Training Data Attribution with Approximate Unrolling

Juhan Bae[1,2], Wu Lin[2], Jonathan Lorraine[1,2,3], Roger Grosse[1,2,4,5]
[1]University of Toronto, [2]Vector Institute, [3]NVIDIA, [4]Anthropic, [5]Schwartz Reisman Institute

VECTOR INSTITUTE

UNIVERSITY OF TORONTO

Training Data Attribution

- Training data attribution (TDA) techniques are motivated by understanding the relationship between training data and the properties of trained models.
- Many TDA methods aim to perform a counterfactual prediction, which estimates how a model's behavior would change if certain data points were removed from (or added to) the training dataset.

Implicit-differentiation-based TDA

- Implicit-differentiation-based TDA (e.g., influence functions) uses the Implicit Function Theorem to estimate the sensitivity of the optimal solution $\boldsymbol{\theta}^\star$ to downweighting a training data point $\mathbf{z}$:

$$\boldsymbol{\theta}^\star_{\text{removed}} \approx \boldsymbol{\theta}^\star + \frac{1}{N}\mathbf{H}_{\boldsymbol{\theta}^\star}^{-1}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}^\star, \mathbf{z}).$$

- These methods provide convenient estimation algorithms that depend solely on the optimal model parameters rather than intermediate checkpoints throughout training.
- However, the classical formulation relies on assumptions such as the uniqueness of and convergence to the optimal solution.

Unrolling-based TDA

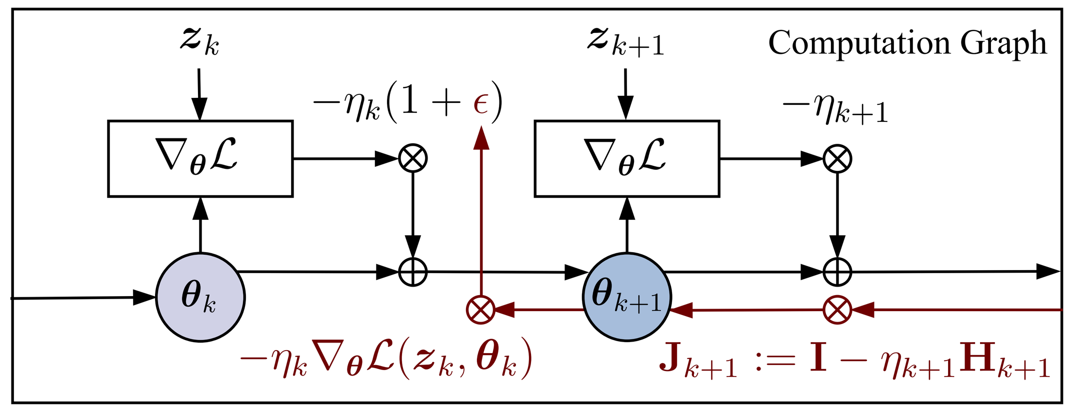
- Consider an update rule at iteration k with a data point weight ϵ:

$$\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k - (1+\epsilon)\eta_k \nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}_k, \mathbf{z}_k).$$

- Unrolling-based TDA methods estimate the effect of removing a data point $\mathbf{z}$ on the final parameters $\boldsymbol{\theta}_T$ by backpropagating through the preceding optimization steps:

$$\begin{aligned}\boldsymbol{\theta}_T^{\text{removed}} &\approx \boldsymbol{\theta}_T - \frac{\mathrm{d}\boldsymbol{\theta}_T}{\mathrm{d}\epsilon}\Big|_{\epsilon=0} \\ &= \boldsymbol{\theta}_T - \frac{\partial \boldsymbol{\theta}_T}{\mathrm{d}\boldsymbol{\theta}_{T-1}}\cdots\frac{\partial \boldsymbol{\theta}_{k+2}}{\mathrm{d}\boldsymbol{\theta}_{k+1}}\frac{\partial \boldsymbol{\theta}_{k+1}}{\partial \epsilon}\Big|_{\epsilon=0} \\ &= \boldsymbol{\theta}_T - (\mathbf{I} - \eta_{T-1}\mathbf{H}_{T-1})\cdots(\mathbf{I} - \eta_{k+1}\mathbf{H}_{k+1})(-\eta_k\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}_k, \mathbf{z}_k)).\end{aligned}$$

- They do not rely on the uniqueness of or convergence to the optimal solution and can incorporate details of training process.
- However, they require storing all intermediate variables during the training process for backpropagation (e.g., parameter vectors for each optimization step).

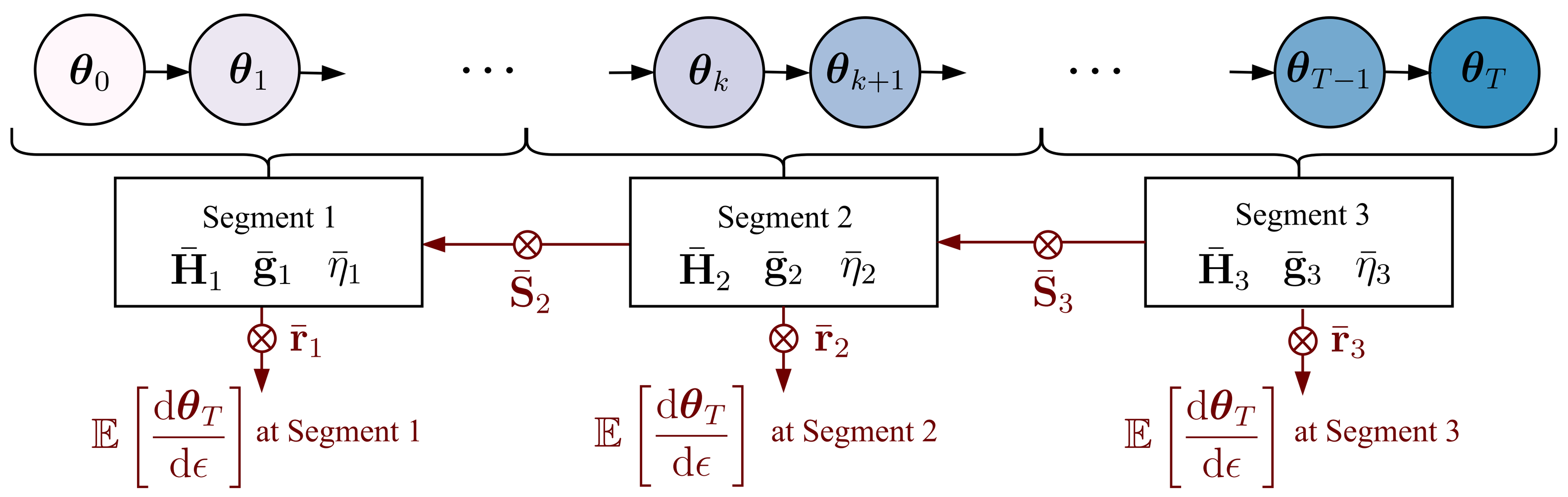


SOURCE (Our Proposed Algorithm)

TDA Strategy	Number of Checkpoints	Allows Non-Convergence	Supports Multi-Stage	Incorporates Optimizer
Implicit Differentiation	1	✗	✗	✗
Unrolling	T	✓	✓	✓
SOURCE **(ours)**	C ($\ll T$)	✓	✓	✓

- In this work, we connect implicit-differentiation-based and unrolling-based TDA approaches and introduce SOURCE that enjoys the advantages of both methods.
- SOURCE inherits three key advantages from unrolling-based methods:
  1. It enables TDA analysis for multi-stage training pipelines (e.g., foundational models and continual learning).
  2. It can incorporate algorithmic choices into the analysis (e.g., SGD vs. Adam).
  3. It maintains a close connection to counterfactual predictions even when implicit-differentiation assumptions fail (e.g., non-converged parameters).
- Unlike previous unrolling approaches, SOURCE achieves these benefits while requiring only a small number of model checkpoints C (e.g., $C=5$) rather than storing the entire training trajectory.

Segmenting the Training Trajectory



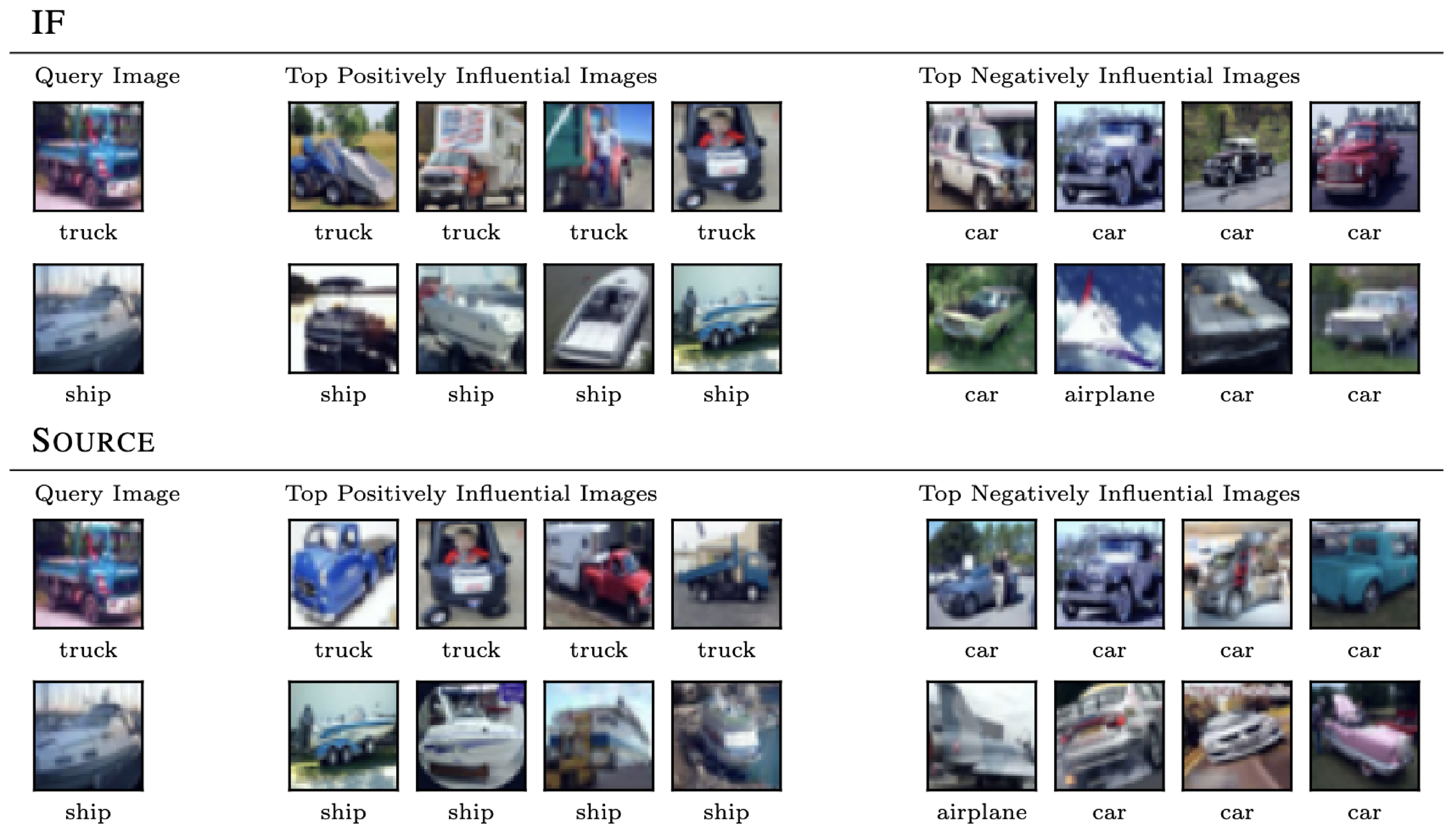
- **Key Idea:** SOURCE partitions the training trajectory into one or more segments and approximates the distributions of gradients and Hessians as stationary within each segment.
- Given L segments, SOURCE approximates the expected total gradient over the data point ordering as:

$$\mathbb{E}\left[\frac{\mathrm{d}\boldsymbol{\theta}_T}{\mathrm{d}\epsilon}\right] \approx -\sum_{\ell=1}^{L}\left(\prod_{\ell'=L}^{\ell+1}\mathbb{E}\left[\mathbf{S}_{\ell'}\right]\right)\mathbb{E}\left[\mathbf{r}_\ell\right],$$

where these quantities are computed using segment-specific averaged Hessians $\bar{\mathbf{H}}_\ell$ and gradients $\bar{\mathbf{g}}_\ell$. Note $\mathbb{E}\left[\mathbf{S}_\ell\right] := \exp(-\bar{\eta}_\ell K_\ell \bar{\mathbf{H}}_\ell)$ and $\mathbb{E}\left[\mathbf{r}_\ell\right] := \frac{1}{N}(-\exp(-\bar{\eta}_\ell K_\ell \bar{\mathbf{H}}_\ell))\bar{\mathbf{H}}_\ell^{-1}\bar{\mathbf{g}}_\ell$, where K_ℓ is the total number of iterations performed in the specified segment.

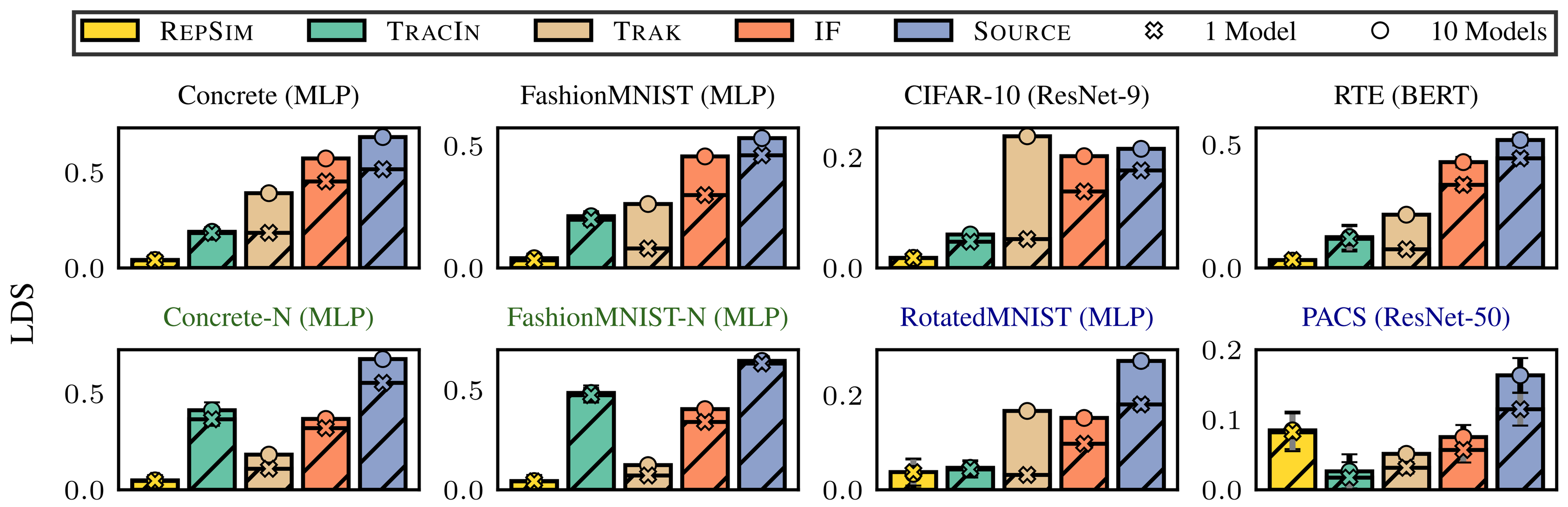
- For practical implementation, we use EK-FAC to efficiently approximate the Hessian, with both averaged Hessian and gradient estimates computed using a set of checkpoints within each segment.
- SOURCE is C times more computationally expensive than EK-FAC influence functions.

Qualitative Results



Linear Datamodeling Score (LDS)

- The LDS measures the Spearman correlation between the estimated quantities after retraining the model without a subset of data points and the predictions made by the TDA method.
- SOURCE especially performs strongly against other baseline techniques on settings that pose challenges to implicit-differentiation-based approaches (e.g., non-converged models and models trained with multiple stages).



Subset Removal Evaluation