FibChat – It makes stuff up, but with structure.

Preprint · July 2025

DOI: 10.13140/RG.2.2.10549.59364/1

1 author:

Gary Nan Tie
Mu Risk LLC
94 PUBLICATIONS **27** CITATIONS

SEE PROFILE

FibChat

- *It makes stuff up, but with structure.*

Query updating via fibration attention

Gary Nan Tie, July 14th, 2025

Abstract

Inspired by Joyal-Tierney symbolic calculus, we present a weak factorization system fibration of neural networks, to update and refine attention queries, in an abductive explainable way, that is coherently coupled to key-value pairing. Hallucinations can be mitigated by perturbing the abductive choice of key for the value outputted by this fibration attention mechanism.

Introduction

In the database retrieval paradigm for transformer attention, we are given a database of key-value pairs and we try to match our query with a suitable key. Note that a value may be paired with multiple keys. Attention is choosing a key whose corresponding value is most germane to our query. The matching process refines a query through a series of updates involving a learnable feed-forward network to introduce non-linearity, otherwise stacked updates would be equivalent to a single update and no refinement (lowering cross-entropy) achieved.

In contrast, we introduce a categorical deep learning paradigm for attention, to update queries utilizing abductive reasoning. As we will see, a certain fibration refines a query in a coherent structured way and explains how a germane key was chosen for the attention output value. The learned fibration lifting couples together query updating and key-value pairing. Moreover, one can steer the continuation by perturbing the abductive choice of key paired to the value outputted by this fibration attention mechanism, thereby mitigating propensity to hallucinate.

Query-Key-Value database retrieval paradigm

In latent space $\mathbb{R}^d$, let subsets Q, K and V denote queries, keys and values respectively, and let $Q' \subseteq \mathbb{R}^d$ denote updated queries, to be defined.

We are given data: $q_1, \ldots, q_n \in Q$,

$k_1, \ldots, k_m \in K$, $v_1, \ldots, v_m \in V$.

Let $L : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ be a loss function,

For $\varepsilon > 0$, define $x =_\varepsilon y$ iff $L(x,y) < \varepsilon$.

Suppose there exists $\varepsilon > 0$ such that we can learn a neural network $p : K \to V$ with $p(k_j) =_\varepsilon v_j$, $j \in [m]$, a key-value pairing.

Let score be a similarity function, eg a kernel.

Fix i, define $a_{ij} = \text{score}(q_i, k_j)$ and $\alpha_{ij} = \text{soft max } a_{ij}$

Define initial updated query $q_i^0 = \sum_{j=1}^{m} \alpha_{ij} v_j \in Q'$, $i \in [n]$,

and $j_0 \triangleq \arg\max_j \text{score}(q_i^0, k_j)$.

Transformer Attention via Fibration

Template diagram of neural networks, commuting up to $\equiv_\varepsilon$, that updates and refines a query in an attention layer:

$$\begin{array}{ccc} Q \otimes K & \xrightarrow{m \otimes 1} & Q' \otimes K \\ {\scriptstyle 1 \otimes \rho}\downarrow & \overset{\sigma}{\dashrightarrow} & \downarrow{\scriptstyle 1 \otimes \rho} \\ Q \otimes V & \xrightarrow[m \otimes 1]{} & Q' \otimes V \end{array} \qquad (\dagger)$$

Suppose there exists $\varepsilon > 0$ so that from $(\dagger)$ we can learn:

$$\sigma^0 : Q \otimes V \to Q' \otimes K \text{ with } \sigma^0 \begin{bmatrix} q_i \\ p(k_j) \end{bmatrix} \equiv_\varepsilon \begin{bmatrix} q_i^0 \\ k_j \end{bmatrix}, \; j \in [m]$$

$$\sigma^1 : Q \otimes V \to Q' \otimes K \text{ with } \sigma^1 \begin{bmatrix} q_i \\ p(k_j) \end{bmatrix} \equiv_\varepsilon \begin{bmatrix} q_i^1 \\ k_j \end{bmatrix}, \; j \in [m]$$

where $q_i^1 \triangleq \pi_1 \sigma^0 \begin{bmatrix} q_i \\ p(k_{j_0}) \end{bmatrix}$

$$\vdots$$

$$\sigma^N : Q \otimes V \to Q' \otimes K \text{ with } \sigma^N \begin{bmatrix} q_i \\ p(k_j) \end{bmatrix} \equiv_\varepsilon \begin{bmatrix} q_i^N \\ k_j \end{bmatrix}, \; j \in [m]$$

where $q_i^N = \pi_1 \sigma^{N-1} \begin{bmatrix} q_i \\ p(k_{j_{N-1}}) \end{bmatrix}$, $j_{N-1} \triangleq \arg\max_j \text{score}(q_i^{N-1}, k_j)$

is the refined query returned.

Choose stacking hyperparameter $N > 6$ large enough so that resulting perplexity is acceptable.

As before, fix i, let $b_{ij} = \text{score}(q_i^N, k_j)$ and $\beta_{ij} = \text{soft max}\, b_{ij}$

Define $\beta_{i\hat{j}} = \max_j \beta_{ij}$ and attention output value be $v_{\hat{j}} =_\varepsilon \rho(k_{\hat{j}})$.

Let $V \xrightarrow{1} Q \otimes V$ and define $\tau \triangleq \pi_2 \sigma^N 1 : V \to K$

$$v \longmapsto \begin{bmatrix} q_i \\ v \end{bmatrix}$$

then $\rho(\tau(v_j)) =_\varepsilon v_j$, $j \in [m]$, ie for affect ρ,

key $\tau(v_j)$ explains value v_j, a form of abduction.

To summarize, given query q_i, lifting σ^N (learnt)

yields a refined query q_i^N, abduction in turn yields

a key $\tau(v_{\hat{j}})$ whose value $\rho(\tau(v_{\hat{j}})) =_\varepsilon v_{\hat{j}}$

is the attention output returned.

$$\sigma^N : Q \otimes V \to Q' \otimes K, \quad \sigma^N \begin{bmatrix} q_i \\ v_j \end{bmatrix} =_\varepsilon \begin{bmatrix} q_i^N \\ \tau(v_j) \end{bmatrix}$$

and $\sigma^N \begin{bmatrix} q_i \\ \rho(k_j) \end{bmatrix} =_\varepsilon \begin{bmatrix} q_i^N \\ k_j \end{bmatrix}$ with respect to (†).

and $q_i^N = \pi_1 \sigma^{N-1} \begin{bmatrix} q_i \\ \rho(k_{j_{N-1}}) \end{bmatrix}$, $j_{N-1} = \arg\max_j \text{score}(q_i^{N-1}, k_j)$.

Hallucination mitigation

Suppose we were dissatisfied with continuation $v_{\hat{j}}$.

We can perturb key $\tau(v_{\hat{j}})$ as follows:

Let $e_i \triangleq (0,..,\underset{i\text{th position}}{1},..,0) \in \mathbb{R}^d$ and $\mathcal{U}$ be a uniform distribution on $[d]$, and k a random draw from $[d]$.

Let $\lambda > 0$, $\tilde{v}(\lambda) \triangleq \rho(\tau(v_{\hat{j}}) + \lambda e_k) \in V$

is a perturbed continuation to consider as alternative to $v_{\hat{j}}$. By varying the direction and magnitude of perturbations a satisfactory perplexity might be achieved.

Conclusion

FibChat - fibration structured attention.

Fibrations are a coherent structured way to refine queries and to explain how a germane key was chosen. Hallucinations are potentially mitigated by perturbing our abductive choice of key.

References

[1] 'Fibrations explain all you need!
- Fibrations, Abduction, Attention
Gary Nan Tie, Mar 4, 2025
DOI: 10.13140/RG.2.2.35984.52488
To conceptualize and perform abductive machine learning, categorical fibrations are proposed, as they coherently give both context and structure for abductive reasoning via an attention mechanism. Moreover, 2-categorical lifting compatibility conditions ensure consistent hierarchical and parallel explanations.

[2] 'Neural Network Abduction'
Gary Nan Tie, Jun 13, 2025
DOI: 10.13140/RG.2.2.18506.07360/1
Abductive reasoning has a fibration semantics that can be implemented by neural networks; a step towards artificial general intelligence.

[3] 'Parsimonious Neural Network Abduction'
Gary Nan Tie, Jun22, 2025
DOI: 10.13140/RG.2.2.15695.80804
For hypotheses whose effect is manifested by observations, abduction seeks to explain a given observation by finding a hypothesis whose effect is that observation. Abductive reasoning has a fibration semantics that can be implemented by neural networks. In this note we introduce a parsimonious choice of explanation according to one's utility function.

[4] 'Attention Fibration Abduction'
Gary Nan Tie, Jul 2, 2025
DOI: 10.13140/RG.2.2.29060.23680/1
Transformer attention and abductive reasoning are both instances of a weak factorization system fibration, implemented by neural networks.