

研究発表

文献

背景

近年大規模言語モデル(LLM)の数が増加

- ChatGPT(OpenAI), LLaMa(Meta)
- LLMの中でも、大規模のものと小規模のものがある.

パラメータ

- 「パラメータ数」：言語モデルの規模を表す指標の一つ.
- モデルが学習する際に調整される変数のこと.

多様なLLM

- 大規模なLLM：計算コストやリソースが高いが、ハイスペック.
- 小規模なLLM：
モデルのサイズを抑えることができるが、高度な質問等の精度は低い.

問題

生成言語モデル多すぎ問題

- 簡単な質問に答えるのにハイスペックモデルを使うのは、リソースの無駄遣い.
- ユーザが簡単な質問をするか、複雑な質問をするかを予想するのは困難.

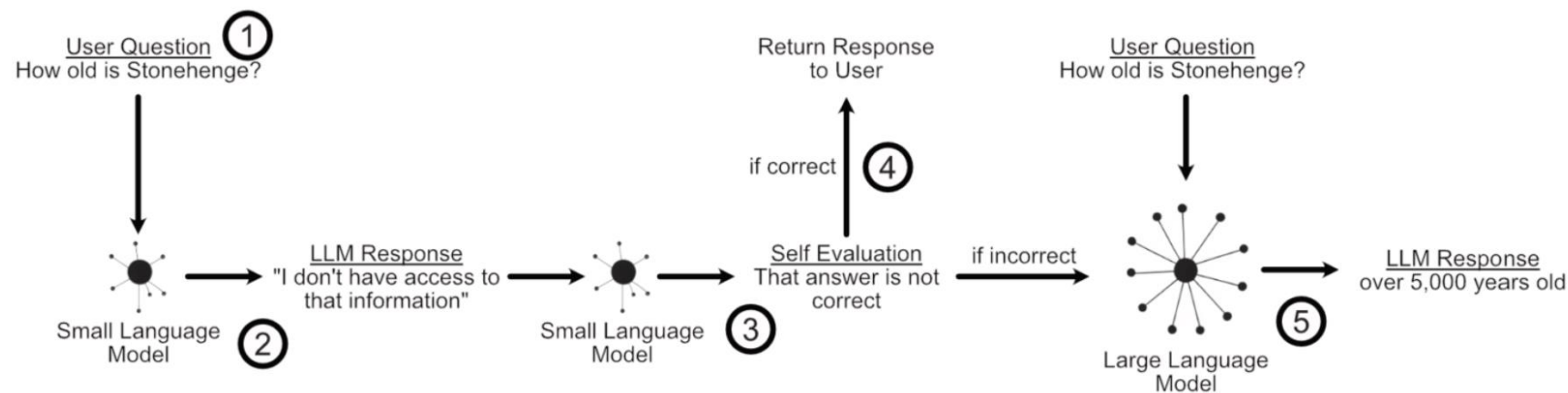
LLMルーティング

- 入力されたクエリを分析し、そのクエリに適したモデルをルーティングする.
- パフォーマンスを維持しながら、コストを削減する.

先行研究

LLMカスケード

- 小さいモデルと大きいモデルのどちらを使用するかを判別できる.
- AutoMix, FrugalGPT



(1) 小さな LLM にクエリを渡し、(2) 応答を生成します。(3) 次に、同じ小さな LLM に質問が正しく回答されたかどうかを尋ねます。(4) 正しく回答された場合は、結果を返します。(5) 小さな LLM が正しくなかった場合は、より大きな LLM を使用して再試行します。

課題点

- 数多くあるLLMからどのLLMを選んでくるかがの方法が確立されていない.

目的・アプローチ

大量にあるLLMから候補となるLLMを選ぶ

- 人がLLMの候補を選ぶより，機械的に候補を選別できる.

アプローチ：モデルのメタ情報等の活用

- モデルのメタ情報
 - モデルの構成要素の情報
- 必要な情報？
 - モデルのメタ情報，モデルの評価...等？
- Hugging face(AI技術のGit Hubのようなもの)に登録されているモデルから，いくつか選び，それを表示させる.
- 例：text generationの分野から評価の高いものをとってくる.

Hugging face

AI技術のために提供されているオープンソースプラットフォーム

- 「Transformers」がベース
- AIモデルやデータを共有したり利用したりできる.
- LLMのGitHubのようなもの.

このプラットフォームの中からLLMを引っ張っていき たい

- どのような情報が入っているか

Hugging face ライブラリ

huggingface_hub

- Hugging Face Hubと接続できる公式ライブラリ
- `hf_hub_download()`
 - モデルやデータの特定ファイルをダウンロード

モデルの構成情報 (config.json)

- モデルの構成情報は事前に「config.json」に保存されている.
- `config.json`
 - モデルをどう組み立てるかを記述したメタ情報ファイル

config.json

architectures	使用されるモデルのクラス名（例: BertForMaskedLM）
model_type	モデルのタイプ（例: bert, gpt2, llama など）
vocab_size	使用する語彙数（トークナイザーと一致する必要がある）
hidden_size	隠れ層のユニット数（ベースモデルでは通常 768）
num_hidden_layers	トランスフォーマーブロック（エンコーダ層）の数
num_attention_heads	マルチヘッドアテンションのヘッド数
intermediate_size	FFN層（フィードフォワードネットワーク）の中間層のサイズ
hidden_act	隠れ層に使う活性化関数（例: "gelu", "relu"）
initializer_range	重み初期化の分布の範囲（例: 0.02）
layer_norm_eps	レイヤーノルムで使用する epsilon（小さな値）
max_position_embeddings	最大トークン長（例: 512）
type_vocab_size	セグメント ID の語彙サイズ（例: 2 → A文/B文）
pad_token_id	パディングトークンの ID
attention_probs_dropout_prob	アテンション確率のドロップアウト率
hidden_dropout_prob	隠れ層出力のドロップアウト率
use_cache	キャッシュを使用するかどうか（高速化用）
position_embedding_type	位置埋め込みの種類（例: "absolute"）
gradient_checkpointing	GPUメモリ節約のためのチェックポイント機能の有効化
transformers_version	作成に使われた transformers ライブラリのバージョン

今後の研究指針

| config.jsonに入っている情報から，ルーティングに使用できそうな情報の選定.

| LLMの情報を違う手法で収集できるか.

| 選んできたモデルに対しての性能の評価.

