On the Inductive Bias of Stacking Towards Improving Reasoning

Nikunj Saunshi, Stefani Karp, Shankar Krishnan, Sobhan Miryoosefi, Sashank J. Reddi, Sanjiv Kumar
Google Research

Presentation at NeurIPS 2024

LLM pretraining efficiency

Scale is important, but makes **training very expensive**

- Time, Resources, $$, Emissions

Better optimizer

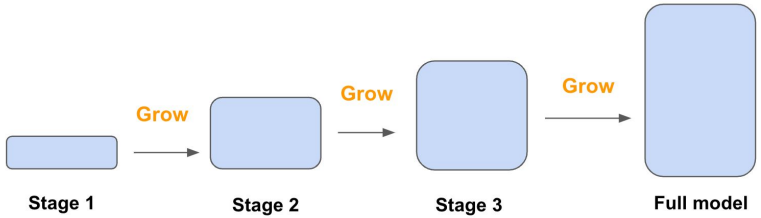
(E.g. AdamW, Shampoo, SOAP, Sophia, …)

- **Fewer steps** to optimize the loss
- Implicit biases: simplicity, sparsity, flatness

Stagewise growing

(E.g. progressive/gradual stacking, bert2bert, LiGO, MSG, …)

- **Lesser walltime** & FLOPs for fixed #steps
- **Biases:** Unknown



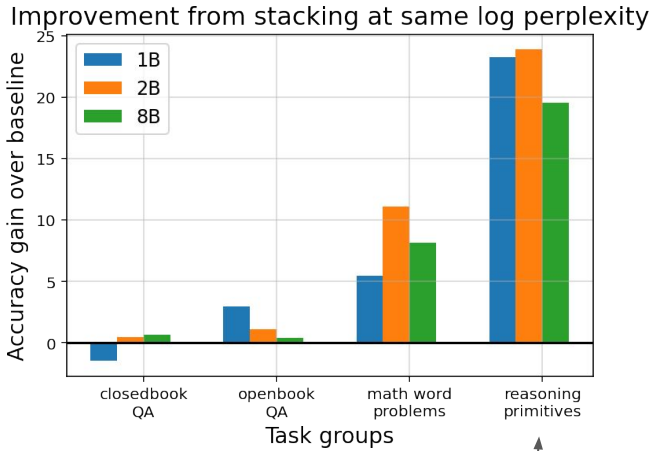
(1) Train small model
(2) Use it to initialize larger model
(3) Repeat for multiple stages

Stagewise growing

Speeds up BERT. **Doesn't scale** to language modeling

- Need tricks (see also KKNMK23)

Bias unknown



This work

- **MIDAS**: New and better stacking approach to grow in depth
- **Training efficiency**: Speeds up 1B,2B,8B LM pretraining by upto 40%
- **Inductive bias**: Significantly improves reasoning at same perplexity! (Connection to looped models)

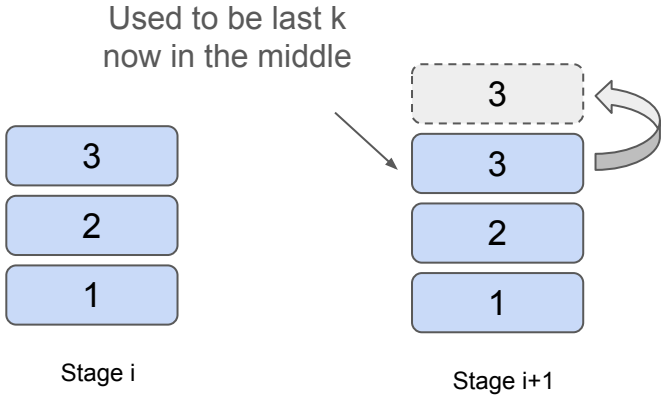
Improves **+20%** on reasoning primitives!

KKNMK23: No Train No Gain: Revisiting Efficient Training Algorithms For Transformer-based Language Models

How to stack layers?

Gradual stacking: Duplicate last k layers

- Better than random init



Insight: Copying last k messes with the **role of layers** at init

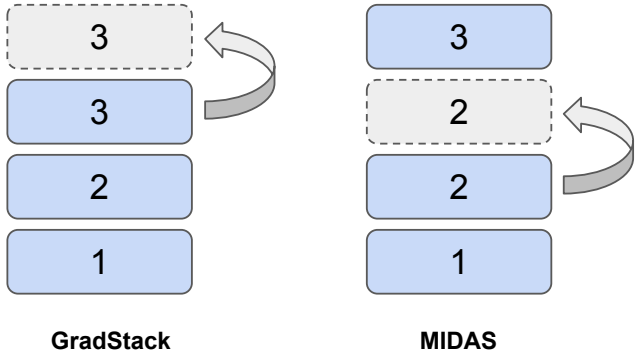
- First and last layers typically play a special role (encoding/decoding)

Proposal: Stack the middle k layers

- Layers play a more similar role

RMKKKKK23: Efficient Training of Language Models using Few-Shot Learning

MIDAS: Middle Layer Stacking



MIDAS Algorithm

1) Partition training steps in L/k stages
2) Duplicates the middle k layers in each stage
3) Follow all hyperparameters as standard training

Experiments with 1B,2B,8B Transformer models on language modeling

Baseline vs GradStack vs MIDAS

MIDAS: Key findings

Training Efficiency

Result 1: MIDAS >> GradStack in all settings

Result 2: MIDAS >= Baseline with ~25-40% speedup

Inductive bias

Result 3: Better downstream evals at the same validation perplexity

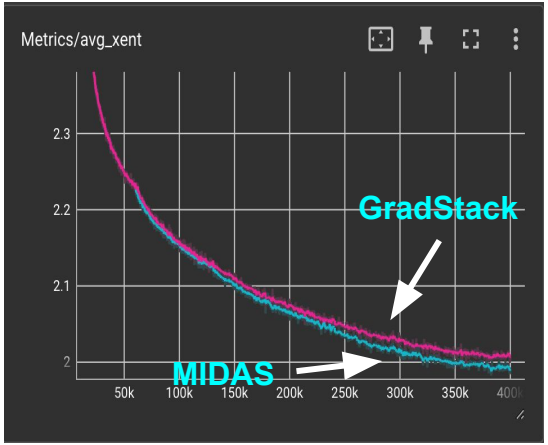
Result 4: Significantly improves tasks that require reasoning

MIDAS: Training efficiency

Training Efficiency

Result 1: MIDAS >> GradStack in all settings

Result 2: MIDAS >= Baseline with ~25-40% speedup



Gap increases in each stage

MIDAS: Training efficiency

Training Efficiency

Result 1: MIDAS >> GradStack in all settings

Result 2: MIDAS >= Baseline with ~25-40% speedup

Evaluate on a suite of 15 downstream tasks

(including closed/open book QA, math problems)

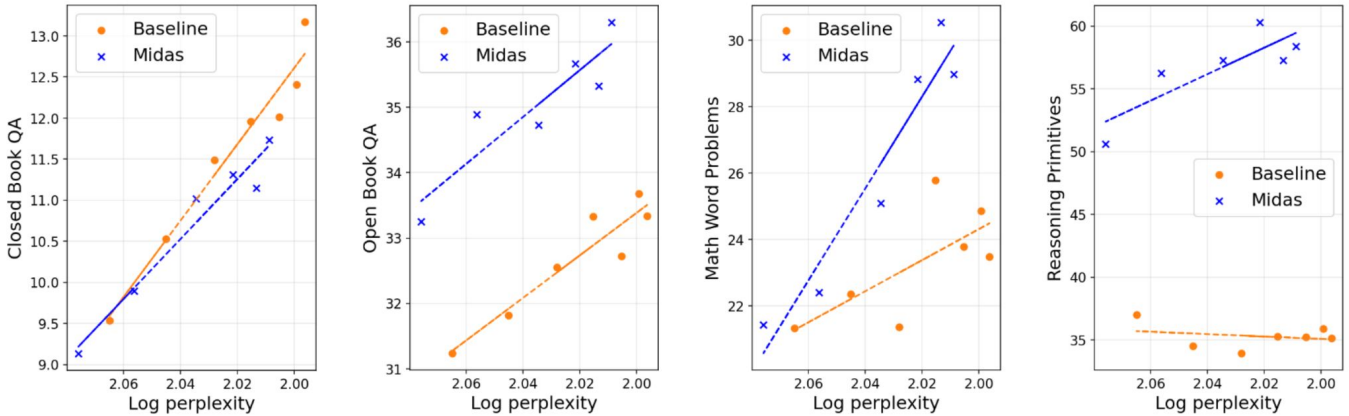
		Speedup	Loss (validation)	Downstream Task Average (15 tasks)
1B	**Baseline**	1x	2.00	24.0
	MIDAS	1.39x	2.03	26.7
	MIDAS	1.24x	2.01	26.8
	MIDAS	1.16x	2.00	28.3
2B	**Baseline**	1x	1.93	28.0
	MIDAS	1.39x	1.95	29.5
	MIDAS	1.24x	1.93	32.9
8B	**Baseline**	1x	1.84	32.8
	MIDAS	1.26x	1.84	36.4

MIDAS: Inductive bias

Inductive bias

Result 3: Better downstream evals at the same validation perplexity

Plot **downstream eval** vs **validation pretraining loss** as training proceeds

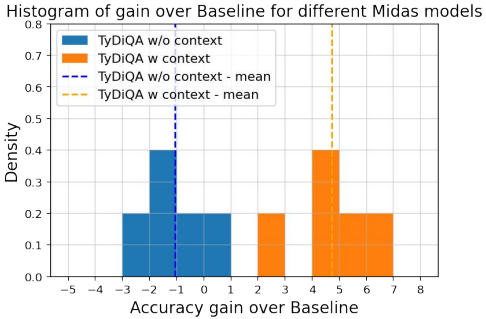


MIDAS extracts more skills at the same pretraining ability

MIDAS: Inductive bias

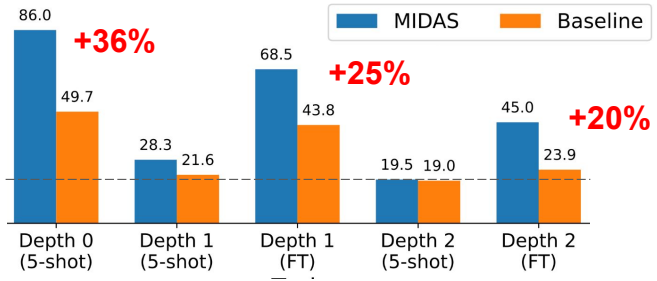
Inductive bias

Result 4: Significantly improves tasks that require reasoning



	Model	Math WPs (5-shot)	GSM8k (Finetune)
2B	Baseline	27.1	8.5
	MIDAS	**38.3**	**14.5**
8B	Baseline	34.9	15.8
	MIDAS	**43.1**	**18.7**

+10%



Improvements on Open book QA >>
Improvements on Closed book QA

Large improvements on math
(with and without finetuning)

Construct **reasoning primitives**
Even larger improvements

Depth 2: a=5, b=**3**, c=b, d=b, e=c, e=__ -> Ans: **3**

Key takeaways

Training Efficiency

Inductive bias

Result 1: MIDAS >> GradStack in all settings

Result 3: Better downstream evals at the same validation perplexity

more skills from same pretraining ability

Result 2: MIDAS >= Baseline with ~25-40% speedup

Result 4: Significantly improves tasks that require reasoning

connection to looped models

arxiv.org/pdf/2409.19044

nsaunshi@google.com