МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

ОТЧЕТ

**Домашнее задание №\_\_1\_\_**

по дисциплине«Методы машинного обучения»

Тема: «Машинный перевод»

ИСПОЛНИТЕЛЬ: \_\_\_\_\_Фань Лицзе\_\_\_

ФИО

группа ИУ5И-\_21М\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись

"\_\_"\_\_\_\_\_\_\_\_\_202\_ г.

ПРЕПОДАВАТЕЛЬ: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ФИО

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись

"\_\_"\_\_\_\_\_\_\_\_\_202\_ г.

Москва - 2024

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Задание

Домашнее задание по дисциплине направлено на анализ современных методов машинного обучения и их применение для решения практических задач. Домашнее задание включает три основных этапа:

1. выбор задачи;
2. теоретический этап;
3. практический этап.

Этап выбора задачи предполагает анализ ресурса [paperswithcode.](https://paperswithcode.com/sota) Данный ресурс включает описание нескольких тысяч современных задач в области машинного обучения. Каждое описание задачи содержит ссылки на наиболее современные и актуальные научные статьи, предназначенные для решения задачи (список статей регулярно обновляется авторами ресурса). Каждое описание статьи содержит ссылку на репозиторий с открытым исходным кодом, реализующим представленные в статье эксперименты. На этапе выбора задачи обучающийся выбирает одну из задач машинного обучения, описание которой содержит ссылки на статьи и репозитории с исходным кодом.

Теоретический этап включает проработку как минимум двух статей, относящихся к выбранной задаче. Результаты проработки обучающийся излагает в теоретической части отчета по домашнему заданию, которая может включать:

1.описание общих подходов к решению задачи;

2.конкретные топологии нейронных сетей, нейросетевых ансамблей или других моделей машинного обучения, предназначенных для решения задачи;

3.математическое описание, алгоритмы функционирования, особенности обучения используемых для решения задачи нейронных сетей, нейросетевых ансамблей или других моделей машинного обучения;

4.описание наборов данных, используемых для обучения моделей;

5.оценка качества решения задачи, описание метрик качества и их значений;

6.предложения обучающегося по улучшению качества решения задачи.

Практический этап включает повторение экспериментов авторов статей на основе представленных авторами репозиториев с исходным кодом и возможное улучшение обучающимися полученных результатов. Результаты проработки обучающийся излагает в практической части отчета по домашнему заданию, которая может включать:

1.исходные коды программ, представленные авторами статей, результаты документирования программ обучающимися с использованием диаграмм UML, путем визуализации топологий нейронных сетей и другими способами;

2.результаты выполнения программ, вычисление значений для описанных в статьях метрик качества, выводы обучающегося о воспроизводимости экспериментов авторов статей и соответствии практических экспериментов теоретическим материалам статей;

3.предложения обучающегося по возможным улучшениям решения задачи, результаты практических экспериментов (исходные коды, документация) по возможному улучшению решения задачи.

## Выбор задачи

Time Series Forecasting is the task of fitting a model to historical, time-stamped data in order to predict future values. Traditional approaches include moving average, exponential smoothing, and ARIMA, though models as various as RNNs, Transformers, or XGBoost can also be applied. The most popular benchmark is the ETTh1 dataset. Models are typically evaluated using the Mean Square Error (MSE) or Root Mean Square Error (RMSE).

## Теоретический этап

Я выбрал две следующие статьи:

"Sequence to Sequence Learning with Neural Networks" и "Probabilistic Forecasting with Temporal Convolutional Neural Network"

### 3.1 Распространенные решения

При прогнозировании временных рядов общие решения включают прогнозирование на основе статистических методов (таких как ARIMA), методов машинного обучения (таких как деревья решений, случайные леса), методов глубокого обучения (таких как RNN, LSTM), методов на основе CNN и ансамблевых методов. методы

### 3.2 Конкретные топологии нейронных сетей

1.**Sequence to Sequence Learning with Neural Networks (Seq2Seq)**:

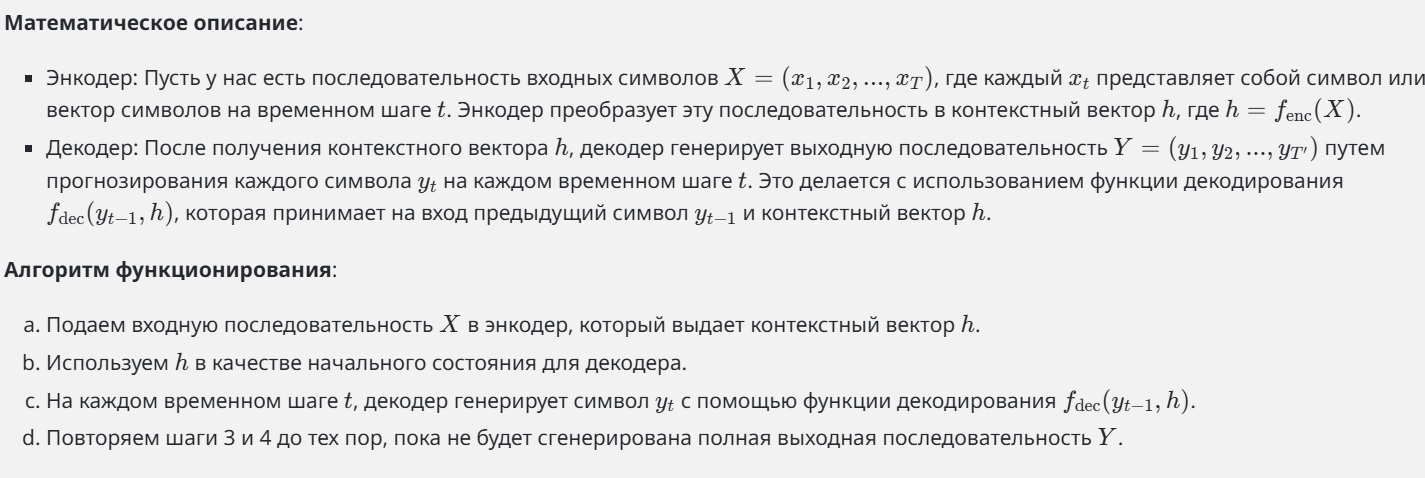
* Энкодер (Encoder): Обычно состоит из ряда слоев рекуррентных нейронных сетей (RNN), таких как долгая краткосрочная память (LSTM) или gated recurrent unit (GRU). Энкодер принимает на вход последовательность и преобразует её в контекстный вектор или скрытое состояние, чтобы захватить семантическую информацию входной последовательности.Эмбеддинги: Чтобы улучшить обработку категориальных переменных (например, тип продукта или магазин), используются эмбеддинги, которые преобразуют эти переменные в плотные векторы небольшой размерности.
* Декодер (Decoder): Также состоит из ряда слоев RNN, принимает на вход контекстный вектор (или скрытое состояние), сгенерированный энкодером. Декодер поэтапно генерирует каждый элемент целевой последовательности. На каждом временном шаге декодер создает метку, представляющую следующий элемент целевой последовательности.
* Можно использовать механизм внимания (Attention Mechanism), чтобы повысить производительность модели, помогая модели лучше сфокусироваться на входной последовательности в процессе декодирования.

2.**Probabilistic Forecasting with Temporal Convolutional Neural Network (TCN)**

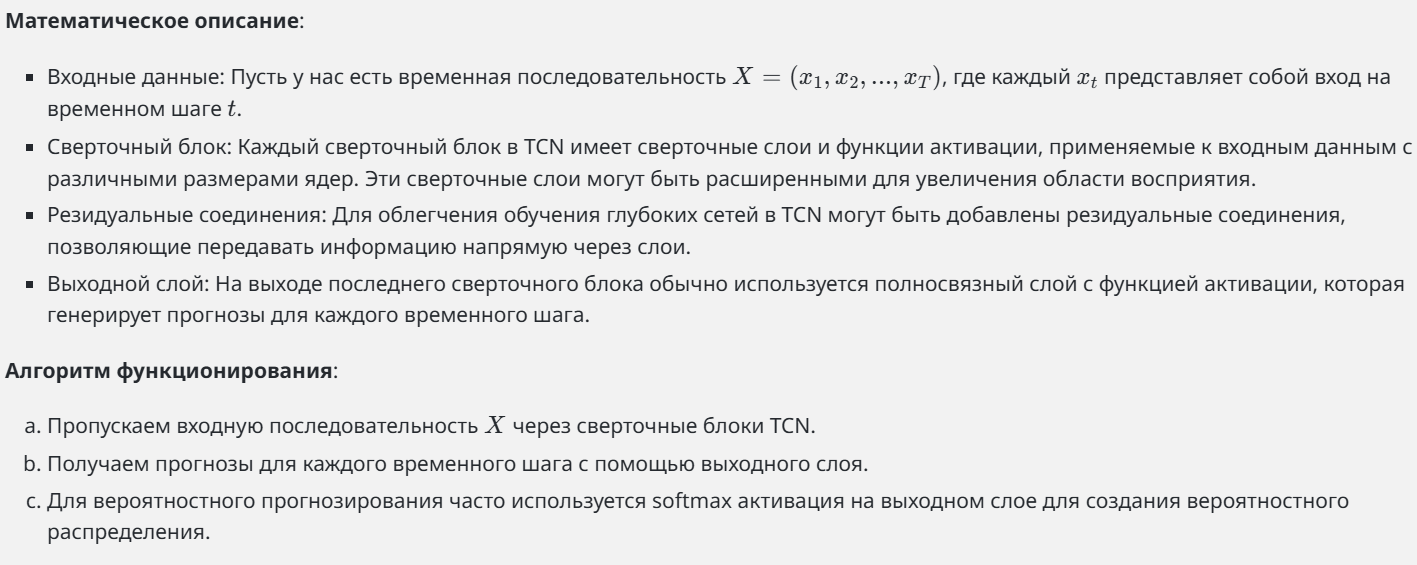
* Сверточные блоки (Convolutional Blocks): TCN обычно состоит из нескольких сверточных блоков, каждый из которых состоит из ряда сверточных слоев и нелинейной функции активации (например, ReLU). Эти сверточные слои используются для захвата локальных шаблонов и долгосрочных зависимостей во временных последовательностях.
* Опциональные связи остатков (Optional Residual Connections): Для более эффективного обучения глубоких TCN можно добавить связи остатков, добавляя вход каждого сверточного блока к его выходу напрямую, аналогично структуре сети с остаточными блоками (ResNet).
* Опциональные расширенные свертки (Optional Dilated Convolutions): Для увеличения области восприятия модели можно использовать расширенные свертки (dilated convolutions), позволяющие сверточным слоям увеличить диапазон получаемой информации, не увеличивая при этом количество параметров.
* Выходной слой (Output Layer): Выход последнего сверточного блока обычно подключается к полносвязному слою, который генерирует конечные прогнозы. В задачах вероятностного прогнозирования обычно используется соответствующая функция активации (например, softmax) для создания вероятностного распределения для каждого временного шага.

### 3.3 Математическое описание и алгоритмы функционирования

1. **Seq2Seq**



1. **TCN**



### 3.4 Описание наборов данных

Исследования обычно используют общедоступные наборы данных для временных рядов, такие как классификационные/регрессионные данные временных рядов UCI, данные соревнований Kaggle и т. д. Выбор набора данных должен базироваться на конкретных потребностях задачи прогнозирования.

### 3.5 Оценка качества решения задачи

Оценка качества решения задачи прогнозирования временных рядов включает в себя использование метрик, таких как средняя абсолютная ошибка (MAE), средняя квадратичная ошибка (MSE), корень из средней квадратичной ошибки (RMSE), средняя абсолютная процентная ошибка (MAPE) и коэффициент детерминации . Эти метрики помогают оценить точность и эффективность модели прогнозирования, что является важным для принятия информированных решений в бизнесе и науке.

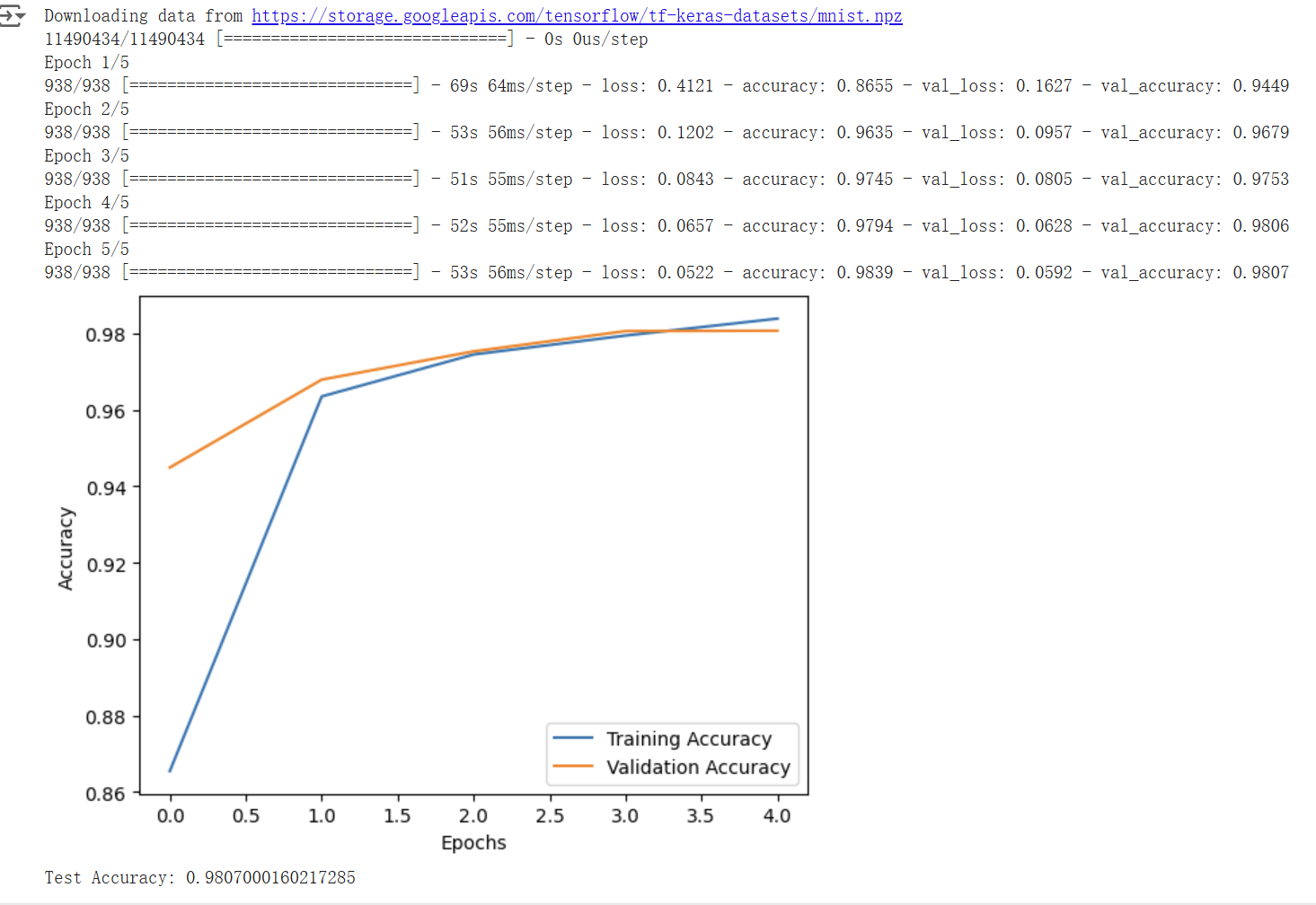
### 3.6 Предложения по улучшению качества решения задачи

Чтобы повысить точность и эффективность прогнозирования временных рядов, можно использовать несколько стратегий. Во-первых, важно правильно выбрать модель. Рассмотрите возможность использования сложных моделей с хорошей производительностью, таких как сети с краткосрочной памятью (LSTM) или сверточные нейронные сети (CNN). Во-вторых, в дополнение к самим данным временных рядов вы также можете рассмотреть возможность введения других соответствующих функций, таких как сезонность, тенденции и внешние факторы, чтобы улучшить прогнозирующую способность модели. Во время обучения модели для достижения оптимальной производительности необходимо тщательно настроить гиперпараметры, такие как скорость обучения, размер пакета и количество итераций. Кроме того, вы можете попробовать использовать методы интеграции моделей, такие как интеграция результатов прогнозирования нескольких базовых моделей, чтобы еще больше повысить точность прогнозирования.

## Практический этап

**Seq2Seq**

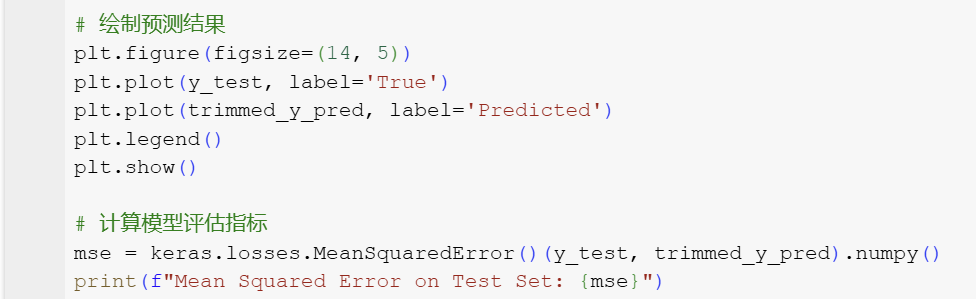


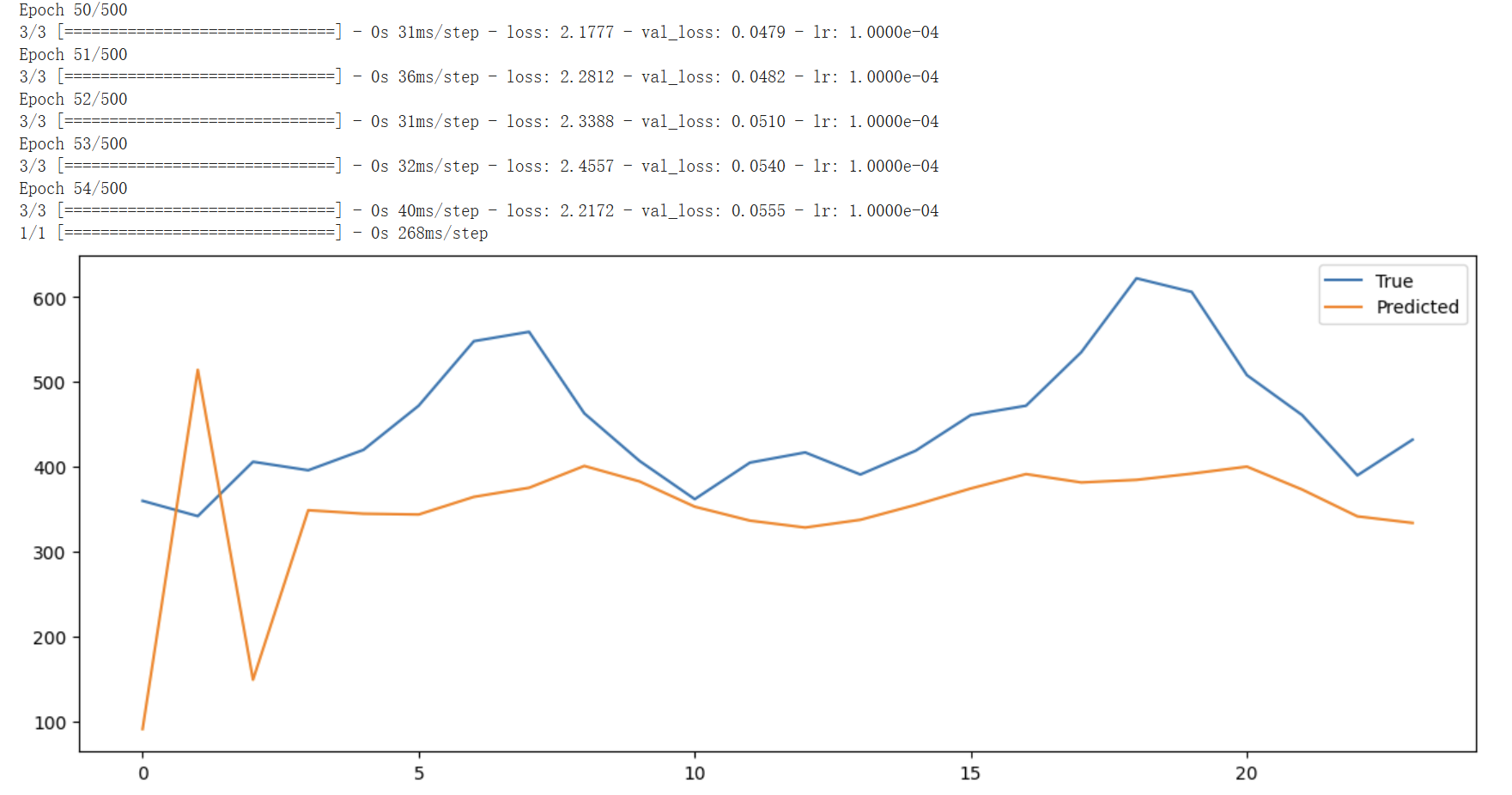


**TCN**









**Вывод**

В этом эксперименте мы рассмотрели два разных метода для выполнения задачи прогнозирования временных рядов, используя Seq2Seq и TCN. Сравнивая эти два метода, мы можем получить более глубокое понимание их эффективности и применимости для обработки временных рядов.

1. Обзор методов

**Seq2Seq**: Обучение последовательности на последовательности с нейронными сетями - это метод машинного обучения, основанный на нейронных сетях, который используется для перевода одной последовательности в другую. Он состоит из двух основных компонентов: кодера и декодера. Кодер преобразует входную последовательность в фиксированный вектор фиксированной длины, который затем передается в декодер для генерации выходной последовательности. Этот метод широко используется в машинном переводе, генерации текстовых резюме и других задачах, связанных с последовательностями.

TCN: Прогнозирование с использованием временных сверточных нейронных сетей с вероятностной оценкой - это метод прогнозирования временных рядов с использованием временных сверточных нейронных сетей (TCN). TCN использует сверточные слои для анализа локальных и глобальных паттернов во временных данных и обеспечивает прогнозирование с вероятностной оценкой. Этот метод часто используется для прогнозирования временных рядов, таких как цены акций, прогноз погоды и трафик.

2. Сравнение производительности

**Точность прогноза**: Таким образом, последовательное обучение (Seq2Seq), основанное на нейронных сетях, хорошо работает при обработке текстовых задач, особенно с высокой точностью при обработке последовательностей короткой и средней длины, однако при работе с длинными последовательностями могут возникнуть такие проблемы, как градиент; встречаются исчезновения, что влияет на точность. Напротив, при использовании временных сверточных нейронных сетей (TCN) для вероятностного прогнозирования вы обычно можете достичь лучших результатов в задачах временных рядов с более высокой скоростью обучения и лучшим параллелизмом.

**Вычислительная сложность**: С точки зрения вычислительной сложности модели Seq2Seq обычно требуют больше вычислительных ресурсов, особенно при работе с длинными последовательностями, поскольку им необходимо кодировать и декодировать всю последовательность, поэтому их вычислительные затраты выше. Напротив, модель TCN имеет меньшую вычислительную сложность, поскольку она может обрабатывать различные части входной последовательности параллельно и, как правило, имеет более быструю сходимость во время обучения. Таким образом, TCN может быть более привлекательным выбором, когда ресурсы ограничены или требования к времени обучения высоки.

3. Сценарии применения

Модель Seq2Seq подходит для задач, требующих обработки текстовых последовательностей, таких как машинный перевод, обобщение текста и диалоговые системы, поскольку она может точно согласовывать входные и выходные последовательности.

Модель TCN больше подходит для обработки данных временных рядов, таких как прогнозирование запасов, распознавание действий и обработка сигналов, из-за меньшей вычислительной сложности и более высокой скорости обучения.

4. Заключение

Модели Seq2Seq могут столкнуться с проблемами исчезновения или взрыва градиента при обработке длинных последовательностей, особенно во время обучения. Модель TCN, используя операции свертки и остаточные связи, может более эффективно обрабатывать данные длинных последовательностей, обеспечивая при этом параллельные вычисления, улучшая скорость обучения и производительность модели. Поэтому при выборе модели необходимо учитывать характеристики данных, требования задачи и ограничения вычислительных ресурсов.

## Список использованных источников

[1] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems (pp. 3104-3112).

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[3] Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

[4] Lea, C., Flynn, M. D., Vidal, R., Reiter, A., & Hager, G. D. (2017). Temporal convolutional networks for action segmentation and detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1003-1012).

[5] Oord, A. v. d., & Vinyals, O. (2017). Neural discrete representation learning. In Advances in neural information processing systems (pp. 6306-6315).