

# Implicit Regularization in Hierarchical Tensor Factorization and Deep Convolutional Neural Networks

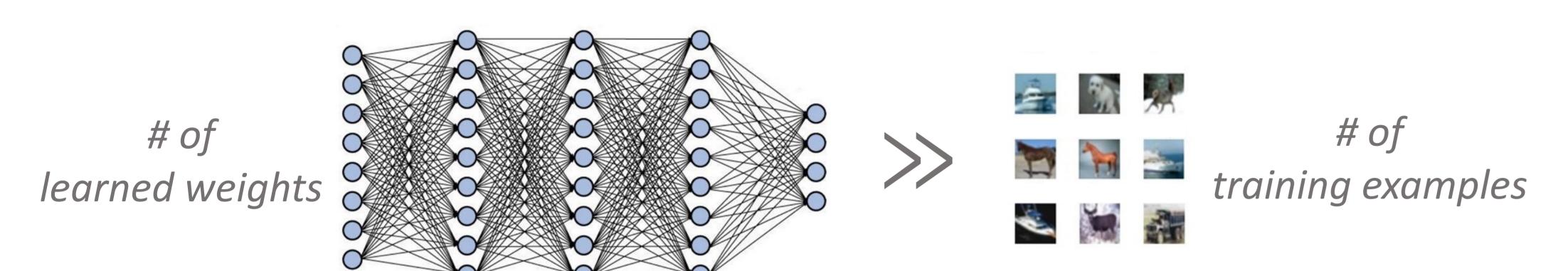
ICML 2022

Noam Razin Asaf Maman Nadav Cohen

Tel Aviv University

# I) Implicit Regularization in Deep Learning

Neural networks (NNs) generalize well despite being overparameterized



#### **Conventional Wisdom**

Gradient descent (GD) induces an implicit regularization towards "simplicity"

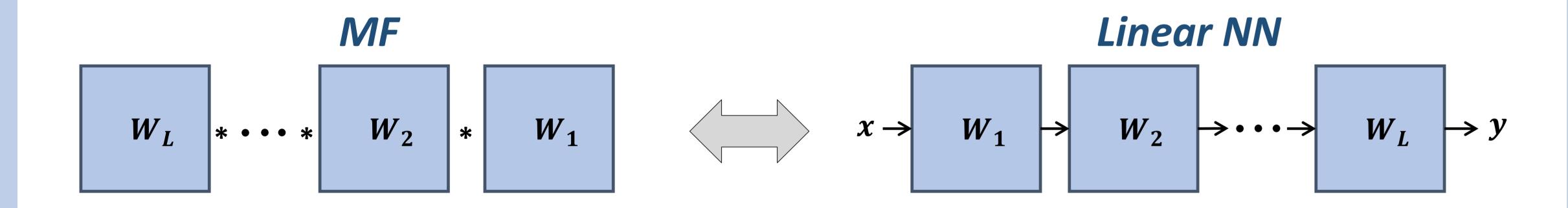
Common testbeds for formalizing this intuition: matrix and tensor factorizations

## II) Background: Matrix Factorization (MF)

Consider minimizing loss  $\mathcal{L}$  over matrices (e.g. matrix completion loss)

MF: parameterize solution as product of matrices and minimize loss with GD

$$\min_{\{W_l\}_l} \mathcal{L}(W_L \cdots W_1)$$



Past Work: Dynamical Characterization (Arora et al. 2019)

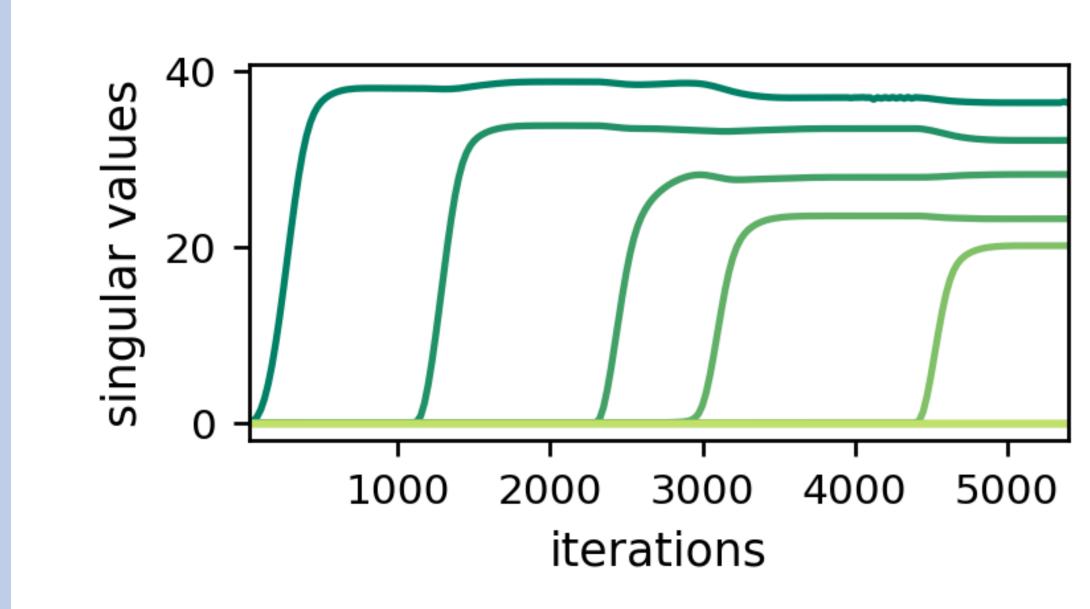
$$\sigma_{\mathsf{M}}^{(r)}$$
 — r'th singular value of  $W_{L:1} := W_L \cdots W_1$ 

Theorem: GD (w/ small step size) over MF leads to  $\frac{d}{dt}\sigma_M^{(r)}(t)\propto\sigma_M^{(r)}(t)^{2-2/L}$ 

#### Implications:

- Singular values move slower when small & faster when large!
- ► Small init ⇒ incremental learning of singular values

Experiment: completion of low rank matrix via MF



Incremental learning of singular values leads to low matrix rank

Limitation of MF as theoretical model for NNs: lacks non-linearity

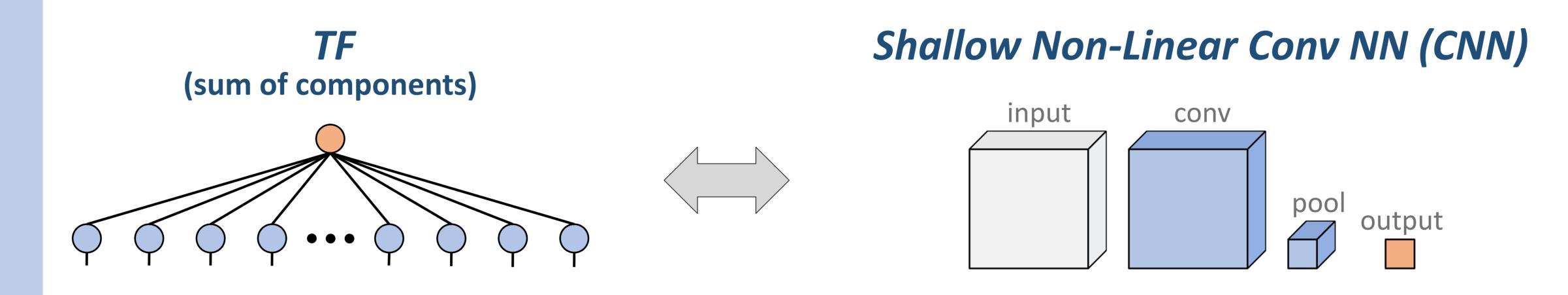
### III) Background: Tensor Factorization (TF)

Consider minimizing loss  $\mathcal{L}$  over tensors (e.g. tensor completion loss)

TF: parameterize solution as sum of outer products and minimize loss with GD

$$\min_{\{\mathbf{w}_r^n\}_{r,n}} \mathcal{L}(\sum_{r=1}^R \mathbf{w}_r^1 \otimes \cdots \otimes \mathbf{w}_r^N)$$

Tensor rank: min # of components required to express a tensor



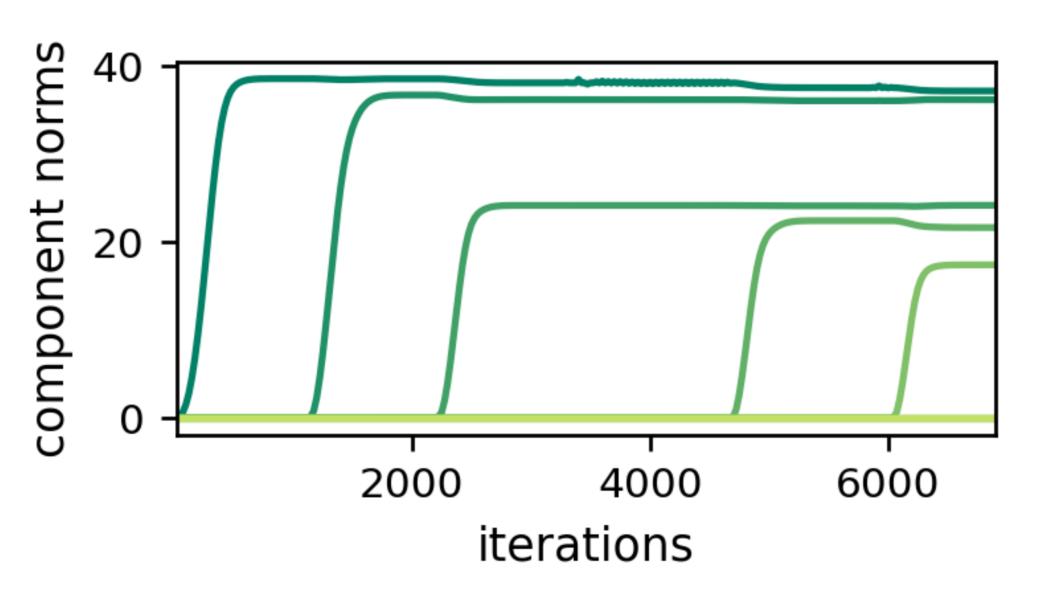
Past Work: Dynamical Characterization (Razin et al. 2021)

$$\sigma_{\mathsf{T}}^{(r)} := \| \otimes_{n=1}^{N} \mathbf{w}_{r}^{n} \|$$
 — norm of  $r$ 'th component

Theorem: GD (w/ small step size) over TF leads to  $\frac{d}{dt}\sigma_T^{(r)}(t) \propto \sigma_T^{(r)}(t)^{2-2/N}$ 

- Dynamics structurally identical to that in MF
- Component norms move slower when small & faster when large!

Experiment: completion of low tensor rank tensor via TF

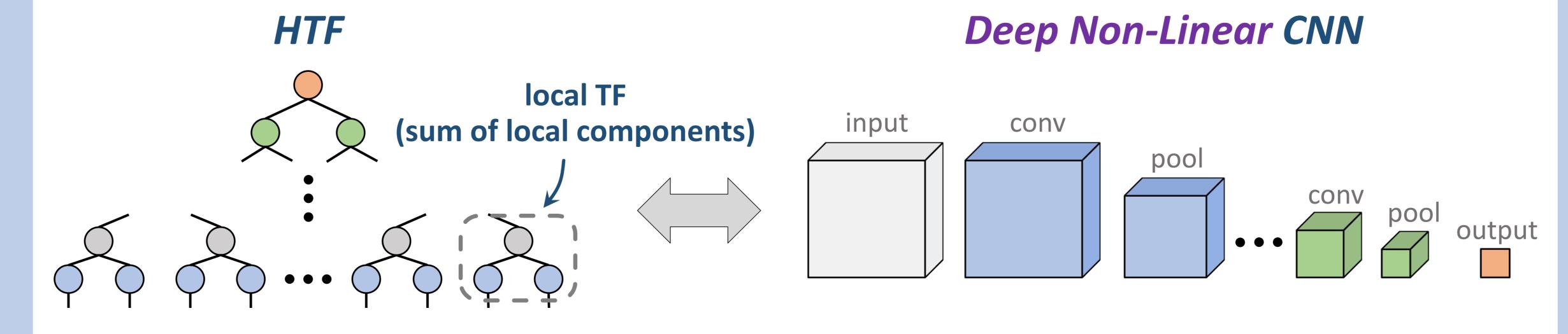


Incremental learning of components leads to low tensor rank

Limitation of TF as theoretical model for NNs: lacks depth

# IV) Hierarchical Tensor Factorization (HTF)

Accounts for both non-linearity and depth



Equivalence studied extensively (e.g. Cohen et al. 2016, Levine et al. 2018)

Representation w/ few local components  $\implies$  low hierarchical tensor (HT) rank

# V) Analysis: Implicit Regularization to Low HT Rank

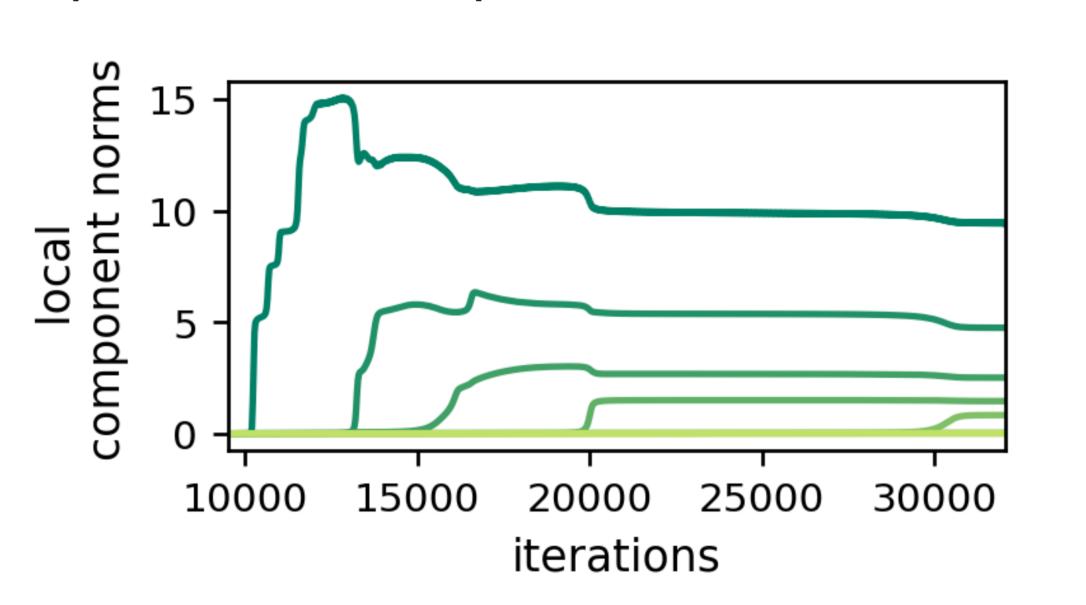
#### Our Work: Dynamical Characterization

 $\sigma_{\rm H}^{(r)}$  — norm of r'th local component at a location, K — # axes of local component

Theorem: GD (w/ small step size) over HTF leads to  $\frac{d}{dt}\sigma_H^{(r)}(t) \propto \sigma_H^{(r)}(t)^{2-2/K}$ 

- Dynamics structurally identical to those in MF & TF
- ► Local component norms move slower when small & faster when large!

Experiment: completion of low hierarchical tensor rank tensor via HTF

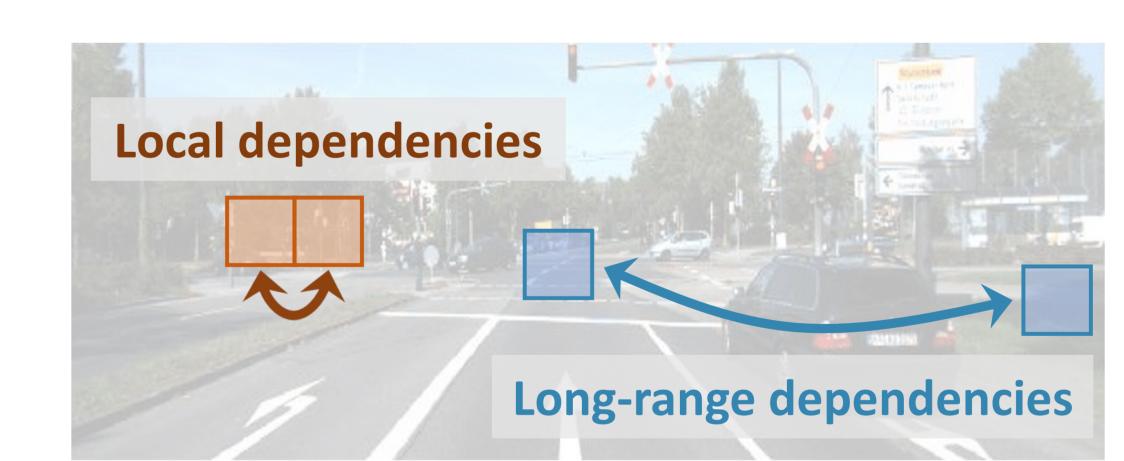


Incremental learning of local components leads to low hierarchical tensor rank!

# VI) Application: Countering Locality of CNNs via Regularization

Fact (Cohen & Shashua 2017, Levine et al. 2018)

Hierarchical tensor rank measures long-range dependencies

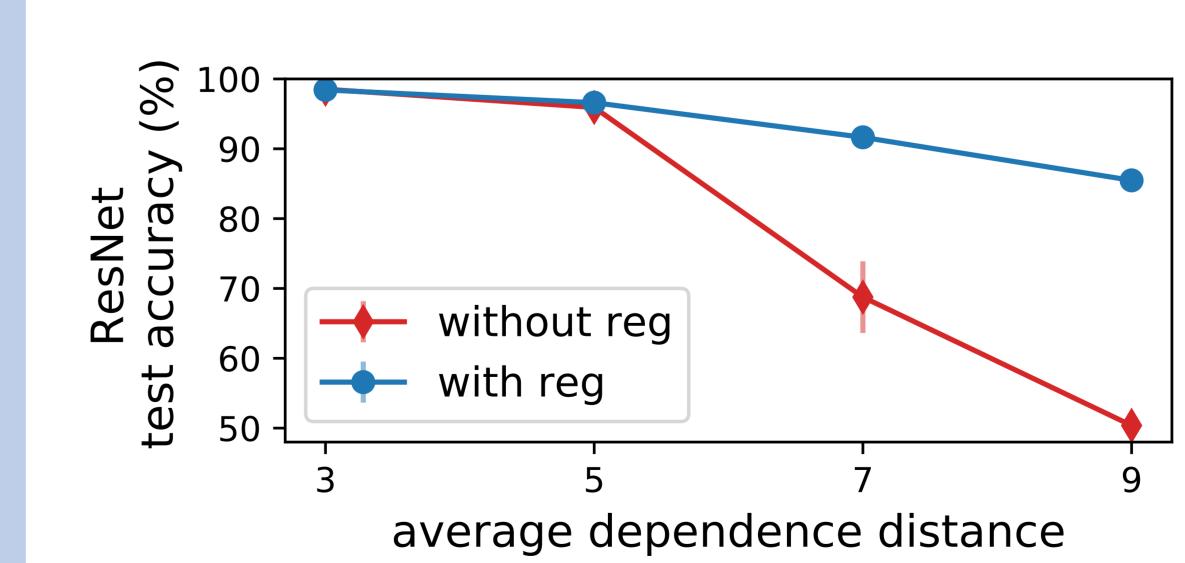


Implicit lowering of hierarchical tensor rank in HTF

Implicit lowering of long-range dependencies in CNNs!

## Can explicit regularization improve CNNs on long-range tasks?

Experiment: regularization promoting high hierarchical tensor rank



Locality of CNNs can be countered via explicit regularization!

#### VII) Takeaways

- ► Implicit reg in HTF lowers HT rank (just as in MF & TF it lowers notions of rank)
- ► This implies implicit reg towards locality in CNNs
- Specialized explicit reg improves performance of CNNs on long-range tasks!