Hallmarks of Optimization Trajectories

in Neural Networks & LLMs: Directional Exploration and Redundancy

Sidak Pal Singh, Bobby He, Thomas Hofmann, Bernhard Schölkopf

ETH zürich

"Implicit bias" **Optimization Trajectories $\Longrightarrow$ Loss Landscapes**

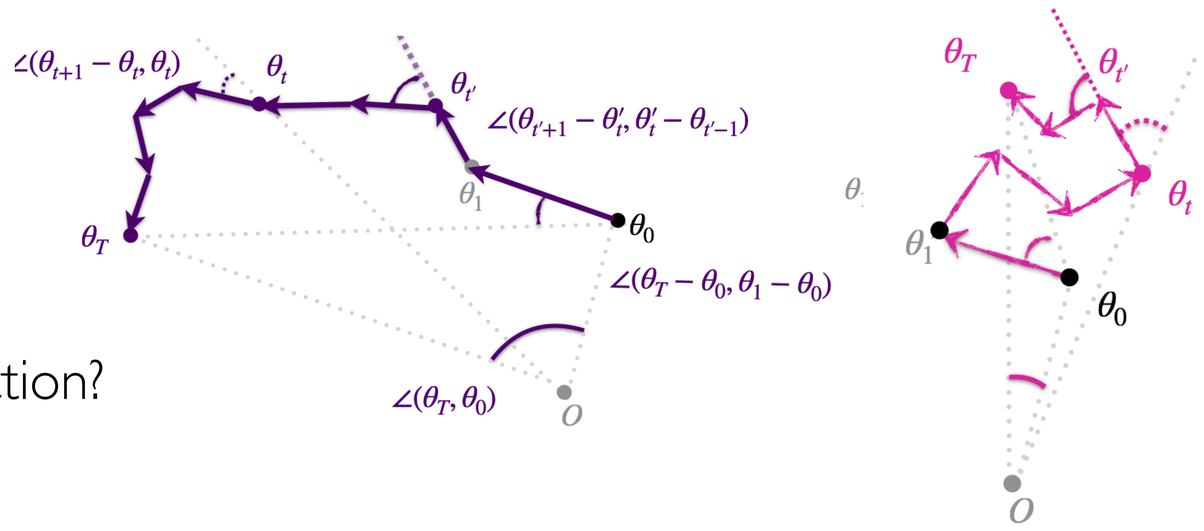
Are marks of regularity visible in the optimization trajectories?

Fundamental Questions:

Q. How are the optimization trajectories structured?

Q. Do they have a lot of twists and turns, or are they straight and direct?

Q. And does this depend on the phase of optimization?



Methodology: Trajectory Map

Take T checkpoints & arrange them in a matrix $\Theta \in \mathbb{R}^{T\times p}$, where p is # parameters

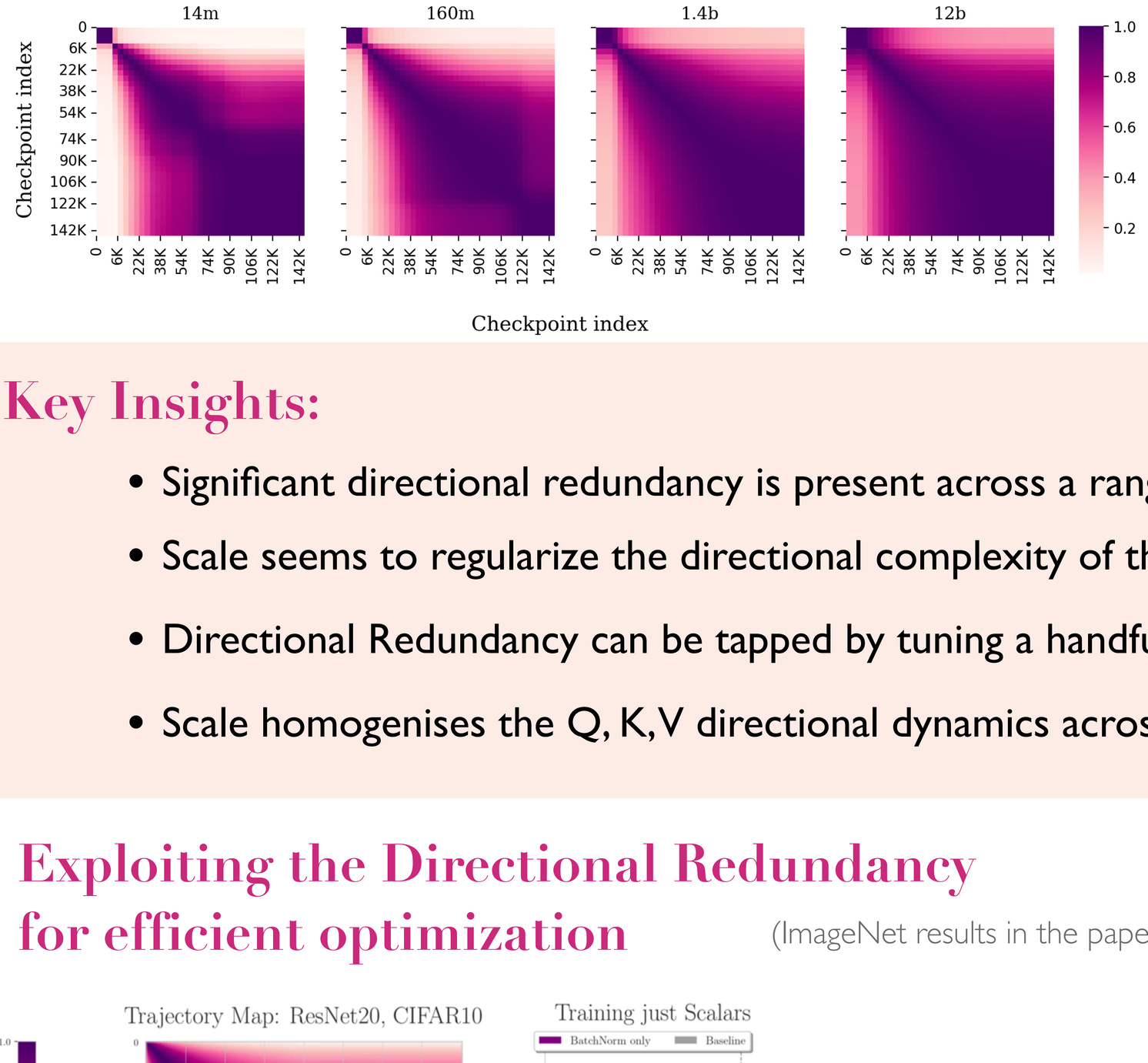
Form a matrix $\mathbf{C}$ of cosine-similarities

$$(\mathbf{C})_{ij} = \frac{\langle \theta_i, \theta_j \rangle}{\|\theta_i\| \, \|\theta_j\|}$$

Compute the Mean Directional Similarity

$$\omega := \frac{1}{T^2} \mathbf{1}_T^\top \mathbf{C} \, \mathbf{1}_T$$

Trajectory Maps for LLMs of increasing size



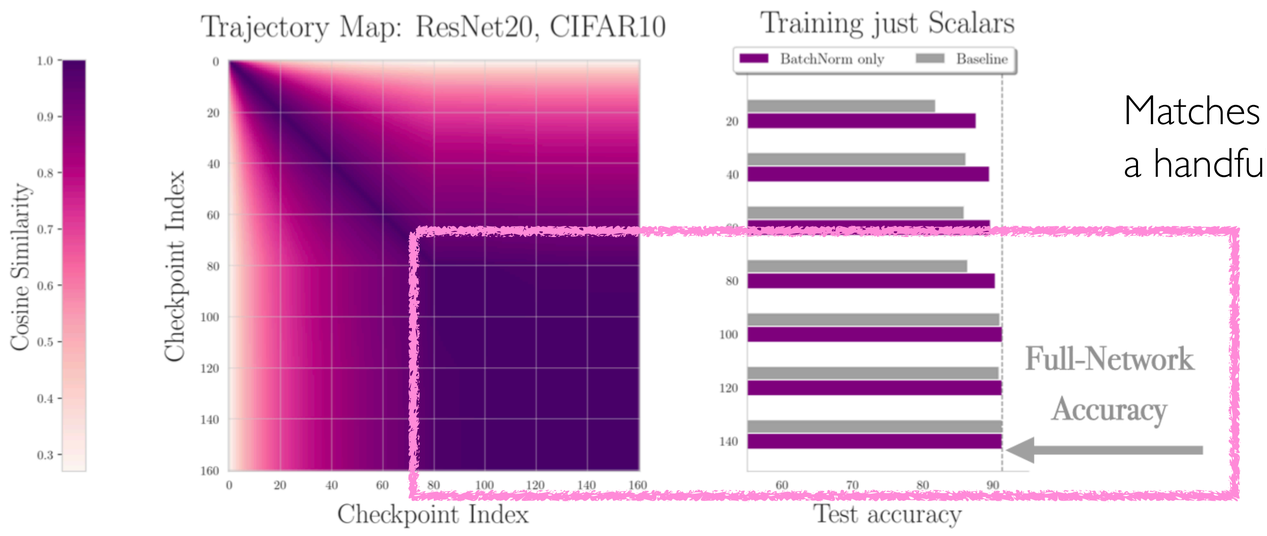
- GPT Neo-X models from Pythia: across 3 orders of magnitude
- MDS increases from 0.65 to 0.82

Key Insights:

- Significant directional redundancy is present across a range of LLMs (as well as Vision models)
- Scale seems to regularize the directional complexity of the trajectories
- Directional Redundancy can be tapped by tuning a handful of scalars (BN/LN parameters)
- Scale homogenises the Q, K, V directional dynamics across depth

Exploiting the Directional Redundancy for efficient optimization

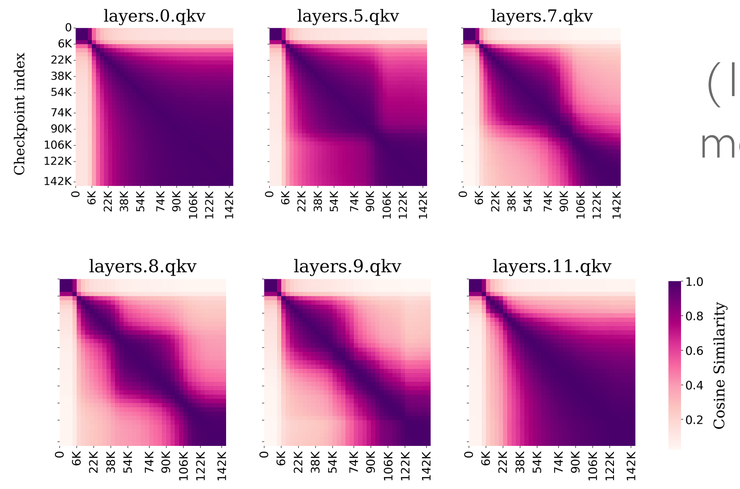
(ImageNet results in the paper)



Matches performance with a handful parameters;

while affording savings in compute & memory

Layerwise Q, K, V dynamics



(160M model)

Middle Layers converge the last directionally