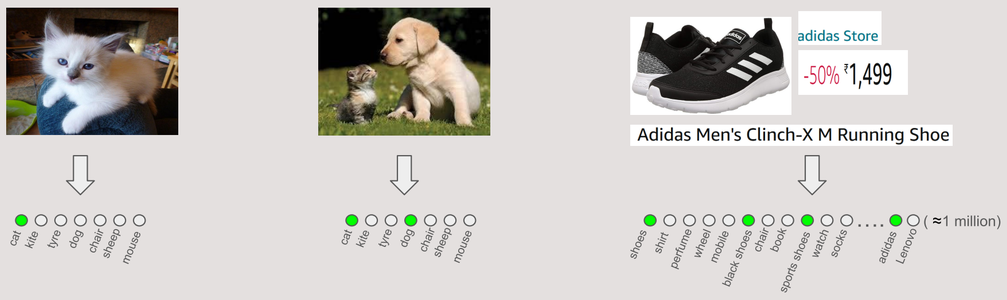
A Multi-Level Cluster-Search Approach for XMC

Mahim Jain, Gourav Pathak*, Prakhar Verma*, Mahesh Mohan M R*
SPAI Group, Department of AI, IIT Kharagpur
{gourvavpathak2002, mjain, prakharverma}@kgpian.iitkgp.ac.in

01. Introduction and Motivation

Extreme Multi-label Classification (XMC) is a task where we have a large number of possible labels (sometimes in the range of millions) and the goal is to predict a subset of relevant labels for each input.

Classification: Multiclass Vs Multilabel vs Extreme Multilabel



Motivation for XMC: To retrieve products and/or contents based on tags

- Amazon: XMC helps recommend products by matching items to millions of possible categories based on customer preferences and search history.
- Wikipedia: XMC can be used to tag Wikipedia articles with relevant topics from a huge set of categories, making it easier for users to find related content.
- Google, Bing, etc: Search engines can similarly be optimized using XMC.

02. Challenges

- When we have millions of labels, it's really hard to train a model efficiently. Too many labels makes it slow or even impossible to train the model in an "end-to-end" way (meaning training everything together without breaking it into smaller parts) [2].
- One another challenge is to retrieve tail products, which are rare and infrequently represented in the dataset. The scarcity of data makes it difficult for models to learn and accurately classify these niche items, impacting retrieval efficiency and overall performance in recommending or categorizing rare products [2].

03. Existing Solution

- Modular approaches: Breaking the big problem into smaller pieces and solving them separately. This makes it easier to manage but could reduce accuracy [2].
- Sampling techniques: Instead of looking at all the labels, they choose a subset of labels for each data point. This speeds up training but could result in accuracy loss if the wrong labels are chosen for sampling [2].
- Earlier methods used simpler sparse linear models (models that assume most of the features are zero) to solve this problem. However, these methods were not as effective, and deep learning models started to perform better [2].
- The problem with using deep learning for XMC is that training these models in an end-to-end manner (training everything together) has been considered impossible because of memory and computational constraints [2].

04. Methodology: 'Multi-Level' & 'Cluster-Search'

We specifically design and explain our approach by dividing it into two simpler parts i.e. explaining Multi-Level dataset hierarchy and Cluster-Search separately using schematic diagrams.

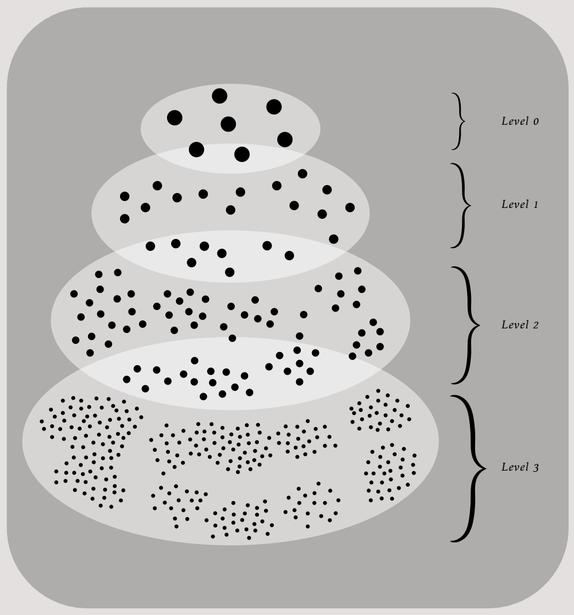


Fig. 4.1: Schematic of Multi-Level Target Labels representation

When we talk about a multi-level dataset visualization, we focus on the following assumptions about the dataset in particular:

- The entire set of 'label' targets can be separated into different subsets
- Each subset will form a 'level' where the data points falling into that 'level' will be related to its predecessor and successor levels by a relationship that they are the 'labels' for a single 'document' or 'query'
- The data points falling into a particular 'level' are mutually disjoint with respect to the 'documents' or 'queries' they are the 'labels' of

Now, having defined these conditions, we partition our dataset into such levels and search one-one by for the most relevant 'labels' to the given 'query'

We follow the same structural hierarchy of the data 'labels' as defined while searching for the relevant targets. The 'Cluster Search' algorithm we use contains the following parts:

- The initial input to the search algorithm is the Level-0 data labels and we retrieve the most relevant label out of them
- Next, we feed this Level-0 search output to our model which then finds for itself the most relevant potential Level-1 labels 'cluster' for the 'query'
- Now that our search space is reduced, we search on this reduced 'cluster' using our encoder model to extract the most relevant Level-1 'label'
- We cyclically feed the search output of our previous level to find out the search space for the next level

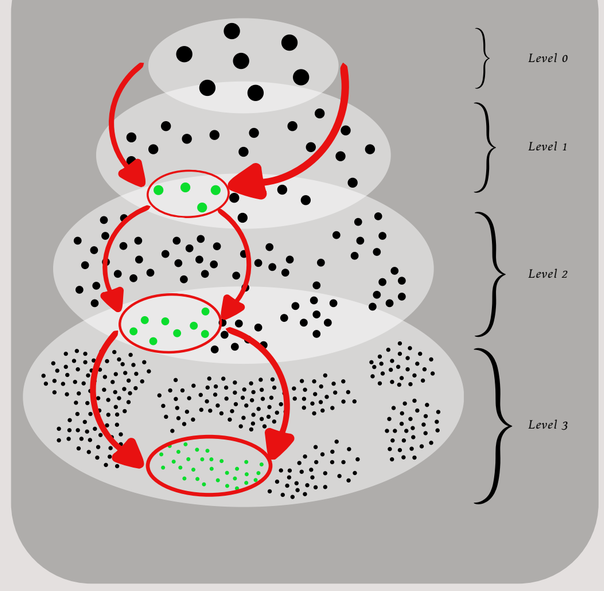


Fig. 4.2: Schematic of Cluster Search on different dataset levels

ⓘ Size of the data points shown on the schematics represents the number of 'documents' it represents as a label

05. T5 (Text-To-Text-Transfer-Transformer)

- The basic idea underlying T5 (Text-To-Text-Transfer-Transformer) model is to treat every text processing problem as a "text-to-text" problem, i.e. taking text as input and producing new text as output [1].
- T5 encoder-decoder Transformer implementation closely follows its originally proposed form [3].
- We use a simplified form of position embeddings where each "embedding" is simply a scalar that is added to the corresponding logit used for computing the attention weights [1].
- The primary building block of the Transformer is self-attention that processes a sequence by replacing each element by a weighted average of the rest of the sequence [4].

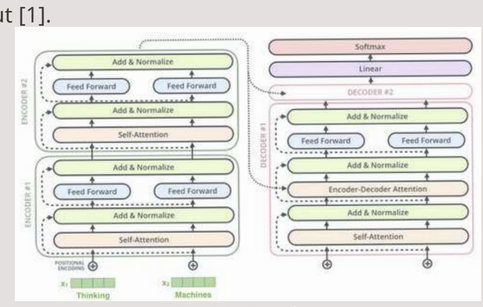


Fig. 5.1 T5 Architecture

06. Results and Findings

- Introduction to the Amazon dataset that we are working with
- The Input Features are shown in Table 6.1 :

S. No.	File_name	Product_ID	Title	Price	Store	Manufacturer
1	Automotive	B000C9FJ1Y	GM 15-8535 Heating and Air Conditioning Blower...	NaN	GM	ACDelco
2	Electronics	B075HRNB8K	Polaroid PIF-300 Instant Film - Twin Pack	NaN	Polaroid	Polaroid
3	Automotive	B07J9ZFLT8	CoolingCare Radiator for 1992-2004 Chevy GMC C...	NaN	Cooling Care	Cooling Care
4	Electronics	B086384SF5	RM-GD014 Remote Control Replacement for Sony ...	7.95	Elekpia	Elekpia Factory
...	...	...	...	...	...	...
1680013	Home_and_Kitchen	B004L8V95C	Urnex Cafiza Espresso Machine Cleaning Tablets...	7.34	Urnex	Urnex

Table 6.1

Fig. 6.1 Automotive

Fig. 6.2 Home & Kitchen

Fig. 6.3 Sports

- Figures 6.1, 6.2 and 6.3 show Word Cloud analysis for Automotive, Home & Kitchen and Sports associated titles respectively.
- Automotive category word cloud analysis shows that most frequently occuring words in the titles of product are Passenger Side, Driver Side, Brake Pad, Compatible etc.
- Home & Kitchen category word cloud analysis shows that the most frequently occuring words are Wall art, White, Art Print etc.
- Sports category word cloud analysis shows that the most frequently occuring words are Black, T Shirt, One Size etc.

- The Labels to be predicted are shown in Table 6.2:

S. No.	Brand_category	L0_category	L1_category	L2_category	L3_category	L4_category
1	GM (1934)	Automotive (484633)	Replacement Parts (253381)	Engine Cooling & Climate Control (18199)	Heating (2081)	Blower Motors (1718)
2	Polaroid (213)	Electronics (166486)	Camera & Photo (25077)	Film Photography (1030)	Film (413)	na (858841)
3	Cooling Care (62)	Automotive (484633)	Replacement Parts (253381)	Engine Cooling & Climate Control(18199)	Radiators (4484)	na (858841)
...	...	...	...	...	...	...
1680012	Remo (143)	Instrument Accessories (48)	Drum & Percussion Accessories (48)	Concert Percussion Accessories (48)	Drum Accessories (48)	Snare Drum Accessories (22)
1680013	Urnex (32)	Kitchen & Dining (54)	Small Appliance Parts & Accessories (54)	Coffee & Espresso Machine Parts & Accessories(54)	Coffee & Espresso Machine Cleaning Products (32)	na (858841)
Total Labels	14262	31	198	911	2199	1852
Tail Labels	1752	0	3	47	133	133

Table 6.2 (Number of samples is provided inside bracket)

- Tail Labels are labels which have less than or equal to 25 training samples in the dataset.
- The F1 score of the T5 Model on various categories of labels in the dataset is shown in Fig. 6.4.
- The high F1 score for Brand_category is attributed to its strong correlation with the Manufacturer attribute in the input data.
- The elevated F1 score observed in the L4 category suggests that the model has predominantly learned to output "NA" for this label, rather than demonstrating genuine classification accuracy.

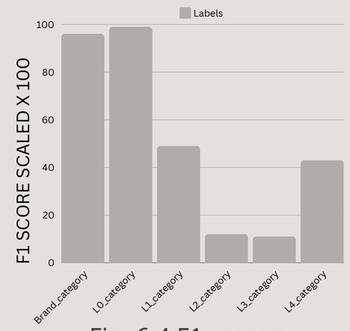


Fig. 6.4 F1 scores

07. Conclusion and Future Work

- The model exhibited robust performance at higher label levels, with an F1 score of 0.96 for the brand category and an impressive 0.99 for the L0 category.
- A notable decline in F1 scores was observed in L1, L2, L3 and L4 labels.
- These results highlight both strengths and areas for improvement in the model's classification capabilities.
- For future work, we will focus on improving model performance for deeper label categories such as L2 and L3, enhancing F1 scores across all classification hierarchies (L0 to L4).

08. References

[1] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, JMLR, 2020.
[2] Vidit Jain, Jatin Prakash, Deepak Saini, Jian Jiao. Renee: End-To-End Training of Extreme Classification Models, MLSys, 2023.
[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit. Attention Is All You Need, NeuriPS, 2017.
[4] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long Short-Term Memory-Networks for Machine Reading, EMNLP, 2016.

** - Indicates Equal Contribution*