Towards Neural Mixture Recommender for Long Range Dependent User Sequences

Jiaxi Tang*, Francois Belletti*, Sagar Jain, Minmin Chen, Alex Beutel, Can Xu and Ed H. Chi

1. Abstract

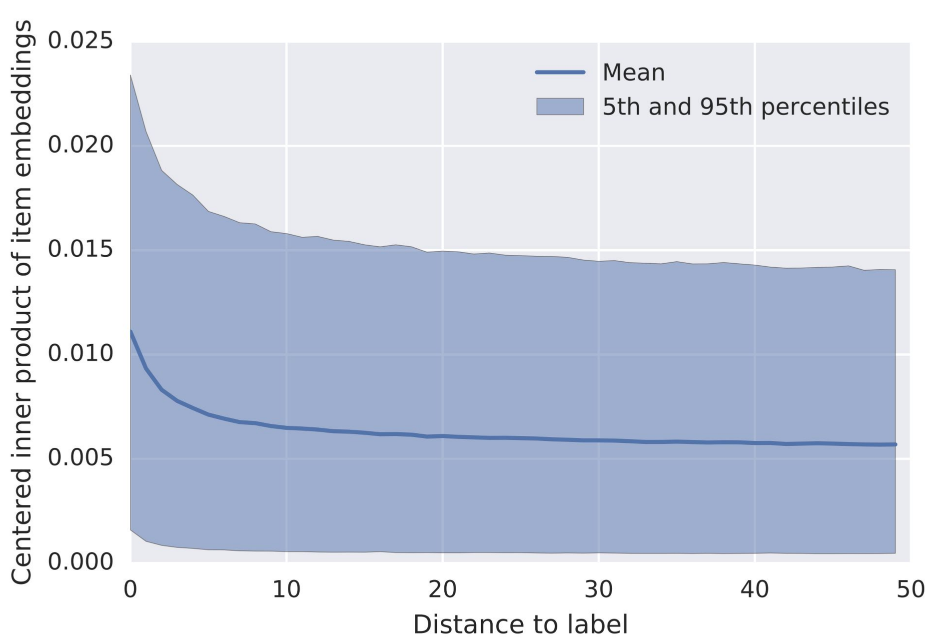
- We empirically analyze the **temporal dependencies** in YouTube data, and find statistically significant **Long Range Dependence (LRD)**.
- We propose a tailored solution to predict which item will be viewed that can **model** temporal dependencies **with different ranges** within the same neural model.
- **Experiments** on both public dataset (MovieLens 20M) and production dataset (YouTube) **demonstrate the effectiveness** of our proposed method.

2. Temporal Dependencies in YouTube

- LRD in sequential recommendations: users' history from long ago may still **influence** their current preference.
- **What are statistical indications that sequences in our data are LRD**?
- We examine the trace of the covariance matrix of embedding sequences as a measurement of dependency, i.e. the **decay of item similarity with time**.

$$\text{Dep}_L = \text{tr}\left(\text{Cov}\left(Q_{e_N}, Q_{e_{N-L}}\right)\right)$$

- Results from *YouTube* dataset: dependencies decay slowly (power-law rate) in user behavior showing LRD.



3. Limitations of Previous works

- *Limitations of some existing Sequential Models:*

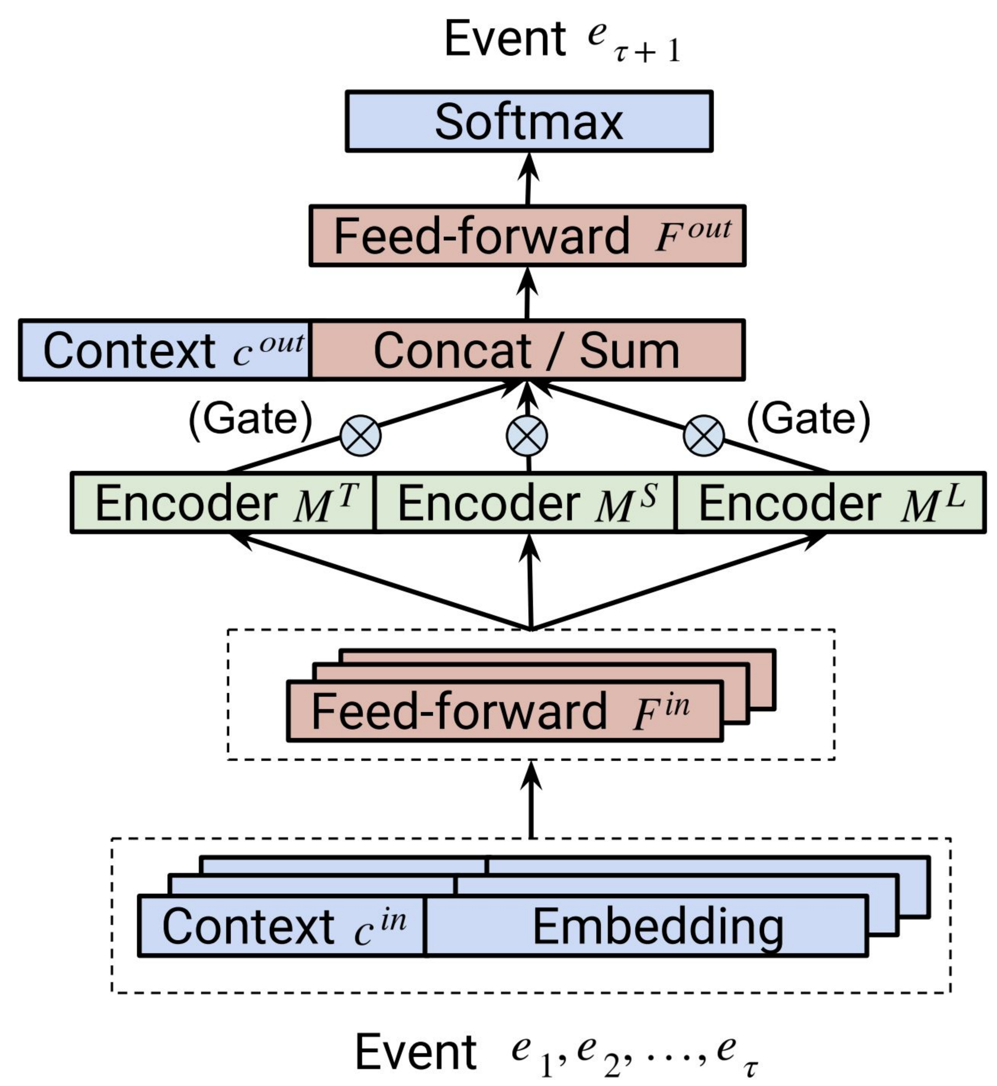
1. Temporal dependencies are limited to a short window (e.g., Caser, Fossil, *etc*).
2. RNNs tend to have difficulties leveraging the information contained far into the past due to gradient propagation issues (e.g., GRU4Rec)
3. Maintaining user latent factors for extended periods of time is challenging (privacy issues, storage issue, etc)

- *Limitations of Single Monolithic Models*

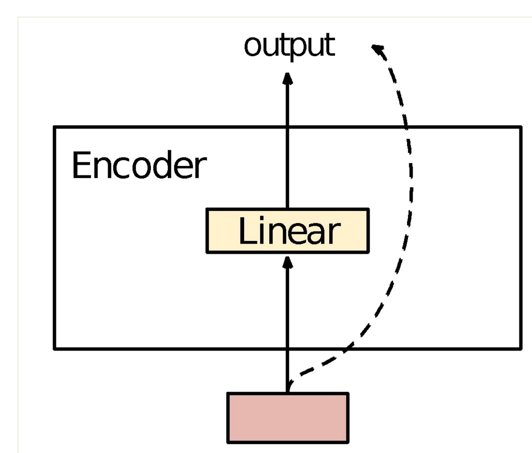
1. Temporal dependencies are noisier and sequential order matters less when looking further into the past.
2. Different scales of temporal dependencies co-exist, each of them best captured by a different architecture.

4. Multi-temporal-range Mixture Model (M3)

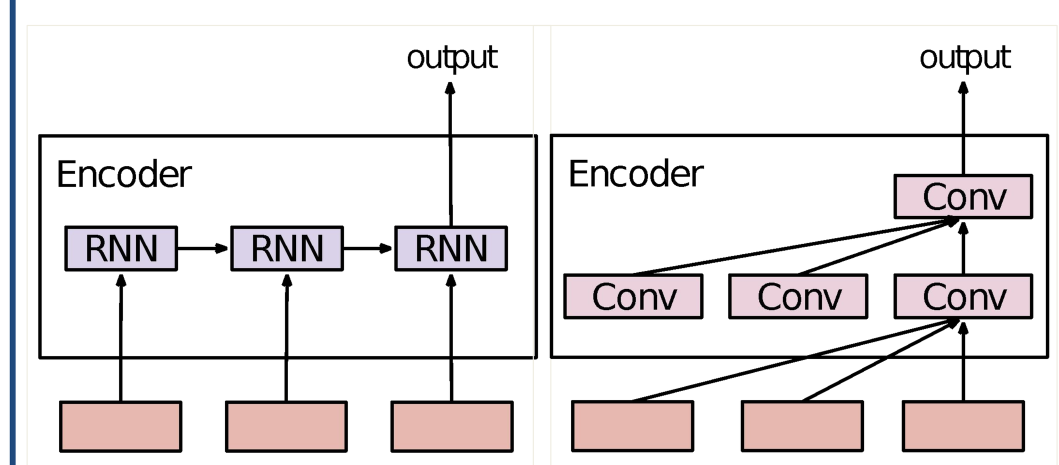
- In M3, we jointly employ three different sequence models (encoders). Each of them focus on **different ranges of temporal dependencies** in user sequences.
- We regard the three encoders as a **Mixture-of-Experts** (MOE) trained end-to-end as a single model. This structure allows the model **adapt to different recommendation scenarios** and provide insightful **interpretability.**



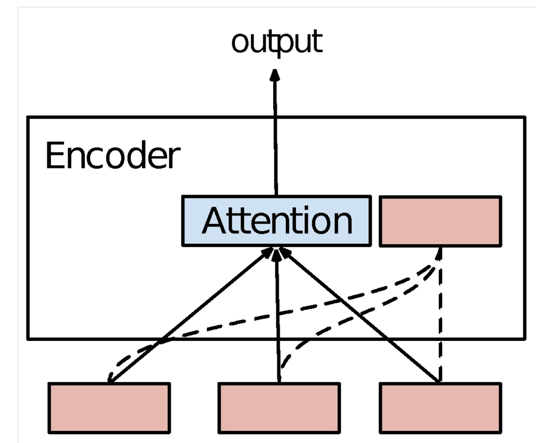
5. Temporal encoders and interpretability



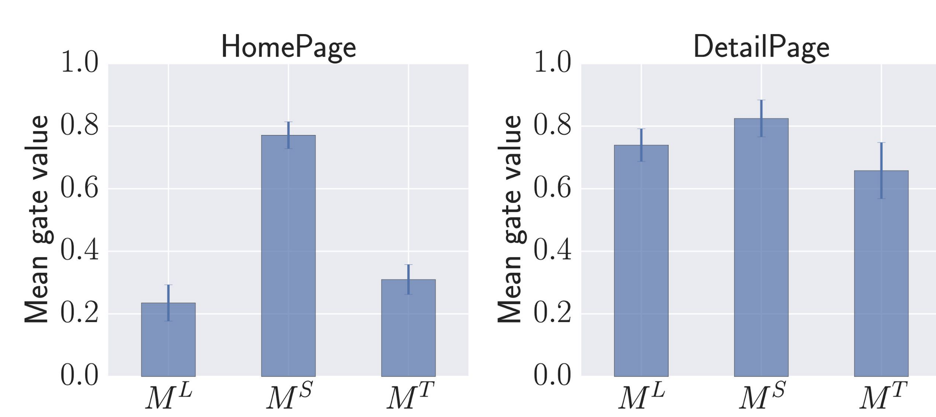
- The **Tiny-range encoder** only focuses on the user's **last event**, ignoring all previous events. It learns the **item-to-item direct co-occurrence pattern**.



- The **Short-range encoder** (a **GRU, LSTM, Temporal Convolution Network**) is **highly sensitive to order and carries more information from recent interactions** in the user sequence.



- The **Long-range sequence encoder** consists of an **Attention Model** which has a **potentially unlimited temporal range**, is robust to noise but **is not sensitive to sequential ordering**.



- **Monitoring** the average activation of the **gate** enables some **interpretability**

7. Experimental Results

- **FMC**: Factorizing model for the first-order Markov chain.
- **DeepBow**: Deep Bag-of-word model representing user by averaging item embeddings from all past events and making predictions through a feed-forward layer.
- **GRU4Rec**: Using a GRU-RNN over user sequences.
- **Caser**: Applying horizontal and vertical convolutional filters over the embedding matrix.
- **Context-FMC**: **contextual** version of *FMC*.
- **DeepYouTube**: concatenating: (1) item embedding from users' last event, (2) item embeddings averaged by all past events and (3) **context** features and makes predictions through a feed-forward layer.
- **Context-GRU**: **contextual** version of *GRU4Rec*.

- **Overall performance:**

M3R and M3C provide ***significant improvements*** over the baselines on two standard datasets for recommendations.

Results on MovieLens 20M:
Only sequential information, no context feature.

	mAP@5	mAP@10	mAP@20
FMC	0.0256	0.0291	0.0317
DeepBoW	0.0065	0.0079	0.0093
GRU4Rec	0.0256	0.0304	0.0343
Caser	0.0225	0.0269	0.0304
M3C	0.0295	0.0342	0.0379
M3R	**0.0315**	**0.0367**	**0.0421**
Improv.	+23.4%	+20.7%	+22.7%

Overall performance on YouTube:
Sequential information + context features

	mAP@5	mAP@10	mAP@20
Context-FMC	0.1103	0.119	0.1240
DeepYouTube	0.1295	0.1399	0.1455
Context-GRU	0.1319	0.1438	0.1503
M3C	0.1469	0.1591	0.1654
M3R	**0.1541**	**01670**	**0.1743**
Improv.	+16.8%	+16.1%	+16.0%

Ablation study
Encoders can address each other's shortcomings.
M3R-TSL performs best on both datasets.

	MovieLens 20M	YouTube Dataset
M3R-T	0.0269	0.1406
M3R-S	0.0363	0.1673
M3R-L	0.0266	0.1359
M3R-TS	0.0412	0.1700
M3R-TL	0.0293	0.1485
M3R-SL	0.0403	0.1702
M3R-TSL	0.0421	0.1743

The Web Conference (WWW 2019), San Francisco, CA