Compact Language Models via Pruning & Knowledge Distillation

Saurav Muralidharan*, Sharath Turuvekere Sreenivas*, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, Pavlo Molchanov

Introduction

Training LLM Model Families

- Model providers often train a family of LLMs, where each model targets a specific deployment scale/size

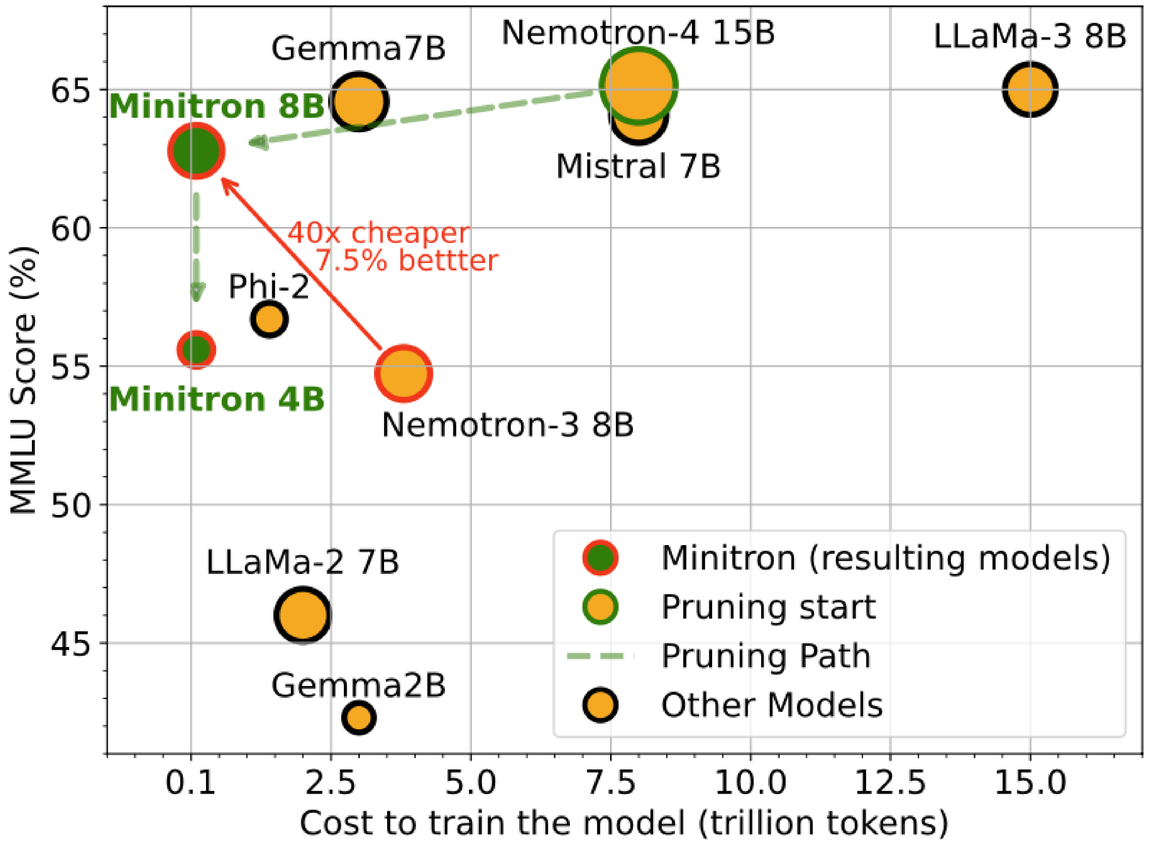
7B, 8x7B, 8x22B, Small, Medium, Large

Meta-3.1 8B, 70B, 405B

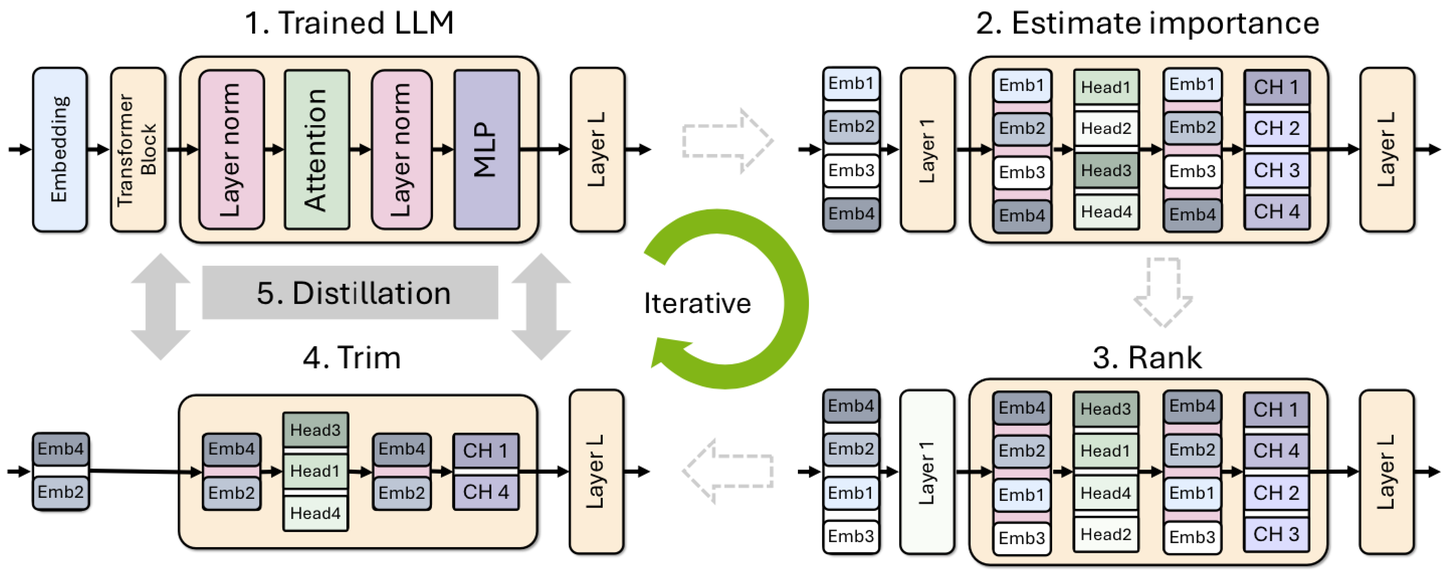
- Each model in the family is trained from scratch – expensive in compute, data, memory, etc.

"Can we train one big model, and obtain smaller, more accurate models from it through a combination of weight pruning and retraining, while only using a small fraction of the original training data?"

Minitron Performance Preview



System Overview



Importance Estimation & Ranking

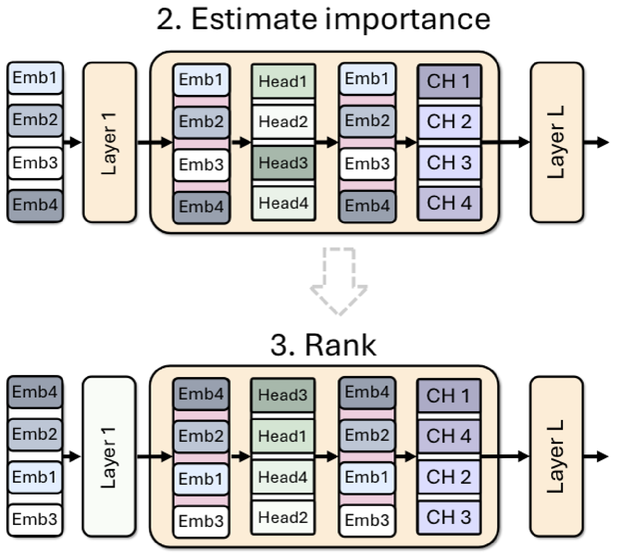
- Activation-based importance of each neuron, head, layer, and embedding channel
- Pass a small calibration dataset (1024 samples) through the network, and obtain rankings for all axes

$$F_{\text{head}}^{(i)} = \sum_{\mathbf{B},\mathbf{S}} \| \text{Attn}(\mathbf{X}\boldsymbol{W}^{Q,i}, \mathbf{X}\boldsymbol{W}^{K,i}, \mathbf{X}\boldsymbol{W}^{V,i}) \|_2$$

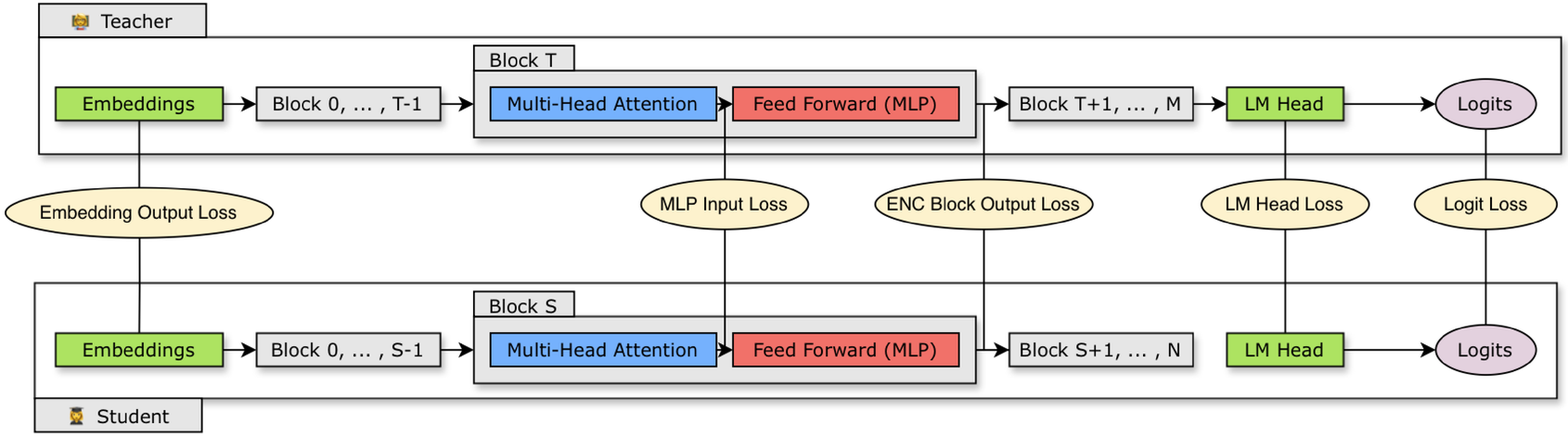
$$F_{\text{neuron}}^{(i)} = \sum_{\mathbf{B},\mathbf{S}} \mathbf{X}(\boldsymbol{W}_1^i)^T$$

$$F_{\text{emb}}^{(i)} = \sum_{\mathbf{B},\mathbf{S}} LN(\mathbf{X})_i$$

$$\text{BI}_i = 1 - \mathbb{E}_{X,t} \frac{\mathbf{X}_{i,t}^T \mathbf{X}_{i+1,t}}{\|\mathbf{X}_{i,t}\|_2 \|\mathbf{X}_{i+1,t}\|_2}$$



Retraining with Distillation



$$L = L_{\text{CLM}} + L_{logits} + \alpha \times L_{is}$$

$$L_{\text{logits}} = \frac{1}{l} \sum_{k=1}^{l} \text{Loss}(p_t^k(x,\tau), p_s^k(x,\tau))$$

$$L_{is} = \frac{1}{l} \sum_{k \in H} \sum_{i=1}^{l} Loss_k(h_t^{ki}, h_s^{ki})$$

Structured Compression Best Practices

1. To train a family of LLMs, train the largest one and prune+distill iteratively to smaller LLMs.
2. Use `(batch=L2, seq=mean)` importance estimation for width axes and PPL/BI for depth.
3. Use single-shot importance estimation; iterative provides no benefit.
4. Prefer width pruning over depth for the model scales we consider ($\leq$ 15B).
5. Retrain exclusively with distillation loss using KLD instead of conventional training.
6. Use (logit+intermediate state+embedding) distillation when depth is reduced significantly.
7. Use logit-only distillation when depth isn't reduced significantly.
8. Prune a model closest to the target size.
9. Perform lightweight retraining to stabilize the rankings of searched pruned candidates.
10. If the largest model is trained using a multi-phase training strategy, it is best to prune and retrain the model obtained from the final stage of training.

Minitron 8B and 4B Accuracy

			Models						
	Benchmark	**Metric**	**Llama-3**	**Llama-2**	**Mistral**	**Gemma**	**Nemotron-4**	**Nemotron-3**	**MINITRON**
	# Parameters		8B	6.7B	7.3B	8.5B	15.6B	8.5B	8.3B
	# Non-Emb. Params		5.9B	6.4B	7B	7.7B	12.5B	6.4B	6.2B
	# Training Tokens		>15T	2T	8T	6T	8T	3.8T	**94B**
KnowledgeLogic	winogrande (5)	acc	77.6	74	78.5	78	83.6	75.9	**79.0**
	arc_challenge (25)	acc_norm	57.8	53	60.3	**61**	58.8	52.8	52.6
	MMLU(5)	acc	**65.3**	46	64.1	64	66.6	54.7	63.8
	hellaswag(10)	acc_norm	82.1	79	**83.2**	82	84.6	78.5	80.7
	gsm8k(5)	acc	50.3	14	37	50	48.5	24.0	**51.3**
	truthfulqa(0)	mc2	43.9	39	42.6	**45**	40.7	36.5	42.6
	XLSum en (20)(3)	rougeL	30.9	31	4.80	17	32	30.9	**31.2**
Coding	MBPP(0)	pass@1	**42.4**	20	38.8	39	38	27.04	35.2
	humaneval (n=20)(0)	pass@1	28.1	12	28.7	**32**	35.4	20.7	31.6

			Models					
	Benchmark	**Metric**	**Phi-2**	**Gemma**	**Gemma2***	**Qwen2***	**MiniCPM***	**MINITRON**
# Parameters			2.7B	2.5B	2.6B	1.5B	2.7B	4.2B
# Non-Emb. Params			2.5B	2B	2B	1.3B	2.4B	2.6B
# Training Tokens			1.4T	3T	2T	7T	1.1T	**94B**
Knowledge, Logic	winogrande (5)	acc	**74**	67	70.9	66.2	-	**74.0**
	arc_challenge (25)	acc_norm	**61**	48	55.4	43.9	-	50.9
	MMLU(5)	acc	57.5	42	51.3	56.5	53.5	**58.6**
	hellaswag(10)	acc_norm	**75.2**	72	73.0	66.6	68.3	75.0
	gsm8k(5)	acc	55	18	23.9	**58.5**	53.8	24.1
	truthfulqa(0)	mc2	44	33	-	**45.9**	-	42.9
	XLSum en (20)(3)	rougeL	1	11	-	-	-	**29.5**
Coding	MBPP(0)	pass@1	**47**	29	29.6	37.4	-	28.2
	humaneval (n=20)(0)	pass@1	**50**	24	17.7	31.1	-	23.3

Minitron vs. Other Compressed Models

			Models					
	Benchmark	**Metric**	**LLMPruner**	**SliceGPT**	**LaCo**	**ShortGPT**	**Sheared LLaMa**	**MINITRON**
8 Billion	# Parameters		9.8B	9.9B	9.8B	9.8B	-	8.3B
	# Non-Emb. Params		9.5B	9.5B	9.5B	9.5B	-	6.2B
	MMLU(5)	acc	25.2	37.1	45.9	54.7	-	**62.8**
	hellaswag(10)	acc_norm	67.8	55.7	64.4	66.6	-	**79.7**
4 Billion	# Parameters		4.8B	4.9B	4.9B	4.9B	2.7B	4.2B
	# Non-Emb. Params		4.5B	4.6B	4.6B	4.6B	2.5B	2.6B
	winogrande (5)	acc	-	-	-	-	64.2	**73.84**
	arc_challenge (25)	acc_norm	-	-	-	-	41.2	**44.97**
	MMLU(5)	acc	23.33	28.92	26.45	43.96	26.4	**55.59**
	hellaswag(10)	acc_norm	56.46	50.27	55.69	53.02	70.8	**73.13**

Minitron Resources

Poster Session: Fri 13 Dec 11 a.m. — 2 p.m. PST

NeurIPS Poster Page

Minitron Website

HuggingFace Models


NVIDIA