Structured Gradient Methods for Deep Learning

Tetiana Parshakova

CERN Alumni & Flatiron Institute, NY

06/25/2026

LLMs success: scaling model $\times$ data $\times$ compute

why the recipe scales

- **transformer:** training parallelizes over sequence positions
- **next-token prediction:** each token acts as a training target
- **scale:** more parameters, tokens, and compute give predictable reductions in loss

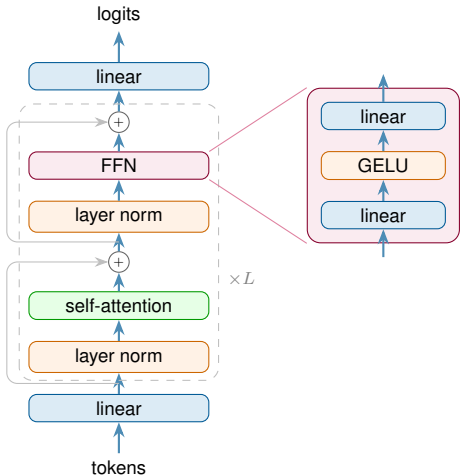
where resources go

- **pretraining:** GPU compute, communication, power, storage, and data processing
- **post-training:** demonstrations, preferences, rollouts, and evaluations
- **inference:** memory and compute for every prompt and generated token

optimization goal: reach a given target loss with fewer FLOPs

Transformers are structured maps

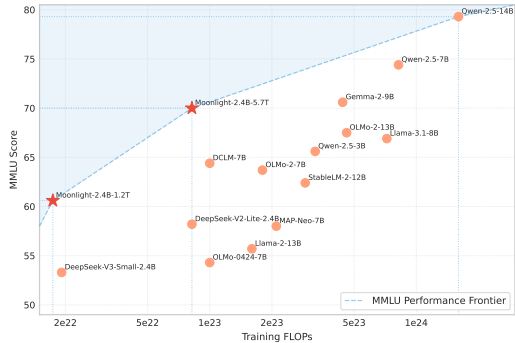
- Adam optimizer is oblivious to structure of transformer
- transformers is compositions of structured blocks
  - linear
  - normalization (RMSNorm)
  - self-attention
  - FFN
  - mixture of experts (MoE)



design principle: use the structure of each block

Muon a new optimizer for deep learning

- MMLU score
  - 57 multiple choice exams
  - math, history, law, CS, etc
- Muon advances Pareto frontier of performance vs training FLOPs
  - Moonlight trained with Muon
- 1T parameters Kimi model [Kimi team, 2025]
- 1.6T parameters Deepseek-V4-Pro [Deepseek-AI, 2026]
- 753B parameters GLM-5.2 [GLM team, 2026]



[Liu et al. (2025)]

Linear layers

- linear layer is a map $x \mapsto Wx$, update changes the output by

$$(W + \Delta W)x - Wx = \Delta W x$$

- worst-case change in output

$$\sup_{\|x\|_2=1} \|\Delta W x\|_2 = \|\Delta W\|_2$$

- stable training requires controlling scale of change in outputs

$$\|\Delta W\|_2 \leq \eta \implies \|\Delta W x\|_2 \leq \eta \|x\|_2 \quad \text{for all } x$$

- spectral stepest descent

$$\Delta W \in \operatorname*{arg\,min}_{\|\Delta W\|_2 \leq \eta} \left(f(W_t) + \langle \nabla f(W_t), \Delta W \rangle \right)$$

conservative view: control operator norm of the update

Beyond linear layers

block operation $\Longrightarrow$ control the change in the output $\Longrightarrow$ optimizer geometry

block	mapping g_θ	worst-case change $\sup_{\|x\|=1} d\left(g_{\theta_{t+1}}(x), g_{\theta_t}(x)\right)$
linear	$g_\theta(x) = Wx$	$\|W_{t+1} - W_t\|_2$
attention	$g_\theta(X) = \mathrm{softmax}_{\mathrm{row}}\left(\frac{X^\top W_Q^\top W_K X}{\sqrt{d_{\mathrm{att}}}}\right)$	$\sup_{\|X\|=1} \mathrm{KL}\left(g_{\theta_{t+1}}(X) \,\middle\|\, g_{\theta_t}(X)\right)$
MoE	$g_\theta(x) = \sum_{e \in \mathcal{S}_\theta(x)} r_{\theta,e}(x)\, \mathtt{FFN}_{\theta,e}(x)$	$\sup_{\|x\|_2=1} \left\|g_{\theta_{t+1}}(x) - g_{\theta_t}(x)\right\|_2$

choosing the metric based on the block structure