7A-02

カープローブの大規模シンセティックデータの生成と共有

水野　貴之†　　藤本　祥二‡　　石川　温‡

国立情報学研究所†　　金沢学院大学経済情報学部‡

はじめに

人々の位置情報履歴（連続的に記録された位置情報）は個人情報に該当するため，「災害対策，テロ対策，公衆衛生，地域差別，マーケティング，混雑対策」に重要だけれども，研究者が容易にアクセスできるデータが十分に存在しない．この問題を解決する１つの手段が，本物そっくりのシンセティック（フェイク）移動軌跡データを人工知能で生成して公開することである．本研究では，大規模言語モデル GPT アーキテクチャを利用して，日本全国を網羅する自動車の移動軌跡の大規模シンセティックデータを作成し，公開する．位置情報を空間的な絶対座標と相対座標を表す文字情報に変換することで，許諾の取れたカープローブのデータを GPT アーキテクチャで学習することができる[Mizuno 2022, Fujimoto2023, Horikomi2023]．学習後のモデルが生成する移動軌跡は，本物の移動軌跡と同じ特徴を示す．

学習用データ

日本全国の自動車の移動軌跡の生成モデルの学習には，パイオニア株式会社が提供する，2014 年 7 月の走行履歴データ(1 日あたり約 7 万台)を利用する．このデータは，パイオニア製カーナビから収集されたデータであり，各移動に関して，「カーナビ ID，出発日時，出発地，到着日時，到着地，走行距離」等が収録されている．プライバシー処理として，日毎にカーナビ ID は変更になる．また，出発地と到着地は，一辺の長さが約 500m の２分の１地域メッシュコードで抽象化されている．

我々は，出発地と到着地に関して，それぞれ対応するメッシュ内の座標をランダムに選び，仮想的な出発座標と到着座標を設定した．

次に，我々が OSS ルーティングエンジン GraphHopper に，出発座標と到着座標を入力し，出発座標から到着座標までの５パターンの移動経路とその移動時間を算出した．そして，オリジナルの走行距離と移動時間に最も近いパターンを１つ選び，それを仮想的な移動軌跡とした．

上記のようにして，我々は，日本全国１ヶ月分の延べ 217 万台の自動車の移動軌跡を作成した．これらの移動軌跡を GPT アーキテクチャで学習する．

モデル

我々は，EleutherAI が開発した自己回帰言語モデル GPT-NeoX-20B のアーキテクチャを用いて，自動車の 1 日の移動軌跡をゼロから学習する．GPT-NeoX-20B は，BERT と同様に，Self-attention 層と Projection 層からなる複数の Transformer 層で構成されている．そのパラメータ数は，GPT-3 に匹敵する 200 億パラメータである．GPT-NeoX は，Transformer 層で処理する位置より前の入力トークン列のみを参照し，前のトークンから次のトークン，すなわち，過去の位置から次の位置を逐次予測することができる．

移動時間間隔と移動場所を表す文字変数

移動軌跡は，移動時間間隔と位置座標の時系列として表現できる．しかし，GPT-NeoX のような言語モデルは，時間や座標などの数値の学習にはあまり向いていない．そこで，我々は，これらの値を一意な文字に変換することを選択した．まず，時間間隔Δtを$\tau = int(\log \Delta t)$として離散化し，離散化した時間間隔$\tau$に一意な文字を割り当てる．

$$r(\tau) \in \{R|\text{離散時間間隔を表すユニークな文字}\}.$$

次に，日本の地域メッシュコード JIS X 0410 を応用して，緯度・経度で表された座標を再帰的に空間を細分化する固有の文字に変換する．例えば，”$\zeta_1\zeta_2\zeta_3\zeta_4\zeta_5$”は 250m の解像度での領域を表現する．ここで，文字変数ζ_iは，JIS X 0410 の i 次メッシュコードに対応する．ζ_1は緯度 40 分，経度 1 度の正方形で囲まれた一意の領域を表す．

Generating and Sharing Synthetic Big Data for Car Probes
† Takayuki Mizuno, National Institute of Informatics
‡ Shouji Fujimoto, Kanazawa Gakuin University
‡ Atushi Ishikawa, Kanazawa Gakuin University

日本の全陸地は 176 個の 1 次メッシュコードで表すことができる．ζ_2は１次メッシュを緯度・経度方向に，それぞれ 8 等分した領域を，ζ_3は 2 次メッシュを緯度・経度方向に，それぞれ 10 等分した領域を示す．それ以降の分割は再帰的に緯度・経度ともに 2 等分され，それぞれのζ_iに固有の文字が割り当てられる．

$$\zeta_i \in \{Z_i | \textit{領域を表すユニークな文字}\},$$

ここで，$R \cap Z_i = \emptyset$ と $Z_i \cap Z_j = \emptyset \ (i \neq j)$である．250m の解像度での日本の全陸地は，わずか 348 文字$(= 176 + 8^2 + 10^2 + 2^2 + 2^2)$の組み合わせで表現することができる．

移動軌跡の表現

各自動車の１日の移動軌跡は，移動時間間隔と移動中の位置の座標に，それぞれ割り当てられた文字rとXによって，以下のように表現できる．

$$X(t_0)\_r(\tau_1)X(t_0 + \Delta t_1)\_r(\tau_2)X\left(t_0 + \sum_{k=1}^{2} \Delta t_k\right)\_\cdots$$

ここで，t_0は与えられた日における最初の時刻，Δt_kとτ_kは，移動中の$(k-1)$番目の位置からk番目の位置までの，移動時間間隔と離散化した移動時間間隔を，それぞれ表す．領域を表す文字Xは，250m の解像度の場合，$X(t) = \zeta_1(t)\zeta_2(t)\zeta_3(t)\zeta_4(t)\zeta_5(t)$となる．一日の終りにはピリオドトークンを"."を加える．前の位置と次の位置を"_"で接続して移動軌跡を表す．このテキスト形式で表現された自動車の１日の移動軌跡を GPT-NeoX-20B のアーキテクチャで学習する．

移動軌跡のシンセティックデータ

図は，モデルにより生成されたある平日の大都市圏における自動車の分次の移動軌跡の，朝７時点のスナップショットである．各黒色の点が自動車であり，本物とそっくりに移動する．

我々は，単位時間あたりの移動距離分布，１回の移動における移動距離分布と移動時間分布，移動距離と移動時間の関係，移動距離の自己相関，移動距離と移動角度の関係，出発地点に戻ってくる再起確率など，移動における時空間の統計的な特徴を，オリジナルとモデルにより生成されたシンセティックデータとで比較した．

移動距離分布は裾野が厚い，移動距離と移動時間には非線形性がある．移動距離には正の自己相関が，長距離移動ほど直線的な移動パターンを示す．一日の終わりに近づくほど，その日の出発地付近に戻ってくる．このような統計性がシンセティックデータでも高精度に再現される．

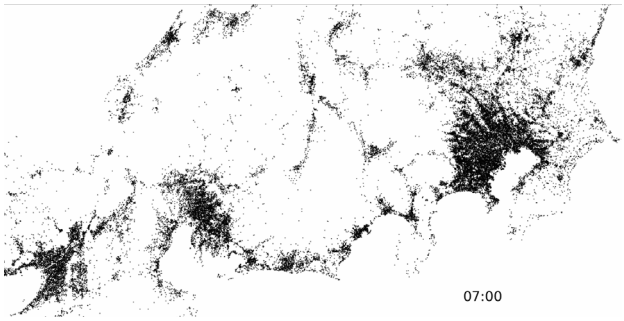

図　自動車の移動軌跡のシンセティックデータのスナップショット

まとめ

GPT-NeoX-20B アーキテクチャを用いて自動車の移動軌跡を生成した．移動時間間隔と移動場所を，ユニークな文字に変換することで，数値情報を学習することが得意ではない GPT-NeoX-20B でも，移動軌跡を十分に学習できる．生成された移動軌跡は，現実の移動軌跡の特徴を再現している．このモデルにより生成された高精度のシンセティック移動軌跡データは，ジオプライバシーを保護しながら，災害，テロ，治安，感染症，空間的分断，マーケティング，交通渋滞などの社会問題に取り組むことに貢献できる．

本講演では，天気や気温，曜日などの環境情報などをトークン化して GPT-NeoX-20B に加えることにより，より精度の高い環境に合わせた自動車の移動軌跡が生成できることも合わせて報告する．また，これらのジンセティックデータをオープンデータとして公開する．

参考文献

[Mizuno 2022] T. Mizuno, S. Fujimoto, A. Ishikawa (2022) Front. Phys. 10, 1021176.
[Fujimoto 2023] S. Fujimoto, A. Ishikawa, T. Mizuo (2023) Proceedings of WI-IAT.
[Horikomi 2023] T. Horikomi, S. Fujimoto, A. Ishikawa, T. Mizuno (2023) arXiv:2308.07940.