

BERT Goes Shopping: Comparing Distributional Models for Product Representations

Federico Bianchi

Bocconi University
Milano, Italy

f.bianchi@unibocconi.it*

Bingqing Yu

Coveo
Montreal, CA

cyu2@coveo.com

Jacopo Tagliabue

Coveo AI Labs
New York, United States

jtagliabue@coveo.com

Abstract

Word embeddings (e.g., word2vec) have been applied successfully to eCommerce products through *prod2vec*. Inspired by the recent performance improvements on several NLP tasks brought by contextualized embeddings, we propose to transfer BERT-like architectures to eCommerce: our model – *ProdBERT* – is trained to generate representations of products through masked session modeling. Through extensive experiments over multiple shops, different tasks, and a range of design choices, we systematically compare the accuracy of *ProdBERT* and *prod2vec* embeddings: while *ProdBERT* is found to be superior to traditional methods in several scenarios, we highlight the importance of resources and hyperparameters in the best performing models. Finally, we conclude by providing guidelines for training embeddings under a variety of computational and data constraints.

1 Introduction

Distributional semantics (Landauer and Dumais, 1997) (hence, DS) is built on the assumption that the meaning of a word is given by the contexts in which it appears: word embeddings obtained from co-occurrence patterns through *word2vec* (Mikolov et al., 2013), proved to be both accurate by themselves in representing lexical meaning, and very useful as components of larger Natural Language Processing (NLP) architectures (Conneau et al., 2017). The empirical success and practical scalability of *word2vec* objective gave rise to many domain-specific models (Ng, 2017; Grover and Leskovec, 2016; Yan et al., 2017): in eCommerce, *prod2vec* is trained replacing words in a sentence with product interactions in a shopping session (Grbovic et al., 2015), eventually generating a vector representation of the products. The model enjoyed immediate success in the field and it is now important in NLP

and Information Retrieval (IR) use cases for eCommerce (Vasile et al., 2018; Bianchi et al., 2020).

As a key improvement over *word2vec*, the NLP community has recently introduced *contextualized representations*, in which a word like *play* would have different embeddings depending on the general topic (e.g. a sentence about *theater* vs *soccer*), where as in *word2vec* the word *play* is going to have only one vector. Transformer-based architectures (Vaswani et al., 2017) in large-scale models - such as BERT (Devlin et al., 2019) - achieved SOTA results in many tasks (Nozza et al., 2020; Rogers et al., 2020). As Transformers are being applied outside of NLP (Chen et al., 2020), it is natural to ask whether we are missing a fruitful analogy with *product representations*. It is *a priori* reasonable to think that a pair of sneakers can have different representations depending on the shopping context: is the user interested in buying these shoes because they are running shoes, or because these shoes are made by her favorite brand?

In *this* work, we explore the adaptation of *BERT*-like architectures to eCommerce: through extensive experimentation on downstream tasks and empirical benchmarks on typical digital retailers, we discuss advantages and disadvantages of contextualized embeddings when compared to traditional *prod2vec*. We summarize our main contributions as follows:

1. We propose and implement a BERT-based contextualized product embeddings model (hence, **ProdBERT**), which can be trained with online shopper behavioral data and produce product embeddings to be leveraged by downstream tasks.
2. We benchmark ProdBERT against *prod2vec* embeddings, showing the potential accuracy gain of contextual representations across different shops and data requirements.

*Corresponding Author. Federico and Bingqing contributed equally to this research.

3. We perform extensive experiments with several hyperparameters, architectures and fine-tuning strategies. We report detailed results from various evaluation tasks, and finally provide guidelines on optimal model designs and recommendations on how to best trade off accuracy with training cost.

1.1 Product Embeddings: an Industry Perspective

The eCommerce industry has been steadily growing in recent years: according to [U.S. Department of Commerce \(2020\)](#), 16% of all retail transactions now occur online; worldwide eCommerce is estimated to turn into a \$4.5 trillion industry in 2021 ([Statista Research Department, 2020](#)). Interest from researchers has been growing at the same pace ([Tsagkias et al., 2020](#)), stimulated by challenging problems and by the large-scale impact that machine learning systems have in the space ([Pichestapong, 2019](#)).

Within the fast adoption of deep learning methods in the field ([Ma et al., 2020](#); [Zhang et al., 2020](#); [Yuan et al., 2020](#)), product representations obtained through *prod2vec* play a key role in many neural architectures: after training, a product space can be used directly ([Vasile et al., 2016](#)), as a part of larger systems for recommendation ([Tagliabue et al., 2020b](#)), or in downstream NLP/IR tasks ([Bianchi et al., 2020](#); [Tagliabue and Yu, 2020](#)). Combining the size of the market with the past success of NLP models in the space, investigating whether Transformer-based architectures result in superior product representations is both theoretically interesting and practically important.

Anticipating some of the themes below, it is worth mentioning that our study sits at the intersection of two important trends: on one side, neural models typically show significant improvements at *very large* scale ([Kaplan et al., 2020](#)) – quantifying expected gains for “reasonable-sized” shops makes our findings relevant even outside public companies and allows principled trade-off between accuracy and ethical considerations ([Strubell et al., 2019](#)); on the other side, the rise of multi-tenant players¹ makes sophisticated models potentially applicable to an unprecedented number of shops

– in this regard, we design our methodology to include *multiple* shops in our benchmarks, and report how training resources and accuracy scale across deployments. For these reasons, we believe our findings will be interesting to a wide range of researchers and practitioners.

2 Related Work

Distributional Models. *Word2vec* ([Mikolov et al., 2013](#)) enjoyed great success in NLP thanks to its computational efficiency, unsupervised nature and accurate semantic content ([Levy et al., 2015](#); [Al-Saqqa and Awajan, 2019](#); [Conneau et al., 2017](#)). Recently, models such as BERT ([Devlin et al., 2019](#)) and RoBERTa ([Liu et al., 2019](#)) shifted much of the community attention to Transformer architectures and their performance ([Talmor and Berant, 2019](#); [Vilares et al., 2020](#)), while it is increasingly clear that big datasets ([Kaplan et al., 2020](#)) and substantial computing resources play a role in the overall accuracy of these architectures; in our experiments, we explicitly address robustness by *i*) varying model designs, together with other hyperparameters; and *ii*) test on multiple shops, differing in traffic, industry and product catalog.

Product Embeddings. *Prod2vec* is a straightforward adaptation to eCommerce of *word2vec* ([Grabovic et al., 2015](#)). Product embeddings quickly became a fundamental component for recommendation and personalization systems ([Caselles-Dupré et al., 2018](#); [Tagliabue et al., 2020a](#)), as well as NLP-based predictions ([Bianchi et al., 2020](#)). To the best of our knowledge, *this* work is the first to explicitly carry the NLP-to-product analogy one step further, and investigate whether Transformer-based architectures deliver higher-quality representations compared to non-contextual embeddings. The closest work in the literature as far as model architecture goes is the recent *BERT4Rec* ([Sun et al., 2019](#)), i.e. a model based on Transformers trained end-to-end for recommendations. The focus of our work is not so much the gains induced by Transformers in sequence modelling, but instead is the quality of the representations obtained through unsupervised pre-training – while recommendations are certainly important, *prod2vec* literature shows the need for a more thorough and general assessment. While our initial findings can be seen as a confirmation on real-world datasets of what is reported by [Sun et al. \(2019\)](#) for sequential prediction, our research uncovers a tighter-than-expected

¹As an indication of the market opportunity, only in 2019 and only in the space of AI-powered search and recommendations, we witnessed Algolia raising USD110M ([Techcrunch, 2019a](#)), Lucidworks raising USD100M ([Techcrunch, 2019c](#)) and Coveo raising CAD227M ([Techcrunch, 2019b](#)).

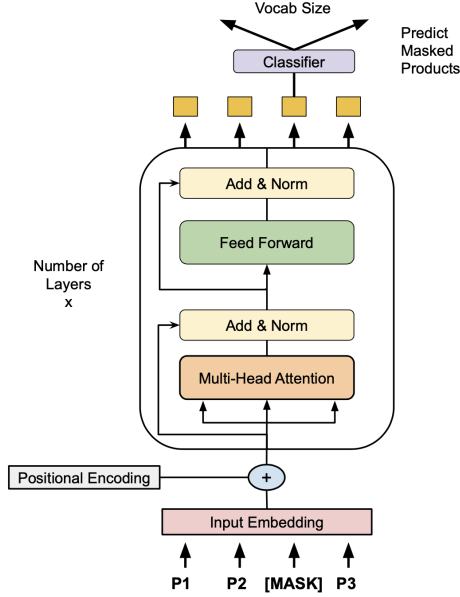


Figure 1: Overall architecture of ProdBERT pre-trained on MLM task.

gap between the models as far as downstream tasks go. Moreover, our extensive tuning and industry-specific benchmarks allow us to draw novel conclusions on optimal design for embeddings model across a variety of scenarios, and to give practitioners insights and practical suggestions for industry deployment.

3 ProdBERT

3.1 Overview

The ProdBERT model is taking inspiration from BERT architecture and aims to learn context-dependent vector representation of products from online session logs. By considering a shopping session as a “sentence” and the products shoppers interact with as “words”, we can transfer masked language modeling (MLM) from NLP to eCommerce. Once trained, ProdBERT becomes capable of predicting masked tokens, as well as providing context-specific product embeddings for downstream tasks.

3.2 Model Architecture

As shown in Figure 1, ProdBERT is based on BERT model architecture, which corresponds to the encoder part of Transformer from Vaswani et al. (2017). Different from BERT’s original implementation, a white-space tokenizer is first used to split an input session into tokens, each one representing a product ID; token are combined with positional

encodings via addition and fed into a stack of self-attention layers, where each layer contains a block for multi-head attention, followed by a simple feed forward network. After obtaining the output from the last self-attention layer, the vectors corresponding to the masked tokens pass through a softmax to generate the final predictions.

3.3 Training Objective

Similar to Liu et al. (2019) and Sun et al. (2019), we train ProdBERT from scratch with the MLM objective. A random portion of the tokens (i.e., the product IDs) in the original sequence is chosen for possible replacements with the *MASK* token; and the masked version of the sequence is fed into the model as input. The target output sequence is exactly the original sequence without any masking, thus the training objective is to predict the original value of the masked tokens, based on the context provided by their surrounding unmasked tokens. The model learns to minimize categorical cross-entropy loss, taking into account only the predicted masked tokens, i.e. the output of the non-masked tokens is discarded for back-propagation.

3.4 Hyperparameters and Design Choices

There is growing literature investigating how different hyperparameters and architectural choices can affect Transformer-based models. For example, Lan et al. (2020) observed diminishing returns when increasing the number of layers after a certain point; Liu et al. (2019) showed improved performance when modifying masking strategy and using duplicated data; finally, Kaplan et al. (2020) reported slightly different findings from previous studies on factors influencing Transformers performance. Hence, it is worth studying the role of hyperparameters and model designs of ProdBERT, in order to narrow down which settings are the best given the specific target of our work, i.e. *product representations*.

Table 1 shows the relevant hyperparameter and design variants for ProdBERT; following improvement in data generalization reported by Liu et al. (2019), when *duplicated* = 1 we augmented the original dataset repeating each session 5 times.² We set the embedding size to 64 after preliminary optimizations: as other values offered no improvements, we report results only for one size.

²This procedure ensures that each sequence can be masked in 5 different ways during training.

Parameter	Values
# epochs [e]	10, 20, 50, 100
# layers [l]	4, 8
masking probability [m]	0.15, 0.25
duplicated [d]	1, 0

Table 1: Hyperparameters and their ranges.

4 Methods

4.1 Prod2vec: a Baseline Model

We benchmark ProdBERT against the industry standard *prod2vec* (Grbovic et al., 2015). More specifically, we train a CBOW model with negative sampling over shopping sessions (Mikolov et al., 2013). Since the role of hyperparameters in *prod2vec* has been extensively studied before (Caselles-Dupré et al., 2018), we prepare embeddings according to the best practices in Bianchi et al. (2020) and employ the following configuration: *window* = 15, *iterations* = 30, *ns_exponent* = 0.75, *dimensions* = [48, 100].

4.2 Dataset

We collected search logs and detailed shopping sessions (anonymized user interactions) from two partnering shops, **Shop A** and **Shop B**. They are mid-sized digital shops, with revenues between 25 and 100 millions USD/year. Shop A and Shop B differ in many aspects, from traffic to the catalog structure: Shop A is in the sport apparel category, whereas Shop B is in home improvement. Sessions for training are sampled with undisclosed probability from the period of March-December 2019; testing sessions are a completely disjoint dataset from January 2020. After pre-processing³, detailed descriptive statistics for the training dataset can be found in Table 2. For fairness of comparison, the exact same datasets are used for both ProdBERT and *prod2vec*.

5 Experiments

5.1 Experiment #1: Next Event Prediction

We choose Next Event Prediction (NEP) as our first evaluation task for ProdBERT. NEP is a standard way to evaluate the quality of product representations (Letham et al., 2013; Caselles-Dupré et al.,

³We only keep sessions that have between 3 and 20 product interactions, to eliminate unreasonably short sessions and ensure computation efficiency.

Shop	Sessions	Products	50/75 pct
Shop A	1,970,832	38,486	5, 7
Shop B	3,992,794	102,942	5, 7

Table 2: Descriptive statistics for the training dataset. *pct* shows 50th and 75th percentiles of the session length.

2018): briefly, NEP consists in predicting the next action the shopper is going to perform given her past actions. Hence, in the case of ProdBERT, we mask the last item of every session and fit the sequence as input to a pre-trained ProdBERT model (this is similar to the standard word prediction task for cloze sentences in the NLP literature (Petroni et al., 2019)). Provided with the model’s output sequence, we take the top K most likely values for the masked token, and perform comparison with the true interaction. As for *prod2vec*, we perform the NEP task by following industry best practices (Ludewig et al., 2019; Bianchi et al., 2020): given a trained *prod2vec*, we take all the before-last items in a session to construct a session vector by averaging their embedding vectors, and use k-NN to predict the last item⁴. Following industry standards, $nDCG@K$ (Mitra and Craswell, 2018) is the chosen metric⁵ with $K = 10$, and all tests ran on 10,000 testing cases (test set is randomly sampled first, and then shared across ProdBERT and *prod2vec* to guarantee a fair comparison).

5.1.1 Results

Table 3 reports results on the NEP task by highlighting some key configurations that led to competitive performances. ProdBERT is significantly superior to *prod2vec*, scoring up to 40% higher than the best *prod2vec* configurations. Since shopping sessions are significantly shorter than sentence lengths in Devlin et al. (2019), we found that changing masking probability from 0.15 (value from standard BERT) to 0.25 consistently improved performance by making the training more effective. As for the number of layers, similar to Lan et al. (2020), we found that adding layers help only up

⁴Previous experiments using LSTM in NEP (Tagliabue et al., 2020b) showed some improvements over k-NN; however, the differences are not significant enough to narrow the gap we have found between *prod2vec* and ProdBERT. Hence, k-NN is chosen in this work for consistency with the standards in the relevant literature.

⁵We also tracked $HR@10$, but given insights were similar, we omitted it for brevity in what follows.

<i>Model</i>	<i>Config</i>	Shop A	Shop B
ProdBERT	$e = 10, l = 4,$ $m = 0.25, d = 0$	0.433	0.259
ProdBERT	$e = 5, l = 4,$ $m = 0.25, d = 1$	0.458	0.282
ProdBERT	$e = 10, l = 8,$ $m = 0.25, d = 0$	0.027	0.260
ProdBERT	$e = 100, l = 4,$ $m = 0.25, d = 0$	0.427	0.255
ProdBERT	$e = 10, l = 4,$ $m = 0.15, d = 0$	0.416	0.242
<i>prod2vec</i>	<i>dimension</i> = 48	<u>0.326</u>	0.214
<i>prod2vec</i>	<i>dimension</i> = 100	0.326	<u>0.218</u>

Table 3: $nDCG@10$ on NEP task for both shops with ProdBERT and *prod2vec* (**bold** are best scores for ProdBERT; underline are best scores for *prod2vec*).

until a point: with $l = 8$, training ProdBERT with more layers resulted in a catastrophic drop in model performance for the smaller Shop A; however, the same model trained on the bigger Shop B got a small boost. Finally, duplicating training data has been shown to bring consistent improvements: while keeping all other hyperparameters constant, using duplicated data results in a 6% increase in $nDCG@10$, not to mention that after only 5 training epochs the model outperforms other configurations trained for 10 epochs or more.

While encouraging, the performance gap between ProdBERT and *prod2vec* is consistent with Transformers performance on sequential tasks (Sun et al., 2019). However, as argued in Section 1.1, product representations are used as input to many downstream systems (Tagliabue et al., 2020a), making it essential to evaluate how the learned embeddings generalize outside of the pure sequential setting.

Our second experiment is therefore designed to test how well contextual representations transfer to other eCommerce tasks, helping us to assess the accuracy/cost trade-off when difference in training resources between the two models is significant – the best-performing ProdBERT model requires 24 minutes per epoch for Shop A, and 2 hours per epoch for Shop B; however full *prod2vec* training can be completed within 4 minutes for Shop A and 20 minutes for Shop B (training on a *Tesla V100 16GB* GPU and with *batch size* of 256).

5.2 Experiment #2: Intent Prediction

A crucial element in the success of Transformer-based language model is the possibility of adapting the representation learned through pre-training to new tasks: for example, the original Devlin et al. (2019) fine-tuned the pre-trained model on 11 downstream NLP tasks. However, the practical significance of these results across a wider range of application is still unclear: on one hand, Li et al. (2020); Reimers and Gurevych (2019) observed that sometimes BERT contextual embeddings can underperform a simple GloVe (Pennington et al., 2014) model; on the other, practitioners are reporting instability issues in fine-tuning – for example, Mosbach et al. (2020) highlights catastrophic forgetting, vanishing gradients and data variance as important factors in practical failures. Hence, given the range of downstream applications featuring product embeddings and the active debate on transferability in NLP, we investigate how ProdBERT representations perform when used in the *intent prediction* task.

Intent prediction is the task of guessing whether a shopping session will eventually ends in the user adding items to the cart (signaling purchasing intention). Since small increases in conversion can be translated into massive revenue boosting, this task is both a crucial problem in the industry and an active area of academic research (Toth et al., 2017; Requena et al., 2020). To implement the intent prediction task, we randomly sample from our dataset 20,000 sessions ending with an add-to-cart actions and 20,000 sessions without add-to-cart, and split the resulting dataset for training, validation and test. Hence, given the list of previous products that a user has interacted with, the goal of the intent model is to predict whether an add-to-cart event will happen or not.

We experimented with several adaptation techniques inspired by best practices from the most recent NLP literature (Peters et al., 2019; Li et al., 2020):

1. *Feature extraction (static)*: we extract the contextual representations from a target hidden layer of pre-trained ProdBERT, and through average pooling, feed them as input to a multi-layer perceptron (MLP) classifier to generate the binary prediction. In addition to alternating between the first hidden layer (*enc_0*) to the last hidden layer (*enc_3*), we also tried concatenation (*concat*), i.e. combining em-

Model	Method	Shop	Accuracy
ProdBERT	<i>enc_0</i>	Shop B	0.567
ProdBERT	<i>enc_3</i>	Shop B	0.547
ProdBERT	<i>concat</i>	Shop B	0.553
ProdBERT	<i>wal</i>	Shop B	0.543
ProdBERT	<i>fine-tune</i>	Shop B	0.560
<i>prod2vec</i>	-	Shop B	0.558
ProdBERT	<i>enc_0</i>	Shop A	0.593
<i>prod2vec</i>	-	Shop A	0.602

Table 4: Accuracy scores in the intent prediction task (**bold** are best scores for each shop).

beddings of all hidden layers via concatenation before average pooling.

2. *Feature extraction (learned)*: we implement a linear weighted combination of all hidden layers (*wal*), with learnable parameters, as input features to the MLP model (Peters et al., 2019).
3. *Fine-tuning*: we take the pre-trained model up until the last hidden layer and add the MLP classifier on top for intent prediction (*fine-tune*). During training, both ProdBERT and task-specific parameters are trainable.

As for our baseline, i.e. *prod2vec*, we implement the intent prediction task by encoding each product within a session with their *prod2vec* embeddings, and feed them to a LSTM network (so that it can learn sequential information) followed by a binary classifier to obtain the final prediction.

5.2.1 Results

From our experiments, Table 4 highlights the most interesting results obtained adapting to the new task the best-performing ProdBERT and *prod2vec* models from NEP. As a first consideration, the shallowest layer of ProdBERT for feature extraction outperforms all other layers, and even beats concatenation and weighted average strategies⁶. Second, the quality of contextual representations of ProdBERT is highly dependent on the amount of data used in the pre-training phase. Comparing Table 3 with Table 4, even though the model delivers strong results in the NEP task on Shop A, its performance on the

⁶This is consistent with Peters et al. (2019), which states that inner layers of a pre-trained BERT encode more transferable features.

transfer learning task is weak, as it remains inferior to *prod2vec* across all settings. In other words, the limited amount of traffic from Shop A is not enough to let ProdBERT form high-quality product representations⁷; however thanks to the particular architecture design of ProdBERT, the model can still effectively leverage the data and perform well on the NEP task, especially since the nature of NEP is closely aligned with the pre-training task, MLM. Third, fine-tuning instability is encountered and has a severe impact on model performance. Since the amount of data available for intent prediction is not nearly as important as the data utilized for pre-training ProdBERT, overfitting has been observed throughout our fine-tuning experiments. In addition, even with optimized hyperparameters, e.g. decreased learning rate, the fine-tuned models still fail to deliver performance equivalent to the best-performing method, i.e. *enc_0*. Fourth, by comparing the results of our best method against the model learnt with *prod2vec* embeddings, we observed *prod2vec* embeddings can only provide limited values for intent estimation and the LSTM-based model stops to improve very quickly; in contrast, the features provided by ProdBERT embeddings seem to encode more valuable information, allowing the model to be trained for longer epochs and eventually reaching a higher accuracy score. Despite the superiority of ProdBERT embeddings revealed from our studies, it is worth noting that the performance gap between ProdBERT and *prod2vec* embeddings is extremely small, and extensive hyperparameter and model architectural search are needed to fully take advantage of the contextual embeddings and deliver competitive values for new tasks.

6 Conclusions and Future Work

Transformer-based models have gained increasing popularity and achieved great success in the NLP community. However, how much these gains translate to other use cases in global industries – such as eCommerce products – is still a topic of research worthy of investigation. In *this* work, we explored the performance of contextualized product representations as trained through a BERT-inspired neural network, *ProdBERT*. By thoroughly benchmarking ProdBERT against *prod2vec* in a multi-shop setting, we were able to uncover important insights

⁷Please, see Appendix A for some data visualization and qualitative considerations.

on the relationship between hyperparameters, adaptation strategies and eCommerce performances on one side, and we could quantify for the first time quality gains across different deployment scenarios, on the other. If we were to sum up our findings for interested practitioners, these are our highlights:

1. Generally speaking, our experimental setting and real-world datasets proved that pre-training ProdBERT with Mask Language Modeling can be applied successfully to sequential prediction problems in eCommerce. These results provide independent confirmation for the findings in Sun et al. (2019), where BERT was used for in-session recommendations over academic datasets. However, the tighter gap on downstream tasks (Table 4) suggests that Transformers’ ability to model long-range dependencies may be more important than pure representational quality in the NEP task.
2. Our investigation on adapting pre-trained contextual embeddings for downstream tasks featured several strategies in feature extraction and fine-tuning (Peters et al., 2019). Our analysis showed that feature-based adaptation leads to the peak performance, as compared to its fine-tuning counterpart.
3. Dataset size *does* indeed matter: as evident from the performance difference in Table 4, ProdBERT shows bigger gains with the largest amount of training data available. Considering the amount of resources needed to train and optimize ProdBERT (Section 5.1.1), the performance gains of contextualized embedding may not be worth the investment for many digital shops with ranking above 5k in the Alexa list; on the other hand, our results demonstrate that with careful and strategic model optimization, shops with a large user base and significant computational resources may achieve superior results with ProdBERT.

While our findings are encouraging, we believe that there are still many interesting questions to tackle when pushing ProdBERT further. In particular, our results require a more detailed discussion with respect to the success of BERT for textual representations, with a focus on two differences between words and products: first, an important aspect of BERT is the tokenizer, that splits words into

subwords – this component is absent in our setting because there exists no straightforward concept of “sub-product”. Second, while product representations may well be context dependent *to some extent*, it may be hard to find genuine cases of multiple meanings, such as those typical of lexical items.

We leave the answer to these questions – as well as the possibility of adapting ProdBERT to even more downstream tasks inspired by the versatility of *prod2vec* – to the next iteration of this project.

Acknowledgements

The authors wish to thank *Coveo* for providing the computational resources used to carry out the experiments, and Luca Bigon, whose engineering skills aptly minimized our “frustration function” when working with Big Data. Federico would like to thank Debora Nozza for the insightful comments on a previous version of this work. Finally, Jacopo wishes to thank the owners of *Shop A* and *Shop B*, for believing in the transformative power of Artificial Intelligence in retail, even before he could show them any proof.

References

- Samar Al-Saqqa and Arafat Awajan. 2019. The use of word2vec model in sentiment analysis: A survey. In *Proceedings of the 2019 International Conference on Artificial Intelligence, Robotics and Control*, pages 39–43.
- Federico Bianchi, Jacopo Tagliabue, Bingqing Yu, Luca Bigon, and Ciro Greco. 2020. *Fantastic embeddings and how to align them: Zero-shot inference in a multi-shop scenario*. In *Proceedings of the SIGIR 2020 eCom workshop, July 2020, Virtual Event, published at http://ceur-ws.org (to appear)*.
- Hugo Caselles-Dupré, Florian Lesaint, and Jimena Royo-Letelier. 2018. Word2vec applied to recommendation: hyperparameters matter. *Proceedings of the 12th ACM Conference on Recommender Systems*.
- Hugo Caselles-Dupré, Florian Lesaint, and Jimena Royo-Letelier. 2018. *Word2vec applied to recommendation: Hyperparameters matter*. In *Proceedings of RecSys ’18*.
- Mark Chen, A. Radford, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. 2020. Generative pretraining from pixels. In *ICML*.
- Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.
- Mihajlo Grbovic, Vladan Radosavljevic, Nemanja Djuric, Narayan Bhamidipati, Jaikit Savla, Varun Bhagwan, and Doug Sharp. 2015. [E-commerce in your inbox: Product recommendations at scale](#). In *Proceedings of KDD '15*.
- Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- J. Kaplan, Sam McCandlish, T. Henighan, T. Brown, Benjamin Chess, R. Child, Scott Gray, A. Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv*, abs/2001.08361.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.
- Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.
- Benjamin Letham, Cynthia Rudin, and David Madigan. 2013. Sequential event prediction. *Machine learning*, 93(2-3):357–380.
- Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. [Improving distributional similarity with lessons learned from word embeddings](#). *Transactions of the Association for Computational Linguistics*, 3:211–225.
- Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. *ArXiv*, abs/2011.05864.
- Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.
- Malte Ludewig, Noemi Mauro, Sara Latifi, and Dietmar Jannach. 2019. [Performance comparison of neural and non-neural approaches to session-based recommendation](#). In *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys '19*, page 462–466, New York, NY, USA. Association for Computing Machinery.
- Yifei Ma, Balakrishnan (Murali) Narayanaswamy, Haibin Lin, and Hao Ding. 2020. [Temporal-contextual recommendation in real-time](#). In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '20*, page 2291–2299, New York, NY, USA. Association for Computing Machinery.
- Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.
- Bhaskar Mitra and Nick Craswell. 2018. [An introduction to neural information retrieval](#). *Foundations and Trends® in Information Retrieval*, 13(1):1–126.
- Marius Mosbach, Maksym Andriushchenko, and D. Klakow. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *ArXiv*, abs/2006.04884.
- Patrick Ng. 2017. dna2vec: Consistent vector representations of variable-length k-mers. *ArXiv*, abs/1701.06279.
- Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [mask]? making sense of language-specific bert models. *arXiv preprint arXiv:2003.02912*.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. In *RepL4NLP@ACL*.
- Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.
- Ann Pichestapong. 2019. [Website personalization: Improving conversion with personalized shopping experiences](#).
- Nils Reimers and Iryna Gurevych. 2019. Sentencebert: Sentence embeddings using siamese bert-networks. In *EMNLP/IJCNLP*.
- Borja Requena, Giovanni Cassani, Jacopo Tagliabue, Ciro Greco, and Lucas Lacasa. 2020. [Shopper intent prediction from clickstream e-commerce data with minimal browsing information](#). *Scientific Reports*, 2020:16983.
- Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *arXiv preprint arXiv:2002.12327*.
- Statista Research Department. 2020. [Global retail e-commerce sales 2014-2023](#).
- Emma Strubell, Ananya Ganesh, and A. McCallum. 2019. Energy and policy considerations for deep learning in nlp. *ArXiv*, abs/1906.02243.

- Fei Sun, Jun Liu, J. Wu, Changhua Pei, X. Lin, Wenwu Ou, and P. Jiang. 2019. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*.
- Jacopo Tagliabue and Bingqing Yu. 2020. Shopping in the multiverse: A counterfactual approach to in-session attribution. *ArXiv*, abs/2007.10087.
- Jacopo Tagliabue, Bingqing Yu, and Marie Beaulieu. 2020a. How to grow a (product) tree. personalized category suggestions for ecommerce type-ahead. In *Companion Proceedings of ACL*, New York, NY, USA. Association for Computing Machinery.
- Jacopo Tagliabue, Bingqing Yu, and Federico Bianchi. 2020b. The embeddings that came in from the cold: Improving vectors for new and rare products with content-based inference. In *Fourteenth ACM Conference on Recommender Systems, RecSys '20*, page 577–578, New York, NY, USA. Association for Computing Machinery.
- Alon Talmor and Jonathan Berant. 2019. Multiqua: An empirical investigation of generalization and transfer in reading comprehension. *ArXiv*, abs/1905.13453.
- Techcrunch. 2019a. [Algolia finds \\$110m from accel and salesforce](#).
- Techcrunch. 2019b. [coveo-raises-227m-at-1b-valuation](#).
- Techcrunch. 2019c. [Lucidworks raises \\$100m to expand in ai finds](#).
- Arthur Toth, L. Tan, G. Fabbri, and Ankur Datta. 2017. Predicting shopping behavior with mixture of rnns. In *eCOM@SIGIR*.
- Manos Tsagkias, Tracy Holloway King, Surya Kallumadi, Vanessa Murdock, and Maarten de Rijke. 2020. Challenges and research opportunities in ecommerce search and recommendations. In *SIGIR Forum*, volume 54.
- U.S. Department of Commerce. 2020. [U.s. census bureau news](#).
- Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.
- Flavian Vasile, Elena Smirnova, and Alexis Conneau. 2016. Meta-prod2vec: Product embeddings using side-information for recommendation. *Proceedings of the 10th ACM Conference on Recommender Systems*.
- Flavian Vasile, Elena Smirnova, and Alexis Conneau. 2018. [Meta-prod2vec - product embeddings using side-information for recommendation](#). In *Proceedings of RecSys '16*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, L. Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.
- David Vilares, Michalina Strzyz, Anders Søgaard, and Carlos Gómez-Rodríguez. 2020. Parsing as pre-training. *ArXiv*, abs/2002.01685.
- Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. 2017. [From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts](#). In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL '17*, New York, NY, USA. Association for Computing Machinery.
- F. Yuan, X. He, Alexandros Karatzoglou, and Liguang Zhang. 2020. Parameter-efficient transfer from sequential behaviors for user modeling and recommendation. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Han Zhang, Songlin Wang, Kang Zhang, Zhi-Ling Tang, Y. Jiang, Y. Xiao, W. Yan, and Wenyun Yang. 2020. Towards personalized and semantic retrieval: An end-to-end solution for e-commerce search via embedding learning. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

A Visualization of Session Embeddings

Figures 2 to 5 represent browsing sessions projected in two-dimensions with t-SNE (van der Maaten and Hinton, 2008): for each browsing session, we retrieve the corresponding type (e.g. shoes, pants, etc.) of each product in the session, and use majority voting to assign the most frequent product

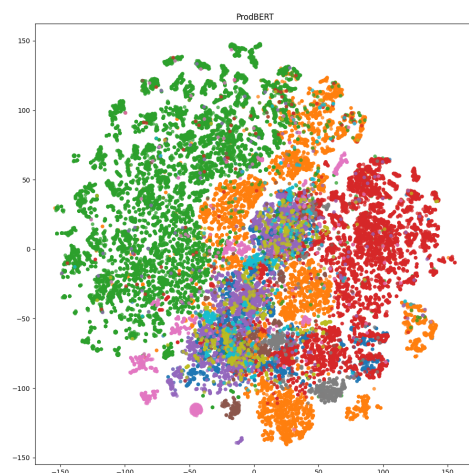


Figure 2: T-SNE plot of browsing session vector space from Shop A and built with the first hidden layer of pre-trained ProdBERT.

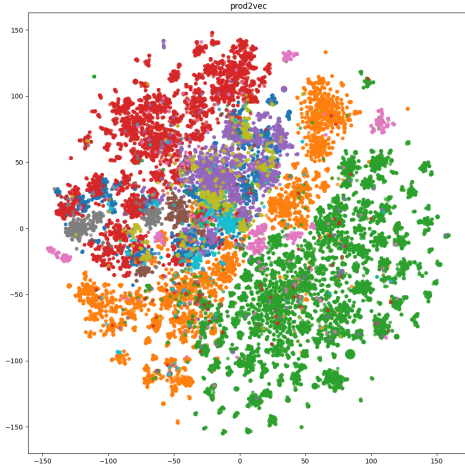


Figure 3: T-SNE plot of browsing session vector space from Shop A and built with *prod2vec* embeddings.

type to the session. Hence, the dots are color-coded by product type and each dot represents a unique session from our logs. It is easy to notice that, first, both contextual and non-contextual embeddings built with a smaller amount of data, i.e. Figures 2 and 3 from Shop A, have a less clear separation between clusters; moreover, the quality of ProdBERT seems even lower than *prod2vec*, as there exists a larger central area where all types are heavily overlapping. Second, comparing Figure 4 with Figure 5, both ProdBERT and *prod2vec* improve, which confirms ProdBERT, given enough pre-training data, is able to deliver better separations in terms of product types and more meaningful representations.

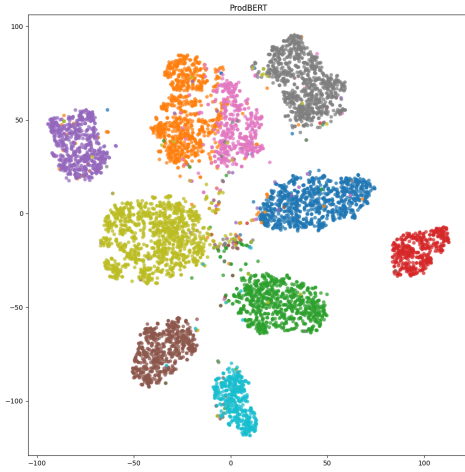


Figure 4: T-SNE plot of browsing session vector space from Shop B and built with the first hidden layer of pre-trained ProdBERT.

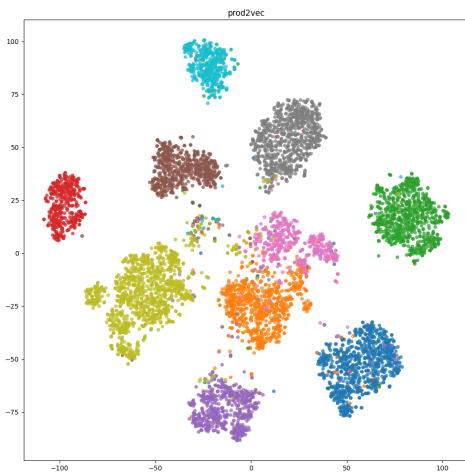


Figure 5: T-SNE plot of browsing session vector space from Shop B and built with *prod2vec* embeddings.