INFLUENCE FUNCTIONS IN DEEP LEARNING ARE FRAGILE

Anonymous authors
Paper under double-blind review

ABSTRACT

Influence functions approximate the effect of training samples in test-time predictions and have a wide variety of applications in machine learning interpretability and uncertainty estimation. A commonly-used (first-order) influence function can be implemented efficiently as a post-hoc method requiring access only to the gradients and Hessian of the model. For linear models, influence functions are well-defined due to the convexity of the underlying loss function and are generally accurate even across difficult settings where model changes are fairly large such as estimating group influences. Influence functions, however, are not well-understood in the context of deep learning with non-convex loss functions. In this paper, we provide a comprehensive and large-scale empirical study of successes and failures of influence functions in neural network models trained on datasets such as Iris, MNIST, CIFAR-10 and ImageNet. Through our extensive experiments, we show that the network architecture, its depth and width, as well as the extent of model parameterization and regularization techniques have strong effects in the accuracy of influence functions. In particular, we find that (i) influence estimates are fairly accurate for shallow networks, while for deeper networks the estimates are often erroneous; (ii) for certain network architectures and datasets, training with weight-decay regularization is important to get high-quality influence estimates; and (iii) the accuracy of influence estimates can vary significantly depending on the examined test points. These results suggest that in general influence functions in deep learning are fragile and call for developing improved influence estimation methods to mitigate these issues in non-convex setups.

1 INTRODUCTION

In machine learning, influence functions (Cook & Weisberg, 1980) can be used to estimate the change in model parameters when the empirical weight distribution of the training samples is perturbed infinitesimally. This approximation is cheaper to compute compared to the expensive process of repeatedly re-training the model to retrieve the exact parameter changes. Influence functions could thus be used to understand the effect of removing an individual training point (or, groups of training samples) on the model predictions at the test-time. Leveraging a first-order Taylor's approximation of the loss function, (Koh & Liang, 2017) has shown that a (first-order) influence function, computed using the gradient and the Hessian of the loss function, can be useful to interpret machine learning models, fix mislabelled training samples and create data poisoning attacks.

Influence functions are in general well-defined and studied for models such as logistic regression (Koh & Liang, 2017), where the underlying loss-function is convex. For convex loss functions, influence functions are also accurate even when the model perturbations are fairly large (e.g. in the group influence case (Koh et al., 2019b; Basu et al., 2019)). However, when the convexity assumption of the underlying loss function is violated, which is the case in deep learning, the behaviour of influence functions is not well understood and is still an open area of research. With recent advances in computer vision (Szeliski, 2010), natural language processing (Sebastiani, 2002), high-stakes applications such as medicine (Lundervold & Lundervold, 2018), it has become particularly important to interpret deep model predictions. This makes it critical to understand influence functions in the context of deep learning, which is the main focus of our paper.

Despite their non-convexity, it is sometimes believed that influence functions would work for deep networks. The excellent work of (Koh & Liang, 2017) successfully demonstrated one example of

influence estimation for a deep network, a small (2600 parameters), "all-convolutional" network (Springenberg et al., 2015). To the best of our knowledge, this is the one of the *few* cases for deep networks where influence estimation has been shown to work. A question of key importance to practitioners then arises: for what other classes of deep networks does influence estimation work? In this work, we provide a comprehensive study of this question and find a pessimistic answer: *influence estimation is quite fragile for a variety of deep networks.*

In the case of deep networks, several factors might have an impact on influence estimates: (i) due to non-convexity of the loss function, different initializations of the perturbed model can lead to significantly different model parameters (with approximately similar loss values); (ii) even if the initialization of the model is fixed, the curvature values of the network (i.e. eigenvalues of the Hessian matrix) at optimal model parameters might be very large in very deep networks, leading to a significant Taylor's approximation error of the loss function and thus resulting in poor influence estimates; (iii) for large neural networks, computing the exact inverse-Hessian Vector product, required in computation of influence estimates, can be computationally very expensive. Thus, one needs to use approximate inverse-Hessian Vector product techniques which might be erroneous; resulting in low quality influence estimates; and finally (iv) different architectures can have different loss landscape geometries near the optimal model parameters, leading to varying influence estimates.

In this paper, we study aforementioned issues of using influence functions in deep learning through an extensive experimental study on progressively-growing complex models and datasets. We first start our analysis with a case study of a small neural network for the Iris dataset where the exact Hessian matrix can be computed. We then progressively increase the complexity of the network and analyse a CNN architecture (depth of 6) trained on 10% of MNIST dataset, similar to (Koh & Liang, 2017). Next, we evaluate the accuracy of influence estimates for more complex deep architectures (e.g. ResNets) trained on MNIST and CIFAR-10. Finally, we compute influence estimates on the ImageNet dataset using ResNet-50.

We make the following observations through our analysis:

- We find that the network depth and width have a strong impact on influence estimates. In particular, we show that influence estimates are fairly accurate when the network is shallow, while for deeper models, influence estimates are often erroneous. We attribute this partially to the increasing curvature values of the network as the depth increases.
- We observe that the weight decay regularization is important to obtain high quality influence estimates in certain architectures and datasets.
- We show that the inverse-Hessian Vector product approximation techniques such as stochastic estimation (Agarwal et al., 2016) are erroneous, especially when the network is deep. This can contribute to the low quality of influence estimates in deep models.
- We observe that the choice of test-point has a significant impact on the quality of influence estimates, across different datasets and architectures.
- In very large-scale datasets such as ImageNet, we have found that even ground-truth influence estimates (obtained by leave-one-out re-training) can be inaccurate and noisy partially due to the model's training and convergence.

These results highlight sensitivity of current influence functions in deep learning and call for developing robust influence estimators to be used in large-scale machine learning applications.

2 Related Works

Influence functions are primarily used to identify important training samples for test-time predictions and debug machine learning models (Koh & Liang, 2017). In recent times, there is an increase in the applications of influence functions for tasks other than interpretability. For e.g.(Schulam & Saria, 2019) has used influence functions to audit the reliability of test-predictions. In NLP, influence functions have been used to detect biases in word-embeddings (Brunet et al., 2018) whereas in the domain of ML security, influence functions have been shown to be effective in crafting stronger data-poisoning attacks (Koh et al., 2019a). Influence functions are also effective in the identification of important training groups (rather than an individual sample) (Basu et al., 2019; Koh et al., 2019b). Prior theoretical work (Giordano et al., 2018; 2019) have focused on quantifying finite sample error-bounds for influence estimates when compared to the ground-truth re-training procedures. Recently,

alternative methods to find influential samples in deep networks have been proposed. In (Yeh et al., 2018), test-time predictions are explained by a kernel function evaluated at the training samples. Influential training examples can also be obtained by tracking the change in loss for a test-prediction through model-checkpoints, which are stored during the training time (Pruthi et al., 2020). While these alternative methods (Yeh et al., 2018; Pruthi et al., 2020) work well for deep networks in interpreting model predictions, they lack the "jackknife" like ability of influence functions which makes it useful in multiple applications other than interpretability (e.g. uncertainty estimation).

3 Basics of Influence Function

Consider h to be a function parameterized by θ which maps from an input feature space $\mathcal{X}$ to an output space denoted by $\mathcal{Y}$. The training samples are denoted by the set $\mathcal{S} = \{z_i : (x_i, y_i)\}_{i=1}^n$, while the loss function is represented by $\ell(h_\theta(z))$ for a particular training example z. The standard empirical risk minimization solves the following optimization problem:

$$\theta^* = \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} \ell(h_\theta(z_i)). \tag{1}$$

Up-weighting a training example z by an infinitesimal amount ϵ leads to a new set of model parameters denoted by $\theta^\epsilon_{\{z\}}$. This set of new model parameters $\theta^\epsilon_{\{z\}}$ is obtained by solving:

$$\theta^\epsilon_{\{z\}} = \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} \ell(h_\theta(z_i)) + \epsilon\ell(h_\theta(z)). \tag{2}$$

Removing a training point z is similar to up-weighting its corresponding weight by $\epsilon = -1/n$ in Equation(2). The main idea used by (Koh & Liang, 2017) is to approximate $\theta^\epsilon_{\{z\}}$ by the first-order Taylor series expansion around the optimal model parameters represented by θ^*, which leads to:

$$\theta^\epsilon_{\{z\}} \approx \theta^* - \epsilon H^{-1}_{\theta^*} \nabla_\theta \ell(h_{\theta^*}(z)), \tag{3}$$

where H_{θ^*} represents the Hessian with respect to model parameters θ^*. Following the classical result of (Cook & Weisberg, 1980), the change in the model parameters ($\Delta\theta = \theta^\epsilon_{\{z\}} - \theta^*$) on up-weighting the training example z can be approximated by the influence function ($\mathcal{I}(z)$) as follows:

$$\mathcal{I}(z) = \frac{d\theta^\epsilon_{\{z\}}}{d\epsilon}|_{\epsilon=0} = -H^{-1}_{\theta^*} \nabla_\theta \ell\left(h_{\theta^*}(z)\right). \tag{4}$$

The change in the loss value for a particular test point z_t when a training point z is up-weighted can be approximated as a closed form expression by the chain rule (Koh & Liang, 2017):

$$\mathcal{I}(z, z_t) = -\nabla\ell(h_{\theta^*}(z_t))^T H^{-1}_{\theta^*} \nabla\ell(h_{\theta^*}(z)). \tag{5}$$

$\mathcal{I}(z, z_t)/n$ is approximately the change in the loss for the test-sample z_t when a training sample z is removed from the training set. This result is, however, based on the assumption that the underlying loss function is strictly convex in the model parameters θ and the Hessian H_{θ^*} is a positive-definite matrix (Koh & Liang, 2017). For large models, inverting the exact Hessian H_{θ^*} is expensive. In such cases, the inverse-Hessian Vector product can be computed efficiently with a combination of Hessian-vector product (Pearlmutter, 1994) and optimization techniques (see Appendix for details).

4 What Can Go Wrong for Influence Functions In Deep Learning?

First-order influence functions (Koh & Liang, 2017) assume that the underlying loss function is convex and the change in model parameters is small when the empirical weight distribution of the training data is infinitesimally perturbed. In essence, this denotes the Taylor's gap in Equation (3) to be small for an accurate influence estimate. However in the case of non-convex loss functions, this assumption is *not* generally true. Empirically, we find that the Taylor's gap is strongly affected by common hyper-parameters for deep networks. For example, in Fig. (1)-(a,b), we find that for networks trained without a weight-decay regularization on Iris, the Taylor's gap is large resulting in low quality influence estimates. In a similar vein, when the network depth and width is significantly large (i.e. the over-parameterized regime), the Taylor's gap increases and substantially degrades the quality of influence estimates (Fig. (2)). Empirically this increase in Taylor's gap strongly correlates

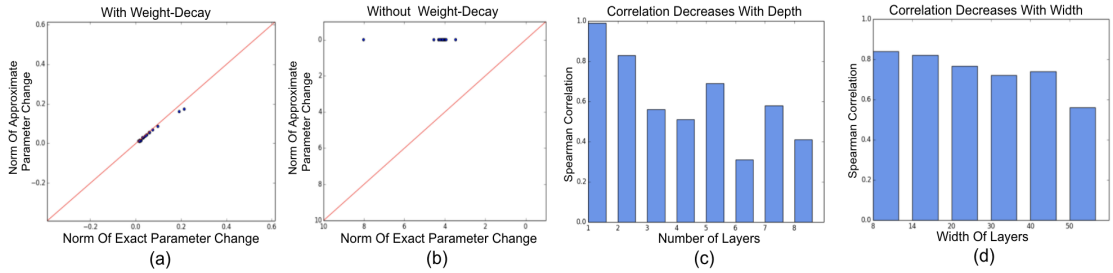

Figure 1: Iris dataset experimental results - (a,b) Comparison of norm of parameter changes computed with influence function vs re-training; (a) trained with weight-decay; (b) trained without weight-decay. (c) Spearman correlation vs. network depth. (d) Spearman correlation vs. network width.

with the curvature values of the loss function evaluated at the optimal model parameters as observed in Fig. (2-(b)).

Further complications may arise for larger models, where influence estimations in such settings require an additional approximation to compute the inverse-Hessian vector product. Nonetheless, we observe in Fig. (2)-(a), that on Iris this approximation has only a marginal impact on the influence estimation. These results show that that network architecture, hyper-parameters, and loss curvatures are significant factors for proper influence estimations. In the next section, we discuss these issues in details through controlled experiments on datasets and models of increasing complexity.

5 Experiments

Datasets: We first study the behaviour of influence functions in a small Iris dataset (Anderson, 1936), where the exact Hessian can be computed. Further, we progressively increase the complexity of the model and datasets: we use small MNIST (Koh & Liang, 2017) to evaluate the accuracy of influence functions in a small CNN architecture with a depth of 6. Next, we study influence functions on modern deep architectures trained on the standard MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky et al., 2000) datasets. Finally, to understand how influence functions scale to large datasets, we use ImageNet (Deng et al., 2009) to compute the influence estimates.

Evaluation Metrics: We evaluate the accuracy of influence estimates at a given test point z_t using both Pearson (Kirch, 2008) and Spearman rank-order correlation (Spearman, 1904) with the ground-truth (obtained by re-training the model) across a set of training points. Most of the existing interpretability methods desire that influential examples are ranked in the correct order of their importance (Ghorbani et al., 2017). Therefore, to evaluate the accuracy of influence estimates, Spearman correlation is often a better choice.

5.1 Understanding Influence Functions when the Exact Hessian Can be Computed

Setup: Computing influence estimates with the exact Hessian has certain advantages in our study: a) it bypasses inverse-Hessian Vector product approximation techniques which induce errors in computing influence estimates. Thus, we can compare influence estimates computed with exact vs. approximate inverse-Hessian Vector products to quantify this type of error; b) The deviation of the parameters computed with the influence function from the exact parameters can be computed exactly. This information can be useful to further quantify the error incurred by (first-order) influence estimates in the non-convex setup. However, computations of the exact Hessian matrix and its inverse are only computationally feasible for models with small number of parameters. Thus, we use the Iris dataset along with a small feed-forward neural network to analyse the behaviour of influence function computed with the exact Hessian in a non-convex setup. We train models to convergence for 60k iterations with full-batch gradient descent. To obtain the ground-truth estimates, we re-train the models for 7.5k steps, starting from the optimal model parameters. For our analysis, we choose the test-point with the maximum loss and evaluate the accuracy of influence estimates with the ground-truth amongst of the top 16.6% of the training points. Through our experiments with the exact Hessian, we answer some relevant questions related to how properties of the network such as depth, width and regularizers (e.g. weight-decay) affect the influence estimates.

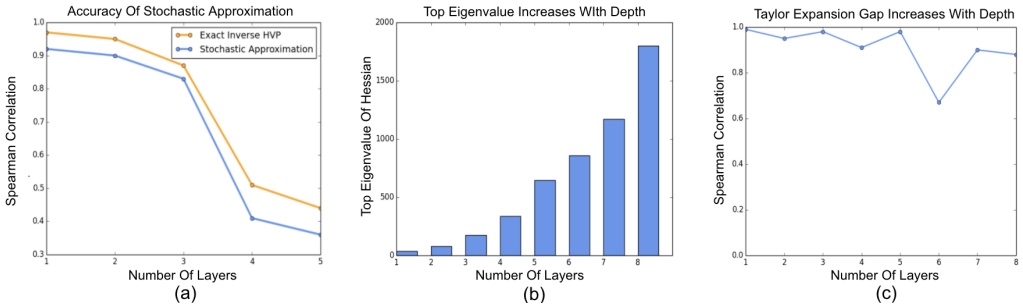

Figure 2: Iris dataset experimental results; (a) Spearman correlation of influence estimates with the ground-truth estimates computed with stochastic estimation vs. exact inverse-Hessian vector product. (b) Top eigenvalue of the Hessian vs. the network depth. (c) Spearman correlation between the norm of parameter changes computed with influence function vs. re-training.

The Effect of Weight-Decay: One of the simpler and common regularization techniques used to train neural networks is weight-decay regularization. In particular, a term $\lambda\|\theta\|_2^2$, penalizing the scaled norm of the model parameters is added to the objective function, during training, where λ is a hyperparameter which needs to be tuned. We train a simple feed-forward network[1] with and without weight-decay regularization. For the network trained with weight-decay, we observe a Spearman correlation of 0.97 between the influence estimates and the ground-truth estimates. In comparison, for the network trained without a weight-decay regularization, the Spearman correlation estimates decrease to 0.508. In this case, we notice that the Hessian matrix is singular, thus a damping factor of 0.001 is added to the Hessian matrix, to make it invertible. To further understand the reason for this decrease in the quality of influence estimates, we compare the following metric across all training examples: a) Norm of the model parameter changes computed by re-training; b) Norm of the model parameter changes computed using the influence function (i.e. $\|H_{\theta^*}^{-1}\nabla\ell(z_i)\|_2 \quad \forall i \in [1, n]$) (Fig. 1-(a,b)). We observe that when the network is trained without weight-decay, changes in model parameters computed with the influence function have a significantly larger deviation from those computed using re-training. This essentially suggests that the gap in Taylor expansion, using (first-order) influence estimates is large, when the model is trained without weight-decay. We observe similar results with smooth activation functions such as tanh (see the Appendix for details).

The Effect Of Network Depth: From Fig. 1-(c), we see that network depth has a dramatic effect on the quality of influence estimates. For example, when the depth of the network is increased to 8, we notice a significant decrease in the Spearman correlation estimates. To further our understanding about the decrease in the quality of influence estimates when the network is deeper, we compute the gap in the approximation between the ground-truth parameter changes (computed by re-training) and the approximate parameter changes (computed using the influence function). To quantify the error gap, we compute the Spearman correlation estimates between the norm of true and approximate parameter changes across the top 16.6% of the influential examples. We find that with increasing depth, the Spearman correlation estimates between the norm of the true and approximate parameter changes decrease. From Fig. 2-(c), we see that the approximation error gap is particularly large when the depth of the network is more than 5. We also notice a consistent increase in the curvature of the loss function (Fig. 2-(b)), as the network becomes deeper. This possibly suggests that the curvature information of the network can be an upper bound in the approximation error gap between the true parameters and the ones computed using the influence function. Even in case of non-smooth activation functions like ReLU, we have a similar observation. (see the Appendix for more details).

The Effect Of Network Width: To see the effect of the network width on the quality of influence estimates, we evaluate the influence estimates for a feed-forward network of constant depth, by progressively increasing its width. From Fig. 1-(d), we observe that with an increase in network width, the Spearman correlation decreases consistently. For example, we find that the Spearman correlation decreases from 0.82 to 0.56, when the width of the network is increased from 8 to 50. This observation suggests that over-parameterizing a network by increasing its width has a strong impact in the quality of influence estimates.

[1] With width of 5, depth of 1 and ReLU activations

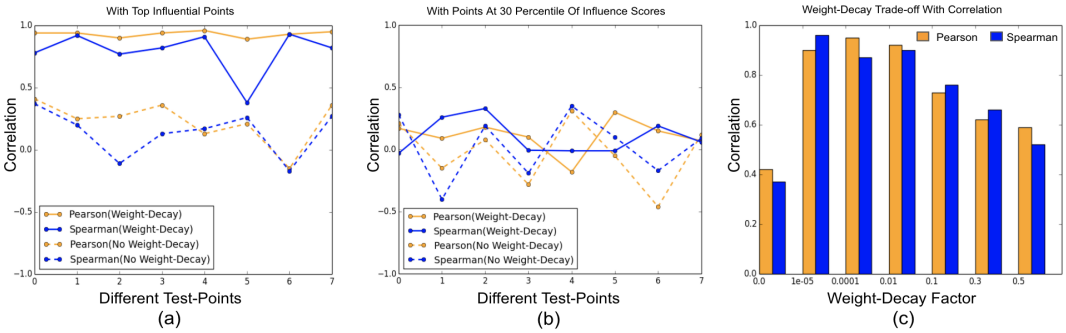

Figure 3: Experiments on small MNIST using a CNN architecture. (a) Estimation of influence function with and without weight decay on (a) the top influential points, (b) training points at 30th percentile of influence score distribution. (c) Correlation vs the weight decay factor (evaluated on the top influential points).

The Effect of Stochastic Estimation on inverse-Hessian Vector Product: For large deep networks, the inverse-Hessian Vector product is computed using stochastic estimation(Agarwal et al., 2016), as the exact Hessian matrix cannot be computed and inverted. To understand the effectiveness of stochastic approximation, we compute the influence estimates with both the exact Hessian and stochastic estimation. We observe that across different network depths, the influence estimates computed with stochastic estimation have a marginally lower Spearman correlation when compared to the ones computed with the exact Hessian. From Fig. 2-(a), we find that the error in the approximation is more, when the network is deeper.

5.2 Understanding Influence Functions in Shallow CNN Architectures

Setup: In this section, we perform a case study using a CNN architecture[2] on the small MNIST dataset (i.e. 10% of MNIST); a similar setup used in (Koh & Liang, 2017). To assess the accuracy of influence estimates, we select a set of test-points with high test-losses computed at the optimal model parameters. For each of the test points, we select 100 training samples with the highest influence scores and compute the ground-truth influence by re-training the model. We also select 100 training points with influence scores at the 30th percentile of the entire influence score distribution. These training points have low influence scores and a lower variance in their scores when compared to the top influential points. The model is trained with and without weight-decay regularization.

When trained with a weight-decay and evaluated based on the top influential points, we find that the correlation estimates are consistently significant (Fig. 3-(a)). This is consistent with the results reported in (Koh & Liang, 2017). However, when the evaluation is done with the set of training samples at the 30th percentile of the influence score distribution, the correlation estimates decrease significantly (Fig. 3-(b)). This shows that influence estimates of only the top influential points are precise when compared to ground-truth re-trainings. Furthermore, without the weight-decay regularization, influence estimates in both cases are poor across all the test-points (Fig. 3-(a,b)).

To further understand the impact of weight-decay on influence estimates, we train the network with different weight-decay regularization factors. From Fig. 3-(c), we see that the selection of weight-decay factor is important in getting high-quality influence estimates. For this specific CNN architecture, we notice that the correlations start decreasing when the weight-decay factor is greater than 0.01. Moreover, from Fig. 3-(a,b), we find that the selection of test-point also has a strong impact on the quality of influence estimates. For example, when the network is trained with weight-decay and the influence estimates are computed for top influential training points, we notice that the Spearman correlation estimates range from 0.92 to 0.38 across different test-points and have a high variance.

These results show that despite some successful applications of influence functions in this non-convex setup, as reported in (Koh & Liang, 2017), their performances are very sensitive to hyper-parameters of the experiment as well as to the training procedure. In the next two sections, we assess the quality of influence estimates on more complex architectures and datasets including MNIST,

[2]The model has 2600 parameters and is trained for 500k iterations to reach convergence with the optimal model parameters θ^*. The ground-truth estimates are obtained by re-training the models from the optimal parameter set θ^* for 30k iterations. When trained with a weight-decay, a regularization factor of 0.001 is used.

Dataset	**MNIST**						**CIFAR-10**					
	A (With Decay)		**B (With Decay)**		**A (Without Decay)**		**A (With Decay)**		**B (With Decay)**		**A (Without Decay)**	
Architecture	P	S	P	S	P	S	P	S	P	S	P	S
Small CNN	**0.95**	**0.87**	**0.92**	**0.82**	**0.41**	0.35	-	-	-	-	-	-
LeNet	**0.83**	**0.51**	0.28	0.29	0.18	0.12	**0.81**	**0.69**	0.45	0.46	0.19	0.09
VGG13	0.34	**0.44**	0.29	0.18	0.38	0.31	**0.67**	**0.63**	**0.66**	**0.63**	**0.79**	**0.73**
VGG14	0.32	0.26	0.28	0.22	0.21	0.11	**0.61**	**0.59**	**0.49**	0.41	**0.75**	**0.64**
ResNet18	**0.49**	0.26	0.39	0.35	0.14	0.11	**0.64**	**0.42**	0.25	0.26	**0.72**	**0.69**
ResNet50	0.24	0.22	0.29	0.19	0.08	0.13	**0.46**	0.36	0.24	0.09	0.32	0.14

Table 1: Correlation estimates on MNIST And CIFAR-10 ; A=Test-point with highest loss; B=Test-point at the 50th percentile of test-loss spectrum; P=Pearson correlation; S=Spearman correlation

CIFAR-10 and ImageNet. In particular, we desire to understand, if the insights gained from experiments on smaller networks can be generalized to more complex networks and datasets.

5.3 Understanding Influence Functions in Deep Architectures

Setup: In this section, we evaluate the accuracy of influence estimates using MNIST and CIFAR-10 datasets across different network architectures including small CNN(Koh & Liang, 2017), LeNet (Lecun et al., 1998), ResNets (He et al., 2015), and VGGNets (Simonyan & Zisserman, 2015)[3]. To compute influence estimates, we choose two test points for each architecture: a) the test-point with the highest loss, and b) the test-point at the 50th percentile of the losses of all test points. For each of these two test points, we select the top 40 influential training samples and compute the correlation of their influence estimates with the ground-truth estimates. To compute the ground-truth influence estimates, we follow the strategy of (Koh & Liang, 2017), where we re-train the models from optimal parameters for 6% of the steps used for training the optimal model. When the networks are trained with a weight-decay regularization, we use a constant weight-decay factor of 0.001 across all the architectures (see Appendix for more details).

Results On MNIST: From Table 1, we observe that for the test-point with the highest loss, the influence estimates in the small CNN and LeNet architectures (trained with the weight-decay regularization) have high qualities. These networks have 2.6k and 44k parameters, respectively, and are relatively smaller and less deep than the other networks used in our experimental setup. As the depth of the network increases, we observe a consistent decrease in quality of influence estimates. For the test-point with a loss at the 50th percentile of test-point losses, we observe that influence estimates *only* in the small CNN architecture have good qualities.

Results On CIFAR-10: For CIFAR-10, across all architectures trained with the weight-decay regularization, we observe that the correlation estimates for the test-point with the highest loss are highly significant. For example, the correlation estimates are above 0.6 for a majority of the network architectures. However, for the test-point evaluated at the 50th percentile of the loss, the correlations decrease marginally across most of the architectures. We find that on CIFAR-10, even architectures trained without weight-decay regularization have highly significant correlation estimates when evaluated with the test-point which incurs the highest loss.

In case of MNIST, we have found that in shallow networks, the influence estimates are fairly accurate while for deeper networks, the quality of influence estimates decrease. For CIFAR-10, although the influence estimates are significant, we found that the correlations are marginally lower in deeper networks such as ResNet-50. The improved quality of influence estimates in CIFAR-10 can be attributed to the fact that for a similar depth, architectures trained on CIFAR-10 are less over-parameterized compared to architectures trained on MNIST. Note that, in Section 5.1, where the exact Hessian matrix can be computed, we observed that over-parameterization decreases the quality of influence estimates. From Table(1), we also observed that the selection of test-point has a significant impact on the quality of influence estimates. Furthermore, we noticed large variations in the quality of influence estimates across different architectures. In general we found that influence estimates for small CNN and LeNet are reasonably accurate, while for ResNet-50, the quality of estimates decrease across both MNIST and CIFAR-10. Precise reasons for these variations are difficult

[3]For CIFAR-10, evaluations on small CNN have not been performed due to the poor test accuracy.

to establish. We hypothesize that it can be due to the following factors: (i) Different architectures trained on different datasets have contrasting characteristics of loss landscapes at the optimal parameters which can have an impact on influence estimates. (ii) The weight-decay factor may need to be set differently in various architectures, to obtain high quality influence estimates.

5.4 Is Scaling Influence Estimates To ImageNet Possible?

The application of influence functions to ImageNet scale models provides an appealing yet challenging opportunity. It is appealing because, if successful, it opens a range of applications to large-scale image models, including interpretability, robustness, data poisoning, and uncertainty estimation. It is challenging for a number of reasons. Notable among these is the high computational cost of training and re-training, which limits the number of ground truth evaluations. In addition, all of the previously discussed difficulties in influence estimations still remain, including (i) non-convexity of the loss, (ii) selection of scaling and damping hyperparameters in the stochastic estimation of the Hessian, and (iii) the lack of convergence of the model parameters. The scale of ImageNet raises additional questions about the feasibility of leave-one-out retraining as the ground truth estimator. Given that there are 1.2M images in the training set, *is it even possible that the removal of one image can significantly alter the model?* In other words, we question whether or not reliable ground truth estimates may be obtained through leave-one-out re-training at this scale.

To illustrate this, we conduct an additional influence estimation on ImageNet. After training an initial model to 92.302% top5 test accuracy, we select two test points at random, calculate influence over the entire training set, and then select the top 50 points by their influences as candidates for re-training. We then use the re-training procedure suggested by (Koh & Liang, 2017), which starts leave-one-out re-training from the parameter set obtained after the initial training. We re-train for an additional 2 epochs, approximately 5% of the original training time, and calculate the correlations. We observe that for both test points, both Pearson and Spearman correlations are very low (less than 0.15, see details in the Appendix).

In our experiments, we observe high variability among ground-truth estimates obtained by re-training the model (see the appendix for details). We conjecture that this may be partially due to the fact that the original model has not be fully converged. To study this, we train the original model with *all* training points for an additional 2 epochs and measure the change in the test loss. We find that the overall top5 test accuracy has improved slightly to 92.336 % (+0.034) and the loss for one of the considered test points has decreased by relatively a significant amount of 0.679. However, the loss for the other point has increased slightly by 0.066. Such changes in loss values can therefore out-power the effect of leave-one-out re-training procedure. Second, we calculate the 2-norm of the weight gradients, which should be close to zero near an optimal point, and compare it to a standard pre-trained ImageNet ResNet-50 model as a baseline. We find these norms to be 20.18 and and 15.89, respectively, showing our model has similar weight gradient norm to the baseline. Although these norms are relatively small given that there are 25.5M parameters, further re-training the model still changes loss values for some samples significantly, making the ground-truth estimates noisy. We suggest that one way to obtain reliable ground-truth influence estimates in such large models can be through assessing the influence of a group of samples, rather than a single one.

6 Conclusion

In this paper, we present a comprehensive analysis of the successes and failures of influence functions in deep learning. Through our experiments on datasets including Iris, MNIST, CIFAR-10, ImageNet and architectures including LeNet, VGGNets, ResNets, we have demonstrated that influence functions in deep learning are fragile in general. We have shown that several factors such as the weight-decay, depth and width of the network, the network architecture, stochastic approximation and the selection of test points, all have strong effects in the quality of influence estimates. In general, we have observed that influence estimates are fairly accurate in shallow architectures such as small CNN(Koh & Liang, 2017) and LeNet, while in very deep and wide architectures such as ResNet-50, the estimates are often erroneous. Additionally, we have scaled up influence computations to the ImageNet scale, where we have observed influence estimates are highly imprecise. These results call for developing robust influence estimators in the non-convex setups of deep learning.

REFERENCES

Naman Agarwal, Brian Bullins, and Elad Hazan. Second order stochastic optimization in linear time. *ArXiv*, abs/1602.03943, 2016.

Anderson. Iris flower dataset. In -, 1936.

Samyadeep Basu, Xuchen You, and Soheil Feizi. Second-order group influence functions for black-box predictions. *ArXiv*, abs/1911.00418, 2019.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard S. Zemel. Understanding the origins of bias in word embeddings. *CoRR*, abs/1810.03611, 2018. URL `http://arxiv.org/abs/1810.03611`.

R. Dennis Cook and Sanford Weisberg. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22(4):495–508, 1980. ISSN 0040-1706. doi: 10.1080/00401706.1980.10486199.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Amirata Ghorbani, Abubakar Abid, and James Y. Zou. Interpretation of neural networks is fragile. In *AAAI*, 2017.

Ryan Giordano, Will Stephenson, Runjing Liu, Michael I. Jordan, and Tamara Broderick. A swiss army infinitesimal jackknife. In *AISTATS*, 2018.

Ryan Giordano, Michael I. Jordan, and Tamara Broderick. A higher-order swiss army infinitesimal jackknife. *ArXiv*, abs/1907.12116, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL `http://arxiv.org/abs/1512.03385`.

Wilhelm Kirch (ed.). *Pearson's Correlation Coefficient*, pp. 1090–1091. Springer Netherlands, Dordrecht, 2008. ISBN 978-1-4020-5614-7. doi: 10.1007/978-1-4020-5614-7_2569. URL `https://doi.org/10.1007/978-1-4020-5614-7_2569`.

P. W. Koh, J. Steinhardt, and P. Liang. Stronger data poisoning attacks break data sanitization defenses. *arXiv preprint arXiv:1811.00741*, 2019a.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL `http://proceedings.mlr.press/v70/koh17a.html`.

Pang Wei Koh, Kai-Siang Ang, Hubert H. K. Teo, and Percy Liang. On the accuracy of influence functions for measuring group effects. *CoRR*, abs/1905.13289, 2019b. URL `http://arxiv.org/abs/1905.13289`.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). *CVPR*, 2000. URL `cifar.com`.

Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pp. 2278–2324, 1998.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pp. 2278–2324, 1998. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.7665`.

Alexander Selvikvåg Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on MRI. *CoRR*, abs/1811.10052, 2018. URL `http://arxiv.org/abs/1811.10052`.

Barak A. Pearlmutter. Fast exact multiplication by the hessian. *Neural Comput.*, 6(1):147–160, January 1994. ISSN 0899-7667. doi: 10.1162/neco.1994.6.1.147. URL http://dx.doi.org/10.1162/neco.1994.6.1.147.

Garima Pruthi, Frederick Liu, Mukund Sundararajan, and Satyen Kale. Estimating training data influence by tracking gradient descent. *ArXiv*, abs/2002.08484, 2020.

Peter G. Schulam and Suchi Saria. Can you trust this prediction? auditing pointwise reliability after learning. In *AISTATS*, 2019.

Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1): 1–47, March 2002. ISSN 0360-0300. doi: 10.1145/505282.505283. URL http://doi.acm.org/10.1145/505282.505283.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:88–103, 1904.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6806.

Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer-Verlag, Berlin, Heidelberg, 1st edition, 2010. ISBN 1848829345, 9781848829343.

Chih-Kuan Yeh, Joon Sik Kim, Ian En-Hsu Yen, and Pradeep Ravikumar. Representer point selection for explaining deep neural networks. *CoRR*, abs/1811.09720, 2018. URL http://arxiv.org/abs/1811.09720.