ICML

International Conference
On Machine Learning

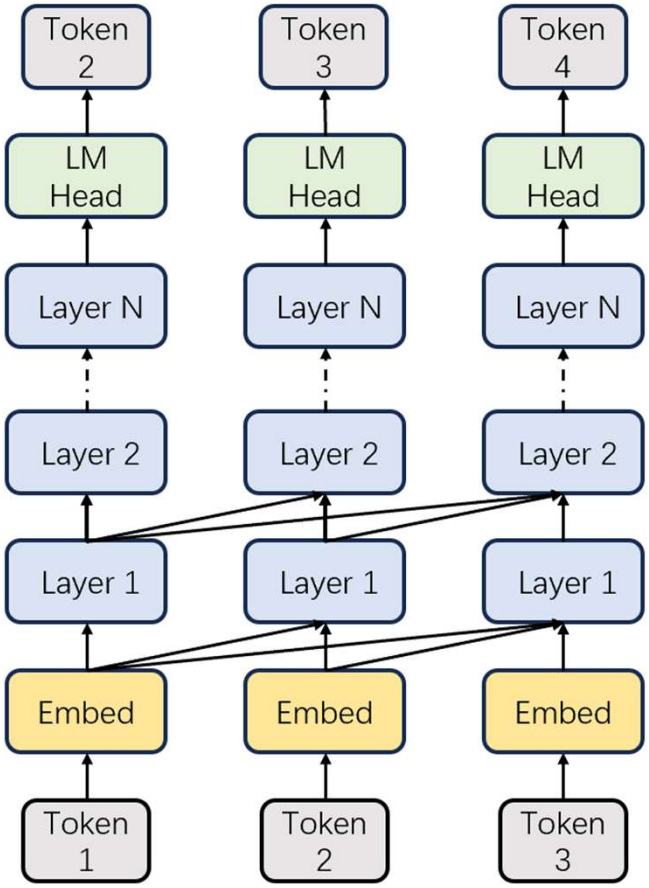
Turbo Connection: Reasoning as Information Flow from Higher to Lower Layers

Mohan Tang, Sidi Lu

Background

In the standard Transformer, a layer at a specific token can only read from states at earlier tokens at lower layers.

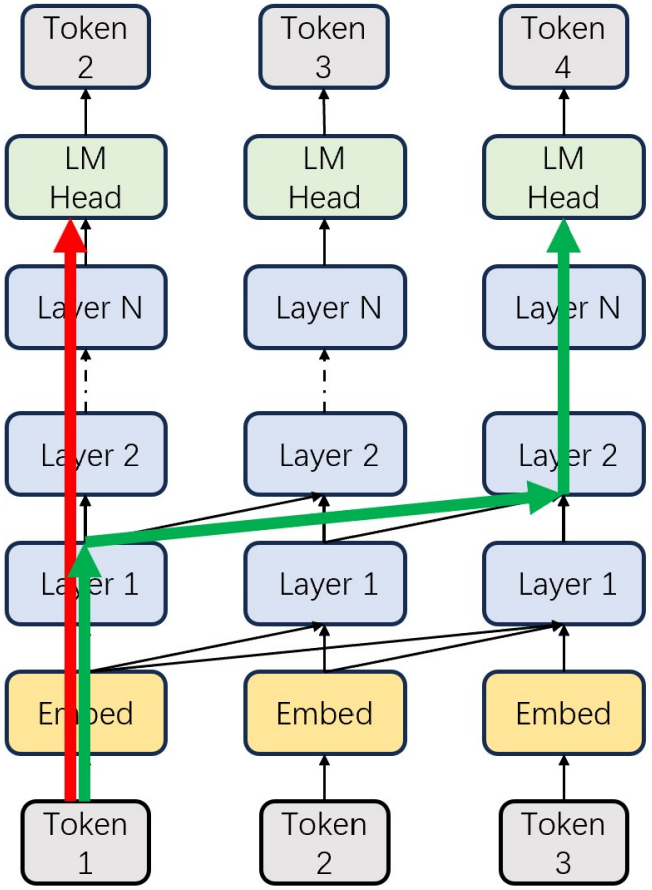
Therefore, when following the path of information flow, the number of steps along the path is always bounded by a fixed number.



Background

In the standard Transformer, a layer at a specific token can only read from states at earlier tokens at lower layers.

Therefore, when following the path of information flow, the number of steps along the path is always bounded by a fixed value.

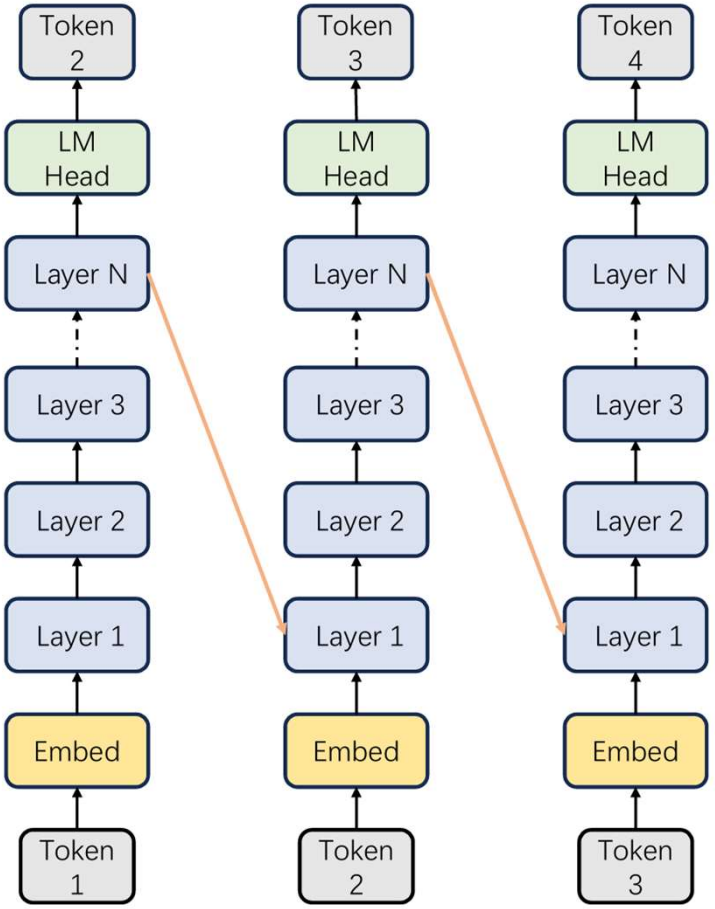


Our Method

We introduce a novel architecture called TurboConn to overcome this fixed-depth constraint. We add zero-initialized residual connections from higher layers at token t to lower layers at token $t + 1$.

Properties of this architecture:

- The effective depth of reasoning grows linearly with sequence length.
- The t to $t + 1$ design enables increased reasoning depth without increased total computation.

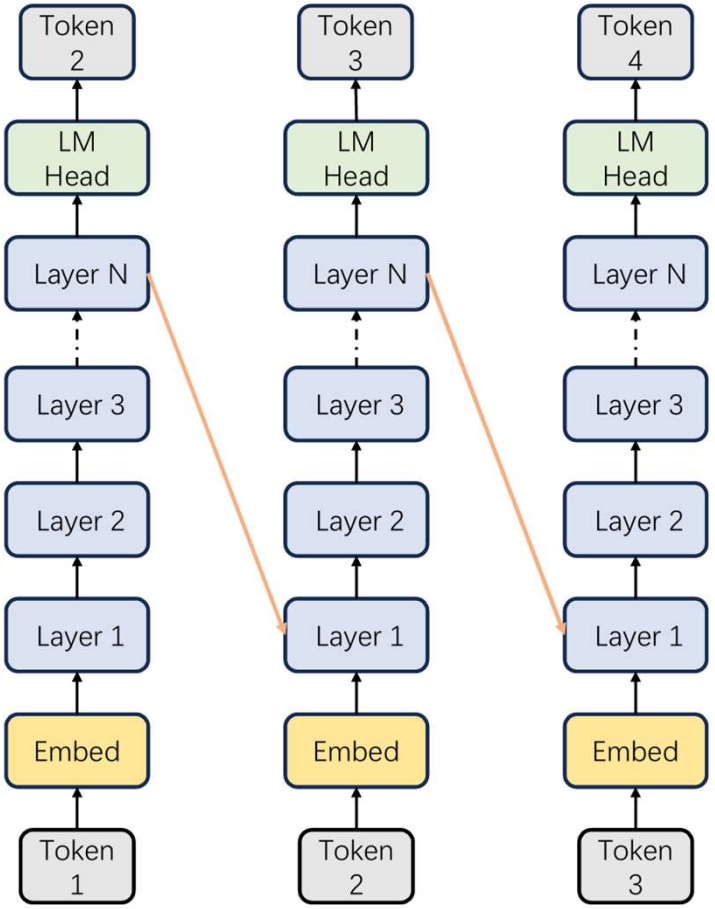


Our Method

Chain-of-Thought is also a method to pass information from higher to lower layers, by sending a single discrete token back to the embedding layer. In contrast, our method sends continuous vectors, which carry significantly more information.

Our method preserves the standard language modeling objective, making it compatible with pretraining.

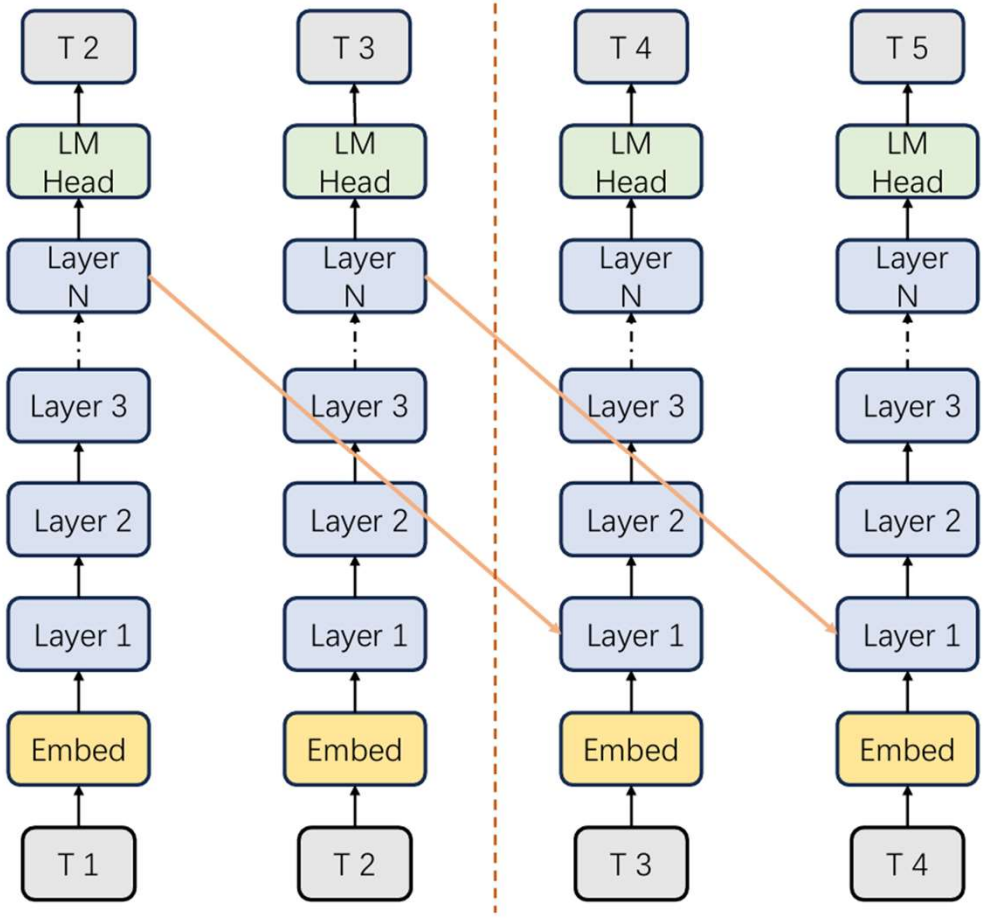
$$\mathcal{L} = -\sum_{i=1}^{k} \log P(t_i \mid t_0, t_1, \ldots, t_{i-1})$$



Grouping

Major tradeoff of TurboConn: loss of full parallelism during training and the prefilling stage of inference.

This is mitigated by “grouping”: group tokens together and send information back in groups. Intra group parallelism is preserved.



Fine-tuning Performance

Model	Dataset	Method	Acc (%)
Llama 3.2 1B	GSM8K (No CoT)	Baseline	7.20
		TurboConn	**8.32**
	Multi-step-arithmetic	Baseline	38.16
		TurboConn	**42.66**
	Parity	Baseline	92.87
		TurboConn	**100.0**
Llama 3.1 8B	GSM8K (No CoT)	Baseline	23.92
		TurboConn	**24.82**
	Multi-step-arithmetic	Baseline	48.32
		TurboConn	**51.86**
	Parity	Baseline	89.40
		TurboConn	**100.0**
Qwen 3 1.7B	GSM8K (No CoT)	Baseline	15.90
		TurboConn	**20.31**
	Multi-step-arithmetic	Baseline	36.10
		TurboConn	**45.81**
	Parity	Baseline	53.78
		TurboConn	**100.0**

Dataset (Avg. Seq. Length)	Method	# Recursions	Time/Step (seconds) (× Transformer)	Acc (%) ($\alpha = 1$)	Acc (%) ($\alpha = 100$)
Parity (154.69 tokens)	Baseline	1.00	1.18 (1.00×)	92.87	–
	TurboConn (Group 4)	38.81	5.75 (4.87×)	98.94	**100.00**
	TurboConn (Group 6)	25.94	3.07 (2.60×)	98.68	**100.00**
	TurboConn (Group 8)	19.84	2.66 (2.25×)	98.83	51.42
	TurboConn (Group 16)	10.00	1.61 (1.36×)	98.85	51.40
Multi-step Arithmetic (125.75 tokens)	Baseline	1.00	1.01 (1.00×)	38.16	–
	TurboConn (Group 4)	31.81	4.46 (4.42×)	40.13	**42.66**
	TurboConn (Group 6)	21.37	2.29 (2.27×)	39.94	41.74
	TurboConn (Group 8)	16.13	1.82 (1.80×)	39.79	41.48
	TurboConn (Group 16)	8.17	1.45 (1.44×)	38.79	38.79
GSM8K (No CoT) (101.92 tokens)	Baseline	1.00	0.77 (1.00×)	7.20	–
	TurboConn (Group 4)	25.86	3.22 (4.18×)	7.67	**8.32**
	TurboConn (Group 6)	17.40	1.73 (2.25×)	7.87	7.95
	TurboConn (Group 8)	13.18	1.39 (1.81×)	7.55	8.17
	TurboConn (Group 16)	6.84	1.38 (1.80×)	7.41	6.94

ICML

International Conference
On Machine Learning

Thank you