EMERGENCY DATAHACK

ИНИД

МЧС РОССИИ

Команда “Звёздочка”

Решение Трека №1

Предсказание заторов льда на реке Лена

EMERGENCY DATAHACK

ИНИД

МЧС РОССИИ

Контекст

- Лена - десятая река в мире по протяженности - более 4400 км
- Лена полностью замерзает зимой
- Бассейн реки занимает более 50% азиатской части России
- Ее ледяные заторы и сильные разливы представляют большую опасность для населения

Задача:

Предсказать образование ледяных заторов на гидропостах в период весеннего ледохода по дням

Качество предсказания оценивается по метрике F1-score

Визуализация

EMERGENCY DATAHACK

ИНИД

МЧС РОССИИ

Вызовы задачи

- Мало данных
  - ежесуточные и интервальные гидро- и метео- измерения за 35 летний период (не слишком много с точки зрения машинного обучения)
- Сложные данные
  - Много отдельных наблюдений, интуитивно связанных с задачей
  - Географически слабо связные объекты
  - Редкость заторов - целевых событий
  - Пропуски данных и ошибки измерений

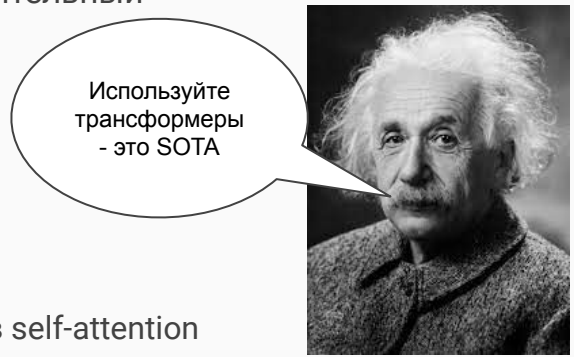
EMERGENCY DATAHACK

ИНИД

МЧС РОССИИ

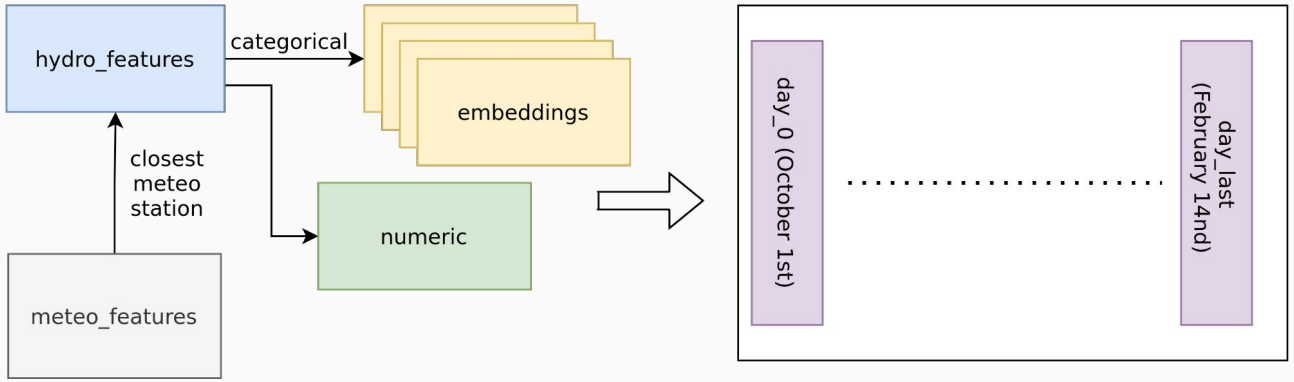
Как решали задачу

- Выявляем зависимости при помощи Deep Learning
  - Основной вызов - подготовить данные на вход NN модели
    - На вход сети подаются события за 4х месячный интервал для данного поста
  - Попробовали рекуррентную сеть (GRU)
  - Попробовали географическую связность гидропостов как дополнительный признак, но это не дало прироста качества
  - **Остановились на двойном трансформере на основе Albert**
- В чем преимущества подхода
  - Сложные формульные **эвристики** модель выводит **сама**
  - **Модульность** и **гибкость** нейросетевых архитектур
  - Способность работать с данными **разной** природы (снимки, ряды)
  - **Интерпретация** важности признаков за счет встроенных механизмов self-attention



Работа с данными

- Для каждого гидропоста добавляются наблюдения с ближайшего метеопоста
- Для каждого категориального признака создается эмбеддинг
- Оценка качества данных используется как категориальный признак
- Модель принимает последовательность признаков за определенный период (139 дней)



EMERGENCY DATAHACK

ИНИД

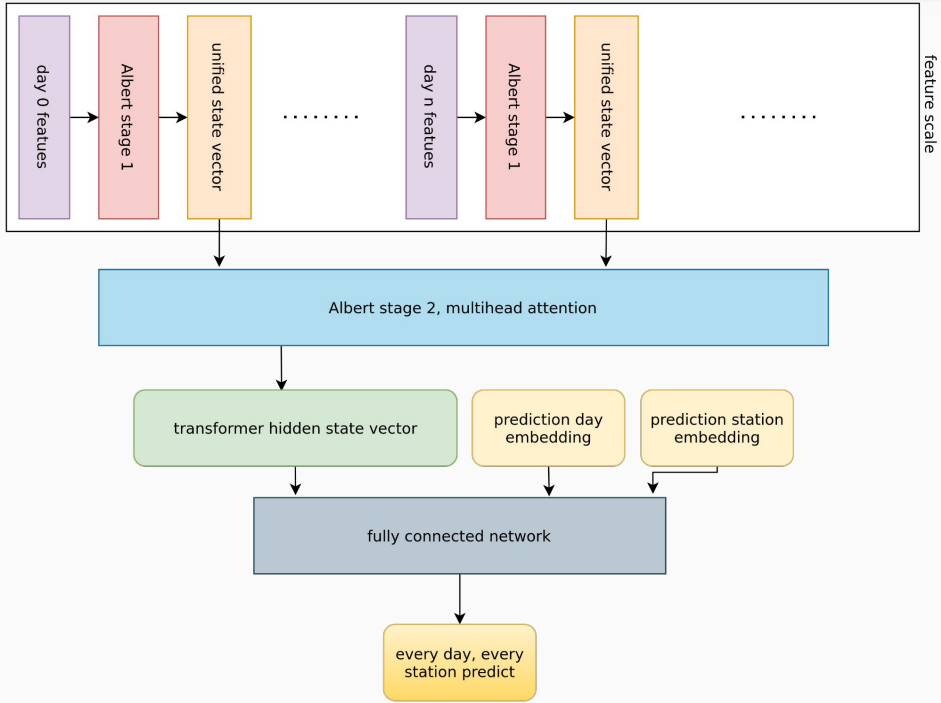
МЧС РОССИИ

Модель

LenaBiTrans
трансформер в трансформере

- Модель принимает последовательность признаков за определенный период (139 дней)
- Внутренний трансформер кодирует информацию по оси признаков
- Внешний трансформер кодирует информацию по оси времени

Такая архитектура осуществляет регуляризацию и позволяет значительно снизить переобучение. Вдохновлена публикацией [Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting](#).



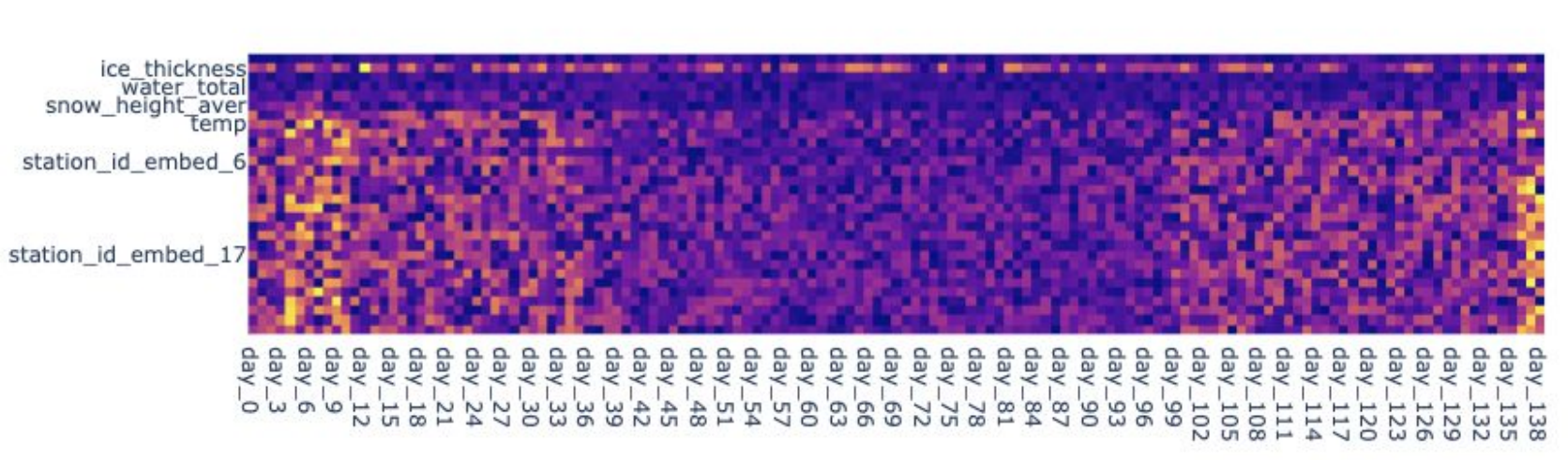
EMERGENCY DATAHACK

ИНИД

МЧС РОССИИ

Что подсказала модель

- Самые важные периоды для предсказания заторов - ноябрь и февраль
- Толщина льда - самый важный признак
- Тепловые признаки из прошлого для долгосрочного прогноза - второстепенная роль



EMERGENCY DATAHACK

ИНИД

МЧС РОССИИ

Выводы

- Модель легко расширять и изменять
- Анализ важности признаков подтверждает обоснованность применения существующих эмпирических признаков и подходов
- Результат можно существенно улучшить при помощи self-ensemble
- Потенциал для роста при использовании признаков теплового баланса

EMERGENCY DATAHACK

ИНИД

МЧС РОССИИ

Сделано с ❤ на открытом ПО

Poetry

pre-commit

[Репозиторий с кодом решения]

СБЕР МАРКЕТ

Команда “Звездочка”

Сергей
Lead ML

Даниил
ML

Глеб
Lead ML