ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

Why state-of-the-art deep learning barely works as good as a linear classifier in extreme multi-label text classification

Mohammadreza Qaraei[1], Sujay Khandagale[2] and Rohit Babbar[1]

1- Aalto University, CS Department
Helsinki, Finland

2- Columbia University, CS Department
New York, USA

Abstract. Extreme Multi-label Text Classification (XMTC) refers to supervised learning of a classifier which can predict a small subset of relevant labels for a document from an extremely large set. Even though deep learning algorithms have surpassed linear and kernel methods for most natural language processing tasks over the last decade; recent works show that state-of-the-art deep learning methods can only barely manage to work as well as a linear classifier for the XMTC task. The goal of this work is twofold : (i) to investigate the reasons for the comparable performance of these two strands of methods for XMTC, and (ii) to document this observation explicitly, as the efficacy of linear classifiers in this regime, has been ignored in many relevant recent works.

1 Introduction

Extreme Multi-label Text Classification (**XMTC**) refers to supervised learning of a classifier that can automatically label a document with a small subset of relevant labels from an *extremely large set of all possible target labels*. Machine learning problems consisting of hundreds of thousands of labels are common in various domains such as product categorization for e-commerce [1, 2], hash-tag suggestion in social media [3], and annotating web-scale encyclopedia [4]. It has been demonstrated that, in addition to automatic labelling, the framework of XMTC can be leveraged to effectively address learning problems arising in bid-phrase suggestion in web-advertising and recommendation systems [4].

In terms of algorithmic advances in machine learning, deep learning methods, by achieving significant performance improvement over classical methods such as SVMs and conditional random fields, have become the *de facto* methods of choice for vision, language, and audio processing tasks. Leveraging deep architectures for representation learning, coupled with techniques such as bi-directional LSTM [5] and Attention mechanism [6], significant advances have been made in various domains such as language modeling [7], machine translation [8], sentiment analysis [9], syntactic parsing, named-entity recognition [10].

Deep learning techniques have also been proposed for the XMTC task. For instance, a convolutional neural network was proposed in [11]. Motivated by the success of sequence-to-sequence models in machine translation, a similar approach was also proposed in [12] for large-scale multi-label classification. On the XMTC task however, contrary to many of those in Natural Language Processing (NLP), deep learning methods have achieved limited success compared to using a distributed linear classifier trained on bag of words data representation [13]. Recently, an attention-based mechanism for XMTC, has been able to slightly surpass linear methods [14].

Our contributions. In this work, we take an investigative approach rather than an algorithmic one. By first summarizing the results from recent works, we then discuss the reasons for the comparable performance of linear classifiers and deep learning methods for XMTC. We argue that the primary reason for this is *labeled data scarcity* in XMTC due to the fat-tailed behavior of distribution of training instance among labels. Another goal of our work is to reiterate the efficacy of linear methods compared to deep learning for XMTC tasks, which has been overlooked in recent works [15].

2 Problem setup and data representation

Let the training data, given by $\mathcal{T} = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\}$, consists of input feature vectors $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^D$ and respective output vectors $\mathbf{y}_i \in \mathcal{Y} \subseteq \{0,1\}^L$ such that $\mathbf{y}_{i_\ell} = 1$ iff the ℓ-th label belongs to the training instance $\mathbf{x}_i$. The goal in XMTC is to learn a multi-label classifier in the form of a vector-valued function $f : \mathbb{R}^D \rightarrow \{0,1\}^L$. The desired function f can be a shallow network (such as a linear SVM) (as shown in Figure 1a) or an attention-based deep neural network from [14] (as shown in Figure 1b).

In XMTC setup, training instance $\mathbf{x}_i$ is represented in two forms, (i) Bag-of-Words (BOW) which is sparse and works in conjunction with linear/shallow classifiers, and (ii) word embeddings, which are dense and form the input layer of a deep learning pipeline.

With its applications in recommendation systems and web-advertising, where correct prediction at top-k slots is of value, the metric for comparison is precision@k (denoted P@k). For a true label vector $\mathbf{y} \in \{0,1\}^L$ and predicted label vector $\hat{\mathbf{y}} \in \mathbb{R}^L$: P@k := $\frac{1}{k} \sum_{l \in rank_k(\hat{\mathbf{y}})} \mathbf{y}_l$ such that $rank_k(\mathbf{y})$ returns the k largest indices of $\mathbf{y}$ ranked in descending order. Propensity-scored variants which capture the classifier performance on tail-labels have also been used in recent works [16].

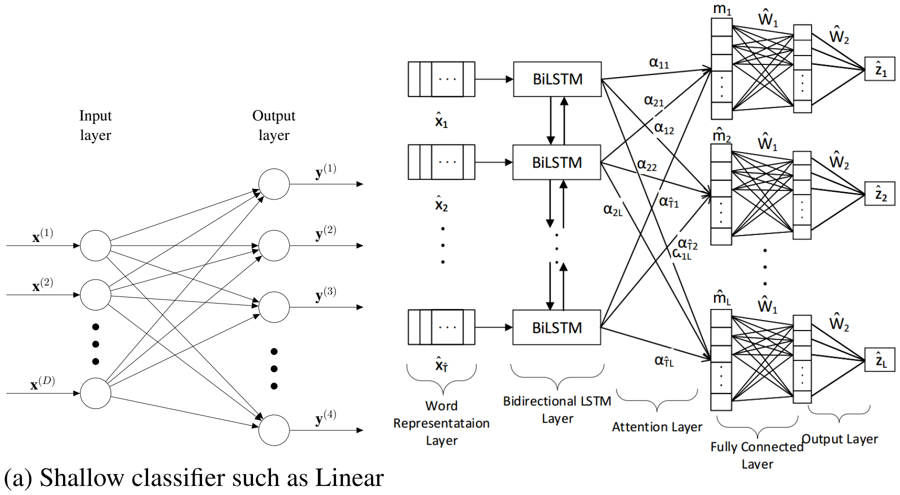
3 Deep learning versus linear classification for XMTC

Dataset	N	N_{test}	L	D	LPI	IPL
Eurlex-4K	15,539	3,809	3,993	5,000	5.32	15.59
AmazonCat-13K	1,186,239	306,782	13,330	203,882	5.04	448.57
Wiki10-31K	14,146	6,616	30,938	101,938	18.74	8.11
Amazon-670K	490,449	153,025	670,091	135,909	5.45	3.99

Table 1: Data Statistics

In this section, we show the comparison of state-of-the-art XMTC approaches on five benchmark datasets from the Extreme Classification Repository [1]. The statistics of

[1] http://manikvarma.org/downloads/XC/XMLRepository.html



(a) Shallow classifier such as Linear SVM

(b) Attention-based deep architecture [14]

Fig. 1: Pictorial depiction of shallow versus deep architectures for XMTC based on Bag-of-Words and word embeddings respectively.

datasets is shown in Table 1, where N is number of training instances, N_{test} is number of test instances, L is the label set dimension, D is the dimensionality of the feature space for BOW representation, LPI is the number of labels-per-instance and IPL is number of instances-per-label. Along with the BOW representation, the repository also provides the raw text for using pre-trained word embeddings and task specific fine-tuning. We compared four state-of-the-art methods, two each from linear classification and deep learning.

For deep learning methods, XML-CNN [11] extends the multi-class convolutional architecture proposed in [17], for multi-label learning. It uses a dynamic max-pooling operation over a convolutional layer using pre-trained 300-dimensional GloVe embeddings. A recently proposed AttentionXML [14] (to appear in upcoming NeurIPS 2019) employs a BiLSTM layer over pretrained 300-dimensional GloVe embeddings followed by an attention layer and one or two fully connected layers. The label-wise attention mechanism enhances the prediction performance by focussing on relevant input words for each positive label in an instance. The architecture is shown in Figure 1b.

For linear methods, state-of-the-art DiSMEC [13] classifier employs a one-vs-rest scheme in a distributed setup for concurrent training across labels with squared hinge loss. Another recently proposed method, Bonsai [18], employs linear SVM in a shallow tree architecture for faster training compared to DiSMEC. As a result, it retains better properties of one-vs-rest and tree-based methods, high predictive accuracy from the former and fast training and prediction from the latter. Other linear methods, such as Parabel [19], perform worse compared to DiSMEC and Bonsai, have been ignored in the interest of space.

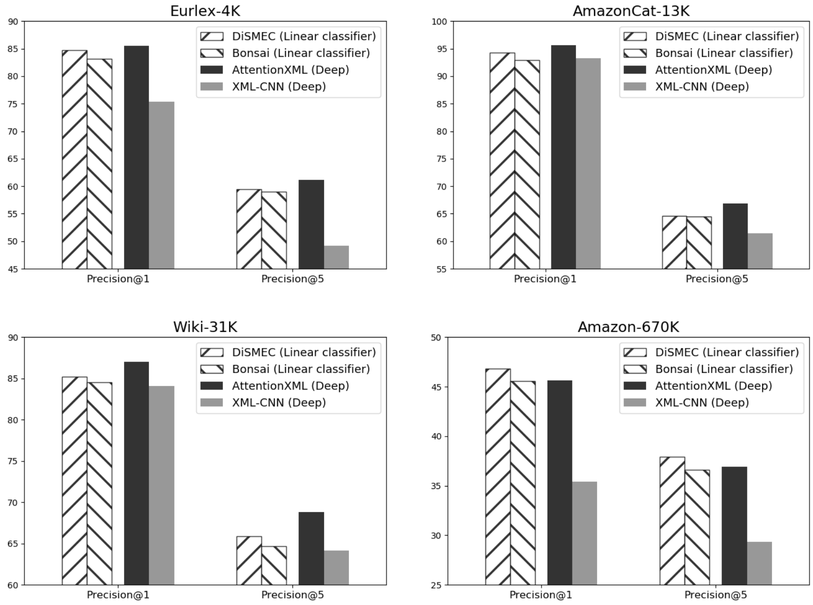


Fig. 2: Comparison of shallow classifiers with BOW vs deep classifiers with embeddings.

3.1 Comparison and discussion

The comparison of deep methods versus linear classifier is shown in Figure 2, where precision@1 and precision@5 metrics are shown as bars. It is clear that, the linear methods, DiSMEC and Bonsai (shown as shaded with white background), are significantly better than XML-CNN (gray colored bars). In comparison to AttentionXML (black-colored bars), the linear DiSMEC model shows better performance on Amazon-670K, and together with Bonsai, obtains quite comparable results on other datasets.

It is interesting to note from the above observations that even though deep learning methods have been largely successful in many NLP and vision tasks, their success for the XMTC task is rather limited. We discuss below possible reasons for this phenomenon

1. **Data scarcity.** The distribution of training instances among labels for the rest four datasets, exhibit fit to power-law distribution. This distribution is shown in Figure 3 for **Amazon-670K** dataset. It shows that only $\sim$100,000 out of 670,000 labels have more than 5 training instances in them [20, 21]. This means that a large fraction of labels have very few training instances that belong to them. On the other hand, the success of deep learning methods depends heavily on the abundance of training data. As a result, deep learning approaches trained on word-embeddings perform sub-optimally compared to the shallow methods on bag-of-words data representation. This is also evident from the last column of

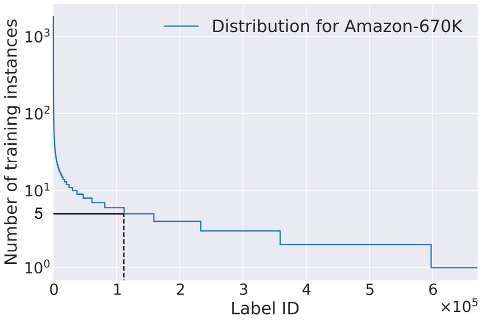


Fig. 3: Power-law distribution. Y-axis is on log-scale. X-axis represents labels sorted according to decreasing number of training instance that belong to them.

Table 1 (IPL), which shows the average number of training instances per label for different datasets.

2. **Nature of the XMTC task.** Firstly, deep learning approaches employing word-embeddings capture strong semantic and contextual information. This is crucial for tasks such as sentiment analysis [10, 22]. However, the nature of XMTC task is quite different, it mainly involves picking the right keyword for each document such as keywords for a News story on the BBC or an article on Wikipedia. The polarity or sentiment of the article matters much less in this context as compared to the main keywords which could be about personalities, place, countries, events, etc. For such scenarios, the signal captured by classical bag-of-words representation and learnt using shallow classifier is strong enough to get correct classification.

4 Conclusion

In this paper, we highlighted the fact that state-of-the-art deep learning methods perform at par with shallow classifiers which employ bag-of-words for the task of large-scale multi-label classification in XMTC. This is quite contrary to the success of these methods on a variety of other NLP and vision tasks. We hope that theoretical and algorithmic progress of deep learning methods to use unlabeled data and other side information to alleviate the data-scarcity problem in XMTC is crucial to its success in other scenarios with limited labeled data such as the situation encountered in self driving-cars.

References

[1] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *RecSys*, pages 165–172. ACM, 2013.

[2] Samy Bengio, Jason Weston, and David Grangier. Label embedding trees for large multi-class tasks. In *NIPS*, pages 163–171, 2010.

[3] Emily Denton, Jason Weston, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. User conditional hashtag prediction for images. In *KDD*, 2015.

[4] Yashoteja Prabhu and Manik Varma. Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *KDD*, pages 263–272. ACM, 2014.

[5] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[7] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh annual conference of the ISCA*, 2010.

[8] I Sutskever, O Vinyals, and QV Le. Sequence to sequence learning with neural networks. *NIPS*, 2014.

[9] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th ICML*, pages 513–520, 2011.

[10] Rie Johnson and Tong Zhang. Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058*, 2014.

[11] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. Deep learning for extreme multi-label text classification. In *SIGIR*, pages 115–124. ACM, 2017.

[12] Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. Maximizing subset accuracy with recurrent neural networks in multi-label classification. In *NIPS*, pages 5413–5423, 2017.

[13] Rohit Babbar and Bernhard Schölkopf. Dismec: Distributed sparse machines for extreme multi-label classification. In *WSDM*, pages 721–729, 2017.

[14] Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. In *NeurIPS*, 2019.

[15] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Extreme multi-label legal text classification: A case study in eu legislation. *arXiv preprint arXiv:1905.10892*, 2019.

[16] Rohit Babbar and Bernhard Schölkopf. Data scarcity, robustness and extreme multi-label classification. *Machine Learning*, pages 1329–1351, 2019.

[17] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[18] Sujay Khandagale, Han Xiao, and Rohit Babbar. Bonsai-diverse and shallow trees for extreme multi-label classification. *arXiv preprint arXiv:1904.08249*, 2019.

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

[20] Rohit Babbar, Krikamol Maundet, and Bernhard Schölkopf. Tersesvm: A scalable approach for learning compact models in large-scale classification. In *SDM*, pages 234–242. SIAM, 2016.

[21] Rohit Babbar, Ioannis Partalas, Eric Gaussier, and Massih-Reza Amini. Re-ranking approach to classification in large-scale power-law distributed category systems. In *ACM SIGIR*, pages 1059–1062, 2014.

[22] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, pages 649–657, 2015.