Recombee

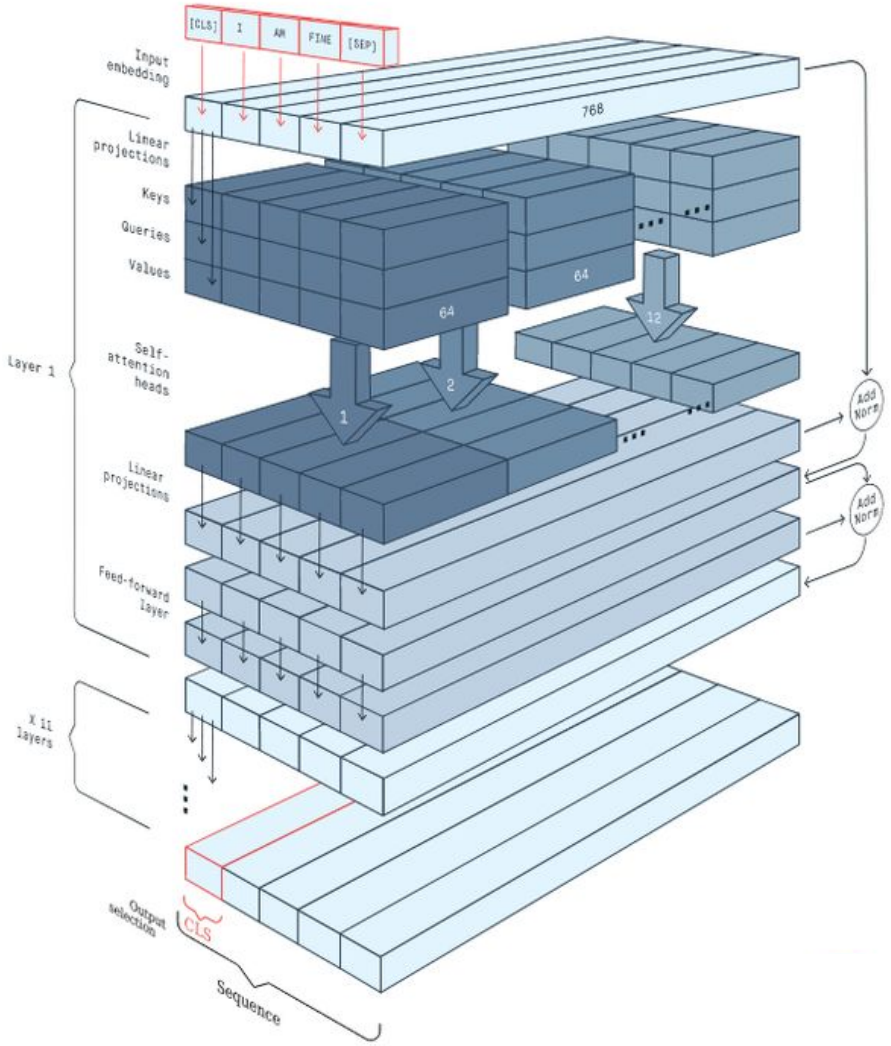
Jak velké jazykové modely pomáhají uživatelům s doporučeními relevantních položek?

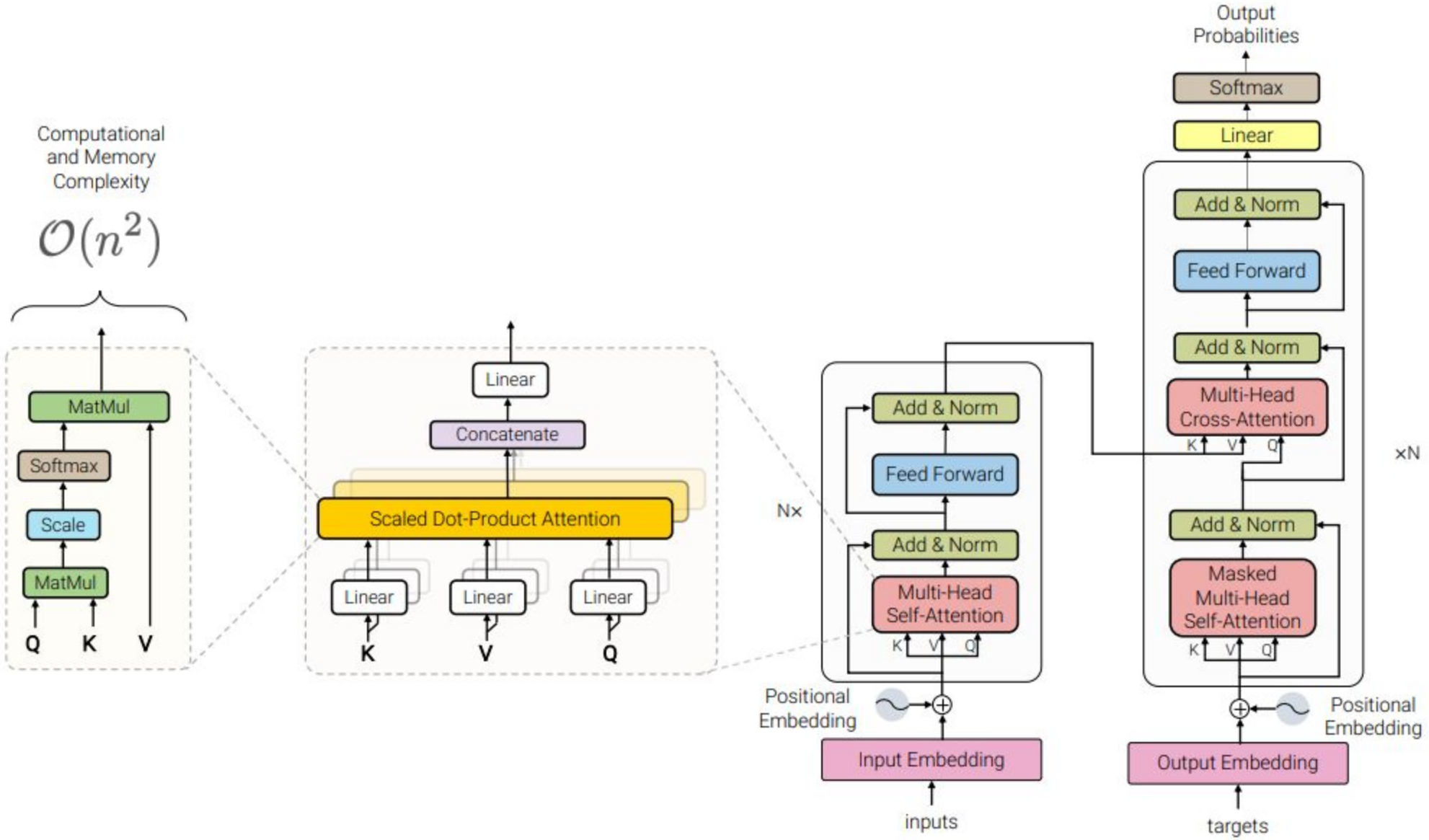
Pavel Kordík

Aldays, 25. ledna 2024

Jak fungují velké jazykové modely?

- Princip autoencodéru
- Trénovaní na textech
- Pozornostní mechanismus
- Dedodér = generátor
- Délka kontextu




Computational and Memory Complexity
$\mathcal{O}(n^2)$
MatMul
Softmax
Scale
MatMul
Q
K
V
Linear
Concatenate
Scaled Dot-Product Attention
Linear
Linear
Linear
K
V
Q
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Self-Attention
K V Q
Positional Embedding
Input Embedding
inputs
Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Cross-Attention
K V Q
Add & Norm
Masked Multi-Head Self-Attention
K V Q
×N
Positional Embedding
Output Embedding
targets

Cíl: porozumět přirozenému jazyku

A doporučit odpovídající položky z katalogu …

My wife who loves shopping summer fashion for her grandchildren

$1 — No Limit

No Limit

First Impressions Baby Girls Gingham Tunic $3, Macy's Kids Tiptoe...

$3 ~~$10~~

Macy's

First Impressions Baby Girls Dotted Cotton Sunsuit, Created for...

$8.40 ~~$10~~

Macy's

First Impressions Baby Girls Striped One Piece Swimsuit and Hat, 2...

$26.88 ~~$35~~

Macy's

Tommy Hilfiger Big Girls Overall - Med Wash

$34.88 ~~$41~~

Macy's

Polo Ralph Lauren Big Girls Plaid Cotton Twill Shirtdress - Red, Green...

$39.75 ~~$47~~

Macy's

Polo Ralph Lauren Big Girls Plaid Cotton Fun Shirtdress - Multi

$39.75 ~~$46~~

Macy's

Style & Co Women's Belted Denim Shirtdress, Created for Macy's -...

$27.80 ~~$32~~

Macy's

First Impressions Baby Girls Strawberry Swim Shirt, Shorts and Hat, 3...

$26.88 ~~$34~~

Macy's

Bluey Girls T-Shirt and French Terry Shorts Outfit Set Toddler |Chil...

$25.92 ~~$48~~

Macy's

Carter's Big Girls Plaid Cotton Flannel Shirtdress - Pink

$22.40 ~~$32~~

Macy's

Tommy Hilfiger Big Girls Flip-Sequin Long Sleeve T-shirt - Oat Heather

$18.88 ~~$26~~

Macy's

Ny Collection Petite Sweetheart Neck Chambray Sundress -...

$36.72 ~~$46~~

Macy's

Cíl: porozumět přirozenému jazyku

A doporučit odpovídající položky z katalogu …

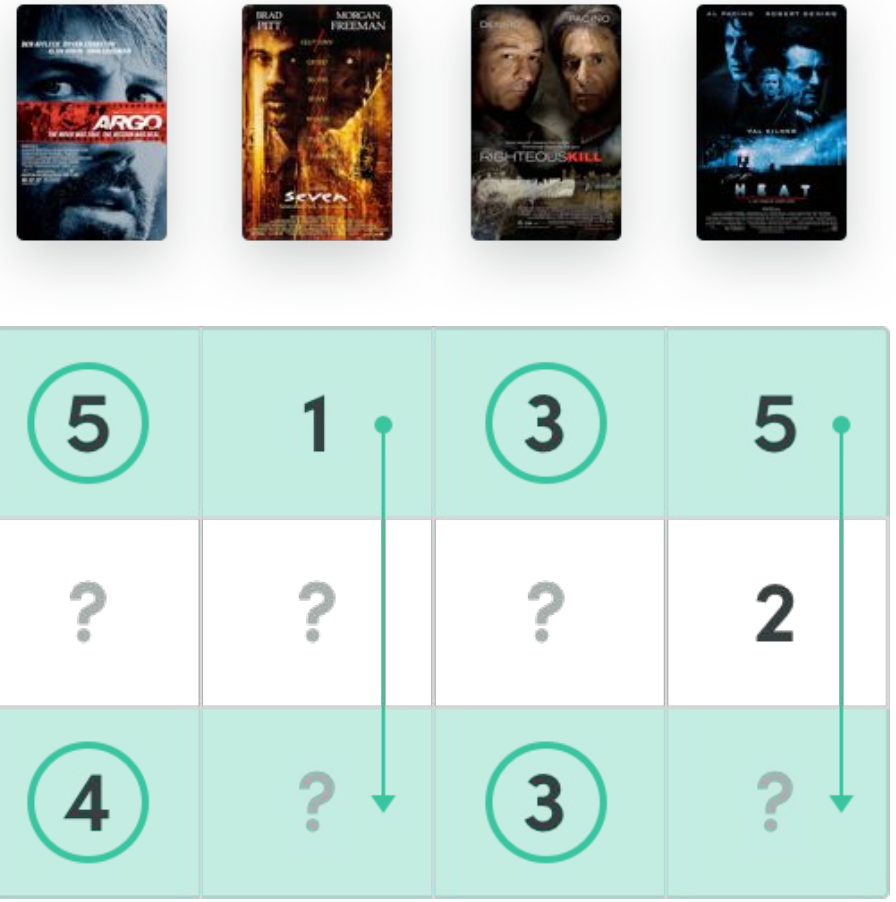
Nutno zkombinovat velké jazykové modely a doporučovací systémy!

Doporučovací systémy

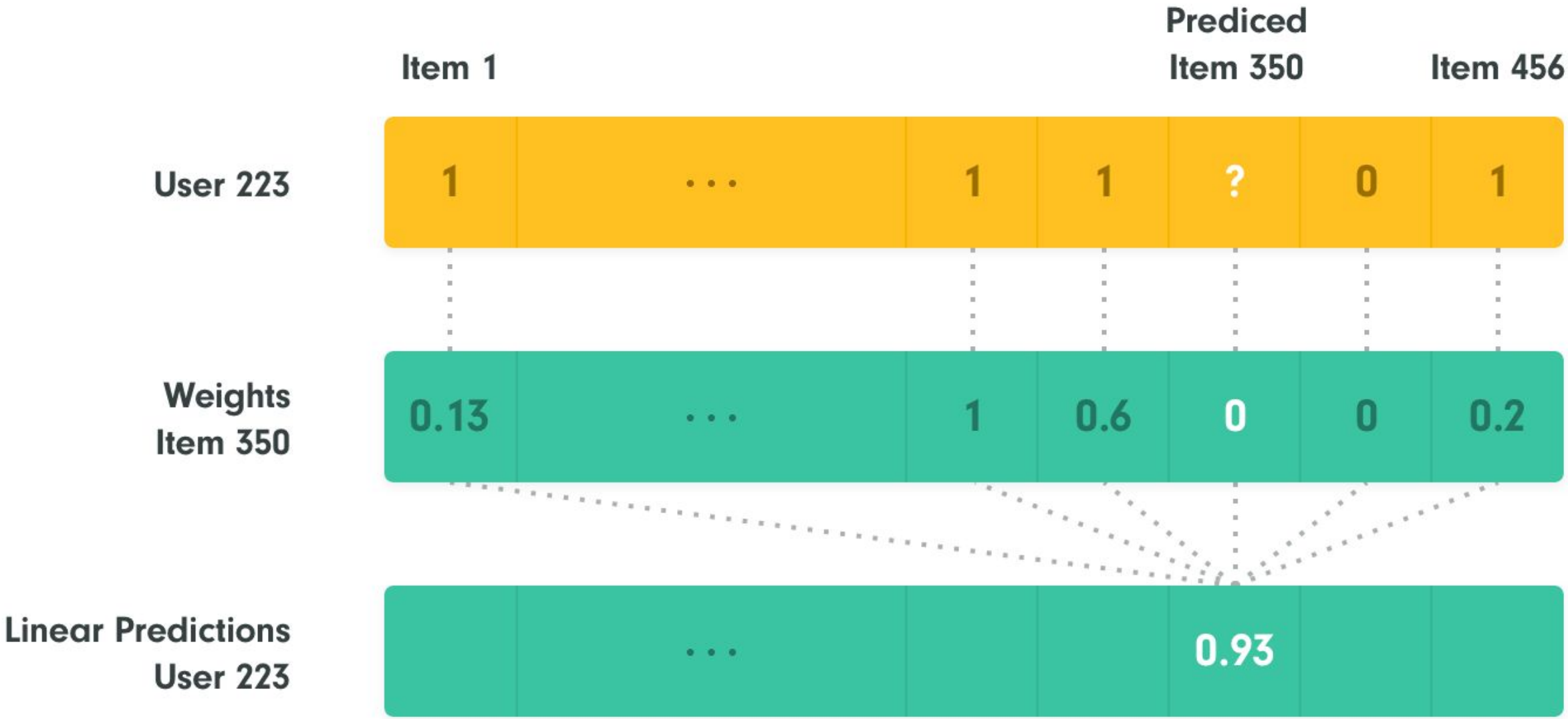
Kolaborativní filtrování

- Z pohledu uživatele
- Z pohledu položky

Podobnosti na základě atributů



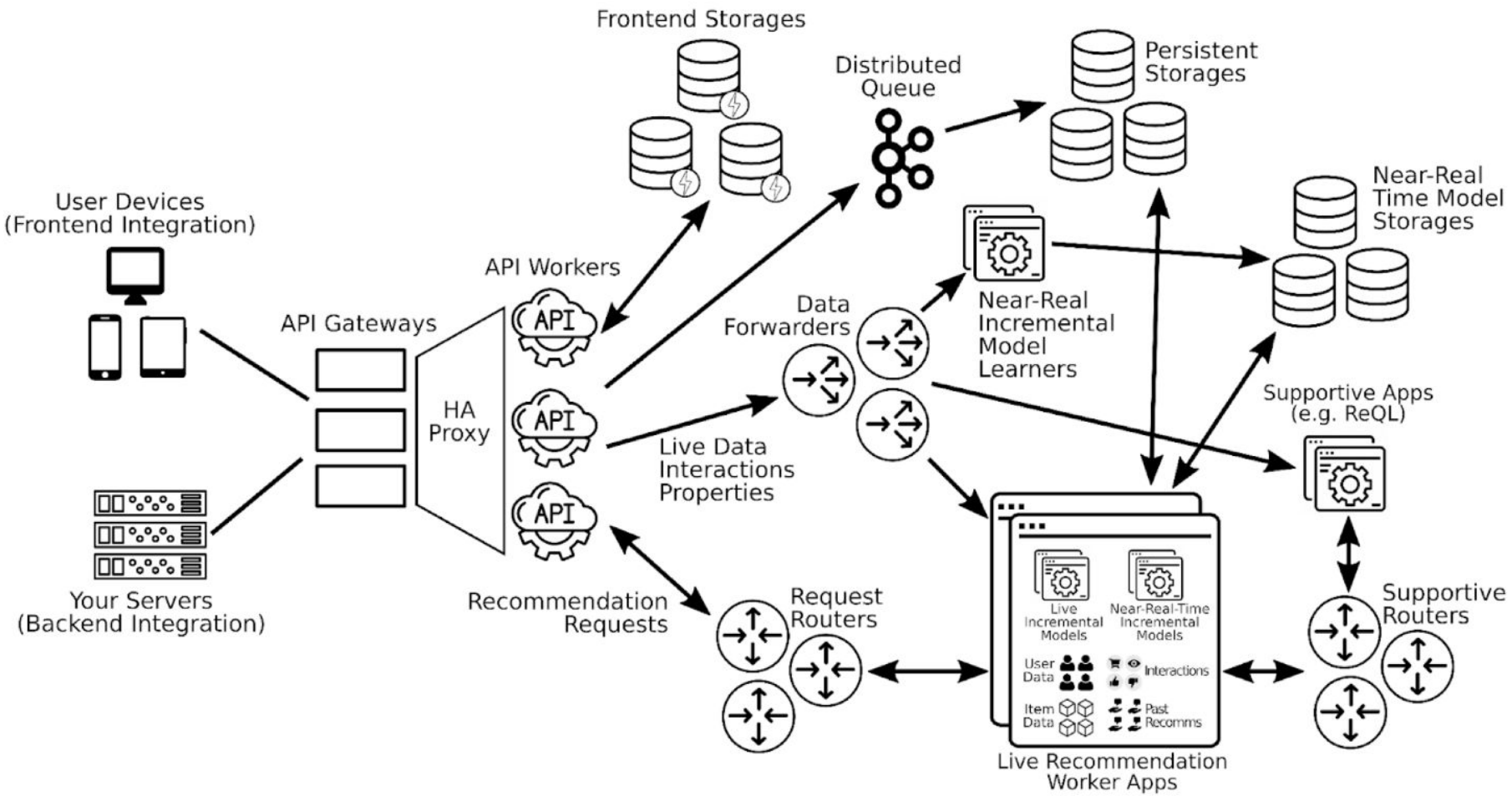
Lineární modely



Výzvy

- Umět odpovědět rychle.
- Okamžitě reagovat na změny v katalogu položek.
- Zvládnout obsloužit miliony uživatelů a stovky tisíc položek.
- Vypočítat odpověď za rozumnou cenu.

Recombee Production Cloud Infrastructure for Universal RSaS



Recombee

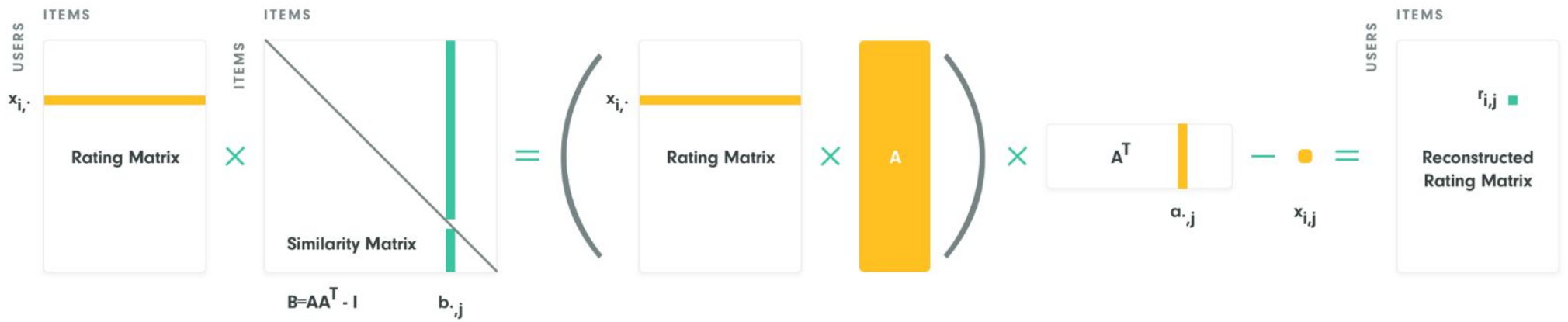
PUBLICATION

Making Linear Autoencoders Work for Large Scale Recommendation Systems

Weight matrix is factorized into low-rank plus sparse structure.

Given the order of parenthesis, neither for optimizing the model nor predicting the recommendations, it will be necessary to store a dense matrix with dimensions superior to i×d.

ELSA



https://www.recombee.com/blog/making-linear-autoencoders-work-for-large-scale-recommendation-systems.html

Recombee

A co škálování velkých jazykových modelů?

- Efektivní transformery
- LoRA
- QLoRA

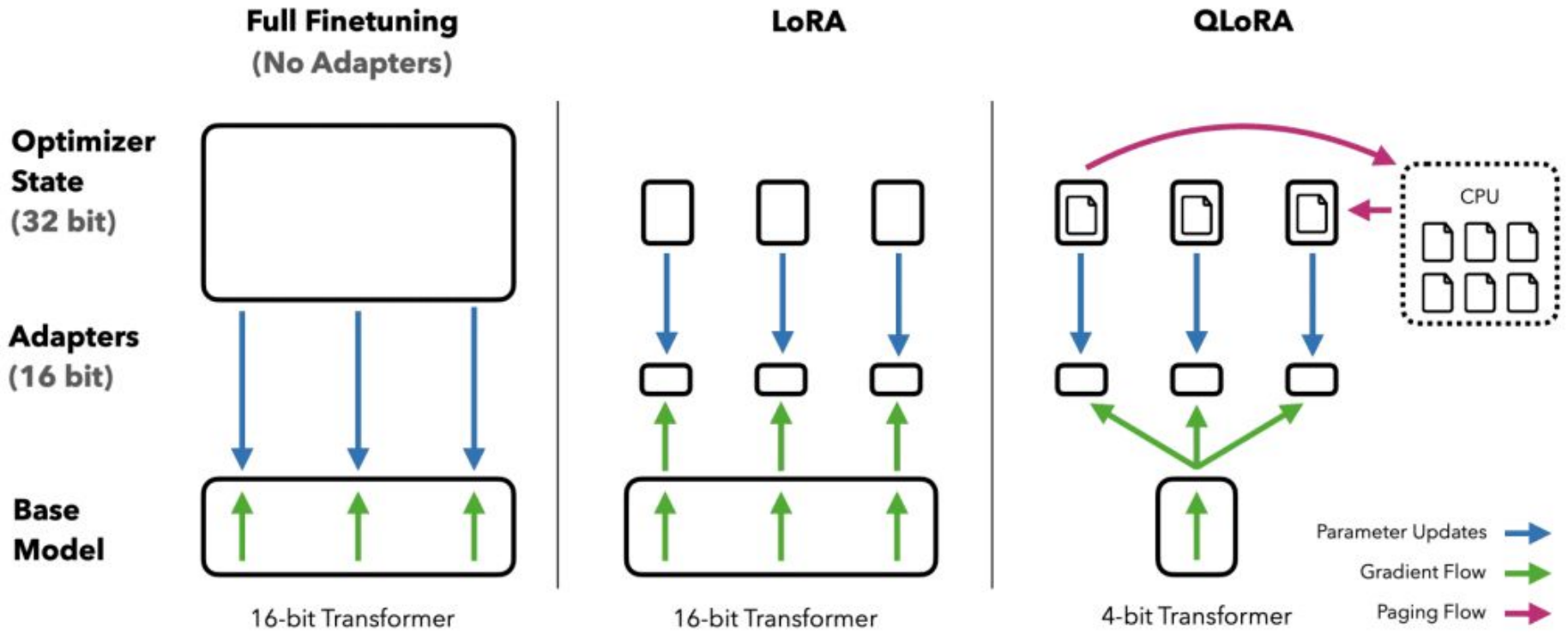


Figure 1: Different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.

NVidia DGX v RecombeeLabu

- Společná výzkumná laboratoř Recombee a ČVUT FIT
- Trénování a evaluace menších efektivních modelů (např. ELSA)
- Studentské práce a projekty

Příklad studentského projektu

Refining SBERT-based Semantic Search: Ethical Controls and Content Precision

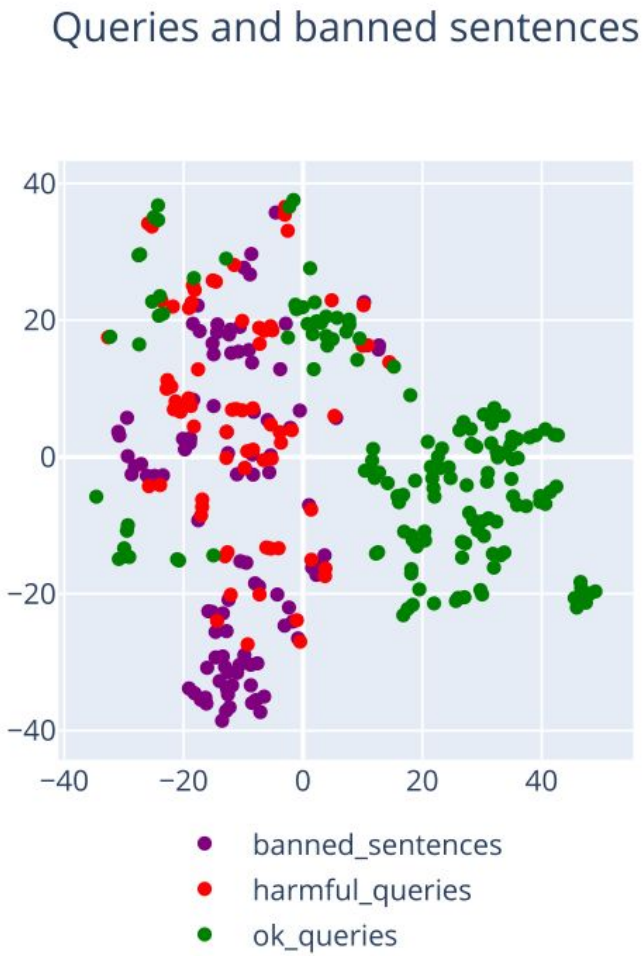
3 Data and model

Data for training the SBERT model were obtained from the client. The various queries used for testing the ethical filters were designed by the author, using several sources such as [2] or [1].

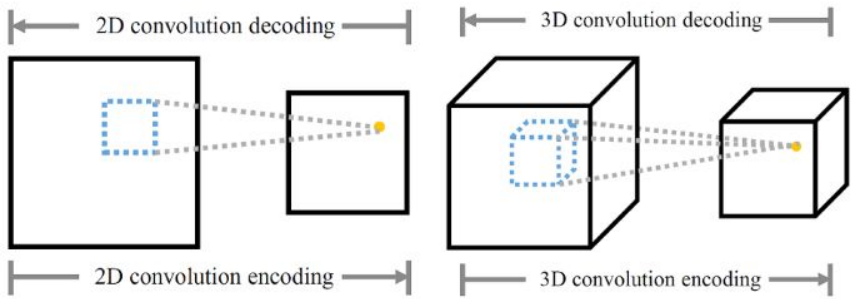
The SBERT model is derived from a pre-trained general MPNetModel from the SentenceTransformers library. It was then trained using the data provided by the customer to better perform on this specific task. It outputs 512-dimensional embedding vectors.

4 Ethical controls and content precision

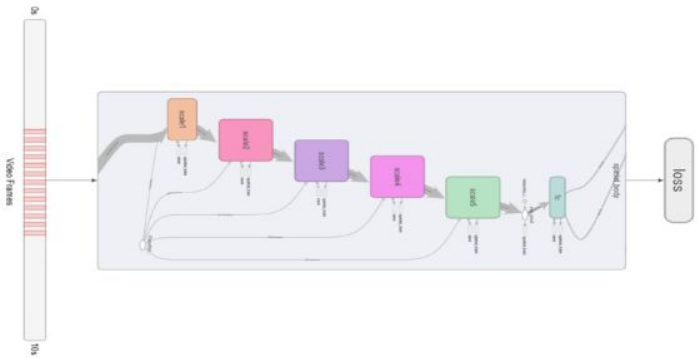
The idea of ethical control is the follows: Create a list of banned sentences, that would cover harmful topics. If a query similarity to some banned sentence is higher than a given threshold, it should be ignored. To measure the similarity between two queries, it is first necessary to get the embedding



Neural Video Embeddings for Better Movie Similarities



3D Convolutional Networks



Movie catalogue

mp4

all feature data (HDF5)

GPU servers

- Audio / video preprocessing
- Shot segmentation
- Scene detection
- Feature extraction

Joint project with Showmax engineering in 2018

Recombee

A to je vše …

Děkuji za pozornost

pavel.kordik@recombee.com