

# Label Attention Network для последовательной классификации по нескольким меткам

Галина Боева

Московский физико-технический институт

*Курс:* Моя первая научная статья /M05-304

*Эксперт:* к.ф-м.н А. Зайцев

2024

## Проблема

Современные подходы фокусируются на архитектуре преобразования последовательных данных, вводящей self-attention к элементам в последовательности. В этом случае мы учитываем временные взаимодействия событий, но теряем информацию о взаимозависимостях меток.

## Цель работы

Создание подхода, основанного на механизме собственного внимания над метками, предшествующими прогнозируемому шагу.

## Задачи работы

- 1) изучение существующих моделей, работающих в области предсказаний множества меток
- 2) разработка метода на основе внимания для предсказания множества меток
- 3) валидация разработанных методов
- 4) обоснование причинно-следственных связей с помощью построения графа на основе внимания

- ▶ Классификация с несколькими метками.  
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.  
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.
- ▶ Основные подходы для задачи классификации с несколькими метками.  
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.  
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.

Пусть  $\mathcal{U} = \{u_1, u_2, \dots, u_N\}$  - это набор из  $N$  элементов.

Каждый элемент  $u_i, 1 \leq i \leq N$ , связан с последовательностью временных множеств  $\mathcal{S}_i = \{s_i^1, s_i^2, \dots, s_i^T\}$ , 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  $\mathcal{Y} = \{y_1, y_2, \dots, y_L\}$  размера  $L$ .

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

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

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

## Предложенный метод

Пусть  $\mathbf{X} \in \mathbb{R}^{L \times D}$  — матрица представлений всех меток из словаря  $\mathcal{Y} = \{y_1, y_2, \dots, y_L\}$ . Для каждой временной метки  $j, 1 \leq j \leq T$  создается временное представление  $\mathbf{t}_j \in \mathbb{R}^D$ , как это сделано в [1]. 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:

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

Тогда:

$$\mathbf{G} = \mathbf{X} \oplus \mathbf{Z}. \quad (3)$$

Для выявления зависимостей меток  $\tilde{\mathbf{G}}$ :

$$\tilde{\mathbf{G}} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{2D}}\right)\mathbf{V}. \quad (4)$$

Слой предсказания:

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

Рассматриваемая функция потерь:

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

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  $\hat{\mathbf{f}}^{(l)}$ .

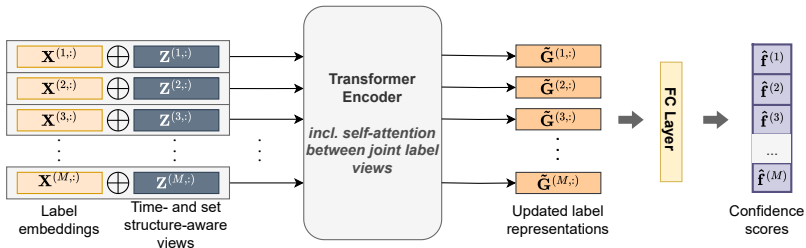


Рис.: Общий пайплайн получения глобальных представлений

Таблица: 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.

## Вычислительный эксперимент: Основные результаты

**Таблица:** 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 $\uparrow$	Weighted ROC-AUC $\uparrow$	Hamming Loss $\downarrow$
Mim	SFCNTSP	0.3791 $\pm$ 0.0081	0.7034 $\pm$ 0.0024	0.0377 $\pm$ 0.0004
	DNNTSP	0.3928 $\pm$ 0.0030	0.6926 $\pm$ 0.0003	0.0365 $\pm$ 0.0003
	GPTopFreq	0.4291 $\pm$ 0.0073	0.6912 $\pm$ 0.0028	0.0398 $\pm$ 0.0005
	TCMBN	0.4979 $\pm$ 0.0180	0.8670 $\pm$ 0.0095	0.0305 $\pm$ 0.0008
	LANET(ours)	<b>0.8214 <math>\pm</math> 0.0224</b>	<b>0.9852 <math>\pm</math> 0.0023</b>	<b>0.0220 <math>\pm</math> 0.0001</b>
Ins	SFCNTSP	0.1672 $\pm$ 0.0112	0.6852 $\pm$ 0.0448	0.0581 $\pm$ 0.0004
	DNNTSP	0.4160 $\pm$ 0.0009	0.7913 $\pm$ 0.0004	0.0541 $\pm$ 0.0002
	GPTopFreq	0.4087 $\pm$ 0.0079	0.7736 $\pm$ 0.0039	0.0529 $\pm$ 0.0008
	TCMBN	0.3687 $\pm$ 0.0065	0.8187 $\pm$ 0.0030	0.0530 $\pm$ 0.0005
	LANET(ours)	<b>0.6159 <math>\pm</math> 0.0029</b>	<b>0.9445 <math>\pm</math> 0.0008</b>	<b>0.0474 <math>\pm</math> 0.0003</b>



# Вычислительный эксперимент: Дополнительные исследования

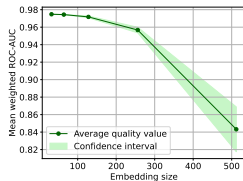


Рис.: Зависимость качества LANET от размера векторных представлений.

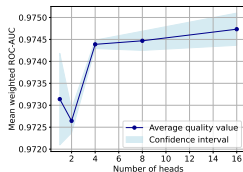


Рис.: Зависимость качества LANET от количества голов во внимании.

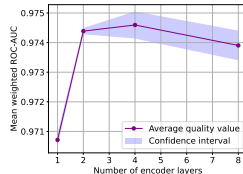
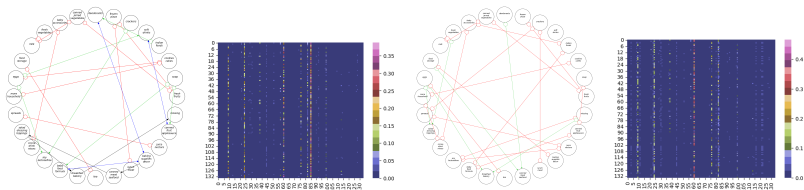


Рис.: Зависимость качества LANET от количества слоев энкодера.

# Graph attention



**Рис.:** Интерпретация взаимосвязи [2] 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.

Статья подана в апреле 2024 года **на конференцию ECAI**.

- ▶ Проведены исследования по анализу различных наборов данных, используемых при сравнении реализованной модели LANET.
- ▶ Проведены ряд экспериментов для задачи классификации с несколькими метками на двух различных выборках и сравнение с базовыми подходами в данной области.
- ▶ Проведен анализ причинно-следственных связей в self-attention, где используется графовый подход на основе построения PAG для взаимосвязи меток.
- ▶ Проведена оценка метрики в зависимости от гиперпараметра, отвечающего за размер входных представлений, количество голов во внимании и также количества слоев энкодера.

1. Kovtun E., Boeva G. *Label Attention Network for sequential multi-label classification: you were looking at a wrong self-attention* // arXiv — 2023.
2. Zhuzhel, V., Grabar, V., Boeva, G., Zabolotnyi, A., Stepikin, A., Zholobov, V., Ivanova, M., Orlov, M., Kireev, I., Burnaev, E., Rivera-Castro, R., Zaytsev, A.: *Continuous-time convolutions model of event sequences* (2023)



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.*



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.*