Relevance under the Iceberg: Reasonable Prediction for Extreme Multi-label Classification

Jyun-Yu Jiang[†], Wei-Cheng Chang[†], Jiong Zhang[†], Cho-Jui Hsieh[‡] and Hsiang-Fu Yu[†]
[†]Amazon Search, Palo Alto, CA, USA
[‡]University of California, Los Angeles, CA, USA
{jyunyu,chanweic,jiongz,hsiangfu}@amazon.com,chohsieh@cs.ucla.edu

ABSTRACT

In the era of big data, eXtreme Multi-label Classification (XMC) has already become one of the most essential research tasks to deal with enormous label spaces in machine learning applications. Instead of assessing every individual label, most XMC methods rely on label trees or filters to derive short ranked label lists as prediction, thereby reducing computational overhead. Specifically, existing studies obtain ranked label lists with a fixed length for prediction and evaluation. However, these predictions are unreasonable since data points have varied numbers of relevant labels. The greatly small and large list lengths in evaluation, such as Precision@5 and Recall@100, can also lead to the ignorance of other relevant labels or the tolerance of many irrelevant labels. In this paper, we aim to provide reasonable prediction for extreme multi-label classification with dynamic numbers of predicted labels. In particular, we propose a novel framework, Model-Agnostic List Truncation with Ordinal Regression (MALTOR), to leverage the ranking properties and truncate long ranked label lists for better accuracy. Extensive experiments conducted on six large-scale real-world benchmark datasets demonstrate that MALTOR significantly outperforms statistical baseline methods and conventional ranked list truncation methods in ad-hoc retrieval with both linear and deep XMC models. The results of an ablation study also shows the effectiveness of each individual component in our proposed MALTOR.

KEYWORDS

extreme multi-label classification, ranked list truncation, ordinal regression, cost-sensitive learning.

ACM Reference Format:
Jyun-Yu Jiang[†], Wei-Cheng Chang[†], Jiong Zhang[†], Cho-Jui Hsieh[‡] and Hsiang-Fu Yu[†]. 2022. Relevance under the Iceberg: Reasonable Prediction for Extreme Multi-label Classification. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3477495.3531767

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531767

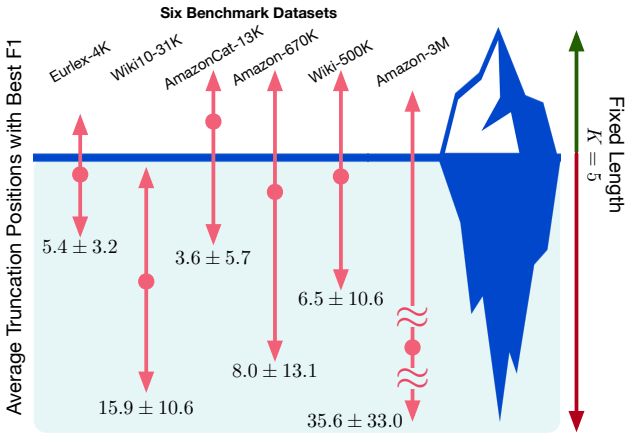


Figure 1: Illustration of average truncation positions with best F1 scores in the testing sets of six datasets based prediction derived by well-trained XR-LINEAR models [27].

1 INTRODUCTION

Extreme multi-label classification (XMC) is a machine learning task to obtain relevant labels from an enormous space for a data input. With more and more large-scale machine learning applications coming into our sight, XMC has already become one of the most indispensable research problems, and has benefits various disciplines, such as e-commerce [7], search engine [6], biology [23], medicine [5], and public health [29]. With more than a decade of research progress, state-of-the art XMC approaches can not only learn satisfactory XMC models with one-versus-rest methods [2, 3, 25, 27] and deep learning [26, 28], but also efficiently derive prediction for a data input based on hierarchical label trees [13, 19, 27] and label filters [16, 18]. However, using beam search [27] and label filtering [18], most of the existing XMC methods only support to provide a truncated ranked list as relevant labels with a fixed (and usually short) length, thereby potentially ignoring many relevant labels and resulting unreasonable prediction.

Figure 1 depicts the illustration of average truncation positions with best F1 scores, which simultaneously consider precision and recall of relevant labels, for prediction of well-trained XMC models. The small fixed length $K = 5$ represents a comparative cutoff that is widely used in evaluation of XMC [27] and recommender systems [12] to ensure high precision. Obviously, some datasets, such as Wiki10-31K and Amazon-3M, fairly favor to keep more labels while others have different preferences. This not only shows how evaluation and inference processes of previous studies using a

fixed small cutoff is unreasonable, but also demonstrates the need of choosing different truncation positions or datasets. Moreover, the other observation is that the variance of best truncation positions is also huge, even for data in the same dataset. Any shared and fixed truncation position could fail to serve all instances in the dataset. In other words, the length of the truncated ranked list for every instance should be dynamically decided to obtain satisfactory prediction and secure those relevant labels under the iceberg.

In this paper, the framework, Model-Agnostic List Truncation with Ordinal Regression (MALTOR), is proposed to address the above challenges for arbitrary XMC models and their prediction. Given a trained XMC model and a testing data input, we first utilize the encoder in the model to derive the data representation. For candidate positions, we consider the predictive scores inferred by the model as position features. After concatenating the data representation and position features as truncation features, we treat the task of list truncation as an ordinal regression problem. Finally, we conduct experiments to demonstrate the effectiveness of MALTOR for both linear and deep XMC models. We also demonstrate the capability of each individual component and feature set in an ablation study.

In literature, to the best of our knowledge, none of the existing XMC studies focus on dynamic prediction lengths during inference. In other fields, some studies [1, 24] attempt to determine the number of predicted classes in traditional multi-label classification, but they cannot be easily applied to XMC. Ranked list truncation in ad-hoc retrieval can be considered a line of related work. Lien et al. [15] propose a deep learning framework, BiCut, to predict the best truncation position based on recurrent neural networks. Bahri et al. [4] and Wu et al. [22] further improve the approach by leveraging the Transformers and attention mechanism [20]. However, these methods cannot be directly applied into XMC since most XMC problems have no label features. Although positive instance feature aggregation (PIFA) [27] can be a feasible solution to derive label features, these features can be incapable of precisely representing the labels. Moreover, these methods treat the task as a multi-class classification problem to identify the best cutoff without considering the ordinal relations among positions. In our experiments, we treat ranked list truncation methods in ad-hod retrieval as comparative baseline methods.

2 REASONABLE PREDICTION FOR XMC

In this section, we first formally define the task of reasonable prediction for XMC as ranked list truncation in this paper, and then introduce our proposed framework, Model-Agnostic List Truncation with Ordinal Regression (MALTOR), as shown in Figure 2. The data input $\boldsymbol{x}$ is first extracted as the data representation $\boldsymbol{r_x}$. The truncation features $\boldsymbol{u_x}$ are then derived by concatenating the data representation $\boldsymbol{r_x}$ and K model predictive scores over the top-K positions. Based on the truncation features $\boldsymbol{u_x}$, MALTOR applies the ordinal regression model to determine the truncation position k for reasonable prediction.

2.1 Problem Statement

Suppose we have a well-trained XMC model $M(\boldsymbol{x})$ and a testing data $\boldsymbol{x}$. We aim to deliver reasonable prediction by deciding a best

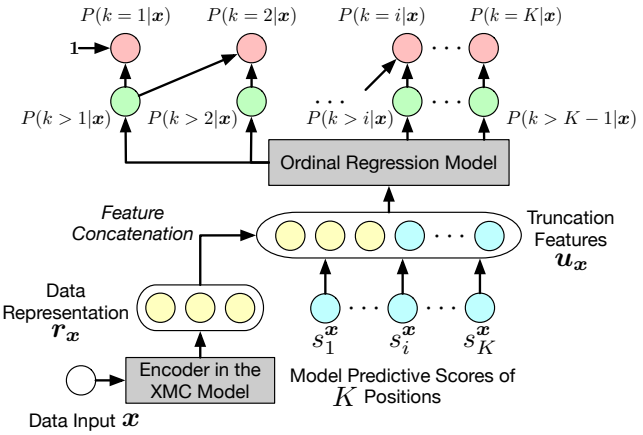


Figure 2: Illustration of our proposed MALTOR framework.

truncation position $k \in \mathcal{K} = \{1, \ldots, K\}$ from long enough K candidate positions so that the target metric based on the remaining top-k labels can be as satisfactory as possible. Note that target metrics can be arbitrary measurements based on different applications. Specifically, in this work, we follow related work in ad-hoc retrieval [4] to focus on F1 scores at k (F1@k) and discounted cumulative gain at k (DCG@k) as:

$$\text{F1@}k = \frac{2 \times \text{Precision@}k \times \text{Recall@}k}{\text{Precision@}k + \text{Recall@}k},$$

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)},$$

where Precision@k and Recall@k state the percentage of relevant labels in the top-k labels and their coverage over all relevant labels; $\text{rel}_i = 1$ if the i-th label is relevant; otherwise, $\text{rel}_i = -0.5$ plays as a penalty for each irrelevant label.

2.2 Truncation Features

MALTOR extracts useful features for list truncation from two perspectives, including *data* and *prediction*.

Data Representation. For the data perspective, we leverage the encoder in the XMC model M, such as tfidf vectorizers and BERT in linear and deep XMC models, to derive data representation $\boldsymbol{r_x}$

Model Predictive Scores. For the prediction perspective, we utilize the model predictive score $s_i^{\boldsymbol{x}} = M(\boldsymbol{x})_i$ of each position i as a feature, where $M(\boldsymbol{x})_i$ is the predictive score for the i-th label; scores $s_1^{\boldsymbol{x}} \geq s_2^{\boldsymbol{x}} \geq \cdots \geq s_K^{\boldsymbol{x}}$ are also descending as ranking results.

Finally, the ultimate truncation features $\boldsymbol{u_x}$ can be obtained by concatenating the data representation and model predictive scores of all position candidates as $\boldsymbol{u_x} = \{\boldsymbol{r_x}; s_1^{\boldsymbol{x}}; \ldots; s_K^{\boldsymbol{x}}\}$.

2.3 List Truncation via Ordinal Regression

Different from conventional list truncation methods that optimize multi-class classification, MALTOR leverages the ordinal relations among candidate positions by treating the task as an ordinal regression problem [21]. Note that numerical forms of positions could be large integers from 1 to K and make regression-based

models[17] unstable. Instead, we estimate the cumulative probability $\Pr(k > i \mid \boldsymbol{x})$ [11] with model weights $\boldsymbol{w}_i$ and θ_i to for each position i as:

$$\Pr(k > i \mid \boldsymbol{x}) = \sum_{j=i+1}^{K} \Pr(k = j \mid \boldsymbol{x}) = \frac{1}{1 + \exp(\boldsymbol{w}_i{}^\top \boldsymbol{u}_{\boldsymbol{x}} - \theta_i)}, \quad (1)$$

where $1 \le i \le K-1$; $\theta_1 \le \cdots \le \theta_{K-1}$ are ascending to justify the cumulative property. Without losing ordinal information of positions, the marginal probability $\Pr(k = i \mid \boldsymbol{x})$ for every position i can be computed as:

$$\Pr(k = i \mid \boldsymbol{x}) = \Pr(k > i - 1 \mid \boldsymbol{x}) - \Pr(k > i \mid \boldsymbol{x}),$$

where $\Pr(k = 1) = 1 - \Pr(k > 1)$ and $\Pr(k = K) = \Pr(k > K)$ are two special cases. Finally, MALTOR derives the predicted truncation position $\hat{k}$ can be computed as

$$\hat{k} = \arg\max_i \Pr(k = i).$$

2.4 Mislabeling Costs and Optimization

To optimize model weights, we follow previous studies [8, 14] to consider a mislabeling cost matrix $C \in \mathbb{R}^{K \times K}$, where $C_{k,i}$ is the cost of predicting the truncation position k as i. Intuitively, farther mispredictions deserve higher costs, so $C_{k,j}$ would satisfy a V-shape, such as absolute ($|k-i|$), and squared ($|k-i|^2$) costs. Moreover, the ordinal regression problem with V-shaped costs can be reduced to a binary classification problem about the cumulative probabilities described in Eq (1). Specifically, $\boldsymbol{w}_i{}^\top \boldsymbol{u}_x - \theta_i$ can be reduced into $\boldsymbol{o}_i{}^\top \hat{\boldsymbol{w}}^\top \hat{\boldsymbol{u}}_x - \theta_i$, where $\hat{\boldsymbol{w}}$ is a block weight sparse mautrix combining $\{\boldsymbol{w}_j \mid 1 \le j \le K\}$ in a diagonal manner; $\hat{\boldsymbol{u}}_x$ tiles the features $\boldsymbol{u}_x$ K times for computations; $\boldsymbol{o}_i$ is an one-hot vector with the i-th bit.

Theorem 2.1. *(Li and Lin [14]) If the costs $C_{k,i}$ are V-shaped, the ordinal regression task is equivalent to a binary classification task that learns $\Pr(k > i \mid \boldsymbol{x})$ with adjusted costs $c_{k,i} = |C_{k,i} - C_{k,i+1}|$. Learned θ_i will be also ordered if the loss is non-increasing to θ_i.*

Based on Theorem 2.1, we solve a linear L2-SVM [10] with L2-regularization with the overall loss $\mathcal{L}$ as:

$$\mathcal{L} = \sum_{\{\boldsymbol{x}, \boldsymbol{u}_{\boldsymbol{x}}, k\} \in \mathcal{D}} \sum_{i=1}^{K-1} \left[c_{k,i} \max(0, 1 - y_{\boldsymbol{x},i} \cdot (\boldsymbol{w}_i{}^\top \boldsymbol{u}_{\boldsymbol{x}} - \theta_i))^2 + \frac{\lambda}{2} ||\boldsymbol{w}_i||^2 \right],$$

where $\mathcal{D}$ is a training dataset for list truncation; $y_{\boldsymbol{x},i} \in \{+1, -1\}$ indicates if $k > i$ for $\boldsymbol{x}$; λ is a hyper-parameter.

2.5 Theoretical Analysis

The theoretical generalization bound of MALTOR can be also shown in Theorem 2.2, and demonstrate our reduction can well generalize.

Theorem 2.2. *Let $c_k = C_{k,1} + C_{k,K}$; $(\boldsymbol{x}, k)$ are drawn i.i.d. from a distribution $P(\boldsymbol{x}, k)$ on $\mathcal{X} \times \mathcal{K}$. If $C_{k,i}$ are V-shaped, there exists a distribution $\hat{P}$ on $(\boldsymbol{X}, Y)$ such that MALTOR's generalization error is*

$$\mathop{\mathbb{E}}_{(\boldsymbol{x},k)\sim P} C_{k,\hat{k}} \le c \cdot \mathop{\mathbb{E}}_{(\boldsymbol{X},Y)\sim \hat{P}} [Y f(\boldsymbol{X})],$$

where $\boldsymbol{X}$ encodes $(\boldsymbol{x}, i)$; $f(\boldsymbol{x}, i) = \boldsymbol{w}_i{}^\top \boldsymbol{u}_{\boldsymbol{x}} - \theta_i$; $c = \max_k [C_{k,1} + C_{k,K}]$; Y is the binary label in the loss function $\mathcal{L}$.

Proof. Since $C_{k,i}$ are V-shaped and $c_k = \sum_{i=1}^{K-1} c_{k,i}$, we have

$$C_{k,\hat{k}} \le \sum_{i=1}^{K-1} c_{k,j} \mathbb{1}[y_{\boldsymbol{x},i} f(\boldsymbol{x}, i) \le 0] = c_k \cdot \mathop{\mathbb{E}}_{i\sim P_i} \mathbb{1}[y_{\boldsymbol{x},i} f(\boldsymbol{x}, i)].$$

$\hat{P}(\boldsymbol{x}, y_{\boldsymbol{x},i})$ then can be constructed by drawing $(\boldsymbol{x}, k, i)$ from P and $P_i(i \mid k)$ so that the generalization error of MALTOR is

$$\begin{aligned} \mathop{\mathbb{E}}_{(\boldsymbol{x},k)\sim P} C_{k,\hat{k}} &\le \mathop{\mathbb{E}}_{(\boldsymbol{x},k)\sim P} c_k \cdot \mathop{\mathbb{E}}_{i\sim P_k} \mathbb{1}[y_{\boldsymbol{x},i} f(\boldsymbol{x}, i) \le 0] \\ &\le c \cdot \mathop{\mathbb{E}}_{(\boldsymbol{x}, y_{\boldsymbol{x},i})\sim \hat{P}} \mathbb{1}[y_{\boldsymbol{x},i} f(\boldsymbol{x}, i) \le 0]. \end{aligned}$$

□

3 EXPERIMENTS

3.1 Experimental Settings

Datasets. In this paper, we adopt six public benchmark extreme multi-label text classification datasets [26], including Eurlex-4k, Wiki10-31K, Amazon-670K, AmazonCat-13K, Wiki-500K, and Amazon-3M. The training datasets in the original partitions are utilized to train XMC models. We randomly split the testing datasets in the original partitions into 80% and 20% instances to train and test MALTOR and baselines. Table 1 shows the detailed statistics of datasets.

XMC Models. To verify MALTOR can be applied to arbitrary XMC models, we adopt XR-LINEAR [27] and XR-TRANSFORMER [28] as the state-of-the-art linear and deep XMC models. We follow the original papers to obtain well-trained XMC models with satisfactory performance in conventional evaluation. The features of XR-LINEAR are tfidf vectors while XR-TRANSFORMER utilizes BERT [9] as a text encoder to derive representations.

Baselines. We compare MALTOR against four statistical baseline methods and two neural ranked list truncation approaches in ad-hoc retrieval. Fixed-1 and Fixed-10 use 1 and 10 as a shared and fixed cutoff. Greedy-K adopts a greedy strategy on training data to determine a fixed truncation position. Similarly, Greedy-Threshold makes the decision based on a fixed threshold of model predictive scores. For neural ranked list truncation methods, BiCut [15] and Choppy [4] represent deep learning approaches using recurrent neural networks and Transformers [20]. For these two sequential neural methods, we apply the PIFA method [27] to aggregate features of positive instances as label features.

Implementation Details. The number of the candidate positions K is 100. For all methods, including MALTOR and baselines, we fine-tune all hyper-parameters to obtain best performance and have fair comparisons. Specifically, for MALTOR, we conduct a grid search on $\lambda \in \{2^{-5}, \ldots, 2^5\}$ with a 5-fold cross-validation to determine the best λ and train MALTOR with the entire training set. We choose the absolute costs $C_{k,j} = |k - j|$ as the mislabeling costs in optimization.

3.2 Experimental Results

Overall Performance. Table 2 shows the performance of different methods on six dastasets with two different XMC models. Interestingly, Greedy-K and Greedy-Threshold are very competitive

Dataset	Eurlex-4K	Wiki10-31K	Amazon-670K	AmazonCat-13K	Wiki-500K	Amazon-3M
n_{train}	3,092	5,292	122,420	245,425	615,536	594,005
n_{test}	883	1,324	30,605	61,357	153,885	148,502
d	186,104	101,938	135,909	203,882	2,381,304	337,067

Table 1: Statistics of six experimental datasets. n_{train} and n_{test} are the numbers of training and testing datasets. d is the dimension of tfidf feature vectors.

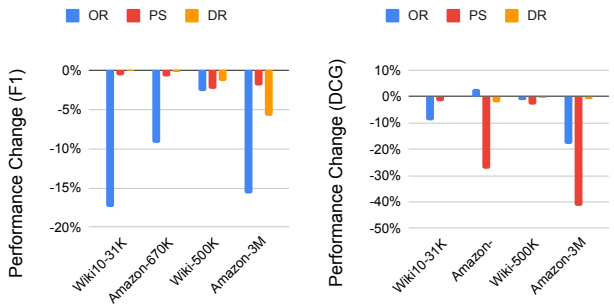
	Eurlex-4K		Wiki10-31K		AmazonCat-13K		Amazon-670K		Wiki-500K		Amazon-3M	
Method	F1	DCG	F1	DCG	F1	DCG	F1	DCG	F1	DCG	F1	DCG
XR-LINEAR												
Oracle	0.6909	1.7911	0.412	2.1072	0.8339	2.155	0.4339	0.8322	0.5599	1.0968	0.3409	2.4347
Fixed-1	0.2711	0.7419	0.0910	0.7440	0.3810	0.8883	0.1442	0.1606	0.2934	0.4978	0.0471	0.1914
Fixed-10	0.4900	1.0095	0.3301	1.5137	0.4985	1.2118	0.2997	-0.3217	0.3026	-0.0708	0.1771	0.4105
Greedy-K	0.5740	1.3953	0.3386	1.5691	0.6383	1.6500	0.3239	0.2220	0.3540	0.5945	0.2288	0.2655
Greedy-Threshold	0.5330	1.3555	0.3134	1.5406	**0.6842**	**1.7369**	0.2882	0.5107	0.3944	0.7652	0.2169	**1.3234**
BiCut [15]	0.4883	1.1859	0.2439	1.295	0.5956	1.4957	0.2790	0.1538	0.3749	0.6077	0.1220	0.4835
Choppy [4]	0.5166	1.2889	0.3386	1.2673	0.6501	1.6500	**0.3437**	0.1337	0.3666	0.4521	0.2368	0.4914
MALTOR	**0.5764**	**1.4508**	**0.3413**	**1.6062**	0.6638	1.7240	0.3368	**0.4920**	**0.4213**	**0.7814**	**0.2435**	1.2510
XR-TRANSFORMER												
Oracle	0.7107	1.8946	0.4250	2.2764	0.8636	2.2946	0.4333	0.8920	0.6061	1.3206	0.2975	1.9501
Fixed-1	0.2827	0.7885	0.0945	0.7825	0.3961	0.9350	0.1533	0.2023	0.3382	0.6362	0.0558	0.2674
Fixed-10	0.4961	1.0952	0.3547	1.7376	0.5149	1.3511	0.2976	-0.2892	0.3186	0.1326	0.2011	0.7059
Greedy-K	0.5860	1.5032	0.3590	1.7585	0.6834	1.7921	0.3212	0.2649	0.4288	0.8173	0.2073	0.6904
Greedy-Threshold	0.5473	1.4598	0.3299	1.6934	0.6863	1.7957	0.2872	0.5592	0.4482	0.9760	0.1826	1.2303
BiCut [15]	0.5088	1.3050	0.2530	1.4057	0.6202	1.6183	0.2883	0.2191	0.4219	0.8212	0.1416	0.5328
Choppy [4]	0.5441	1.4073	0.3539	1.3098	0.6834	1.6845	0.3088	0.2023	0.4382	0.8362	0.2061	0.5633
MALTOR	**0.6038**	**1.5687**	**0.3647**	**1.7930**	**0.7035**	**1.8904**	**0.3368**	**0.4920**	**0.4757**	**1.0109**	**0.2367**	**1.2573**

Table 2: Performance of different methods on six datasets with XR-LINEAR and XR-TRANSFORMER XMC models. Oracle indicates the upperbound performance with the ground truth truncation positions.

baseline methods, although they are statistical approaches using fixed cutoffs.

Neural ranked list truncation approaches (i.e., BiCut and Choppy) have unstable performance across different datasets and XMC models. This can be because PIFA features can be incapable of precisely representing the labels. As our proposed framework, MALTOR significantly outperforms all baseline methods in all datasets and target metrics with both linear and deep XMC models, except the F1 score in Amazon-670K with XR-LINEAR.

Ablation Study. Figure 3 shows the performance changes in F1 and DCG on four datasets after removing each component and each feature set from MALTOR. If a component is more important, the performance loss after its removal would be also huger. For the F1 metric, ordinal regression is the most important components. This is because ordinal regression obtains ordering knowledge. In contrast, predictive scores play the most impactful role for the DCG metric. The reason could be that these scores can be helpful to detect irrelevant labels ranked in high positions to avoid penalty in the DCG metric. The only negative impact is made by ordinal regression for DCG on the Amazon-670K dataset. It could be because the average truncation positions with best DCG on Amazon-670K is only 2.1 ± 2.6 so that the model only needs to focus on top-ranked



(a) Performance Changes in F1 **(b) Performance Changes in DCG**

Figure 3: Perofrmance changes after removing each component and features from MALTOR. OR, PS, DR abbreviate ordinal regression, predictive score, and data representation, respectively.

labels. The ordinal information about relations among all positions can introduce more noises in this case.

Cost Selection. The mislabeling cost matrix $C_{k,i}$ plays an important role in both task reduction and optimization. Table 3 shows

Dataset	F1		DCG	
	Absolute	Squared	Absolute	Squared
Eurlex-4K	**0.5764**	0.5685	**1.4508**	1.4342
Wiki10-31K	**0.3413**	0.3401	**1.6062**	1.5959
Amazon-670K	**0.6638**	0.6592	**1.7240**	1.6991
AmazonCat-13K	**0.3368**	0.3258	0.4920	**0.4921**
Wiki-500K	**0.4213**	0.4114	**0.7814**	0.7655
Amazon-3M	**0.2435**	0.2386	**1.2510**	1.1894

Table 3: Performance of MALTOR with absolute and squared costs as mislabeling costs $C_{k,i}$.

the performance of MALTOR with absolute costs $C_{k,i} = |k - i|$ and squared costs $C_{k,i} = |k - i|^2$. We can see that absolute costs lead to more satisfactory performance in most of the datasets, compared to squared costs. This can be because the number of candidate positions K is 100 so that squared costs could be too sensitive to weigh both close and far mispredictions. As a result, we choose the absolute costs to formulate the mislabeling cost matrix for optimization.

4 CONCLUSION

In this paper, we are the first study to tackle unreasonable prediction for conventional eXtreme Multi-label Classification (XMC) methods that usually only consider a fixed and short length for ranked label lists. To conduct reasonable prediction, we propose the novel framework, Model-Agnostic List Truncation with Ordinal Regression (MALTOR), to derive ranked label lists with dynamic lengths for different data inputs. We first leverage ordinal regression to consider relations among labels. By using V-shaped mislabeling costs, we reduce the task of ordinal regression into an extended binary classification problem with a theoretically guaranteed generalization bound. Experimental results on six benchmarks demonstrate that MALTOR not only can be applied to any XMC models and prediction, but also significantly outperforms statistical and neural baseline methods. We also show the effectiveness of each component and feature set in MALTOR with an in-depth ablation study.

REFERENCES

[1] Hosein Azarbonyad, Mostafa Dehghani, Maarten Marx, and Jaap Kamps. 2021. Learning to rank for multi-label text classification: combining different sources of information. *Natural Language Engineering* 27, 1 (2021), 89–111.

[2] Rohit Babbar and Bernhard Schölkopf. 2017. DiSMEC: Distributed sparse machines for extreme multi-label classification. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 721–729.

[3] Rohit Babbar and Bernhard Schölkopf. 2019. Data scarcity, robustness and extreme multi-label classification. *Machine Learning* 108, 8 (2019), 1329–1351.

[4] Dara Bahri, Yi Tay, Che Zheng, Donald Metzler, and Andrew Tomkins. 2020. Choppy: Cut transformer for ranked list truncation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1513–1516.

[5] Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noémie Elhadad. 2018. Multi-label classification of patient notes: case study on ICD code assignment. In *Workshops at the thirty-second AAAI conference on artificial intelligence*.

[6] Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. 2015. Sparse Local Embeddings for Extreme Multi-label Classification.. In *NIPS*, Vol. 29. 730–738.

[7] Wei-Cheng Chang, Daniel Jiang, Hsiang-Fu Yu, Choon Hui Teo, Jiong Zhang, Kai Zhong, Kedarnath Kolluri, Qie Hu, Nikhil Shandilya, Vyacheslav Ievgrafov, et al. 2021. Extreme multi-label learning for semantic matching in product search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2643–2651.

[8] Wei Chu and S Sathiya Keerthi. 2005. New approaches to support vector ordinal regression. In *Proceedings of the 22nd international conference on Machine learning*. 145–152.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171—-4186.

[10] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *the Journal of machine Learning research* 9 (2008), 1871–1874.

[11] Eibe Frank and Mark Hall. 2001. A simple approach to ordinal classification. In *European conference on machine learning*. Springer, 145–156.

[12] Jyun-Yu Jiang, Patrick H Chen, Cho-Jui Hsieh, and Wei Wang. 2020. Clustering and constructing user coresets to accelerate large-scale top-k recommender systems. In *Proceedings of The Web Conference 2020*. 2177–2187.

[13] Sujay Khandagale, Han Xiao, and Rohit Babbar. 2020. Bonsai: diverse and shallow trees for extreme multi-label classification. *Machine Learning* 109, 11 (2020), 2099–2119.

[14] Ling Li and Hsuan-Tien Lin. 2006. Ordinal regression by extended binary classification. *Advances in neural information processing systems* 19 (2006).

[15] Yen-Chieh Lien, Daniel Cohen, and W Bruce Croft. 2019. An assumption-free approach to the dynamic truncation of ranked lists. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*. 79–82.

[16] Weiwei Liu, Donna Xu, Ivor W Tsang, and Wenjie Zhang. 2018. Metric learning for multi-output tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 408–422.

[17] Peter McCullagh. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)* 42, 2 (1980), 109–127.

[18] Alexandru Niculescu-Mizil and Ehsan Abbasnejad. 2017. Label filters for large scale multilabel classification. In *Artificial intelligence and statistics*. PMLR, 1448–1457.

[19] Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In *Proceedings of the 2018 World Wide Web Conference*. 993–1002.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

[21] Christopher Winship and Robert D Mare. 1984. Regression models with ordinal variables. *American sociological review* (1984), 512–525.

[22] Chen Wu, Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. 2021. Learning to Truncate Ranked Lists for Information Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 4453–4461.

[23] Shunyao Wu, Yuzhu Chen, Zhiruo Li, Jian Li, Fengyang Zhao, and Xiaoquan Su. 2021. Towards multi-label classification: Next step of machine learning for microbiome research. *Computational and Structural Biotechnology Journal* (2021).

[24] Yiming Yang and Siddharth Gopal. 2012. Multilabel classification with meta-level features in a learning-to-rank framework. *Machine Learning* 88, 1 (2012), 47–68.

[25] Ian EH Yen, Xiangru Huang, Wei Dai, Pradeep Ravikumar, Inderjit Dhillon, and Eric Xing. 2017. Ppdsparse: A parallel primal-dual sparse method for extreme classification. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 545–553.

[26] Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2019. AttentionXML: Label Tree-based Attention-Aware ee Model for High-Performance Extreme Multi-Label Text Classification. *Advances in Neural Information Processing Systems* 32 (2019), 5820–5830.

[27] Hsiang-Fu Yu, Kai Zhong, Jiong Zhang, Wei-Cheng Chang, and Inderjit S Dhillon. 2022. PECOS: Prediction for enormous and correlated output spaces. *the Journal of machine Learning research* (2022).

[28] Jiong Zhang, Wei-Cheng Chang, Hsiang-Fu Yu, and Inderjit S Dhillon. 2021. Fast Multi-Resolution Transformer Fine-tuning for Extreme Multi-label Text Classification. In *Advances in Neural Information Processing Systems*.

[29] Wenbin Zheng, Xiaping Fu, and Yibin Ying. 2014. Spectroscopy-based food classification with extreme learning machine. *Chemometrics and Intelligent Laboratory Systems* 139 (2014), 42–47.