

GENIAC 中間報告

ストックマーク株式会社

近江崇宏



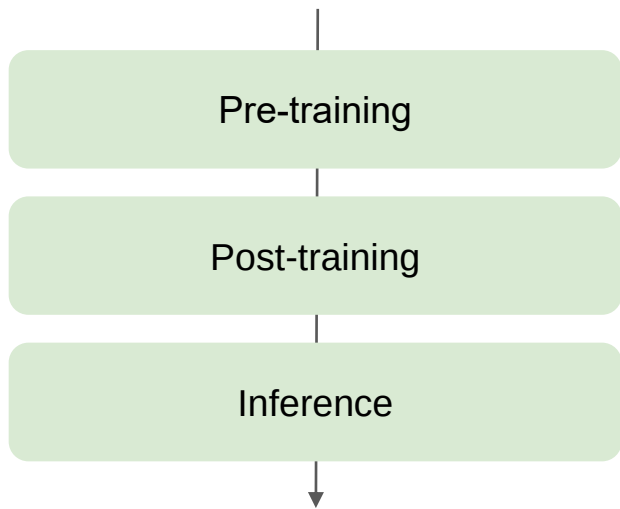


概要

開発テーマ：厳密さが要求される**ビジネス**用途における**ハルシネーション**を大幅抑止した基盤モデル

ステージ毎にハルシネーション抑制のための施策を行う。事前学習では、ドメイン知識を獲得することで、そのドメインの質問回答における正確性の向上を目指す。

LLM life cycle



2024年2月中旬~4月中旬

- 48 nodeのa3インスタンスで、1000億パラメータのLLMの事前学習を行う。
- 公開されているデータセットだけではなく、特許や弊社が独自に収集しているビジネスドメインのデータセットを含めて学習し、ビジネスのドメイン知識の獲得を行う。



学習：やったこと

- 48ノードのa3インスタンスで、1000億パラメータ規模のLlama 2のモデルを用いてゼロから事前学習をおこなった。（環境：GPUDirect-TCPX + COS OS, Megatron-LM）
- 日本語のデータセットはビジネスドメインのデータを含めて3050億トークン。英語はRedPajamaのデータセット（Books除く）からサンプルした。
- 最終的には日本語・英語合わせて9100億トークンを学習した（計画は1兆トークン）。最初の6000億トークンは英語3, 日本語1の割合、残りの3100億トークンは英語1, 日本語9の割合で学習をおこなった。

日本語データの構成

corpus	tokens after preprocessing
Stockmark Web Corpus (This dataset will not be released)	8.8 billion
Patent	37.5 billion
Wikipedia	1.5 billion
mC4	52.6 billion
CommonCrawl (snapshot: 2020-50 ~ 2024-10)	203.7 billion

ビジネスドメイン



学習：課題など

- 前半での学習の進み方（lossの減少）が想定よりも遅かった。
 - 後半は日本語の割合を多くして学習をおこなった。
 - 日本語と英語のデータの混合割合、ハイパーパラメータ、学習計画をどう決めるか？
- 途中でプロセスがエラーで止まってしまうことが多かった。
 - 期間全体で18回発生した（メンテナンスは除く）。おそらく、「悪い」ノードが含まれてしまっていると思われる。
 - プロセスの停止が起こることを前提として運用を行なった。
 - プロセスが停止した場合に、それを自動で検知して復旧する仕組みの整備。
 - メンテナンス通知時に、自動的にチェックポイントを保存する仕組みの整備。
 - ノード不良で全ノードが使えない場合の対応を決めておく（問題のあるノードを特定し、それ以外のノードのみで学習を継続するなど）。
- 試行錯誤をするのが難しい。



学習：13bと100bのモデルの開発の違い

- 100bの学習ではより多くの計算資源を必要とする。計算ノードが増えるとノード間通信がボトルネックとなりうる。
 - 今回は、GPUDirect-TCPXを利用し、高速なノード間通信が可能な環境を利用した。
- 100bのモデルだとcheckpoint (optimizer stateなどを含む)のサイズが1.3TBほどあり、とても大きいため、ディスクのスループットもより重要になる。
 - 今回はNFSとしてFilestoreを使用。容量が大きくするとスループットが上がるので、大きな容量（80TB）を設定した。この場合のcheckpointの保存時間は8-10分ほど。



評価

ドメイン知識の評価：

時事問題、企業動向、社会課題、技術トレンドなどに関する50題のQA（Stockmark business questions）により評価。

モデル	正解率
stockmark-100b	0.90
stockmark-13b	0.80
GPT-4-turbo	0.64
GPT-3.5-turbo	0.44

一般の言語理解・生成能力の評価：

Japanese Vicuna QA Benchmarkにより評価。

モデル	平均スコア
stockmark-100b	6.0
Swallow-70b-instruct-hf	5.6
GPT-3.5 (text-davinci-003)	5.1

* 計算やコーディングを要する問題を除く60題で評価

今後、指示学習での指示追従性の改善も引き続き試みる。



モデルの公開 (5/16)

事前学習モデル：<https://huggingface.co/stockmark/stockmark-100b>

指示学習モデル：<https://huggingface.co/stockmark/stockmark-100b-instruct-v0.1>

デモ：<https://huggingface.co/spaces/stockmark/stockmark-LLM>

