RABANT: Lottery Ticket Compression of Genomic Transformers for Clinically Oriented Variant Pathogenicity Classification

Anirudha M, Ashwin Acharya, Rajashree Shettar and K. Badari Nath

Department of Computer Science and Engineering, R V College of Engineering, Bengaluru, 560059, Karnataka, India.

Supplementary Information

This file contains supplementary methods, notes and tables for the article above. Supplementary Tables S1–S5 report quantities that underpin the pruned-weight analysis, the mask construction and the Dynamic Rank Fusion (DRF) check. All numerical values are either computed directly from the checkpoints and tables described in the article or transcribed from it; no additional experiments are reported.

Supplementary Methods: parameter accounting and prunable set

Supplementary Table S1 lists every parameter group of the dual-sequence classifier (NT-v2-100M backbone with the classification head) and how it was treated by pruning. The prunable set $\mathcal{P}$ comprises the two-dimensional weight matrices of the 22 encoder layers (query, key, value, attention output, and the two feed-forward matrices of each layer; 132 matrices) and the two linear layers of the classification head, giving $N_{\mathcal{P}} = 92{,}668{,}416$ weights. The pre-trained checkpoint also contains a 352-weight contact-prediction projection that the classifier does not use. It was masked together with the other two-dimensional matrices but is not counted in $\mathcal{P}$; its contribution is negligible ($< 0.0004\%$ of all parameters).

At 50% sparsity, 46,334,384 weights are removed, which is 48.3% of all $N = 95{,}915{,}971$ parameters. The RABANT checkpoints therefore contain 49,581,074 non-zero parameters; this count also reflects 513 parameters that are exactly zero in the dense checkpoint, namely the padding-position embedding row and one bias. The figure of 48 M active parameters used to name the model in the article is the nominal value $(1 - s)N$ with $N \approx 96$ M. The difference from the exact count arises because 3,247,203 parameters (3.4%), namely the embeddings, layer-normalisation parameters, biases and rotary-frequency buffers, are never pruned.

Supplementary Methods: mask construction and verification

For every criterion the order-statistic threshold was applied separately to each masked weight matrix, so that exactly 50% of the weights of every matrix are removed (the sparsity s of the article is therefore a per-matrix sparsity). This was verified directly on the 50%-sparsity checkpoints of all four criteria: each of the 135 masked matrices contains exactly half zeros. As a further check, the magnitude mask was compared with thresholding the absolute weights of the dense fine-tuned checkpoint (Supplementary Table S2). The magnitude checkpoint reproduces the per-matrix top-50% selection for 99.996% of weights on average, whereas a single global threshold over all matrices would reproduce only 92.2% on average. Under a global threshold, the fraction of weights retained would range from 11.6% to 77.5% across matrices, which is not observed in any checkpoint.

Supplementary Table S1. Parameter accounting of the dual-sequence classifier. $N = 95{,}915{,}971$ parameters in total; all tensors are stored in 32-bit floating point.

Parameter group	Count	Treatment
Encoder weight matrices (132 matrices: Q, K, V, attention output, two feed-forward)	92,274,688	pruned
Classification-head linear weights (2 matrices: $1536 \to 256$, $256 \to 2$)	393,728	pruned
Prunable set $\mathcal{P}$	**92,668,416**	
Contact-prediction projection weights (unused by the classifier)	352	masked, not in $\mathcal{P}$
Word embeddings ($4{,}107 \times 512$)	2,102,784	retained
Position embeddings ($2{,}050 \times 512$)	1,049,600	retained
Layer-normalisation parameters	49,152	retained
Biases (attention, feed-forward, head; contact-projection bias)	45,315	retained
Rotary inverse-frequency buffers (non-trainable)	352	retained
Total	95,915,971	

Supplementary Table S2. Agreement between the magnitude checkpoint and thresholding of the dense fine-tuned weights, computed over the 135 masked matrices. Agreement is the fraction of weights in a matrix that are classified identically (retained or removed).

Reference selection rule applied to the dense weights	Mean agreement (%)	Minimum over matrices (%)
Top 50% of $\lvert\theta_i\rvert$ within each matrix	99.996	99.72
Top 50% of $\lvert\theta_i\rvert$ over all masked matrices (single global threshold)	92.19	61.65

Supplementary Methods: pruned-weight diagnostics

Supplementary Tables S3 and S4 give the exact values behind the pruned-weight analysis (Fig. 4 of the main text). All quantities are computed over $\mathcal{P}$ (92,668,416 weights) using the dense fine-tuned weights θ_i. The pruned energy fraction is $E(M)/\|\boldsymbol{\theta}_{\mathcal{P}}\|_2^2$ with $E(M) = \sum_{i \in \mathcal{Z}} \theta_i^2$. The largest 10% of weights are defined within each matrix by $|\theta_i|$. The Jaccard index of two retained sets is $|M \cap M'|/|M \cup M'|$; for two independent random masks that each retain 50% of the weights its expected value is 1/3.

Supplementary Table S3. Weight mass removed by each importance criterion at 50% sparsity (dense fine-tuned model). Random pruning at 50% sparsity removes 50% of the weight energy and 50% of the largest 10% of weights in expectation.

Criterion	Weights removed (%)	Energy removed (%)	Largest 10% removed (%)	Energy relative to magnitude
Absolute magnitude	50.0	7.1	0.0	1.0×
Diagonal Fisher	50.0	50.8	51.0	7.2×
EMA gradient	50.0	51.1	51.8	7.2×
Weight movement	50.0	47.3	46.5	6.7×

Supplementary Table S4. Jaccard index between the retained sets of the four criteria at 50% sparsity, computed over $\mathcal{P}$. The chance level for two independent masks is 1/3.

	Magnitude	Fisher	EMA gradient	Movement
Magnitude	1.00	0.33	0.33	0.34
Fisher	0.33	1.00	0.55	0.31
EMA gradient	0.33	0.55	1.00	0.32
Movement	0.34	0.31	0.32	1.00

Supplementary Methods: Dynamic Rank Fusion weights

Supplementary Table S5 gives the DRF weights $\alpha_k(s)$ for all five sparsity levels, computed from the validation AUROC values of the exploratory phase with the blending and temperature settings stated in the Methods (0.6/0.4 blend, $T = 2.0$ AUROC percentage points). The values at 50% and 60% sparsity (0.902 and 0.987 for magnitude) are those quoted in the Results. Because the weights are a deterministic function of the validation AUROC values, they restate the exploratory results on a common scale and do not constitute independent evidence.

Supplementary Table S5. DRF weights $\alpha_k(s)$ by criterion and sparsity level (each column sums to one, up to rounding).

Criterion	10%	20%	30%	50%	60%
Absolute magnitude	0.609	0.698	0.713	0.902	0.987
Diagonal Fisher	0.314	0.286	0.285	0.098	0.012
EMA gradient	0.044	0.011	0.001	0.000	0.000
Weight movement	0.034	0.004	0.001	0.000	0.001

Supplementary Note: Proposition 1 under per-matrix removal counts

The masks used in all experiments remove a fixed number $K^{(m)}$ of weights from each prunable matrix $\mathcal{P}_m$, rather than a single global number $K = \sum_m K^{(m)}$. The minimax result of the article continues to hold under this constraint, as follows.

Claim. Let $\mathcal{M}_{K^{(1)},\dots,K^{(M)}}$ be the set of masks with $|\mathcal{Z} \cap \mathcal{P}_m| = K^{(m)}$ for every matrix m. Then, for constants $G, \Lambda \geq 0$ with $G + \Lambda > 0$, the mask in $\mathcal{M}_{K^{(1)},\dots,K^{(M)}}$ that minimises the worst-case second-order loss increment $G\sqrt{E(M)} + \frac{1}{2}\Lambda E(M)$ removes, in each matrix, the $K^{(m)}$ weights of smallest $|\theta_i|$. Moreover, every mask in this set satisfies $E(M) \geq E(M^{\mathrm{mag}})$, where M^{mag} is the per-matrix magnitude mask.

Proof. The worst-case expression is derived exactly as in the main text and is a strictly increasing function of $E(M)$, so minimising it over $\mathcal{M}_{K^{(1)},\dots,K^{(M)}}$ is equivalent to minimising $E(M) = \sum_m \sum_{i \in \mathcal{Z} \cap \mathcal{P}_m} \theta_i^2$. This sum separates across matrices, and the removal counts are fixed, so each inner sum can be minimised independently. A sum of $K^{(m)}$ squares drawn without replacement from $\{\theta_i^2\}_{i \in \mathcal{P}_m}$ is smallest when the $K^{(m)}$ smallest squares are chosen, that is, when $\mathcal{Z} \cap \mathcal{P}_m$ holds the $K^{(m)}$ smallest values of $|\theta_i|$. The final statement is the same minimisation. □

Remark. The guarantee is relative to masks with the same per-matrix counts, which is the setting in which all four criteria are compared. The per-matrix magnitude mask does not in

general coincide with the set of the K globally smallest weights, and its energy is therefore at least that of the global magnitude mask.

Supplementary Methods: statistical reporting

Confirmatory comparisons. For each metric, the two models being compared were represented by five fold-level estimates (one per cross-validation fold; $n = 5$ per model), each a macro-average over the three unseen test sets (or the value on a single source for the per-source comparisons). The comparisons of interest are RABANT against the dense (uncompressed) model and RABANT against the native NT-v2-50M model. AUROC is the primary endpoint; accuracy, precision and specificity at the decision threshold $c = 0.5$ are secondary endpoints. Each comparison used Welch's unequal-variance t-test (two-sided, significance level 0.05), with Welch–Satterthwaite degrees of freedom, a 95% confidence interval for the difference of means and Cohen's d with the pooled standard deviation. Exact p-values are reported in the main-text tables. The symbol $\pm$ always denotes the standard deviation over the five folds.

Justification and limits. Welch's test was chosen because the two groups are small and their variances differ markedly for several metrics (for example the specificity standard deviations range from 2.74 to 4.83 percentage points). With $n = 5$ the normality of fold-level estimates cannot be meaningfully assessed. The five folds share training data and the same pre-trained checkpoint, so fold-level estimates are not strictly independent; the reported p-values and intervals should therefore be regarded as approximate. The variability reported reflects training variability across folds and seeds on fixed test sets, and does not include sampling variability of the test sets themselves.

Multiple comparisons. The p-values reported for the secondary endpoints and for the per-source comparisons are unadjusted for multiple comparisons. AUROC was designated as the primary endpoint, and conclusions that rest on the threshold-dependent metrics should be read as exploratory. Readers who wish to apply a correction can do so from the exact p-values given in the main text.

Exploratory phase. The values in the exploratory-phase tables are single runs on one fixed validation split and carry no statistical testing. Differences of a few tenths of a percentage point between single runs (for example between neighbouring sparsity levels of magnitude pruning) are of the same size as the fold-to-fold standard deviation of macro-averaged AUROC in the confirmatory phase (0.33–0.35 percentage points) and should not be interpreted.