Beggars Can't Be Choosers

Augmenting Sparse Data for Embedding-Based Product Recommendations in Retail Stores

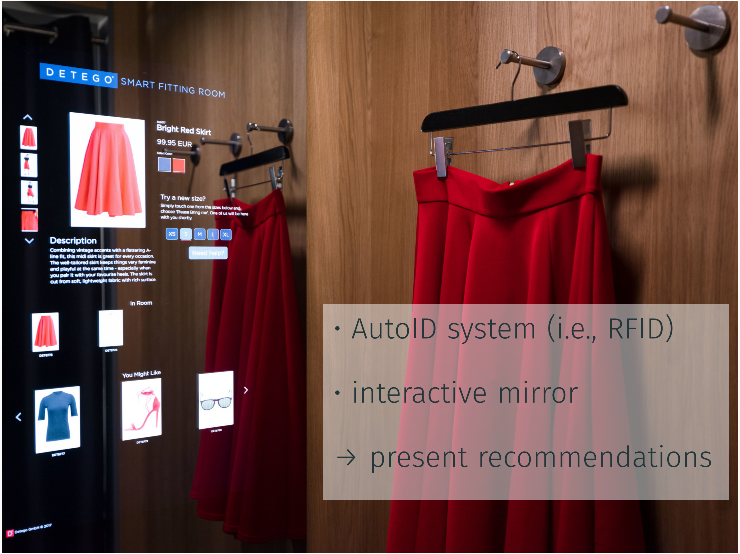
Matthias Wölbitsch[1,2], Simon Walk[1], Michael Goller[1], and Denis Helic[2]

June 11, 2019

[1] Detego GmbH
[2] ISDS, Graz University of Technology

Motivation

Recommender systems (RS) in brick-and-mortar stores



- AutoID system (i.e., RFID)
- interactive mirror

→ present recommendations

Motivation

Challenges related to RS for fashion retailer

- (often) no customer purchase histories
- feedback for recommended products
- limited, sparse, and location-dependent data

→ Traditional RS approaches are often not applicable

Goal: tackle outlined issues related to RS for (fashion) retailer

Approach

Recommendations leveraging shopping baskets

- embedding-based approach (prod2vec[1])
- low-dimensional vector representation of products p_i

$$\mathcal{L} = \sum_{B \in \mathcal{B}} \sum_{\substack{p_i, p_j \in B \\ p_i \neq p_j}} \log \Pr(p_j \mid p_i)$$

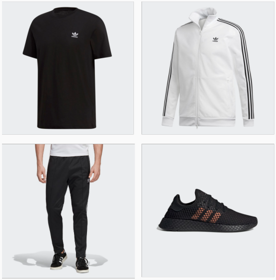
predict remaining products in shopping basket $B \in \mathcal{B}$
(i.e., skip-gram architecture)

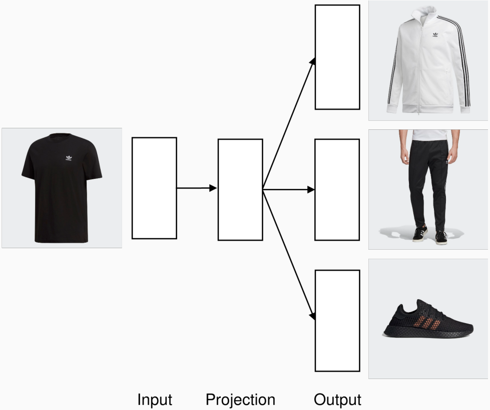
[1]Grbovic et al., “E-commerce in Your Inbox: Product Recommendations at Scale”.

Approach

Training using the skip-gram model

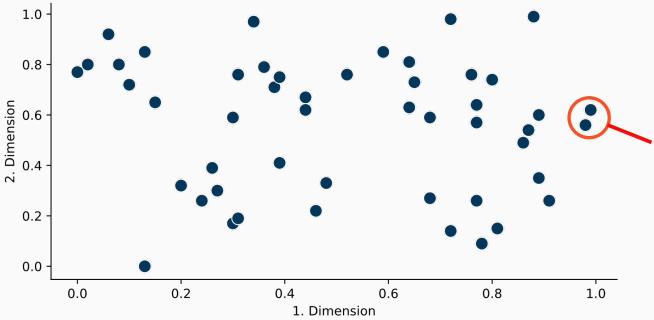
Shopping Basket

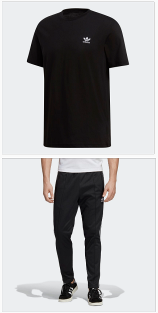




Computing recommendations

- related products are close in the embedding space
- find k nearest neighbours for recommendations





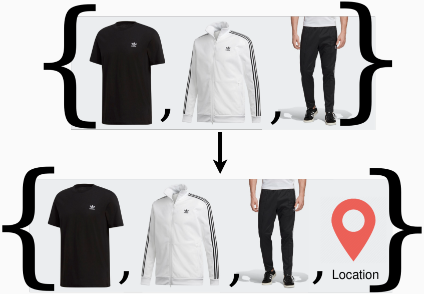
Approach

Problem

- environment and demographics affect purchase behaviour
- product assortments often vary drastically across stores

Point-of-sale (POS) extension

- extend shopping baskets with point-of-sale information
- general model which also captures local specialties



Approach

Problem

- only very limited number of shopping baskets available
- due to changing assortments, few sales,...

Data augmentation

- generate new training examples based on existing baskets
- strategies: repetitions, combinations,...

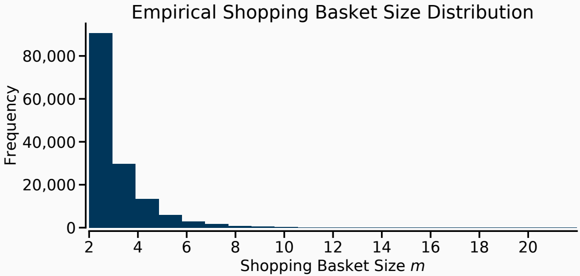
Experimental Setup

Dataset

- 20 stores across 4 cities
- 150,000 shopping baskets
- 17,000 distinct products

Evaluation

- predict remaining products in shopping basket
- metrics: $Recall_k$ and $NDCG_k$



https://github.com/detegoDS/shopping_basket_dataset

Results | POS Extension

		mean $NDCG_{k=20}$	mean $Recall_{k=20}$
	individual models		
(a)	co-purchase baseline	0.0637	0.1938
(b)	prod2vec baseline	0.1022	0.1544
	general models		
(c)	co-purchase baseline	0.0767	**0.2308**
(d)	prod2vec baseline	0.1292	0.1937
(e)	**prod2vec w/ POS**	**0.1323**	0.1968

Findings:

- general model benefits from additional POS information
- tackle location-based cold-start problem

Results | Data Augmentation

		mean $NDCG_{k=20}$	mean $Recall_{k=20}$
	baseline		
(a)	prod2vec baseline	0.1292	0.1937
	data augmentation		
(b)	replicated baskets prod2vec	0.1279	0.1868
(c)	pair & triple augmented prod2vec	0.1285	0.1866
(d)	pair augmented prod2vec	**0.1334**	**0.1972**

Findings:

- augmentation based on extracted pairs performs best
- introduction of additional & novel training contexts
- captures latent pairwise relationships of products

Results | Combined Approaches

		mean $NDCG_{k=20}$	mean $Recall_{k=20}$
	baselines		
(a)	co-purchase baseline	0.0767	**0.2308**
(b)	prod2vec baseline	0.1292	0.1937
	combined approaches		
(c)	pair augmented POS prod2vec	0.1351	0.1999
(d)	ensemble	**0.1382**	0.2152

Findings:

- augmentation and POS extension complement each other
- ensemble using co-purchase information further improves performance

Conclusion

Contributions

- RS framework for brick-and-mortar (fashion) retailer
  - point-of-sale extension & data augmentation
  - overall 7% improvement in recommendation performance
- real-world dataset[2]

Future Work

- recommendations adapted for different settings/domains
- leverage smart fitting room data

[2] https://github.com/detegoDS/shopping_basket_dataset

Dataset Properties

	# shopping baskets	# products	mean product overlap
City A	31,583 (21.53%)	11,819	0.47
City B	30,269 (20.63%)	9,318	0.51
City C	53,108 (36.20%)	11,313	0.54
City D	31,760 (21.65%)	9,597	0.53
Total	146,720 (100.0%)	17,392	