TEXAS
The University of Texas at Austin

DeepSeek-V3

Presented by Bozhi You

DeepSeek-AI, DeepSeek-V3 Technical Report

Agenda

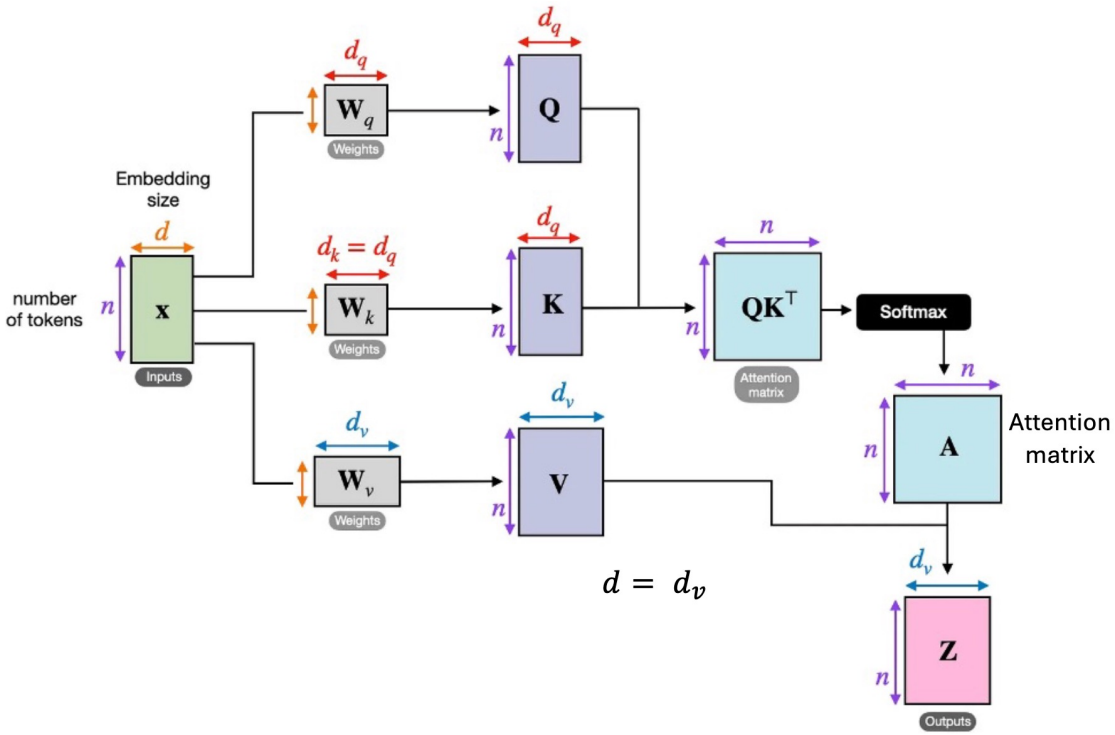
TEXAS
The University of Texas at Austin

- Multi-Head Latent Attention (MLA)
- Auxiliary-Loss-Free MoE
- Multi-Token Prediction (MTP)
- DualPipe
- NOT covered
  - FP8 Mixed Precision Training, etc.
  - Evaluation
  - DeepSeek-V2/V3.1/R1

Multi-Head Latent Attention (MLA)

TEXAS
The University of Texas at Austin

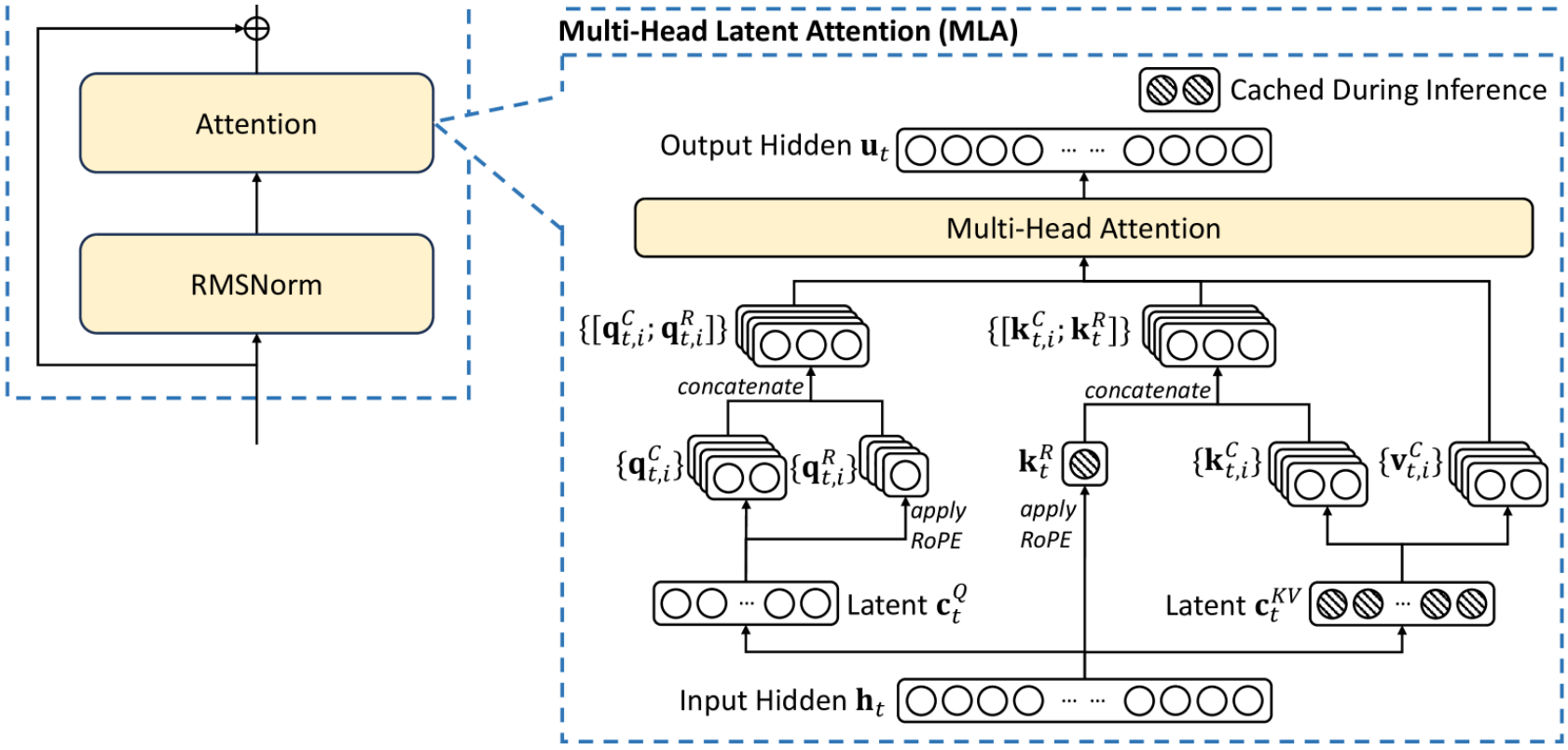
- KV cache is large – how to reduce?



Multi-Head Latent Attention (MLA)

TEXAS
The University of Texas at Austin

- KV cache is large – how to reduce?
- $q = xW^Q, k = xW^K, v = xW^V$ — x is source of truth
  - Compress x: $c = xW$ (compression rate = aspect ratio of W)
  - $q = cW^{Q'}, k = cW^{K'}, v = cW^{V'}$
- Details
  - Separate compression path for Q and KV
  - Positional information not compressed


Attention
RMSNorm
Multi-Head Latent Attention (MLA)
Cached During Inference
Output Hidden $\mathbf{u}_t$
Multi-Head Attention
$\{[\mathbf{q}_{t,i}^C; \mathbf{q}_{t,i}^R]\}$
$\{[\mathbf{k}_{t,i}^C; \mathbf{k}_t^R]\}$
concatenate
concatenate
$\{\mathbf{q}_{t,i}^C\}$
$\{\mathbf{q}_{t,i}^R\}$
$\mathbf{k}_t^R$
$\{\mathbf{k}_{t,i}^C\}$
$\{\mathbf{v}_{t,i}^C\}$
apply RoPE
apply RoPE
Latent $\mathbf{c}_t^Q$
Latent $\mathbf{c}_t^{KV}$
Input Hidden $\mathbf{h}_t$

Agenda

TEXAS
The University of Texas at Austin

- Multi-Head Latent Attention (MLA)
- **Auxiliary-Loss-Free MoE**
- Multi-Token Prediction (MTP)
- DualPipe

Background: Mixture-of-Experts (MoE)

TEXAS
The University of Texas at Austin

- DeepSeek is a large MoE model
- Intuition: a big model → a *set* of smaller models
  - Ensemble learning: run all of them, aggregate on outputs
  - DSMoE: pick and run a *subset* of them, aggregate on outputs
- Details
  - Conceptually many *expert*s but trained jointly + parameter sharing
  - Some are standing, others picked by *Router*
  - Weighted aggregation

Revisiting Mixture of Experts (MoE)

- Sparse MoE layers instead of dense feed-forward network (FFN) layers.
- Routers determines which tokens are sent to which expert.

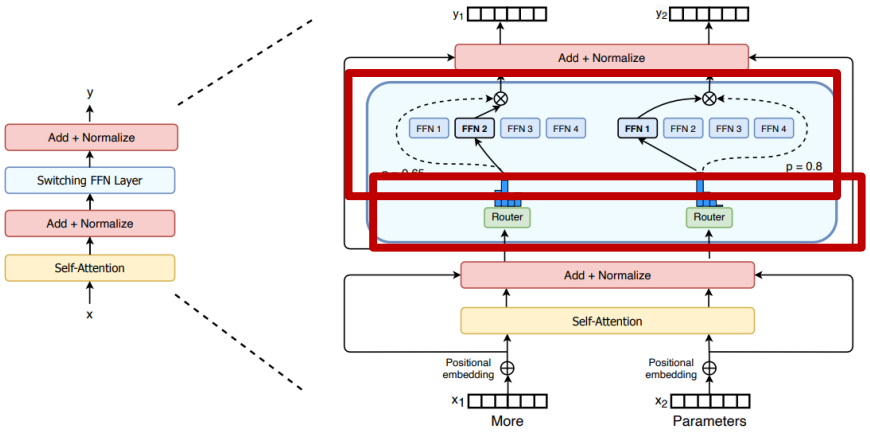


Figure 2: Illustration of a Switch Transformer encoder block. We replace the dense feed forward network (FFN) layer present in the Transformer with a sparse Switch FFN layer (light blue). The layer operates independently on the tokens in the sequence. We diagram two tokens (x_1 = "More" and x_2 = "Parameters" below) being routed (solid lines) across four FFN experts, where the router independently routes each token. The switch FFN layer returns the output of the selected FFN multiplied by the router gate value (dotted-line).

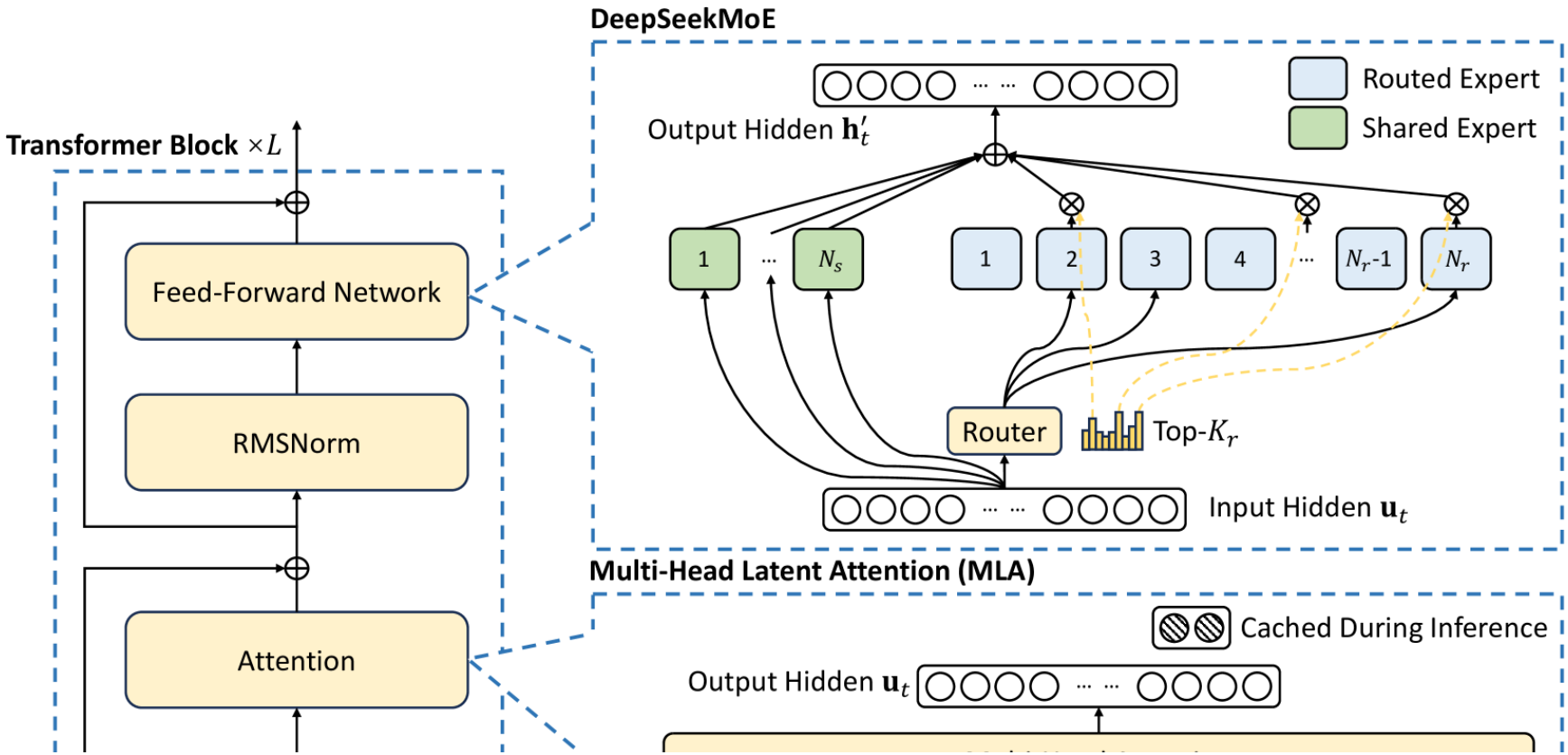
https://huggingface.co/blog/moe
Shazeer et al. 2022

Background: Mixture-of-Experts (MoE)

TEXAS
The University of Texas at Austin

- DeepSeek is a large MoE model
- Intuition: a big model → a *set* of smaller models
  - Ensemble learning: run all of them, aggregate on outputs
  - DSMoE: pick and run a *subset* of them, aggregate on outputs
- Details
  - Conceptually many *expert*s but trained jointly + parameter sharing
  - Some are standing, others picked by *Router*
  - Weighted aggregation
- *671B total parameters with 37B activated for each token*
  - *1 shared expert and 8 of 256 routed*

Background: Mixture-of-Experts (MoE)



Auxiliary-Loss-Free MoE

TEXAS
The University of Texas at Austin

- Auxiliary loss for load balance
  - A lazy router may always pick its favorite experts – load imbalance!
  - To discourage this, loss = task-specific loss + penalty on choosing same expert
  - Penalty might be emphasized during training
- Don't penalize it, prevent it
  - Biased weight when routing
  - Bias reflects workload

$$g'_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} + b_i \in \text{Topk}(\{s_{j,t} + b_j | 1 \leqslant j \leqslant N_r\}, K_r), \\ 0, & \text{otherwise.} \end{cases}$$

Agenda

TEXAS
The University of Texas at Austin

- Multi-Head Latent Attention (MLA)
- Auxiliary-Loss-Free MoE
- **Multi-Token Prediction (MTP)**
- DualPipe

Multi-Token Prediction (MTP)

TEXAS
The University of Texas at Austin

- What needs to be changed?
  - Output
  - Loss
- Multi-head output layer, one head for each stride
- Loss aggregation: cross-entropy

$$\mathcal{L}_{\text{MTP}}^{k} = \text{CrossEntropy}(P_{2+k:T+1}^{k}, t_{2+k:T+1}) = -\frac{1}{T}\sum_{i=2+k}^{T+1} \log P_i^k[t_i]$$

Agenda

TEXAS
The University of Texas at Austin

- Multi-Head Latent Attention (MLA)
- Auxiliary-Loss-Free MoE
- Multi-Token Prediction (MTP)
- **DualPipe**

DualPipe

TEXAS
The University of Texas at Austin

- How to leverage multi-GPU systems?
  - Data parallelism, model parallelism, …
- Pipeline layers over GPUs
  - Issue?
- DualPipe: overlapping forward and backward pipelines

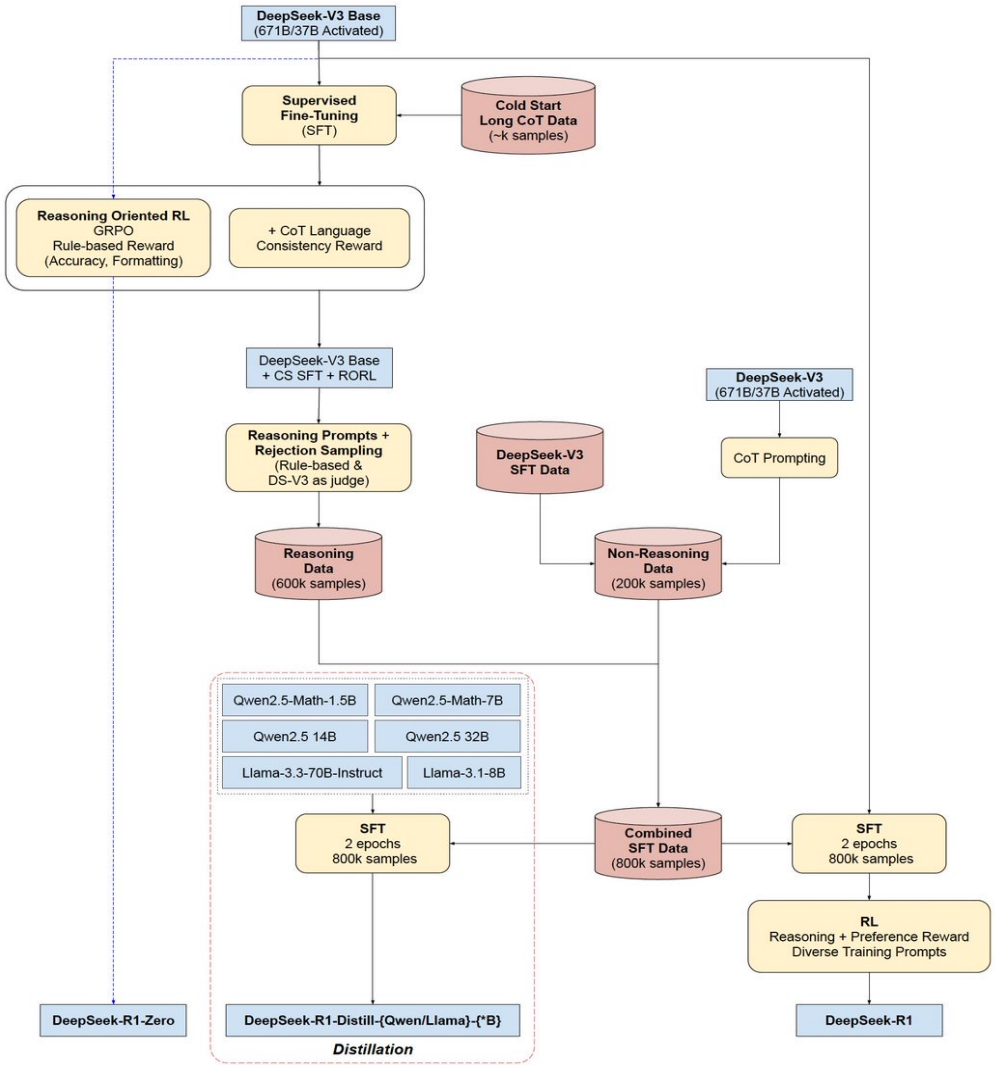
Takeaways

TEXAS
The University of Texas at Austin

- Multi-Head Latent Attention (MLA)
  - Reduce large KV cache: compress input x to latent c , then derive q, k, v from c
- Auxiliary-Loss-Free MoE
  - Eliminate the auxiliary loss; use a biased weight during routing
- Multi-Token Prediction (MTP)
  - Multi-head output layer + loss aggregated via cross-entropy
- DualPipe
  - Multi-GPU; pipeline parallelism; overlaps the forward and backward pipelines

DeepSeek Training Paradigm

- From V3 to R1 (Jan 2025)
- V3.1 (Aug 2025)



https://x.com/SirrahChan/status/1881488738473357753/photo/1
https://www.youtube.com/watch?v=QdEuh2UVbu0&list=WL&index=10