事前学習済みTransformerモデルにおける隠れ層間冗長性の解析とバイパス機構の探索

静岡大学）馬場海好　狩野芳伸

Kano Laboratory

導入

問題点

- 言語モデルの大規模化によるモデル内部の処理効率に疑問
  → 冗長性が存在するのでは無いかという仮説
- モデルの隠れ状態に対する定量的な分析手法が無い

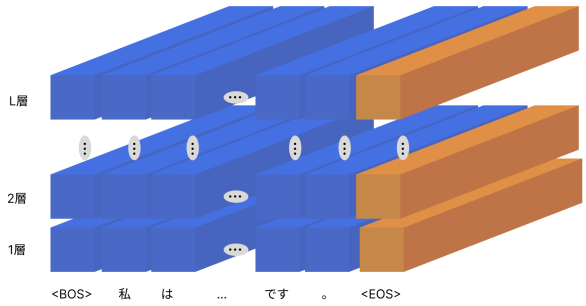
この問題に対して

- Transformerモデルの隠れ状態に対する分析手法の提案
- 隠れ層間の冗長性の解析

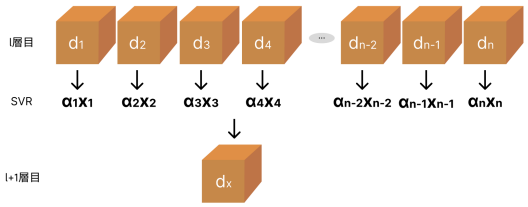
を行った

提案手法

1. Transformerモデルにテキストを入力し、層ごとの隠れ状態を得る



2. 特定のトークンに対するn層目の隠れ状態ベクトルからn+1層目のd次元の値を予測するタスクに対して各層・次元ペアでサポートベクター回帰(SVR)を用いて学習・推論



3. 推論結果(R^2 *Score*)から高い精度で推論が行えている層ペアを特定し、学習されたSVRのパラメータから、大きい重みに対応する次元を特定し、隠れ層間の冗長性を解析する

SVRによって求められる式

$$y_{d(n+1)} = a_{1(n)}x_{1(n)} + a_{2(n)}x_{2(n)} + \cdots + a_{\mathrm{d}-1(n)}x_{\mathrm{d}-1(n)} + a_{\mathrm{d}(n)}x_{\mathrm{d}(n)}$$

d: 次元数　a: *SVR*の重み　x_d: n入力

実験

＊ 使用モデル

言語モデル：rinna/llama-3-youko-8b

- 隠れ状態の次元数　：4096
- 層数　：32

＊ データセット

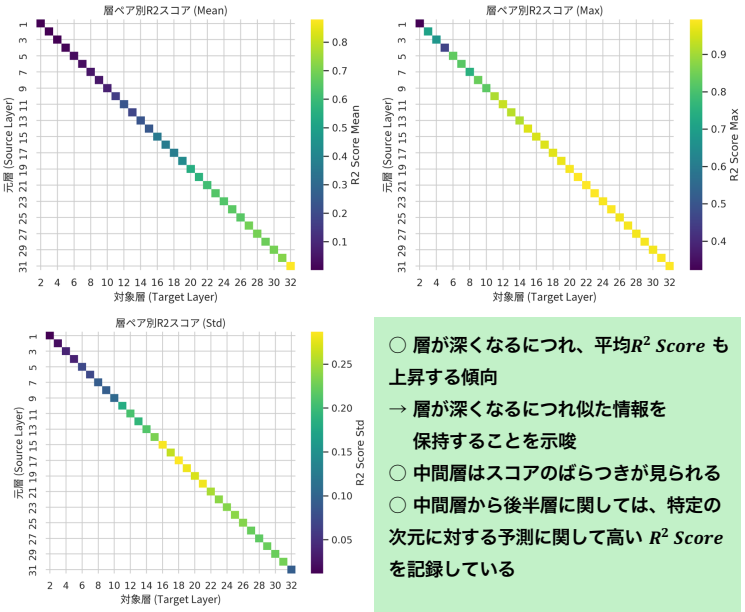
JSTSデータセットのsentence1からランダムに 10000件

＊ SVRの学習・推論で用いる隠れ状態

データ数：10000件からランダムに取得された1000件

- うち学習/推論　：800件/200件
- 注目トークン　：<EOS>

実験: 寄与度分析



○ 層が深くなるにつれ、平均R^2 *Score* も上昇する傾向
→ 層が深くなるにつれ似た情報を保持することを示唆
○ 中間層はスコアのばらつきが見られる
○ 中間層から後半層に関しては、特定の次元に対する予測に関して高い R^2 *Score* を記録している

図1・2・3：層別 R^2 *Score* のヒートマップ(Mean・Max・Std)

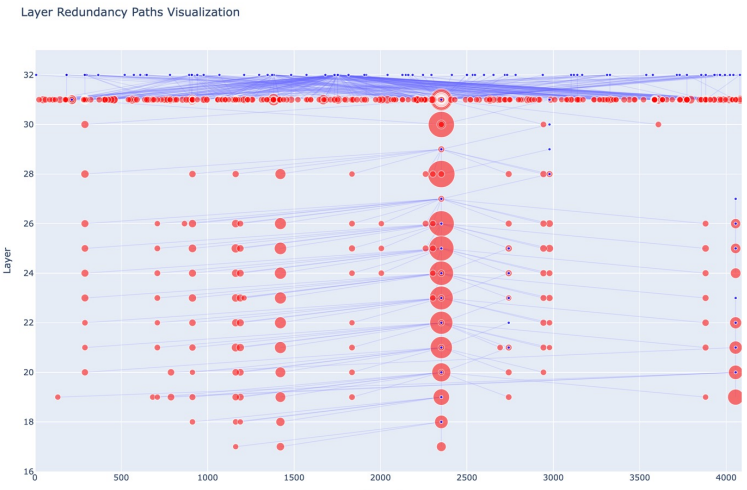


図4： R^2 *Score* 上位500件の寄与次元の可視化

線で繋がった先の青いノードの推論に寄与している次元を赤いノードで表現

○ 2352次元に対する推論結果が上位500件に多く見られる
また、1421・2352次元の寄与が大きく見られ、層ごとに見た際にも、寄与している次元に違いが見られない。
→ 機械的な情報の流れ：冗長性を示唆？
○ 最終層に関しては寄与次元の分布に他の層と異なる特徴が見られる。
他の層とは異なる情報処理が行われていることを示唆

展望

- 他のモデル・他のトークンにおける同様の分析
- 言語モデルに対する具体的なバイパス機構の提案