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

M. Ferrari Dacrema, P. Cremonesi – Politecnico di Milano, Italy

D. Jannach – University of Klagenfurt, Austria

ACM Conference on Recommender Systems (RecSys ’19),

Motivation

Observation: In other fields of ML, DL SoA algorithms
not as strong as expected

Goal: Evaluation and reproducibility study of recent top-n DL algorithms for RS

How easy is to reproduce the published results?

How competitive DL is against heuristic baselines?

Methodology: Collecting Relevant Papers

- Conferences: RecSys, WWW, KDD, SIGIR
- Long paper, from 2015 to 2018
- DL applied to traditional Top-n
- Evaluation with accuracy metrics

Methodology: Collecting Reproducible Papers

- Source code available and runnable
- Public dataset (original split preferably)
- If not available, ask the authors and wait a month

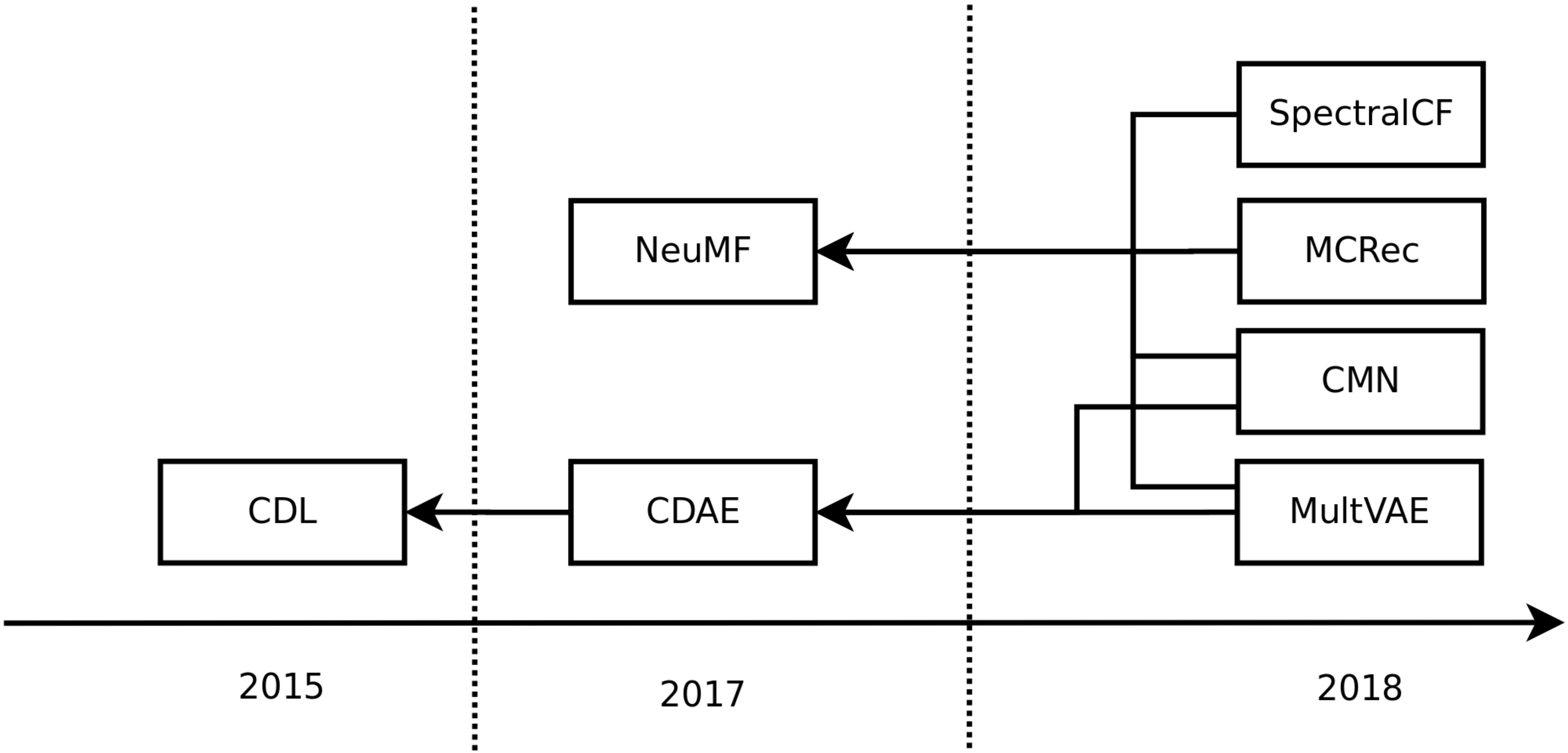
Reproducible papers statistics

Conference	Repr. Ratio
KDD	3/4 (75 %)
WWW	2/4 (50 %)
SIGIR	1/3 (30 %)
RecSys	1/7 (14 %)
Total	7/18 (39 %)

Reproducible papers list

SIGIR ‘18	CMN	Collaborative memory networks
KDD ’18	MCRec	Metapath based context for rec.
KDD ’17	CVAE	Collaborative variational autoencoder
KDD ’15	CDL	Collaborative deep learning
WWW ’17	NeuMF	Neural collaborative filtering
WWW ’18	Mult-VAE	Variational autoencoder for CF
RecSys’18	SpectralCF	Spectral collaborative filtering

Reproducible papers used as baseline in later ones



Most common issues for lack of reproducibility

Source Code

(1) Lost

(2) NDA

(3) Not working

Data

(1) Lost

(2) NDA

(8) No reply

Methodology: Experimental evaluation

Same experimental procedure as the original paper:
same data, train/test split, metrics, cutoffs

Hyperparameters

- DL: Use original hyperparameters
- Baselines: Bayesian search, 40 cases

Baselines

Non personalized:	Top Popular
Collaborative Filtering:	ItemKNN, UserKNN, P3alpha, RP3beta
Hybrid:	ItemKNN CF + CBF
Machine Learning:	SLIM

Result summary - DL algorithms outperforming baselines

Algorithm	CF + CBF
MCRec	-
SpectralCF	-
CMN	4/12 - 30%
NeuMF	6/12 - 50%
CDL	9/24 - 37%
CVAE	9/24 - 37%
Mult-VAE	12/12 - 100%

Result summary - DL algorithms outperforming baselines

Algorithm	CF + CBF	CF + CBF + NP
MCRec	-	-
SpectralCF	-	-
CMN	4/12 - 30%	-
NeuMF	6/12 - 50%	6/12 - 50%
CDL	9/24 - 37%	9/24 - 37%
CVAE	9/24 - 37%	9/24 - 37%
Mult-VAE	12/12 - 100%	12/12 - 100%

Result summary - DL algorithms outperforming baselines

Algorithm	CF + CBF	CF + CBF + NP	CF + CBF + NP + SLIM
MCRec	-	-	-
SpectralCF	-	-	-
CMN	4/12 - 30%	-	-
NeuMF	6/12 - 50%	6/12 - 50%	-
CDL	9/24 - 37%	9/24 - 37%	9/24 - 37%
CVAE	9/24 - 37%	9/24 - 37%	9/24 - 37%
Mult-VAE	12/12 - 100%	12/12 - 100%	10/12 - 83%

Why this discrepancy in our results vs the original ones?

- Weak baselines
- Poor tuning of baseline hyperparameters

Methodological issues:

- Number of epochs selected with test data
- All sorts of experimental procedures

How can we move forward?

Add simple baselines

Improve reproducibility:

- Virtualization technology
- Include preprocessing and tuning code

Improve motivation of experimental design

Disclaimer

We tried to be as fair as possible, if you believe there are methodological issues in our work, please contact us.

Extended version on the way

- More conferences
- 26 relevant articles
- 12 reproducible

- 41.010 total experiments
- 253 days of Amazon AWS

Follow our lab on ResearchGate !

Q & A

Follow our lab on ResearchGate !