Generalization in Deep Learning Through the Lens of Implicit Rank Lowering

Noam Razin

Tel Aviv University

ICTP Youth in High Dimensions

30 June 2022

Sources

Implicit Regularization in Deep Learning May Not Be Explainable by Norms

R + Cohen

NeurIPS 2020

Implicit Regularization in Tensor Factorization

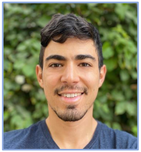
R*** + Maman* + Cohen

ICML 2021

Implicit Regularization in Hierarchical Tensor Factorization and Deep Convolutional Neural Networks

R + Maman + Cohen

ICML 2022



Asaf Maman

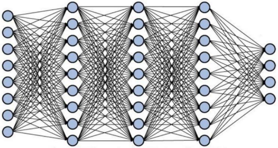
Nadav Cohen

*Equal contribution

Implicit Regularization in Deep Learning

Neural networks (NNs) generalize with no explicit regularization despite:

# of learned weights

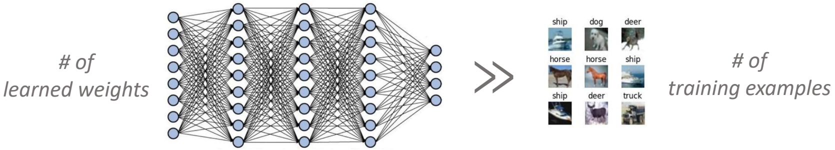


$\gg$

# of training examples

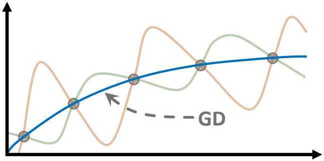
Implicit Regularization in Deep Learning

Neural networks (NNs) generalize with no explicit regularization despite:



Conventional Wisdom

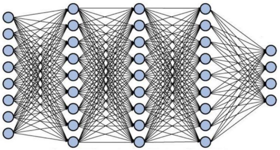
Gradient descent (GD) induces implicit regularization towards "simplicity"



Implicit Regularization in Deep Learning

Neural networks (NNs) generalize with no explicit regularization despite:

# of learned weights

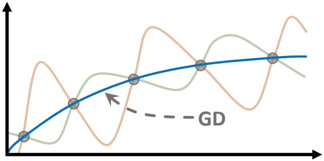


$\gg$

# of training examples

Conventional Wisdom

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

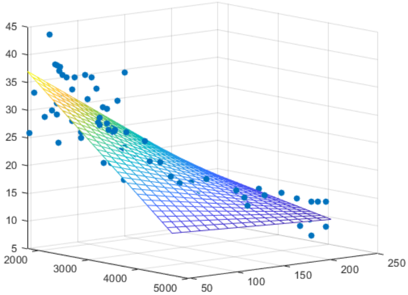


Goal

Mathematically characterize this implicit regularization

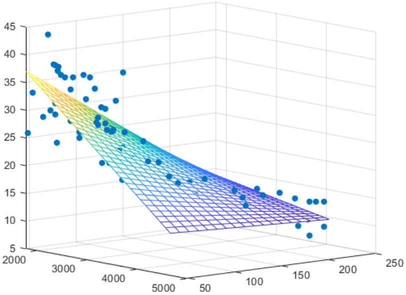
Warm Up: Linear Models

Linear Regression



Warm Up: Linear Models

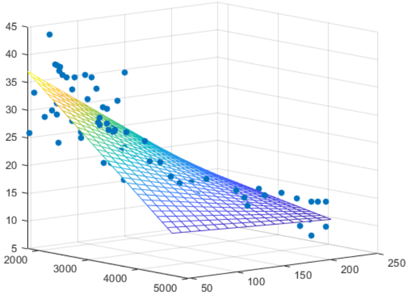
Linear Regression



When # of learned weights > # of training examples:

Warm Up: Linear Models

Linear Regression



When $\#$ of learned weights $>$ $\#$ of training examples:

GD initialized at 0 converges to min ℓ_2 norm solution

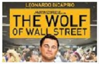
$$\underset{\mathbf{w}}{\mathrm{argmin}} \, \|\mathbf{w}\|_2 \text{ s.t. } \mathbf{w} \text{ is global min}$$

Outline

1. Matrix Factorization
   - Implicit Regularization $\neq$ Norm Minimization
2. Tensor Factorization
3. Hierarchical Tensor Factorization
4. Implications for Modern Deep Learning
5. Conclusion

Matrix Completion $\longleftrightarrow$ Two-Dimensional Prediction

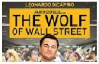
Matrix completion: recover unknown matrix given subset of entries

				
Bob	4	?	?	4
Alice	?	5	4	?
Joe	?	5	?	?

observations $\{y_{i,j}\}_{(i,j)\in\Omega}$

Matrix Completion $\longleftrightarrow$ Two-Dimensional Prediction

Matrix completion: recover unknown matrix given subset of entries

				
Bob	4	?	?	4
Alice	?	5	4	?
Joe	?	5	?	?

observations $\{y_{i,j}\}_{(i,j)\in\Omega}$

$d \times d'$ matrix completion $\longleftrightarrow$ prediction from $\{1, ..., d\} \times \{1, ..., d'\}$ to $\mathbb{R}$

value of entry (i, j) $\longleftrightarrow$ label of input (i, j)

observed entries $\longleftrightarrow$ train data

unobserved entries $\longleftrightarrow$ test data

matrix $\longleftrightarrow$ predictor

MF $\longleftrightarrow$ Linear NN

Matrix Factorization (MF):

Parameterize solution as product of matrices and fit observations via GD

$$\min_{W_1,\ldots,W_L} \ \sum\nolimits_{(i,j)\in\Omega} \left([W_L W_{L-1} \cdots W_1]_{i,j} - y_{i,j}\right)^2$$

MF $\longleftrightarrow$ Linear NN

Matrix Factorization (MF):

Parameterize solution as product of matrices and fit observations via GD

$$\min_{W_1,\ldots,W_L} \sum_{(i,j)\in\Omega} \big([W_L W_{L-1} \cdots W_1]_{i,j} - y_{i,j}\big)^2$$

$$\uparrow$$

hidden dimensions large enough to not limit rank

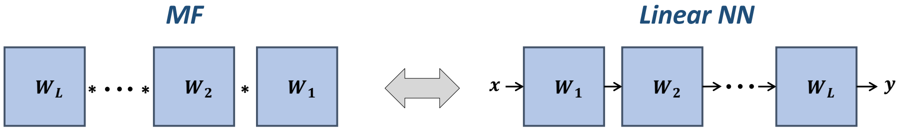
MF $\longleftrightarrow$ Linear NN

Matrix Factorization (MF):

Parameterize solution as product of matrices and fit observations via GD

$$\min_{W_1,\ldots,W_L} \sum_{(i,j)\in\Omega} \left([W_L W_{L-1} \cdots W_1]_{i,j} - y_{i,j}\right)^2$$

hidden dimensions large enough to not limit rank



MF $\longleftrightarrow$ Linear NN

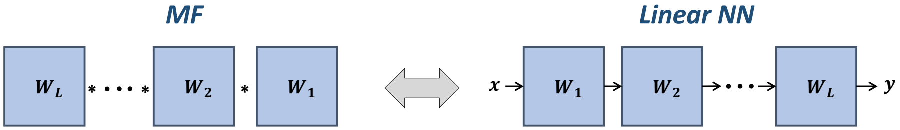
Matrix Factorization (MF):

Parameterize solution as product of matrices and fit observations via GD

$$\min_{W_1,\ldots,W_L} \sum_{(i,j)\in\Omega} \left([W_L W_{L-1} \cdots W_1]_{i,j} - y_{i,j} \right)^2$$

$\uparrow$

hidden dimensions large enough to not limit rank



Empirical Phenomenon (Gunasekar et al. 2017)

MF (with small init and step size) accurately recovers low rank matrices

MF $\longleftrightarrow$ Linear NN

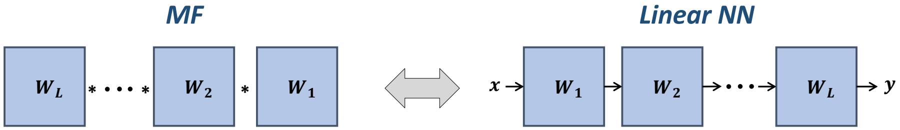
Matrix Factorization (MF):

Parameterize solution as product of matrices and fit observations via GD

$$\min_{W_1,\ldots,W_L} \sum_{(i,j)\in\Omega} \left([W_L W_{L-1} \cdots W_1]_{i,j} - y_{i,j} \right)^2$$

$\uparrow$

hidden dimensions large enough to not limit rank



Empirical Phenomenon (Gunasekar et al. 2017)

MF (with small init and step size) accurately recovers low rank matrices

Conjecture (Gunasekar et al. 2017)

Gradient flow over MF with small init $\Longrightarrow$ min nuclear norm solution

Dynamical Analysis of Implicit Regularization in MF

Dynamical Analysis of Implicit Regularization in MF

$W_e := W_L \cdots W_1$ — end matrix $\left\{\sigma_M^{(r)}\right\}_r$ — singular values of W_e

Dynamical Analysis of Implicit Regularization in MF

$W_e := W_L \cdots W_1$ — end matrix $\left\{\sigma_M^{(r)}\right\}_r$ — singular values of W_e

Theorem (Arora et al. 2019)

When training MF with near-zero init: $\frac{d}{dt}\sigma_M^{(r)}(t) \propto \sigma_M^{(r)}(t)^{2-\frac{2}{L}}$

Dynamical Analysis of Implicit Regularization in MF

$W_e := W_L \cdots W_1$ — end matrix $\left\{\sigma_M^{(r)}\right\}_r$ — singular values of W_e

Theorem (Arora et al. 2019)

When training MF with near-zero init: $\frac{d}{dt}\sigma_M^{(r)}(t) \propto \sigma_M^{(r)}(t)^{2-\frac{2}{L}}$

Singular values move slower when small and faster when large!

Dynamical Analysis of Implicit Regularization in MF

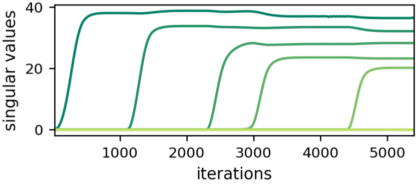
$W_e := W_L \cdots W_1$ — end matrix $\left\{\sigma_M^{(r)}\right\}_r$ — singular values of W_e

Theorem (Arora et al. 2019)

When training MF with near-zero init: $\frac{d}{dt}\sigma_M^{(r)}(t) \propto \sigma_M^{(r)}(t)^{2-\frac{2}{L}}$

Singular values move slower when small and faster when large!

Experiment: completion of low rank matrix via MF



Dynamical Analysis of Implicit Regularization in MF

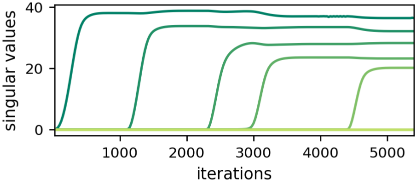
$W_e := W_L \cdots W_1$ — end matrix $\left\{ \sigma_M^{(r)} \right\}_r$ — singular values of W_e

Theorem (Arora et al. 2019)

When training MF with near-zero init: $\frac{d}{dt}\sigma_M^{(r)}(t) \propto \sigma_M^{(r)}(t)^{2-\frac{2}{L}}$

Singular values move slower when small and faster when large!

Experiment: completion of low rank matrix via MF



Incremental learning of singular values leads to low rank

Dynamical Analysis of Implicit Regularization in MF

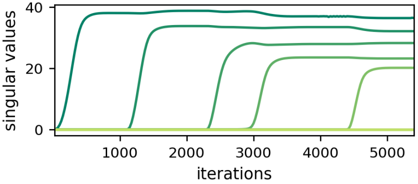
$W_e := W_L \cdots W_1$ — end matrix $\left\{\sigma_M^{(r)}\right\}_r$ — singular values of W_e

Theorem (Arora et al. 2019)

When training MF with near-zero init: $\frac{d}{dt}\sigma_M^{(r)}(t) \propto \sigma_M^{(r)}(t)^{2-\frac{2}{L}}$

Singular values move slower when small and faster when large!

Experiment: completion of low rank matrix via MF



Incremental learning of singular values leads to low rank

Conjecture (Arora et al. 2019)

For any $\|\cdot\|$, *exist observations for which MF* $\not\Rightarrow$ *min* $\|\cdot\|$ *solution*

Outline

Implicit Regularization in MF $\neq$ Norm Minimization

Does the implicit regularization in MF minimize a norm?

Implicit Regularization in MF $\neq$ Norm Minimization

Does the implicit regularization in MF minimize a norm?

Theorem

There exist matrix completion settings where MF drives all norms to ∞ while effective rank is minimized

Implicit Regularization in MF $\neq$ Norm Minimization

Does the implicit regularization in MF minimize a norm?

Theorem

There exist matrix completion settings where MF drives all norms to ∞ while effective rank is minimized

Implicit regularization in MF$\neq$ norm minimization

Implicit Regularization in MF $\neq$ Norm Minimization

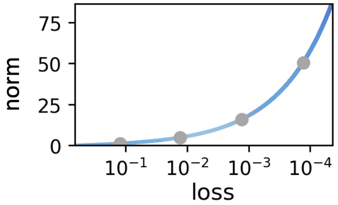
Does the implicit regularization in MF minimize a norm?

Theorem

There exist matrix completion settings where MF drives all norms to ∞ while effective rank is minimized

Implicit regularization in MF$\neq$ norm minimization

Experiment



Implicit Regularization in MF $\neq$ Norm Minimization

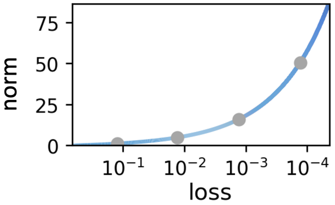
Does the implicit regularization in MF minimize a norm?

Theorem

There exist matrix completion settings where MF drives all norms to ∞ while effective rank is minimized

Implicit regularization in MF$\neq$ norm minimization

Experiment



Chou et al. 2020, Li et al. 2021: further support for implicit rank minimization

Outline

Drawbacks of Studying MF

$$\begin{bmatrix} 4 & ? & ? & 4 \\ ? & 5 & 4 & ? \\ ? & 5 & ? & ? \end{bmatrix} = \boldsymbol{W_L} * \cdots * \boldsymbol{W_2} * \boldsymbol{W_1}$$

As a surrogate for deep learning, MF is limited:

Drawbacks of Studying MF

$$\begin{bmatrix} 4 & ? & ? & 4 \\ ? & 5 & 4 & ? \\ ? & 5 & ? & ? \end{bmatrix} = \boldsymbol{W_L} * \cdots * \boldsymbol{W_2} * \boldsymbol{W_1}$$

As a surrogate for deep learning, MF is limited:

(1) Misses non-linearity

Drawbacks of Studying MF

$$\begin{bmatrix} 4 & ? & ? & 4 \\ ? & 5 & 4 & ? \\ ? & 5 & ? & ? \end{bmatrix} = \boxed{W_L} * \cdots * \boxed{W_2} * \boxed{W_1}$$

As a surrogate for deep learning, MF is limited:

(1) Misses non-linearity

(2) Does not capture prediction with more than 2 input variables

Drawbacks of Studying MF

$$\begin{bmatrix} 4 & ? & ? & 4 \\ ? & 5 & 4 & ? \\ ? & 5 & ? & ? \end{bmatrix} = \boldsymbol{W_L} * \cdots * \boldsymbol{W_2} * \boldsymbol{W_1}$$

As a surrogate for deep learning, MF is limited:

(1) Misses non-linearity

(2) Does not capture prediction with more than 2 input variables

> Tensor factorization accounts for both **(1)** and **(2)**

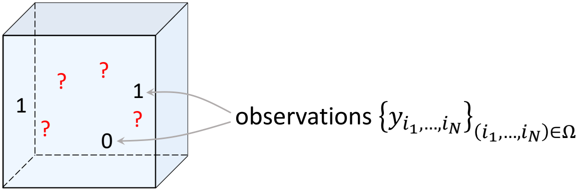
Tensor Completion $\longleftrightarrow$ Multi-Dimensional Prediction

Tensor: N-dimensional array ($N =$ order of tensor)

Tensor Completion ⟷ Multi-Dimensional Prediction

Tensor: N-dimensional array (N = order of tensor)

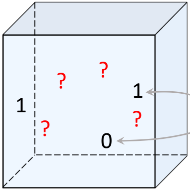
Tensor completion: recover unknown tensor given subset of entries



Tensor Completion $\longleftrightarrow$ Multi-Dimensional Prediction

Tensor: N-dimensional array ($N =$ order of tensor)

Tensor completion: recover unknown tensor given subset of entries



observations $\{y_{i_1,\ldots,i_N}\}_{(i_1,\ldots,i_N)\in\Omega}$

$[d_j] := \{1, \ldots, d_j\}$

$d_1 \times \cdots \times d_N$ tensor completion $\longleftrightarrow$ prediction from $[d_1] \times \cdots \times [d_N]$ to $\mathbb{R}$

value of entry $(i_1, \ldots, i_N)$	$\longleftrightarrow$	label of input $(i_1, \ldots, i_N)$
observed entries	$\longleftrightarrow$	train data
unobserved entries	$\longleftrightarrow$	test data
tensor	$\longleftrightarrow$	predictor

TF $\longleftrightarrow$ Shallow Non-Linear Convolutional NN

Tensor Factorization (TF):

Parameterize solution as sum of outer products and fit observations via GD

$$\min_{\{\mathbf{w}_r^n\}_{r,n}} \sum_{(i_1,\ldots,i_N)\in\Omega} \left(\left[\sum_{r=1}^{R} \mathbf{w}_r^1 \otimes \cdots \otimes \mathbf{w}_r^N \right]_{i_1,\ldots,i_N} - y_{i_1,\ldots,i_N} \right)^2$$

TF $\longleftrightarrow$ Shallow Non-Linear Convolutional NN

Tensor Factorization (TF):

Parameterize solution as sum of outer products and fit observations via GD

$$\min_{\{\mathbf{w}_r^n\}_{r,n}} \sum\nolimits_{(i_1,\ldots,i_N)\in\Omega} \Big(\big[\sum\nolimits_{r=1}^{R} \mathbf{w}_r^1 \otimes \cdots \otimes \mathbf{w}_r^N \big]_{i_1,\ldots,i_N} - y_{i_1,\ldots,i_N} \Big)^2$$

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

TF $\longleftrightarrow$ Shallow Non-Linear Convolutional NN

Tensor Factorization (TF):

Parameterize solution as sum of outer products and fit observations via GD

$$\min_{\{\mathbf{w}_r^n\}_{r,n}} \sum\nolimits_{(i_1,\ldots,i_N)\in\Omega} \Big(\big[\underset{\uparrow}{\textstyle\sum_{r=1}^{R}} \mathbf{w}_r^1 \otimes \cdots \otimes \mathbf{w}_r^N \big]_{i_1,\ldots,i_N} - y_{i_1,\ldots,i_N} \Big)^2$$

R large enough to not constrain tensor rank

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

TF $\longleftrightarrow$ Shallow Non-Linear Convolutional NN

Tensor Factorization (TF):

Parameterize solution as sum of outer products and fit observations via GD

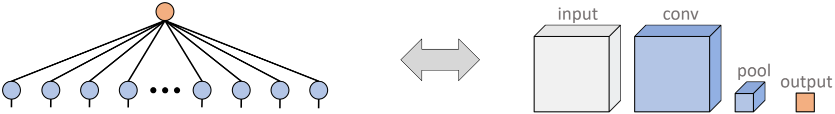
$$\min_{\{\mathbf{w}_r^n\}_{r,n}} \sum_{(i_1,\ldots,i_N)\in\Omega} \Big(\big[\underset{\substack{\uparrow}}{\textstyle\sum_{r=1}^{R}} \mathbf{w}_r^1 \otimes \cdots \otimes \mathbf{w}_r^N \big]_{i_1,\ldots,i_N} - y_{i_1,\ldots,i_N} \Big)^2$$

R large enough to not constrain tensor rank

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

TF

Shallow Non-Linear Convolutional NN (CNN)



TF $\longleftrightarrow$ Shallow Non-Linear Convolutional NN

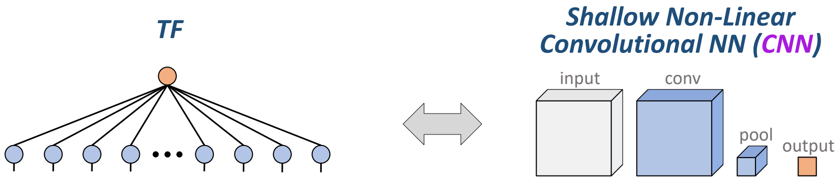
Tensor Factorization (TF):

Parameterize solution as sum of outer products and fit observations via GD

$$\min_{\{\mathbf{w}_r^n\}_{r,n}} \sum\nolimits_{(i_1,\ldots,i_N)\in\Omega} \Big(\big[\underbrace{\textstyle\sum_{r=1}^{R}}_{\uparrow} \mathbf{w}_r^1 \otimes \cdots \otimes \mathbf{w}_r^N \big]_{i_1,\ldots,i_N} - y_{i_1,\ldots,i_N} \Big)^2$$

R large enough to not constrain tensor rank

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



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

Dynamical Analysis of Implicit Regularization in TF

$\sigma_T^{(r)}(t) := \|\otimes_{n=1}^{N} \mathbf{w}_r^n(t)\|_F$ — Frobenius norm of r'th component

Dynamical Analysis of Implicit Regularization in TF

$\sigma_T^{(r)}(t) := \|\otimes_{n=1}^N \mathbf{w}_r^n(t)\|_F$ — Frobenius norm of r'th component

Theorem

When training TF with near-zero init: $\frac{d}{dt}\sigma_T^{(r)}(t) \propto \sigma_T^{(r)}(t)^{2-\frac{2}{N}}$

Dynamical Analysis of Implicit Regularization in TF

$\sigma_T^{(r)}(t) := \|\otimes_{n=1}^N \mathbf{w}_r^n(t)\|_F$ — Frobenius norm of r'th component

Theorem

When training TF with near-zero init: $\frac{d}{dt}\sigma_T^{(r)}(t) \propto \sigma_T^{(r)}(t)^{2-\frac{2}{N}}$

Component norms move slower when small and faster when large!

Dynamical Analysis of Implicit Regularization in TF

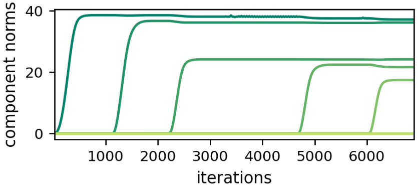
$\sigma_T^{(r)}(t) := \|\otimes_{n=1}^{N} \mathbf{w}_r^n(t)\|_F$ — Frobenius norm of r'th component

Theorem

When training TF with near-zero init: $\frac{d}{dt}\sigma_T^{(r)}(t) \propto \sigma_T^{(r)}(t)^{2-\frac{2}{N}}$

Component norms move slower when small and faster when large!

Experiment: completion of low tensor rank tensor via TF



Dynamical Analysis of Implicit Regularization in TF

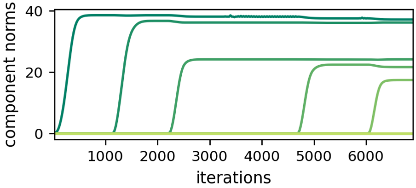
$\sigma_T^{(r)}(t) := \|\otimes_{n=1}^N \mathbf{w}_r^n(t)\|_F$ — Frobenius norm of r'th component

Theorem

When training TF with near-zero init: $\frac{d}{dt}\sigma_T^{(r)}(t) \propto \sigma_T^{(r)}(t)^{2-\frac{2}{N}}$

Component norms move slower when small and faster when large!

Experiment: completion of low tensor rank tensor via TF



Incremental learning of components leads to low tensor rank!

Dynamical Analysis of Implicit Regularization in TF

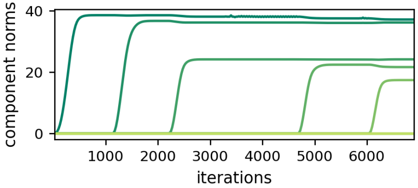
$\sigma_T^{(r)}(t) := \|\otimes_{n=1}^{N} \mathbf{w}_r^n(t)\|_F$ — Frobenius norm of r'th component

Theorem

When training TF with near-zero init: $\frac{d}{dt}\sigma_T^{(r)}(t) \propto \sigma_T^{(r)}(t)^{2-\frac{2}{N}}$

Component norms move slower when small and faster when large!

Experiment: completion of low tensor rank tensor via TF



Incremental learning of components leads to low tensor rank!

Theorem (under technical conditions)

If tensor completion has tensor rank 1 solution, then TF will reach it

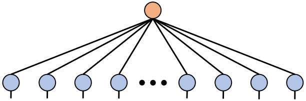
Outline

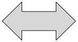
1. Matrix Factorization
   - Implicit Regularization $\neq$ Norm Minimization
2. Tensor Factorization
3. **Hierarchical Tensor Factorization**
4. Implications for Modern Deep Learning
5. Conclusion

HTF $\longleftrightarrow$ Deep Non-Linear CNN

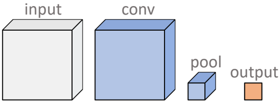
Limitation of TF: does not account for depth

TF
(sum of components)



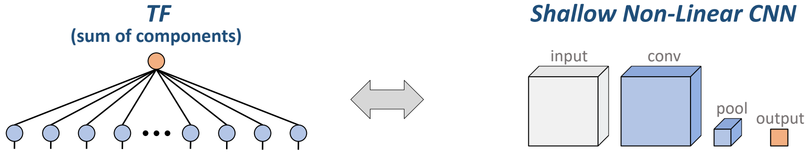


Shallow Non-Linear CNN

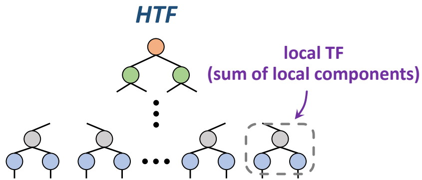


HTF $\longleftrightarrow$ Deep Non-Linear CNN

Limitation of TF: does not account for depth

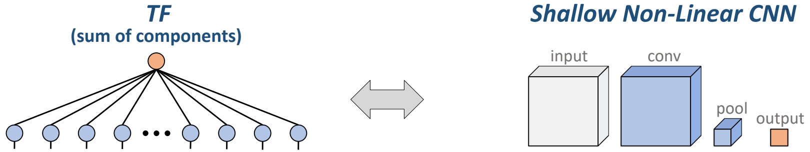


Hierarchical Tensor Factorization (HTF):

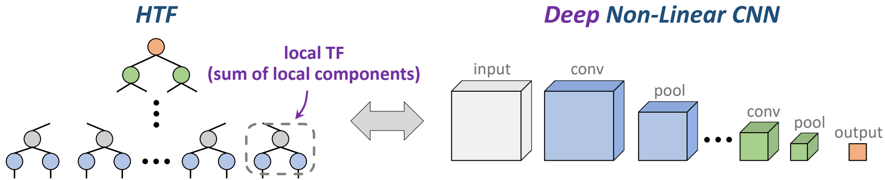


HTF $\longleftrightarrow$ Deep Non-Linear CNN

Limitation of TF: does not account for depth



Hierarchical Tensor Factorization (HTF):

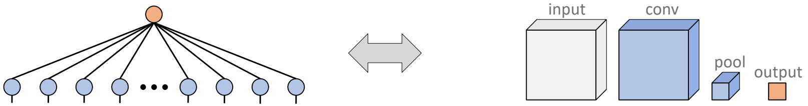


HTF $\longleftrightarrow$ Deep Non-Linear CNN

Limitation of TF: does not account for depth

TF
(sum of components)

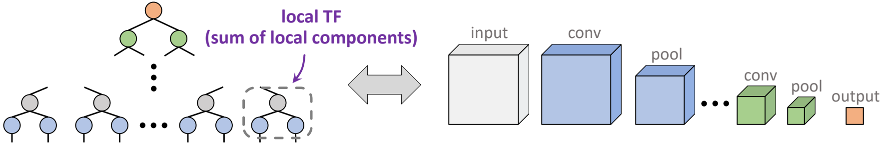
Shallow Non-Linear CNN



Hierarchical Tensor Factorization (HTF):

HTF

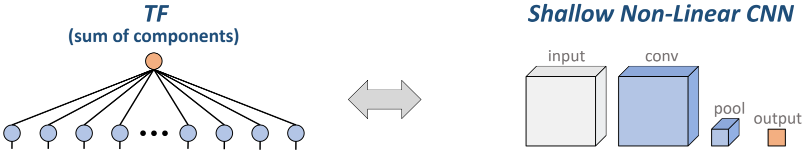
Deep Non-Linear CNN



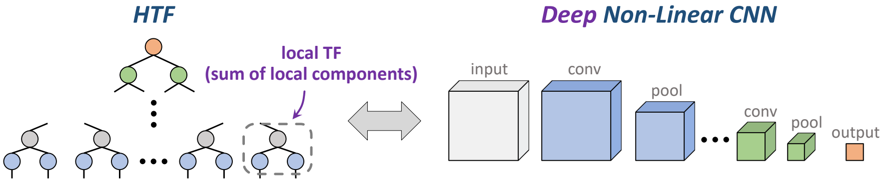
Equivalence studied extensively (e.g. Cohen et al. 2016, Levine et al. 2018, Khrulkov et al. 2018)

HTF $\longleftrightarrow$ Deep Non-Linear CNN

Limitation of TF: does not account for depth



Hierarchical Tensor Factorization (HTF):



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

Representation w/ few local components $\Longrightarrow$ low hierarchical tensor rank

Dynamical Analysis of Implicit Regularization in HTF

$\sigma_H^{(r)}(t)$ — Frobenius norm of r'th local component in a location of HTF

Dynamical Analysis of Implicit Regularization in HTF

$\sigma_H^{(r)}(t)$ — Frobenius norm of r'th local component in a location of HTF

K — order of local component

Dynamical Analysis of Implicit Regularization in HTF

$\sigma_H^{(r)}(t)$ — Frobenius norm of r'th local component in a location of HTF

K — order of local component

Theorem

When training HTF with near-zero init: $\frac{d}{dt}\sigma_H^{(r)}(t) \propto \sigma_H^{(r)}(t)^{2-\frac{2}{K}}$

Dynamical Analysis of Implicit Regularization in HTF

$\sigma_H^{(r)}(t)$ — Frobenius norm of r'th local component in a location of HTF

K — order of local component

Theorem

When training HTF with near-zero init: $\frac{d}{dt}\sigma_H^{(r)}(t) \propto \sigma_H^{(r)}(t)^{2-\frac{2}{K}}$

Local component norms move slower when small and faster when large!

Dynamical Analysis of Implicit Regularization in HTF

$\sigma_H^{(r)}(t)$ — Frobenius norm of r'th local component in a location of HTF

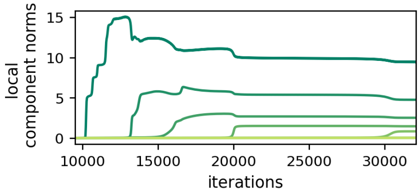
K — order of local component

Theorem

When training HTF with near-zero init: $\frac{d}{dt}\sigma_H^{(r)}(t) \propto \sigma_H^{(r)}(t)^{2-\frac{2}{K}}$

Local component norms move slower when small and faster when large!

Experiment: completion of low hierarchical tensor rank tensor via HTF



Dynamical Analysis of Implicit Regularization in HTF

$\sigma_H^{(r)}(t)$ — Frobenius norm of r'th local component in a location of HTF

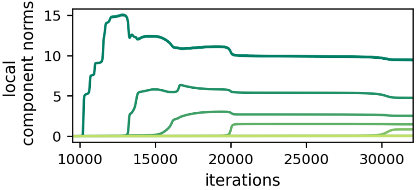
K — order of local component

Theorem

When training HTF with near-zero init: $\frac{d}{dt}\sigma_H^{(r)}(t) \propto \sigma_H^{(r)}(t)^{2-\frac{2}{K}}$

Local component norms move slower when small and faster when large!

Experiment: completion of low hierarchical tensor rank tensor via HTF

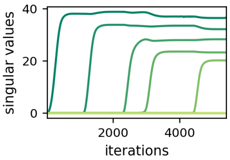


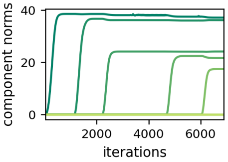
Incremental learning of local components leads to low hierarchical tensor rank!

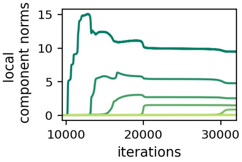
Implicit Regularization in MF / TF / HTF Are Analogous

	MF	**TF**	**HTF**
Quantity	singular values	component norms	local component norms
Dynamics	$\frac{d}{dt}\sigma_M^{(r)}(t) \propto \sigma_M^{(r)}(t)^{2-\frac{2}{L}}$	$\frac{d}{dt}\sigma_T^{(r)}(t) \propto \sigma_T^{(r)}(t)^{2-\frac{2}{N}}$	$\frac{d}{dt}\sigma_H^{(r)}(t) \propto \sigma_H^{(r)}(t)^{2-\frac{2}{K}}$

Experiment





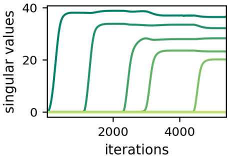


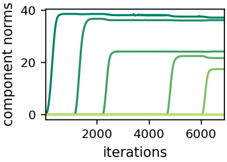
	MF	**TF**	**HTF**
Incremental learning lowers	**rank**	**tensor rank**	**hierarchical tensor rank**

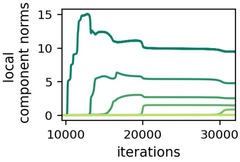
Implicit Regularization in MF / TF / HTF Are Analogous

	MF	**TF**	**HTF**
Quantity	singular values	component norms	local component norms
Dynamics	$\frac{d}{dt}\sigma_M^{(r)}(t) \propto \sigma_M^{(r)}(t)^{2-\frac{2}{L}}$	$\frac{d}{dt}\sigma_T^{(r)}(t) \propto \sigma_T^{(r)}(t)^{2-\frac{2}{N}}$	$\frac{d}{dt}\sigma_H^{(r)}(t) \propto \sigma_H^{(r)}(t)^{2-\frac{2}{K}}$

Experiment







	MF	**TF**	**HTF**
Incremental learning lowers	**rank**	**tensor rank**	**hierarchical tensor rank**

Implicit regularization in MF / TF / HTF are structurally identical!

Outline

1. Matrix Factorization
   - Implicit Regularization $\neq$ Norm Minimization
2. Tensor Factorization
3. Hierarchical Tensor Factorization
4. Implications for Modern Deep Learning
5. Conclusion

Practical Application: Rank Lowering in NN Layers

Practical Application: Rank Lowering in NN Layers

Parameterize layers of NN as MF / TF / HTF

Practical Application: Rank Lowering in NN Layers

Parameterize layers of NN as MF / TF / HTF

$\Longrightarrow$ implicit rank lowering induces compressibility and generalization

Practical Application: Rank Lowering in NN Layers

Parameterize layers of NN as MF / TF / HTF

$\Longrightarrow$ implicit rank lowering induces compressibility and generalization

Implicit Rank-Minimizing Autoencoder

Li Jing
Facebook AI Research
New York

Jure Zbontar
Facebook AI Research
New York

Yann LeCun
Facebook AI Research
New York

ExpandNets: Linear Over-parameterization to Train Compact Convolutional Networks

Shuxuan Guo
CVLab, EPFL

Jose M. Alvarez
NVIDIA

Mathieu Salzmann
CVLab, EPFL

THE LOW-RANK SIMPLICITY BIAS IN DEEP NETWORKS

Minyoung Huh
MIT CSAIL

Hossein Mobahi
Google Research

Richard Zhang
Adobe Research

Brian Cheung
MIT CSAIL & BCS

Pulkit Agrawal
MIT CSAIL

Phillip Isola
MIT CSAIL

Understanding Generalization in Deep Learning via Tensor Methods

Jingling Li[1,3] **Yanchao Sun**[1] **Jiahao Su**[4] **Taiji Suzuki**[2,3] **Furong Huang**[1]

[1]Department of Computer Science, University of Maryland, College Park
[2]Graduate School of Information Science and Technology, The University of Tokyo
[3]Center for Advanced Intelligence Project, RIKEN
[4]Department of Electrical and Computer Engineering, University of Maryland, College Park

Potential Explanation for Generalization on Natural Data

Potential Explanation for Generalization on Natural Data

Challenge

Find complexity measures that:

Potential Explanation for Generalization on Natural Data

Challenge

Find complexity measures that:

- Are implicitly lowered by GD over NNs

Potential Explanation for Generalization on Natural Data

Challenge

Find complexity measures that:

- Are implicitly lowered by GD over NNs
- Capture essence of natural data (allow its fit with low complexity)

Potential Explanation for Generalization on Natural Data

Challenge

Find complexity measures that:

- Are implicitly lowered by GD over NNs
- Capture essence of natural data (allow its fit with low complexity)

Can ranks serve as measures of complexity?

Potential Explanation for Generalization on Natural Data

Challenge

Find complexity measures that:

- Are implicitly lowered by GD over NNs
- Capture essence of natural data (allow its fit with low complexity)

Can ranks serve as measures of complexity?

Experiment

MNIST & FMNIST can be fit with low (hierarchical) tensor rank

Potential Explanation for Generalization on Natural Data

Challenge

Find complexity measures that:

- Are implicitly lowered by GD over NNs
- Capture essence of natural data (allow its fit with low complexity)

Can ranks serve as measures of complexity?

Experiment

MNIST & FMNIST can be fit with low (hierarchical) tensor rank

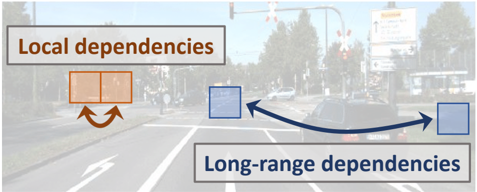
Implicit lowering of ranks may explain generalization on natural data!

Countering Locality of CNNs via Regularization

Countering Locality of CNNs via Regularization

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

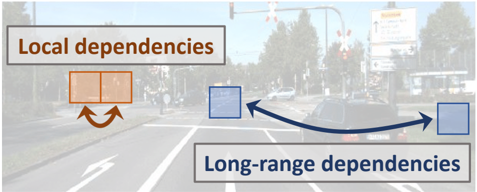
Hierarchical tensor rank measures long-range dependencies



Countering Locality of CNNs via Regularization

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

Hierarchical tensor rank measures long-range dependencies

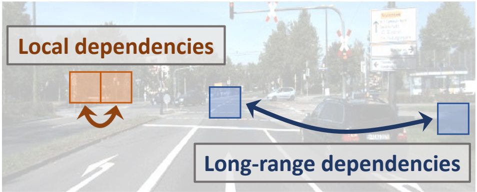


Implicit lowering of
hierarchical tensor rank in HTF
$\Updownarrow$
Implicit lowering of
long-range dependencies in CNNs!

Countering Locality of CNNs via Regularization

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

Hierarchical tensor rank measures long-range dependencies



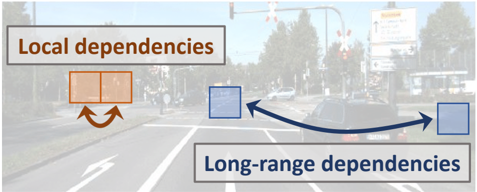
Implicit lowering of
hierarchical tensor rank in HTF
$\Updownarrow$
Implicit lowering of
long-range dependencies in CNNs!

Can explicit regularization improve CNNs on long-range tasks?

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

$\Updownarrow$

Implicit lowering of
long-range dependencies in CNNs!

Can explicit regularization improve CNNs on long-range tasks?

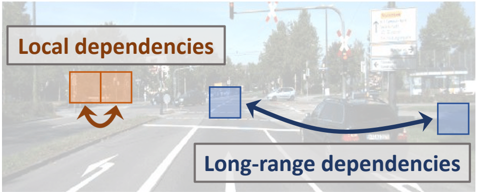
Experiments

Regularization promoting high hierarchical tensor rank

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
$\Updownarrow$
Implicit lowering of
long-range dependencies in CNNs!

Can explicit regularization improve CNNs on long-range tasks?

Experiments

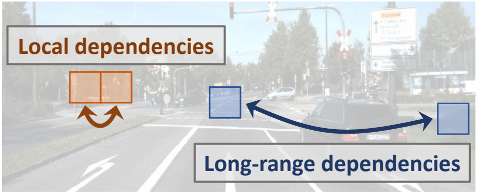
Regularization promoting high hierarchical tensor rank

$\Longrightarrow$ improves CNNs (e.g. ResNets) on long-range tasks!

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
$\Updownarrow$
Implicit lowering of
long-range dependencies in CNNs!

Can explicit regularization improve CNNs on long-range tasks?

Experiments

Regularization promoting high hierarchical tensor rank

$\Longrightarrow$ improves CNNs (e.g. ResNets) on long-range tasks!

Locality of CNNs can be countered via explicit regularization!

Outline

Recap

Goal: understand implicit regularization in deep learning

Recap

Goal: understand implicit regularization in deep learning

Matrix Factorization (Linear NN):

- *Existing conjecture:* implicit regularization minimizes norm

Recap

Goal: understand implicit regularization in deep learning

Matrix Factorization (Linear NN):

- *Existing conjecture:* implicit regularization minimizes norm
- *We showed:* it can drive all norms to ∞ while minimizing rank

Recap

Goal: understand implicit regularization in deep learning

Matrix Factorization (Linear NN):

- *Existing conjecture:* implicit regularization minimizes norm
- *We showed:* it can drive all norms to ∞ while minimizing rank

Tensor and Hierarchical Tensor Factorizations (Non-Linear CNNs):

- *We showed:* implicit regularization lowers tensorial ranks

Recap

Goal: understand implicit regularization in deep learning

Matrix Factorization (Linear NN):

- *Existing conjecture:* implicit regularization minimizes norm
- *We showed:* it can drive all norms to ∞ while minimizing rank

Tensor and Hierarchical Tensor Factorizations (Non-Linear CNNs):

- *We showed:* implicit regularization lowers tensorial ranks

Implications to Modern Deep Learning:

Recap

Goal: understand implicit regularization in deep learning

Matrix Factorization (Linear NN):

- *Existing conjecture:* implicit regularization minimizes norm
- *We showed:* it can drive all norms to ∞ while minimizing rank

Tensor and Hierarchical Tensor Factorizations (Non-Linear CNNs):

- *We showed:* implicit regularization lowers tensorial ranks

Implications to Modern Deep Learning:

- Parameterizing layers of NN as MF / TF/ HTF $\Longrightarrow$ compression

Recap

Goal: understand implicit regularization in deep learning

Matrix Factorization (Linear NN):

- *Existing conjecture:* implicit regularization minimizes norm
- *We showed:* it can drive all norms to ∞ while minimizing rank

Tensor and Hierarchical Tensor Factorizations (Non-Linear CNNs):

- *We showed:* implicit regularization lowers tensorial ranks

Implications to Modern Deep Learning:

- Parameterizing layers of NN as MF / TF/ HTF $\Longrightarrow$ compression
- Rank lowering may explain generalization on natural data

Recap

Goal: understand implicit regularization in deep learning

Matrix Factorization (Linear NN):

- *Existing conjecture:* implicit regularization minimizes norm
- *We showed:* it can drive all norms to ∞ while minimizing rank

Tensor and Hierarchical Tensor Factorizations (Non-Linear CNNs):

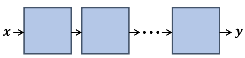
- *We showed:* implicit regularization lowers tensorial ranks

Implications to Modern Deep Learning:

- Parameterizing layers of NN as MF / TF/ HTF $\Longrightarrow$ compression
- Rank lowering may explain generalization on natural data
- One may counter locality of CNNs via explicit regularization!

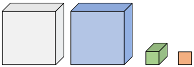
Implicit Rank Lowering in Deep Learning

Linear NN



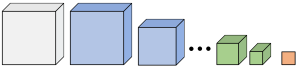
matrix rank lowering

Shallow Non-Linear CNN



tensor rank lowering

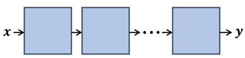
Deep Non-Linear CNN



hierarchical tensor rank lowering

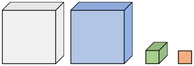
Implicit Rank Lowering in Deep Learning

Linear NN



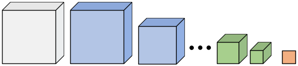
matrix rank lowering

Shallow Non-Linear CNN

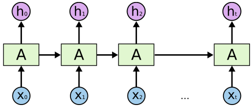


tensor rank lowering

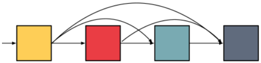
Deep Non-Linear CNN



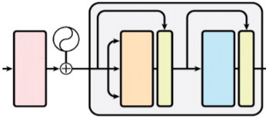
hierarchical tensor rank lowering



??? rank lowering



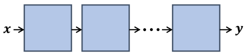
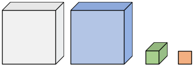
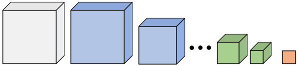
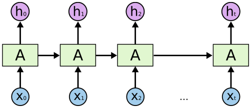
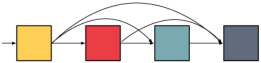
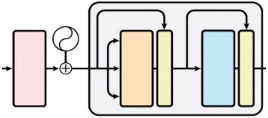
??? rank lowering



??? rank lowering

Hypothesis: in each NN architecture implicit regularization lowers corresponding notion of rank

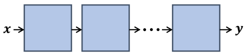
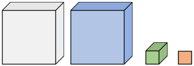
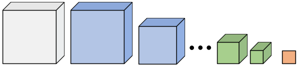
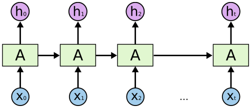
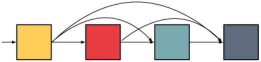
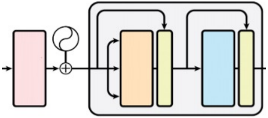
Implicit Rank Lowering in Deep Learning

Linear NN	*Shallow Non-Linear CNN*	*Deep Non-Linear CNN*
		
matrix rank lowering	**tensor rank lowering**	**hierarchical tensor rank lowering**
		
??? rank lowering	**??? rank lowering**	**??? rank lowering**

Hypothesis: in each NN architecture implicit regularization lowers corresponding notion of rank

Discovering lowered notions of rank may pave way to:

Implicit Rank Lowering in Deep Learning

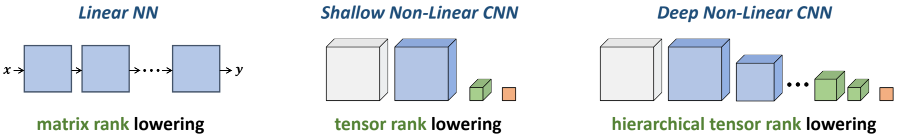
Linear NN	*Shallow Non-Linear CNN*	*Deep Non-Linear CNN*
		
matrix rank lowering	**tensor rank lowering**	**hierarchical tensor rank lowering**
		
??? rank lowering	**??? rank lowering**	**??? rank lowering**

Hypothesis: in each NN architecture implicit regularization lowers corresponding notion of rank

Discovering lowered notions of rank may pave way to:

- Explaining generalization

Implicit Rank Lowering in Deep Learning



Hypothesis: in each NN architecture implicit regularization lowers corresponding notion of rank

Discovering lowered notions of rank may pave way to:

- Explaining generalization
- Enhancing performance via regularization and architecture design

Thank You!

Work supported by:
Apple Scholars in AI/ML PhD fellowship, Google Research Scholar Award, Google Research Gift, the Yandex Initiative in Machine Learning, the Israel Science Foundation (grant 1780/21), Len Blavatnik and the Blavatnik Family Foundation, Tel Aviv University Center for AI and Data Science, and Amnon and Anat Shashua.