UNIVERSITÄT KLAGENFURT

Why Are Deep Learning Models Not Consistently Winning Recommender Systems Competitions Yet?

Dietmar Jannach, University of Klagenfurt, Austria
Gabriel De Souza P. Moreira, NVIDIA
Even Oldrige, NVIDIA

Presented at the ACM RecSys Challenge Workshop 2020, Online
dietmar.jannach@aau.at

www.aau.at

The Big Success of Deep Learning

- Deep Learning is used everywhere in the Recommender Systems literature
  - Rating prediction, personalized ranking, predicting next items or baskets, learning representations of users, items and side information, combining different types of information in hybrids, building conversational systems, creating attacks, …

- Disclaimer: Also by us

The Big Success of Deep Learning

- Deep Learning methods are almost always better than everything we had before
  - virtually no one uses Matrix Factorization, Learning-to-Rank techniques, Decision Trees and Random Forests, Nearest Neighbors, Logistic Regression, SVMs, GBMs etc., except as baselines to beat (exaggerating)

- Deep Learning methods even outperform other Deep Learning methods!

The Big Success of Deep Learning

- Deep Learning is successfully deployed in industry
  - There are many published reports on the success of Deep Learning for Recommender Systems
  - Alibaba, Baidu, Google, Pinterest, Facebook, YouTube and many, many more.

But Who Wins in Competitions?

- Winning solutions RecSys Challenges 2017-2019:
  - Substantial feature engineering
    - i.e., built on domain knowledge
  - Gradient Boosting
    - i.e., building ensembles of weak learners (late 1990s)
- Winning solution RecSys Challenge 2020 (NVIDIA):
  - Substantial feature engineering
  - Gradient Boosting

Why this Discrepancy?

- Are the problems different?
- Are the researchers different?
- Is the evaluation methodology different?

Are the Problems Different?

- Academic setting vs. Challenges (vs. Real-World)
- Datasets sizes
  - Challenge datasets can be relatively large, e.g., from XING (320M interactions).
  - Not uncommon in academia, though (Netflix, 100M)
- DL:
  - DL in academia: starting with ML100k and smaller
  - DL should in principle profit from large datasets, but sometimes difficult to hold giant data tables

Are the Problems Different?

- Dataset Characteristics
  - Challenge datasets are often large, but collected during a small time window (e.g., a few days).
  - Real-world datasets can contain rich interaction histories over extended periods of time
- DL Success in Practice:
  - Advantages of DL maybe only unfold once more data (per user) is available?

Are the Problems Different?

- Prediction Problem and Targets
  - Academia:
    - Traditional assumption of long-term user profiles, several interactions per user; predict relevance or ranking
  - Challenges:
    - Different problem settings
      - More cold-start and session-based recommendations
      - Difficulties of DL in session-based recommendation
    - Side information often available
    - Sometimes different metrics, sometimes created for the challenge

Ludewig, M., Latifi, S., Mauro, N. and Jannach, D.: *"Empirical Analysis of Session-Based Recommendation Algorithms"*. User Modeling and User-Adapted Interaction, Vol. forthcoming. 2020

Why this Discrepancy?

- Are the problems different?
- Are the researchers different?
- Is the evaluation methodology different?

Are the Researchers Different?

- Some Hypotheses:
  1. Maybe they are not experts in DL
  2. Maybe they do not have GPU-powered hardware
  3. Maybe they just prefer GBMs because they worked in the past
- H1 and H2 are difficult to maintain
  - when looking at profiles of participants

Are the Researchers Different?

- However
  - Contestants might have different objectives, preferences, and work processes
  - GBMs have some advantages in this process
    - stable to overfitting, lightweight preprocessing, automatic feature selection
  - Neural networks sometimes need more care, tuning effort, and computing time
    - Which is a disadvantage when the challenge duration is short

Why this Discrepancy?

- Are the problems different?
- Are the researchers different?
- Is the evaluation methodology different?

Is the Evaluation Process Different?

Aspect	Academic Research	Machine Learning Competition
Selection of dataset	By researcher	By competition host
Selection and optimization of baselines	By researcher	Represented by other participants
Selection of evaluation metric(s)	By researcher	By competition host
Selection of protocol specifics[10]	By researcher	By competition host
Execution of measurement	By researcher	By competition host
Publication of measurement results	By researcher	By competition host
Test data	Available to researcher	Never available to participant
Source code sharing	Researcher may share	Must be shared sometimes
Dataset sharing	Researcher may share	Training data available to all participants

Is the Evaluation Process Different?

- Academic researchers have much more freedom
  - Which, of course, is important
- They are usually the only ones who evaluate their own method before publication
  - Researcher biases might emerge
  - Leading to choice of weak baselines, the use of non-optimized baselines, or the choice of specific experimental configurations
    - Some “wins” of DL methods in the academic literature showed to be actually non-existent

Ferrari Dacrema, M., Cremonesi, P. and Jannach, D.: *"Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches"*. In: Proceedings of the 2019 ACM Conference on Recommender Systems (RecSys 2019). Copenhagen, 2019

Discussion & Conclusion

- We highlighted a number of potential reasons for the observed discrepancy
  - No clear answer yet
- We believe that the potential of DL methods for recommendations is not fully exploited yet
  - Better combination of different information sources
  - Better tools for DL to work “out-of-the-box” needed
- We should not forget about non-DL methods in academic research

Thanks for the Attention (which is all we need)

- Questions?
- dietmar.jannach@aau.at