Interpreting Recommender Models with Sparse Autoencoders

Lyudmila Zavadskaya,
Thanakrit Lerdmatayakul,
Kseniia Kuvshinova

Skolkovo
Institute of Science
and Technology

Skoltech

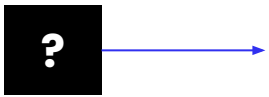
Project Description

Motivation

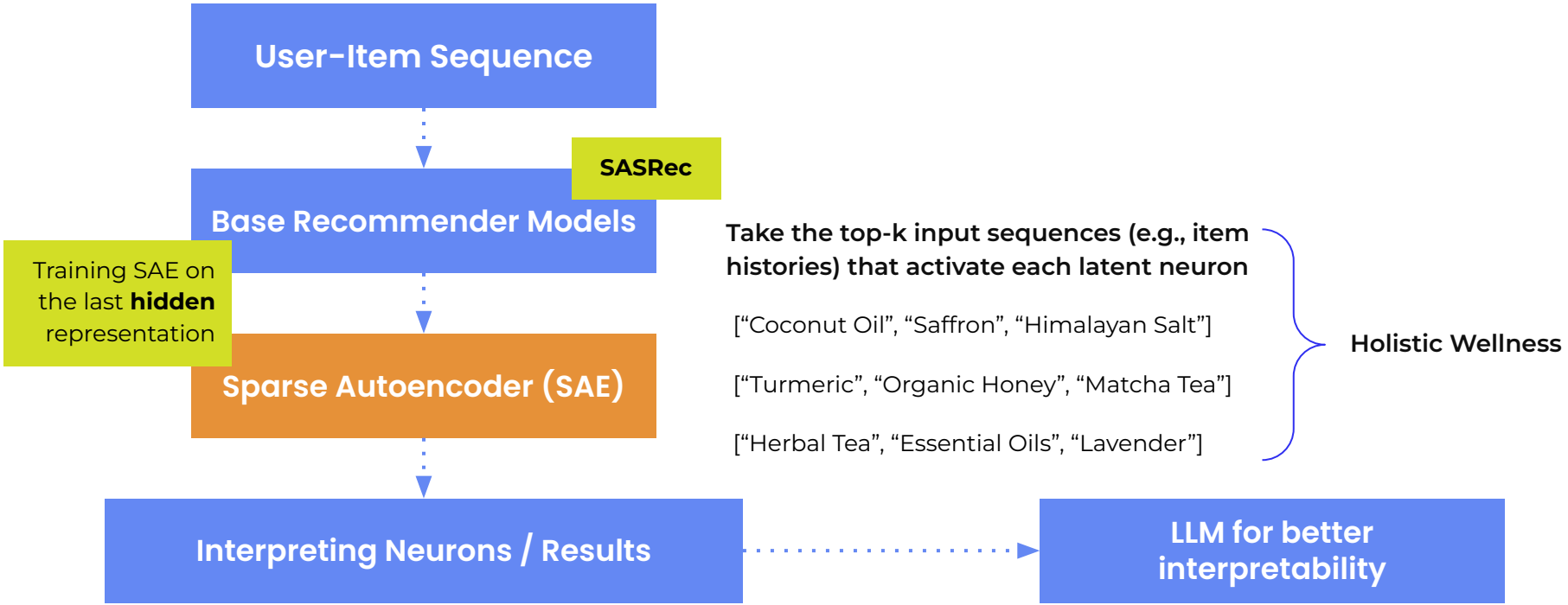
Recommender systems are powerful but often operate as black boxes, making it difficult to understand why a model made a certain recommendation.

Goals

- ➔ Apply a **Sparse Autoencoder (SAE)** to capture and interpret the internal neuron activations of a recommender model.
- ➔ Extract interpretable latent features that influence model behavior and generate human-readable explanations.

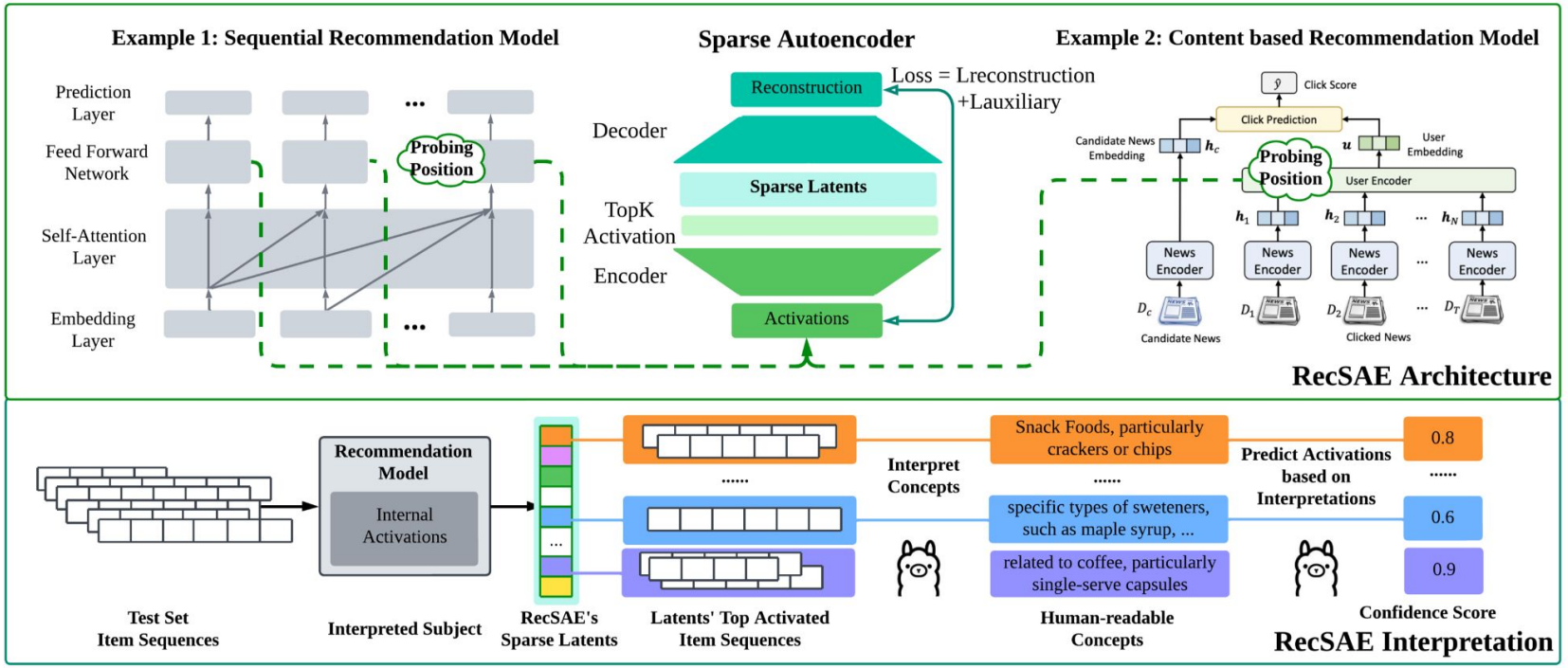


Model Architecture



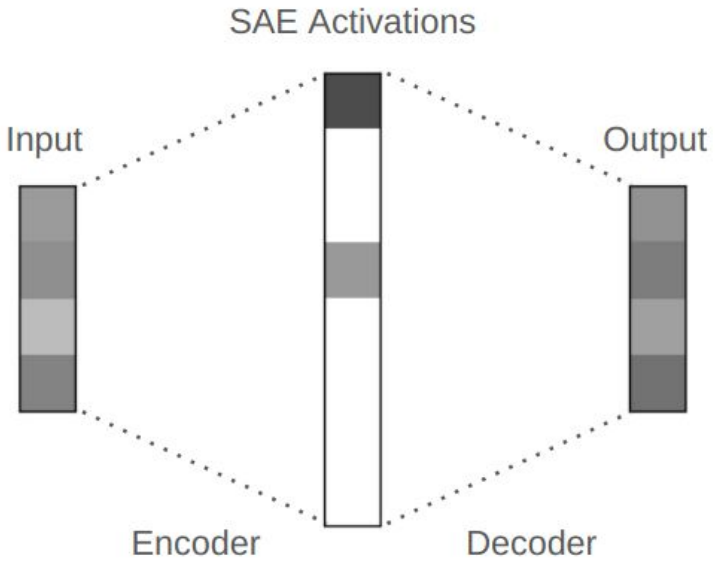
Skoltech

Model Architecture



Skoltech

SAE Loss Function



Reconstruction Loss

$$\mathcal{L}_{\text{recon}} = \left\| \mathbf{h}_u - \hat{\mathbf{h}}_u \right\|_2^2$$

Auxiliary Loss

$$\mathcal{L}_{\text{aux}} = \left\| \mathbf{e} - \mathbf{e}' \right\|_2^2$$

Total Loss

$\mathbf{e} = \mathbf{h}_u - \hat{\mathbf{h}}_u$: residual (reconstruction error)

Skoltech

Experimental setup

Task

Next item prediction - sequential recommendations

Base model

SASRec (Self-Attentive Sequential Recommendation) + ReChorus framework

Interpreting model

RecSAE (Interpret the Internal States of Recommendation Model with Sparse Autoencoder)

Dataset

MovieLens-1m

Evaluation metrics

Hit rate, NDCG

Hyperparameters

Default for ReChorus framework

$$NDCG = \frac{DCG}{IDCG}$$

$$DCG = \sum_{i=1}^{n} \frac{rel_i}{\log_2(i+1)}$$

$$IDCG = \sum_{i=1}^{n} \frac{1}{\log_2(i+1)}$$

RecSAE meets Attention (SASRec)

Dataset

MovieLens-1m

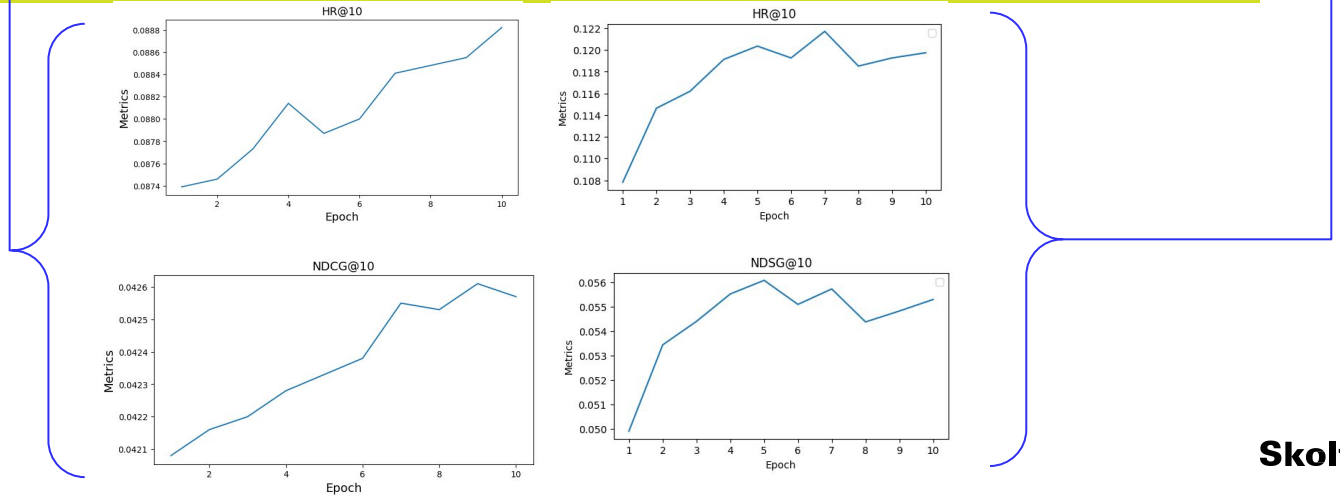
Epoch	Loss	HR@5	NDCG@5	HR@10	NDCG@10	HR@20	NDCG@20	HR@50	NDCG@50
0	SASRec_test	0.1100	0.0717	0.1628	0.0885	0.2303	0.1056	0.3546	0.1303
12	RecSAE_test	0.1086	0.0708	0.1608	0.0874	0.2342	0.1060	0.3563	0.1300

RecSAE meets Attention (SASRec)

Dataset

Grocery_and_Gourmet

Epoch		HR@5	NDCG@5	HR@10	NDCG@10	HR@20	NDCG@20	HR@50	NDCG@50
	RecSAE_test	0.0192	0.0092	0.0299	0.0135	0.0526	0.0192	0.0895	0.0264
10	RecSAE_Attention_test	0.0332	0.0206	0.0655	0.0309	0.1413	0.0498	0.47490	0.11427
10	RecSAE_Contrastive_test	0.0621	0.0369	0.1198	0.0553	0.2234	0.0811	0.53375	0.14181



Skoltech

Interpreting Latent Dimensions Through Movie Activations

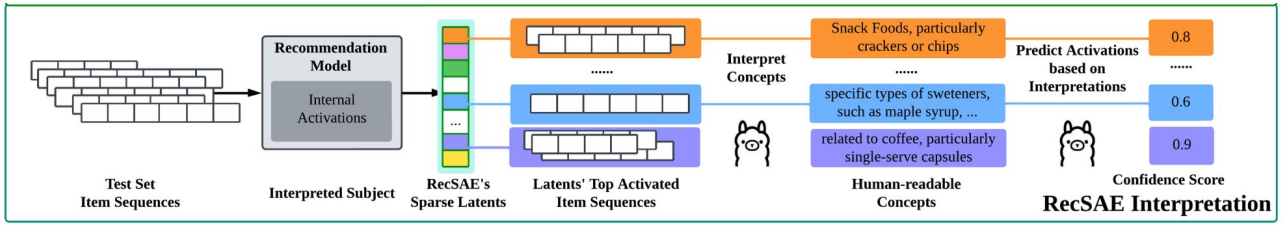
User_id 36

history "[7 1093 932 1020 1694 2340 1644 1450 2365 170 894 506 2497 488 923 2649 485 1012 3028 2192]"

Indices "[1746 2000 1637 447 1915 61 1895 1195 1263 1230 1839 1010 1131 1032 823 1462 1402 406 139 1144 1803 1166 302 1634 1820 803 1567 1834 1832 235 1682 626]"

values"[7.549864 7.1386633 6.4357576 6.0744743 5.8772316 5.746204 5.260926 5.0798707 4.398423 3.8670907 3.6049898 3.3875067 3.226031 3.2155035 3.176168 2.9819226 2.9206958 2.8021443 2.6873546 2.5557065 2.413963 2.4033062 2.2532837 2.1736991 2.138201 1.9582134 1.8653014 1.860255 1.8487558 1.8289199 1.7823701 1.76725]"

Interpreting Latent Dimensions Through Movie Activations



Concept ID 662

id	Movie	Activation Counts	
2620	Wonder Boys (2000)	[Comedy, Drama]	5
3093	Best in Show (2000)	[Comedy]	4
3083	Almost Famous (2000)	[Comedy, Drama]	4
3079	Nurse Betty (2000)	[Comedy, Thriller]	4
3030	Saving Grace (2000)	[Comedy]	4

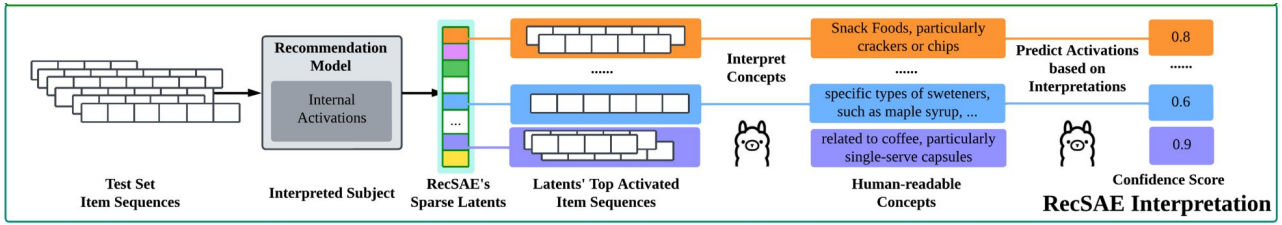
LLM

🧩 Concept 662 likely represents: From chatgpt

"Offbeat Character-Driven Comedies or Dramedies"

- A dry sense of humor
- Bittersweet tones
- Some satirical or absurd elements (like Best in Show, Nurse Betty)

Interpreting Latent Dimensions Through Movie Activations



Latent Top activated Sequence

Concept ID 112

id	Movie	Activation Counts	
2824	Gladiator (2000)	[Action, Drama]	165
3083	Almost Famous (2000)	[Comedy, Drama]	164
670	Godfather, The (1972)	[Action, Crime, Drama]	158
3093	Best in Show (2000)	[Comedy]	154

🧩 Concept likely represents:

"Dramatic Action with a Touch of Humor"

- Strong, emotional drama with intense action (e.g., Gladiator, The Godfather)
- Character-driven stories that blend seriousness with moments of levity (e.g., Almost Famous, Best in Show)
- The mix of action, drama, and comedy highlights multi-dimensional characters
- Themes of family, power, and personal growth interspersed with light-hearted humor

Interpreting Latent Dimensions Through Movie Activations

Latent Top activated Sequence

Concept ID 855

id	Movie	Activation Strength	
927	Star Wars: Episode V - The Empire Strikes Back...	[Action, Adventure, Drama, Sci-Fi, War]	534.602145
608	Dr. Strangelove or: How I Learned to Stop Worr...	[Sci-Fi, War]	479.348611
2022	Matrix, The (1999)	[Action, Sci-Fi, Thriller]	404.222781
708	Casablanca (1942)	[Drama, Romance, War]	382.004656

Interpreting Latent Dimensions Through Movie Activations

Concept ID 112: Activation Strength

id	Movie		Activation Strength
49	Usual Suspects, The (1995)	[Crime, Thriller]	824.016994
670	Godfather, The (1972)	[Action, Crime, Drama]	819.087257
253	Pulp Fiction (1994)	[Crime, Drama]	716.190738
2824	Gladiator (2000)	[Action, Drama]	705.189637

Top Concepts that Activated Movie 2620 (wonder boys) by Total Activation Strength:

Concept ID: 1637, Total Activation Strength: 1485.2351986

Concept ID: 1820, Total Activation Strength: 1105.4994348

Concept ID: 803, Total Activation Strength: 676.3621023100004

Concept ID: 1863, Total Activation Strength: 527.87242494

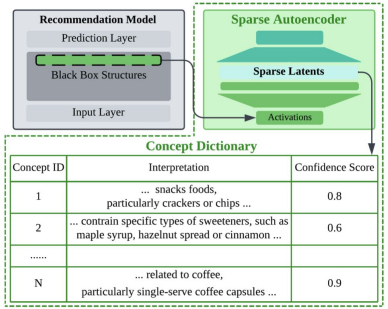
Concept ID: 112, Total Activation Strength: 503.6987066699998

1637, 447, 803, etc activate a lot of movies!

Conclusion

Latent Feature Interpretation

- RecSAE successfully identifies over 150 interpretable concepts from the models.
- Revealing what each model learns and how it represents information, offering potential insights into the unique characteristics of each architecture.
- The RecSAE framework is highly flexible, as the detector module can be integrated with various types of recommendation models, and the interpretation process is both automated and scalable



Skoltech

Thx

Skoltech