Essays

Haochen Ding

August 2025

1 Understanding Black-box Predictions via Influence Functions

The first paper studies how individual training data points affect a model's predictions through adapting influence functions. Mathematically, the influence of one training point on the model is calculated as the derivative of the minimizer after perturbing the data point, which depends on the data shift with respect to the data shift. The paper calculated the effect of both upweighting and perturbing a training input, on both the minimizer and functions (in particular the loss function) of the minimizer. Computing the real Hessian matrix can be challenging, thus the paper proposed two techniques of approximating the influence: conjugate gradients and stochastic estimation. The result of the paper is then validated in comparison with the results of leave-one-out retraining. As for applications, influence functions can be applied to identify mislabeled or outlier training points and explain certain test predictions.

2 Training Data Attribution via Approximate Unrolled Differentiation

The second paper proposed a novel TDA algorithm called SOURCE that enjoys the advantages of both implicit-differentiation-based and unrolling-based approaches. Unrolling keeps track of all the optimization variables and uses the chain rule to calculate the contribution of a single iteration to the total influence by multiplying all the Jacobian matrices. Although this method improves upon influence functions with its capability of dealing with non-convergence and multi-stage training, the fact that it needs to keep all intermediate variables makes it expensive. SOURCE reduces the cost by segmenting the training process into several approximately independent and stationary segments. With this approximation, the Hessians and gradients of each segment can be estimated using only a handful of checkpoints in each segment, and the formula for calculating the influence can be simplified, which greatly

reduces the memory cost of unrolling. The result of SOURCE is an influence-function-like formula, which also simplifies computations. SOURCE is suitable in cases where implicit-differentiation-based approaches struggle, such as in non-converged models and multi-stage training pipelines.

3 TRAK: Attributing Model Behavior at Scale

The third paper proposed a new metric, LDS, for evaluating data attribution methods along with a new data attribution method, TRAK, that aims to be both effective and scalable in large-scale differentiable settings. TRAK first simplifies large-scale non-linear models by linearizing the model output function into logistic regression, transforming the model into parameter space with gradient input features. Then it reduces the dimensions of input gradients using the Johnson-Lindenstrauss method. After these simplifications, TRAK approximates the influence using the One-step Newton approximation. In order to smooth out the impact of non-deterministic training, TRAK takes random subsets of training examples and computes the terms separately. Since for neural networks attribution scores are often sparse, TRAK post-processes the scores via soft thresholding. The paper applies TRAK to both binary and multi-class classification tasks and evaluates TRAK using LDS. For applications, the paper demonstrates the utility of TRAK across various modalities and scales: image classifiers trained on ImageNet, vision-language models (CLIP), and language models (BERT and mT5).

4 Detailed Summary of Training Data Attribution via Approximate Unrolled Differentiation

Unrolling is a TDA method that approximates the impact of downweighting a data point's gradient update on the final model parameters by backpropagating through the preceding optimization steps. Compared to influence functions, it does not rely on the uniqueness of or convergence to the optimal solution and keeps track of multiple training stages, making it more suitable for modern neural networks. However, it is expensive because it requires either storing or recomputing every step of the training trajectory.

The SOURCE algorithm is motivated by the unrolling TDA approach and aims to reduce the cost by segmenting the overall training into different segments that are approximately independent and stationary. This approximation allows us to efficiently calculate the Hessians and gradients using a handful of checkpoints, and it also simplifies the formula of influence

calculation.

According to unrolled differentiation, the contribution of iteration k to the total derivative can be found by multiplying all the Jacobian matrices along the accumulation path to the local gradient. The effect of perturbing a training example is measured by the derivative of the minimizer with respect to the perturbation $\frac{d\theta_T}{d\epsilon}$, and SOURCE averages the derivative across multiple training trajectories since the effect of removing z_m on any single training trajectory may be noisy and idiosyncratic. It calculates the derivative using

$$\mathbb{E}\left[\frac{d\theta_T}{d\epsilon}\right] = -\sum_{k=0}^{T-1} \frac{\eta_k}{B}\, \mathbb{E}\big[\delta_k J_{k+1:T} g_k\big]. \tag{1}$$

After segmentation, the overall formula can be further reduced into calculating terms containing averaged Hessians, gradients, and step sizes. To be more specific,

$$\mathbb{E}\left[\frac{d\theta_T}{d\epsilon}\right] = -\mathbb{E}\left[\sum_{\ell=1}^{L}\Big(\prod_{\ell'=\ell+1}^{L} S_{\ell'}\Big) r_\ell\right] \;\approx\; -\sum_{\ell=1}^{L}\Big(\prod_{\ell'=\ell+1}^{L} \mathbb{E}[S_{\ell'}]\Big)\,\mathbb{E}[r_\ell], \tag{2}$$

where $S_{\ell'}$ and r_ℓ can be calculated using approximated Hessians, gradients, and step sizes:

$$\mathbb{E}[S_\ell] \;\approx\; \exp(-\bar{\eta}_\ell K_\ell \bar{H}_\ell) := \bar{S}_\ell, \tag{3}$$

$$\mathbb{E}[r_\ell] \approx \frac{1}{N}\Big(I - e^{-\bar{\eta}_\ell K_\ell \bar{H}_\ell}\Big)\bar{H}_\ell^{-1}\bar{g}_\ell := \bar{r}_\ell. \tag{4}$$

Due to the approximation that the segments are stationary and independent of one another, calculating these variables only requires a handful of checkpoints. If the stationary approximations are too inaccurate, we can carve the segments smaller, at the expense of computational and memory requirements. The final SOURCE formula is shown below. Being motivated by unrolling approach, SOURCE ends up in a influence-function-like formula that is computationally cheaper to calculate. SOURCE uses EK-FAC parameterization with GNH approximation to further accelerate the computation of Hessians and their exponentials due to the fact that approximated Hessians have explicit eigen decompositions.

$$\mathbb{E}\left[\frac{d\theta_T}{d\epsilon}\right] \;\approx\; -\frac{1}{N}\sum_{\ell=1}^{L}\Big(\prod_{\ell'=\ell+1}^{L} \bar{S}_{\ell'}\Big)\,\bar{r}_\ell. \tag{5}$$