УДК 004.4

Пример применения сети графического внимания GAT для прогнозирования разницы энергий HOMO-LUMO

Ф. Леон[*], Г.А. Ососков[‡], Е.Н. Толочко[‡], Ю.В. Гайдамака[*†]

[*] *Кафедра теории вероятностей и кибербезопасности, Российский университет дружбы народов, ул. Миклухо-Маклая, д.6, Москва, Россия, 117198*

[†] *Федеральный исследовательский центр «Информатика управление» РАН ул. Вавилова, д. 44, корп. 2, Москва, 119333, Россия*

[‡] *Объединенный институт ядерных исследований ул. Жолио-Кюри, д. 6, Дубна, 141980, Россия*

Email: leon.jf@outlook.com, ososkov@jinr.ru, yauheni.talochka@gmail.com, gaydamaka-yuv@rudn.ru

В этом исследовании рассматривается применение архитектуры сети внимания графов (GAT) для задач регрессии, в частности, для прогнозирования разницы в энергии молекул между HOMO и LUMO. Мы разработали модель графовой нейронной сети (GNN), обученную на наборе данных PCQM4Mv2, который содержит примерно 3,8 миллиона молекул и предоставляется тестом Open Graph Benchmark (OGB). Производительность модели GAT оценивалась с использованием стандартных показателей регрессии, что продемонстрировало ее потенциал для точного предсказания квантовых свойств.

Ключевые слова: Графовая сеть внимания (GAT), Графовые нейронные сети (GNNs), разрыв HOMO-LUMO, Молекулярные графы, Прогнозирование молекулярных свойств.

1. Введение

Разрыв HOMO-LUMO является важным электронным свойством в молекулярной химии, влияющим на оптическое и электронное поведение молекул. Точное предсказание этого параметра имеет значительное значение для материаловедения и разработки лекарственных препаратов. Графовые модели глубокого обучения, такие как GAT, продемонстрировали многообещающие результаты в эффективной обработке молекулярных структур [1].

В хемоинформатике традиционные методы машинного обучения часто опираются на вручную созданные молекулярные дескрипторы и признаки, что может ограничивать способность модели к обобщению различных молекулярных структур. С другой стороны, GAT обеспечивают более гибкий и выразительный подход, используя графовые представления молекул. Они применяют механизм внимания, который назначает разную степень значимости различным атомным и химическим связям, позволяя более точно моделировать молекулярные свойства. Это делает GAT особенно полезными для таких приложений, как разработка лекарств, материаловедение и прогнозирование свойств молекул, где важно улавливать сложные молекулярные взаимосвязи. В сравнении с традиционными графовыми сверточными сетями (GCN), GAT улучшают извлечение признаков за счет динамического взвешивания вклада соседних узлов, что приводит к повышенной точности предсказания молекулярных свойств [2].

2. Набор данных и архитектура модели

Мы описываем наш набор данных OGB-LSC (Open Graph Brenchmark - Large Scale Challenge), который охватывает категорию задач прогнозирования на уровне графа ML в графах [3]. Мы подчеркиваем практическую значимость и разделение данных для набора данных.

2.1. Набор данных PCQM4Mv2

Практическая значимость. Точное прогнозирование зазоров HOMO-LUMO имеет решающее значение для продвижения исследований в области материаловедения и химии. Это свойство играет важную роль в определении электронного, оптического и проводящего поведения молекулы. Приложения охватывают различные области, такие как проектирование органических полупроводников, фотоэлектрических мате-риалов и катализаторов. Традиционные вычислительные методы, такие как теория функционала плотности (DFT), хотя и точны, требуют больших вычислительных затрат. Подходы машинного обучения, особенно те, которые используют GNN, предлагают многообещающую альтернативу, предоставляя быстрые и точные прогнозы [4].

Обзор датасета. Набор данных, используемый в этом исследовании, представляет собой набор данных PCQM4Mv2, полученный из Open Graph Benchmark – Large Scale Challenge (OGB-LSC). Этот набор данных специально подобран для задачи прогнозирования на уровне графа, где цель состоит в том, чтобы предсказать молекулярные свойства на основе их графических представлений.

- **Число молекул**: Набор данных содержит более 3,8 миллионов молекулярных графов, что делает его одним из крупнейших общедоступных наборов данных по квантовой химии.
- **Узлы и ребра графического представления**: Узлы соответствуют атомам молекулы с признаками, кодирующими атомные свойства (9 признаков). Края представляют собой химические связи с признаками, отражающими типы связей, порядок связей и стереохимию (3 признака).
- **Целевое свойство**: Целевым свойством служит щель HOMO-LUMO каждой молекулы, которая измеряется в электронвольтах (eV).

Набор данных	Тип задачи	Статистика
PCQM4M	Графический уровень	#graphs: 3,746,619
		#nodes (total): 52,970,652
		#edges (total): 54,546,813

Таблица 1

Базовая статистика набора данных OGB-LSC

2.2. Методология

Модель Архитектуры. Предлагаемая модель представляет собой графовую нейронную сеть (GNN), разработанную для прогнозирования разрыва HOMO-LUMO gap молекул. Архитектура адаптирована для захвата графически структурированной природы молекулярных данных

GNN Layers. Модель использует слои GATv2 для агрегации и распространения информации через молекулярный граф. Четыре слоя GATv2 накладываются друг на друга, за каждым из которых следует функция активации ELU и механизм нормализации BatchNorm и LayerNorm.

Механизм объединения. Операция «global mean pooling» и «global sum pooling» объединяют вложения на уровне узлов в единое представление на уровне графа, что позволяет учитывать разные аспекты молекулярной структуры.

Окончательная регрессия головы. Представление на уровне графа передается через глубокую нейросеть, состоящую из нескольких полностью связанных слоев с активацией ELU и dropout для регуляризации. Итоговый выходной слой формирует один скалярный прогноз, представляющий зазор HOMO-LUMO gap.

Разделение набора данных. Мы разделяем молекулы по их PubChem ID с соотношением 90/2/4/4 (train/validation/test-dev/test-challenge). Однако мы не используем молекулы подмножеств test-dev и test-challenge, поскольку test-dev — это квалифицировать нашу модель и сравнить ее с другими моделями, ранее представленными на сайте OGB-LSC, а с другой стороны, подмножество test-challenge предназначено для использования в конкурсах моделирования.

Data Loaders. Настроено для перемешивания набора данных, что улучшает обобщение модели. Где каждая партия содержит 128 молекулярных графа, параллель-ная загрузка данных с использованием 4 рабочих потоков и оптимизирует передачу данных в память GPU.

Гиперпараметры. Ключевые гиперпараметры модели GNN:

- Hidden Dimension: 256
- Number of Layers: 4
- Heads: 8
- Learning Rate: 0.001
- Batch Size: 128
- Dropout Rate: 0.4
- Number of Passes: 3

2.3. Процесс обучения

Loss Function: Функция потерь, используемая во время обучения, была среднеквадратичной ошибкой (MSE), которая подходит для задач регрессии. Эта функция измеряет среднеквадратичное отклонение между предсказанными и фактическими значениями, сильнее штрафуя большие ошибки. Выбор MSE гарантирует, что модель отдает приоритет минимизации значительных отклонений в прогнозах.

Optimizer/Scheduler: Модель была обучена с использованием оптимизатора Adam, популярного выбора для задач глубокого обучения из-за его механизма адаптивной скорости обучения. Начальная скорость обучения была установлена на уровне 0,001. Это постепенное снижение помогает модели более эффективно сходиться по мере обучения.

Процесс обучения состоял из 30 эпох, со следующими стратегиями для повышения стабильности и производительности. Параметры модели сохранялись всякий раз, когда улучшалась средняя абсолютная ошибка валидации (MAE), гарантируя, что модель с наилучшими показателями сохранялась для дальнейшей оценки.

3. Моделирование и численные результаты

На рисунке 1 представлены графики потерь на обучающем наборе (MSE) и MAE на валидационном наборе в течение 30 эпох. Левый график показывает, что обучающая ошибка постепенно уменьшается, что свидетельствует о том, что модель эффективно минимизирует ошибки во время обучения. На начальных этапах наблюдаются колебания, но по мере обучения значение функции потерь стабилизируется, достигая итогового значения 0.0952.

Аналогично, правый график иллюстрирует динамику MAE на валидационном наборе, который также демонстрирует положительную тенденцию. После начальных колебаний значение MAE постепенно снижается, подтверждая способность модели к обобщению на новых данных. К концу обучения MAE достигает 0.2720, что подтверждает эффективность процесса обучения. Эти результаты указывают на то, что модель на основе GAT успешно улавливает молекулярные закономерности и может быть надежно использована для предсказания разрыва HOMO-LUMO.

Предложенная модель GATv2 демонстрирует высокую точность в прогнозировании разрыва HOMO-LUMO в среднем диапазоне, достигая средней относительной процентной ошибки 5,19%, что свидетельствует о высокой предсказательной способности. Однако, как показывают результаты, модель испытывает трудности с экстремальными значениями HOMO-LUMO, особенно с меньшими разрывами. В этих случаях наблюдается значительно более высокая относительная процентная

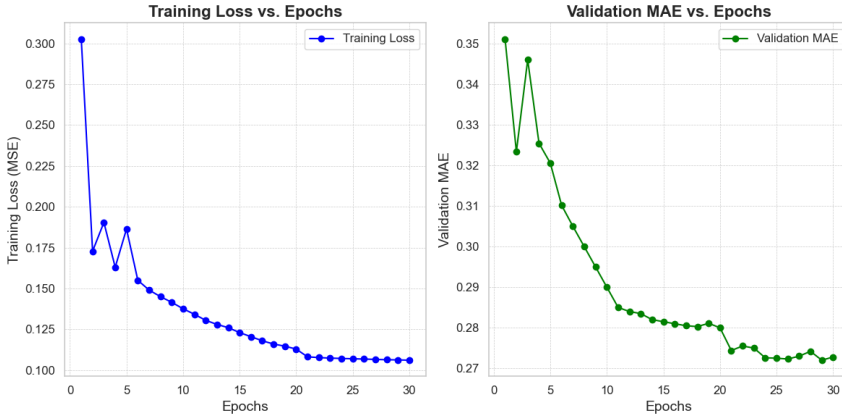

Рис. 1. Эволюция потери обучения (функция потерь MSE) и валидационная MAE во время обучения модели GNN

ошибка, что говорит о сложности захвата распределения менее распространенных молекулярных структур. Такое поведение может быть связано с дисбалансом в наборе данных или сложностью представления молекулярных графов с экстремальными электронными свойствами. На рисунке 2

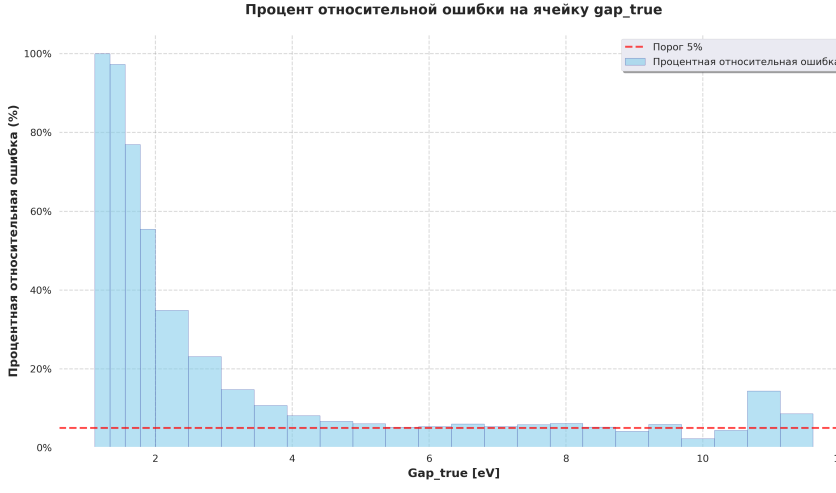

Рис. 2. Диаграмма рассеяния: прогнозируемые значения против реальные ценности.

4. Заключение

Разработанная модель GATv2 демонстрирует высокую точность предсказания разрыва HOMO-LUMO в среднем диапазоне, но испытывает трудности с экстремальными значениями, особенно с низкими, что указывает на ограниченную способность к обобщению для молекул с редкими электронными свойствами. Для повышения точности можно рассмотреть нормализацию целевой переменной для балансировки предсказаний, использование взвешенной функции потерь для акцентирования внимания на редких молекулах, а также модификации архитектуры, включая дополнительные слои или механизмы регуляризации, чтобы лучше улавливать сложные молекулярные закономерности. Дальнейшие исследования должны быть направлены на оценку этих улучшений, чтобы обеспечить стабильные и надежные предсказания во всем диапазоне значений HOMO-LUMO.

Благодарности

Публикация выполнена в рамках соглашения о сотрудничестве в научно- исследовательской деятельности и подготовке кадров между ОИЯИ и РУДН от 06.07.2021 №40-18/48 и соглашения о создании научного консорциума «Аналитика Больших данных для задач естественно-научного профиля» от 13.07.2021 №40-18/46.

Литература

1. Veličković Petar, Cucurull Guillem, Casanova Arantxa, Romero Adriana, Liò Pietro, and Bengio Yoshua. Graph Attention Networks. — 2018. — 1710.10903.
2. Xu Lei, Pan Shourun, Xia Leiming, and Li Zhen. Molecular Property Prediction by Combining LSTM and GAT // Biomolecules. — 2023. — Vol. 13, no. 3. — Access mode: https://www.mdpi.com/2218-273X/13/3/503.
3. Benchmark Stanford Open Graph. PCQM4Mv2: A Benchmark for Learning from Molecular Graphs. — 2021. — Access mode: https://ogb.stanford.edu/docs/lsc/pcqm4mv2/. Accessed: 2025-01-16.
4. Hu Weihua, Fey Matthias, Ren Hongyu, Nakata Maho, Dong Yuxiao, and Leskovec Jure. OGB-LSC: A Large-Scale Challenge for Machine Learning on Graphs // KDD Cup 2021. NeurIPS Datasets and Benchmarks Track. — 2021. — Subjects: Machine Learning (cs.LG). arXiv:2103.09430.

UDC 004.4

An example of using the graphical attention network GAT to predict the HOMO-LUMO energy difference

F. Leon[*], G. A. Ososkov[‡], Y. Talochka[‡], Yu. V. Gaidamaka[*†]

[*] *Department of Probability Theory and Cyber Securitys*
Peoples' Friendship University of Russia
Miklukho-Maklaya str. 6, Moscow, 117198, Russia
[†] *Federal Research Center "Computer Science and Control" of the RAS*
44-2, Vavilova St. Moscow, 119333, Russian Federation
[‡] *Joint Institute for Nuclear Research*
6 Joliot-Curie St. Dubna, 141980, Russia

Email: leon.jf@outlook.com, ososkov@jinr.ru, yauheni.talochka@gmail.com, gaydamaka-yuv@rudn.ru

This study examines the application of the graph attention network (GAT) architecture to regression problems, specifically to predict the energy difference between the HOMO and

LUMO of molecules. We developed a graph neural network (GNN) model trained on the PCQM4Mv2 dataset, which contains approximately 3.8 million molecules and is provided by the Open Graph Benchmark (OGB). The performance of the GAT model was evaluated using standard regression metrics, demonstrating its potential to accurately predict quantum properties.

Key words and phrases: Graph attention network (GAT), Graph neural networks (GNNs), HOMO-LUMO gap, Molecular graphs, Molecular property prediction.