Attention Fibration Abduction

Gary Nan Tie, Jul 2, 2025

Abstract

Transformer attention and abductive reasoning are intimately connected, embodied by fibrations. Suitable attention mechanisms can make a key-value pairing a learnable neural network fibration for abductive reasoning. A fibration lifting explains an attention output value to a query, by identifying the key paired to it; explainable attention through abduction.

Keywords

large language model (LLM), transformer attention, abductive reasoning, fibration, category theory, neural network, deep learning, XAI

Introduction

Words in a large language model are represented by tokens, long vectors of real numbers, that are mapped to a lower dimensional latent space so that related words are close to one another. Transformer attention is modeled as query key-value database retrieval. In a category with objects subsets of the latent space and arrows neural networks between them, we reinterpret the database retrieval paradigm as a categorical fibration, whose lifting induces a retract that is a form of abduction. As we will see, fibrations embody both attention and abduction as learnable neural networks.

Fibrations explain all you need!

In latent space $\mathbb{R}^d$, let subsets Q, K, V denote queries, keys and values respectively, and let

$Q' = \{$ finite sums of probability weighted values $\}$

denote updated queries.

Let $q_1, \ldots, q_n \in Q$ and

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

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

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

Suppose there exists $\varepsilon_1 > 0$ such that we can learn

a neural network $\rho : K \longrightarrow V$ with $\rho(k_j) =_{\varepsilon_1} v_j$, $j = 1, \ldots m$

Let score be a similarity function, for example a kernel.

Define $a_{ij} = \text{score}(q_i, k_j)$, $\alpha_{ij} = \text{softmax}\, a_{ij}$

Define updated query $q_i' = \sum_{j=1}^{m} \alpha_{ij} v_j \in Q'$

for $i \in [n]$.

Let $b_{ij} = \text{score}(q'_i, v_j)$ and $\beta_{ij} = \text{softmax}\, b_{ij}$,

and $a'_{ij} = \text{score}(q'_i, k_j)$ and $\alpha'_{ij} = \text{softmax}\, a'_{ij}$

Let $\alpha_{i\hat{j}} = \max_j \alpha_{ij}$, $\beta_{i\check{j}} = \max_j \beta_{ij}$ and $\alpha'_{i\tilde{j}} = \max_j \alpha'_{ij}$

For query q_i, attention outputs value $\rho(k_{\tilde{j}})$.

Suppose there exist $\varepsilon_2 > 0$ and $\varepsilon_3 > 0$ such that we can

learn neural networks $f: Q \to K$ and $g: Q' \to V$

so that for $i \in [n]$: $f(q_i) =_{\varepsilon_2} k_{\hat{j}}$, $\rho(f(q_i)) =_{\varepsilon_2} \rho(k_{\hat{j}})$

and $g(q'_i) =_{\varepsilon_3} v_{\check{j}}$.

Suppose there exists $\varepsilon_4 > 0$ such that we can learn

neural network $m: Q \to Q'$ so that for $i \in [n]$:

$m(q_i) =_{\varepsilon_4} q'_i$, $g(m(q_i)) =_{\varepsilon_4} g(q'_i)$

and $g \circ m(q_i) =_{\varepsilon_4} \rho \circ f(q_i)$.

Suppose there exists $\varepsilon_5 > 0$ such that we can learn

neural network $h : Q' \to K$ so that for $i \in [n]$:

$h(q_i') =_{\varepsilon_5} k_{\tilde{j}}$, $h \circ m(q_i) =_{\varepsilon_5} f(q_i)$, $\rho \circ h(q_i') =_{\varepsilon_5} g(q_i')$,

and $h(m(q_i)) =_{\varepsilon_5} h(q_i')$, $\rho(h(q_i')) =_{\varepsilon_5} \rho(k_{\tilde{j}})$.

Let $\varepsilon = \max\{\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5\}$, then

$$\begin{array}{ccc} Q & \xrightarrow{f} & K \\ {\scriptstyle m}\downarrow & \nearrow_{h} & \downarrow{\scriptstyle \rho} \\ Q' & \xrightarrow[g]{} & V \end{array} \qquad (\dagger) \text{ commutes up to } =_\varepsilon$$

$$v_{\hat{j}} =_\varepsilon g(q_i') =_\varepsilon g \circ m(q_i) =_\varepsilon \rho \circ f(q_i) =_\varepsilon \rho(k_{\hat{j}}) =_\varepsilon v_{\tilde{j}}$$

$$v_{\tilde{j}} =_\varepsilon \rho(k_{\tilde{j}}) =_\varepsilon \rho \circ h(q_i') =_\varepsilon g(q_i') =_\varepsilon v_{\hat{j}} \qquad k_{\tilde{j}} =_\varepsilon h(q_i') =_\varepsilon h \circ m(q_i) =_\varepsilon f(q_i) =_\varepsilon k_{\hat{j}}$$

If $\varepsilon_1, \ldots, \varepsilon_5$ are acceptable then ρ is a fibration,

with lifting h, which induces partial retract $\tau : V \to K$

via $g \circ \tau = h$. As $\rho \circ \tau|_{g(Q')} = 1_{g(Q')}$

τ is an abduction for affect ρ (see [3] for details).

Fibration diagram $(\dagger)$ can be interpreted as

attention and abduction.

For query q_i, the attention mechanism output is value

$$p(k_j) = p(h(q_i')) = g(q_i') = v_j$$

So for the key-value pairing p the key $k_j = h(q_i')$ explains the choice of value $v_j = g(q_i')$ for the attention output. Note that the existence of a lifting h making effect/pairing p a fibration depends on the learnability of the attention mechanisms f, g, h and neural network m.

Conclusion:

Suitable attention mechanisms can make a key-value pairing a learnable neural network fibration for abductive reasoning. A fibration lifting explains the attention output value to a query, by identifying a key paired to it; explainable attention through abduction.

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.