RepSpec: Structural Re-parameterized Draft Model Training for Speculative Decoding

Feiye Huo*, Jianchao Tan*, Jiahao Liu, Zixu Jiang, Jiacheng Li, Jingang Wang, Xunliang Cai, Shengli Sun

ICLR International Conference On Learning Representations

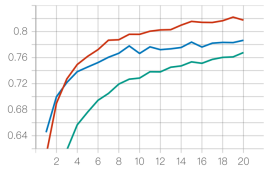
Abstract

As LLMs grow in size, autoregressive inference latency rises. Speculative decoding mitigates this by having a large target model verify the draft tokens, but the inference latency gain is limited by the draft model's capacity (parameter gap).

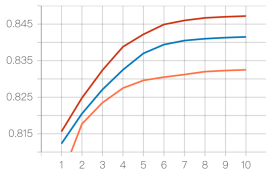
We propose **RepSpec**: adding redundant linear structures during training (merged at inference) to boost draft model effectiveness without extra inference cost. It improves accepted length when applied to EAGLE1/3; Furthermore, a hybrid linear and nonlinear strategy yields further gains.

Training Result

To show RepSpec's enhancement of draft model training, we present step-1 accuracy across training epochs for EAGLE-1 and EAGLE-3. Green (EAGLE-1 baseline), Orange (EAGLE-3 baseline), Blue (Linear method), Red (Hybrid method).



(1). EAGLE-1 w/o RepSpec



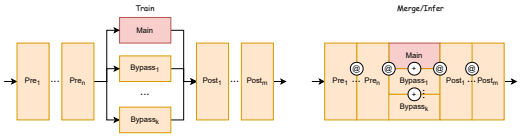
(2). EAGLE-3 w/o RepSpec

RepSpec

1. Pure Linear Method

The simplest method to apply the structural re-parameterization to draft model training is to add mergeable redundant branches to all eligible linear blocks. For the compact Transformer decoders, these include the embedding layer, self-attention projection layers, and MLP linear layers. Empirically, re-parameterizing all linear layers in attention and MLP is sufficient. Specifically, we temporarily add training-only linear layers (Pre-before, Bypass-side, Post-after) to an existing linear layer to construct the reparameterized structure *Aug*.

$$Aug = Post_{m:1} \circ (Main + \sum_{l=1}^{k} Bypass) \circ Pre_{1:n}$$



2. Hybrid Method

While we aim for non-redundant inference overhead of draft model, speculative decoding prioritizes end-to-end speedup, therefore we can afford moderate extra draft model computation if offset by acceptance length gains. Pure linear method has a bounded optimization due to fixed parameter space, so we apply simple non-linear decompositions to some of the linear structures. Although these cannot be fully merged during inference, they raise model capacity upper-bounds. The hybrid approach even outperforms simply adding draft model layers.

Inference Effectiveness

Inference performance of different methods across different models and benchmarks. Here, τ denotes the acceptance length, v denotes the inference speed.

Target	Strategy	Draft	Method	MT τ	MT v	GSM8k τ	GSM8k v	Alpaca τ	Alpaca v	Human τ	Human v	QA τ	QA v	Sum τ	Sum v	Avg τ	Avg v
L31 8B	Chain	E1	Baseline	2.54	48	2.74	52	2.41	46	2.99	58	2.28	42	2.28	42	2.54	48
			Linear	2.67	**52**	2.93	**56**	2.54	**48**	3.22	**63**	2.38	**44**	2.45	**46**	2.70	**52**
			Hybrid	**2.79**	52	**3.00**	53	**2.68**	47	**3.39**	60	**2.48**	45	**2.56**	45	**2.82**	50
		E3	Baseline	3.33	65	3.32	65	3.44	68	3.99	74	2.96	57	3.03	58	3.35	64
			Linear	3.58	**68**	3.60	**69**	3.75	**74**	4.25	**78**	3.08	**59**	3.30	**61**	3.60	**68**
			Hybrid	**3.62**	65	**3.65**	66	**3.78**	70	**4.32**	75	**3.15**	55	**3.42**	59	**3.66**	65
	Tree	E1	Baseline	3.91	70	4.17	77	3.83	73	4.52	82	3.46	62	3.25	57	3.86	70
			Linear	4.16	**74**	4.42	**81**	4.11	**75**	4.82	**88**	3.65	**67**	3.46	**62**	4.10	**75**
			Hybrid	**4.34**	72	**4.62**	79	**4.28**	71	**5.05**	83	**3.87**	65	**3.68**	60	**4.31**	72
		E3	Baseline	5.94	100	5.86	95	6.34	110	6.51	109	5.13	82	5.37	89	5.86	98
			Linear	6.00	**102**	5.95	**99**	6.48	**115**	6.60	**114**	5.25	**87**	5.41	**91**	5.95	**101**
			Hybrid	**6.08**	98	**6.01**	95	**6.61**	114	**6.65**	110	**5.32**	85	**5.55**	90	**6.03**	99
L2 13B	Chain	E1	Baseline	2.87	51	3.01	56	2.75	51	3.41	62	2.51	48	2.65	46	2.87	52
			Linear	2.97	54	3.09	58	2.84	54	3.56	64	2.60	50	2.73	49	2.97	55
			Hybrid	**3.09**	**56**	**3.20**	**60**	**2.96**	**55**	**3.78**	**67**	**2.73**	**53**	**2.93**	**51**	**3.12**	**57**
	Tree	E1	Baseline	4.38	76	4.64	80	4.32	78	5.09	89	3.94	70	3.48	63	4.31	76
			Linear	4.54	79	4.75	81	4.47	80	5.21	92	4.10	72	3.77	66	4.47	78
			Hybrid	**4.71**	**82**	**4.95**	**84**	**4.65**	**82**	**5.42**	**93**	**4.25**	**75**	**4.02**	**66**	**4.67**	**80**
V 7B	Tree	M	Baseline	2.38	44	2.66	48	2.40	45	2.75	50	2.12	37	2.14	38	2.41	44
			Linear	2.55	**47**	2.72	**51**	2.45	**47**	2.88	**53**	2.25	**40**	2.26	**40**	2.52	**46**
			Hybrid	**2.62**	44	**2.89**	49	**2.51**	43	**2.92**	51	**2.38**	37	**2.34**	37	**2.61**	44
		H	Baseline	3.66	60	3.62	53	3.50	47	3.88	58	2.91	40	2.85	39	3.59	49
			Linear	3.89	**64**	3.92	**56**	3.86	**53**	4.03	**60**	2.96	**42**	2.94	**43**	3.60	**53**
			Hybrid	**4.05**	63	**3.99**	54	**4.02**	51	**4.12**	59	**3.02**	38	**2.99**	39	**3.70**	51

Conclusion

- RepSpec is a structural re-parameterization framework designed to enhance draft models during training for speculative decoding.

- RepSpec inserts mergeable linear branches during training to boost model capacity without inference overhead; its hybrid variant achieves better end to end inference performance.

- RepSpec modifies substructures only, no change to overall model design pattern, and effectively enhances parameter-constrained draft models.