# LLM が Hallucination を起こす確率

吹谷和雄

2024.04.07

【現状課題】

LLM を活用するにあたって Hallucination が起きることは、業務系の Task では致命的な欠陥である。最近は、それを回避するために、いろいろな方策（RAG など）が Prompt Engineering で提案されている。しかし、最初の Prompt の Response に対する数回以降の再 Prompt（Few Shot など）の Response の精度が、どんどん落ちていく傾向がある。

その理由は、”文脈を継承”していると言われている Prompt と Response の Context 履歴のすべてを連結 Concatenate して、次の Prompt に付加する為のようである。その根本的な構造的原因が下記にあげる Entangle な”混合状態”にある。

【数理解析】

LLM/Transformer の core である Self-Attention を基に近年の量子コンピューティング Quantum Computing の観点から数理的構造を解明する。

文書/文章を Token/形態素を要素とする 2 階テンソルの”状態”$|\mathrm{D}\rangle$とする。
その Token の線形結合は、

$$|\mathrm{D}\rangle = \sum_{i=1}^{m}\sum_{j=1}^{n}\Lambda_{ij}|i\rangle_U|j\rangle_V \qquad \cdots①$$

で表される。これは部分系 U と V で構成され、U はｍ個の規格化された直交基底$|i\rangle_U$で、V はｎ個の規格化された直交基底$|j\rangle_V$である。$\Lambda_{ij}$は線形結合の係数で、ｍ行ｎ列の行列になっているので、これを特異値分解をすると、

$$|D\rangle = \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\sum_{\rho=1}^{k}U_{i\rho}\lambda_\rho V_{j\rho}^{*}\right]|i\rangle_U|j\rangle_V \qquad \text{where, k=min(n,m)}$$

となり、これを

$$|U_\rho\rangle = \sum_{i=1}^{m}U_{i\rho}|i\rangle_U\ , \qquad |V_\rho\rangle = \sum_{i=1}^{n}V_{j\rho}^{*}|j\rangle_V$$

と新たな基底に定義すると、①式は、

$$|D\rangle = \sum_{\rho=1}^{k}\lambda_\rho|U_\rho\rangle|V_\rho\rangle \qquad \cdots②$$

と書き直せる。

ちなみに、文書/文章$|D\rangle$の偏平度[1]を表すノルム Norm の二乗は、

$$\left[\sum_{\rho=1}^{k}\langle V_\rho|\langle U_\rho|\lambda_\rho\right]\left[\sum_{\sigma=1}^{k}\lambda_\sigma|U_\sigma\rangle|V_\sigma\rangle\right] = \sum_{\rho=1}^{k}\left(\lambda_\rho\right)^2$$

となり、固有値$\lambda_\rho$を使って表せる。

---

[1] 文書/文章の”質”を表す指標数値

以下では、$\langle D|D\rangle = 1$と規格化して扱う。

文書/文章$|D\rangle$の密度行列（Density Matrix）を定義する。

$$\hat{\rho} = |D\rangle\langle D| = \left[\sum_{ij}\Lambda_{ij}|i\rangle_U|j\rangle_V\right]\left[\sum_{i'j'}\langle j'|_V\langle i'|_U\Lambda^*_{i'j'}\right] \quad \cdots③$$

$\hat{\rho}$を使うと、演算子$\hat{O}$の期待値をトレースの形

$$\langle\hat{O}\rangle = \langle D|\hat{O}|D\rangle = Tr\left[\hat{O}\hat{\rho}\right]$$

で表せて、大変便利です。

そこで、基底の表記　$|ij\rangle = |i\rangle_U|j\rangle_V$　を使って、

$$\rho_{(i'j')(ij)} = \langle i'j'|\hat{\rho}|ij\rangle = \langle i'j'|D\rangle\langle D|ij\rangle = \Lambda_{i'j'}\Lambda^*_{ij} \quad \cdots④$$

と密度行列を表示できる。部分系 V についての縮約

$$\sum_j\langle j|\hat{\rho}|j\rangle = \sum_j\left[\sum_i\Lambda_{ij}|i\rangle\right]\left[\sum_{i'}\langle i'|\Lambda^*_{i'j}\right]$$

は、部分系 V についてトレースを取ったと考えられるので$Tr_V\hat{\rho}$と書き表せる。

部分系 U に対する密度行列（Density Matrix）も

$$\hat{\rho}^U = Tr_V\hat{\rho} = \sum_{ii'}\left(\sum_j\Lambda_{ij}\Lambda^*_{i'j}\right)|i\rangle\langle i'|$$

となる。$\hat{\rho}^U$は密度行列になり、基底$|i\rangle_U$を使った行列表示では、

$$\rho^U_{ii'} = \sum_j\Lambda_{ij}\Lambda^*_{i'j}$$

となる。同様に、部分系 U について縮約を取ると、

$$\hat{\rho}^V = Tr_U\hat{\rho}\,, \quad \rho^V_{jj'} = \sum_i\Lambda_{ij}\Lambda^*_{ij'}$$

というように、部分系 V についての密度行列もつくれる。

文書/文章$|D\rangle$に対する特異値分解は、密度行列と深く関係している。

密度行列$\rho^U_{ii'}$の特異値分解

$$\begin{aligned}\rho^U_{ii'} &= \sum_j\Lambda_{ij}\Lambda^*_{i'j} = \sum_j\sum_\rho U_{i\rho}\lambda_\rho V^*_{j\rho}\sum_\sigma U^*_{i'\sigma}\lambda_\sigma V_{j\sigma}\\ &= \sum_{\rho\sigma}U_{i\rho}\lambda_\rho\delta_{\rho\sigma}U^*_{i'\sigma}\lambda_\sigma = \sum_\rho\left(\lambda_\rho\right)^2U_{i\rho}U^*_{i'\rho}\end{aligned}$$

から考える。この関係から、U の列ベクトルが$\rho^U_{ii'}$の固有ベクトルであり、固有値が$\left(\lambda_\rho\right)^2$で与えられることがわかる。従って、密度行列は式③の基底を使って、

$$\hat{\rho}^U = \sum_{ii'}\rho^U_{ii'}|i\rangle\langle i'| = \sum_\rho\left(\lambda_\rho\right)^2\sum_i U_{i\rho}|i\rangle\sum_{i'}\langle i'|U^*_{i'\rho} = \sum_\rho\left(\lambda_\rho\right)^2|U_\rho\rangle\langle U_\rho|$$

と表される。部分系 V についても、同じように密度行列を

$$\rho^V_{jj'} = \Sigma_\rho\left(\lambda_\rho\right)^2V^*_{j\rho}V_{j'\rho}\ , \qquad \hat{\rho}^V = \Sigma_\rho\left(\lambda_\rho\right)^2|V_\rho\rangle\langle V_\rho|$$

と構成できる。密度行列のトレースが|D⟩のノルムを与えると

$$Tr_U\hat{\rho}^U = Tr_V\hat{\rho}^V = \sum_\rho (\lambda_\rho)^2 = \langle D|D\rangle$$

も再確認できる。これらの式から、$|U_\rho\rangle_U$は$\hat{\rho}^U$の固有状態、$|V_\rho\rangle_V$は$\hat{\rho}^V$の固有状態であり、共通の固有値$(\lambda_\rho)^2$を持つことがわかる。

密度行列$\hat{\rho}^U$と$\hat{\rho}^V$の固有値$(\lambda_\rho)^2$には、確率解釈を持ち込むことが可能である。部分系Uについて、基底$|U_\rho\rangle_U$で構成した射影演算子

$$\hat{P}_\rho^U = |U_\rho\rangle\langle U_\rho|$$

の組$\hat{P}_1^U, \hat{P}_2^U, \cdots, \hat{P}_k^U$で表される射影測定を、$\hat{\rho}^U$で記述される状態に対して行うと、終状態として$|U_\rho\rangle_U$が測定される確率が$(\lambda_\rho)^2$となる。部分系Vに対しても、同様に基底$|V_\rho\rangle_V$から射影測定を構成できて、同じ確率$(\lambda_\rho)^2$を得る。

この規格化された確率について、情報エントロピー

$$S = -\sum_\rho (\lambda_\rho)^2 ln(\lambda_\rho)^2 \qquad \cdots⑤$$

が定義できて、エンタングルメント・エントロピー（Entanglement Entropy）と呼ばれる。
直積$|D\rangle = |U_1\rangle\langle V_1|$で状態が表される場合には、部分系 U と V はエンタングルしておらず、$\hat{\rho}^U$も$\hat{\rho}^V$も純粋状態の密度行列になり、$\lambda_1 = 1$以外の特異値は全て 0 となる。従ってこの場合、エンタングルメント・エントロピーは S＝0 である。２つ以上の特異値が正の値となる混合状態では、部分系 U と V はエンタングルしていて、S＞0 となる。
このことから、式⑤の S を、エンタングルメントの定量的な指標とすることが考えられる。
すべての特異値が等しい場合に S は最大値

$$S_{max} = -\sum_{\rho=1}^{k} \frac{1}{k} ln \frac{1}{k} = lnk$$

となり、この場合には部分系 U と V は「最も強くエンタングルしている」（Maximaly Entangled）と言い表せる。
部分系 U だけに作用するユニタリー演算子$\hat{Q}^U$と、部分系 V だけに作用するユニタリー演算子$\hat{Q}^V$を、状態$|D\rangle$に作用させたもの

$$\hat{Q}^U\hat{Q}^V|D\rangle = \sum_\rho \lambda_\rho \hat{Q}^U|U_\rho\rangle\hat{Q}^V|V_\rho\rangle$$

を考える。ユニタリー変換は、直交基底を別の直交基底に写すだけなので、左辺の特異値分解が右辺で与えられるという関係は、$\hat{Q}^U\hat{Q}^V$の作用の下でも保たれる。また、特異値$\lambda_\rho$

も変化しない。従って、
・部分系に作用するユニタリー演算子は、エンタングルメントに影響しない
といえる。ユニタリーではない演算子が部分系に作用したり、部分系 U と部分系 V に「またがって」演算子が作用する場合には、エンタングルメントが増えたり減ったりする。
例えば、式⑤で表される射影演算子の組で表される測定を行うと、終状態はいずれかの射影演算子が作用したものとなり、エンタングルメント・エントロピーが 0 である純粋状態に変化する。

実際的な応用の場面では、往々にして$\lambda_\rho$が速やかに、おおよそ指数関数的に 0 へと減衰する状況がある。例えば、文書/文章$|D\rangle$が事前学習データや Prompt などで基底状態を表していれば、この速やかな$\lambda_\rho$の減衰がみられる。この場合、状態$|D\rangle$の良い近似が、少ない数の項の重ね合わせ

$$|D\rangle = \sum_{\rho=1}^{\chi} \lambda_\rho |U_\rho\rangle |V_\rho\rangle$$

で与えられる。ただし、$\chi$は、$\chi < k = \min(m,n)$を満たす整数である。
・大きな特異値$\chi$個だけを「拾い」、小さな特異値は「捨てる」という近似を行っても、$\lambda_1 \gg \lambda_\chi$が成立している限り、状態はほとんど変化しない。（次元圧縮の妥当性）
元の$|D\rangle$と、近似$|\tilde{D}\rangle$の距離は、

$$\varepsilon^2 = \left||D\rangle - |\tilde{D}\rangle\right|^2 = \sum_{\rho=\chi+1}^{k} (\lambda_\rho)^2$$

と、「捨てた」密度演算子の固有値の和で与えられる。「拾う数」$\chi$を充分に多くしておけば、$\varepsilon^2$の値を小さく保つことが可能である。
$\varepsilon^2 \ll 1$が満たされる場合、$|\tilde{D}\rangle$が持つエンタングルメント・エントロピー

$$\tilde{S} = -\sum_{\rho=1}^{\chi} (\lambda_\rho)^2 ln(\lambda_\rho)^2$$

は、$|D\rangle$のエンタングルメント・エントロピーS より、かすかに小さい値となるだけで、実質上「S と$\tilde{S}$はほぼ等しい」と言ってよい。これがテンソルネットワーク形式で使われる、最も重要な近似である。

結論をまとめると、事前学習データや Prompt などのテキストの Attention 抽出のコア式$softmax\left(\frac{QK^T}{\sqrt{d}}\right)V$は、直積という純粋状態で扱っているが、実際のテキストは Entangle された混合状態になっているので、Hallucination が起きる。その確率は$(\lambda_\rho)^2$に依存され、Entanglement Entropy という指標 S で程度がわかる。 □