Label Attention Network для последовательной классификации по нескольким меткам.

Боева Галина
Антиплагиат
Сколтех
boeva.gl@phystech.edu

Консультант: к.ф.-м.н. Грабовой Андрей
Антиплагиат
grabovoy.av@phystech.edu

Эксперт: к.ф.-м.н. Зайцев Алексей
Сколтех
a.zaytsev@skoltech.ru

15 мая 2024 г.

Аннотация

Рассматривается задача прогнозирования временных наборов для последовательных данных. Современные подходы фокусируются на архитектуре преобразования последовательных данных, используя собственное внимания ("self-attention") к элементам в последовательности. В этом случае учитываются временные взаимодействия событий, но теряется информацию о взаимозависимостях меток. Мотивированные этим недостатком, предлагается использовать механизм собственного внимания("self-attention") к меткам, предшествующим прогнозируемому шагу. Поскольку рассматриваемый подход представляет собой сеть внимания к меткам, то называется LANET(Label Attention NETwork). В данном исследовании обосновывается вывод причинно-следственной связи внимания, которое указывает на важность меток и их взаимозависимости.

Ключевые слова временные ряды · взаимосвязь меток

1 Введение

Классификация с несколькими метками является более естественной, чем бинарная или многоклассовая классификация, поскольку все, что окружает нас в реальном мире, обычно описывается несколькими метками [1]. Та же логика может быть перенесена на последовательность событий с отметками времени. События в последовательности, как правило, характеризуются несколькими категориальными значениями вместо одной. Существует множество подходов к классификации с несколькими метками в компьютерном зрении [2], обработке естественного языка [3] или классической структуре табличных данных [4]. Однако постановке задачи с несколькими метками для последовательностей событий, как правило, уделяется меньше внимания. Итак, основной целью является противостоять такому недостатку внимания и решить проблему предсказания набора меток для последовательных данных с временными метками. Важно отметить, что модель должна предсказывать набор меток, соответствующих следующему шагу, принимая во внимание содержимое предыдущих групп меток для последовательности событий, связанных с объектом(Рисунок 1).

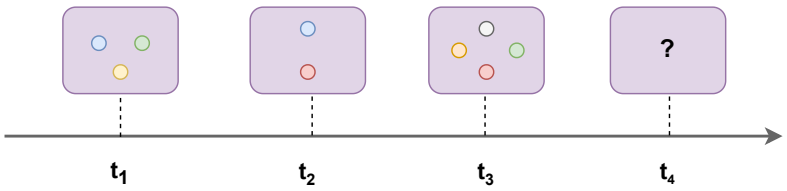

Рис. 1: На рисунке показано визуальное представление постановки задачи. Наша модель должна предсказать метки для момента времени t_4, учитывая историю предыдущих наборов меток. Требуется предсказать несколько меток, так что это определение задачи классификации с несколькими метками.

Взаимодействие между состояниями объекта в разные временные метки имеет важное значение для решения задач с последовательными данными [5]. Следовательно, выразительные и мощные модели должны быть способны изучать такие взаимодействия. Несколько архитектур нейронных сетей, таких как трансформеры [6] или рекуррентные нейронные сети [7], способны делать это. Например, трансформер напрямую определяет механизм внимания, который измеряет, как связаны различные временные метки в последовательности. Однако применение современных методов глубокого обучения ограничено [8], и они в первую очередь сосредоточены на прогнозировании меток для последовательности в целом.

1.1 Основные подходы для задачи классификации с несколькими метками.

Постановка задачи классификации с несколькими метками возникает во многих различных областях, например, при категоризации текста или тегировании изображений, и все они влекут за собой свои собственные особенности и проблемы. В обзоре [9] исследуются основы обучения с использованием нескольких меток, обсуждаются хорошо зарекомендовавшие себя методы, а также самые последние подходы. Возникающие тенденции рассматриваются в более свежем обзоре [1].

В работе [10] рассматривается та же постановка задачи классификации с несколькими метками в потоке событий, что и у нас. Модель авторов нацелена на фиксацию временных и вероятностных зависимостей между типами параллельных событий путем кодирования исторической информации с помощью энкодера, а затем использования условной смеси экспертов Бернулли. В этой статье [11] обсуждается постановка задачи прогнозирования временных наборов для пользователей, она предлагает систему непрерывного обучения, которая позволяет явно фиксировать изменяющиеся пользовательские предпочтения, поддерживая банк памяти, который мог бы хранить состояния всех пользователей и элементов. В этой парадигме авторы строят неубывающую универсальную последовательность, содержащую все пользовательские взаимодействия, а затем в хронологическом порядке извлекают уроки из каждого взаимодействия. Для исследования взаимосвязи между продуктами в корзине был предложен ConvTSP [12], который объединяет динамические интересы пользователей и статистические интересы в единое векторное представление.

1.2 Рекомендательные системы.

В этом разделе мы представим статьи, связанные с проблемой рекомендации следующей корзины. Эта формулировка похожа на нашу, поэтому мы также рассмотрели многие подходы и идеи при анализе нашей области исследований. Авторы в [13] предложили персонализированную модель, которая фиксирует краткосрочные зависимости внутри временного набора продуктов, а также долгосрочную, основанную на исторической информации о пользователях. Также в [14] для соединения локальной и глобальной пользовательской информации предлагается гибридный метод, основанный на автоэнкодере [15] для извлечения контекста и рекуррентные нейронные сети для понимания динамики изменения интересов. Чтобы преодолеть подобные проблемы, для предсказания следующей рекомендации создается сеть внимания на основе графов, использующая hyper-edge подход [16]. При такой постановке задачи возникает сложность работы со словарем товарных категорий, поскольку они насчитывают тысячи значений, в исследовании [17] используется GRU для прогнозирования следующей корзины, которая легко масштабируется до большого ассортимента.

Вклад. Разработана архитектура на основе трансформера на основе собственного внимания между метками для работы над задачей классификации последовательностей событий по нескольким меткам. Основной вклад заключается в следующем:

- Была введена архитектура LANET для прогнозирования набора меток для текущего события, используя информацию из предыдущих событий. Особенностью архитектуры является вычисление собственного внимания между представлениями меток.
- LANET превосходит модели на основе трансформера, которые фокусируются на вычислении собственного внимания между временными метками. Оцениваются все показатели в различных наборах данных.
- Также построен граф причинно-следственных связей для меток, который использует веса внимания LANET.

2 Постановка задачи

В теории последовательности событий каждое событие характеризуется одной категориальной меткой и временной отметкой. Существует множество доступных последовательностей событий, связанных с разными пользователями, с их базовыми моделями развития. При работе с такой структурой данных широко распространенной целью является выявление скрытых закономерностей последовательности на пользовательском и общем уровнях для прогнозирования будущего поведения. В большинстве случаев событию присваивается не одна метка, а некоторый набор меток. Более общей и реалистичной постановкой задачи является рассмотрение возможности одновременной привязки одного и того же момента времени к различным значениям. В дальнейшем будет рассматриваться временные наборы как последовательность связанных с событиями наборов временных меток, состоящих из произвольного числа меток.

Пусть $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ — это набор из N элементов. Каждый элемент $u_i, 1 \leq i \leq N$, связан с последовательностью временных множеств $\mathcal{S}_i = \{s_i^1, s_i^2, \ldots, s_i^T\}$, где T — число наблюдаемых временные метки. Набор $s_i^j, 1 \leq i \leq N, 1 \leq j \leq T$, представляет собой набор произвольного количества меток, выбранных из словаря $\mathcal{Y} = \{y_1, y_2, \ldots, y_L\}$ размера L.

Цель задачи предсказания временных множеств состоит в том, чтобы предсказать последующий набор меток $\hat{s}_i^{T+1}$, то есть,

$$\hat{s}_i^{T+1} = g(s_i^1, s_i^2, , s_i^T, \mathbf{W}), \tag{1}$$

где $\mathbf{W}$ относится к обучаемым параметрам функции g.

3 Архитектура LANET

Большинство моделей, связанных с трансформаторами, используемых для последовательного предсказания с несколькими метками, используют вычисление собственного внимания между последовательными представлениями входных временных меток. Вместо этого LANET использует собственное внимание между представлениями меток. Пусть $\mathbf{X} \in \mathbb{R}^{L \times D}$ — матрица представлений всех меток из словаря $\mathcal{Y} = \{y_1, y_2, \ldots, y_L\}$. Для каждой временной метки $j, 1 \leq j \leq T$ создается временное представление $\mathbf{t}_j \in \mathbb{R}^D$, как это сделано в [10]:

$$\mathbf{t}_j^{(d)} = \begin{cases} \cos\left(t_j / 10000^{\frac{d-1}{D}}\right), & \text{if } d \text{ is odd,} \\ \sin\left(t_j / 10000^{\frac{d}{D}}\right), & \text{if } d \text{ is even,} \end{cases}$$

Для каждого момента времени $t_j, 1 \leq j \leq T$ образуется матрица представлений $\mathbf{Z} \in \mathbb{R}^{L \times D}$. l-я строка, $1 \leq l \leq L$, матрицы $\mathbf{Z}$, обозначаемая как $\mathbf{Z}^{(l,:)}$, равна сумме представлений временных меток, в которых метка $y_l \in \mathcal{Y}$ отображается как элемент набора:

$$\mathbf{Z}^{(l,:)} = \sum_{j | y_l \in s_i^j} \mathbf{t}_j. \tag{2}$$

Объединенное представление последовательности $\mathcal{S}_i = \{s_i^1, s_i^2, \ldots, s_i^T\}$ представляет собой объединение определенных матриц, содержащих информацию о времени и структуре набора:

$$\mathbf{G} = \mathbf{X} \oplus \mathbf{Z}. \tag{3}$$

Для выявления зависимостей меток $\tilde{\mathbf{G}}$ воспользуемся механизмов собственного внимания:

$$\tilde{\mathbf{G}} = \text{softmax}(\frac{\mathbf{QK^T}}{\sqrt{2D}})\mathbf{V}, \tag{4}$$

где $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ - матрицы запроса, ключа и значения, которые являются линейными преобразованиями матрицы $\mathbf{G}$ После полученных обновленных представлений используется слой предсказаний для получения оценок достоверности для каждой из меток:

$$\hat{\mathbf{f}} = \text{sigmoid}(\tilde{\mathbf{G}}\mathbf{W}^{\text{out}} + b^{\text{out}}). \tag{5}$$

LANET обучается end-to-end, принимая историческую последовательность $\mathcal{S}_i = \{s_i^1, s_i^2, \ldots, s_i^T\}$ в качестве входных данных и формируя вектор оценки достоверности $\hat{\mathbf{f}}$ в качестве выходных данных. Мы используем следующую функцию потерь:

$$\mathcal{L}_i = -\frac{1}{L}\sum_{l=1}^{L}\left(\mathbf{I}_l \log \hat{\mathbf{f}}^{(l)} + \mathbf{I}_l^{'} \log\left(1 - \hat{\mathbf{f}}^{(l)}\right)\right), \tag{6}$$

где $\mathbf{I}_l = \mathbf{I}\{y_l \in s_i^{T+1}\}$ является индикаторной функцией метки y_l, которая является членом множества s_i^{T+1}, в то время как $\mathbf{I}_l^{'}$ — это индикаторная функция с противоположным условием $\mathbf{I}_l^{'} = \mathbf{I}\{y_l \notin s_i^{T+1}\}$. Мы обозначим l-ю составляющую прогнозируемого вектора оценки достоверности $\hat{\mathbf{f}}$ как $\hat{\mathbf{f}}^{(l)}$.

4 Вычислительный эксперимент

4.1 Описание данных

Таблица 1: Статистика наборов данных для прогнозирования временных наборов.

Dataset	#Sets	MdnSS	MaxSS	Vocab	MnLen	#Seqs
Mimic III	17 849	5	23	169	2.7	6636
Instacart	115 604	6	43	134	16.5	7000

- **Mimic III** — датасет, состоящий из медицинских карт пациентов из отделения интенсивной терапии. Событие, связанное с пациентом, включает в себя время поступления в больницу и набор классификационных кодов заболеваний.
- **Instacart** — набор данных содержит записи о заказах товаров пользователями. Товары из маркетплейсов и магазинов.

Общая статистика по рассмотренным наборам данных приведена в таблице 1.

4.2 Основные результаты

Показатели для сравнения подхода our PLANET с установленными моделями для решения задачи прогнозирования временных множеств представлены в таблице 2. LANET демонстрирует высочайшую производительность по всем наборам данных, существенно превосходя своих конкурентов. В Mimic III разброс значений между LANET и другими моделями довольно велик. LANET лучше выявляет взаимозависимости и сложные закономерности в данных. В Mimic III меньше всего событий, что влияет на сходимость моделей. В результате LANET работает быстрее, чем остальные. Более того, если мы рассмотрим набор данных Instacart, который содержит наибольшее количество наборов меток и последовательностей, то качество других подходов будет снижаться из-за сложности структуры набора данных. Подводя итог, можно отметить конкурентоспособность модели TCMBN, которая также основана на архитектуре transformer, но в то же время выигрывает правильная работа с этикетками.

4.3 Исследование и интерпретация модели

Зависимость качества модели от размера представлений. Хотя мы используем обучаемые представления для объектов с метками и временем, необходимо определить размерность их представлений. Для каждого объекта мы создаем представления одинакового размера. Размерность вектора

Таблица 2: Сравнение подхода our LANET с существующими моделями для прогнозирования временных наборов на основе четырех наборов данных. Выделены наилучшие значения, а вторые по значению подчеркнуты.

Data	Model	Weighted F1↑	Weighted ROC-AUC↑	Hamming Loss↓
Mim	SFCNTSP	0.3791 ± 0.0081	0.7034 ± 0.0024	0.0377 ± 0.0004
	DNNTSP	0.3928 ± 0.0030	0.6926 ± 0.0003	0.0365 ± 0.0003
	GPTopFreq	0.4291 ± 0.0073	0.6912 ± 0.0028	0.0398 ± 0.0005
	TCMBN	0.4979 ± 0.0180	0.8670 ± 0.0095	0.0305 ± 0.0008
	LANET(ours)	**0.8214 ± 0.0224**	**0.9852 ± 0.0023**	**0.0220 ± 0.0001**
Ins	SFCNTSP	0.1672 ± 0.0112	0.6852 ± 0.0448	0.0581 ± 0.0004
	DNNTSP	0.4160 ± 0.0009	0.7913 ± 0.0004	0.0541 ± 0.0002
	GPTopFreq	0.4087 ± 0.0079	0.7736 ± 0.0039	0.0529 ± 0.0008
	TCMBN	0.3687 ± 0.0065	0.8187 ± 0.0030	0.0530 ± 0.0005
	LANET(ours)	**0.6159 ± 0.0029**	**0.9445 ± 0.0008**	**0.0474 ± 0.0003**

представлений определяет объем информации, которую может воспринять модель, и общее количество изучаемых параметров. Результаты представлены в 2. Заметно, что модель с трудом справляется с эффективным изучением представлений с высокой размерностью.

Зависимость качества модели от количества голов во внимании. Гиперпараметром модели является количество голов на уровне внимания. Этот гиперпараметр позволяет модели учитывать различные закономерности в последовательности и концентрироваться на ее отдельных разделах. 3 показывает, что с увеличением их количества повышается качество. Но при большом количестве голов потребление ресурсов становится более обременительным, поэтому мы выбрали оптимальный параметр, равный четырем.

Зависимость производительности модели от количества уровней энкодера. Важным для рассмотрения является гиперпараметр количества уровней энкодера в модели transformer. Этот гиперпараметр отвечает за идентификацию абстрактных представлений, а также за распознавание сложных взаимосвязей между метками. 4 показывает, что мы достигли оптимума при количестве слоев, равном четырем. При дальнейшем улучшении качество снижается, но производительность модели также снижается из-за затухания градиентов.

Графовая интерпретация весов внимания. Важной частью результирующей архитектуры является уровень кодирования, который включает в себя уровень внимания. Внимание, в свою очередь, указывает на степень значимости взаимосвязи между метками, которая важна для дальнейшего прогнозирования модели.Мы выбираем наиболее подходящие надписи для подборки в Instacart, чтобы определить причинно-следственные связи между предсказаниями надписей. На рисунке 5 слева показана тепловая карта для их взаимосвязи. Мы замечаем, что в матрице внимания метки, встречающиеся в последовательности, явно преобладают над теми, которых в ней нет, что четко выражено с помощью весовых коэффициентов. Вглядываясь глубже, мы видим, что небольшие различия во внимании описывают связь между конкретными типами событий.

Кроме того, мы рассматриваем наиболее релевантные метки для выборки. На рисунке слева показана тепловая карта для их взаимосвязи. Для создания причинно-следственных связей нам потребовалась графическая визуализация шкал внимания для отдельных меток. Эта идея лежит в основе фреймворка CLEANN [18], который предлагает метод извлечения причинно-следственных связей в виде частичного графа предков (PAG) [19]. Итак, чтобы составить график, мы рассмотрели одного из пользователей и соответствующую историческую информацию о ярлыках. Используя предварительно обученную модель LANET, мы получили значения коэффициента внимания, введенные в алгоритм CLEANN.

Левая визуализация графика на рисунке 5 содержит несколько типов связей:

- Красные линии указывают на близость меток внутри графика;
- Синие связи более сложны, это двунаправленное взаимодействие между метками на графике;
- Черный цвет означает, что метка является родительской для последующих;
- Зеленые, напротив, — это дочерние.

В первом случае между этикетками возникли сложные и запутанные отношения. Например, если исходным кодом является ‘мясные консервы с морепродуктами вы создадите ‘заправки для салатов".

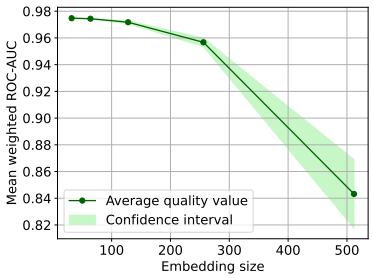

Рис. 2: The dependence of LANET quality on the embedding size.

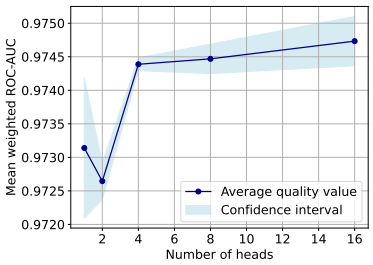

Рис. 3: The dependence of LANET quality on the number of heads.

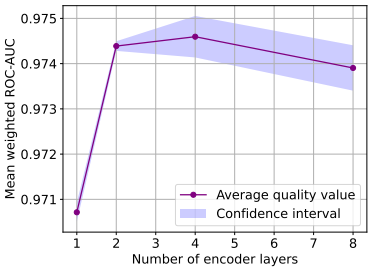

Рис. 4: The dependence of LANET quality on the number of encoder layers.

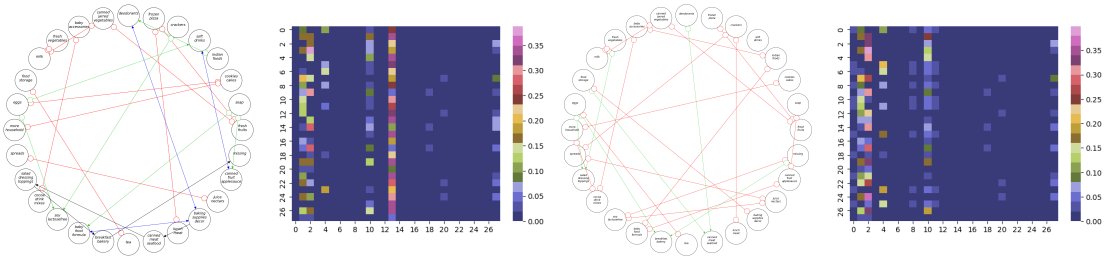

Рис. 5: Интерпретация взаимосвязи надписей с помощью слоя attention. Слева приведен рисунок, показывающий взаимосвязь между подмножеством надписей и их вербальной интерпретацией. Рядом с графиком приведена тепловая карта, которая иллюстрирует взаимосвязь всех возможных надписей в наборе данных Instacart. Справа представлены измененные графики, которые получены в результате удаления метки с наибольшим весом внимания из всех возможных значений и соответствующего распределения весов на тепловой карте. Данные получены из набора данных Instacart.

Некоторые связи могут показаться нам нелогичными, но эта история индивидуальна для каждого пользователя при покупке товара в магазине.

Более того, чтобы выяснить и идентифицировать связи, мы решили удалить метку с наибольшим общим весом в матрице внимания и посмотреть на перераспределение весов в этом случае (рисунок 5 справа). Модель обратила внимание на множество других меток. PAG демонстрирует измененную картинку, на которой исчезли все синие и черные края графика, что соответствует более сложной и ориентированной связи, чем простая ‘соседская". Корреляция между метками стала ниже. Более того, ‘мясные консервы с морепродуктами"изменили свое поведение. Она стала дочерней компанией и больше ни с кем не связана, что влияет на прогнозную способность этой этикетки для следующего временного шага. Это исследование показывает, что наилучшие прогностические возможности LANET в основном зависят от способности модели обнаруживать взаимосвязи между ярлыками, а не от построения работы со временем и порядка размещения корзин.

5 Заключение

В этой работе рассматривается проблему прогнозирования временных наборов: учитывая историю наборов с временными метками, состоящих из произвольного числа категориальных меток, цель состоит в том, чтобы предсказать набор меток для следующего события. Для решения этой проблемы мы предлагаем модель LANET, который агрегирует историческую информацию в векторные представления, чего нет в других существующих моделях. Особый взгляд на доступную информацию позволяет более эффективно фиксировать временные зависимости и метки. Наш метод демонстрирует наилучшую производительность на четырех рассмотренных наборах данных, превосходя подход SOTA и обеспечивая улучшение на 65% с точки зрения взвешенного F1 на одном из наборов данных. Что касается ограничений,

то LANET показывает стабильно высокие результаты, особенно в наборах данных с объемом словаря меток до 200.

Список литературы

[1] Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor Tsang. The emerging trends of multi-label learning. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[2] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 647–657, 2019.

[3] Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. Label-specific document representation for multi-label text classification. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 466–475, 2019.

[4] Adane Nega Tarekegn, Mario Giacobini, and Krzysztof Michalak. A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965, 2021.

[5] Thomas Hartvigsen, Cansu Sen, Xiangnan Kong, and Elke Rundensteiner. Recurrent halting chain for early multi-label classification. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1382–1392, 2020.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[7] Stephen Grossberg. Recurrent neural networks. *Scholarpedia*, 8(2):1888, 2013.

[8] Wenyu Zhang, Devesh K Jha, Emil Laftchiev, and Daniel Nikovski. Multi-label prediction in time series data using deep neural networks. *arXiv preprint arXiv:2001.10098*, 2020.

[9] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013.

[10] Xiao Shou, Tian Gao, Shankar Subramaniam, Debarun Bhattacharjya, and Kristin Bennett. Concurrent multi-label prediction in event streams. In *AAAI Conference on Artificial Intelligence*, 2023.

[11] Le Yu, Zihang Liu, Leilei Sun, Bowen Du, Chuanren Liu, and Weifeng Lv. Continuous-time user preference modelling for temporal sets prediction. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[12] Fan Zhang, Shuai Wang, Yongjie Qin, and Hong Qu. Conv-based temporal sets prediction for next-basket recommendation. In *2023 International Conference on Frontiers of Robotics and Software Engineering (FRSE)*, pages 419–425. IEEE, 2023.

[13] Mozhdeh Ariannezhad, Ming Li, Sebastian Schelter, and Maarten de Rijke. A personalized neighborhood-based model for within-basket recommendation in grocery shopping. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 87–95, 2023.

[14] V Ramanjaneyulu Yannam, Jitendra Kumar, Tejaswini Vankayala, and Korra Sathya Babu. Hybrid approach for next basket recommendation system. *International Journal of Information Technology*, 15(3):1733–1740, 2023.

[15] Cheng-Yuan Liou, Wei-Chen Cheng, Jiun-Wei Liou, and Daw-Ran Liou. Autoencoder for words. *Neurocomputing*, 139:84–96, 2014.

[16] Tengshuo Song, Feng Guo, Haoran Jiang, Wenyun Ma, Zhenbao Feng, and Lei Guo. Hgat-br: Hyperedge-based graph attention network for basket recommendation. *Applied Intelligence*, 53(2):1435–1451, 2023.

[17] Luuk Van Maasakkers, Dennis Fok, and Bas Donkers. Next-basket prediction in a high-dimensional setting using gated recurrent units. *Expert Systems with Applications*, 212:118795, 2023.

[18] Raanan Y Rohekar, Yaniv Gurwicz, and Shami Nisimov. Causal interpretation of self-attention in pre-trained transformers. *Advances in Neural Information Processing Systems*, 36, 2024.

[19] Thomas Richardson and Peter Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.