ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Applying the GAN Framework to Recommender Systems

Master Thesis

Mohammed Ajil

September 13, 2019

Advisors: Prof. Ce Zhang, Bojan Karlas
Department of Computer Science, ETH Zürich

Abstract

This example thesis briefly shows the main features of our thesis style, and how to use it for your purposes.

Contents

Contents **iii**

1 Introduction **1**

2 Background **3**
2.1 Recommendation Systems 3
2.1.1 Problem Statement 3
2.1.2 Variants 3
2.1.3 Well-Known Systems 3
2.2 Concepts 3
2.2.1 Recurrent Neural Networks 3
2.2.2 Generative Adversarial Network Framework 4
2.2.3 Teacher Forcing 4
2.3 Previous Work 4
2.3.1 Professor Forcing 4
2.3.2 Meta-Prod2Vec 4
2.3.3 Hierarchical RNNs for personalized Recommendations 5
2.4 KPIs 5
2.4.1 Click-Through-Rate 5
2.4.2 Conversion Rate 5

3 Dataset **7**
3.1 Data Collection 7
3.1.1 Data Generation Mechanism 7
3.1.2 Data Storage 7
3.2 Data Extraction 7
3.3 User Parallel Batches 8

4 System Overview **9**
4.1 Model Architecture 9

4.2 Implementation 9

5 Experiments **11**
5.1 Offline 11
5.1.1 Experiment Setup 11
5.1.2 Measurements 11
5.2 Online/Production Experiment Setup 11
5.2.1 Experiment Setup 11
5.2.2 Measurements 12

6 Results **13**

A Dummy Appendix **15**

Bibliography **17**

Chapter 1

Introduction

- Recommender System are a field in machine learning that get more and more attention
- In principle the goal of a recommender system is to recommend items (in the case of an e-commerce system products) to users
- The idea is that the system helps the user navigate the mass of products available and to show the user what is relevant
- A recomender system can be tuned to optimize for different metrics such as click through rate, conversion rate etc.
- In the past there have been mainly to approaches to recommender systems user based and item based
- In the last years there have been more and more proposals for session based approaches (refer to paper that does the survey)
- Specifically GRU4Rec has gained attention as a RNN based approach to solving this problem
- In a session based setting we try to model the sequence of events a user makes when he browses the product catalog instead of items or users themselves.
- This has the following advantages: Usually we have a lot of users that only visit the site once or twice a year, these users are very difficult to model. The same goes for products, there are a lot of products that have limited attention from the userbase, therefore it is difficult to get a useful representation for products like this.
- In a session based setting we model the sequence of events, which is more useful since there is a lot more data, and sessions are comparable even if we do not know which user is online.

- Session based approaches work in both settings where we dont know the user and where we know the user
- Item based are for unkown users and user based for known users
- The goal of the thesis is twofold:
  - First we want to explore the capabilities of such a system in a production setting with real world data. Are we really better than simple approaches like "often bought together"?
  - Also we want to compare it to a more sophisticated system that is a blackbox and consumed as a service provided by Google.
  - Third, since this is an RNN approach we want to find out if we can improve the system by applying the GAN framework in form of professor forcing
  - Fourth, the system inherently produces embeddings for products and users, we want to find out if we can improve the overall performance if we use a pretrained product embedding which captures the semantic similarity of products

Chapter 2

Background

2.1 Recommendation Systems

2.1.1 Problem Statement

- The general form of a recommendation system is that we assign scores to items.
- A personalized recommendation system uses the context of the user to produce scores dependant of the active user.
- Write out a function that describes a general recommendation system

2.1.2 Variants

- User Based
- Item Based
- Session Based

2.1.3 Well-Known Systems

- Describe a classical user based approach
- Describe a classical item based approach

2.2 Concepts

2.2.1 Recurrent Neural Networks

- Explain the concept of neural Networks

- What are the problems (vanishing gradients etc)
- Explain what a GRU is and why does it handle vanishing gradient better than simple RNN cells?
- How does a layered RNN work?

2.2.2 Generative Adversarial Network Framework

- Explain the way GANs work
- Loss function is learnable
- The discriminator is part of the loss function
- Especially good if we want generate stuff

2.2.3 Teacher Forcing

- RNN have two modes: Teacher forced and Free running
- When training we always use the ground truth as input
- This generates two different distributions of hidden states, which in turn generates two different distributions for the output
- This is a problem since this makes RNNs in production less predictable

2.3 Previous Work

2.3.1 Professor Forcing

- Here we want to mitigate the teacher forcing problem, by applying the gan framework
- Add graphic showing how it should work
- Describe the model as we did in the slides, show the loss function

[1]

2.3.2 Meta-Prod2Vec

- This is a model that allows us to capture the semantic similarity between products
- Explain how it works by relating to word2vec (refer to word2vec paper)

[3]

2.3.3 Hierarchical RNNs for personalized Recommendations

- This is the starting point for the thesis
- Which components does the system have?
- Show the graphic from the paper

[2]

2.4 KPIs

2.4.1 Click-Through-Rate

2.4.2 Conversion Rate

- Describe different KPIs
- What do they meaure, how to optimize for it

Chapter 3

Dataset

3.1 Data Collection

3.1.1 Data Generation Mechanism

- Each click the user makes on the site is tracked
- An event is generated containing the information and context of the click of the user
- This event is sent to a message queue
- From the message queue there is a subscriber that archives this event in Google BigQuery
- The most important fields tracked are: UserId, SessionId, ProductId, Timestamp, LastLoggedInUserId
- Google BigQuery is a Data Warehousing solution consisting of append-only tables

3.1.2 Data Storage

- Data is stored in BigQuery in a denormalized form
- One row describes one event

3.2 Data Extraction

- First we extract the events on ProductDetail Pages from BigQuery
- This is done in shards, where each shard approximately corresponds to one day of events

- We only use the data where we know which user produced the event
- The events are filtered first by removing all events generated by known bot user agents
- In a second step we merge the fields UserId and LastLoggedInUserId, if the latter is set we know that the user ”passively” logged out of the shop (long time without a session)
- Then the events are grouped by sessions
- Then the sessions are grouped by users
- After that we filter out users that have too few sessions
- Then we filter products that have too few events
- After that we have a dataset that contains all users with enough sessions and all events on products with enough events
- Show the structure of one user in the json

3.3 User Parallel Batches

- Show how the user parallel batches should look like
- Explain how we transform the json dataset to these user parallel batches

Chapter 4

System Overview

4.1 Model Architecture

- Describe Model Architecture
- Describe different Components of Model Architecture
- Describe different variants of the model (with/without pf, with/without embeddings)

4.2 Implementation

- Describe Class Diagram
- Describe Code Structure
- Describe Deployment
- Describe prediction mode/training mode

Chapter 5

Experiments

For each of the model variants evaluate the experiments offline and online.

5.1 Offline

5.1.1 Experiment Setup

- Describe the last session out split (test and eval)
- Describe how we train the model and how we evaluate it

5.1.2 Measurements

- Here we want to confirm a few things: user rnn helps, embeddings helps, profforcing helps
- The evaluation set is always the same
- Show the relevant KPIs (described in the chapter before)

5.2 Online/Production Experiment Setup

5.2.1 Experiment Setup

- Describe the online shop
- Where do we have recommendations, general description
- Explain SOFI
- Where do we put our recommendation system? Against what does it compete?

5.2.2 Measurements

- Show results for AutoML, Often Bought Together and our system in the different variants
- Show the relevant KPIs

Chapter 6

Results

- bring everything together
- compare best performing model to automl from google

Appendix A

Dummy Appendix

You can defer lengthy calculations that would otherwise only interrupt the flow of your thesis to an appendix.

Bibliography

[1] Alex Lamb, Anirudh Goyal, Ying Zhang, Saizheng Zhang, Aaron Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. http://arxiv.org/abs/1610.09038, 2016.

[2] Massimo Quadrana, Alexandros Karatzoglou, Balázs Hidasi, and Paolo Cremonesi. Personalizing session-based recommendations with hierarchical recurrent neural networks. http://arxiv.org/abs/1706.04148, 2017.

[3] Flavian Vasile, Elena Smirnova, and Alexis Conneau. Meta-prod2vec - product embeddings using side-information for recommendation. http://arxiv.org/abs/1607.07326, 2016.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.