BERT on a Data Diet: Finding Important Examples by Gradient-Based Pruning

Anonymous ACL submission

Abstract

Current pre-trained language models rely on large datasets for achieving state-of-the-art performance. However, past research has shown that not all examples in a dataset are equally important during training. In fact, it is sometimes possible to prune a considerable fraction of the training set while maintaining the test performance. Established on standard vision benchmarks, two gradient-based scoring metrics for finding important examples are *GraNd* and its estimated version, *EL2N*. In this work, we employ these two metrics for the first time in NLP. We demonstrate that these metrics need to be computed after at least one epoch of fine-tuning and they are not reliable in early steps. Furthermore, we show that by pruning a small portion of the examples with the highest GraNd/EL2N scores, we can not only preserve the test accuracy, but also surpass it. This paper details adjustments and implementation choices which enable *GraNd* and *EL2N* to be applied to NLP.

1 Introduction

Large datasets have made the phenomenal performance of Transformer-based (Vaswani et al., 2017) pre-trained language models (PLMs) possible (Devlin et al., 2019; Sanh et al., 2019; Liu et al., 2019; Clark et al., 2020; Bommasani et al., 2021). However, recent studies show that a significant fraction of examples in a training dataset can be omitted without sacrificing test accuracy. To this end, many metrics have been introduced for ranking the examples in a dataset based on their importance. One of these metrics is *Forgetting Score* (Toneva et al., 2019; Yaghoobzadeh et al., 2021) which recognizes the examples that are misclassified after being correctly classified during training or are always misclassifed. *Datamap* (Swayamdipta et al., 2020) is another technique for diagnosing datasets which uses *confidence* and *variability* metrics.

GraNd and its approximation, *EL2N*, are two recently introduced metrics that have only been studied for image classification models and tasks (Paul et al., 2021). The goal of this study is to adapt these metrics for PLMs and apply them in NLP tasks. We select a topic classification and a natural language inference dataset to run our experiments. We find that training with PLMs instead of randomly initialized models, used in Paul et al. (2021), brings new challenges. Besides, adapting these metrics to NLP is non-trivial because of the inherent differences between the visual and language modalities.

In summary, our contributions are threefold: (1) we adapt GraNd and EL2N metrics to the language domain to identify important examples in a dataset; (2) we show that in contrary to the results in computer vision, early score computation steps are not sufficient for finding a proper subset of the data in NLP; and (3) we observe that pruning a small fraction of the examples with the highest EL2N/GraNd scores will result in better performance and in some cases even better than fine-tuning on the whole dataset.

2 Background

In this section, we describe the two metrics introduced by Paul et al. (2021) for pruning the training data: **GraNd** and its estimated variant, **EL2N**.

2.1 GraNd

Consider $\mathbf{X} = \{x_i, y_i\}_{i=1}^{N}$ to be a training dataset for a given classification task with K classes, where x_i is the input (i.e. a sequence of tokens) and y_i is its corresponding label. To estimate the importance of each training sample (x_i, y_i), Paul et al. (2021) propose utilizing the expected value of the loss gradient norm denoted as GraNd:

$$\text{GraNd}(x_i, y_i) = \mathbb{E}_{\boldsymbol{w}} \left\| \boldsymbol{g}(x_i, y_i) \right\|_2 \qquad (1)$$

The vector $\boldsymbol{g}$ is the loss gradient with respect to the model's weights. This method is based on the assumption that the expected impact of a sample on

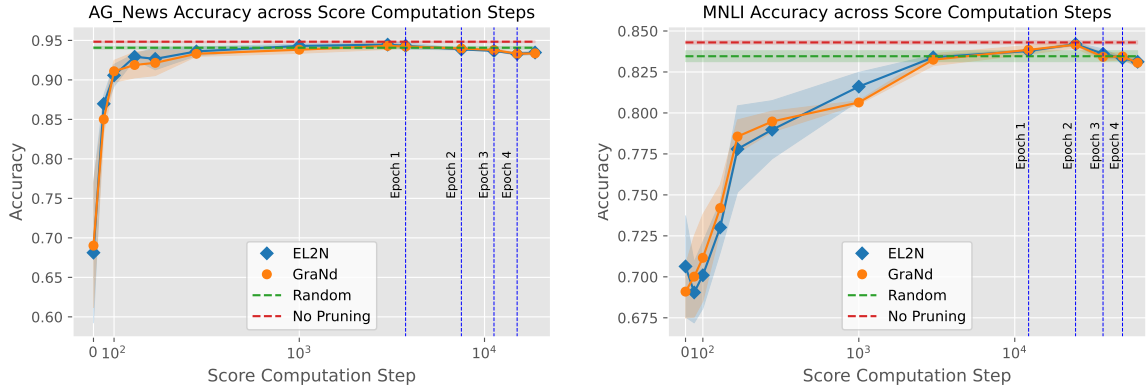

Figure 1: MNLI/AGNews accuracy of training BERT-base on the top 70%/50% of examples with the highest EL2N or GraNd scores computed at various steps of fine-tuning. Each point is the average of three runs and the standard deviation is shown as the shaded area. Early score computation steps are shown to result in lower accuracies.

the network weights ($\boldsymbol{w}$) denotes the importance of that sample. By sorting and ranking the training data, a top subset could be used for the training process, while pruning out the remaining data with lower scores.

Note that, unlike the original method stated by Paul et al. (2021), which initializes the entire network with random weights, we start our training with a pre-trained language model, a common practice in current NLP. Therefore, we compute the GraNd scores only for the randomly initialized classifier layer on top of the PLM.

2.2 EL2N

By defining $\psi^{(k)}(x_i) = \nabla_{\boldsymbol{w}} f^{(k)}(x_i)$ for the k^{th} logit, the loss gradient ($\boldsymbol{g}$) can be written as:

$$\boldsymbol{g}(x_i, y_i) = \sum_{k=1}^{K} \nabla_{f^{(k)}} \mathcal{L}(f(x_i), y_i)^T \psi^{(k)}(x_i) \quad (2)$$

Since the $\mathcal{L}(f(x_i), y_i)$ is a cross entropy loss, $\nabla_{f^{(k)}} \mathcal{L}(f(x_i), y_i)^T = p(x_i)^{(k)} - y_i^{(k)}$, where $p(x_i)$ is the output probability vector of the model. By assuming orthogonality and uniform sizes across logits for ψ over the training examples, the norm in Eq. 1 is approximately equal to $\|p(x_i) - \boldsymbol{y}_i\|_2$ ($\boldsymbol{y}_i$ is the one-hot vector of the true label). As a result, an estimate of GraND is **EL2N**, which is defined as follows:

$$\text{EL2N}(x_i, y_i) = \mathbb{E}_{\boldsymbol{w}} \|p(x_i) - \boldsymbol{y}_i\|_2 \quad (3)$$

It is worth noting that this formulation is similar to the confidence metric defined by Swayamdipta et al. (2020). When computing confidence, the expected value is obtained by averaging the model output probabilities across the training epochs. However, as stated in Eq. 3, the EL2N expectation is over multiple weight initializations ($\boldsymbol{w}$).

3 Experiments

In this section, we verify the effectiveness of GraNd and EL2N (Paul et al., 2021) in the NLP domain. Our experimental setup is similar to Paul et al. (2021) for the followings: (1) models are trained on the subsets of the data with higher GraNd and EL2N scores; (2) based on the expectation over multiple weight initializations mentioned in Eq. 1 and Eq. 3, we average the scores over five independent training runs.[1]; and, (3) for putting our results in context, we utilized random pruning as the baseline[2] and the accuracy of training on the whole dataset with no pruning as the target performance.

Two major differences between these two setups are mentioned here: (1) we used a pre-trained model, i.e., BERT (Devlin et al., 2019), standard in the NLP domain, whereas Paul et al. (2021) uses vision models initialized with random weights; and (2) as fine-tuning requires few epochs of training, we computed the metrics over fine-grained steps rather than epochs.

3.1 Datasets and Setup

Datasets. We evaluate our methods on two different classification tasks. For natural language inference, we used MNLI dataset (Williams et al., 2018), and for topic classification, we used AG's News

[1]It is ten independent training runs in Paul et al. (2021)
[2]Random selection with the same proportion

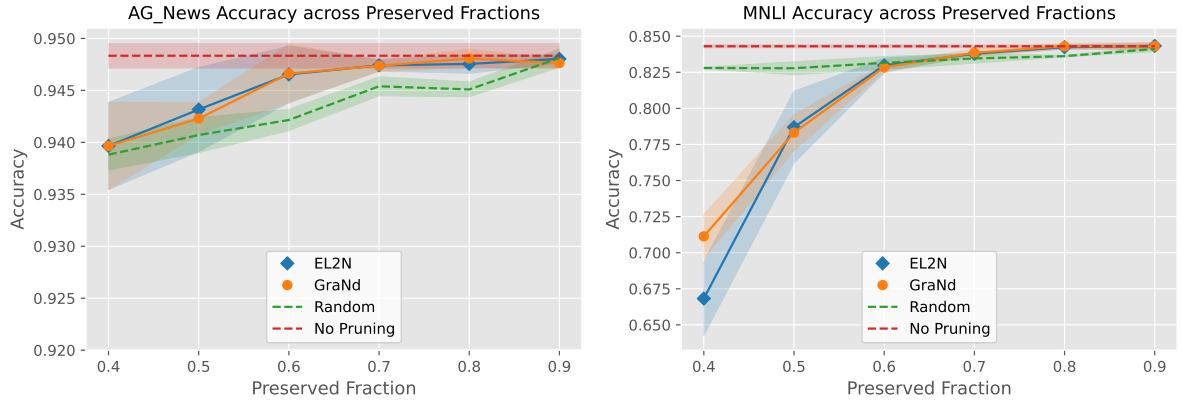

Figure 2: MNLI/AGNews accuracy of training BERT-base on the top k% of the examples with the highest EL2N/GraNd scores computed after one epoch of fine-tuning. Each point is the average of three runs and the standard deviation is shown as the shaded area. The performance grows as the preserved fraction increases.

(Zhang et al., 2015). We report the validation set (matched) and test set accuracy, respectively.

Setup. We used the Transformers library from HuggingFace (Wolf et al., 2020) and BERT-base-uncased as our pre-trained language model.[3] To fine-tune BERT, we trained the model for five epochs and selected the best performance, with 3e-5 as the learning rate. For calculating GraNd and EL2N scores, we used batch sizes of 12 and 32 respectively executed on a Tesla P100 GPU.

3.2 Results

Score computation step. Figure 1 demonstrates BERT's performance on MNLI and AG's News datasets after fine-tuning on the respective top 70% and 50% of examples, ranked by EL2N and GraNd scores. Most notably, early score computation steps are shown not to be reliable for obtaining a representative subset of the dataset to the extent that they may result in a lower performance than even random sampling. To further investigate this, Figure 3 shows the distribution of labels in the top examples chosen by EL2N across score computation steps. The distribution is extremely unbalanced in early steps which can explain the low performance of fine-tuning on high-scoring examples at the early stages. Moreover, the two datasets show different behaviours. AG's News seems to provide acceptable scores after only 500 steps, while MNLI takes longer to reach the random baseline. For further

[3]Our evaluations are based on the BERT's base version (12-layer, 768-hidden size, 12-attention head, 110M parameters) due to the massive number of experiments and resource limitations.

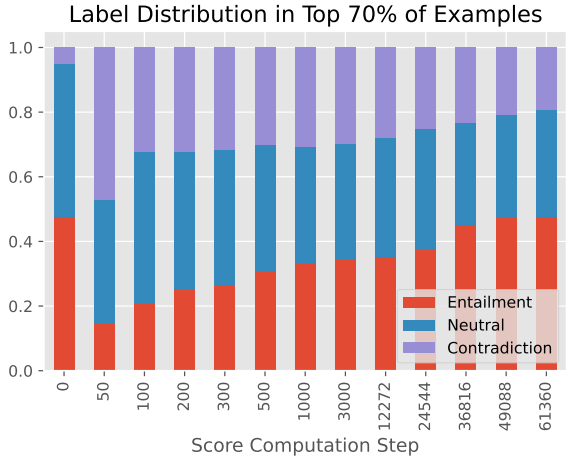

Figure 3: The distribution of labels of top-70% examples with highest EL2N scores in MNLI across score computation steps. The distribution is extremely unbalanced in early steps.

experiments we used the first epoch as the score computation step which is 3,750 and 12,272 for AG's News and MNLI, respectively.

Preserved fraction. To examine how the size of training set can affect the performance, we fine-tuned the model using different percentages of the entire dataset. Figure 2 shows the performance of BERT-base trained on the top k% of the data points with the highest EL2N/GraNd scores calculated after one epoch of fine-tuning. As shown in Figure 2, the models fine-tuned on smaller subsets of the dataset could not perform as well as the one fine-tuned on the whole dataset in both tasks and scoring metrics. Albeit, we see the growing performance as the preserved fraction increases.

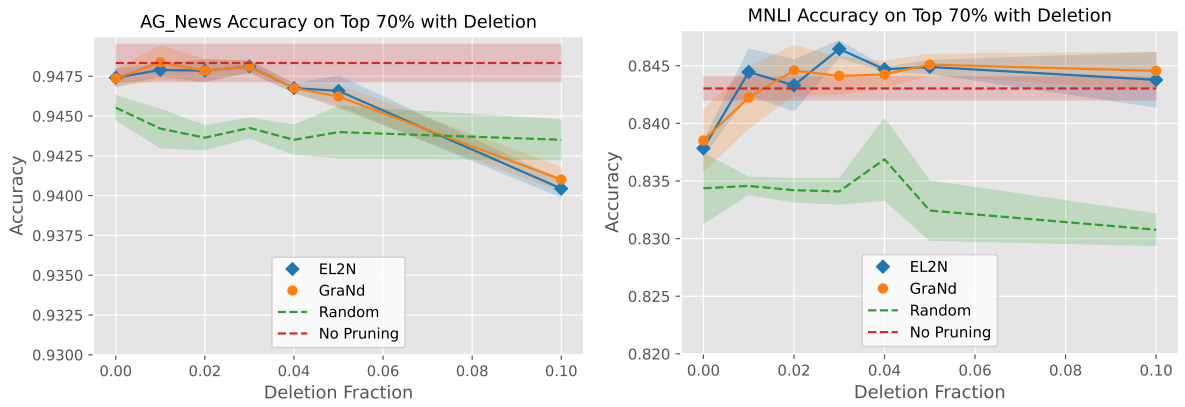

Figure 4: MNLI/AGNews accuracy of training BERT-base on the top 70% of examples with the highest EL2N/GraNd scores computed at first epoch of fine-tuning with top k% deleted. An increase in performance is observed in both datasets when deleting a small portion of highest scoring examples.

Interestingly, in MNLI, a model trained on the chosen subset performs worse than a random subset until 60% of data is preserved. This is contradicting with AG's News where even in small portions, EL2N and GraNd could outperform random baseline. Nevertheless, the EL2N and GraNd scores are shown to be better than random pruning when we preserve 70% or more of the dataset.

Noise examples. As discussed in recent research, datasets often include noisy examples, detection of which has been of interest (Swayamdipta et al., 2020; Jindal et al., 2019; Chen et al., 2022). We conducted an experiment to see whether removing the highest scoring examples in EL2N and GraNd, which might correspond to noisy examples (Paul et al., 2021), can improve the final performance. Figure 4 illustrates the performance of BERT-base after being trained on the top 70% of data and when the top k% of it is removed. An increase in performance can be seen in both datasets initially which can explain that deleting a conservative amount of the highest scoring examples can help the model to learn better. Remarkably, in the MNLI dataset we see that removing the top scoring examples achieves even higher accuracies than training on the whole dataset. This shows that, on the MNLI dataset, it is possible to reach higher performance when training on about 66% (70%-4%) than the whole 100%. On the AG's News, however, pruning does not result in performance gain, but comparable performance can be obtained with only 67% of the data.

Computation concerns. For computing EL2N scores, we employed the average across five seeds, and each seed is trained for one epoch. It is computationally equivalent to total of five epochs of training. However, we argue that it is possible to use fewer seeds to achieve scores which highly correlate to the mean of five seeds. Average spearman correlation of the scores between each of those seeds and the mean of five seeds is 0.9311 with standard deviation of 0.0042 which shows that using only one seed can yield very similar results to five seeds. The correlation increases to 0.9722 between average of two seeds and five seeds. This sheds light on the fact that even with limited resources, EL2N may find the important examples by only fine-tuning a few seeds, each for one epoch.

4 Conclusions

We adapted two data pruning metrics from computer vision, called EL2N and GraNd, to NLP. We showed that despite the major differences between the two fields, we can find subsets of data that maintain or, in some cases, even improve the performance. We demonstrated that unlike in vision, neither GraNd nor EL2N can yield acceptable performances in early steps of fine-tuning. In summary, at least one epoch of fine-tuning is necessary for a reliable computation of either of the two scores. Finally, we explained that, despite being dataset dependent, pruning the highest scoring examples, which may be related to noise, can achieve higher accuracies than when training on the whole dataset. A potential future work could be an end-to-end pruning mechanism with a single fine-tuning procedure.

References

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Derek Chen, Yu, and Zhou Samuel R. Bowman. 2022. Clean or annotate: How to spend a limited data collection budget. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 152–168, Hybrid. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ishan Jindal, Daniel Pressel, Brian Lester, and Matthew Nokleby. 2019. An effective label noise model for DNN text classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3246–3256, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar S. Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke S. Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. ArXiv:1907.11692.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems*, volume 34, pages 20596–20607. Curran Associates, Inc.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC^2 Workshop*.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. 2019. An empirical study of example forgetting during deep neural network learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yadollah Yaghoobzadeh, Soroush Mehri, Remi Tachet des Combes, T. J. Hazen, and Alessandro Sordoni. 2021. Increasing robustness to spurious correlations using forgettable examples. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3319–3332, Online. Association for Computational Linguistics.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*.