What Does It Mean to Be a Transformer?

Insights from a Theoretical Hessian Analysis

ICLR 2025

Joint Work

Weronika Ormaniec

Felix Dangel

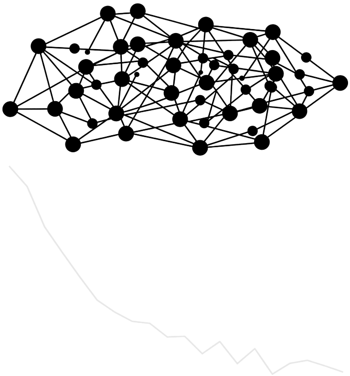
Sidak Pal Singh

ETH zürich

VECTOR INSTITUTE

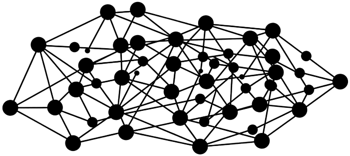
Why Study the Transformer Hessian?

Why Study the Transformer Hessian?

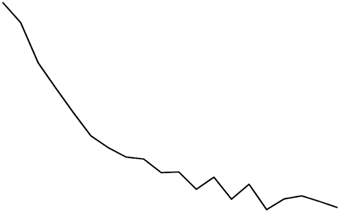


Unique Architecture

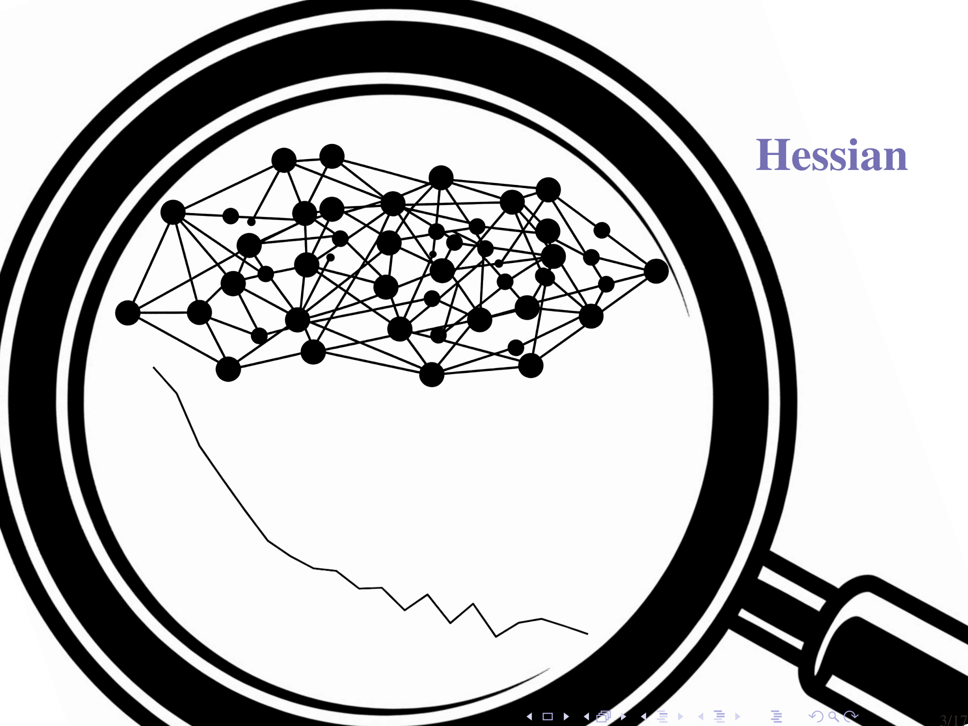
Why Study the Transformer Hessian?



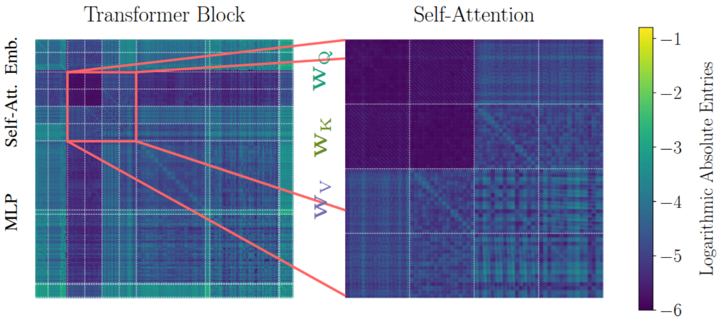
Unique Architecture



Tricky Optimization



Hessian



On the left, the Hessian of a minimal Transformer, and, on the right the zoomed-in block w.r.t. query, key and value parameters.

We Want to Understand What Makes The Transformer Hessian Special

We Want to Understand What Makes The Transformer Hessian Special

Exact Hessian Structure

We Want to Understand What Makes The Transformer Hessian Special

Exact Hessian Structure

We Want to Understand What Makes The Transformer Hessian Special

Exact Hessian Structure

Impact of Design Components

We Want to Understand What Makes The Transformer Hessian Special

Exact Hessian Structure

Impact of Design Components

Contrast with Other Models

Setup

Gauss-Newton & Block Decomposition

$$\mathbf{H} = \mathbf{H}_{\mathrm{o}} + \mathbf{H}_{\mathrm{f}}$$

We split the Hessian into two terms, then we analyze their blocks.

Dependence on Data

Data Dependence Varies Across Hessian Blocks

Data Dependence Varies Across Hessian Blocks

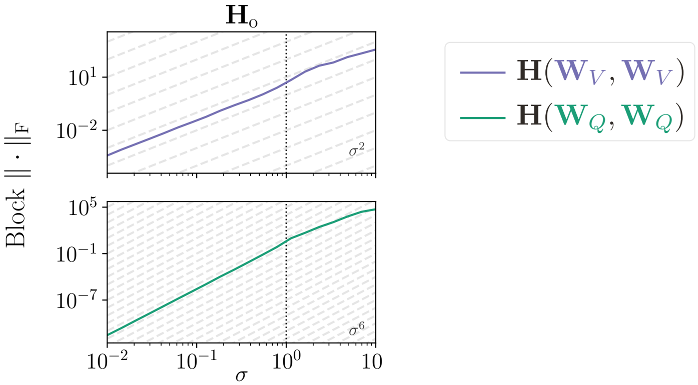
$$\mathbf{H}_{\mathrm{o}} \in \begin{array}{c c} & \begin{array}{ccc} \mathrm{Q} & \mathrm{K} & \mathrm{V} \end{array} \\ \begin{array}{c} \mathrm{Q} \\ \mathrm{K} \\ \mathrm{V} \end{array} & \begin{bmatrix} \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^4) \\ \cdot & \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^4) \\ \cdot & \cdot & \mathcal{O}(\mathbf{X}^2) \end{bmatrix} \end{array}$$

Data Dependence Varies Across Hessian Blocks

$$\mathbf{H}_{\mathrm{o}} \in \begin{array}{c} \\ \mathrm{Q} \\ \mathrm{K} \\ \mathrm{V} \end{array} \begin{array}{c} \begin{array}{ccc} \mathrm{Q} & \mathrm{K} & \mathrm{V} \end{array} \\ \begin{bmatrix} \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^4) \\ \cdot & \mathcal{O}(\mathbf{X}^6) & \mathcal{O}(\mathbf{X}^4) \\ \cdot & \cdot & \mathcal{O}(\mathbf{X}^2) \end{bmatrix} \end{array}$$

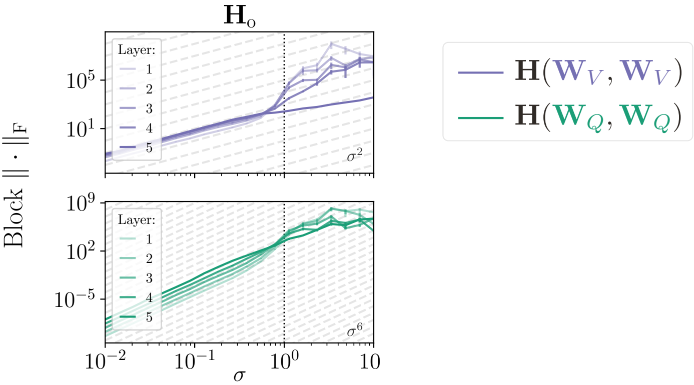
Query and key blocks are more data-dependent.

Data Dependence Varies Across Hessian Blocks



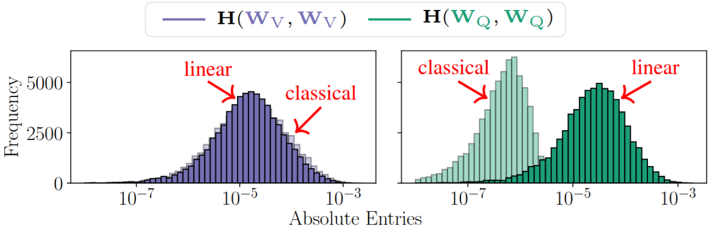
Growth rates of block Frobenius norms w.r.t. the magnitude σ of $\mathbf{X}$ confirm our theoretical predictions.

Data Dependence Varies Across Hessian Blocks



Theoretical growth rates of block Frobenius norms w.r.t. the magnitude σ of $\mathbf{X}$ hold also for deeper networks and $\sigma < 1$.

Data Dependence Varies Across Blocks Because of Softmax



Softmax results in heterogeneity in magnitudes of Hessian block entries.

Self-Attention vs MLP Hessian

Model Family	Transformer
$\mathbf{H}_{\mathbf{O}}^{\text{lin}}$	$\mathcal{O}(\mathbf{\Sigma}_{\mathbf{xx}}^3)$

Dependence of the Hessian of linear layers on the intra-sequence covariance matrix $\mathbf{\Sigma}_{\mathbf{xx}} = \frac{1}{L}\mathbf{X}^\top\mathbf{X}$ in a big $\mathcal{O}$ notation.

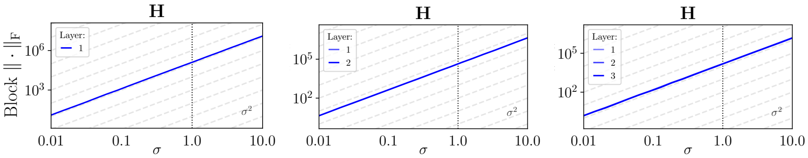
Self-Attention vs MLP Hessian

Model Family	Transformer	MLP/CNN
$\mathbf{H}_{\mathbf{O}}^{\text{lin}}$	$\mathcal{O}(\mathbf{\Sigma}_{\mathbf{xx}}^3)$	$\mathcal{O}(\mathbf{\Sigma}_{\mathbf{xx}})$

Dependence of the Hessian of linear layers on the intra-sequence covariance matrix $\mathbf{\Sigma}_{\mathbf{xx}} = \frac{1}{L}\mathbf{X}^\top\mathbf{X}$ in a big $\mathcal{O}$ notation.[a]

[a] Source of MLP Hessian formulas: Singh et al. "Analytic Insights into Structure and Rank of Neural Network Hessian Maps." In NeurIPS (2021).

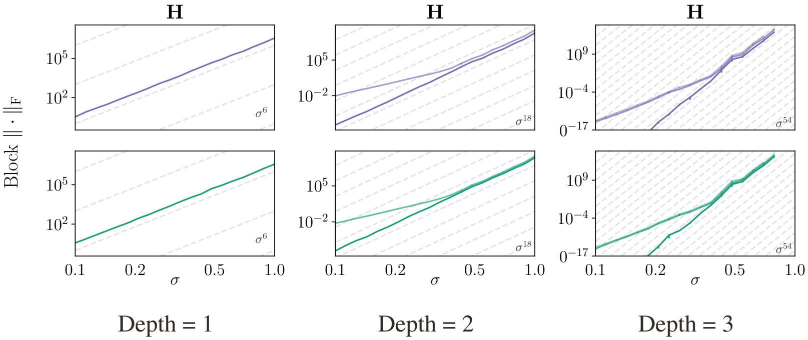
Multilayer Linear MLP Hessian Growth Rates



Depth = 1 Depth = 2 Depth = 3

Diagonal blocks of a linear MLP grow the same with σ irrespective of network depth.

Multilayer Linear Transformer Hessian Growth Rates



Diagonal blocks of a linear Transformer grow super-exponentially with depth.

Conclusion

Conclusion

Summary:

- Exact Hessian of the self-attention layer
- Block-heterogeneity in terms of data dependence
- Influence of softmax on the Hessian
- Differences compared to MLPs/CNNs

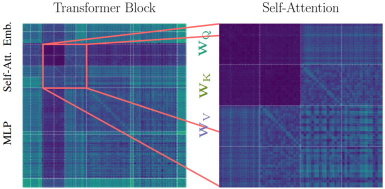
Conclusion

Summary:

- Exact Hessian of the self-attention layer
- Block-heterogeneity in terms of data dependence
- Influence of softmax on the Hessian
- Differences compared to MLPs/CNNs

In the paper you will also find the discussion of:

- Block-heterogeneity in terms of weights and attention moments
- Influence of the query-key parametrization of the self-attention
- Influence of multi-head self-attention



Thank you!

What Does It Mean to Be a Transformer?
Insights from a Theoretical Hessian Analysis

Weronika Ormaniec
Felix Dangel
Sidak Pal Singh

Correspondence: wormaniec@ethz.ch