PLATON: Pruning Large Transformer Models with Upper Confidence Bound of Weight Importance

Qingru Zhang[1], Simiao Zuo[1], Chen Liang[1], Alexander Bukharin[1], Pengcheng He[2], Weizhu Chen[2], Tuo Zhao[1]

[1]Georgia Institute of Technology & [2]Microsoft Azure

qingru.zhang@gatech.edu

July 20, 2022

Background

Transformer Model: Computational Challenges

Challenges:

- Massive memory footprint (e.g. BERT, ViT, GPT-3)
- High inference latency
- Restricts their deployment on edge devices

Solution:

- Pruning: masking out redundant weights
- By ranking weights' importance score S

Pruning Transformer Model

Iterative Pruning:

$$\widetilde{\boldsymbol{\theta}}^{(t)} = \boldsymbol{\theta}^{(t)} - \alpha \nabla \mathcal{L}(\boldsymbol{\theta}^{(t)}),$$
$$\boldsymbol{\theta}^{(t+1)} = \mathcal{T}(\widetilde{\boldsymbol{\theta}}^{(t)}, S^{(t)}),$$

where

$$[\mathcal{T}(\widetilde{\boldsymbol{\theta}}, S)]_j = \begin{cases} \widetilde{\boldsymbol{\theta}}_j & \text{if } S_j \text{ is in the top } r^{(t)}\% \text{ of } S, \\ 0 & \text{otherwise.} \end{cases}$$

Remaining ratio $r^{(t)}$ follows a schedule. (Sanh et al., 2020; Han et al., 2015; Zhu and Gupta, 2018)

Importance Indicator

Sensitivity approximates the difference of the loss function when masking a parameter with 0.

$$I_j = |\boldsymbol{\theta}_{j,-j}^{\top} \nabla \mathcal{L}(\boldsymbol{\theta})| \approx |\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta} - \boldsymbol{\theta}_{j,-j})|$$

where $\boldsymbol{\theta}_{j,-j} = [0, \ldots, 0, \theta_j, 0, \ldots, 0] \in \mathbb{R}^d$

- A small sensitivity indicates that the weight is not very important.
- Applied in many prior works (Sanh et al., 2020; Liang et al., 2021)

Existing Challenges

Uncertainty of Importance Estimation:

- $I_j^{(t)}$ is computed based on a sampled mini batch of data.
- $I_j^{(t)}$ varies dramatically due to complicated training dynamics
- High variability of $I_j^{(t)} \Rightarrow$ cannot reflect contribution of $\boldsymbol{\theta}_j$.

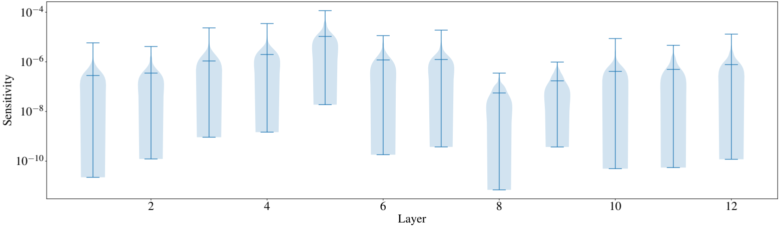


Figure: Violin plot of sampled weight over t

Our Method: PLATON

PLATON: Uncertainty Quantification

Sensitivity Smoothing:

$$\overline{I}_j^{(t)} = \beta_1 \overline{I}_j^{(t-1)} + (1-\beta_1) I_j^{(t)},$$

Uncertainty Quantification:

$$U_j^{(t)} = |I_j^{(t)} - \overline{I}_j^{(t)}|.$$

$$\overline{U}_j^{(t)} = \beta_2 \overline{U}_j^{(t-1)} + (1-\beta_2) U_j^{(t)}.$$

- A large $\overline{U}_j^{(t)}$ indicate $\overline{I}_j^{(t)}$ is not yet a reliable indicator.
- Retain this weight for further exploration.
- $\overline{U}_j^{(t)} \Rightarrow$ upper confidence bound of weight importance.

PLATON: Intuitions

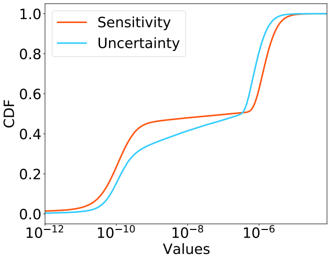


Figure: The CDF of sensitivity and uncertainty when pruning $\text{BERT}_{\text{base}}$ on RTE.

- $\overline{I}_j^{(t)}$ and $\overline{U}_j^{(t)}$ are highly skewed to zero.
  - Apply $\log()$ to distribute them more evenly.
- Define the importance score as

$$S_j^{(t)} = \exp(\log(\overline{I}_j^{(t)}) + \log(\overline{U}_j^{(t)})) \\ = \overline{I}_j^{(t)} \cdot \overline{U}_j^{(t)}$$

- Share the same sprint as UCB.
  - $\overline{I}_j^{(t)} \Rightarrow$ Exploitation on historical importance.
  - $\overline{U}_j^{(t)} \Rightarrow$ Exploration for the uncertain weights.

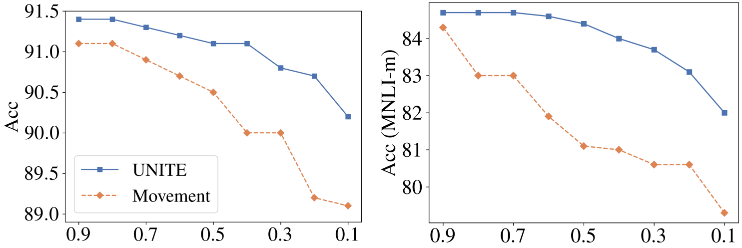
Experimental Results

GLUE Benchmark

Table: Results with $\text{BERT}_{\text{base}}$ on GLUE development set.

Ratio	Method	MNLI m / mm	RTE Acc	QNLI Acc	MRPC Acc / F1	QQP Acc / F1	SST-2 Acc	CoLA Mcc	STS-B P/S Corr
100%	$\text{BERT}_{\text{base}}$	84.6 / 83.4	69.3	91.3	86.4 / 90.3	91.5 / 88.5	92.7	58.3	90.2 / 89.7
20%	ℓ_0 Regularization	80.5 / 81.1	63.2	85.0	75.7 / 80.2	88.5 / 83.3	85.0	*N.A.*	82.8 / 84.7
	Magnitude	81.5 / 82.9	65.7	89.2	79.9 / 86.2	86.0 / 83.8	84.3	42.5	86.8 / 86.6
	Movement	80.6 / 80.8	*N.A.*	81.7	68.4 / 81.1	89.2 / 85.7	82.3	*N.A.*	*N.A.*
	Soft-Movement	81.6 / 82.1	62.8	88.3	80.9 / 86.7	90.6 / **87.5**	89.0	48.5	87.8 / 87.5
	PLATON	**83.1 / 83.4**	**68.6**	**90.1**	**85.5 / 89.8**	**90.7 / 87.5**	**91.3**	**54.5**	**89.0 / 88.5**
15%	ℓ_0 Regularization	79.1 / 79.8	62.5	84.0	74.8 / 79.8	87.9 / 82.3	82.8	*N.A.*	81.8 / 84.2
	Magnitude	80.1 / 80.7	64.6	88.0	69.6 / 79.4	83.6 / 79.2	82.8	*N.A.*	85.4 / 85.0
	Movement	80.1 / 80.3	*N.A.*	81.2	68.4 / 81.0	89.6 / 86.1	81.8	*N.A.*	*N.A.*
	Soft-Movement	81.2 / 81.7	60.2	87.2	81.1 / 87.0	90.4 / 87.1	88.4	40.8	86.9 / 86.6
	PLATON	**82.7 / 83.0**	**65.7**	**89.9**	**85.3 / 89.5**	**90.5 / 87.3**	**91.1**	**52.5**	**88.4 / 87.9**
10%	ℓ_0 Regularization	78.0 / 78.7	59.9	82.8	73.8 / 79.5	87.6 / 82.0	82.5	*N.A.*	82.7 / 83.9
	Magnitude	78.8 / 79.0	57.4	86.6	70.3 / 80.3	78.8 / 77.0	80.7	*N.A.*	83.4 / 83.3
	Movement	79.3 / 79.5	*N.A.*	79.2	68.4 / 81.2	89.1 / 85.4	80.2	*N.A.*	*N.A.*
	Soft-Movement	80.7 / 81.1	58.8	86.6	79.7 / 85.9	**90.2** / 86.7	87.4	*N.A.*	86.5 / 86.3
	PLATON	**82.0 / 82.2**	**65.3**	**88.9**	**84.3 / 88.8**	**90.2 / 86.8**	**90.5**	**44.3**	**87.4 / 87.1**

Experimental Results



(a) QQP **(b)** MNLI-m

Figure: Performance of pruning $\text{BERT}_{\text{base}}$ under different pruning ratio.

Summary

Summary

- Pruning methods suffer from high variability of importance scoring due to stochastic sampling and training dynamics.
- Sensitivity estimated on mini batches may not be an accurate indicator of weight importance.
- PLATON combines both sensitivity smoothing and uncertainty quantification to resolve such variability.
- Uncertainty quantification acts like upper confidence bound of importance estimation and explores weights for a longer time.
- Extensive experimental results demonstrate the effectiveness of PLATON.

References

Han, S., Pool, J., Tran, J. and Dally, W. J. (2015). Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada* (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett, eds.).

Liang, C., Zuo, S., Chen, M., Jiang, H., Liu, X., He, P., Zhao, T. and Chen, W. (2021). Super tickets in pre-trained language models: From model compression to improving generalization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Sanh, V., Wolf, T. and Rush, A. M. (2020). Movement pruning: Adaptive sparsity by fine-tuning.

Zhu, M. and Gupta, S. (2018). To prune, or not to prune: Exploring the efficacy of pruning for model compression.

Thank You!