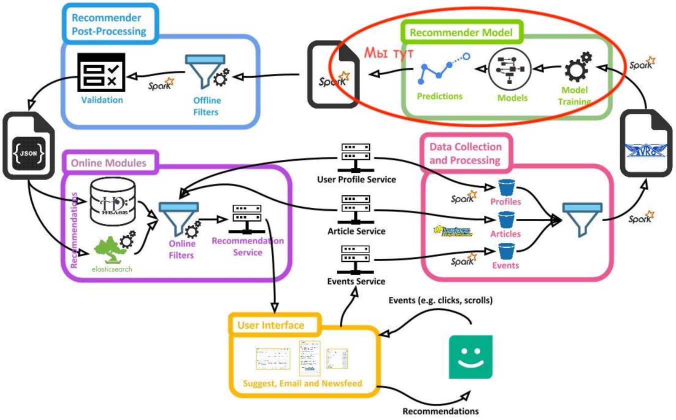
Большие рекомендательные модели

Сергей Малышев

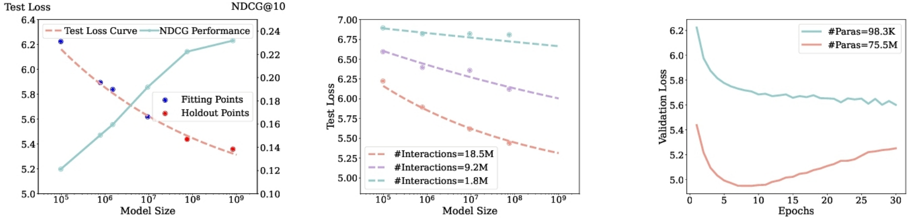
16 апреля 2025 г.

Контекст


Recommender Post-Processing
Validation
Spark
Offline Filters
Мы тут
Spark
Recommender Model
Predictions
Models
Model Training
Spark
JSON
Online Modules
Recommendations
elasticsearch
Online Filters
Recommendation Service
User Profile Service
Article Service
Events Service
Data Collection and Processing
Spark
Profiles
Articles
Events
Spark
Spark
AVRO
User Interface
Suggest, Email and Newsfeed
Events (e.g. clicks, scrolls)
Recommendations

Мотивация

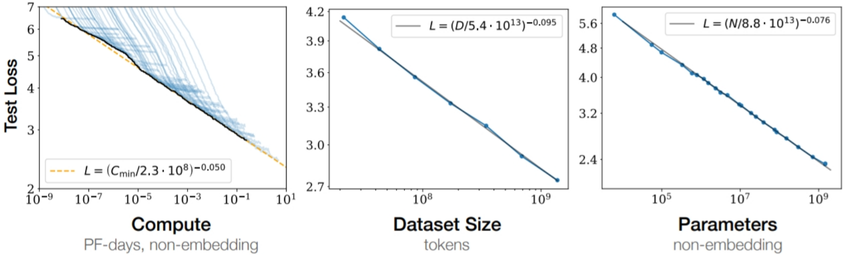
Какие сейчас проблемы у стандартных рекомендательных моделей?



(a) Model Scaling (b) Data Scaling (c) Validation Loss with Data Repetition

Проблема №1: Не можем запихнуть все айтемы в одну модель
Проблема №2: Качество моделей перестает расти при увеличении параметров[ZHL+24] / размера датасета

Чего мы хотим?



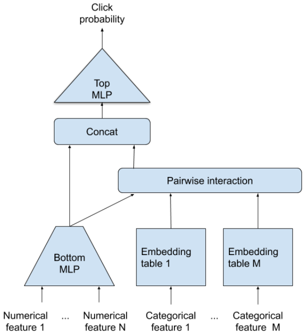
Потребность №1: Хотим передавать все айтемы в модель

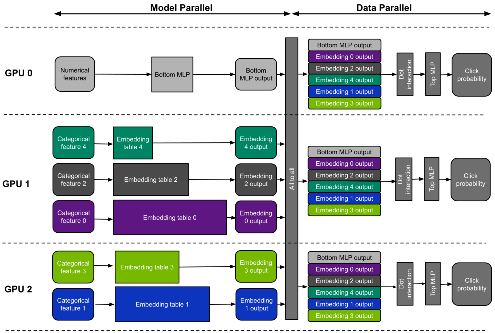
Потребность №2: Хотим растить качество модели на уровне линейного роста по мере роста датасета / увеличения количества параметров [1]

[1] ttps://www.lesswrong.com/w/scaling-laws

Архитектуры больших рек. моделей

Deep Learning Recommendation Model for Personalization and Recommendation Systems [NMS+19]

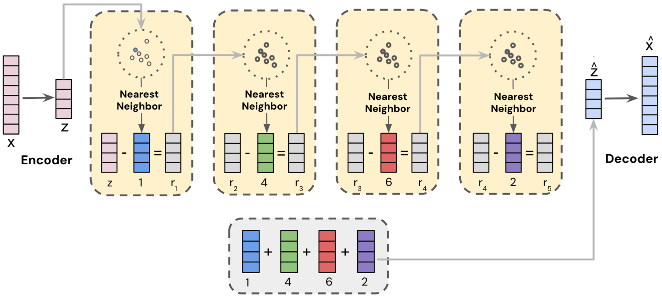




Идея Простая архитектура, но много параметров в эмбеддингах[2]

[2] https://catalog.ngc.nvidia.com/orgs/nvidia/resources/dlrm_for_pytorch

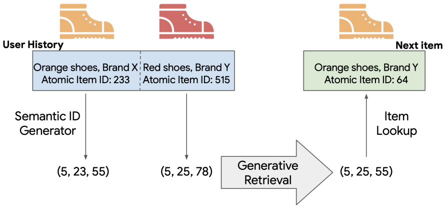
Semantic ID [SVM+24]

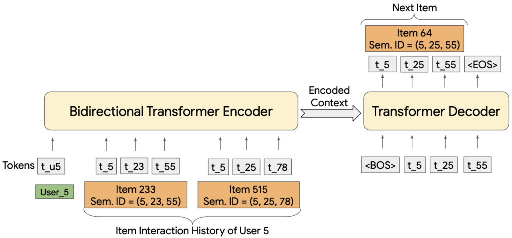


Идея

* С помощью RQ VAE кодируем айтемы последовательностью целых чисел
* Используем как эмбеддинг для другой модели
* Решаем проблему холодного старта

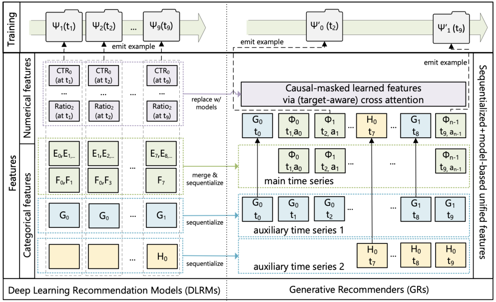
Recommender Systems with Generative Retrieval [RMS+23]

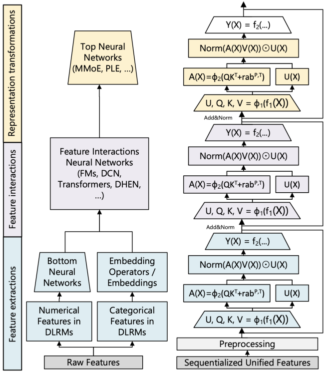




Идея
* Строим seq2seq модель из Semantic-IDs получая end2end подход
* Решаем проблему холодного старта
* Делаем рекомендации разнообразнее

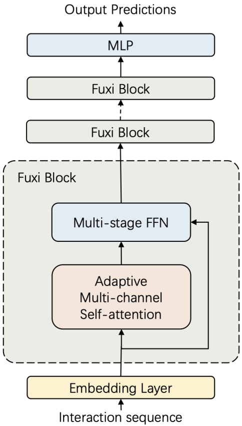
HSTU [ZLL+24]

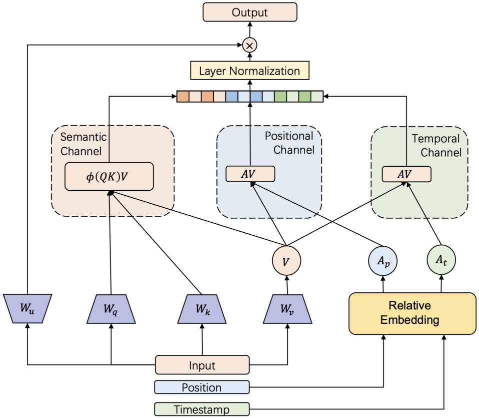




Идея
* Меняем постановку на генеративную
* Используем изменения контекста
* Вместо фичей используем экшены пользователей
* Вместо DLRM используем HSTU блок с attention без FFN

FuXi-alpha [YGC+25]





Идея
* Берем HSTU за основу
* Возвращаем туда FFN
* Свои механизмы attention для позиционной и темпоральной компоненты

Немного про метрики и масштабируемость

Dataset	MovieLens-1M					MovieLens-20M					KuaiRand				
Model	NG@10	NG@50	HR@10	HR@50	MRR	NG@10	NG@50	HR@10	HR@50	MRR	NG@10	NG@50	HR@10	HR@50	MRR
BPRMF	0.0607	0.1027	0.1185	0.3127	0.0556	0.0629	0.1074	0.1241	0.3300	0.0572	0.0248	0.0468	0.0520	0.1560	0.0235
GRU4Rec	0.1015	0.1460	0.1816	0.3864	0.0895	0.0768	0.1155	0.1394	0.3177	0.0689	0.0289	0.0531	0.0597	0.1726	0.0275
NARM	0.1350	0.1894	0.2445	0.4915	0.1165	0.1037	0.1552	0.1926	0.4281	0.0910	0.0411	0.0747	0.0836	0.2399	0.0387
SASRec	0.1594	0.2187	0.2824	0.5500	0.1375	0.1553	0.2119	0.2781	0.5353	0.1330	0.0486	0.0877	0.0978	0.2801	0.0454
LLaMa	0.1620	0.2207	0.2926	0.5591	0.1373	0.1640	0.2206	0.2915	0.5476	0.1402	0.0495	0.0878	0.0973	0.2752	0.0466
HSTU	0.1639	0.2238	0.2969	0.5672	0.1390	0.1642	0.2225	0.2909	0.5553	0.1410	0.0491	0.0861	0.0992	0.2718	0.0451
FuXi-α	0.1835	0.2429	0.3254	0.5941	0.1557	0.1954	0.2533	0.3353	0.5969	0.1677	0.0537	0.0942	0.1067	0.2951	0.0497
SASRec-Large	0.1186	0.1733	0.2183	0.4671	0.0186	0.0206	0.0379	0.0412	0.1209	0.0207	0.0285	0.0428	0.0544	0.1227	0.0258
LLaMa-Large	0.1659	0.2257	0.2990	0.5692	0.1408	0.1842	0.2412	0.3202	0.5776	0.1576	0.0494	0.0878	0.0970	0.2754	0.0466
HSTU-Large	<u>0.1844</u>	<u>0.2437</u>	<u>0.3255</u>	<u>0.5929</u>	<u>0.1568</u>	<u>0.1995</u>	<u>0.2572</u>	<u>0.3407</u>	<u>0.6012</u>	<u>0.1714</u>	<u>0.0494</u>	<u>0.0883</u>	<u>0.0990</u>	<u>0.2799</u>	<u>0.0460</u>
FuXi-α-Large	**0.1934**	**0.2518**	**0.3359**	**0.5983**	**0.1651**	**0.2086**	**0.2658**	**0.3530**	**0.6113**	**0.1792**	**0.0555**	**0.0963**	**0.1105**	**0.2995**	**0.0510**

360Brew [FSE+25]

Instruction:
You are provided a member's profile and a set of jobs, their description, and interactions that the member had with the jobs. For each past job, the member has taken one of the following actions: applied, viewed, dismissed, or did not interact.
Your task is to analyze the job interaction data along with the member's profile to predict whether the member will apply, view, or dismiss a new job referred to as the "Question" job.

Note: Focus on skills, location, and years of experience more than other criteria. In your calculation, assign a 30% weight to the relevance between the member's profile and the job description, and a 70% weight to the member's historical activity.

Member Profile:
Current position: software engineer, current company: Google, Location: Sunnyvale, California.

Past job interaction data:
Member has applied to the following jobs: [Title: Software Engineer, Location: New York, Country: USA, Company: Meta, Description: . . .]
Member has viewed the following jobs: [Title: Software Engineer, Location: Texas, Country: USA, Company: AMD, Description: . . .]

Question:
Will the member apply to the following job: [Title: Software Engineer, Location: Seattle, Country: USA, Company: Apple, Description: . . .]

Answer:
The member will apply

Идея
* Дообучаем предобученную LLM-ку Mixtral на постановку рекомендаций
* Continuous pretraining, instruction fine-tuning, supervised fine-tuning

Итоги

Итоги

Плавненько движемся в сторону генеративной постановки рекомендаций по аналогии с LLM

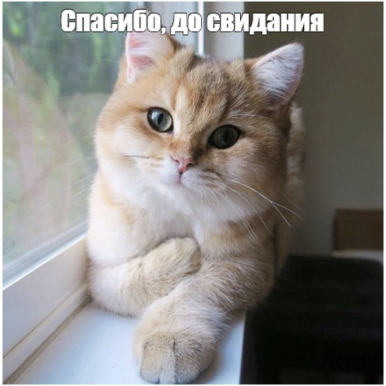
А так же в сторону foundation моделей

Получаем за счет этого буст по качеству

Попутно решаем проблему холодного старта

Но пока все это есть лишь у единиц, подходы должны настояться

До следующего раза [ZLL+24]



https://t.me/mlvok

Литература I

- Hamed Firooz, Maziar Sanjabi, Adrian Englhardt, Aman Gupta, Ben Levine, Dre Olgiati, Gungor Polatkan, Iuliia Melnychuk, Karthik Ramgopal, Kirill Talanine, et al., *360brew: A decoder-only foundation model for personalized ranking and recommendation*, arXiv preprint arXiv:2501.16450 (2025).

- Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G Azzolini, et al., *Deep learning recommendation model for personalization and recommendation systems*, arXiv preprint arXiv:1906.00091 (2019).

- Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, et al., *Recommender systems with generative retrieval*, Advances in Neural Information Processing Systems **36** (2023), 10299–10315.

Литература II

- Anima Singh, Trung Vu, Nikhil Mehta, Raghunandan Keshavan, Maheswaran Sathiamoorthy, Yilin Zheng, Lichan Hong, Lukasz Heldt, Li Wei, Devansh Tandon, et al., *Better generalization with semantic ids: A case study in ranking for recommendations*, Proceedings of the 18th ACM Conference on Recommender Systems, 2024, pp. 1039–1044.
- Yufei Ye, Wei Guo, Jin Yao Chin, Hao Wang, Hong Zhu, Xi Lin, Yuyang Ye, Yong Liu, Ruiming Tang, Defu Lian, et al., *Fuxi-alpha : Scaling recommendation model with feature interaction enhanced transformer*, arXiv preprint arXiv:2502.03036 (2025).
- Gaowei Zhang, Yupeng Hou, Hongyu Lu, Yu Chen, Wayne Xin Zhao, and Ji-Rong Wen, *Scaling law of large sequential recommendation models*, Proceedings of the 18th ACM Conference on Recommender Systems, 2024, pp. 444–453.

Литература III

Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Michael He, et al., *Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations*, arXiv preprint arXiv:2402.17152 (2024).