Correcting the LogQ Correction

Revisiting Sampled Softmax for Large-Scale Retrieval

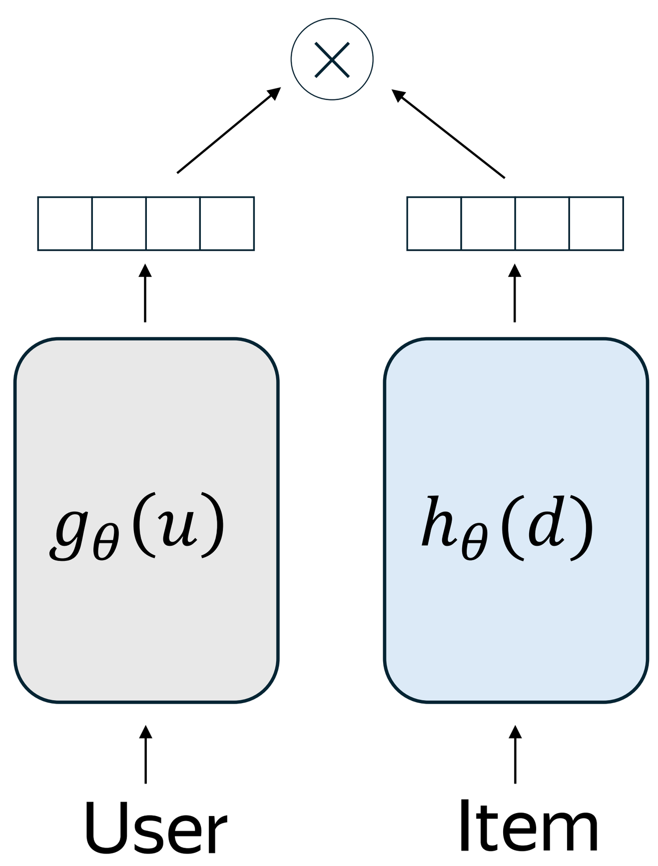
Kirill Khrylchenko, Vladimir Baikalov, Sergei Makeev, Artem Matveev, Sergei Liamaev

Embedding-Based Retrieval

Two-tower models encode users and items separately into embeddings:

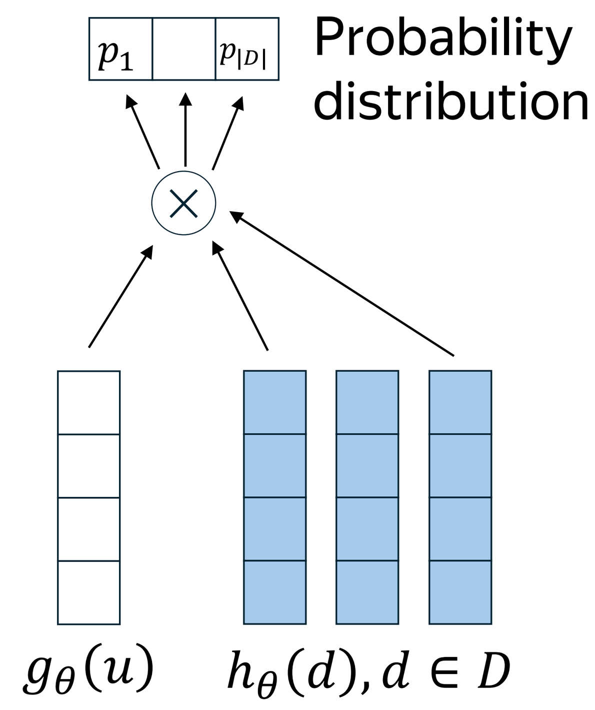
$$f_\theta(u,d) = \langle g_\theta(u), h_\theta(d) \rangle$$

Enables fast ANN search over large catalogs.



Why softmax loss:

- Enables global comparison across catalog
- Avoids folding effects
- Empirically stronger than pairwise or BCE alternatives



$$\mathcal{L}_{\text{softmax}}(u,p) = -\log P_\theta(p \mid u) = -\log \frac{e^{f_\theta(u,p)}}{\sum_{d \in \mathcal{D}} e^{f_\theta(u,d)}}$$

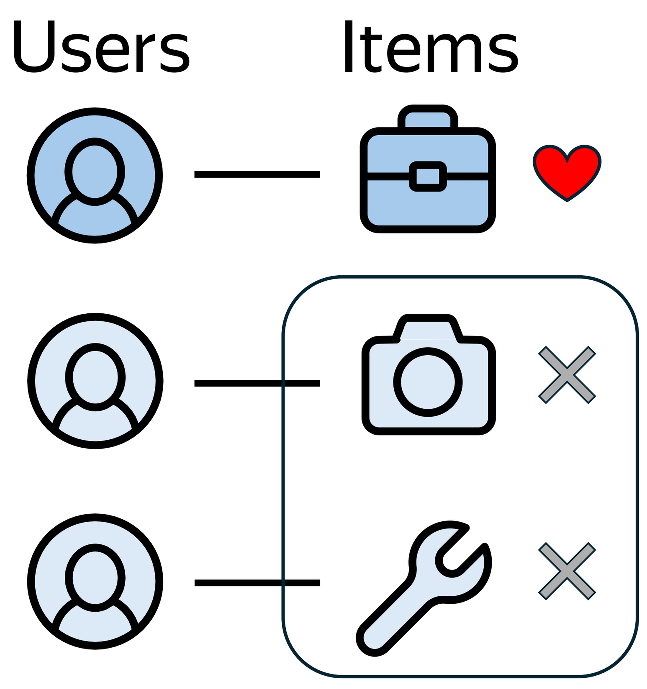
Large-Scale Retrieval

To enable large-scale retrieval, **sampled softmax** approximates denominator via negative sampling:

$$\mathcal{L}_{\text{sampled}}(u,p) = -\log \frac{e^{f_\theta(u,p)}}{e^{f_\theta(u,p)} + \sum_{d \sim Q} e^{f_\theta(u,d)}}$$

Negative sources:

- **Uniform sampling over catalog** requires many samples due to triviality
- **In-batch:** other positives as negatives; ≈ unigram distribution



LogQ Correction

In-batch negatives introduce systemic bias:

- Popular items appear more as negatives
- Using biased estimate of full softmax gradient

LogQ correction addresses this bias by substracting logarithm of item popularity from model logits:

$$f'_\theta(u,d) = f_\theta(u,d) - \log Q(d)$$

- Intuition: manually substracting item popularity so that the model doesn't have to
- Derived via importance sampling applied to full softmax gradient

Correcting the Correction

Shortcoming of standard logQ correction: derivation assumes positives are sampled from Q, but they are present deterministically.

To account for this:

- Decompose full softmax gradient into positive and negative terms
- Apply importance sampling **only** to negatives

$$\mathcal{L}_{\text{ours}}(u,p) = -w_{up} \log \frac{e^{f_\theta(u,p)}}{\sum_{d \sim Q'} e^{f_\theta(u,d) - \log Q'(d)}}$$

- Weight $w_{up} = sg\,(1 - P_\theta(p \mid u))$ reflects model confidence: small when p already scores high
- Positive is **excluded** from softmax denominator
- $d \sim Q'$ excludes p from sampling

Evaluation

- Transformer models (SASRec / ARGUS)
- Sampled softmax loss with mixed negative sampling

Academic datasets, Recall@20

LogQ Correction	Leave-One-Out		Temporal Split	
	ML-1M	*Steam*	*ML-1M*	*Steam*
Without	0.3853	0.1470	0.2514	0.1389
Standard	0.3893	0.1694	**0.2800**	0.1485
Improved	**0.3937**	**0.1730**	0.2792	**0.1609**

Industrial large-scale dataset

LogQ Correction	R@10	R@100	R@1000
Without	0.0280	0.0990	0.2992
Standard	**0.0304**	0.1211	0.4036
Improved	0.0279	**0.1222**	**0.4345**

Code & Data