Implicit Regularization in Hierarchical Tensor Factorization and Deep Convolutional Neural Networks

Noam Razin Asaf Maman Nadav Cohen

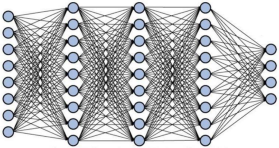
Tel Aviv University

International Conference on Machine Learning (ICML) 2022

Implicit Regularization in Deep Learning

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

# of learned weights

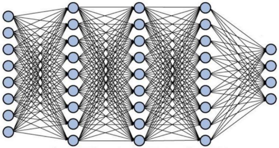


$\gg$

# of training examples

Implicit Regularization in Deep Learning

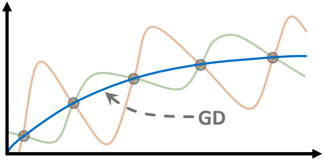
Neural networks (NNs) generalize with no explicit regularization despite:

# of learned weights  $\gg$ 

# of training examples

Conventional Wisdom

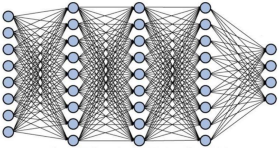
Gradient descent (GD) induces implicit regularization towards "simplicity"



Implicit Regularization in Deep Learning

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

# of learned weights

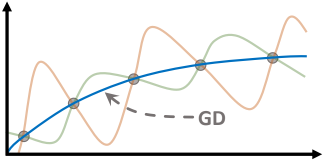


$\gg$

# of training examples

Conventional Wisdom

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



Goal

Mathematically characterize this implicit regularization

Background: Matrix Factorization

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

Background: Matrix Factorization

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

Matrix Factorization (MF)

Parameterize solution as product of matrices and minimize $\mathcal{L}$ via GD

$$\min_{W_1,\ldots,W_L} \mathcal{L}\left(W_L W_{L-1} \cdots W_1\right)$$

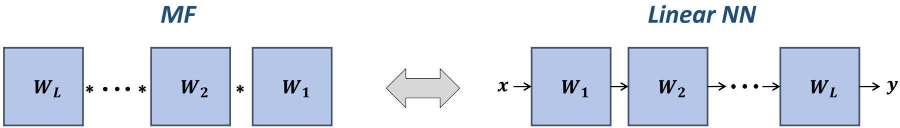
Background: Matrix Factorization

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

Matrix Factorization (MF)

Parameterize solution as product of matrices and minimize $\mathcal{L}$ via GD

$$\min_{W_1,\ldots,W_L} \ \mathcal{L}(W_L W_{L-1} \cdots W_1)$$



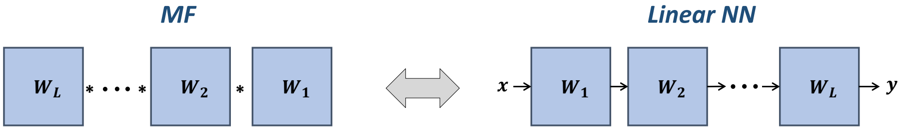
Background: Matrix Factorization

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

Matrix Factorization (MF)

Parameterize solution as product of matrices and minimize $\mathcal{L}$ via GD

$$\min_{W_1,\ldots,W_L} \ \mathcal{L}(W_L W_{L-1} \cdots W_1)$$



Prior Work: Dynamical Analysis (e.g. Arora et al. 2019, Li et al. 2021)

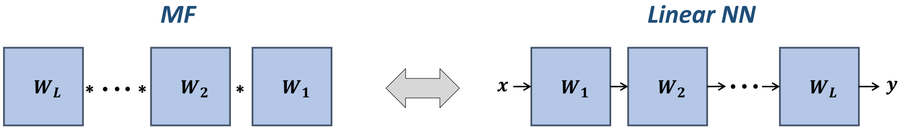
Incremental learning leads to low matrix rank

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

Matrix Factorization (MF)

Parameterize solution as product of matrices and minimize $\mathcal{L}$ via GD

$$\min_{W_1,\ldots,W_L} \ \mathcal{L}(W_L W_{L-1} \cdots W_1)$$



Prior Work: Dynamical Analysis (e.g. Arora et al. 2019, Li et al. 2021)

Incremental learning leads to low matrix rank

Limitation (as surrogate for deep learning): lacks non-linearity

Background: Tensor Factorization

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

Background: Tensor Factorization

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

Tensor Factorization (TF)

Parameterize solution as sum of outer products and minimize $\mathcal{L}$ via GD

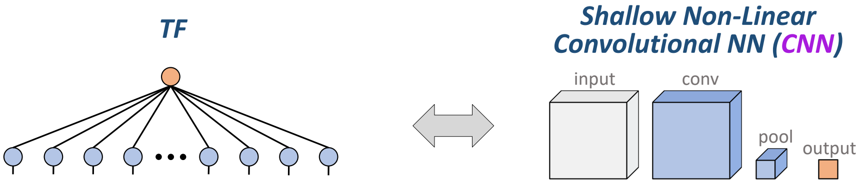
$$\min_{\{\mathbf{w}_r^n\}_{r,n}} \mathcal{L}\left(\sum_{r=1}^{R} \mathbf{w}_r^1 \otimes \cdots \otimes \mathbf{w}_r^N\right)$$

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

Tensor Factorization (TF)

Parameterize solution as sum of outer products and minimize $\mathcal{L}$ via GD

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



Background: Tensor Factorization

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

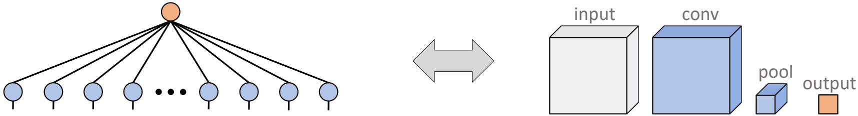
Tensor Factorization (TF)

Parameterize solution as sum of outer products and minimize $\mathcal{L}$ via GD

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

TF

Shallow Non-Linear Convolutional NN (CNN)



Prior Work: Dynamical Analysis (Razin & Cohen 2020, Razin et al. 2021)

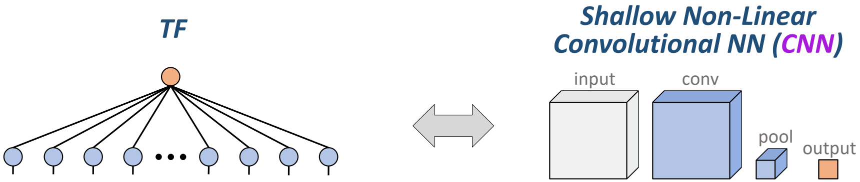
Incremental learning leads to low tensor rank

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

Tensor Factorization (TF)

Parameterize solution as sum of outer products and minimize $\mathcal{L}$ via GD

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



Prior Work: Dynamical Analysis (Razin & Cohen 2020, Razin et al. 2021)

Incremental learning leads to low tensor rank

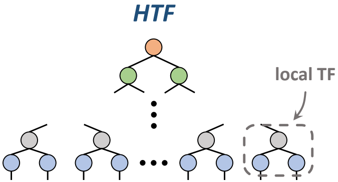
Limitation (as surrogate for deep learning): lacks depth

Hierarchical Tensor Factorization (HTF)

Accounts for both non-linearity and depth

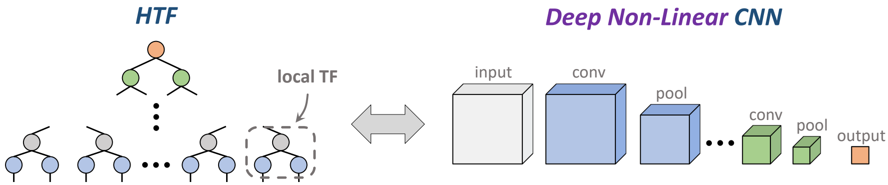
Hierarchical Tensor Factorization (HTF)

Accounts for both non-linearity and depth



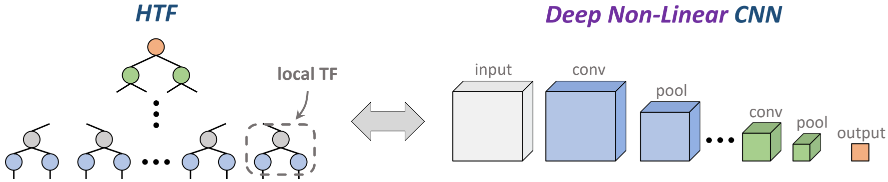
Hierarchical Tensor Factorization (HTF)

Accounts for both non-linearity and depth



Hierarchical Tensor Factorization (HTF)

Accounts for both non-linearity and depth

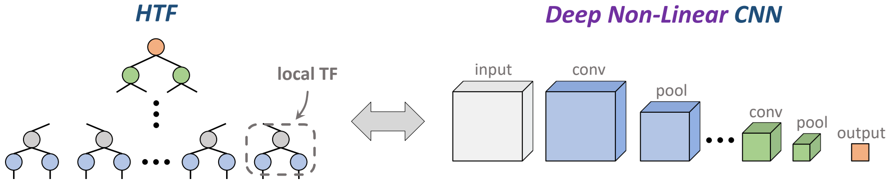


Equivalence studied extensively (e.g. Cohen et al. 2016, Levine et al. 2018, Khrulkov et al. 2018)

Implicit Regularization in Hierarchical Tensor Factorization

Hierarchical Tensor Factorization (HTF)

Accounts for both non-linearity and depth



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

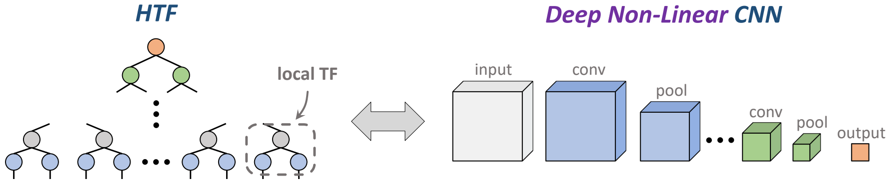
Our Work: Dynamical Analysis

Incremental learning leads to low hierarchical tensor rank

Implicit Regularization in Hierarchical Tensor Factorization

Hierarchical Tensor Factorization (HTF)

Accounts for both non-linearity and depth



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

Our Work: Dynamical Analysis

Incremental learning leads to low hierarchical tensor rank

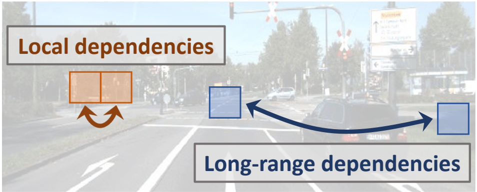
Implicit regularization in HTF is structurally identical to that in MF & TF!

Countering Locality of CNNs via Regularization

Countering Locality of CNNs via Regularization

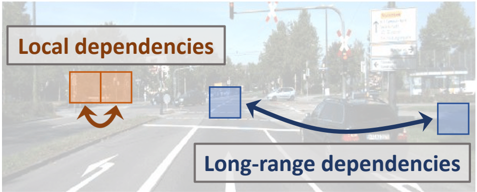
Fact (Cohen & Shashua 2017, Levine et al. 2018)

Hierarchical tensor rank measures long-range dependencies



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

Hierarchical tensor rank measures long-range dependencies

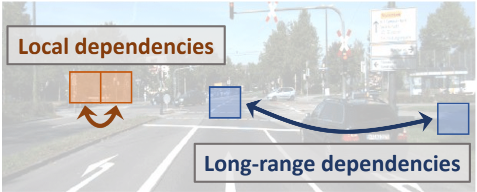


Implicit lowering of
hierarchical tensor rank in HTF
$\Updownarrow$
Implicit lowering of
long-range dependencies in CNNs!

Countering Locality of CNNs via Regularization

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

Hierarchical tensor rank measures long-range dependencies



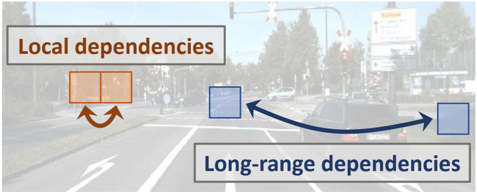
Implicit lowering of
hierarchical tensor rank in HTF
$\Updownarrow$
Implicit lowering of
long-range dependencies in CNNs!

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?

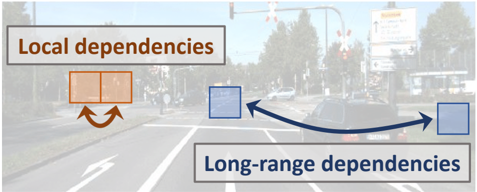
Our Work: Experiments

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?

Our Work: Experiments

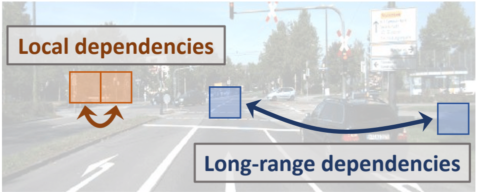
Regularization promoting high hierarchical tensor rank

$\Longrightarrow$ 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?

Our Work: 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!

Conclusion

Takeaways

Conclusion

Takeaways

- Implicit regularization in HTF lowers hierarchical tensor rank

Conclusion

Takeaways

- Implicit regularization in HTF lowers hierarchical tensor rank

  (just as in MF & TF it lowers notions of rank)

Conclusion

Takeaways

- Implicit regularization in HTF lowers hierarchical tensor rank

  (just as in MF & TF it lowers notions of rank)

- This implies implicit regularization towards locality in CNNs

Conclusion

Takeaways

- Implicit regularization in HTF lowers hierarchical tensor rank

  (just as in MF & TF it lowers notions of rank)

- This implies implicit regularization towards locality in CNNs

- Specialized explicit regularization can counter locality of CNNs!

Conclusion

Takeaways

- Implicit regularization in HTF lowers hierarchical tensor rank

  (just as in MF & TF it lowers notions of rank)

- This implies implicit regularization towards locality in CNNs

- Specialized explicit regularization can counter locality of CNNs!

Poster: #1409