2025/7/16　輪講グループ1

B4 野村卓矢

VectorDB-Enabled Transfer-Learning for Time-Series Forecasting

「ベクトルデータベースによる時系列予測のための転移学習」

Alessandro Falcetta, Giulio Cristofaro, Lorenzo Epifani, Manuel Roveri

AILMSystems ‘24: Proceedings of the 4th International Conference on AI-ML Systems
Article No.: 18, Pages 1 - 9

はじめに

- 時系列予測は機械学習分野における伝統的な課題であり、金融、医療、交通など、多くの分野で使われている。
- 様々なディープラーニングアーキテクチャが提案されているが、大量の学習データの必要性から実用的ではない。
- 転移学習が注目を集めているが、転移用のソースデータを特定することは容易ではない。

→そこで、新たな転移学習へのアプローチを提案する。

研究目的

- ベクトルデータベースを用いて各ドメインのベクトル類似度検索とアンサンブル学習モデルを活用した、時系列予測における新たな転移学習モデル「VectorDB Enabled Transfer-learning for Time-series Forecasting (VETT)」を提案する。

- 提案手法の効果性と効率性を多用なドメインのデータセットにて検証する。

関連研究

- 時系列予測について
- 過去の値の集合を学習モデルにかけて将来の値を推定する。
- 本研究ではRlinearをベースモデルとする。【17】
- 可逆正規化層【14】と線形射影を使用したモデル。
- 他の複雑なモデルと同等の性能を持ちつつ、より少ないデータでの訓練が可能で、少ないデータでも効果的な学習が可能。

関連研究

- ベクトルデータベースについて【９】
- データをベクトルとして保存し、以下の特性を持つ。
- 類似度検索
- 複雑で非構造的なデータへの対応
- スケーラビリティと性能

関連研究

- データが不足している際の時系列予測について
- DTWやWasserstein距離での類似判定、LSTMや時間分割手法の活用例

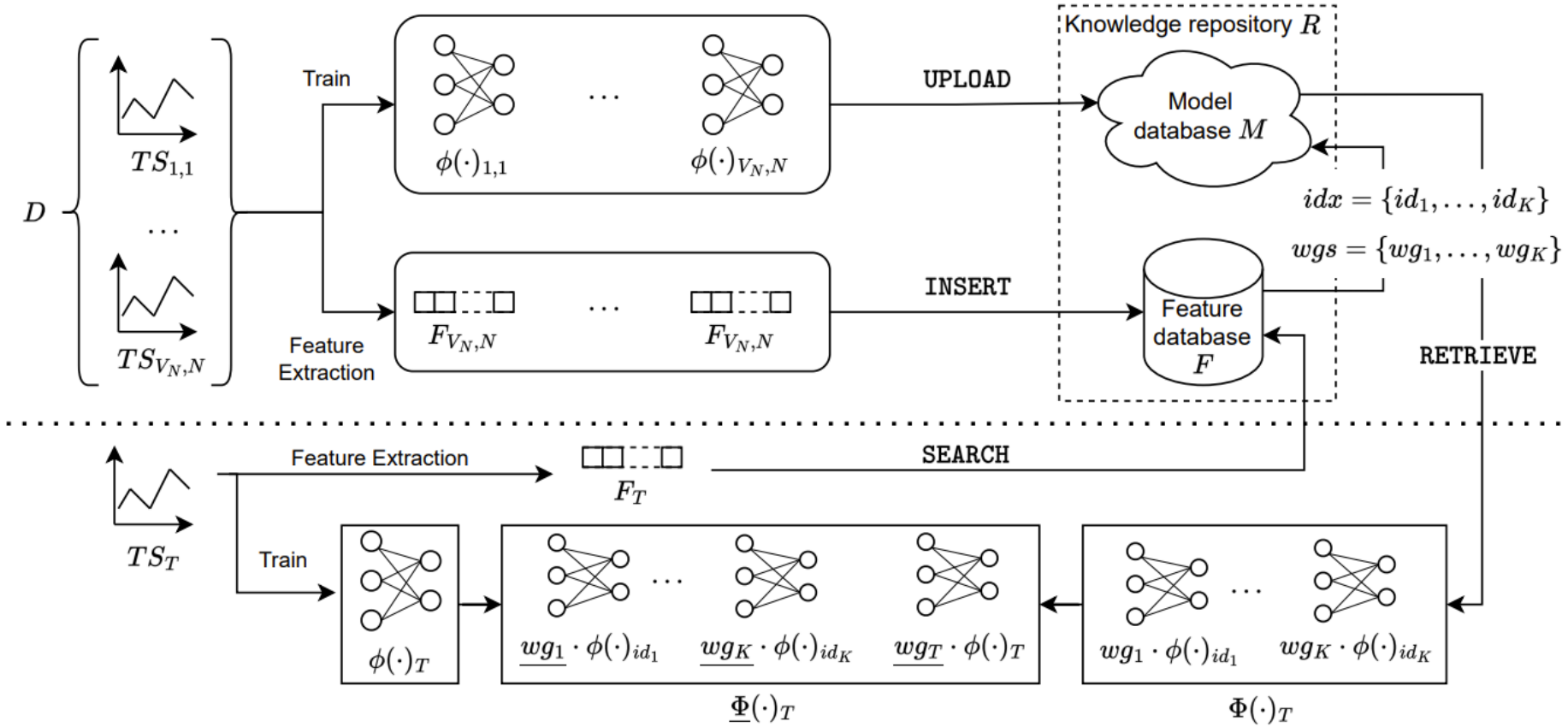
→異なる分野に拡張できない。大規模なナレッジベースの活用が困難。

- エネルギー生産予測分野のファインチューニングを用いるものや、異なる季節における転移学習を用いる例、ドメイン非依存のあらゆる時系列に適用可能な例

→単一ソースの利用のみで、複数のソースを活用しない。

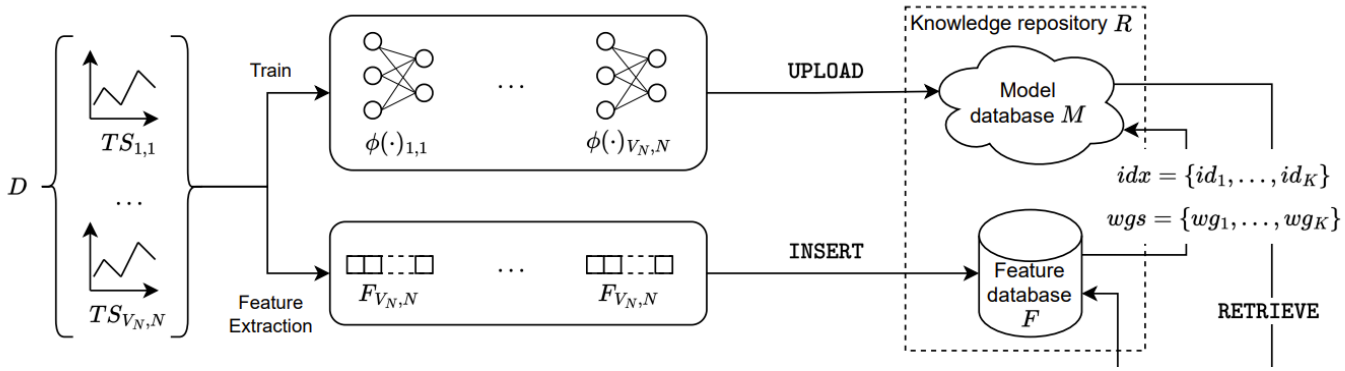
提案手法

- 主にナレッジリポジトリとアンサンブル学習で構成される。



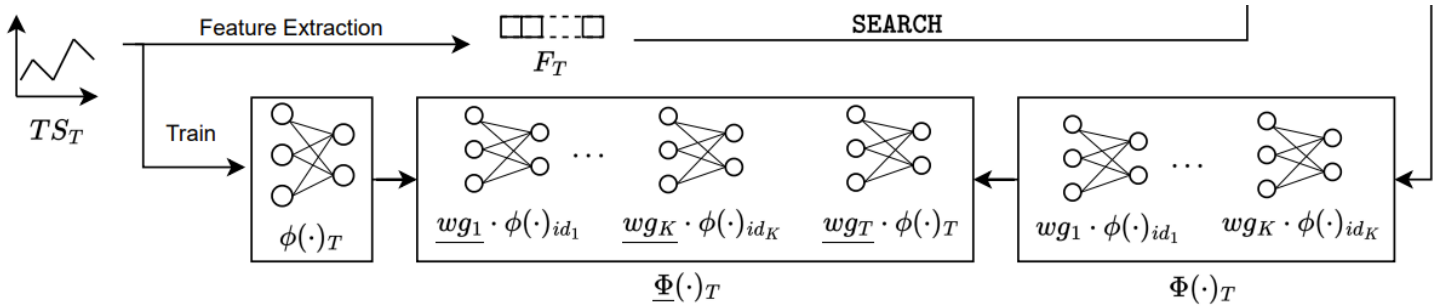
ナレッジリポジトリ

- 2つの異なるデータベースF,Mの集合R={F,M}、
- Fは各時系列から抽出された特徴量を保存する特徴データベース。(ベクトルデータベース)
- Mは各時系列に対して学習したモデル関数を保存する予測モデルデータベースである。
- 類似検索を可能とし、アンサンブル学習にモデルを提供することが目標。



アンサンブル学習

- ターゲット時系列に対して特徴抽出を行う。
- このターゲット特徴ベクトルを用いて類似検索アルゴリズムを実行し、F内で最も類似度の高いK個の時系列を特定する。
- コサイン類似度で算出される類似度スコアが用いられる。この類似度スコアはアンサンブル学習の初期重みとして利用される。
- Kの値は調整可能。



アンサンブル学習

- 検索で得られた情報を基に対応するモデルを取得し、重み付きで合成して初期アンサンブルモデルを構成する。
- このアンサンブルモデルの数エポックをベースモデルと同様の学習アルゴリズムを用いて更新するファインチューニングを行う。
- ターゲット時系列上で訓練されたローカルモデルもアンサンブルモデルに含め、それぞれの検証損失に基づいてアンサンブル重みを更新し、最終アンサンブルモデルを得る。
- 最終アンサンブルモデルにも再度ファインチューニングを実施する。

実装の詳細

- 特徴データベースにはQdrantという最先端のオープンソースベクトルデータベースを採用した。
- 高次元データの格納とインデックス作成、検索の効率性の点で本研究に適している。

- 使用する特徴量についてはPythonライブラリのtsfreshに基づいて、絶対エネルギー、自己相関関数、平均、歪度などを含む700以上の特徴を使用する。

実装の詳細

- モデルデータベースには機能はあまり必要ないのでシンプルなファイルシステムを使用した。

- 時系列予測モデルにも特別な条件は不要なため、RInearを採用した。良好な予測精度と高速な学習時間という利点を持つ。

実験

- 2つのシナリオでアンサンブル学習の有効性と効率性を検証する。
- Internalシナリオ：ターゲット時系列と同じドメインの時系列がナレッジリポジトリに存在する場合。
- Externalシナリオ：ターゲット時系列と同じドメインの時系列がナレッジリポジトリに存在しない場合。

使用データセット

- Air Quality【26】：大気質化学センサ（9種類）の1時間ごとの平均応答（11か月間）
- Energy【1】：電力の消費、発電、価格に関する20個の時系列（4年間、1時間間隔）
- Traffic【5】：高速道路における 862 個の道路占有率時系列（2015～2016年、1時間ごと）
- ETTH【36】：ETTh1 および ETTh2 の2つの電力変圧器データセットを統合したもので、油および負荷に関する14個の時系列（2016年7月～2018年7月）

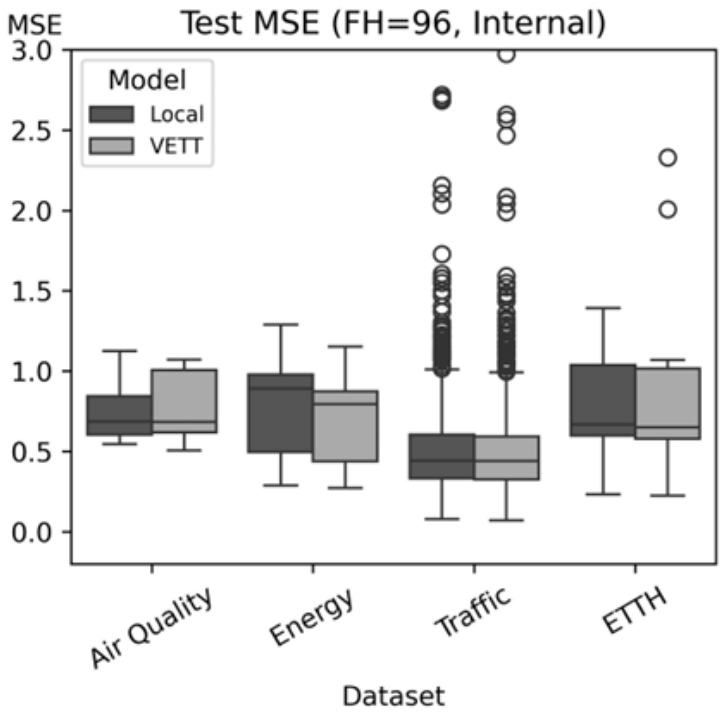
前処理と訓練設定

- データ挿入前に定数のみの時系列や外れ値を除去するデータクリーニングを実施。
- 各時系列は事件軸のはじめから80％を訓練データ、おわりの20%をテストデータとして分割。
- 入力系列長：96
- 学習率：0.001　Weight decay：0.0005
- エポック数：30
- ファインチューニングエポック数：5
- 評価指標：MSE(平均に乗誤差)
- K=5

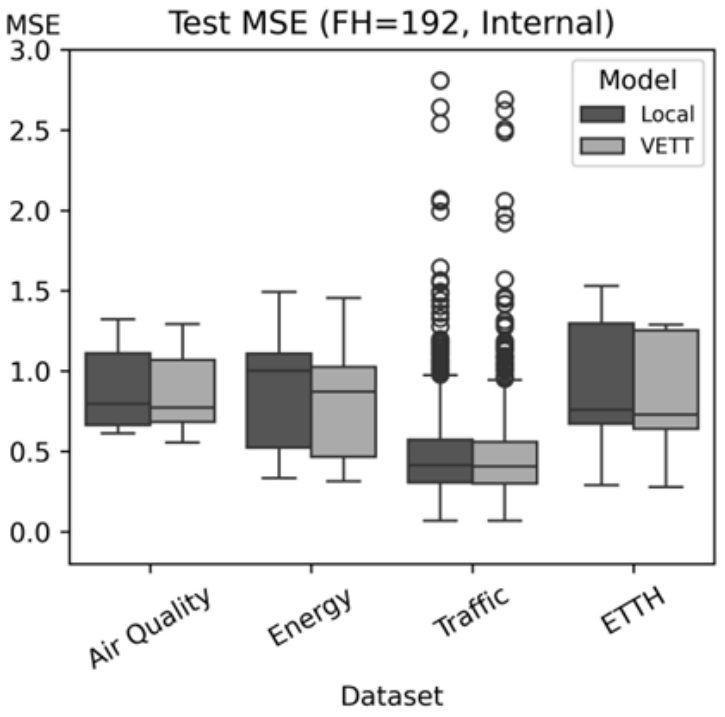
Internalシナリオ

- すべてのデータセットの集合Dを使ってナレッジリポジトリRを構築する。
- Dに含まれる時系列を1つずつターゲット時系列として取り出し、残りのすべてをアンサンブルのソースとして利用した。
- データ不足状況を再現するために各ターゲット時系列の訓練データから最後20%のみを使用した。

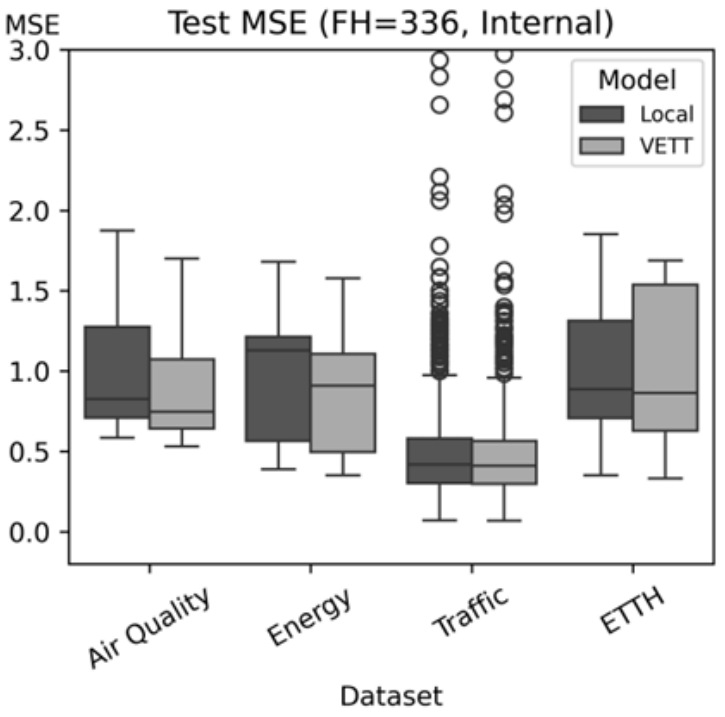
Internalシナリオ



(a) FH=96

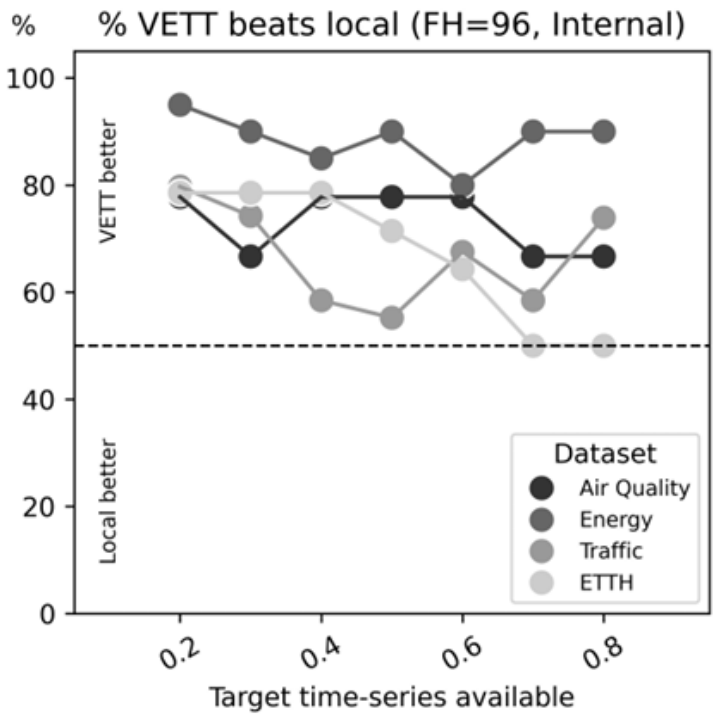


(b) FH=192

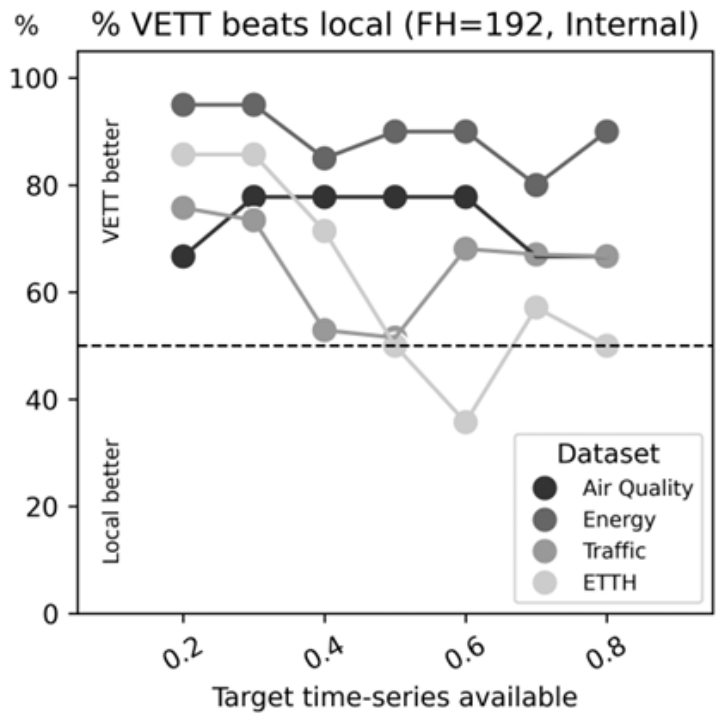


(c) FH=336

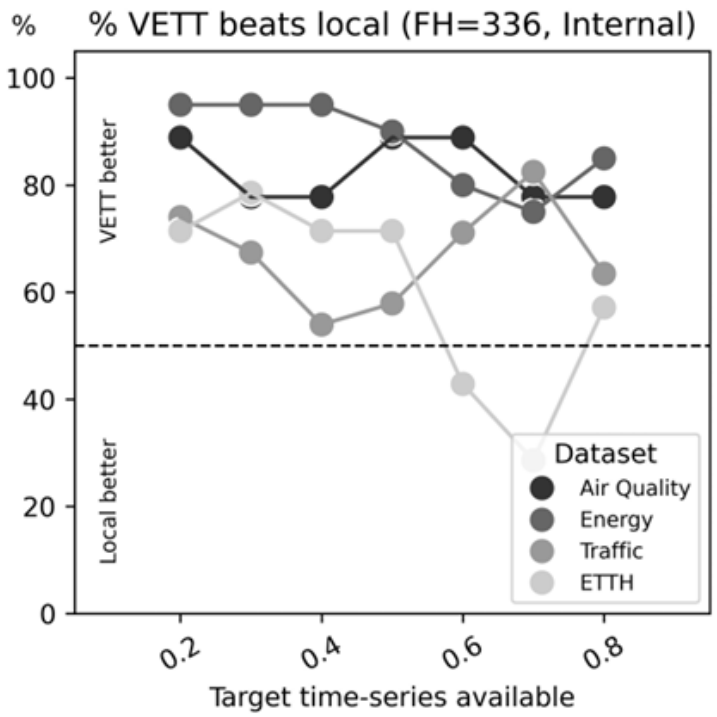
Internalシナリオ



(a) FH=96



(b) FH=192

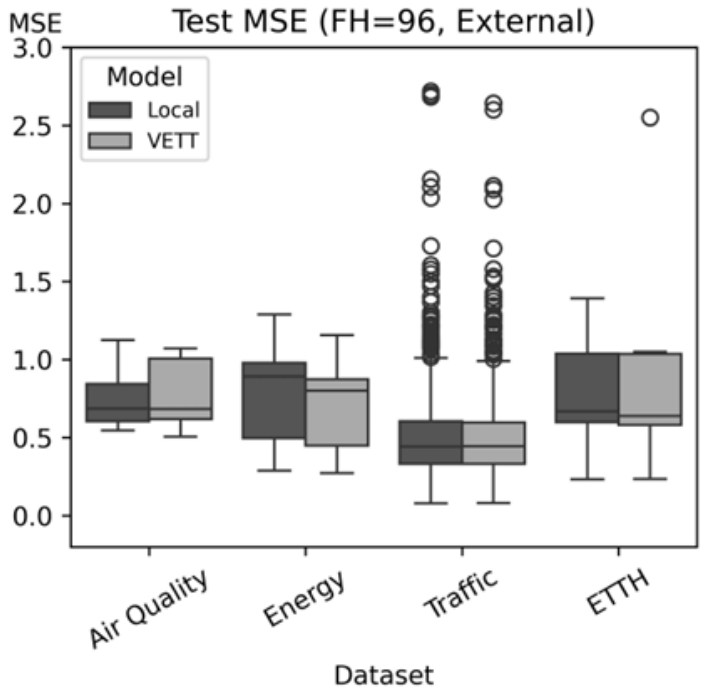


(c) FH=336

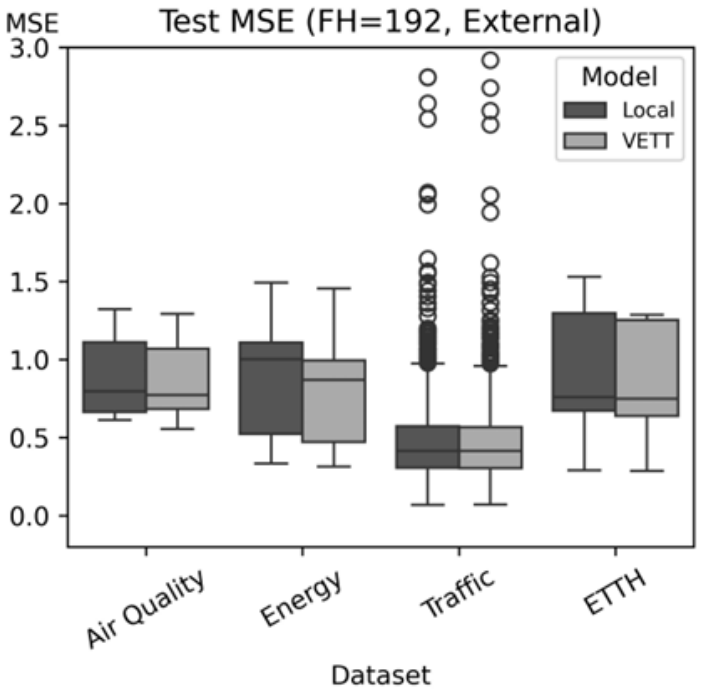
Externalシナリオ

- ナレッジリポジトリRがターゲット時系列と同じドメインの時系列を含まないという仮定で評価する。
- 検索に使用する特徴データベースFから、ターゲットドメイン全体のデータを完全に除外して行う。

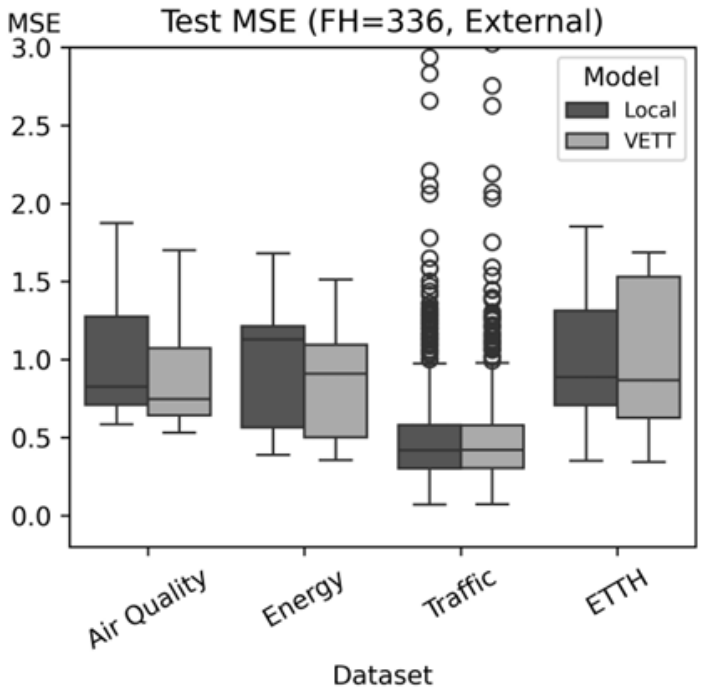
Externalシナリオ



(a) FH=96

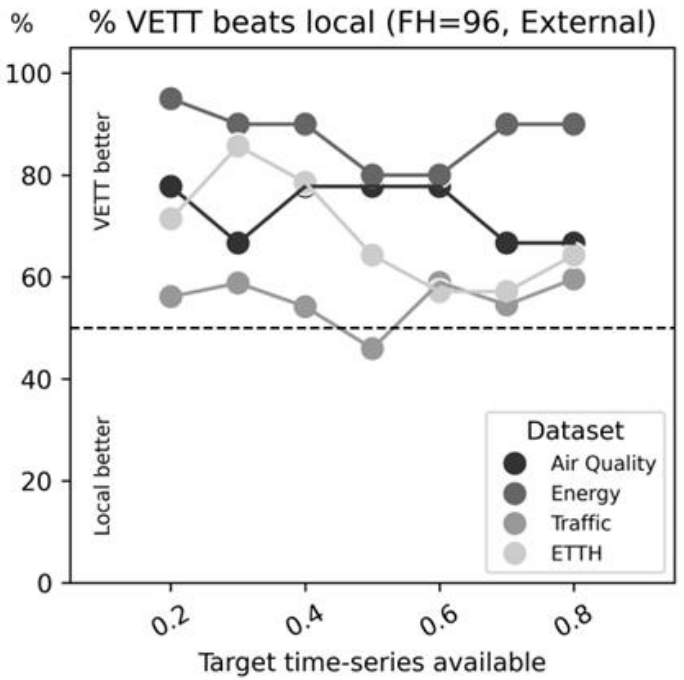


(b) FH=192

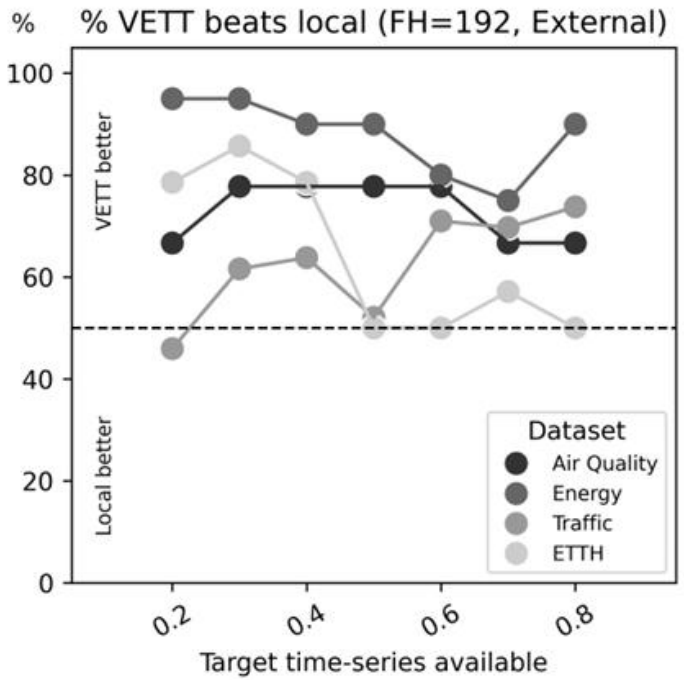


(c) FH=336

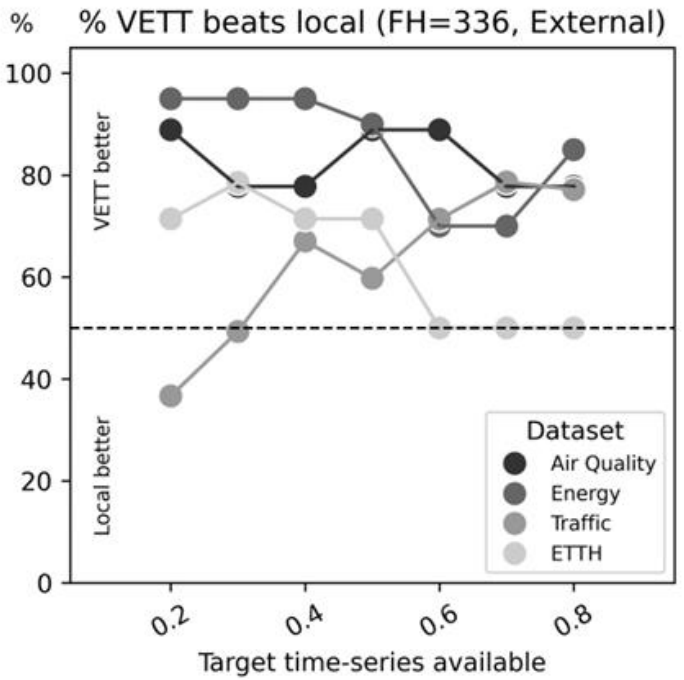
Externalシナリオ



(a) FH=96



(b) FH=192



(c) FH=336

VETTの効率性

- サービスとして提供できるほどの効率性があるか検証。
- 実行環境はクラウドで実行するものとして以下の通り。
- OS：Ubuntu 20.04 LTS
- CPU：Intel Xeon Gold 5318 S（2基）
- RAM：384 GB
- 並列ジョブ数：16
- GPU：未使用

VETTの効率性

- Total time：各データセット内のすべての時系列について、InternalシナリオとExternalシナリオ両方の推定にかかった時間。
- Time for series：1つの時系列あたりにかかった平均処理時間。

Dataset	Number of time-series	Points used	Total time (s)	Time for a series (s)
Air Quality	9	1497	25.7	2.9
Energy	20	5610	190.2	9.5
Traffic	862	2807	2558.0	3.0
ETTH	14	2787	48.7	3.5

結論

- スケーラブルかつ効果的な転移学習の時系列予測フレームワーク、VETTを提案した。
- VETTは高効率性・高速性によってクラウド環境に適している。
- 実験結果により、有効性と効率性の両立が実証された。

今後の展望

①特徴量選定の検討

類似度が高い＝知識転移可能とは限らない。時間依存の特性(周期性や傾向)なども考慮した特徴抽出へのアプローチが必要。

②時系列予測モデルの選定

より複雑で深いモデルはVETTによって大きな恩恵を受ける可能性がある。効果的なモデルをより少ないエポック数で学習可能となるためである。

今後の展望

③クラウド上での実運用と統合

VETTをクラウド上で既存の時系列予測サービスと統合することを計画している。これにはナレッジリポジトリを何百万もの時系列データで構成する必要がある。

また、ゼロショット予測の可能性も探り、これによりさらに高速な応答性能を実現できると期待する。

読後考察

- 天候予測について、本手法を工夫することで、ある地域の天候関連の時系列データを1つのドメインと考え、類似したベクトルに絞られた時系列から学習を行い、ソースドメインでファインチューニングすることで、新たな地域の天候予測の展開に短期間で取り組むことができると考えた。
- 分野間でもクロスドメインすることができるので、初めの内は膨大な量の違う分野の時系列データをソースとしながら、数年かけて段々と類似ドメインのソースデータが増えていけば精度が向上していくのではないかと考えた。