Notes on DeepSeek

Hong Lan

FMRG, CBHB

February 12, 2026

1. Manifold-Constrained Hyper-Connections

2. Conditional Memory

1. Manifold-Constrained Hyper-Connections

2. Conditional Memory

Hyper-Connections

- Let $x_l \in \mathbb{R}^{n\times C}$ be the l-th layer of a ResNet, the recursive structure of hyper-connections across layers proposed by Zhu et al. (2025) writes

$$x_{l+1} = \mathcal{H}_l^{res} x_l + \left(\mathcal{H}_l^{post}\right)^\top \mathcal{F}\left(\mathcal{H}_l^{pre} x_l, \mathcal{W}_l\right)$$

- where $\mathcal{H}_l^{res} \in \mathbb{R}^{n\times n}, \mathcal{H}_l^{post} \in \mathbb{R}^{1\times n}, \mathcal{H}_l^{pre} \in \mathbb{R}^{1\times n}$ are learnable mappings, and $\mathcal{F} : \mathbb{R}^{n\times C} \to \mathbb{R}^{1\times C}$ denotes the residual function. Iterating forward yields

$$x_L = \left(\prod_{i=1}^{L-l} \mathcal{H}_{L-i}^{res}\right) x_l + \sum_{i=l}^{L-1} \left(\prod_{j=1}^{L-1-i} \mathcal{H}_{L-j}^{res}\right) \left(\mathcal{H}_i^{post}\right)^\top \mathcal{F}\left(\mathcal{H}_i^{pre} x_i, \mathcal{W}_i\right)$$

- where $L > l$ indexes the layer deeper than l, and n measures the rate expanded from the original layer of dimension $1 \times C$.

Source of Instability

- As noticed by Xie et al. (2026), the product of the sequence of learnable mapping $\mathcal{H}_i^{res}$ may introduce instability into training

$$\prod_{i=1}^{L-l} \mathcal{H}_{L-i}^{res} = \mathcal{H}_{L-1}^{res}\mathcal{H}_{L-2}^{res} \dots \mathcal{H}_{L-l}^{res}$$

- This can be seen by assuming $\mathcal{H}_i^{res} = \mathcal{H}^{res}$, $\forall\, i$—applying the eigenvalue decomposition, the product above rewrites

$$\left(\mathcal{H}^{res}\right)^{L-l} = D\Lambda^{L-l}D^{-1}$$

- which is stable only if the largest absolute value of any eigenvalue

$$\rho\left(\mathcal{H}^{res}\right) = \max\left\{|\lambda|_1, \dots, |\lambda|_{L-l}\right\}$$

- i.e., the spectral radius of $\mathcal{H}^{res}$ is smaller than 1.

Manifold-Constrained Hyper-Connections

- The mHC proposed by Xie et al. (2026) ensuring stability in training by restraining $\mathcal{H}_l^{res}$ to be a doubly stochastic matrix, i.e., let h_{ij} be the i, jth element of $\mathcal{H}_l^{res}$

$$\sum_i h_{ij} = \sum_j h_{ij} = 1,\ h_{ij} \geq 0\ \forall i, j$$

- The mHC is implemented by employing Sinkhorn and Knopp's (1967) method, in essence

$$M^{(t)} = \mathcal{T}_r\left(\mathcal{T}_c\left(M^{(t-1)}\right)\right),\ M^{(0)} = \exp\left(\mathcal{H}_l^{res}\right)$$

- where $\mathcal{T}_r, \mathcal{T}_c$ denotes the row, column normalization respectively.

1. Manifold-Constrained Hyper-Connections

2. Conditional Memory

Reasoning–Retrieval Decomposition in Language Models

- Cheng et al. (2026) formalize language modeling as comprising two qualitatively distinct sub-tasks: compositional reasoning and knowledge retrieval
- Vaswani et al.'s (2017) Transformer lacks an explicit retrieval mechanism; both reasoning and retrieval are implemented via matrix computation
- Attention-based computation scales as $\mathcal{O}(T^2 d)$, where T denotes sequence length and d model dimension, whereas classical n-gram lookup achieves $\mathcal{O}(1)$ access
- Cheng et al. (2026) introduce the Engram module to decouple retrieval from computation, enabling a more efficient allocation of computational budget across the two sub-tasks

Engram Module 1/2: Retrieving

- $\mathcal{P}$ maps tokens to canonical identifiers, e.g., normalization

$$x'_t = \mathcal{P}(x_t),\ t = 1, 2, \ldots, T$$

- Construct suffix N-grams as local context descriptors

$$g_{t,n} = \left(x'_{t-n+1}, \ldots, x'_t\right),\ n = 2, \ldots, N$$

- Multi-head hashing approximates a large N-gram table without explicit enumeration

$$z_{t,n,k} = \varphi_{n,k}(g_{t,n}),\ k = 1, \ldots, K$$

- Retrieve and concatenate learned memory embeddings

$$\begin{aligned} e_{t,n,k} &= E_{n,k}(z_{t,n,k}),\ \forall n, k \\ \mathbf{e}_t &= \begin{bmatrix} e_{t,2,1} & \cdots & e_{t,2,K} & \cdots & e_{t,N,1} & \cdots & e_{t,N,K} \end{bmatrix} \end{aligned}$$

- Note, the above retrieval via multi-head hashing requires $\mathcal{O}(1)$

Engram Module 2/2: Gating

- Let $\mathbf{h}_t$ denote the hidden state from preceding attention layers, the gate measures semantic alignment between current global context and retrieved memory

$$\alpha_t = \sigma\left(\frac{\mathsf{RMSNorm}(\mathbf{h}_t)^\top \mathsf{RMSNorm}(\mathbf{k}_t)}{\sqrt{d}}\right),\ \mathbf{k}_t = W_K \mathbf{e}_t$$

- If memory is irrelevant, $\alpha_t \to 0$. The context-aware gate filters the retrieved memory

$$\tilde{\mathbf{v}}_t = \alpha_t \mathbf{v}_t,\ \mathbf{v}_t = W_V \mathbf{e}_t$$

- A short convolution expands expressivity and injects local interaction

$$\mathbf{Y} = \mathsf{SiLU}(\mathsf{Conv1D}(\mathsf{RMSNorm}(\tilde{\mathbf{V}}))) + \tilde{\mathbf{V}}$$

- Residual structure preserves stability as usual

Engram-Augmented Transformer

- Let $\mathbf{X}$ be an input sequence, recall the computation graph of the standard Transformer

$$\begin{aligned}\mathbf{B} &= \text{LN}(\mathbf{X} + \text{Attention}(\mathbf{X}))\\ \mathbf{H} &= \text{LN}(\mathbf{B} + \text{FFN}(\mathbf{B}))\end{aligned}$$

- Given $\mathbf{H}$, the Engram module takes as well $\mathbf{X}$ to compute $\mathbf{Y}$, and then augments $\mathbf{H}$ through residual connection

$$\begin{aligned}\mathbf{Y} &= \text{Engram}(\mathbf{H}, \mathbf{X})\\ \mathbf{H} &= \mathbf{H} + \mathbf{Y}\end{aligned}$$

- Followed by the standard attention and MoE. Note that, if the gate suppresses memory, the block reduces to the standard Transformer
- The model can now allocate capacity between computation and retrieval instead of simulating both with depth alone

References

Cheng, Xin, Wangding Zeng, Damai Dai, Qinyu Chen, Bingxuan Wang, Zhenda Xie, Kezhao Huang, Xingkai Yu, Zhewen Hao, Yukun Li, Han Zhang, Huishuai Zhang, Dongyan Zhao, and Wenfeng Liang. 2026. "Conditional Memory via Scalable Lookup: A New Axis of Sparsity for Large Language Models." URL: `https://arxiv.org/abs/2601.07372`.

Sinkhorn, Richard, and Paul Knopp. 1967. "Concerning nonnegative matrices and doubly stochastic matrices." Pacific Journal of Mathematics, 21: 343–348.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." URL: `https://arxiv.org/abs/1706.03762`.

Xie, Zhenda, Yixuan Wei, Huanqi Cao, Chenggang Zhao, Chengqi Deng, Jiashi Li, Damai Dai, Huazuo Gao, Jiang Chang, Kuai Yu, Liang Zhao, Shangyan Zhou, Zhean Xu, Zhengyan Zhang, Wangding Zeng, Shengding Hu, Yuqing Wang, Jingyang Yuan, Lean Wang, and Wenfeng Liang. 2026. "mHC: Manifold-Constrained Hyper-Connections." URL: `https://arxiv.org/abs/2512.24880`.

Zhu, Defa, Hongzhi Huang, Zihao Huang, Yutao Zeng, Yunyao Mao, Banggu Wu, Qiyang Min, and Xun Zhou. 2025. "Hyper-Connections." URL: `https://arxiv.org/abs/2409.19606`.