Utilizing Insights from Optimization Trajectories of Deep Learning

Boyang "Albert" Li

Nanyang Associate Professor & NRF Fellow
School of Computer Science and Engineering,
Nanyang Technological University, Singapore

Insights from Optimization Trajectories of Deep Learning

- Optimizing the loss function $\mathcal{L}(w)$ over network parameters w

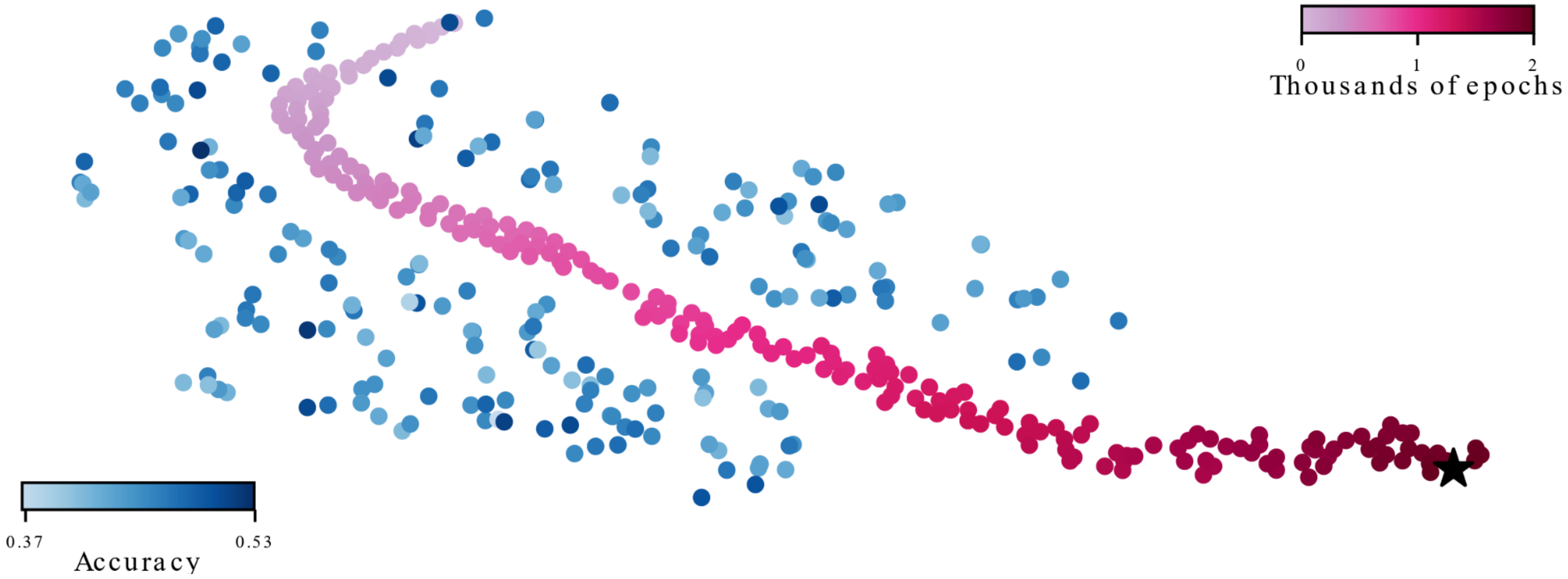
$$w \leftarrow w - \eta \frac{d\mathcal{L}(w)}{dw}$$

$$\mathcal{L}(w) = \sum_i \ell(x^{(i)}, y^{(i)}, w)$$

Summation over training data points

- The simple method above (with some additional details) achieves surprisingly good results.
- How is this possible?
  - What about non-convex losses (multiple local minima)?

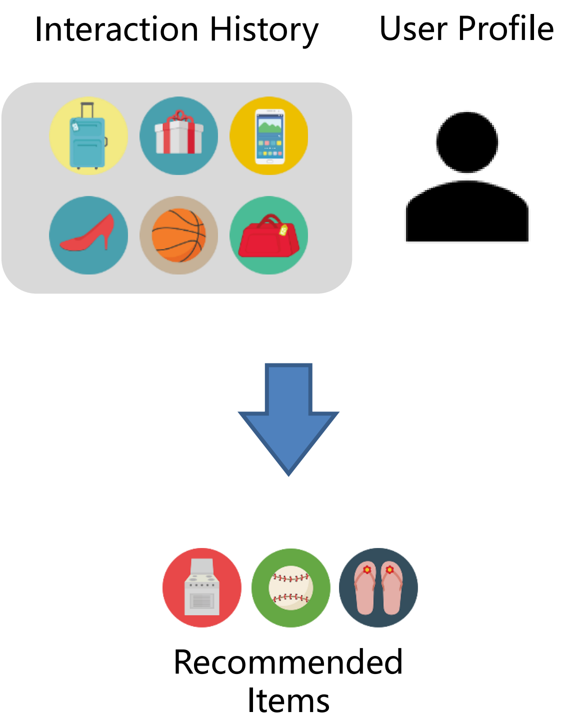
The Optimization Trajectory of Deep Learning



- Red dots: the iterates of SGD after each tenth epoch.
- Blue dots: locations of nearby "bad" minima with perfect train accuracy but poor generalization.
- The final iterate of SGD (black star) also achieves perfect train accuracy, but with 98.5% test accuracy. Miraculously, SGD always finds its way through a landscape full of bad minima, and lands at a minimizer with excellent generalization.

W. Ronny Huang, Zeyad Emam, Micah Goldblum, Liam Fowl, Justin K. Terry, Furong Huang, Tom Goldstein. Understanding Generalization through Visualizations. 2019

Utilize Neighborhood Information? Initialization Matters.



- Many neural recommender systems are outperformed by simple nearest neighbor methods [1].

[1] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches. RecSys 2019.

- Neighborhood-informed initialization boosts multiple deep learning methods above nearest neighbors and other simple baselines [2].

[2] Yinan Zhang, Boyang Li, Yong Liu, Hao Wang, Chunyan Miao. Initialization Matters: Regularizing Manifold-informed Initialization for Neural Recommendation Systems. KDD 2021.

Which training points affect predictions? Trajectory Matters.

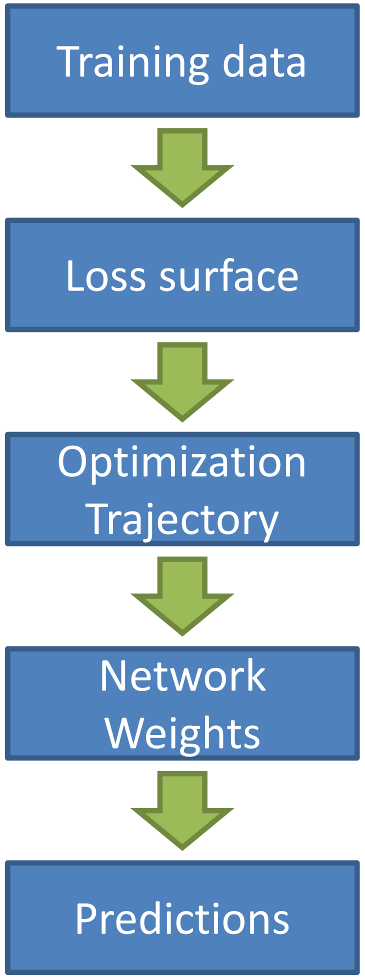
$x^{(i)}, y^{(i)}$

Remove?

$\sum_i \ell(x^{(i)}, y^{(i)}, w)$

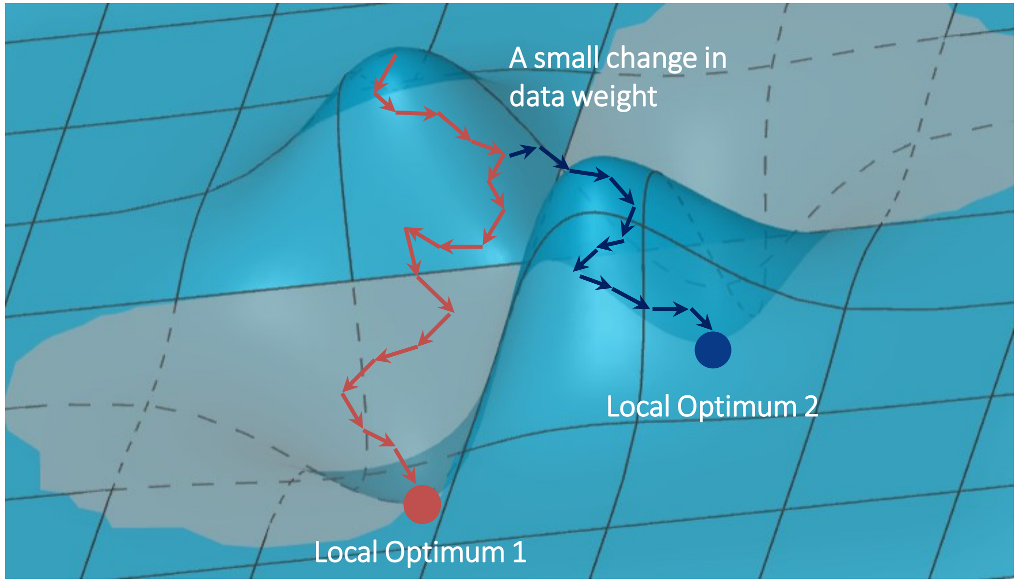
$\left.\frac{d\mathcal{L}(w)}{dw}\right|_{w_t}$

w_T



- Previous works like [3] do not model the change in the entire optimization trajectory.

[3] Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. ICML 2017.



Which training points affect predictions? Trajectory Matters.

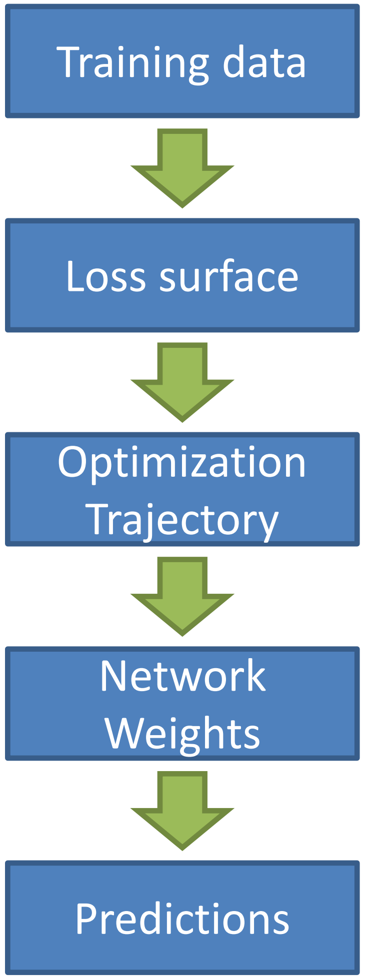
$x^{(i)}, y^{(i)}$

Remove?

$\sum_i \ell(x^{(i)}, y^{(i)}, w)$

$\left.\frac{d\mathcal{L}(w)}{dw}\right|_{w_t}$

w_T



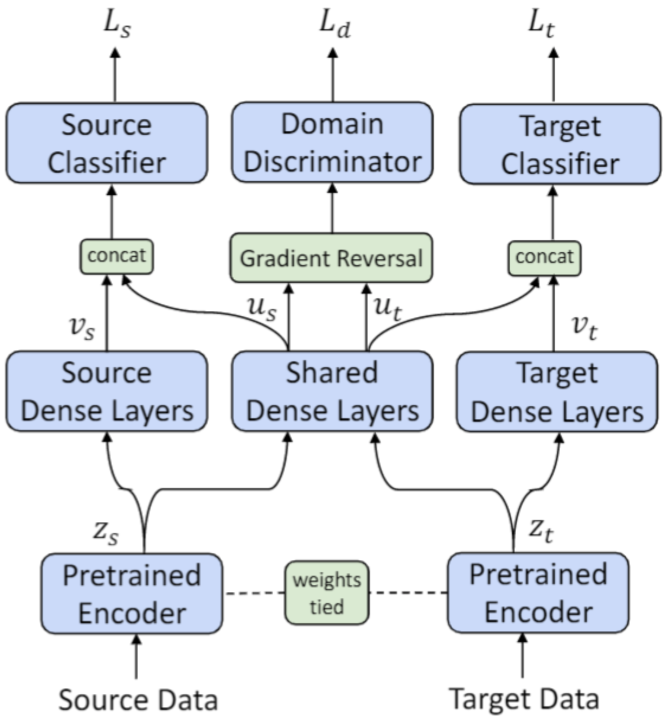
- Previous works like [3] do not model the change in the entire optimization trajectory.

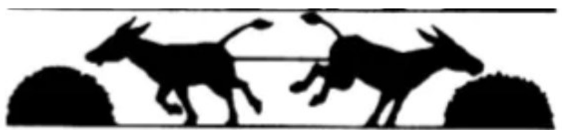
[3] Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. ICML 2017.

- In [4], we explicitly consider the change in the trajectory and propose an approximation algorithm with bounded and diminishing errors.

[4] Yuanyuan Chen, Boyang Li, Han Yu, Pengcheng Wu, and Chunyan Miao. HyDRA: Hypergradient Data Relevance Analysis for Interpreting Deep Neural Networks. AAAI 2021.

Multiple Losses? Their Interaction Matters.

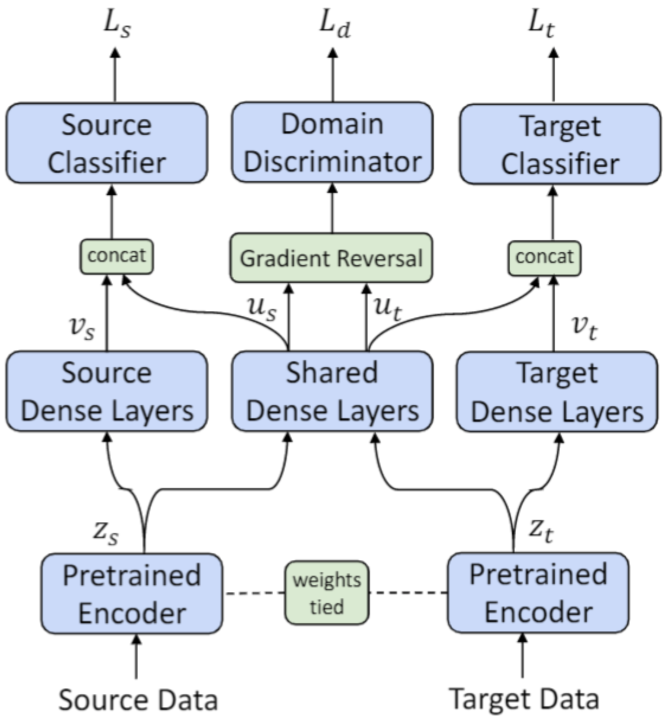


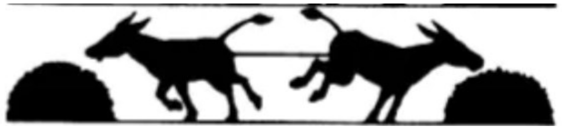


- In this common transfer learning setup, the domain discriminator encourages the source-domain and target-domain features to be similar.
- However, this can create difficulties in optimization.
- We encourage the gradients of different losses to point in the same direction, which improves transfer.

[5] Xu Guo, Boyang Li, Han Yu, and Chunyan Miao. Latent-Optimized Adversarial Neural Transfer for Sarcasm Detection. NAACL 2021.

Multiple Losses? Their Interaction Matters.





- First, take a GD step on L_d with latent representation z_s and z_t

$$z'_s = z_s - \gamma \frac{dL_d}{dz_s}, \qquad z'_t = z_t - \gamma \frac{dL_d}{dz_t}$$

- After that, optimize domain-specific losses on z'_s and z'_t

$$\mathcal{L} = L_s(z'_s) + L_t(z'_t) + L_d(z_s, z_t)$$

- Why does this work? By first-order Taylor expansion

$$L_s(z'_s) \approx L_s(z_s) + \frac{dL_s(z_s)}{dz_s}\left(-\gamma \frac{dL_d}{dz_s}\right)$$

- Minimizing $L_s(z'_s)$ is to encourage $\frac{dL_s(z_s)}{dz_s}$ and $\frac{dL_d}{dz_s}$ to have the similar directions.

[5] Xu Guo, Boyang Li, Han Yu, and Chunyan Miao. Latent-Optimized Adversarial Neural Transfer for Sarcasm Detection. NAACL 2021.