Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches

Kuznetsov Dmitriy
AMI171

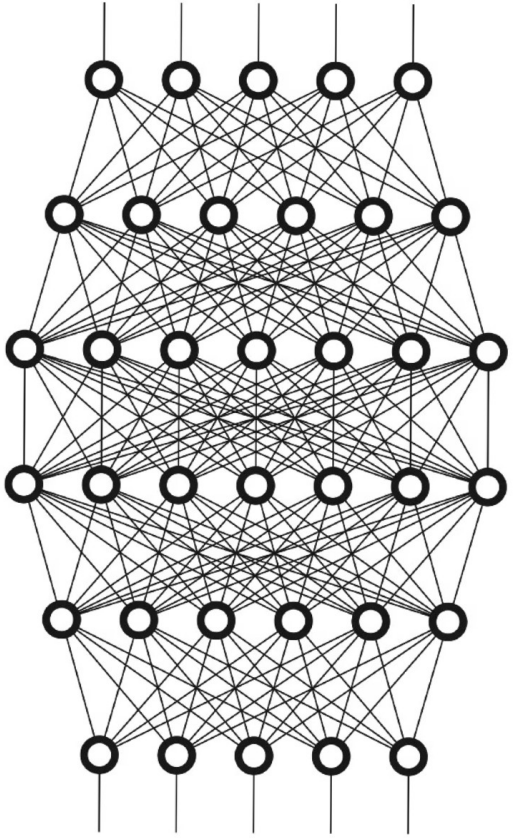
Recomender systems

$u \in U \qquad i \in I$

$R = \{r_{ui}\},\ r_{ui} \in \mathbb{R}$

$\forall u \in U \forall i \in I\ s.t.\ \overline{\exists} r_{ui} : \text{find}\, r_{ui}$

SOTA Recommender Systems



* DL techniques have started to dominate in RS

* Some researchers noticed, that well-tuned heuristic outperform new NN-methods

Shed light on SOTA DL RS

Noticed problems

* weak baselines

* establishment of weak methods as new baselines

* difficulties in comparing or reproducing results

Target goals

* Real ***reproducibility***

* Real ***progress***

Methodology

* Scan DL top-n recommendation papers from scientific conferences
(2015-2018; KDD, SIGIR, WWW, RecSys)

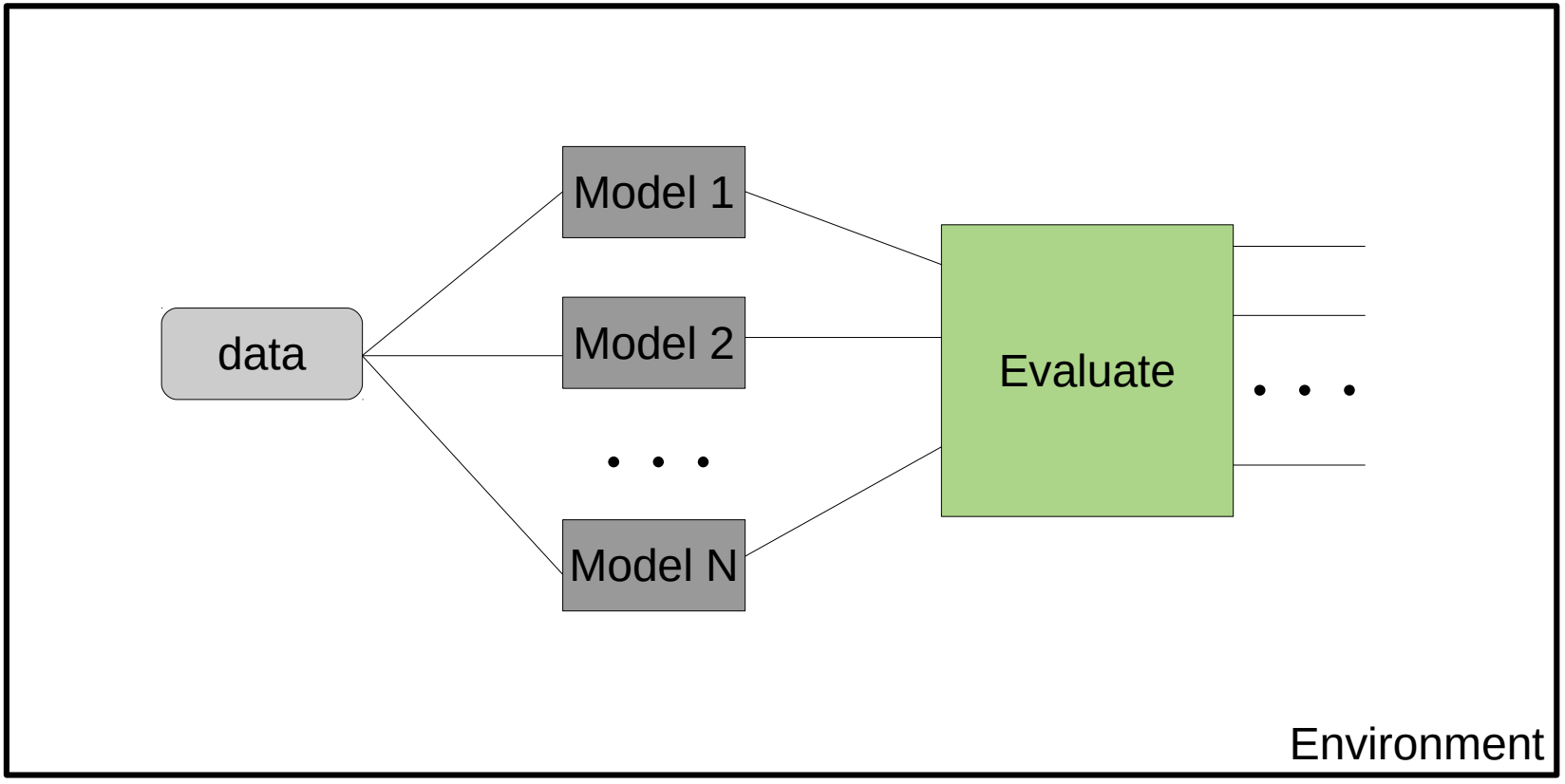
* Split papers into Reproducible and Non-Reproducible

* Reproducible:
  - working sources is available
  - At least one dataset is available

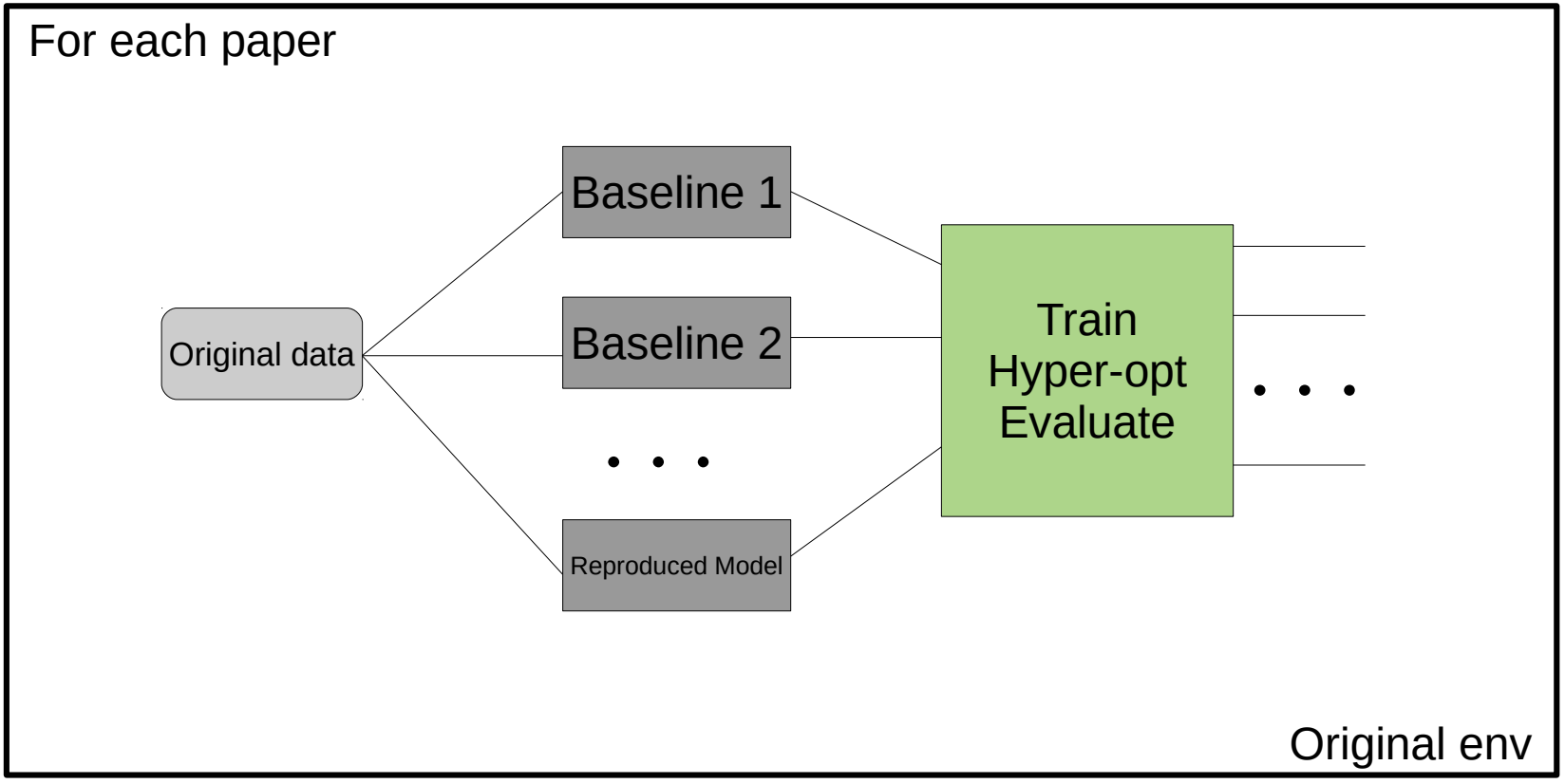
Conference	Rep. ratio	Reproducible
KDD	3/4 (75%)	[17], [23], [48]
RecSys	1/7 (14%)	[53]
SIGIR	1/3 (30%)	[10]
WWW	2/4 (50%)	[14], [24]
Total	7/18 (39%)	

Non-reproducible: KDD: [43], RecSys: [41], [6], [38], [44], [21], [45], SIGIR: [32], [7], WWW: [42], [11]

Evaluation Methodology. Naive



Evaluation Methodology



Baselines

Top-n popular

ItemKNN $r_i, r_j \in \mathbb{R}^{|U|}$

$$s_{ij} = \frac{r_i^T r_j}{\|r_i\|\|r_j\| + h}$$

UserKNN $r_u, r_v \in \mathbb{R}^{|I|}$

$$s_{uv} = \frac{r_u^T r_v}{\|r_u\|\|r_v\| + h}$$

ItemKNN-CBF $f_i, f_j \in \mathbb{R}^{|F|}$

$$s_{ij} = \frac{f_i^T f_j}{\|f_i\|\|f_j\| + h}$$

ItemKNN-CFCBF

$$\hat{r}_i = [r_i, \omega f_i]$$

P³α

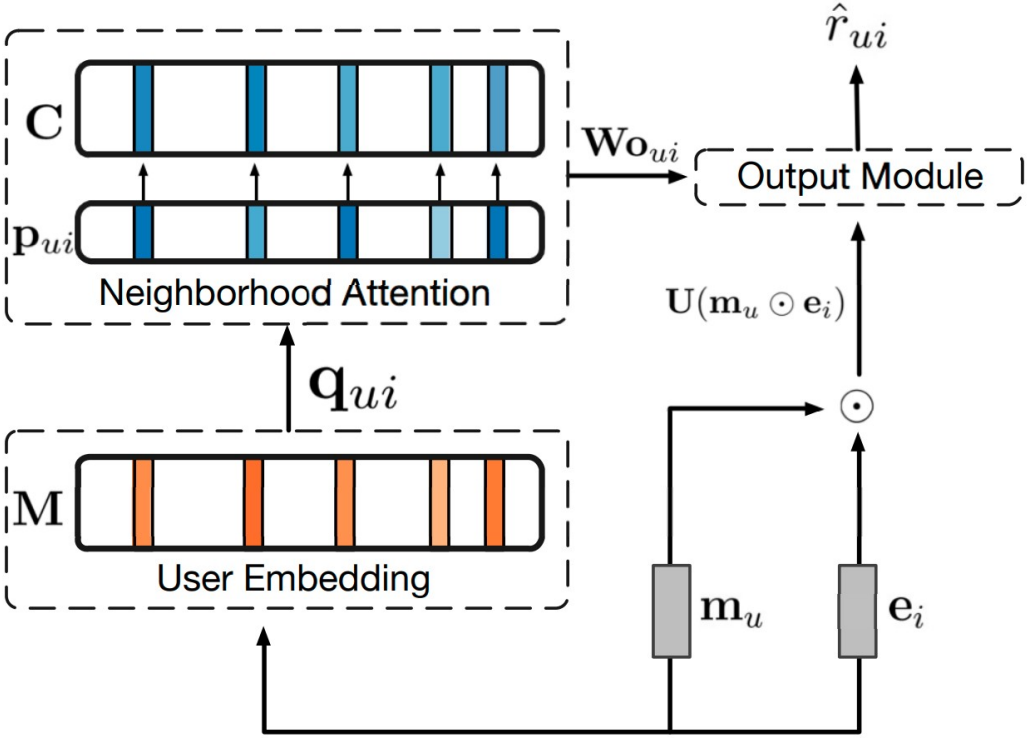
$$p_{ui} = (r_{ui}/N_u)^\alpha$$

$$p_{iu} = (r_{ui}/N_i)^\alpha$$

RP³β

$$\hat{s}_{ij} = \frac{s_{ij}}{sum(r_i \neq 0)^\beta sum(r_j \neq 0)^\beta}$$

Collaboritive Memory Networks

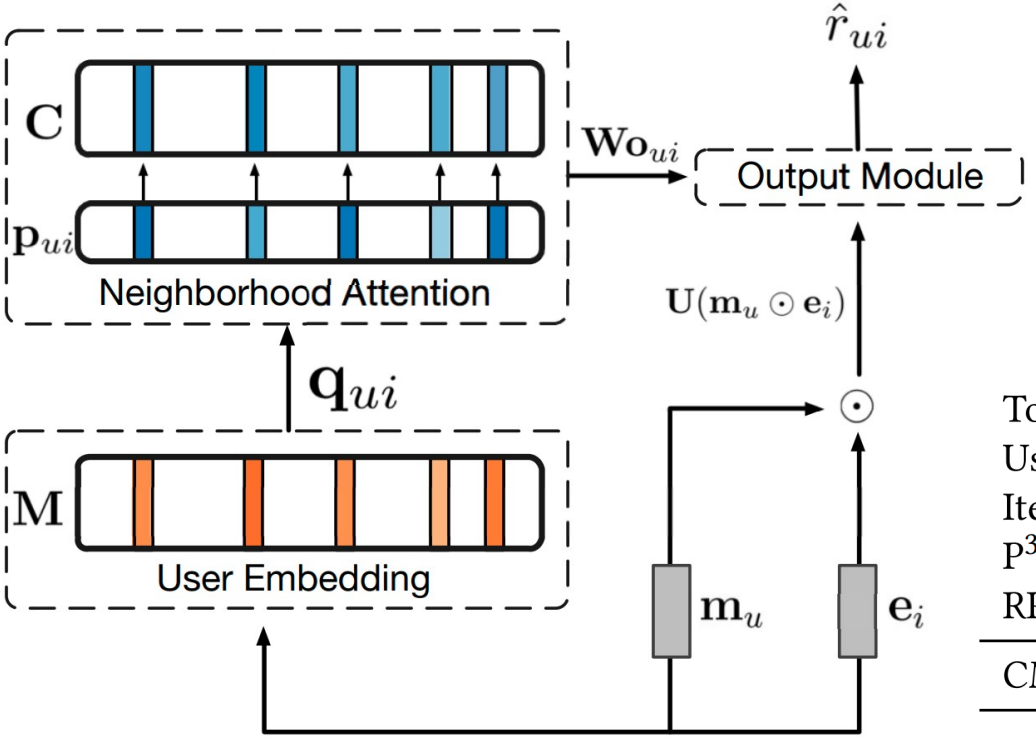


Original baselines:
* ItemKNN
* Matrix Factorization (s.t. SVD++)
* Neural RS (s.t. NeuMF)

* Hyperparams proposed
(LOO by hit-rate and nDCG)

* *Experiments reproducable*

Collaboritive Memory Networks

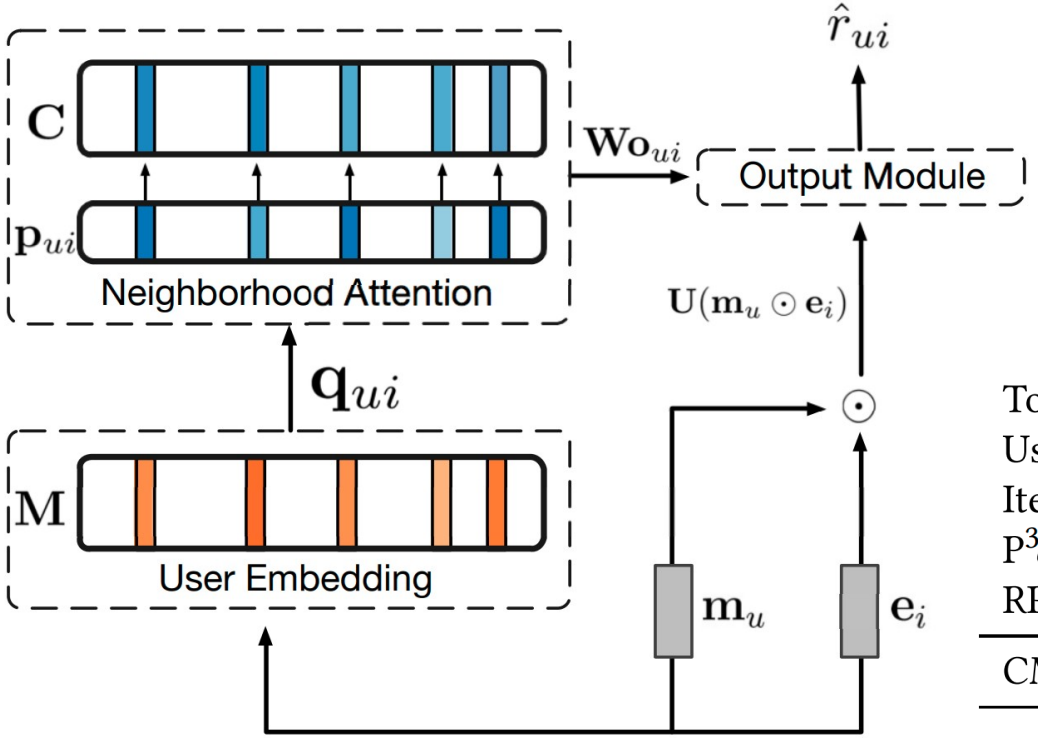


Original baselines:
* ItemKNN
* Matrix Factorization (s.t. SVD++)
* Neural RS (s.t. NeuMF)

	Pinterest			
	HR@5	NDCG@5	HR@10	NDCG@10
TopPopular	0.1668	0.1066	0.2745	0.1411
UserKNN	**0.6886**	**0.4936**	0.8527	**0.5470**
ItemKNN	**0.6966**	**0.4994**	**0.8647**	**0.5542**
$P^3\alpha$	0.6871	**0.4935**	0.8449	**0.5450**
$RP^3\beta$	**0.7018**	**0.5041**	**0.8644**	**0.5571**
CMN	0.6872	0.4883	0.8549	0.5430

$\rightarrow HR(5)$

Collaboritive Memory Networks



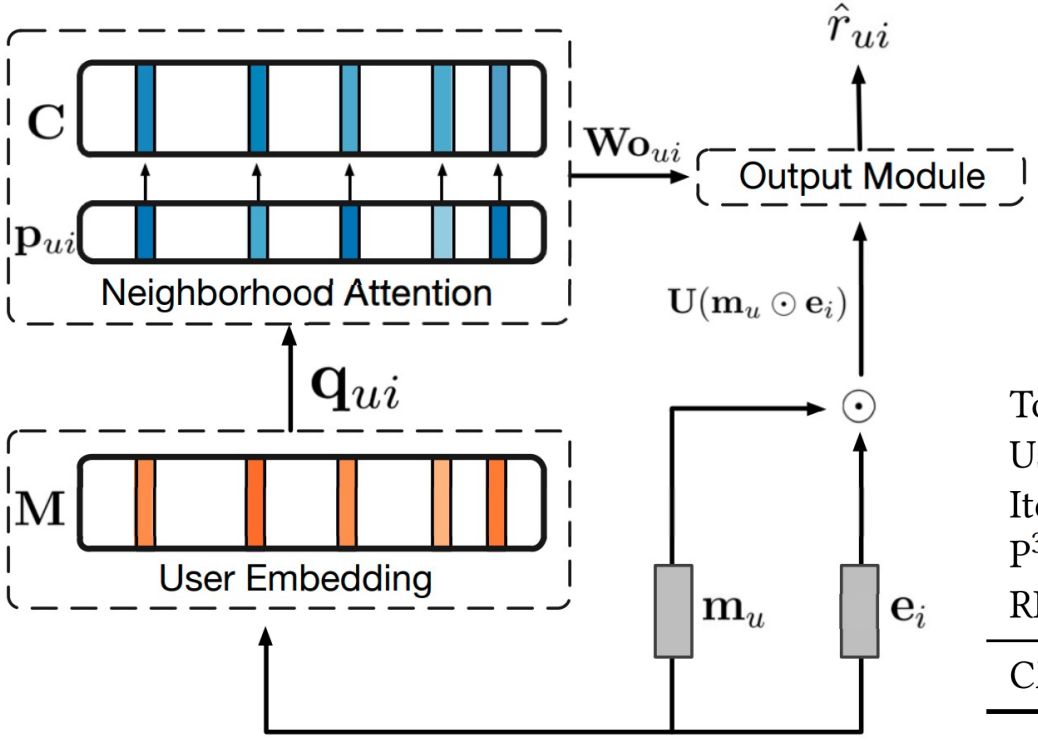
Original baselines:

* ItemKNN
* Matrix Factorization (s.t. SVD++)
* Neural RS (s.t. NeuMF)

	CiteULike-a			
	HR@5	NDCG@5	HR@10	NDCG@10
TopPopular	0.1803	0.1220	0.2783	0.1535
UserKNN	**0.8213**	**0.7033**	**0.8935**	**0.7268**
ItemKNN	**0.8116**	**0.6939**	0.8878	**0.7187**
$P^3\alpha$	**0.8202**	**0.7061**	0.8901	**0.7289**
$RP^3\beta$	**0.8226**	**0.7114**	**0.8941**	**0.7347**
CMN	0.8069	0.6666	0.8910	0.6942

$\rightarrow HR(5)$

Collaboritive Memory Networks



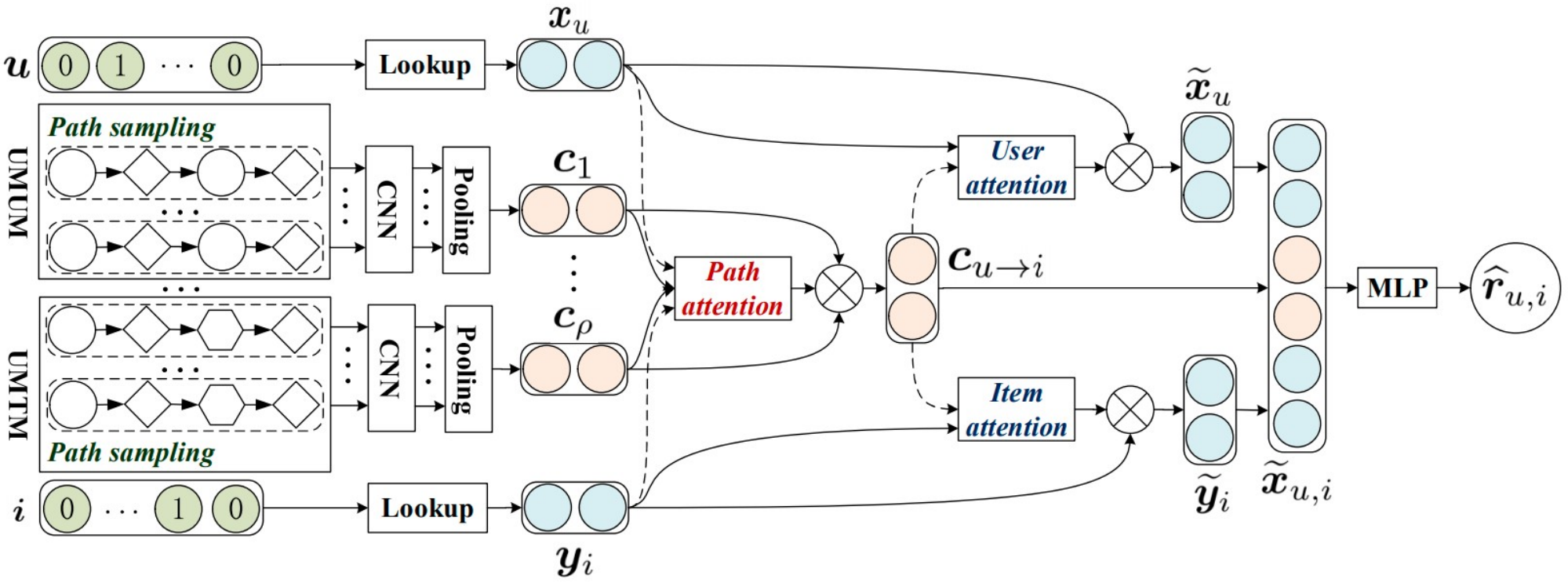
Original baselines:

* ItemKNN
* Matrix Factorization (s.t. SVD++)
* Neural RS (s.t. NeuMF)

	Epinions			
	HR@5	NDCG@5	HR@10	NDCG@10
TopPopular	**0.5429**	**0.4153**	**0.6644**	**0.4547**
UserKNN	0.3506	0.2983	0.3922	0.3117
ItemKNN	0.3821	0.3165	0.4372	0.3343
$P^3\alpha$	0.3510	0.2989	0.3891	0.3112
$RP^3\beta$	0.3511	0.2980	0.3892	0.3103
CMN	0.4195	0.3346	0.4953	0.3592

$\rightarrow HR(5)$

Metapath based Context for RECommendation



Metapath based Context for RECommendation

Movie-Lens100k

	PREC@10	REC@10	NDCG@10
TopPopular	0.1907	0.1180	0.1361
UserKNN	0.2913	0.1802	0.2055
ItemKNN	**0.3327**	**0.2199**	**0.2603**
$P^3\alpha$	0.2137	0.1585	0.1838
$RP^3\beta$	0.2357	0.1684	0.1923
MCRec	0.3077	0.2061	0.2363

$\rightarrow precision(10)$

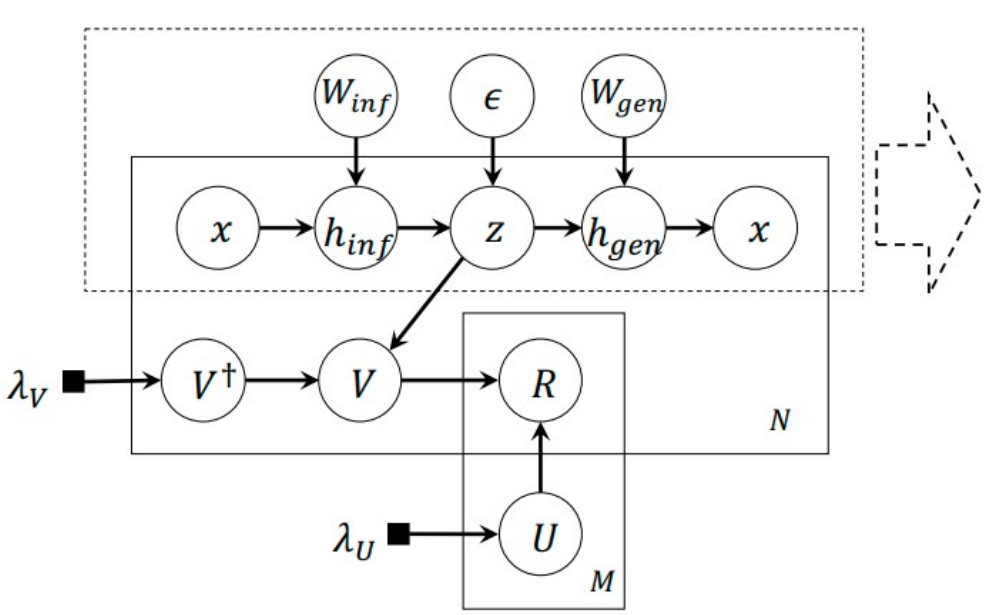
* train/test split provided only on MovieLens

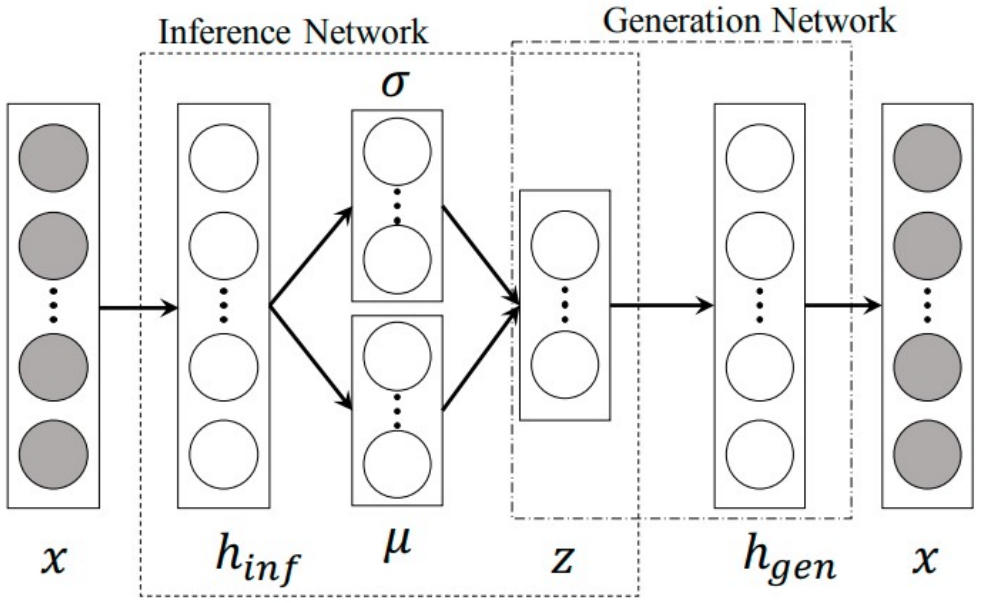
* baselines same as in previous one

* Moreover, NeuMF optimized inappropriate
(hypers from original paper)

* hyper-opt on test set for their method
(source is source code)

Collaborative Variational Autoencoder





Collaborative Variational Autoencoder

	REC@50	REC@100	REC@300
TopPopular	0.0044	0.0081	0.0258
UserKNN	0.0683	0.1016	0.1685
ItemKNN	**0.0788**	0.1153	0.1823
$P^3\alpha$	**0.0788**	0.1151	0.1784
$RP^3\beta$	**0.0811**	0.1184	0.1799
ItemKNN-CFCBF	**0.1837**	**0.2777**	**0.4486**
CVAE	0.0772	0.1548	0.3602

$\rightarrow recall(5)$

* CiteULike (135k and 205k); (sparse and dense)

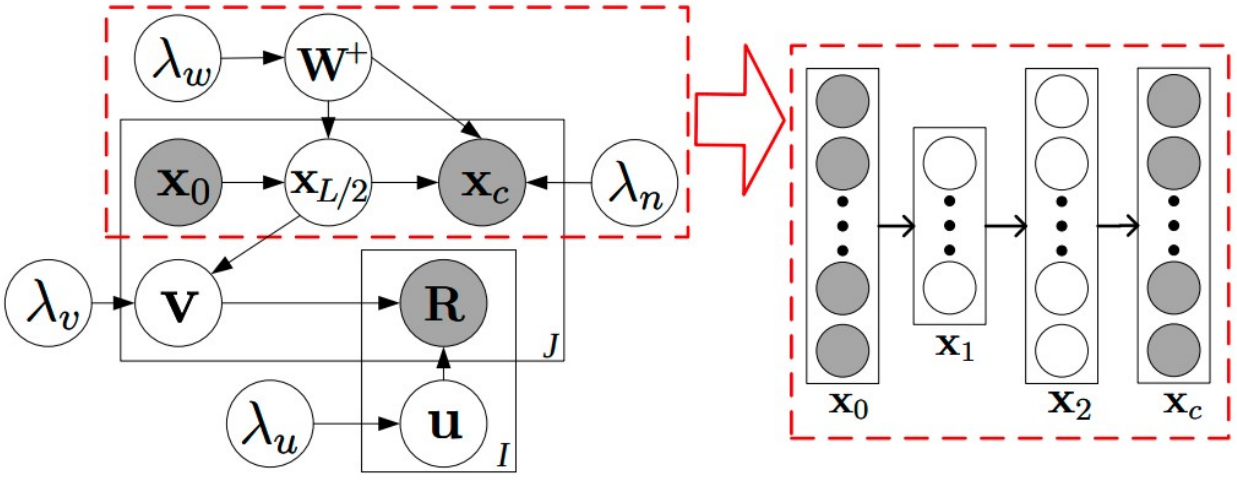
* 5 times train/test split

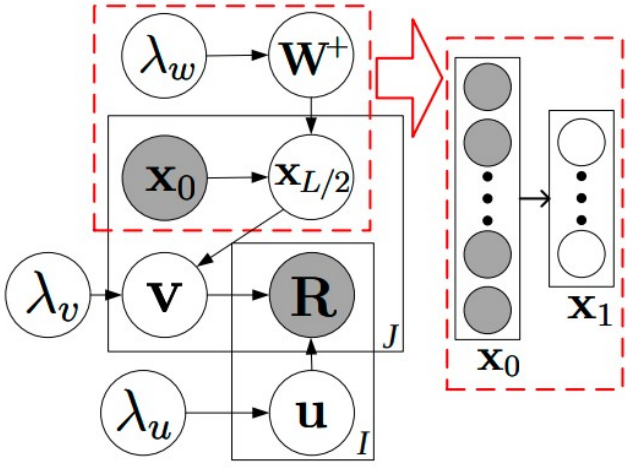
* Opt by val-recall (50 to 300)

* Baselines: 3 SOTA DL models (in particular CDL)

* CVAE outperform on most sets (100+ lists)

Collaborative Deep Learning





Collaborative Deep Learning

CiteULike-a dense

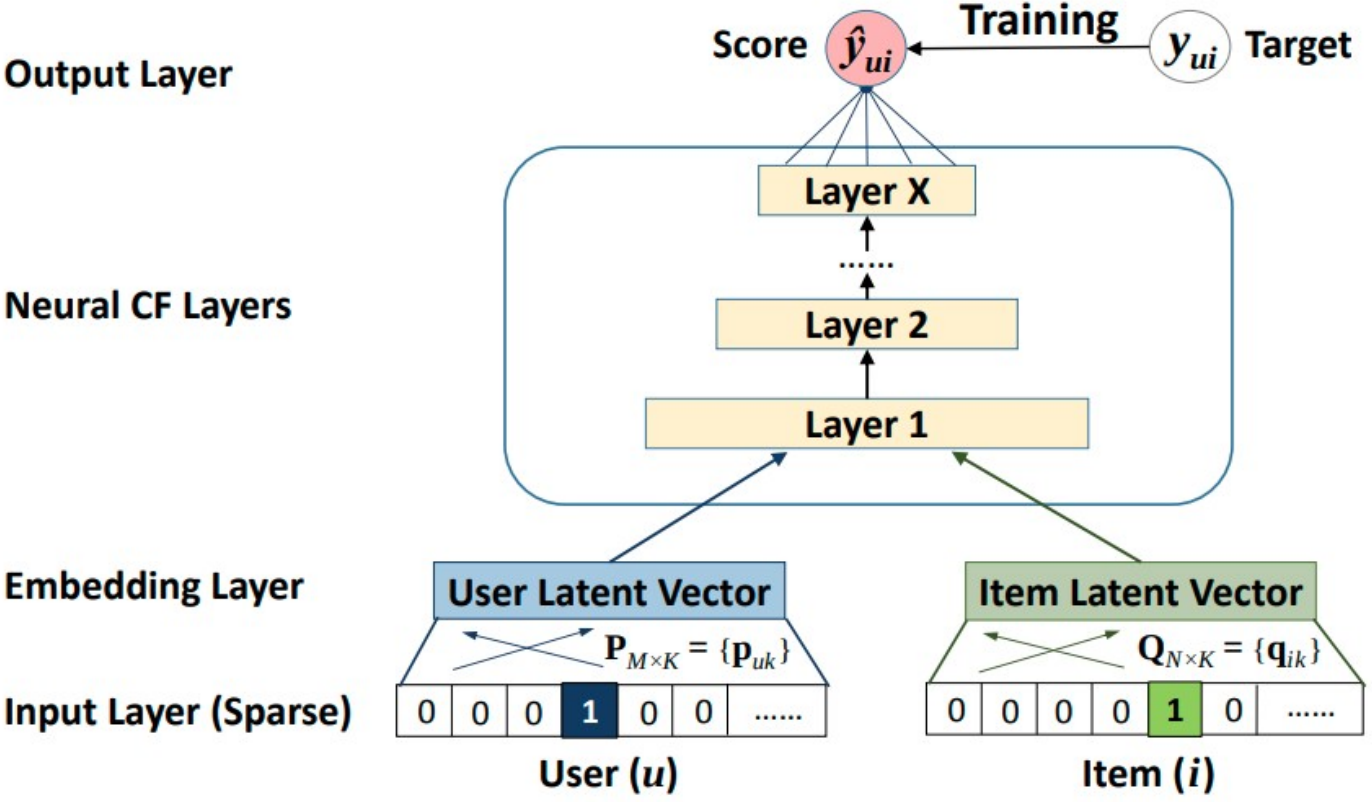
	REC@50	REC@100	REC@300
TopPopular	0.0038	0.0073	0.0258
UserKNN	**0.0685**	0.1028	0.1710
ItemKNN	**0.0846**	**0.1213**	0.1861
$P^3\alpha$	**0.0718**	**0.1079**	0.1777
$RP^3\beta$	**0.0800**	**0.1167**	0.1815
ItemKNN-CBF	**0.2135**	**0.3038**	**0.4707**
ItemKNN-CFCBF	**0.1945**	**0.2896**	**0.4620**
CDL	0.0543	0.1035	0.2627

$\rightarrow recall(5)$

* Opt by val-recall (50 to 300)

* CDL outperform on 2/4 sets (100+ lists)

Neural Collaborative Filtering



Neural Collaborative Filtering

	Pinterest			
	HR@5	NDCG@5	HR@10	NDCG@10
TopPopular	0.1663	0.1065	0.2744	0.1412
UserKNN	0.7001	**0.5033**	0.8610	**0.5557**
ItemKNN	**0.7100**	**0.5092**	**0.8744**	**0.5629**
$P^3\alpha$	0.7008	**0.5018**	0.8667	**0.5559**
$RP^3\beta$	**0.7105**	**0.5116**	**0.8740**	**0.5650**
NeuMF	0.7024	0.4983	0.8719	0.5536

* LOO; splits shared;
* hyper-opt on val

* n_epochs by test val

Neural Collaborative Filtering

	Movielens 1M			
	HR@5	NDCG@5	HR@10	NDCG@10
TopPopular	0.3043	0.2062	0.4531	0.2542
UserKNN	0.4916	0.3328	0.6705	0.3908
ItemKNN	0.4829	0.3328	0.6596	0.3900
$P^3\alpha$	0.4811	0.3331	0.6464	0.3867
$RP^3\beta$	0.4922	0.3409	0.6715	0.3991
NeuMF	0.5486	0.3840	0.7120	0.4369
SLIM	**0.5589**	**0.3961**	**0.7161**	**0.4470**

$$\min_W \frac{1}{2}\|A - AW\|_F^2 + \frac{\beta}{2}\|W\|_f^2 + \lambda\|W\|_1$$

$$\begin{cases} W \geq 0 \\ diag(W) = 0 \end{cases}$$

Spectral Collaborative Filtering

HetRec, Amazon Instant Video

* Train/test split is not provided

* Provided only half of hyperparams
(unknown half optimized on val by authors)

* Evaluation procedure is not provided

* All authors baselines outperform SpectralCF on all measures

Spectral Collaborative Filtering

MovieLens (1M)

random split

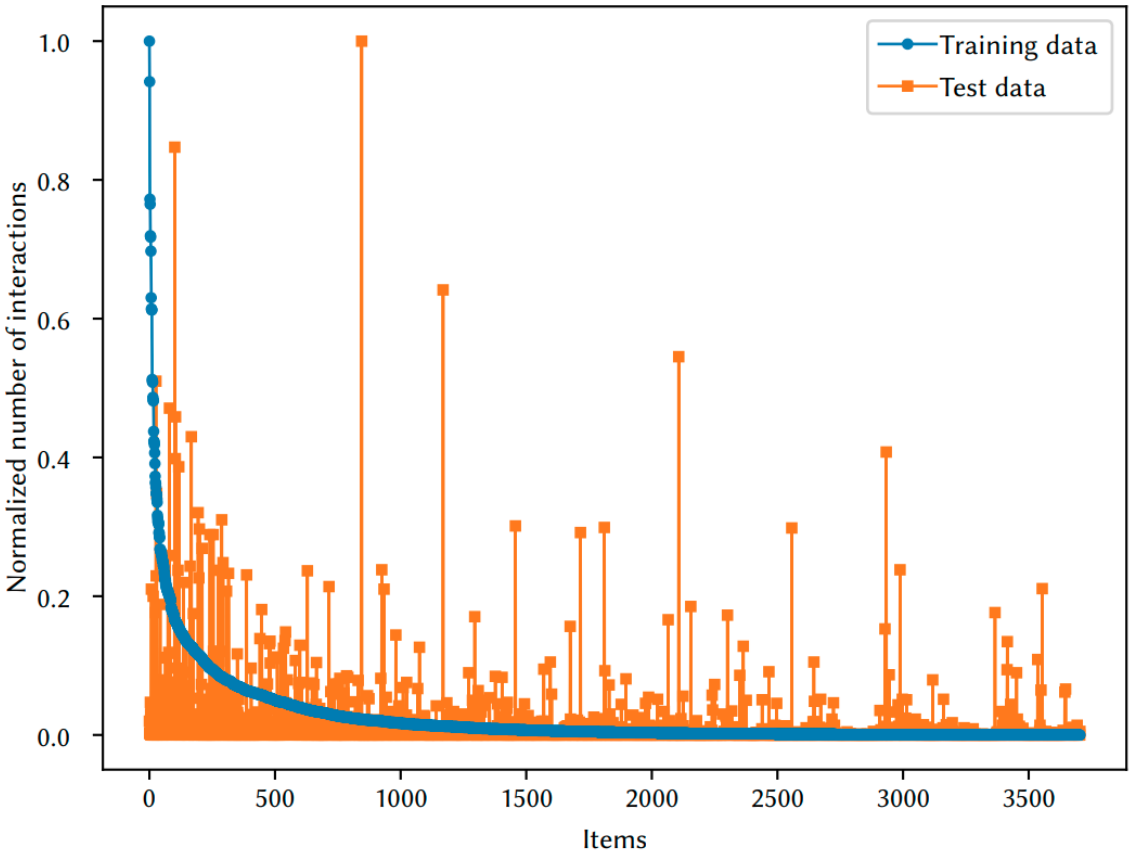
	Cutoff 20		Cutoff 60		Cutoff 100	
	REC	MAP	REC	MAP	REC	MAP
TopPopular	**0.1853**	**0.0576**	**0.3335**	**0.0659**	**0.4244**	**0.0696**
UserKNN CF	**0.2881**	**0.1106**	**0.4780**	**0.1238**	**0.5790**	**0.1290**
ItemKNN CF	**0.2819**	**0.1059**	**0.4712**	**0.1190**	**0.5737**	**0.1243**
$P^3\alpha$	**0.2853**	**0.1051**	**0.4808**	**0.1195**	**0.5760**	**0.1248**
$RP^3\beta$	**0.2910**	**0.1088**	**0.4882**	**0.1233**	**0.5884**	**0.1288**
SpectralCF	0.1843	0.0539	0.3274	0.0618	0.4254	0.0656

* Train/test split provided
* SpectralCF outperform all baselines with a huge margin

Recall@20 50% higher than authors best baseline

* *But...*

Spectral Collaborative Filtering



* Gini index and Shannon entropy diverge largely from a random splits'

* 0.92 gini vs 0.79 gini on random split

Summary and link

* **Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches**
Maurizio Ferrari Dacrema, Paolo Cremonesi, Dietmar Jannach

https://arxiv.org/pdf/1907.06902.pdf

Вопросы

1. Какие исходные бейзлайны рассматривались в статье? Опишите устройство ItemKNN-CFCBF

2. Опишите процесс воспроизведения авторами результатов SOTA-моделей и пайплайн проведения экспериментов.

3. Сформулируйте причины неутешительных результатов SOTA-моделей. Раскройте одну из причин чуть подробнее.