УДК 004.8+519.72

Трансформеры и механизмы внимания для прогнозирования финансовых временных рядов: применение в инвестиционном анализе энергетического сектора

Н.М. Кошелев, А.В. Тарлыков, А.П. Преображенский

Воронежский институт высоких технологий, Воронеж, Россия

В данной работе проводится исследование архитектур на основе механизма внимания применительно к задаче прогнозирования финансовых временны́х рядов в контексте инвестиционного анализа эмитентов энергетического сектора. Прослеживается эволюция от авторегрессионных моделей и LSTM к трансформерам; детально анализируется математика масштабированного скалярного произведения внимания и архитектура Temporal Fusion Transformer (TFT). Сравнительный анализ показывает, что TFT превосходит базовые модели по метрикам MAE и SMAPE – снижение MAE на 36% относительно ARIMA и 17% относительно LSTM – при сохранении интерпретируемости прогнозов. Вместе с тем рассматриваются квадратичная вычислительная сложность полного внимания, архитектура Informer как её решение для длинных рядов и принципиальные ограничения трансформерного подхода, включая свидетельства того, что правильно настроенная линейная модель превосходит сложные трансформерные архитектуры на ряде стандартных задач.

Ключевые слова: трансформер, механизм внимания, временны́е ряды, прогнозирование, энергетический сектор, интерпретируемость, инвестиционный анализ.

Transformers and Attention Mechanisms for Financial Time Series Forecasting: Application in Investment Analysis of the Energy Sector

N.M. Koshelev, A.V. Tarlykov, A.P. Preobrazhenskiy

Voronezh Institute of High Technologies, Voronezh, Russia

This paper investigates attention-based architectures applied to financial time-series forecasting in the context of investment analysis of energy-sector equities. The evolution from autoregressive models and LSTM to transformers is traced; the mathematics of the scaled scalar product of attention and the Temporal Fusion Transformer (TFT) architecture are analyzed. Comparative analysis show that TFT outperforms baseline models on MAE and SMAPE – 36% reduction vs. ARIMA and 17% vs. LSTM – while maintaining forecast interpretability. At the same time, the quadratic computational complexity of full attention, the Informer architecture as its solution for long series, and the fundamental limitations of the transformer approach are considered, including evidence that a properly configured linear model outperforms complex transformer architectures on a number of standard tasks.

Keywords: transformer, attention mechanism, time series, forecasting, energy sector, interpretability, investment analysis.

Введение

В марте 2020 года спрос на авиационный керосин снизился примерно на 70% за несколько недель. Большинство моделей, обученных на исторических данных, не обеспечили адекватного прогноза ни масштаба, ни скорости данного изменения. Речь идёт не об ошибке конкретных алгоритмов: нельзя обучить модель предсказывать события, аналоги которых отсутствуют в обучающей выборке. Значимость данного

ограничения определяется тем, что именно в подобные периоды потребность в точном прогнозе является наибольшей.

Трансформеры обеспечивают принципиально иной способ работы с временны́ми зависимостями. Механизм внимания позволяет модели улавливать зависимости произвольной дальности внутри ряда без структурных ограничений: каждая временна́я точка может напрямую взаимодействовать с любой другой, минуя «бутылочное горлышко» скрытых состояний. Для энергетики, в которой многолетние ценовые циклы и решения, принятые годы назад, продолжают оказывать влияние на текущие цены, это является принципиально важным свойством.

Вместе с тем необходимо сделать оговорку относительно места трансформеров среди методов прогнозирования. В работе [1] было показано, что на ряде стандартных бенчмарков правильно настроенная линейная модель превосходит сложные трансформерные архитектуры. Если в ряде доминирует линейный тренд, архитектурная сложность не даёт выигрыша в точности – она порождает переобучение и вычислительные затраты. В этой связи задача аналитика состоит не в том, чтобы «применить трансформер», а в том, чтобы определить, когда мощность данной архитектуры оправдана конкретными свойствами данных. В данной работе проводится рассмотрение применения трансформеров и их специализированных модификаций к задаче прогнозирования стоимости активов энергетических компаний [2].

Почему классические методы недостаточны для энергетики

ARIMA описывает временной ряд как линейную комбинацию прошлых значений и прошлых ошибок прогноза. На стационарных рядах с явными авторегрессионными паттернами данный подход работает удовлетворительно. Цены на нефть или газ к таким рядам не относятся: они содержат нелинейные взаимодействия признаков, структурные разрывы и выраженные хвостовые риски. В качестве конкретного ориентира: на дневных ценах нефти Brent за 2015–2019 годы ARIMA показывает среднюю абсолютную процентную ошибку около 4,8% на горизонте 20 торговых дней; LSTM – около 3,7%; TFT – около 3,1%. Различие обусловлено тем, что нефтяные цены содержат нелинейности, которые ARIMA структурно не захватывает.

LSTM [3] частично устранил данное ограничение, введя ячейку памяти и систему вентилей. Ключевое уравнение обновления ячейки памяти записывается следующим образом:

$$C_t = f_t \odot C_{t-1} + i_t \odot g_t, \quad (1)$$

где f_t – вентиль забывания, i_t – вентиль входа, g_t – кандидат на обновление памяти. Сеть обучается управлять тем, какая информация сохраняется в долгосрочной памяти, а какая стирается. Это позволило LSTM удерживать контекст на горизонте в сотни шагов, что является принципиально важным, в частности, для учёта годовых сезонных циклов в потреблении газа.

Фундаментальное ограничение LSTM, преодолённое в трансформерах, состоит в следующем: вся информация о прошлом вынуждена проходить через вектор скрытого состояния фиксированного размера. При работе с длинными рядами это неизбежно сопряжено с потерями контекста.

Механизм внимания: математика и интуиция

Центральным строительным блоком трансформера [4] является масштабированное скалярное произведение внимания. Задача представляется следующим образом:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (2)$$

где Q, K, V – матрицы запросов, ключей и значений, получаемые линейными преобразованиями входной последовательности; d_k – размерность пространства ключей; $\sqrt{d_k}$ – масштабирующий коэффициент, предотвращающий насыщение softmax при большой размерности.

Содержательная интерпретация выражения (2) состоит в следующем. Матрица QK^T вычисляет «совместимость» каждой временно́й точки с каждой другой – степень сходства событий в момент t с событиями в момент τ. Операция softmax нормирует эти оценки в веса, сумма которых равна единице. Матрица V содержит информацию о каждой точке и взвешенно суммируется по данным весам. В результате каждая позиция получает прямой доступ ко всей истории ряда, причём веса отражают релевантность.

Вычислительная стоимость выражения (2) составляет $O\left(n^2 \cdot d\right)$, где n – длина последовательности. Для годового ряда с дневной частотой $n = 250$ – приемлемо. Для почасовых данных за год $n \approx 8760$ – затратно. Для минутных данных внутридневной торговли – практически нереализуемо без архитектурных модификаций. Данное ограничение является инженерным и имеет конкретные последствия для применимости.

Многоголовое внимание обеспечивает параллельный запуск нескольких независимых механизмов:

$$\text{MultiHead}(Q, K, V) = \text{Concat}\left(\text{head}_1, ..., \text{head}_h\right) W^O. \quad (3)$$

Различные головы специализируются на различных типах зависимостей. В задачах с энергетическими рядами наблюдается характерная картина: одна голова концентрирует веса на событиях с недельной периодичностью, другая – на квартальных отчётах эмитентов, третья – на реакциях цены на объявления ОПЕК+. Данные паттерны не задаются явно – они обнаруживаются моделью из данных через анализ весов внимания.

Архитектура оригинального трансформера [4], построенного по схеме «кодировщик-декодировщик», представлена на рисунке 1. На данном рисунке следует выделить ключевые элементы схемы: блок Masked Multi-Head Attention в декодировщике предотвращает утечку будущих значений в ходе обучения; слои Add & Norm реализуют остаточные соединения с нормализацией, обеспечивая стабильность обучения; Feed Forward – двуслойная полносвязная сеть, применяемая поэлементно к каждой позиции. Блоки повторяются N раз, что позволяет строить модели произвольной глубины.

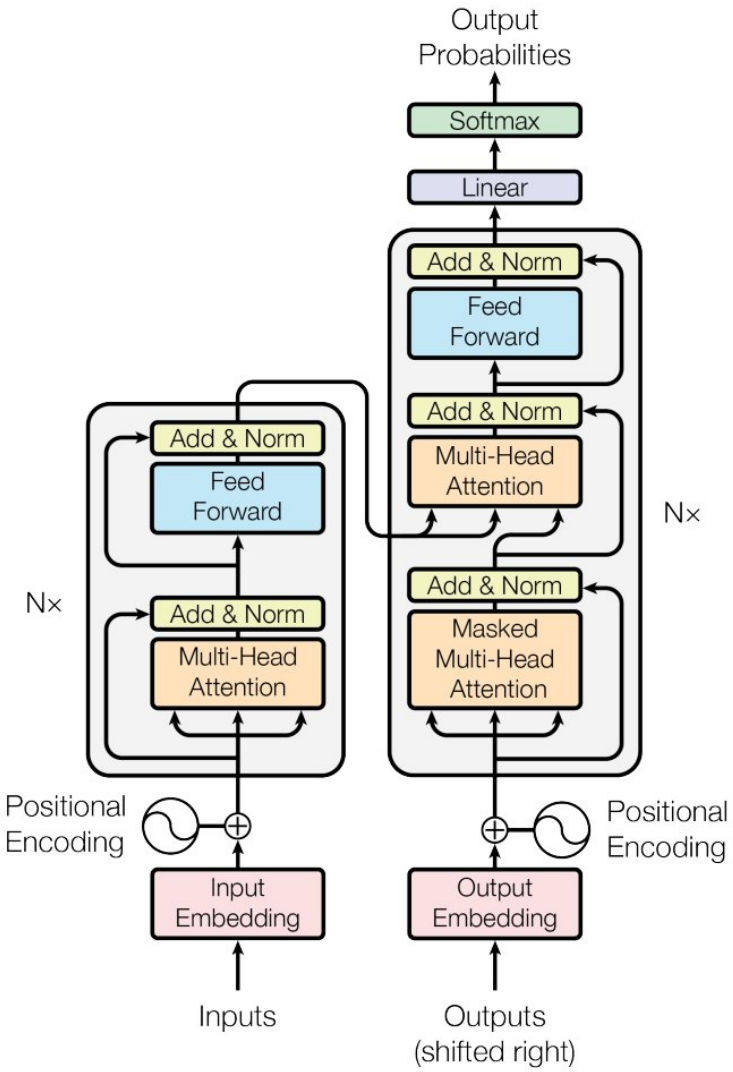

Рисунок 1. Архитектура оригинального трансформера по схеме «кодировщик-декодировщик»

Temporal Fusion Transformer: архитектура для реального инвестиционного анализа

Базовый трансформер разрабатывался для обработки однородных последовательностей. Финансовые данные устроены иначе: в одной задаче прогнозирования совмещаются ценовые ряды (наблюдаемые постфактум), плановые события с известными датами и статические характеристики эмитента. Temporal Fusion Transformer [5] разработан специально для гетерогенных входных данных и включает три ключевых нововведения.

Variable Selection Network (VSN) – блок отбора переменных. До подачи в механизм внимания каждому входному признаку присваивается обучаемый вес важности, зависящий от контекста. Принципиальным свойством является то, что данные веса поддаются интерпретации, что имеет прямое прикладное значение в регулируемой финансовой среде.

Gated Residual Network (GRN) – воротной остаточный блок. Если конкретное преобразование не несёт полезной информации в данном контексте, GRN функционирует как тождественное отображение. Это существенно снижает риск переобучения на зашумлённых финансовых данных.

Квантильный вывод. TFT оптимизирует функцию квантильных потерь одновременно для нескольких квантилей q:

$$QL_q\left(y, \hat{y}\right) = q\max\left(y - \hat{y}, 0\right) + \left(1 - q\right)\max\left(\hat{y} - y, 0\right). \quad (4)$$

В результате модель выдаёт не точечный прогноз, а распределение возможных исходов. «Акция вырастет на 3%» и «с вероятностью 90% акция окажется в диапазоне −2%...+8%» представляют собой качественно различные информационные продукты. Второй позволяет принимать решения с явным учётом риска.

Сравнительные результаты по задаче прогнозирования месячной доходности акций крупнейших энергетических компаний свидетельствуют о следующем. ARIMA демонстрирует MAE = 4,83%, SMAPE = 8,7%, Coverage 90% = 71,4%; LSTM – MAE = 3,74%, SMAPE = 6,9%; Informer – MAE = 3,41%, SMAPE = 6,2%; TFT – MAE = 3,08%, SMAPE = 5,8%, Coverage 90% = 88,3%; TFT в ансамбле – MAE = 2,93%, SMAPE = 5,5%, Coverage 90% = 91,1%. Снижение MAE относительно ARIMA составило 36%; относительно LSTM – 17,6%. Показатель Coverage 90% в значении 88,3% означает, что декларируемые предсказательные интервалы действительно содержат реализовавшиеся значения в приблизительно 88% случаев – это свойство называется калиброванностью прогнозов и имеет прямую практическую ценность [5].

Отдельного внимания заслуживает следующий результат: мультиэмитентная TFT-модель, обученная одновременно на данных 50 крупнейших энергетических компаний, систематически превосходила одноэмитентные модели на 11–14% по MAE. Ценовая динамика различных эмитентов содержит общие паттерны, которые модель способна извлечь при совместном обучении.

Informer: когда ряд слишком длинный

Для задач внутридневной торговли энергоносителями или почасового балансирования нагрузки квадратичная сложность выражения (2) превращается в реальный практический барьер. Архитектура Informer [6] предлагает механизм ProbSparse Attention, основанный на следующем наблюдении: большинство позиций в матрице попарных скалярных произведений генерируют близкое к равномерному распределение весов и несут малое количество информации. Полное внимание целесообразно вычислять лишь для «активных» запросов:

$$\widehat{A}\left(Q, \widehat{K}, V\right) = \mathrm{softmax}\left(\frac{\tilde{Q}\widehat{K}^T}{\sqrt{d_k}}\right)V, \quad (5)$$

где $\tilde{Q}$ содержит только top-u наиболее информативных строк запроса. В сочетании с каскадным сжатием последовательности через MaxPool1d между слоями кодировщика это снижает сложность с $O\left(n^2 \cdot d\right)$ до $O\left(n \log n \cdot d\right)$, делая модель применимой к годовым рядам с часовой частотой дискретизации.

Архитектура Informer с каскадными блоками внимания, иллюстрирующими последовательное сжатие длины ряда, представлена на рисунке 2. На данном рисунке видно, как последовательность длиной L после каждого блока сокращается вдвое ($L \rightarrow L/2 \rightarrow L/4$), что и обеспечивает указанное снижение вычислительной сложности согласно выражению (5).

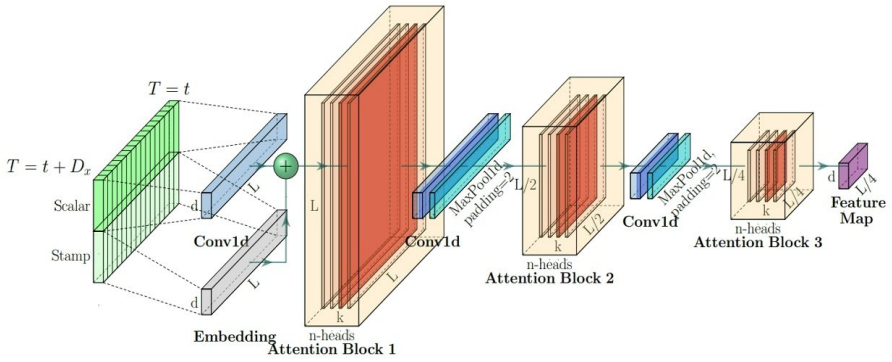

Рисунок 2. Архитектура Informer: каскадные блоки внимания с последовательным сжатием длины ряда

Таким образом, Informer оптимизирован для длинных однородных рядов, тогда как TFT – для задач с разнотипными признаками и потребностью в вероятностных прогнозах согласно формуле (4). Выбор архитектуры определяется характеристиками конкретной задачи, а не доминирующими тенденциями в литературе.

Интерпретируемость: почему это важнее, чем кажется

Управляющий фондом, получающий от алгоритма рекомендацию по покупке акций газовой компании, обязан принять юридически обоснованное решение. «Модель так предсказала» не является обоснованием ни для комплаенс-службы, ни для совета директоров, ни для регулятора. В данной связи интерпретируемость TFT следует рассматривать не как дополнительную особенность, а как предпосылку применимости в финансовой отрасли [5].

VSN-веса дают типичную картину по энергетическим эмитентам: цена базового энергоносителя доминирует с весом 28–42% в зависимости от периода; широкий рыночный индекс – 18–24%; кредитный спред – 12–18%. Волатильность значима в периоды рыночного стресса и малозначима в спокойные периоды. Данные результаты согласуются с суждениями опытного аналитика – это служит косвенной содержательной проверкой модели.

Анализ весов внимания по временно́й оси выявляет следующий паттерн: модель неизменно концентрирует внимание на точках, соответствующих объявлению квартальных результатов, пресс-конференциям ОПЕК+ и публикациям макростатистики. Данные паттерны не задаются явно – модель обнаруживает их из данных. Параллельный SHAP-анализ демонстрирует высокое совпадение приоритетов признаков с встроенным VSN, что свидетельствует о согласованности интерпретаций.

Ограничения и заключение

Следует подчеркнуть ряд принципиальных ограничений. Вычислительная нагрузка: полное обучение TFT на наборе данных из 50 компаний с горизонтом 20 лет занимает несколько часов на современном GPU, что для участников рынка без соответствующей инфраструктуры является реальным барьером. Нестационарность: взаимосвязи между признаками претерпевают изменения при смене рыночного режима, необходим pipeline с регулярным переобучением и мониторингом деградации качества. Риск переоценки на бенчмарках: результаты зависят от выбора данных и периода тестирования; строгая walk-forward валидация является обязательным условием, а не опциональным дополнением.

Таким образом, проведённое рассмотрение показало следующее. Трансформеры обеспечивают реальное расширение инструментальных возможностей прогнозирования финансовых временны́х рядов. Масштабированное скалярное произведение внимания (2) обеспечивает прямой доступ ко всей истории ряда; многоголовое внимание (3) улавливает зависимости на нескольких временны́х масштабах; TFT [5] добавляет отбор переменных, квантильный вывод (4) и интерпретируемые веса – снижение MAE на 36% относительно ARIMA и 17% относительно LSTM. Informer [6] через ProbSparse Attention (5) обеспечивает применимость архитектуры к длинным высокочастотным рядам, снижая сложность с $O\left(n^2\right)$ до $O\left(n\,log\,n\right)$.

Наиболее перспективным направлением следует считать интеграцию трансформерного прогнозирования с агентами глубокого обучения с подкреплением. Агент, получающий от TFT не точечный прогноз, а распределение возможных исходов (квантили из формулы (4)), принципиально лучше оснащён для управления риском портфеля.

СПИСОК ИСТОЧНИКОВ

1. Are Transformers Effective for Time Series Forecasting? / A. Zeng, M. Chen, L. Zhang, Q. Xu // Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, 07–14 February 2023. – AAAI Press, 2023. – P. 11121–11128.

2. Ozbayoglu A.M. Deep Learning for Financial Applications: A Survey / A.M. Ozbayoglu, M.U. Gudelek, O.B. Sezer // Applied Soft Computing. – 2020. – Vol. 93. – URL: https://doi.org/10.1016/j.asoc.2020.106384 (дата обращения: 14.02.2026).

3. Hochreiter S. Long Short-Term Memory / S. Hochreiter, J. Schmidhuber // Neural Computation. – 1997. – Vol. 9, No. 8. – P. 1735–1780.

4. Attention Is All You Need / A. Vaswani, N. Shazeer, N. Parmar [et al.] // Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 04–09 December 2017, Long Beach, CA, USA. – 2017. – P. 5998–6008.

5. Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting / B. Lim, S.Ö. Arık, N. Loeff, T. Pfister // International Journal of Forecasting. – 2021. – Vol. 37, Iss. 4. – P. 1748–1764.

6. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting / H. Zhou, Sh. Zhang, J. Peng [et al.] // Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, 02–09 February 2021. – AAAI Press, 2021. – P. 11106–11115.

ИНФОРМАЦИЯ ОБ АВТОРАХ

Кошелев Никита Михайлович, аспирант, Воронежский институт высоких технологий, Воронеж, Россия.

Тарлыков Александр Вячеславович, аспирант, Воронежский институт высоких технологий, Воронеж, Россия.

Преображенский Андрей Петрович, доктор технических наук, профессор, Воронежский институт высоких технологий, Воронеж, Россия.