Category Recognition in E-Commerce using Sequence-to-Sequence Hierarchical Classification

Idan Hasson
eBay Research
ihasson@ebay.com

Slava Novgorodov
eBay Research
snovgorodov@ebay.com

Gilad Fuchs
eBay Research
gfuchs@ebay.com

Yoni Acriche
eBay Research
yacriche@ebay.com

ABSTRACT

E-commerce platforms often use a predefined structured hierarchy of product categories. Apart from helping buyers sort between different product types, listing categorization is also critical for multiple downstream tasks, including the platform's main listing search. Traditionally, when creating a new listing, sellers need to assign the product they sell to a single category. However, the high diversity of product types in the platform, along with the hierarchy's low level of granularity result in tens of thousands of different possible categories that sellers need to pick from. This, in turn, creates a unique classification challenge, especially for sellers with a large number of listings. Moreover, the expected cost of making a category classification error is high, as it can impact the likelihood that their listing will get discovered by relevant buyers, and eventually sold.

To help with the challenge of category recognition we present CatReComm - an interactive real-time system that is generalized to provide category recommendations in different e-commerce scenarios. We present results from using the system for two main sub-tasks - listing and search-query category recognition, and demonstrate an end-to-end scenario of the former one. The system uses a convolutional sequence-to-sequence approach, and to the best of our knowledge, is the first to use this approach for category recognition. We define a new metric for evaluating this model which captures the hierarchical characteristics of the data and supports displaying multiple classification results. Finally, our experimental results show the effectiveness and efficiency on real-world data.

ACM Reference Format:
Idan Hasson, Slava Novgorodov, Gilad Fuchs, and Yoni Acriche. 2021. Category Recognition in E-Commerce using Sequence-to-Sequence Hierarchical Classification. In *Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining (WSDM '21), March 8–12, 2021, Virtual Event, Israel.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3437963.3441717

1 INTRODUCTION

E-commerce platforms getting more popular in the recent years. Selling on these platforms is the main and sometimes only source of income for a significant number of people. The diversity in consumer needs, as well as the varied areas of sellers' businesses, have led to a growing number of products being offered on these platforms. Recent studies showed the importance of high quality product data (e.g., image, title, attributes and description [12]), and its effect on the listing's likelihood to sell. While many potential customers find their products by performing a keyword based search, others use category tree navigation, or a combination of the two. For example, a user that is interested in a cover for an iPhone X might write "iphone x flipcover" in the search box, or alternatively navigate to the "Smartphones Cases, Covers & Skins" category. Therefore, listing the product in the right category is critical for sellers, as it can have a substantial impact on the number of relevant buyers who view the listing, and eventually buy it.

Although its critical importance, it is not trivial for sellers to find the right category manually. The high diversity in product types that are sold in the platform along with the requirement that the hierarchy will have a relatively low granularity inflates the number of categories. For example, the eBay U.S. site alone has more than 40,000 different categories. While it is easy for sellers not to blunder between very distinctly different categories, the differences between others can be very subtle, like, for example, with the "Home Décor" and "Furniture" categories. Moreover, different e-commerce platforms maintain their own category hierarchies, which increases the complexity for multi-platform sellers. For example, cases for iPhone X can be under "Electronics" → "Smartphones" → "Smartphone Accessories" on one platform, but under "Mobile Phones" → "iPhones" → "iPhone Accessories" on another.

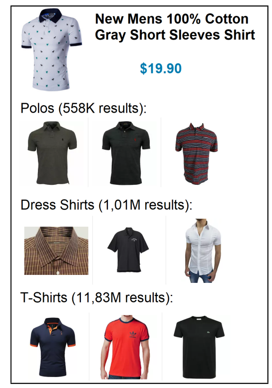

Figure 1: Creating a new shirt listing in the platform

Example 1.1. Consider an e-commerce platform and a seller that wants to sell a single men's shirt. To create the listing the seller adds data such as image, title, description and price. While getting to the "Shirts" category ("Fashion" → "Men" → "Men's Clothing" → "Shirts"), the seller realizes that there are many subcategories, e.g., Polos, Dress Shirts and T-Shirts. In this case, to determine which category the listing belongs to, the seller has to manually inspect

WSDM '21, March 8–12, 2021, Virtual Event, Israel

ACM ISBN 978-1-4503-8297-7/21/03.
https://doi.org/10.1145/3437963.3441717

each possible category. Then, for each category the seller needs to examine what type of shirts it contains, by inspecting the existing listings' titles, pictures and prices. Figure 1 depicts a screenshot of these categories, along with the shirt that the seller wishes to sell. This manual inspection can be time consuming even for a single listing, let alone at scale.

Although the category recognition problem is mostly focused on the listing level (as listing is the only data entity that formally gets assigned to a category), it can be further generalized. For example, by applying it to the platform's main listing search we can ask what is the most relevant product category per query. The definition for category relevancy in this context can be derived per query from the most frequent listing category in the search results page (SRP). Like with the case of listings, such identification can be highlighted in the online experience to help users to further narrow their search results, or alternatively, to be used as a feature in the search backend.

To help with these different types of challenges we introduce CatReComm - a highly generalized interactive real-time system for solving category recognition problems in an e-commerce environment. The framework capabilities include listing and query category recognition. The system provides initial top-k suggestions and allows sellers, in the case of listing category recognition, to interactively refine their listings' classification in the categories tree. The model behind CatReComm is trained on real-world data and is demonstrated in a real-life scenario of creating a new listing in an e-commerce platform.

CatReComm can also be helpful in supporting other Machine Learning models in the platform. This is due to the tendency in large platforms to split the train and inference of models to be per category (usually - high level categories with depth of one or two in the categories taxonomy). Therefore, the category classifier can be also utilized as a coarse classifier for these models, and route to the relevant inference model.

2 TECHNICAL BACKGROUND

Sequence-to-Sequence (seq2seq) models have been widely used for different tasks, e.g. machine translation [1, 15], video captioning [17] and text summarization [11]. These models take one sequence as an input, and generate a new sequence as an output, based on encoder-decoder architecture. A seq2seq model is usually based on recurrent neural networks [15], convolutional neural networks [5] or transformer architectures [16].

Seq2seq models additionally have been used for text classification. In [19], a seq2seq model is utilized for multi-label text classification, in which the labels often exhibit complex dependencies. Similar dependencies also exist in hierarchical classification. When working with taxonomic data, each level (rank) in the hierarchy usually depends on all the higher levels.

There are several common approaches for hierarchical classification. In flat classification, the model considers only the leaf categories, and is trained to predict only them. This approach ignores the hierarchical structure of the data, hence does not utilize the existing similarities between categories that share part of the hierarchical path. In contrast, in the hierarchically-structured local classifiers approach, a different model is trained for each level in the taxonomy. In hierarchies with a large number of levels this approach can inflate the number of models, and increase the complexity of the solution. As an intermediate approach, the global classifiers method learns a single model which considers the entire class hierarchy as a whole [8].

The ability of the seq2seq model to learn the dependencies between the tokens in the target sequence, along with the fact that it is a single model, makes the seq2seq model a good candidate for a global classifier to the hierarchical classification task. More formally, given an input sequence $x_1, x_2, ..., x_n$, and the hierarchical labels path of this sample: $l_1, l_2, ..., l_k$, we can train the seq2seq model with the following input-output couples:
$x_1\ x_2\ ...\ x_n \rightarrow l_1\ l_2\ ...\ l_k$.

Our task is based on eBay categories taxonomy. Every listing in eBay is labeled with at least one category. Every category has two main components: (a) the vertical - one of eight values in U.S. site, e.g. Electronics, Fashion, Collectibles, etc. and (b) up to 7 levels of hierarchical categories. In this demo we demonstrate the ability of a seq2seq model to predict the correct category given the listing title, in the case of listing category recognition, or the user's search query, for query category recognition.

2.1 System Architecture and Performance

For training the seq2seq model we used the Fairseq library [13]. This library implements varied NLP algorithms, including common machine translation seq2seq architectures. Specifically for this work, we chose to use the convolutional seq2seq flavour, due to its superiority in prediction time and performance. The implementation of this model is based on [5].

This convolutional seq2seq architecture consists of a fully convolutional model, i.e. a convolutional encoder and a convolutional decoder. The general components in the encoder and decoder include: embedding layer (including positional embedding), linear layer to project to the size of the convolutional layers, GLU convolutional blocks with residual connections, and a linear layer to project back to the size of the embedding. The decoder includes attention layers after each GLU convolutional block. Dropout is used before each linear and convolutional layer, except of the linear layer after the convolutional component. Specifically in our work, we used the parameters of the "fconv_iwslt_de_en" model architecture, including 256 embedding dimensions, 4 encoder convolutional layers and 3 decoder convolutional layers with kernel width of 3 and 256 output channels (as depicted in Figure 2).

We trained the model with a dropout of 0.2, learning rate of 0.25, max number of epochs of 15, gradient clipping of 0.1 and a maximum number of tokens in a batch of 1500. We used a beam size of 5, to get 5 category suggestions. The optimizer for the training process was the Nesterov Accelerated Gradient (NAG) optimizer. Training was conducted on a single GPU.

To train the listing category recognition model we used 8.5 million of historical listings as a train set. For the query category recognition model the train set was comprised of 2.5 million of historical queries. Unlike with listings, historical search queries do not have a label for what was the right category. Therefore, we tagged each query using the distribution of its SRP listings categories. Since this distribution often includes more than a single category, and in

Table 1: The distribution of the top-level category in the test data.

Top Level Category	Listing Category Recognition Count	Listing Category Recognition %	Query Category Recognition Count	Query Category Recognition %
Electronics	4758	5.79	3691	14.69
Fashion	24950	30.35	6355	25.29
Home & Garden	7864	9.57	3387	13.48
Collectibles	27826	33.84	6566	26.13
Lifestyle	3768	4.58	2249	8.95
Business & Industrial	2111	2.57	1289	5.13
Media	10938	13.31	1587	6.32
Total	82215		25124	

Table 2: The mean TS@r,k results of our category recognition models.

rank (r)	Listing Category Recognition Suggestions (k)					Query Category Recognition Suggestions (k)				
	1	2	3	4	5	1	2	3	4	5
1	93.22	95.95	96.92	97.43	97.81	77.69	83.08	85.17	86.60	87.83
2	91.46	94.67	95.85	96.50	96.98	74.56	80.18	82.46	84.23	85.54
3	87.36	91.91	93.75	94.78	95.54	70.49	76.88	79.43	81.32	82.72
4	84.69	90.01	92.14	93.32	94.19	67.33	74.40	77.15	79.13	80.58
5	82.85	88.63	90.93	92.21	93.11	65.30	72.74	75.65	77.66	79.11
6	82.14	88.06	90.40	91.69	92.58	64.67	72.19	75.13	77.14	78.56
7	82.04	87.95	90.29	91.57	92.46	64.61	72.13	75.07	77.07	78.48
8	82.03	87.94	90.28	91.56	92.45	64.60	72.12	75.06	77.06	78.47

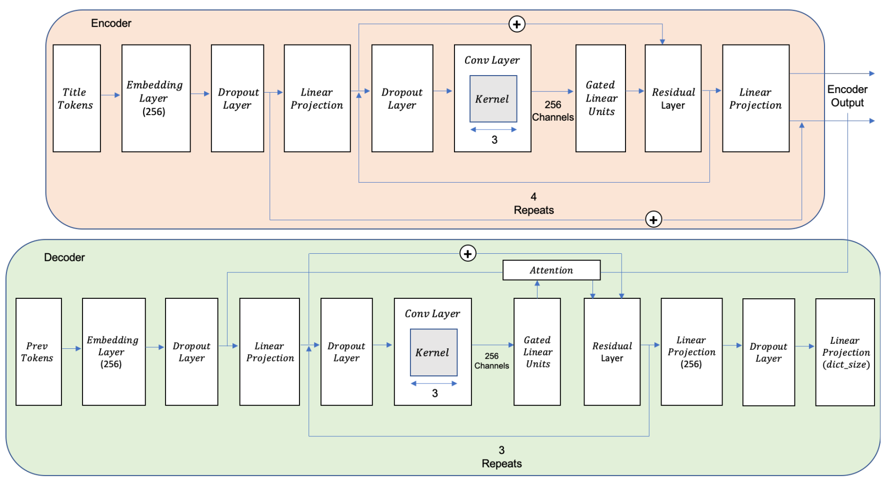

Figure 2: The "fconv_iwslt_de_en" convolutional seq2seq model implemented in the Fairseq library.

order to preserve relations between related queries (e.g. "iphone accessories" and "iphone case"), our training consisted of sample weightening. For example, if a query returned 20 results from the category "Fashion" → "Men" → "Men's Clothing" → "Shirts" → "Polos" and 30 results from the category "Fashion" → "Men" → "Men's Clothing" → "Shirts" → "T-shirts", we labeled this query (only for the training phase) as "Polos" with a probability of 0.4 and as "T-shirts" with a probability of 0.6. In total, we had 11.8 million pairs of (query, category) in the train data.

In order to evaluate the model we define Taxonomy Similarity (TS), a metric that measures the similarity between any two taxonomic labels based on the taxonomy distance metric described in [3]. Given two labels: a, b with the hierarchical paths: $a_1, a_2, ..., a_m$ and $b_1, b_2, ..., b_n$, TS is defined as:

$$TS = \frac{\sum_{l=1}^{min(m,n)} 1[a_l == b_l]}{max(m,n)} \tag{1}$$

Based on TS, we define the novel metric TS@r as the TS when considering $ranks \leq r$ only. This metric is useful to estimate the ability of our model to serve as a first phase routing model. Additionally, we define TS@r,k as the maximal TS@r given the top-k suggestions by the model. This metric fits our need to reduce the number of possible categories from tens of thousands to only a few, while taking into account that there might be more than one possible classification. In this definition, we assume that there exists one ground truth label which we compare to our classifications. For listing category recognition the label would be the main category of the listing. For query category recognition this would be the category with the highest probability in the SRP category distribution.

Our test data consists of 82,215 listings for the listing category recognition model, and 25,124 queries for the query category recognition task. Table 1 depicts the distribution of the top-level category in the test data, and table 2 shows the results of mean TS@r,k over the test data, for $r = 1, .., 8$ and $k = 1, .., 5$.

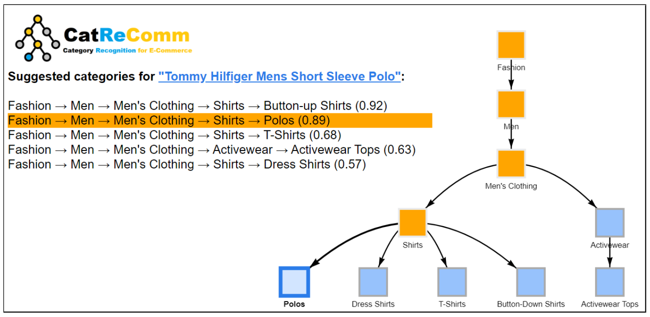

Figure 3: Top-5 Suggestions

Figure 4: Sample listings

3 DEMONSTRATION SCENARIO

We demonstrate the operation of CatReComm - an interactive real-time system that guides the sellers for the best category for their listings. We reenact a real-life scenario where a seller, played by the audience, creates a listing for sale and provides information such as title, image, description and price of her listing. Then, as depicted in Figure 3, the system suggests the top-5 most recommended categories, sorted by the recommendation confidence score (computed by the model described in Section 2.1). The category with the highest confidence score is selected by default and our UI visualizes it using a categories tree. In the presented tree the path to the target category is highlighted, and the seller can see the parent and the sibling categories. By clicking on one of the other suggestions the system highlights the path in the tree from the root to the newly selected category. Moreover, the seller can decide to dismiss the suggestions and decide on any other arbitrary category. The UI is equipped with drag-and-drop and zoom-in/zoom-out capabilities, enabling to explore different categories and to deal with large category trees. Moreover, by default only the nodes which are participating in the suggested categories are drawn. This makes the visualization much more compact, compared to the full tree.

To help the seller make the decision the system also allows to see sample listings from the suggested categories by clicking on the corresponding node in the graph, as shown in Figure 4. This provides sellers who are not sure about two specific suggested categories a way to compare the newly created listing to the listings in the existing catalog. For more detailed analysis the seller may also click on the image to see the corresponding listing page.

In addition to the interactive tool with UI, CatReComm also supports the service mode where it provides the suggestion via API, by getting the title and k (desired number of suggestions) and returning the top-k suggested categories and their corresponding confidence score. Furthermore, CatReComm provides prediction API for queries. This service mode allows to use our models as part of different flows in the platform and unify the suggestion mechanism across the e-commerce platform.

Related Work Category Recognition of listings in e-commerce sites is usually based on different data components, e.g.: titles [14], descriptions [2] and reviews [7]. Many related algorithms for this task take advantage of traditional machine learning techniques, for instance: SVM [4, 7], LDA [9], KNN (usually as a coarse level classifier [7], or together with another model [2]). In the last couple of years, deep learning algorithms have been exploited to solve this task as well. In [18], Attention CNN (ACNN) was used in order to recognize the top-level category (rank = 1) of a listing between 35 possible labels. [6] used multiple RNNs with fully connected layers to classify between 4100 different categories. [10] examined a GRU sequence-to-sequence architecture for listing category recognition.

To the best of our knowledge, our system is the first to use the convolutional sequence-to-sequence approach. We define the novel TS@r,k metric for evaluating the model, fitting the hierarchical characteristics of the data in e-commerce platforms, and show the results for both listings and user-queries.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv* (2014).

[2] Ali Cevahir and Koji Murakami. 2016. Large-scale Multi-class and Hierarchical Product Categorization for an E-commerce Giant. In *Proc. of COLING*. 525–535.

[3] Chung-Yen Chen, Sen-Lin Tang, and Seng-Cho T. Chou. 2019. Taxonomy based performance metrics for evaluating taxonomic assignment methods. In *BMC Bioinformatics*.

[4] Jianfu Chen and David Warren. 2013. Cost-Sensitive Learning for Large-Scale Hierarchical Classification. In *Proc. of CIKM*. 1351–1360.

[5] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional Sequence to Sequence Learning. *CoRR* abs/1705.03122 (2017).

[6] Jung-Woo Ha, Hyuna Pyo, and Jeonghee Kim. 2016. Large-Scale Item Categorization in e-Commerce Using Multiple Recurrent Neural Networks. In *Proc. of the KDD*. 107–115.

[7] S. Huang, X. Liu, X. Peng, and Z. Niu. 2012. Fine-grained Product Features Extraction and Categorization in Reviews Opinion Mining. In *ICDM*. 680–686.

[8] Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. 2013. Evaluation Measures for Hierarchical Classification: a unified view and novel approaches. *Data Mining and Knowledge Discovery* 29 (06 2013).

[9] Zornitsa Kozareva. 2015. Everyone Likes Shopping! Multi-class Product Categorization for e-Commerce. In *Proc. of NAACL*. 1329–1333.

[10] Yundi Maggie Li, Liling Tan, Stanley Kok, and Ewa Szymanska. 2018. Unconstrained Product Categorization with Sequence-to-Sequence Models. In *SIGIR 2018 Workshop on eCommerce (ECOM 18)*.

[11] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar GulÌ‡lçehre, and Bing Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proc. of SIGNLL*. 280–290.

[12] Slava Novgorodov, Ido Guy, Guy Elad, and Kira Radinsky. 2019. Generating Product Descriptions from User Reviews. In *Proc. of WWW*.

[13] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. *CoRR* abs/1904.01038 (2019).

[14] Dan Shen, Jean-David Ruvini, and Badrul Sarwar. 2012. Large-scale item categorization for e-commerce. In *Proc. of CIKM*. 595–604.

[15] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. *CoRR* abs/1409.3215 (2014).

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR* abs/1706.03762 (2017).

[17] Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to Sequence - Video to Text. *CoRR* abs/1505.00487 (2015).

[18] Yandi Xia, Aaron Levine, Pradipto Das, Giuseppe Di Fabbrizio, Keiji Shinzato, and Ankur Datta. 2017. Large-Scale Categorization of Japanese Product Titles Using Neural Attention Models. In *Proc. of EACL*. 663–668.

[19] Z. Yang and G. Liu. 2019. Hierarchical Sequence-to-Sequence Model for Multi-Label Text Classification. *IEEE Access* 7 (2019), 153012–153020.