Further investigating transformer models in recommender systems

Jakub Frąc [1] Dominykas Seputis [2] Marcell Schuh [1]

[1]Vrije Universiteit Amsterdam [2]University of Amsterdam

Motivations

- Transformers4Rec[1] introduced by NVIDIA to bridge NLP transformers models to RecSys.
- While the original paper showed promising results it focused on smaller datasets and simple features like item IDs.
- We want to explore the application of T4Rec on a larger dataset, with more complex features, and validate the assumptions surrounding dataset sizes of transformers models from NLP in the context of RecSys.

Key Idea

- Item recommendations can be represented as a next-in-sequence prediction problem
- Incorporating more expressive features such as as article textual embeddings might improve model performance.
- Using the larger dataset and two smaller subsamplings we are looking the validate the assumptions from NLP on the amount of data required by Transfomer models.

Overview of Transformers4Rec

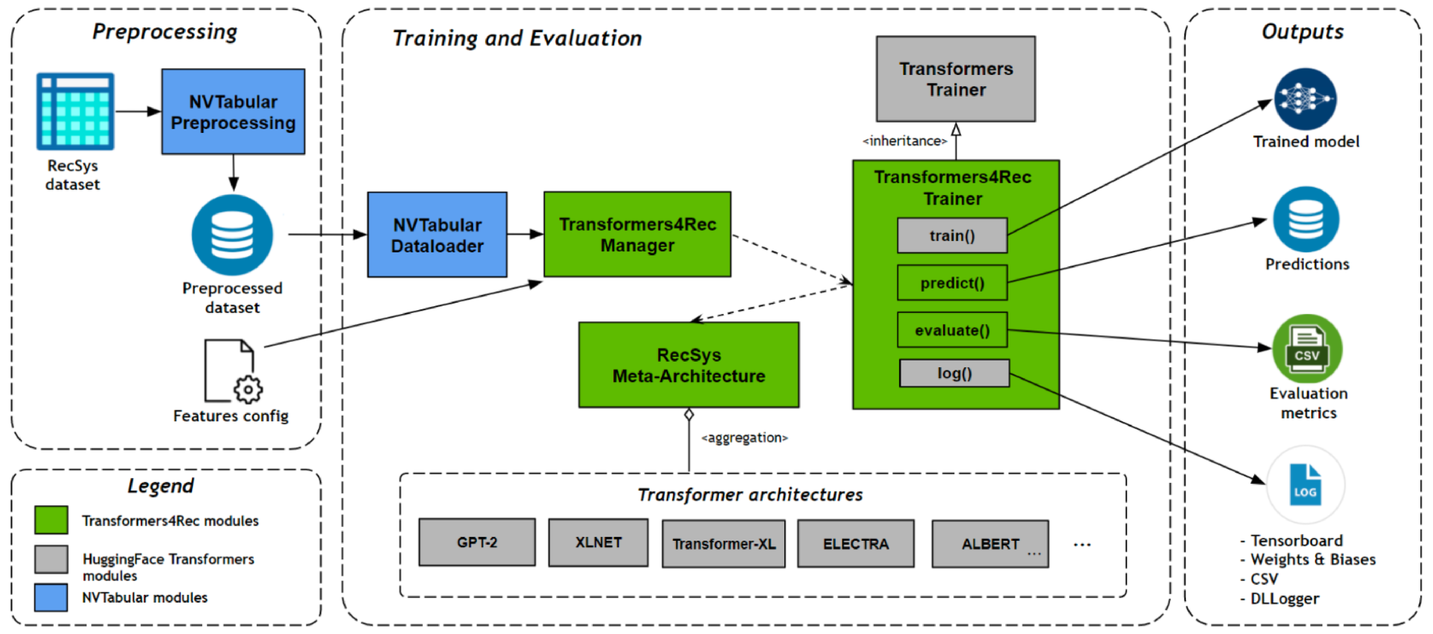


Figure 1. Transformers4Rec Pipeline Overview[1]

Approach

1. Started by preprocessing the Ekstra Bladet News Recommendation Dataset (EB-NeRD), by creating artificial sessions from combining impressions and user history.
2. Divided the resulting sessions into smaller sessions.
3. The final dataset spans 35 days of impressions, we use the first 34 days for training and the final day for reporting performance.
4. Extend Transformers4Rec to support additional article features, and the articles textual embedding along side article ids.

Experiments

We designed several experiments to verify the importance of additional signals.

- Trained models with various combination of features:
  - Article IDs, only
  - Article IDs, and article features
  - Article IDs, and article text embeddings
  - Article IDs, article features, and text embeddings
- Trained models on different subsamplings (10%, 50%, and 100%) of the dataset.
- Tested to model two bases: XLNet, and GPT2. In order compare masked language modeling (MLM) in the encoder setting against next-token prediction in the decoder setting.
- We use Normalized Discounted Cumulative Gain (NDCG) and recall at various top-k stages to assess the prediction of next-click.

Results

Feature Impact on Model Performance

We found that the addition of article metadata only improved metrics by 2%, in contrast the addition of pre-trained textual embeddings resulted in nearly doubled metrics. All this with only a 4.6% growth in model paramter size.

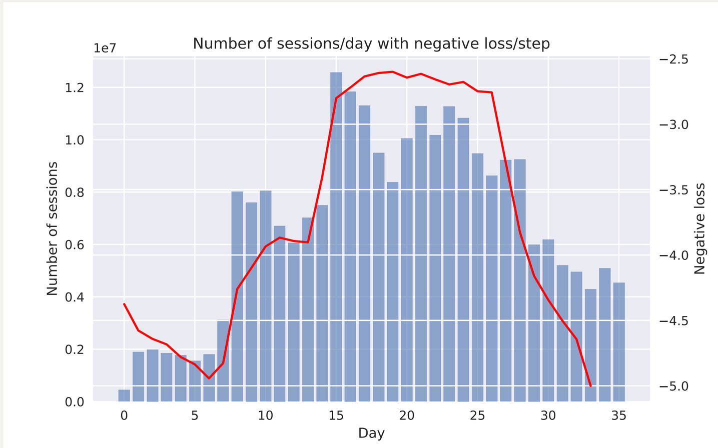
Features combination	Model size (parameters)	NDCG @ 10	Recall @ 10	NDCG @ 20	Recall @ 20
tokens	35,421,632	0.372	0.582	0.402	0.701
tokens+feats	35,427,803	0.379	0.587	0.409	0.706
tokens+emb	37,123,072	0.667	0.693	0.673	0.717
tokens+feats+emb	37,128,917	0.678	0.706	0.685	0.732

Figure 2. Comparison of model performance trained on different feature combinations and model sizes. The models were trained using XLNet with 100% of the training set.

Data Size Impact on Model Performance

To evaluate the importance of dataset size for transformer-based recommender systems, we compare two XLNet configurations trained on *tokens+feats* and *tokens* with 10%, 50%, and 100% of the dataset subsamples.

Train dataset subsample size influence to model performance. The models were trained using XLNet with *tokens* and *tokens+feats* feature combinations.



Influence of sample size to the training loss of XLNet model trained on 100% of data using *tokens* as features.

Results (cont.)

Model's base influence to model's performance

In our preliminary results (fig. 3) GPT-2 clearly outperforms XLNET despite having fewer parameters. It is unclear whether the encoder vs. decoder-only training strategies or the model architectures are influencing this performance discrepancy.

Model	Model size (parameters)	NDCG @ 10	Recall @ 10	NDCG @ 20	Recall @ 20
XLNet tokens+embs	37,123,072	0.169	0.225	0.181	0.272
XLNet tokens+feats+embs	37,128,917	0.205	0.265	0.217	0.312
GPT-2 tokens+embs	36,344,789	0.259	0.333	0.273	0.392
GPT-2 tokens+feats+embs	36,338,944	0.341	0.467	0.362	0.551

Figure 3. Comparison of model performance traind with different model heads (GPT-2 and XLNET).

Limitations

- It is still not entirely clear whether Transformer models can be used for continual training in production setting.
- Large data volume and the model complexity requires large amounts of compute to perform full training.

Future work

- Further investigate Transformer re-training capabilities for the production setting.
- Allow to finetune pretrained text embeddings models with the MLM/NTP downstream objective.

Summary

Our experimental results validated several key insights.

- We found clear indication that more expressive features, such as textual embeddings improves the accuracy.
- Consistent with observations from NLP larger datasets clearly improve transformer based recommendation systems.
- When comparing XLNET and GPT-2 we found that GPT-2 proved superior, despite having fewer parameters. Suggesting generative pre-training might be advantageous.
- While Transformers4Rec is promising we identified several issues while working with the framework around documentation, adaptability and computational requirements.

References

[1] Gabriel de Souza Pereira Moreira, Sara Rabhi, Jeong Min Lee, Ronay Ak, and Even Oldridge. Transformers4rec: Bridging the gap between nlp and sequential / session-based recommendation. In *Fifteenth ACM Conference on Recommender Systems*, RecSys '21. ACM, September 2021.