric two-branch structure combines spatial and channel operations in reverse orders.

Key Advantages:

- **(a) Parameter Efficient: More efficient than a symmetric design.**
- **(b) Purely Multilinear: Maintains the network's polynomial nature.**



(a) MLP-Mixer (b) $\mathcal{R}$-PolyNets (c) MONet **(d) DTTN**

The DTTN Architecture: From Module to Network

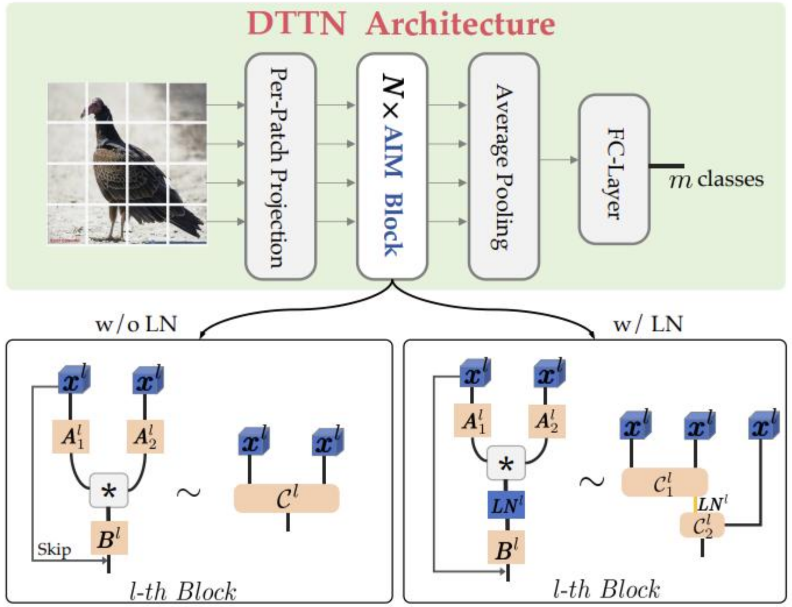


Figure 3: Schematic overview of the DTTN architecture.

$$
\begin{aligned}
\boldsymbol{x}^{l+1} &= \boldsymbol{x}^l + \boldsymbol{B}^l\left((\boldsymbol{A}_1^l \boldsymbol{x}^l) * (\boldsymbol{A}_2^l \boldsymbol{x}^l)\right) \\
&= \boldsymbol{x}^l + \mathsf{Reshape}\left(\boldsymbol{B}^l ({\boldsymbol{A}_1^l}^T \odot {\boldsymbol{A}_2^l}^T)^T\right) \times_{2,3}^{1,2} (\boldsymbol{x}^l \otimes \boldsymbol{x}^l) \\
&= \boldsymbol{\mathcal{C}}^l \times_{2,3}^{1,2} (\boldsymbol{x}^l \otimes \boldsymbol{x}^l)
\end{aligned}
$$

- The DTTN is built by hierarchically stacking our core AIM blocks into a deep, multi-stage architecture.

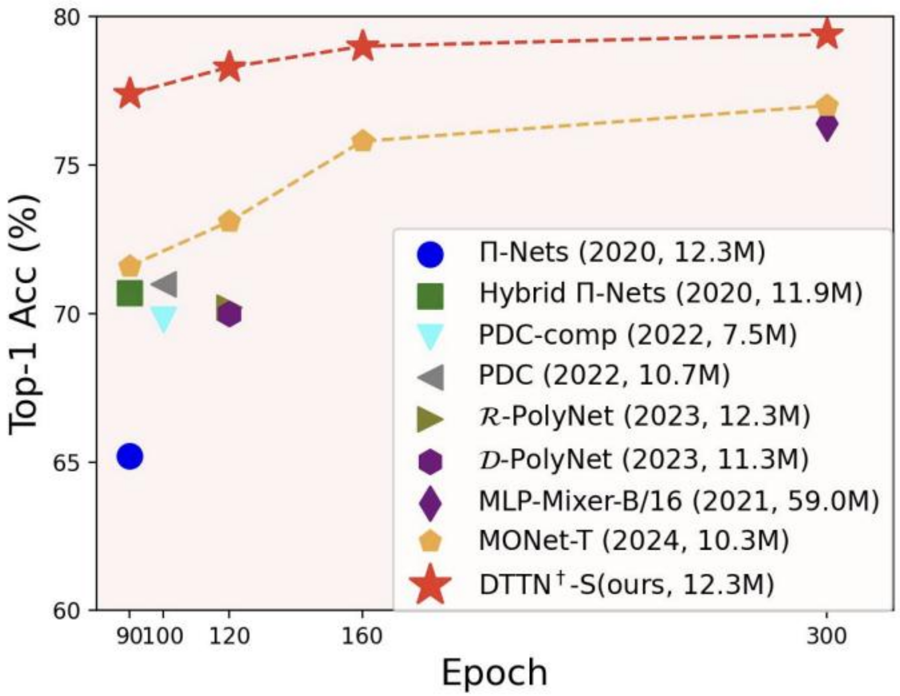
Theoretical Insight: Unfolding into a Tree Tensor Network

Proposition 1. *The DTTN has the capability to capture 2^L multiplicative interactions among input elements, which can be represented in the format of Equation (1) as $\Phi(\boldsymbol{x}) = \otimes^{2^L} \phi(\boldsymbol{x}, \boldsymbol{\Lambda}_\phi)$. Consequently, the elements of $f(\boldsymbol{x})$ are homogeneous polynomials of degree 2^L over the feature map $\phi(\boldsymbol{x}, \boldsymbol{\Lambda}_\phi)$.*

Theorem 1. *Given the local mapping function $\phi^{i_j}(x_j) = [x_j^0, \ldots, x_j^{2^L}]^T$, a polynomial network with the expansion form of Equation (6) can be transformed into a quantum-inspired TNs model with finite bond dimension.*

- A DTTN without Layer Normalization is mathematically equivalent to a Tree Tensor Network.
- Each AIM acts as a binary node in a tree, performing a tensor contraction that fuses features.
- This equivalence allows DTTN to model 2^Lorder interactions, inheriting the exponential expressive power of classic Tensor Networks.

Main Result: State-of-the-Art on ImageNet-1K



Model	Top-1(%)	Params (M)	FLOPs(B)	Epoch	Activation	Attention	Reso.
CNN-based							
ResNet-50 [20]	77.2	25.0	4.1	-	ReLU	×	224^2
A^2Net [4]	77.0	33.4	31.3	-	ReLU	✓	224^2
AA-ResNet-152 [1]	79.1	61.6	23.8	100	ReLU	✓	224^2
RepVGG-B2g4 [15]	79.4	55.7	11.3	200	ReLU	×	224^2
Transformer- and Mamba-based							
ViT-B/16 [16]	77.9	86.0	55.0	300	GeLU	✓	224^2
DeiT-S/16 [53]	81.2	24.0	5.0	300	GeLU	✓	224^2
Swin-T/16 [36]	81.3	29.0	4.5	300	GeLU	✓	224^2
Vim-S [62]	80.5	26.0	-	300	SiLU	✓	224^2
MLP-based							
MLP-Mixer-B/16 [51]	76.4	59.0	11.6	300	GeLU	×	224^2
MLP-Mixer-L/16 [51]	71.8	507.0	44.6	300	GeLU	×	224^2
CycleMLP-T [3]	81.3	28.8	4.4	300	GeLU	×	224^2
Hire-MLP-Tiny [19]	79.8	18.0	2.1	300	GeLU	×	224^2
ResMLP-24 [52]	79.4	6.0	30.0	300	GeLU	×	224^2
S^2MLP-Wide [59]	80.0	71.0	14.0	300	GeLU	×	224^2
S^2MLP-Deep [59]	80.7	10.5	51.0	300	GeLU	×	224^2
ViP-Small/14 [21]	80.5	30.0	6.5	300	GeLU	✓	224^2
AFFNet [24]	79.8	6.0	1.5	300	ReLU	✓	256^2
Polynomial- and Multilinear-based							
Π-Nets [9]	65.2	12.3	1.9	90	-	×	224^2
DTTN-S(ours)	**71.8 / 77.2**	12.3	4.1	90 / 300	-	×	224^2
Hybrid Π-Nets [9]	70.7	11.9	1.9	90	ReLU+Tanh	×	224^2
PDC [8]	71.0	10.7	1.6	100	ReLU+Tanh	×	224^2
PDC-comp [8]	70.2	7.5	1.3	100	ReLU+Tanh	×	224^2
$\mathcal{R}$-PolyNets [10]	70.2	12.3	1.9	120	-	×	224^2
$\mathcal{D}$-PolyNets [10]	70.0	11.3	1.9	120	-	×	224^2
MONet-T [7]	77.0	10.3	2.8	300	-	×	224^2
DTTN†-T(ours)	**77.9**	7.1	2.3	300	-	×	224^2
DTTN†-S(ours)	**79.4**	12.3	4.1	300	-	×	224^2
MONet-S [7]	81.3	32.9	6.8	300	-	×	224^2
DTTN†-L(ours)	**82.4**	35.9	12.3	300	-	×	224^2

- ☐ DTTN significantly outperforms all competing polynomial and multilinear networks.
- ☐ The convergence curve demonstrates faster training and higher final accuracy.

Ablation studies

Table 6: The influence of network depth and width on model performance.

	Top-1 (%)	Params(M)
L=8, d=256	79.2	5.6
L=16,d=256	85.5	10.2
L=24,d=256	86.8	14.8
L=32,d=256	**87.2**	19.4
L=32,d=64	63.4	1.3
L=32,d=128	82.5	4.9
L=32,d=512	**87.9**	76.8

Table 7: The influence of different design choices for AIM on the performance of the DTTN variants.

	Top-1 (%)	Params(M)
SIM-Conv	86.2	9.1
SIM-Linear	84.9	4.8
DTTN†-T	**86.4**	6.9
Sim-Conv	87.8	15.9
Sim-Linear	85.2	8.3
DTTN†-S	87.7	12.1

Table 8: The influence of layer normalization inside AIM on the performance of the DTTN variants.

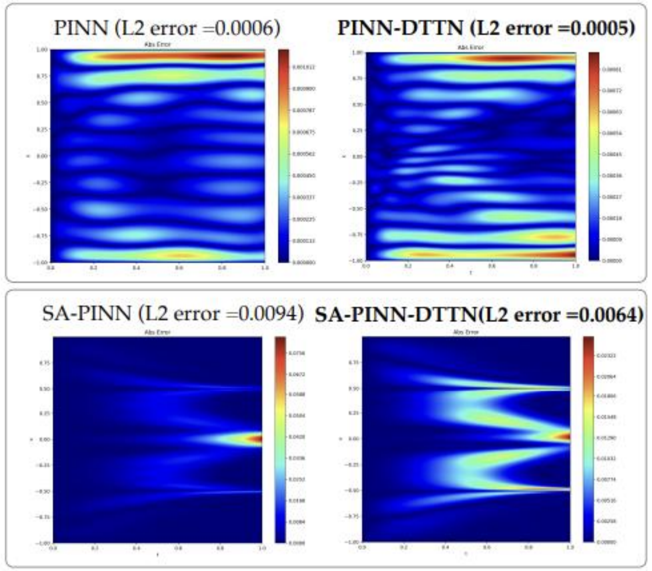
Model	Top-1 (%)	Params(M)
DTTN-T	85.6	6.9
DTTN†-T	**86.4**$_{+1.8}$	6.9
DTTN-S	87.3	12.1
DTTN†-S	**87.7**$_{+0.4}$	12.1
DTTN-L	87.6	35.6
DTTN†-L	**88.1**$_{+0.5}$	35.6

- **Antisymmetric Design:** The AIM's antisymmetric structure is validated to be more parameter-efficient, saving approximately 20% of parameters with minimal trade-off in performance compared to symmetric alternatives.
- **Importance of Layer Normalization (LN):** LN is proven to be a critical component for achieving state-of-the-art results. It significantly boosts final accuracy (by up to +1.8%) and accelerates training convergence.
- **Scalability with Depth & Width:** The model's performance scales effectively and predictably with increasing network depth and width, demonstrating its robustness and behaving like modern deep architectures.

Broader Impact

Table 5: Validating AIM as a pluggable module for enhancing feature interaction in recommendation models with consistent performance gains.

Model	Criteo	Avazu
DeepFM	80.12	75.46
DeepFM+AIM	$80.44_{+0.32}$	$75.73_{+0.27}$
FiBiNet	80.42	76.01
FiBiNet+AIM	$80.97_{+0.55}$	$76.08_{+0.07}$
DCN-V2	80.93	76.14
DCN-V2+AIM	$81.15_{+0.22}$	$76.52_{+0.38}$



$$\begin{cases} \frac{\partial u}{\partial t} = -u \\ u(x,t=0) = sin(\pi x) \\ u(1,t) = u(-1,t) = 0 \end{cases}$$

$$\begin{cases} u_t - 0.0001u_{xx} + 5u^3 - 5u = 0 \\ u(x,t=0) = x^2 cos(\pi x) \\ u(1,t) = u(-1,t) = -1. \end{cases}$$

Figure 4: Performance of PINNs on linear and nonlinear Allen-Cahn PDEs: L2 error and absolute error across the Spatial-Temporal domain.

- Recommendation Systems: The AIM acts as a plug-and-play module, effectively boosting CTR prediction performance in existing models.
- Physics-Informed Neural Networks (PINNs): DTTN serves as a powerful, activation-free core, achieving lower prediction error in solving complex PDEs.

Conclusion

- We proposed DTTN, the first Tensor Network-inspired architecture to achieve promising performance on ImageNet-1K, solving the scalability challenge.

- The hierarchical AIM architecture provides a scalable and efficient realization of a Tree Tensor Network, capturing exponential-order interactions through pure multilinear operations.

- DTTN offers a promising direction for developing powerful and transparent white-box models that combine high performance with clear structural interpretability.