

Generalization in Deep Learning Through the Lens of Implicit Rank Minimization

Noam Razin

Tel Aviv University



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

Outline

1 Implicit Regularization in Deep Learning

2 Matrix Factorization

- Implicit Regularization \neq Norm Minimization

3 Tensor Factorization

4 Hierarchical Tensor Factorization

5 Implications for Modern Deep Learning

6 Conclusion

Generalization via Bias-Variance Tradeoff

Classically, generalization is understood via the bias-variance tradeoff



Generalization via Bias-Variance Tradeoff

Classically, generalization is understood via the bias-variance tradeoff



Tradeoff can be controlled through:

Generalization via Bias-Variance Tradeoff

Classically, generalization is understood via the bias-variance tradeoff



Tradeoff can be controlled through:

- Limiting model size

Generalization via Bias-Variance Tradeoff

Classically, generalization is understood via the bias-variance tradeoff

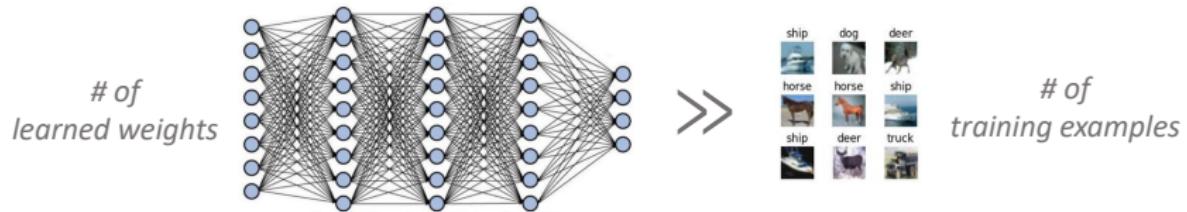


Tradeoff can be controlled through:

- Limiting model size
- Adding regularization (e.g. ℓ_2 penalty)

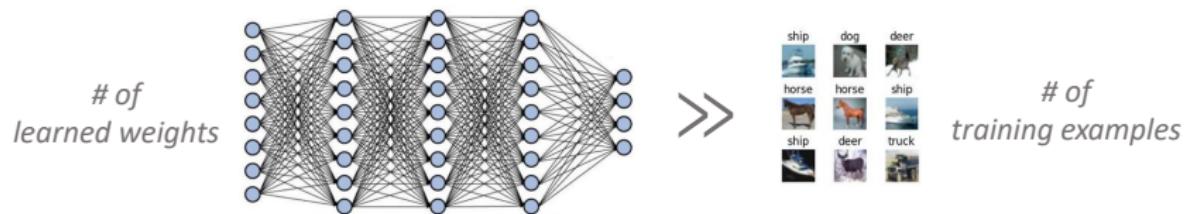
Generalization in Deep Learning

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



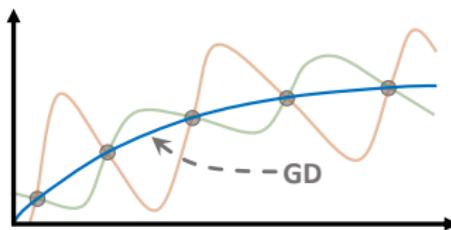
Generalization in Deep Learning

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



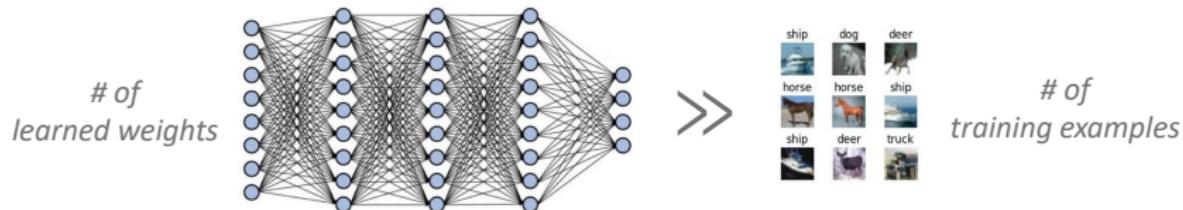
Conventional Wisdom

Gradient descent (**GD**) induces **implicit regularization** towards “simplicity”



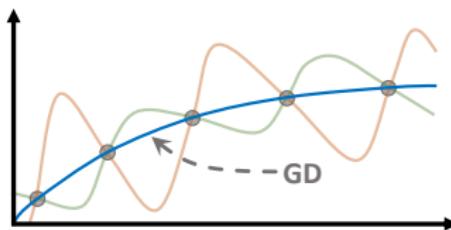
Generalization in Deep Learning

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



Conventional Wisdom

Gradient descent (GD) induces **implicit regularization** towards “simplicity”

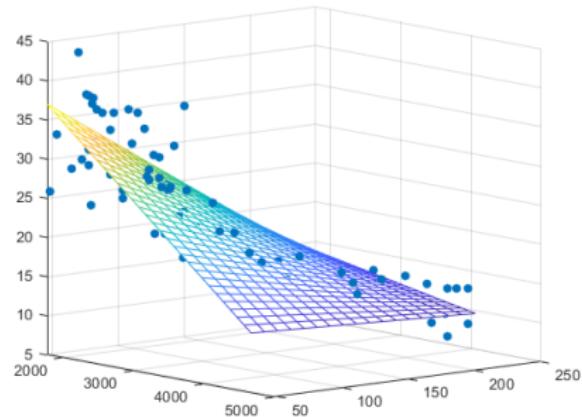


Goal

Mathematically characterize this implicit regularization

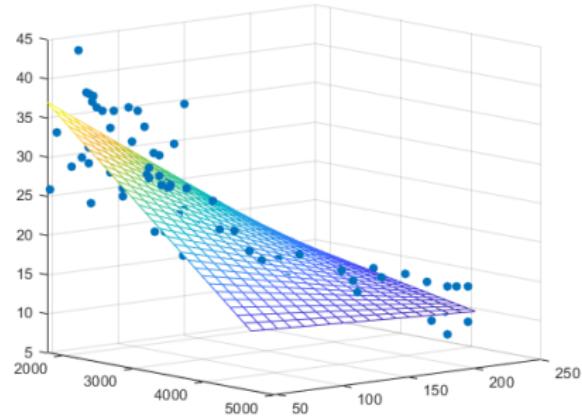
Linear Models: Implicit Norm Minimization

Linear Regression



Linear Models: Implicit Norm Minimization

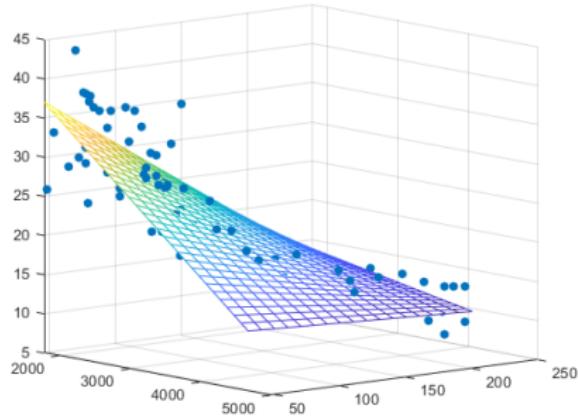
Linear Regression



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

Linear Models: Implicit Norm Minimization

Linear Regression



When # of learned weights > # of training examples:

GD initialized at 0 converges to $\min \ell_2$ norm solution

$$\operatorname{argmin}_{\mathbf{w}} \|\mathbf{w}\|_2 \text{ s.t. } \mathbf{w} \text{ is global min}$$

Implicit Norm Minimization In Deep Learning?

Widespread Hope

In deep learning, GD finds solution with **min norm** (possibly not ℓ_2)

$$\operatorname{argmin}_{\mathbf{w}} \|\mathbf{w}\| \text{ s.t. } \mathbf{w} \text{ is global min}$$

Implicit Norm Minimization In Deep Learning?

Widespread Hope

In deep learning, GD finds solution with **min norm** (possibly not ℓ_2)

$$\operatorname{argmin}_{\mathbf{w}} \|\mathbf{w}\| \text{ s.t. } \mathbf{w} \text{ is global min}$$

Demonstrated in various settings, e.g.:

- Neyshabur et al. 2015
- Gunasekar et al. 2017
- Soudry et al. 2018
- Gunasekar et al. 2018a
- Gunasekar et al. 2018b
- Li et al. 2018
- Jacot et al. 2018
- Ji & Telgarsky 2019a
- Ji & Telgarsky 2019b
- Wu et al. 2019
- Oymak & Soltanolkotabi 2019
- Nacson et al. 2019a
- Nacson et al. 2019b
- Woodworth et al. 2020
- Lyu & Li 2020
- Ali et al. 2020
- Chizat & Bach 2020
- Lyu et al. 2021

Perspective: Implicit Rank Minimization

Perspective

To understand implicit regularization in deep learning:

Perspective: Implicit Rank Minimization

Perspective

To understand implicit regularization in deep learning:

- Language of standard **norm regularizers** might not suffice

Perspective: Implicit Rank Minimization

Perspective

To understand implicit regularization in deep learning:

- Language of standard **norm regularizers** might not suffice
- **Notions of rank** may be key

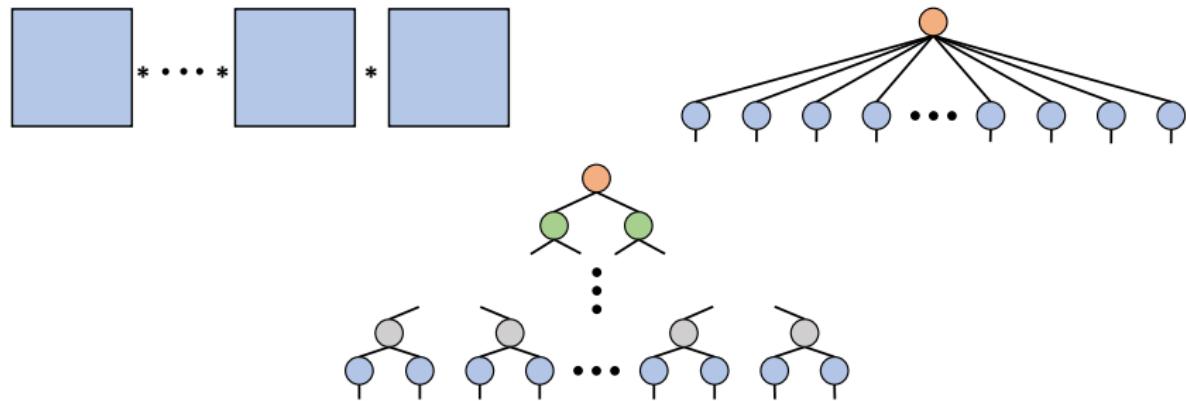
Perspective: Implicit Rank Minimization

Perspective

To understand implicit regularization in deep learning:

- Language of standard **norm regularizers** might not suffice
- **Notions of rank** may be key

Case will be made via matrix and tensor factorizations



Outline

1 Implicit Regularization in Deep Learning

2 Matrix Factorization

- Implicit Regularization \neq Norm Minimization

3 Tensor Factorization

4 Hierarchical Tensor Factorization

5 Implications for Modern Deep Learning

6 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 $\{1, \dots, d\} \times \{1, \dots, d'\}$ to \mathbb{R}

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 $\{1, \dots, d\} \times \{1, \dots, d'\}$ to \mathbb{R}

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

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 $\{1, \dots, d\} \times \{1, \dots, d'\}$ to \mathbb{R}

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

observed entries \longleftrightarrow train data

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 $\{1, \dots, d\} \times \{1, \dots, d'\}$ 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 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 $\{1, \dots, d\} \times \{1, \dots, d'\}$ 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

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

MF \longleftrightarrow Linear NN**Matrix Factorization (MF):**

Parameterize solution as **product of matrices** and fit observations via GD

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

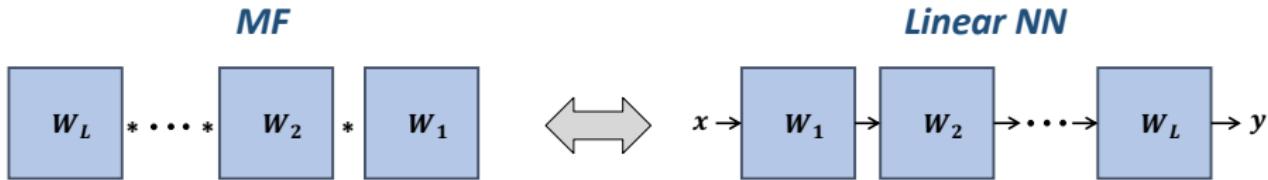
↑
hidden dimensions large enough to **not limit rank**

MF \longleftrightarrow Linear NN**Matrix Factorization (MF):**

Parameterize solution as **product of matrices** and fit observations via GD

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

↑
hidden dimensions large enough to **not limit rank**

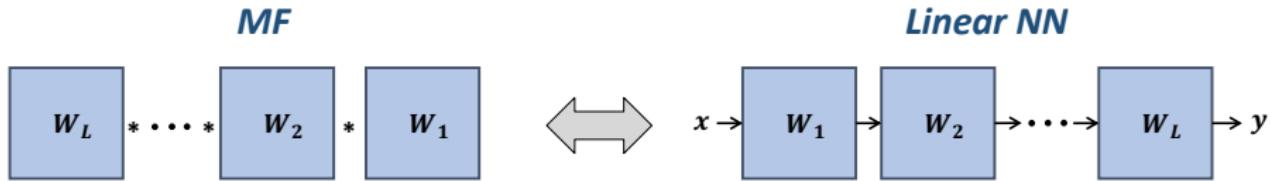


MF \longleftrightarrow Linear NN**Matrix Factorization (MF):**

Parameterize solution as **product of matrices** and fit observations via GD

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

↑
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: Nuclear Norm Minimization

Classic Result (Candes & Recht 2009)

For low **rank** ground truth:

$$\min \|W\|_{nuclear} \quad s.t. \quad [W]_{i,j} = y_{i,j} \quad \forall (i,j) \in \Omega$$

perfectly recovers under certain technical conditions

Conjecture: Nuclear Norm Minimization

Classic Result (Candes & Recht 2009)

For low rank ground truth:

$$\min \|W\|_{nuclear} \quad s.t. \quad [W]_{i,j} = y_{i,j} \quad \forall (i,j) \in \Omega$$

perfectly recovers under certain technical conditions

Conjecture (Gunasekar et al. 2017)

Training MF via gradient flow (GD with step size $\rightarrow 0$) with small init

\implies *min nuclear norm solution*

Conjecture: Nuclear Norm Minimization

Classic Result (Candes & Recht 2009)

For low rank ground truth:

$$\min \|W\|_{nuclear} \quad s.t. \quad [W]_{i,j} = y_{i,j} \quad \forall (i,j) \in \Omega$$

perfectly recovers under certain technical conditions

Conjecture (Gunasekar et al. 2017)

*Training MF via gradient flow (GD with step size $\rightarrow 0$) with small init
 \implies min nuclear norm solution*

Proven in certain restricted cases (Gunasekar et al. 2017, Li et al. 2018, Belabbas 2020)

Conjecture: No Norm is Minimized

Conjecture: No Norm is Minimized

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

Conjecture: No Norm is Minimized

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

Theorem (Arora et al. 2019)

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

Conjecture: No Norm is Minimized

$W_e := W_L \cdots W_1$ — end matrix $\{\sigma_M^{(r)}\}_r$ — 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!

Conjecture: No Norm is Minimized

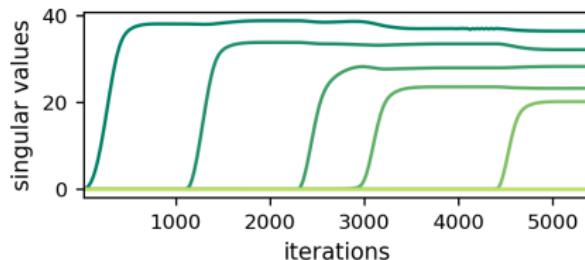
$W_e := W_L \cdots W_1$ — end matrix $\{\sigma_M^{(r)}\}_r$ — singular values of W_e

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



Conjecture: No Norm is Minimized

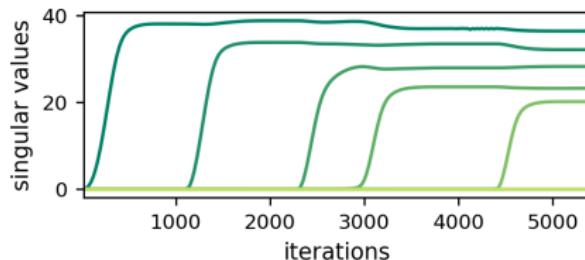
$W_e := W_L \cdots W_1$ — end matrix $\{\sigma_M^{(r)}\}_r$ — singular values of W_e

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: No Norm is Minimized

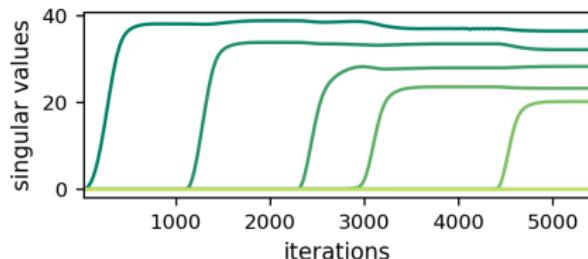
$W_e := W_L \cdots W_1$ — end matrix $\{\sigma_M^{(r)}\}_r$ — singular values of W_e

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

1 Implicit Regularization in Deep Learning

2 Matrix Factorization

- Implicit Regularization \neq Norm Minimization

3 Tensor Factorization

4 Hierarchical Tensor Factorization

5 Implications for Modern Deep Learning

6 Conclusion

Our Work: Implicit Regularization \neq Norm Minimization

Open question: does implicit regularization in MF **minimize a norm?**

Our Work: Implicit Regularization \neq Norm Minimization

Open question: does implicit regularization in MF **minimize a norm?**

Theorem (informal)

There exist matrix completion settings where MF:

Our Work: Implicit Regularization \neq Norm Minimization

Open question: does implicit regularization in MF **minimize a norm?**

Theorem (informal)

There exist matrix completion settings where MF:

- drives all **norms towards ∞**
- essentially **minimizes rank**

Our Work: Implicit Regularization \neq Norm Minimization

Open question: does implicit regularization in MF **minimize a norm?**

Theorem (informal)

There exist matrix completion settings where MF:

- drives all **norms towards ∞**
- essentially **minimizes rank**

Implicit regularization in MF \neq norm minimization

Our Work: Implicit Regularization \neq Norm Minimization

Open question: does implicit regularization in MF **minimize a norm?**

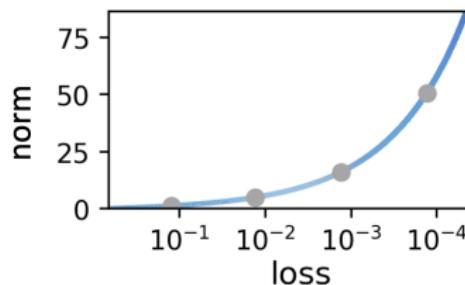
Theorem (informal)

There exist matrix completion settings where MF:

- drives all **norms towards ∞**
- essentially **minimizes rank**

Implicit regularization in MF \neq norm minimization

Experiment



Our Work: Implicit Regularization \neq Norm Minimization

Open question: does implicit regularization in MF **minimize a norm?**

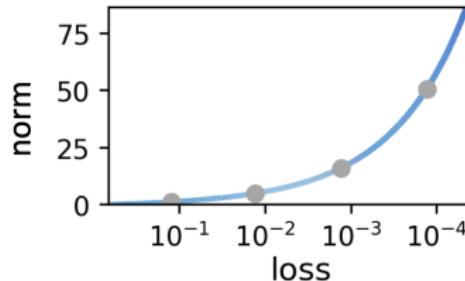
Theorem (informal)

There exist matrix completion settings where MF:

- drives all **norms towards ∞**
- essentially **minimizes rank**

Implicit regularization in MF \neq norm minimization

Experiment



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

Outline

1 Implicit Regularization in Deep Learning

2 Matrix Factorization

- Implicit Regularization \neq Norm Minimization

3 Tensor Factorization

4 Hierarchical Tensor Factorization

5 Implications for Modern Deep Learning

6 Conclusion

Drawbacks of Studying MF

$$\begin{array}{|c|c|c|c|} \hline 4 & ? & ? & 4 \\ \hline ? & 5 & 4 & ? \\ \hline ? & 5 & ? & ? \\ \hline \end{array} = W_L * \dots * W_2 * W_1$$

As a surrogate for deep learning, MF is limited:

Drawbacks of Studying MF

$$\begin{array}{|c|c|c|c|} \hline 4 & ? & ? & 4 \\ \hline ? & 5 & 4 & ? \\ \hline ? & 5 & ? & ? \\ \hline \end{array} = \boxed{W_L} * \cdots * \boxed{W_2} * \boxed{W_1}$$

As a surrogate for deep learning, MF is limited:

- (1) Misses non-linearity

Drawbacks of Studying MF

$$\begin{array}{|c|c|c|c|} \hline 4 & ? & ? & 4 \\ \hline ? & 5 & 4 & ? \\ \hline ? & 5 & ? & ? \\ \hline \end{array} = \boxed{W_L} * \cdots * \boxed{W_2} * \boxed{W_1}$$

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

$$\begin{array}{|c|c|c|c|} \hline 4 & ? & ? & 4 \\ \hline ? & 5 & 4 & ? \\ \hline ? & 5 & ? & ? \\ \hline \end{array} = \boxed{W_L} * \cdots * \boxed{W_2} * \boxed{W_1}$$

As a surrogate for deep learning, MF is limited:

- (1) Misses non-linearity
- (2) Does not capture prediction with more than 2 input variables

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

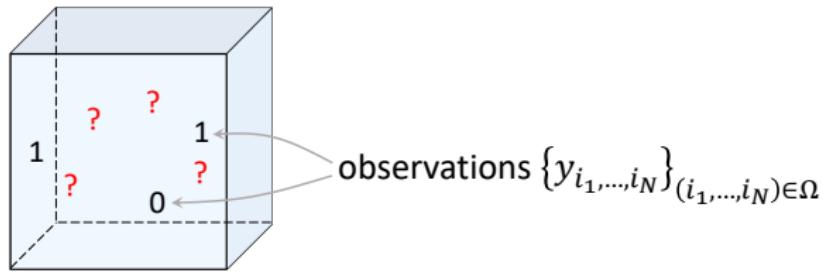
Tensor Completion \longleftrightarrow Multi-Dimensional Prediction

Tensor: N -dimensional array ($N = \text{order}$ of tensor)

Tensor Completion \longleftrightarrow Multi-Dimensional Prediction

Tensor: N -dimensional array ($N = \text{order}$ of tensor)

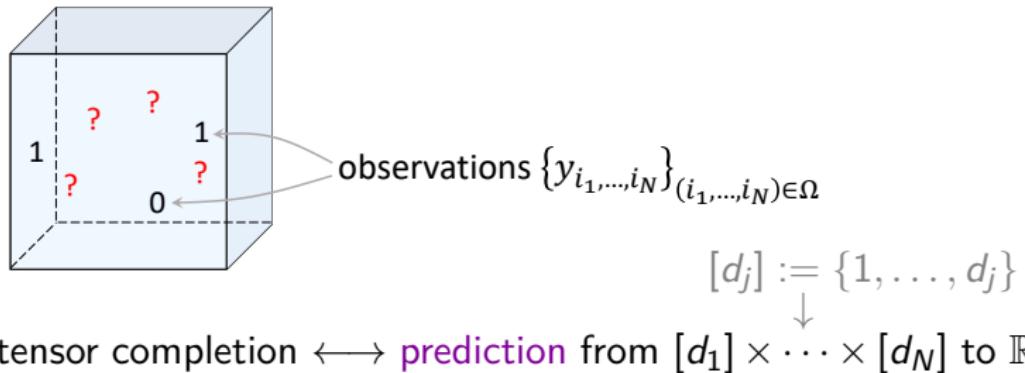
Tensor completion: recover unknown tensor given subset of entries



Tensor Completion \longleftrightarrow Multi-Dimensional Prediction

Tensor: N -dimensional array ($N = \text{order}$ of tensor)

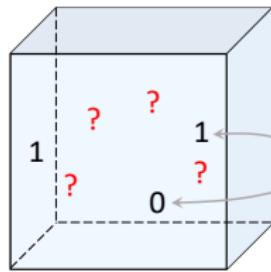
Tensor completion: recover unknown tensor given subset of entries



Tensor Completion \longleftrightarrow Multi-Dimensional Prediction

Tensor: N -dimensional array ($N = \text{order}$ of tensor)

Tensor completion: recover unknown tensor given subset of entries



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

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

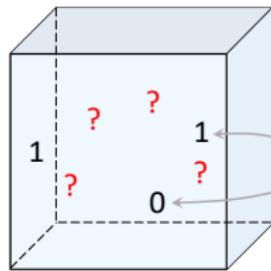
$d_1 \times \dots \times d_N$ tensor completion \longleftrightarrow prediction from $[d_1] \times \dots \times [d_N]$ to \mathbb{R}

value of entry (i_1, \dots, i_N) \longleftrightarrow label of input (i_1, \dots, i_N)

Tensor Completion \longleftrightarrow Multi-Dimensional Prediction

Tensor: N -dimensional array ($N = \text{order}$ of tensor)

Tensor completion: recover unknown tensor given subset of entries



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

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

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

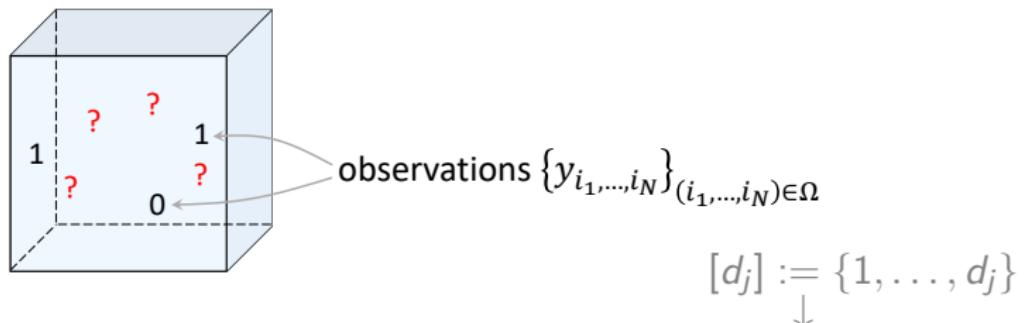
value of entry (i_1, \dots, i_N) \longleftrightarrow label of input (i_1, \dots, i_N)

observed entries \longleftrightarrow train data

Tensor Completion \longleftrightarrow Multi-Dimensional Prediction

Tensor: N -dimensional array ($N = \text{order}$ of tensor)

Tensor completion: recover unknown tensor given subset of entries



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

value of entry (i_1, \dots, i_N) \longleftrightarrow label of input (i_1, \dots, i_N)

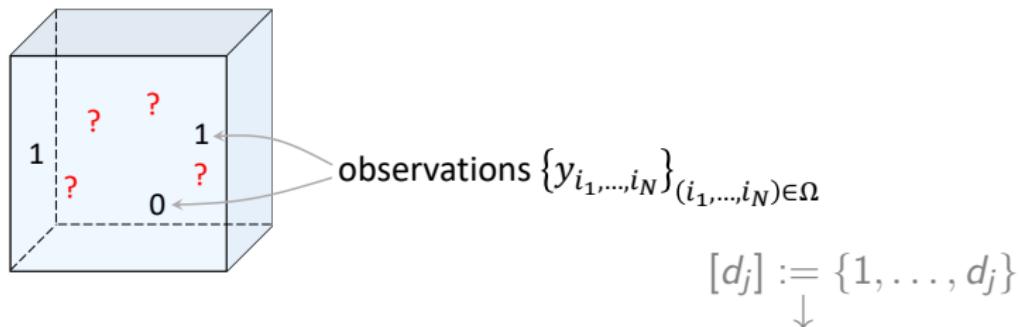
observed entries \longleftrightarrow train data

unobserved entries \longleftrightarrow test data

Tensor Completion \longleftrightarrow Multi-Dimensional Prediction

Tensor: N -dimensional array ($N = \text{order}$ of tensor)

Tensor completion: recover unknown tensor given subset of entries



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

value of entry (i_1, \dots, i_N) \longleftrightarrow label of input (i_1, \dots, 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

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

TF \longleftrightarrow Shallow Non-Linear Convolutional NN**Tensor Factorization (TF):**

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

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

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

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

\uparrow

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

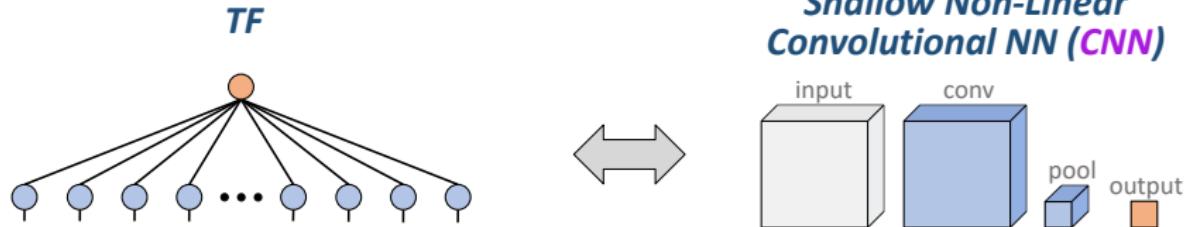
Tensor Factorization (TF):

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

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

↑
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):

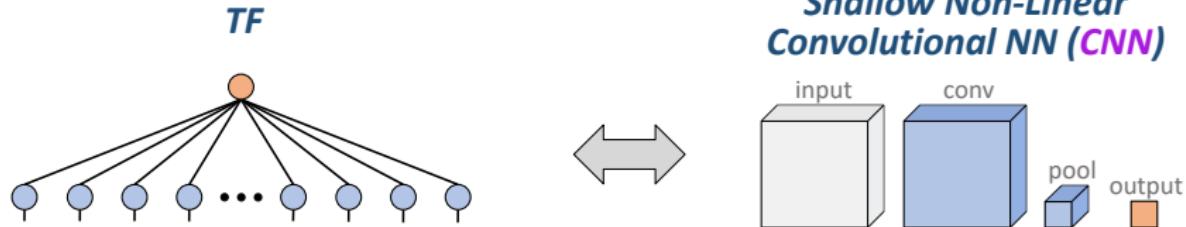
Parameterize solution as **sum of outer products** and fit observations via GD

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

\uparrow

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

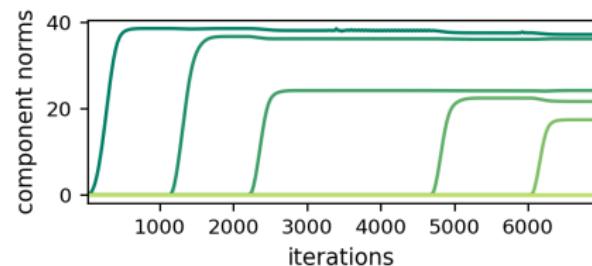
$\sigma_T^{(r)}(t) := \|\otimes_{n=1}^N \mathbf{w}_r^n(t)\|_F$ — Frobenius norm of r 'th component

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

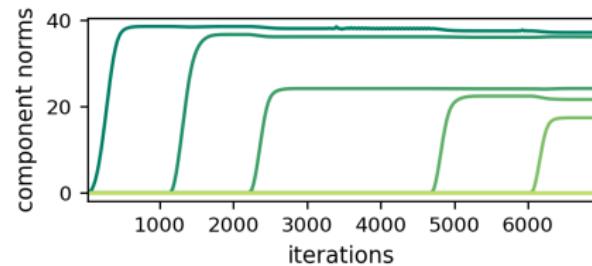
$\sigma_T^{(r)}(t) := \|\otimes_{n=1}^N \mathbf{w}_r^n(t)\|_F$ — Frobenius norm of r 'th component

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

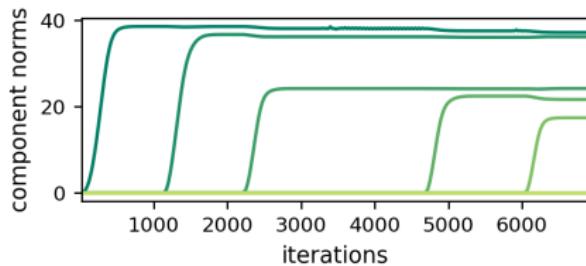
$\sigma_T^{(r)}(t) := \|\otimes_{n=1}^N \mathbf{w}_r^n(t)\|_F$ — Frobenius norm of r 'th component

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

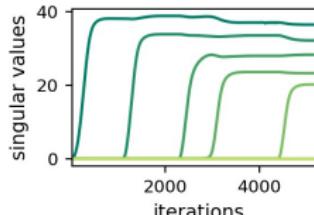
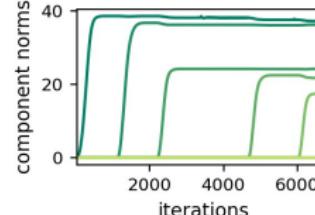


Incremental learning of components leads to low tensor rank!

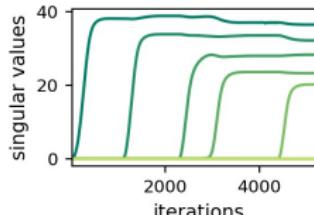
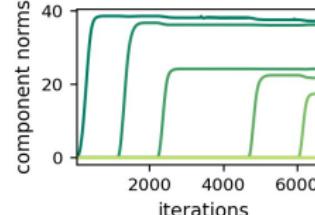
Theorem (under technical conditions)

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

Analogy Between Implicit Regularizations

	MF	TF
Quantity	singular values	component norms
Dynamics	$\frac{d}{dt} \sigma_M^{(r)}(t) \propto \sigma_M^{(r)}(t)^{2-\frac{2}{L}}$	$\frac{d}{dt} \sigma_T^{(r)}(t) \propto \sigma_T^{(r)}(t)^{2-\frac{2}{N}}$
Experiment	 <p>singular values</p> <p>iterations</p>	 <p>component norms</p> <p>iterations</p>
Incremental learning minimizes	rank	tensor rank

Analogy Between Implicit Regularizations

	MF	TF
Quantity	singular values	component norms
Dynamics	$\frac{d}{dt} \sigma_M^{(r)}(t) \propto \sigma_M^{(r)}(t)^{2-\frac{2}{L}}$	$\frac{d}{dt} \sigma_T^{(r)}(t) \propto \sigma_T^{(r)}(t)^{2-\frac{2}{N}}$
Experiment	 singular values	 component norms
Incremental learning minimizes	rank	tensor rank

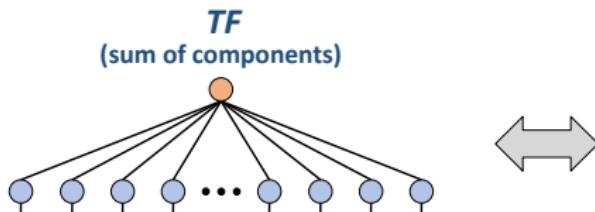
Implicit regularizations in MF and TF have identical structure!

Outline

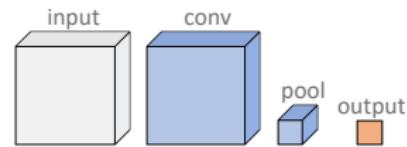
- 1 Implicit Regularization in Deep Learning
- 2 Matrix Factorization
 - Implicit Regularization \neq Norm Minimization
- 3 Tensor Factorization
- 4 Hierarchical Tensor Factorization
- 5 Implications for Modern Deep Learning
- 6 Conclusion

HTF \longleftrightarrow Deep Non-Linear CNN

TF does not account for **depth**

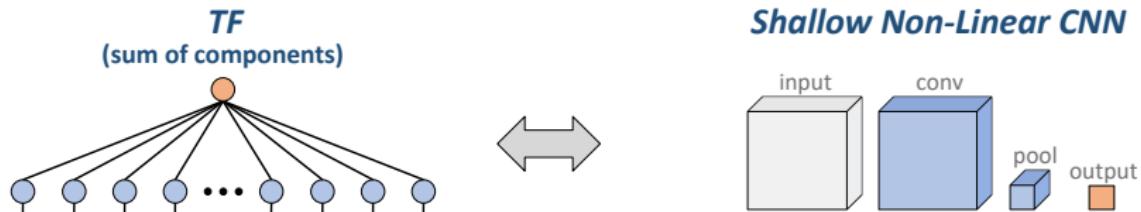


Shallow Non-Linear CNN

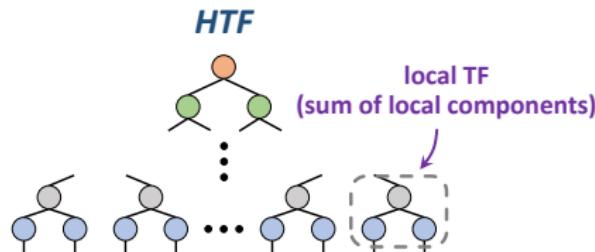


HTF \longleftrightarrow Deep Non-Linear CNN

TF does not account for **depth**

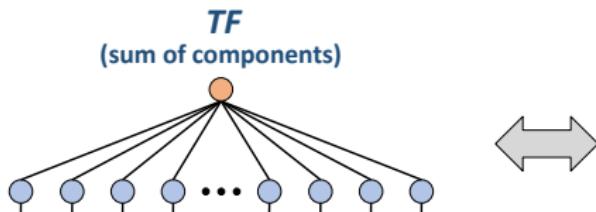


Hierarchical Tensor Factorization (HTF):

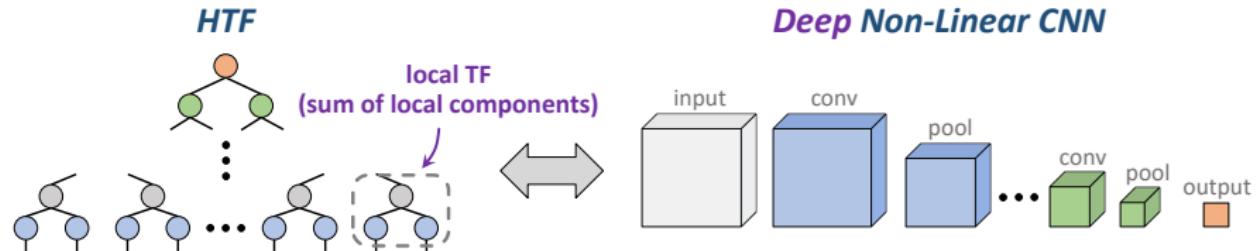


HTF \longleftrightarrow Deep Non-Linear CNN

TF does not account for **depth**

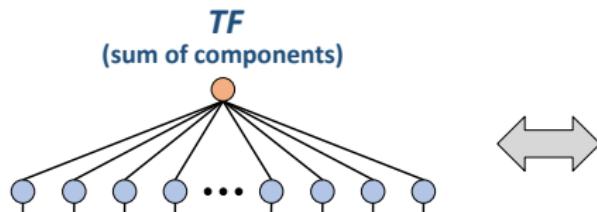


Hierarchical Tensor Factorization (HTF):

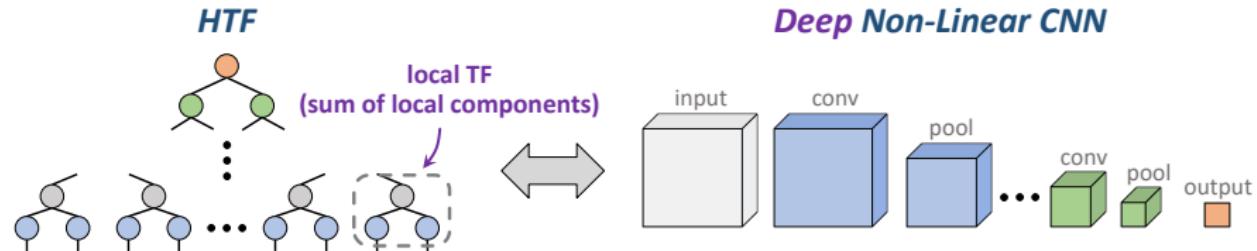


HTF \longleftrightarrow Deep Non-Linear CNN

TF does not account for **depth**



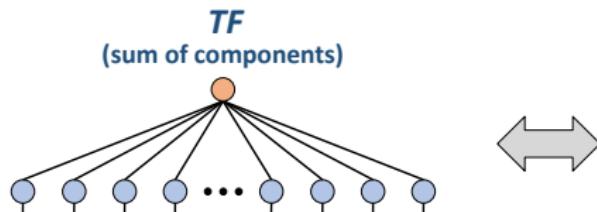
Hierarchical Tensor Factorization (HTF):



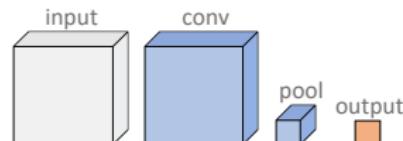
Equivalence studied extensively (e.g. Cohen et al. 2016, Levine et al. 2018, Khrulkov et al. 2018)

HTF \longleftrightarrow Deep Non-Linear CNN

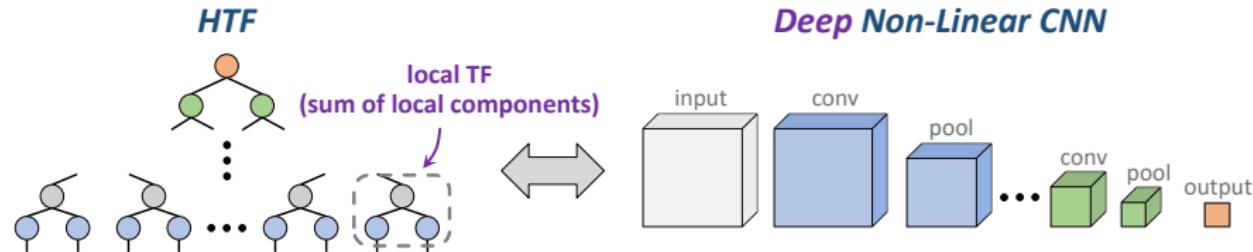
TF does not account for **depth**



Shallow Non-Linear CNN



Hierarchical Tensor Factorization (HTF):



Deep Non-Linear CNN

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

Representation w/ few local components \implies 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

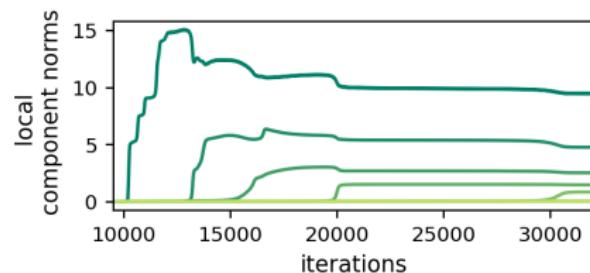
K — order of local component

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

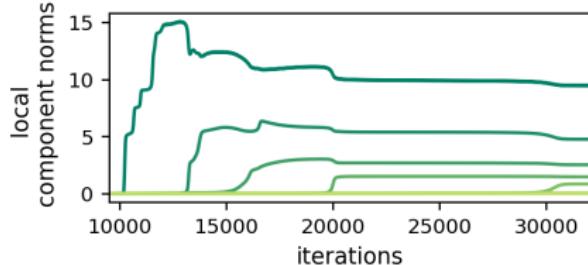
$\sigma_H^{(r)}(t)$ — Frobenius norm of r 'th local component in a location of HTF
 K — order of local component

Theorem

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

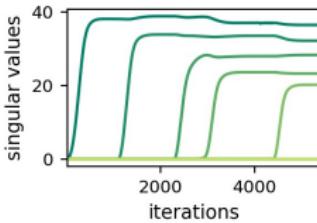
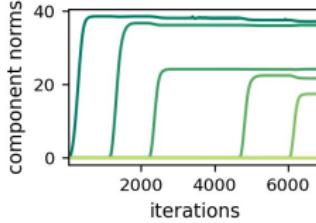
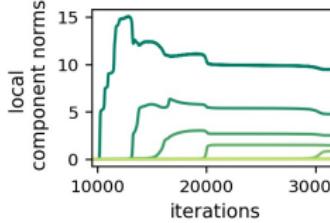
Local component norms move slower when small and faster when large!

Experiment: completion of low hierarchical tensor rank tensor via HTF

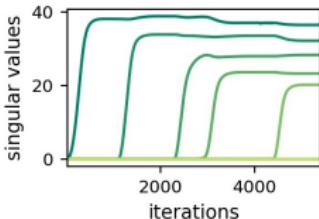
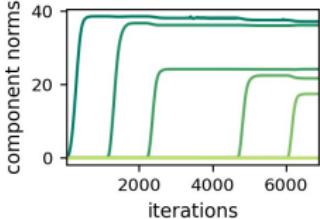
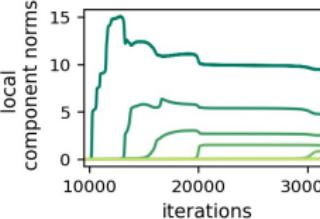


Incremental learning of local components leads to low hierarchical tensor rank!

Analogy Between Implicit Regularizations

	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	 <p>singular values</p> <p>iterations</p>	 <p>component norms</p> <p>iterations</p>	 <p>local component norms</p> <p>iterations</p>
Incremental learning minimizes	rank	tensor rank	hierarchical tensor rank

Analogy Between Implicit Regularizations

	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	 <p>singular values</p> <p>iterations</p>	 <p>component norms</p> <p>iterations</p>	 <p>local component norms</p> <p>iterations</p>
Incremental learning minimizes	rank	tensor rank	hierarchical tensor rank

Implicit regularizations in MF, TF and HTF have identical structure!

Outline

1 Implicit Regularization in Deep Learning

2 Matrix Factorization

- Implicit Regularization \neq Norm Minimization

3 Tensor Factorization

4 Hierarchical Tensor Factorization

5 Implications for Modern Deep Learning

6 Conclusion

Practical Application: Rank Minimization in NN Layers

Practical Application: Rank Minimization in NN Layers

Parameterize layers of NN as MF / TF / HTF

Practical Application: Rank Minimization in NN Layers

Parameterize layers of NN as MF / TF / HTF

⇒ implicit rank minimization induces compressibility and generalization

Practical Application: Rank Minimization in NN Layers

Parameterize layers of NN as MF / TF / HTF

⇒ implicit rank minimization 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
Brian Cheung
MIT CSAIL & BCS

Hossein Mobahi
Google Research
Pulkit Agrawal
MIT CSAIL

Richard Zhang
Adobe Research
Phillip Isola
MIT CSAIL

Understanding Generalization in Deep Learning via Tensor Methods

Jingling Li^{1,3} **Yanchao Sun¹** **Jiahao Su⁴** **Taiji Suzuki^{2,3}** **Furong Huang¹**

¹Department of Computer Science, University of Maryland, College Park

²Graduate School of Information Science and Technology, The University of Tokyo

³Center for Advanced Intelligence Project, RIKEN

⁴Department 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 minimized by GD over NNs

Potential Explanation for Generalization on Natural Data

Challenge

Find complexity measures that:

- Are implicitly minimized 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 minimized 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 minimized 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 minimized 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**



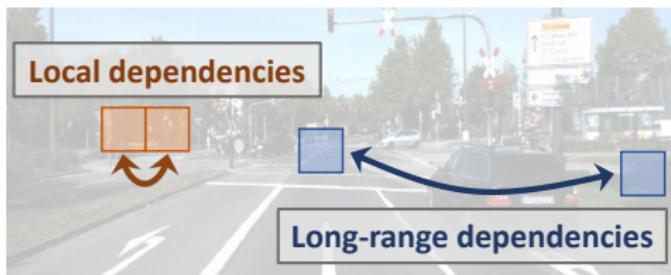
**Implicit minimization of ranks may
explain generalization on natural data!**

Counteracting Locality of CNNs via Regularization

Counteracting Locality of CNNs via Regularization

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

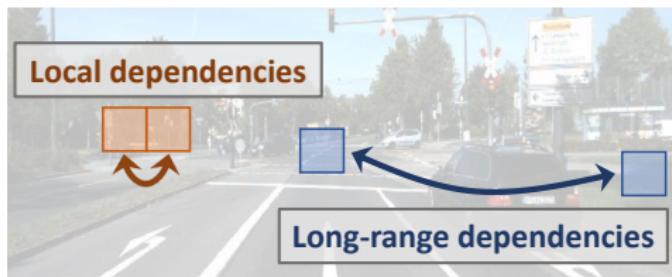
Hierarchical tensor rank measures long-range dependencies



Countering Locality of CNNs via Regularization

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

Hierarchical tensor rank measures long-range dependencies



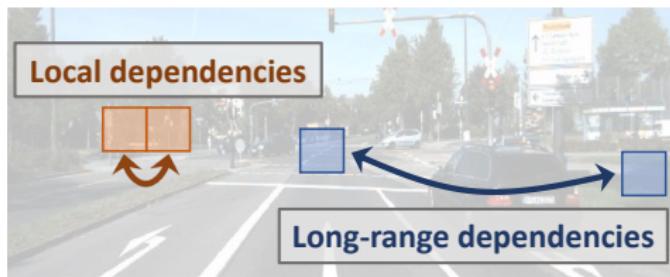
Implicit minimization of
hierarchical tensor rank in HTF

↔
Implicit minimization of
long-range dependencies in CNNs!

Counteracting Locality of CNNs via Regularization

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

Hierarchical tensor rank measures long-range dependencies



Implicit minimization of
hierarchical tensor rank in HTF

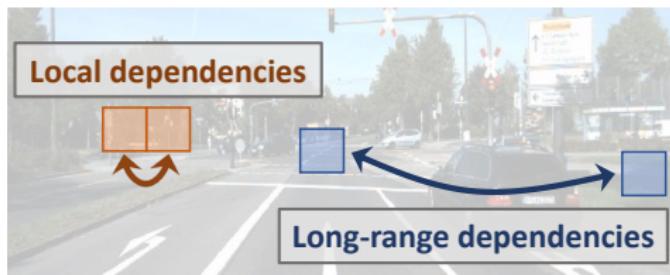
↔
Implicit minimization of
long-range dependencies in CNNs!

CNNs are not suitable for long-range tasks

Counteracting Locality of CNNs via Regularization

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

Hierarchical tensor rank measures long-range dependencies



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

CNNs are not suitable for long-range tasks

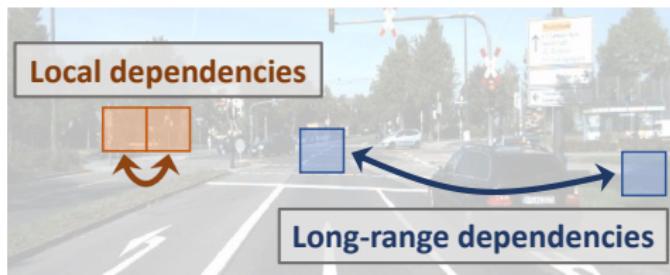
- Conventional wisdom: due to expressiveness

(Cohen & Shashua 2017, Linsley et al. 2018, Kim et al. 2020)

Counteracting Locality of CNNs via Regularization

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

Hierarchical tensor rank measures long-range dependencies



Implicit minimization of
hierarchical tensor rank in HTF

↑
Implicit minimization of
long-range dependencies in CNNs!

CNNs are not suitable for long-range tasks

- Conventional wisdom: due to expressiveness

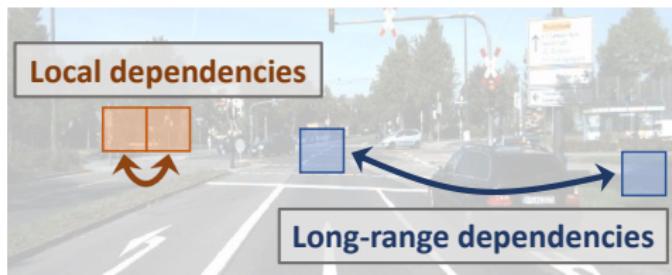
(Cohen & Shashua 2017, Linsley et al. 2018, Kim et al. 2020)

- Our analysis: implicit regularization is also a cause

Counteracting Locality of CNNs via Regularization

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

Hierarchical tensor rank measures long-range dependencies



Implicit minimization of
hierarchical tensor rank in HTF

Implicit minimization of
long-range dependencies in CNNs!

CNNs are not suitable for long-range tasks

- Conventional wisdom: due to expressiveness

(Cohen & Shashua 2017, Linsley et al. 2018, Kim et al. 2020)

- Our analysis: implicit regularization is also a cause

Can explicit regularization improve CNNs on long-range tasks?

Countering Locality of CNNs via Regularization

Experiment

Tasks: “Is Same Class” and Pathfinder ([Linsley et al. 2018](#), [Tay et al. 2021](#))

Countering Locality of CNNs via Regularization

Experiment

Tasks: “Is Same Class” and Pathfinder ([Linsley et al. 2018](#), [Tay et al. 2021](#))

distance: 0 ✓



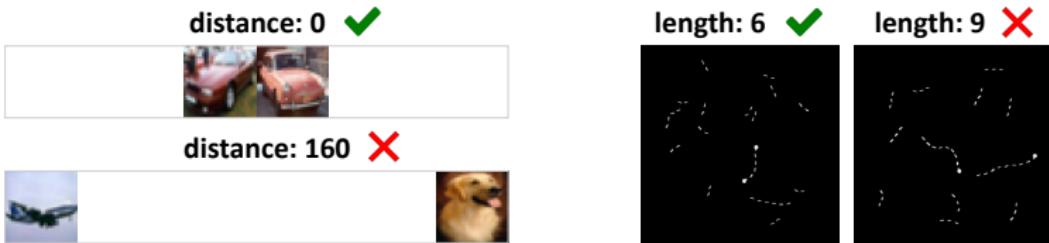
distance: 160 ✗



Countering Locality of CNNs via Regularization

Experiment

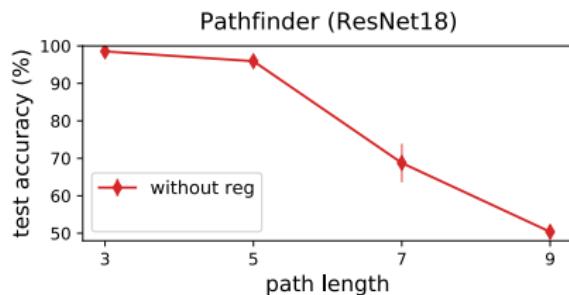
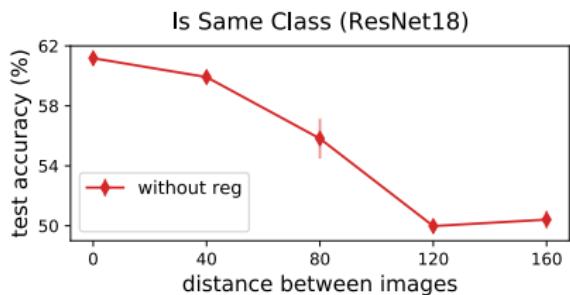
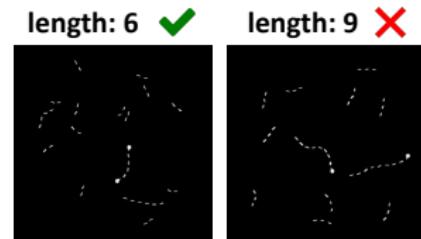
Tasks: “Is Same Class” and Pathfinder ([Linsley et al. 2018](#), [Tay et al. 2021](#))



Countering Locality of CNNs via Regularization

Experiment

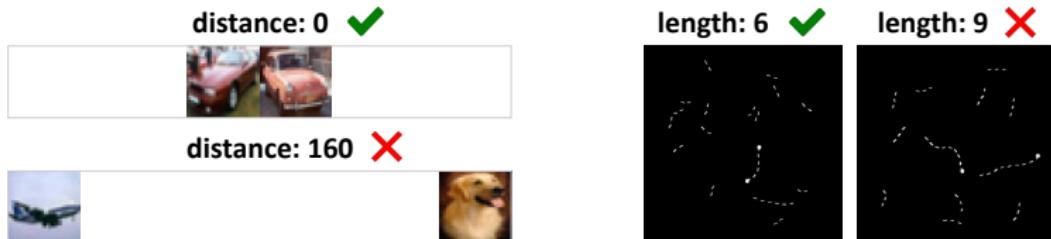
Tasks: “Is Same Class” and Pathfinder ([Linsley et al. 2018](#), [Tay et al. 2021](#))



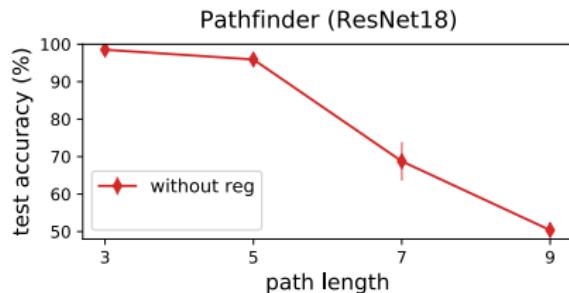
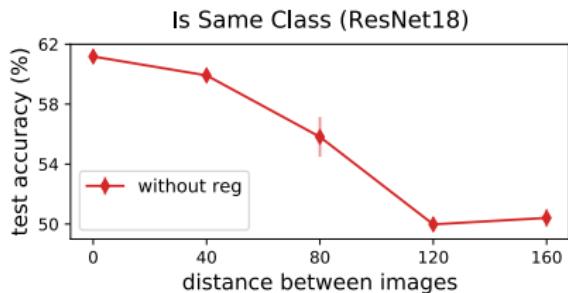
Countering Locality of CNNs via Regularization

Experiment

Tasks: “Is Same Class” and Pathfinder ([Linsley et al. 2018](#), [Tay et al. 2021](#))



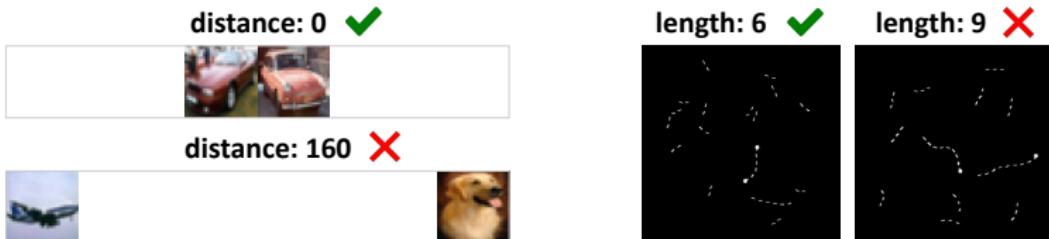
Regularization: promotes **high hierarchical tensor rank**



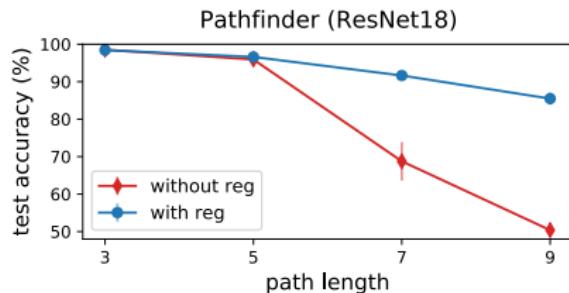
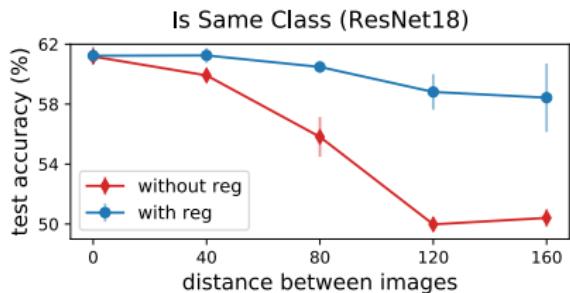
Countering Locality of CNNs via Regularization

Experiment

Tasks: “Is Same Class” and Pathfinder ([Linsley et al. 2018](#), [Tay et al. 2021](#))



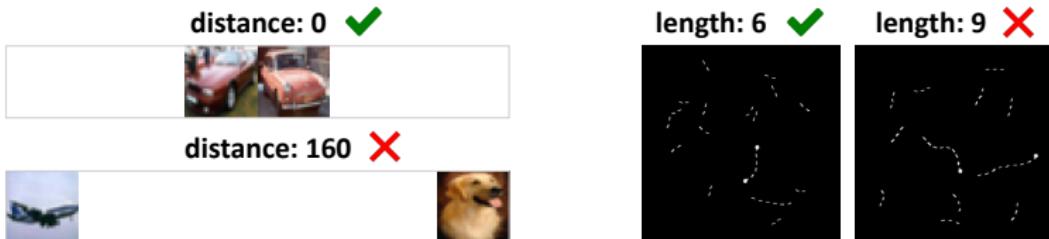
Regularization: promotes **high hierarchical tensor rank**



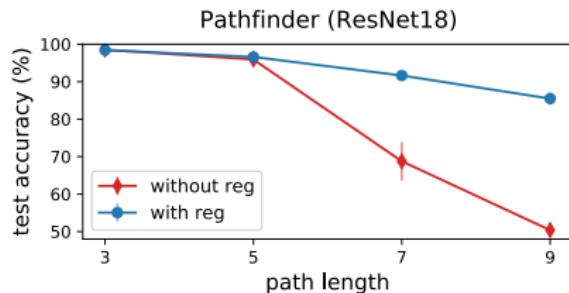
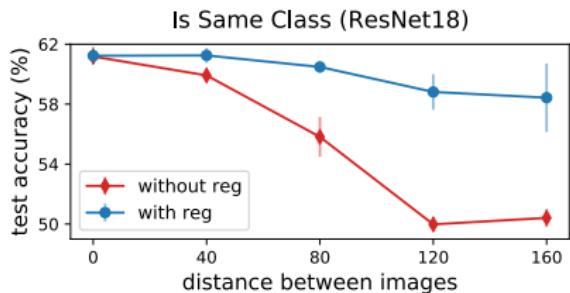
Countering Locality of CNNs via Regularization

Experiment

Tasks: “Is Same Class” and Pathfinder (Linsley et al. 2018, Tay et al. 2021)



Regularization: promotes **high hierarchical tensor rank**



Explicit regularization can improve CNNs on long-range tasks!

Outline

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

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:* GD 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:* GD can drive **all norms to ∞** while **minimizing rank**

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

- *We showed:* GD **minimizes tensor and hierarchical tensor ranks**

Recap

Goal: understand implicit regularization in deep learning

Matrix Factorization (Linear NN):

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

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

- *We showed:* GD **minimizes tensor and hierarchical tensor 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:* GD can drive all norms to ∞ while minimizing rank

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

- *We showed:* GD minimizes tensor and hierarchical tensor ranks

Implications to Modern Deep Learning:

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

Recap

Goal: understand implicit regularization in deep learning

Matrix Factorization (Linear NN):

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

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

- *We showed:* GD minimizes tensor and hierarchical tensor ranks

Implications to Modern Deep Learning:

- Parameterizing layers of NN as MF / TF/ HTF \implies compression
- Rank minimization 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:* GD can drive all norms to ∞ while minimizing rank

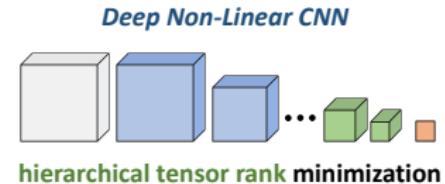
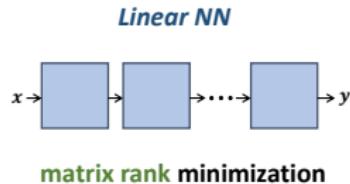
Tensor and Hierarchical Tensor Factorizations (Non-Linear CNNs):

- *We showed:* GD minimizes tensor and hierarchical tensor ranks

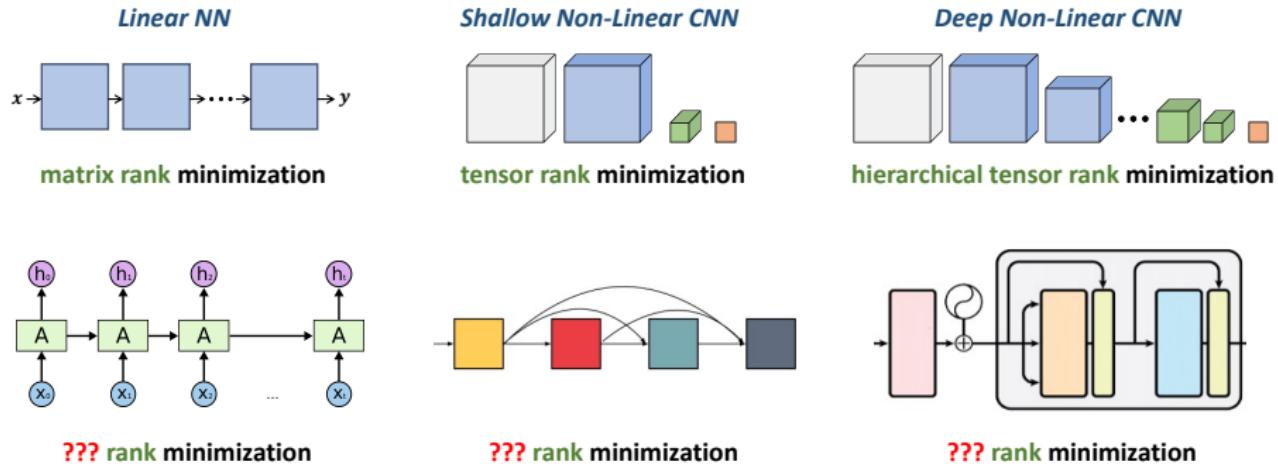
Implications to Modern Deep Learning:

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

Implicit Rank Minimization in Deep Learning

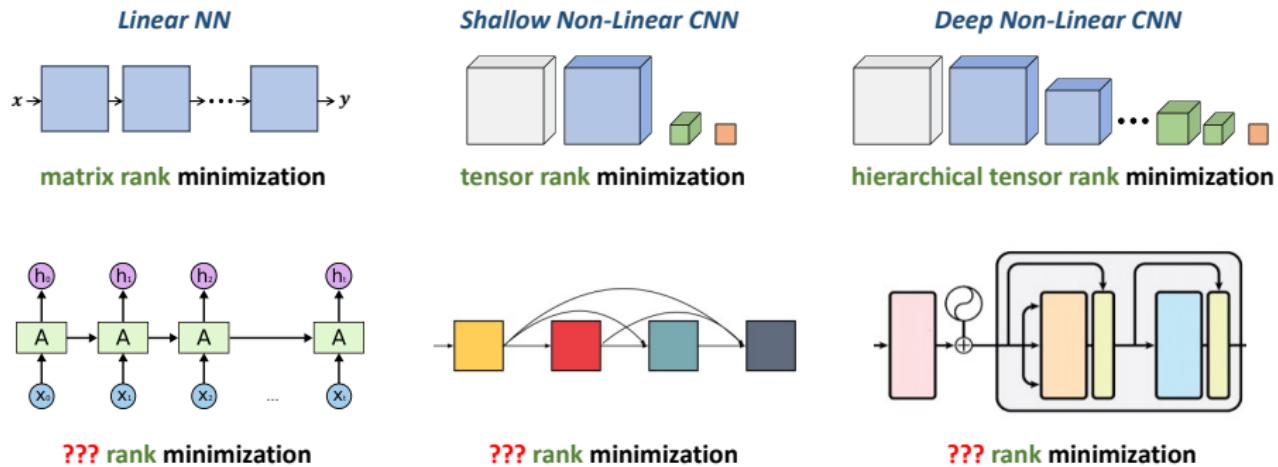


Implicit Rank Minimization in Deep Learning



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

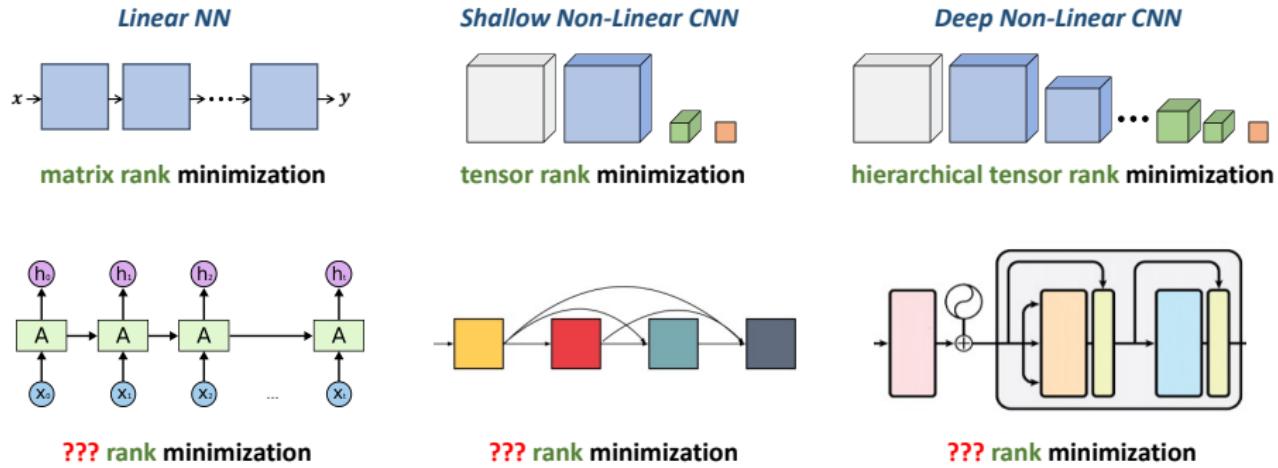
Implicit Rank Minimization in Deep Learning



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

Discovering minimized **notions of rank** may pave way to:

Implicit Rank Minimization in Deep Learning

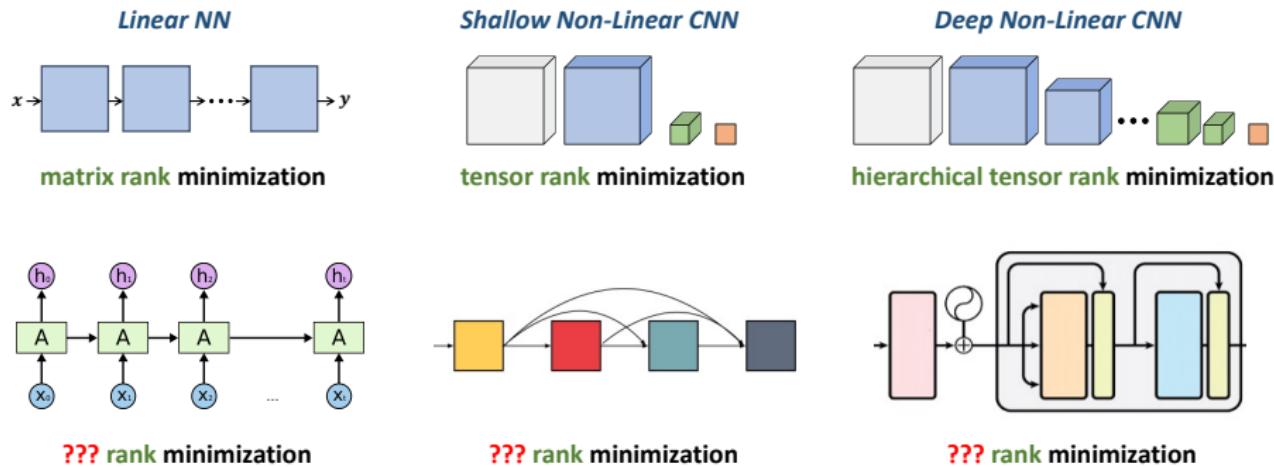


Hypothesis: in each NN architecture implicit regularization minimizes corresponding notion of rank

Discovering minimized **notions of rank** may pave way to:

- Explaining generalization

Implicit Rank Minimization in Deep Learning



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

Discovering minimized 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.