PMTL: A Progressive Multi-level Training Framework for Retail Taxonomy Classification

Gaurab Bhattacharya, Gaurav Sharma, Kallol Chatterjee, Chakrapani, Bagya Lakshmi V, Jayavardhana Gubbi
Arpan Pal and Ramachandran Rajagopalan
Tata Consultancy Services, India

***Abstract*—Retail taxonomy classification provides hierarchical labelling of items and it has widespread applications, ranging from product on-boarding, product arrangement and faster retrieval. It is fundamental to both physical space as well as e-commerce. Manual processing based on meta-data was adopted and more recently, image based approaches have emerged. Traditionally, hierarchical classification in retail domain is performed using feature extractors and using different classifier branches for different levels. There are two challenges with this approach: error propagation from previous levels which affects the decision-making of the model and the label inconsistency within levels creating unlikely taxonomy tree. Further, the training frameworks rely on large datasets for generalized performance. To address these challenges, we propose PMTL, a progressive multi-level training framework with logit-masking strategy for retail taxonomy classification. PMTL employs a level-wise training framework using cumulative global representation to enhance and generalize output at every level and minimize error propagation. Also, we have proposed logit masking strategy to mask all irrelevant logits of a level and enforce the model to train using only the relevant logits, thereby minimizing label inconsistency. Further, PMTL is a generalized framework that can be employed to any full-shot and few-shot learning scheme without bells and whistles. Our experiments with three datasets with varied complexity in full-shot and few-shot scenario demonstrates the effectiveness of our proposed method compared to the state-of-the-art.**

I. Introduction

Hierarchical labelling of objects is a natural and frequent phenomenon for categorization irrespective of the domain. This is predominant in retail sector where millions of products are organized to support hierarchical labelling, *e.g.,* biscuits will be placed under the snacks section of grocery unit. Due to the recent exponential growth in large retail shops and e-commerce sector, the problem becomes more fundamental for search, retrieve and planogram design. Hence, in retail industry, taxonomy of objects play a major role as far as product alignment, association and customer experience are concerned. Out of the different objects, apparel taxonomy classification using images is an important aspect due to its large variation, high inter-class similarity, inter-relationship of labels and significant global market share. Hence, hierarchical fashion taxonomy classification framework is a key component to ensure automatic internal mapping and association of products, faster retrieval with few clicks and improved customer satisfaction.

In recent years, many research works have explored hierarchical taxonomy classification in fashion domain [2], [4], [10], [18] and beyond [9], [17], [19]. Traditionally, hierarchical training is performed using a global classifier [2], [19] or by level-based or parent-node based local classifier [4], [17]. The level-based local classifier trains separate model for each level, hence label inconsistency problem crops up where outputs of different levels create an impossible combination (*e.g.*, menswear $\rightarrow$ top-wear $\rightarrow$ leggings). To mitigate this, parent node-based local classifier can be trained where the model in one level is selected based on the decision by its predecessor. However, this is computationally expensive, especially in retail scenario having large number of classes. Moreover, error propagation from previous level output can significantly impact its performance. Also, existing hierarchical taxonomy classification frameworks are suited for large-scale datasets and they are not directly applicable where there is lack of data.

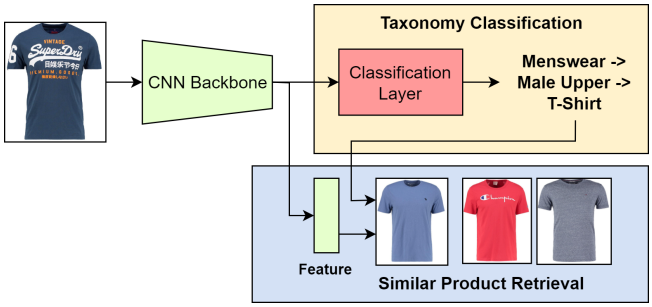

Fig. 1: Hierarchical taxonomy classification and similar product retrieval using the proposed PMTL framework.

To address these three challenges, we propose **P**rogressive **M**ulti-level **T**raining with **L**ogit masking (PMTL), a generalized hierarchical taxonomy classification framework for few-shot and full-shot data. PMTL enables the models to be trained separately for each level, hence reducing error propagation problem during training. To further enhance the model's performance at each level and get the label-wise constraint from the previous level, we augment the global representation from the model of the previous level. During the training, we use logit masking strategy to restrict the model to learn only relevant classes through part of final classification layer, thereby addressing the label inconsistency issue and incorporating the benefit of parent node-based local classifier. This framework is generalized irrespective of dataset size and can be attached to any hierarchical classification network,

including few-shot methods such as [11], [12], [14] without bells and whistles. The capabilities of PMTL framework are depicted in Figure 1, where hierarchical taxonomy classification and similar item retrieval are performed. To the best of our knowledge, this is the first attempt to solve these challenges irrespective of dataset size in retail scenario. To validate the efficacy of the proposed PMTL, we have experimented with three publicly available large scale real-world datasets and our method outperformed all existing methods by a significant margin. The contributions of this work are given below:

- We propose PMTL, a generalized progressive training framework for image based hierarchical classification which consolidates the benefit of global feature-based, parent node-based and level-based classifier training.
- To address label inconsistency issue, we have proposed logit masking strategy to use only the relevant part of classification layer at every level to perform training, thereby increasing scalability of the model.
- The proposed PMTL framework can be adopted for both few-shot and full-shot scenario and we conduct extensive experiments to validate the novel components in our approach.

II. Related works

A. Hierarchical Image Classification

In visual recognition, category hierarchy exploits the relationship between coarse and fine-grained classes [13] and has demonstrated improvement in classification performance [9], [17]. Traditionally, hierarchical image classification frameworks either train a global network followed by separate classification branches [2], [19] for different levels or they go for multi-step training process using local classifiers per level or per parent node [4], [17]. However, global classifiers often give inferior performance to local classifiers since features of global classifier is not specialized for each level. On the other hand, local classifier-based approaches face two problem: label inconsistency in local classifier per level and error propagation in local classifier per node. To address this, we have proposed PMTL where we progressively train different levels of hierarchical classifier by considering the proposed label masking strategy to overcome the label inconsistency issue. By this, we enforce the model to only focus on relevant classes depending on its previous label while training. Also, contrary to local classifier per node, the final output of a level in our method do not determine the model's decision of choosing a model in the next level and addresses the error propagation problem.

B. Few-shot Learning

In recent years, few-shot learning framework has seen widespread development across different domains due to its use of very less data per class, improved performance in unseen classes and efficient embedding creation mechanism. Matching network [14] creates embeddings from few-shot examples and then uses a new image to match feature similarity between them. Prototypical network [11] creates average embedding vector (*i.e.,* prototypes) for each class and returns the class with minimum distance from its corresponding prototype for a new image. Network relation module [12] is used to compare the features from the support images with that of a new image and generate relation score. Although these methods have been widely popularized in biometric recognition and medical applications due to their constrained nature, they have not been widely used in retail scenario due to its large variation in data. To the best of our knowledge, our paper is the first attempt to explore the effect of few-shot learning in a hierarchical retail image classification.

C. Retail Taxonomy Classification

Automatic retail taxonomy classification is challenging in multiple ways: large variations in image, high inter-class similarity, hierarchical labels and their inter-relationships and large number of classes. Authors in [2] proposed a novel hierarchical fashion image classification model using HMCN-F [15] as backbone. Hierarchical classification on Fashion-MNIST [16] is performed in [10]. In [4], authors proposed Add-Net and Concat-Net to address this problem. Recently, [18] introduced hierarchy-preserving losses for taxonomy classification. Authors in [8] fuse features from multiple levels using visual attention for hierarchical classification.However, [4], [8], [10] only used datasets such as Fashion-MNIST [16] and CIFAR-100 [6] which do not possess the complexity of real-world retail datasets. [2] uses global classifier, giving inferior performance, as evident in Sections 4 and 5 of our paper and all these methods [2], [4], [8], [10], [18] do not address the crucial challenges, *i.e.,* label inconsistency and error propagation. On the contrary, our proposed PMTL aims to address these problems while building a general hierarchical training framework for both full-shot and few-shot scenario.

III. Proposed method

PMTL is a generalized framework for hierarchical multi-label classification which addresses three crucial challenges of hierarchical classification: error propagation from previous levels, label inconsistency between levels and poor performance due to small dataset. To address error propagation, we have proposed progressive multi-label training and then logit masking strategy is proposed to circumvent label inconsistency. Finally, we employ PMTL in full-shot and few-shot scenario for three datasets to analyze its performance with less data in training set.

A. Progressive Multi-level Training

The hierarchical classification models are trained by one of the following three ways: (a) global network followed by classification layers for each levels; (b) level-wise local classifier; and (c) parent node-wise local classifier. In Progressive Multi-level Training (PMT), we consolidate features from all these training methods to enhance the taxonomy classification performance and alleviate the challenges. The training framework for 3-level taxonomy classification is depicted in Figure 2. PMT is a hierarchical training process where models are separately trained for each level. The level-wise

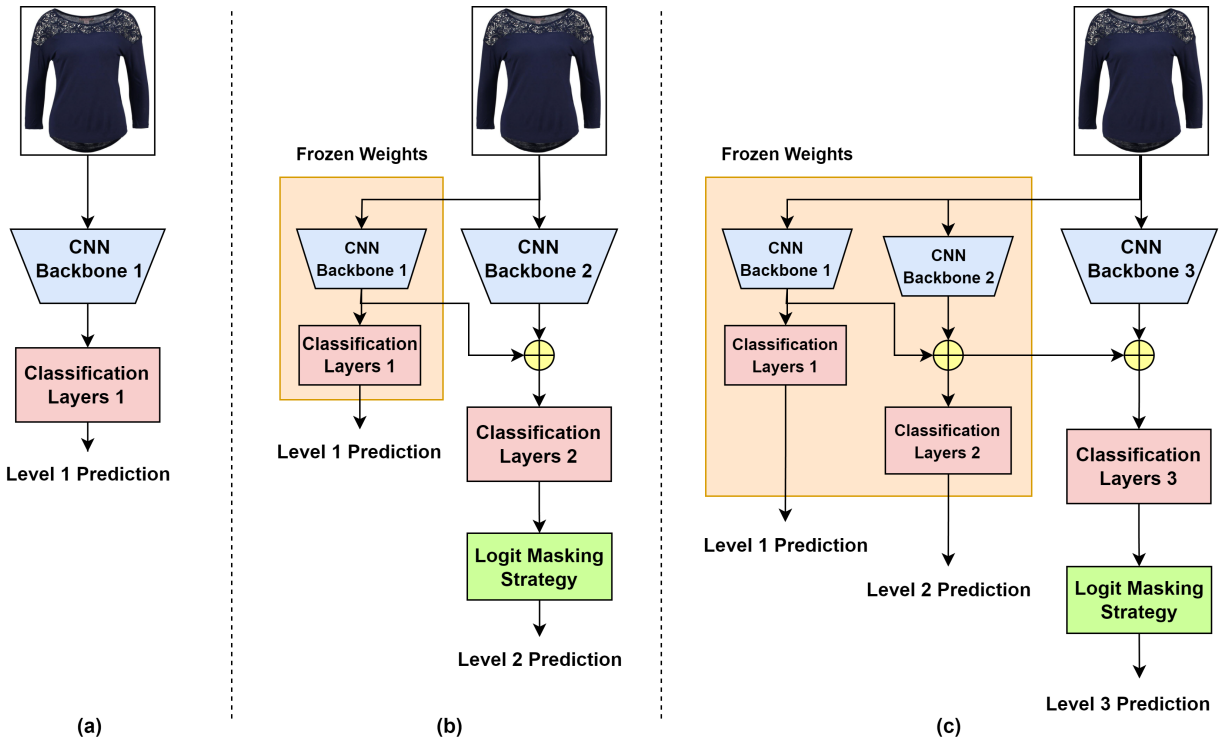

Fig. 2: PMTL Framework for a three-level hierarchical taxonomy classification. (a). Level-1 training. (b). Level-2 training which uses global description from frozen weights of level 1 model and logit masking strategy is employed. (c). Level-3 training using cumulative global representation of all previous levels and global masking strategy.

approach is adopted to reduce error propagation in subsequent levels. In the first level (root node training), the model is trained independently using cross-entropy loss. However, from the second level, we fuse the global representation of the predecessor model to the response of the model at that level. In this way, the global representation of the first level is added to the feature of second level before classification layers. This insertion enhances the representation in subsequent layers by providing global features. However, it does not directly participate in decision-making on choice of the model in a level, as opposed to the parent-node wise local classifier and hence do not propagate error. Also, it should be noted that the global representation changes as we go deep into the finer levels, contrary to global network based training approach, where same global representation is fed to all branches corresponding to coarse and fine levels. Further, while adding features from the predecessor, we freeze the weights of the predecessor model to prevent its weights aligning to new task using labels of a new level.

In our training framework, we hypothesize that the global representation plays a critical role in enhancing the representation of subsequent layers and hence its insertion improves the taxonomy classification performance. Also, we hypothesize that the global representation should change for different levels by incorporating more information from all predecessors. We validate our hypotheses during ablation study experiments (TP 1 and TP 3, respectively, in Table VII) for both full-shot and few-shot training method, where we observe that our method supports both hypotheses by giving improved performance than model without global representation (TP 1) and model with same global representation (TP 3).

B. Logit Masking Strategy

Although PMT utilizes progressive level-wise training and progressive enhancement of global representation for better classification, it does not address the label inconsistency problem. Traditionally, this problem is resolved by training separate model for each parent node and select one of the models based on the decision of its predecessor. However, it has two problems: 1) the results of the subsequent layers can go wrong if the predecessor gives incorrect response; and 2) training one model for each parent node is a time-consuming and resource-exhaustive operation.

To alleviate these, we propose logit masking strategy. Here, during training, we consider predecessor level ground truth as an input and mask a part of the logit and train the model with the remaining part. Hence, model weights are updated only using the relevant logits in the classification layer and irrelevant logits are masked and have no impact on training. Hence, we can make one model 'mimic' the behaviour of a set of models, each for a parent node. Also, we can keep one model at one level, thereby reducing error propagation and resource consumption while incorporating label consistency across the levels. In logit masking, we keep only the relevant

logits according to previous level ground truth and mask all irrelevant logits before loss computation. Assume that the previous level ground truth annotation GT_{prev} is a one-hot encoding vector of d classes and training level logit L is a vector of n classes. Then, it can be represented as:

$$GT_{prev} \in \{0,1\}^d : \sum_{i=1}^{d} GT_{prev}(i) = 1. \tag{1}$$

Assuming that the i^{th} class in previous level has n_i child nodes, total number of classes in the training level is $n = \sum_{i=1}^{d} n_i$. Using this the logit mask $Mask_{prev}$ for GT_{prev} can be represented as follows:

$$\begin{gathered} Mask_{prev} = [mask_1,\ mask_2,\ \ldots,\ mask_d] \\ mask_i = GT_{prev}(i) \times \mathbf{1}^{n_i}. \end{gathered} \tag{2}$$

Here, $\mathbf{1}^{n_i}$ represents a n_i-dimension vector of all ones. Here, $Mask_{prev} \in \{0,1\}^n$, where its values are one only when $GT_{prev}(i)$ is one, *i.e.,* the child nodes to the correct node in the previous level. This is then multiplied with the classification layer to make all irrelevant logit to zero and the relevant part to retain their values. After multiplication, this is used as the predicted logit vector for loss computation, thereby restricting the model to learn only using the relevant classes.

While incorporating the logit masking strategy, we hypothesize that different model for each class in training level is not needed since logit masking can mimic the behaviour. To validate this, we have run ablation study experiments by using multiple models in each level with or without global representation and using logit masking in full-shot and few-shot scenario (TP 2 and TP 4 in Table VII). The results validate our hypothesis that using multiple models for each level is not improving performance. Rather, the training loss converges very fast while testing result is less than our proposed method, signifying overfitting.

C. PMTL for Full Dataset

PMTL is designed to perform hierarchical taxonomy classification irrespective of dataset size. To analyze this, we have used three full large-scale datasets with varied complexity and challenges [1], [7], [16]. For experiment, we have considered ImageNet-pretrained Resnet-18 [3] model as a backbone for all levels. The global representation in subsequent levels are extracted from the global average pooling layer of Resnet-18 model of predecessor level. In classification layers, we have used two dense layers of size 128 and number of classes, respectively.

D. PMTL for Few-shot Dataset

The crucial part of few-shot learning is the training scheme, which is able to generalize well with very less number of instances per class. We observe the performance of PMTL in Prototypical network [11], which is regarded as a standard few-shot learning method. Contrary to Resnet-18, prototypical network produces embeddings for each image and we create the 'prototype' as the average embedding of all images from the same class in support set. The embedding of an unseen image is compared with the prototypes and the class corresponding to the prototype having minimum distance from the embedding of the unseen image is considered to be the predicted class. The output embedding of the model in previous level is considered as the global representation for the model in training level.

In few-shot learning using prototypical network [11], there is one modification in logit masking strategy, which sets it apart from full-shot training. In full-shot training, we seek the maximum similarity and hence mask irrelevant logits with zero. However, in prototypical network, we seek the minimum distance. Hence, we need to mask the irrelevant logit with a large value, preferably more than maximum value out of all distances from prototype. This can be done by replacing zeros to this high value after logit masking.

IV. Experimental Results

We evaluate the proposed method on two downstream applications: fashion taxonomy classification and similar item retrieval. We study the effectiveness of our training and loss computation strategy by performing extensive ablation experiments. Further, we compare the class-specific performance of our proposed method with the state-of-the-art. All experiments are performed in full-shot and few-shot setup for our proposed method to be considered a generalized framework for hierarchical multi-label classification.

A. Experimental Setup

Datasets. For fashion taxonomy classification, we have considered three datasets: DeepFashion [7], Shopping100k [1] and Fashion-MNIST [16]. For each dataset, we consider three-level hierarchy with each super-category having one or multiple sub-categories[1]. DeepFashion [7] provides fashion images worn by human models with variations in poses, occlusions and illuminations. For this work, we have considered the In-Store subset of the dataset and perform taxonomy classifications using three levels: gender (male, female), clothing type (upper-wear, bottom-wear, full-body and outer-wear) and product category (shirt, trouser, *etc.*). We use query subset as testing image for taxonomy classification and gallery subset as retrieval gallery for similar item retrieval. Shopping100k [1] provides fashion images with background having large variations in style. Here, we have considered similar levels as in DeepFashion. Fashion-MNIST [16] provides images with smaller resolution and less variations. We create a three-level taxonomy with level 1 having two classes (clothing, non-clothing), 6 classes for level 2 (top-wear, bottom-wear, outer-wear, one-piece, shoes, accessories) and 10 classes in level 3 as per annotations.

Baselines and performance metrics. To compare our proposed method for both applications, we have used four state-of-the-art methods: HMCNF-augmented hierarchical classifier (HierC) [2], Add-Net [4], Concat-Net [4], H-CNN [10] and recently proposed HiMulConE [18] and BA-CNN [8]. For

[1]A detailed explanation of these three levels for all datasets can be found at the link: https://tinyurl.com/p24pen3m

TABLE I: Hierarchical Taxonomy Classification Performance of DeepFashion Dataset [7] and Comparison with the state-of-the-art.

Method	Train data	Test data	L1 Acc.	L2 Acc.	L3 Acc.
HierC [2]	Full	Full	78.21	41.62	11.47
HierC [2]	Few-shot	Full	83.53	40.37	10.88
Add-Net [4]	Full	Full	84.26	52.14	24.37
Add-Net [4]	Few-shot	Full	84.21	31.46	7.42
Concat-Net [4]	Full	Full	84.24	48.81	21.98
Concat-Net [4]	Few-shot	Full	83.72	18.41	6.82
H-CNN [10]	Full	Full	85.89	74.94	49.99
H-CNN [10]	Few-shot	Full	81.22	47.37	15.43
HiMulConE [18]	Full	Full	89.07	55.19	25.38
HiMulConE [18]	Few-shot	Full	75.91	47.39	10.96
BA-CNN [8]	Full	Full	96.42	82.82	54.29
BA-CNN [8]	Few-shot	Full	80.56	41.02	20.81
PMTL	**Full**	**Full**	**98.68**	**93.05**	**77.30**
PMTL	**Few-shot**	**Full**	**83.47**	**65.18**	**39.37**

TABLE II: Hierarchical Taxonomy Classification Performance of Shopping100k Dataset [1] and Comparison with the state-of-the-art.

Method	Train data	Test data	L1 Acc.	L2 Acc.	L3 Acc.
HierC [2]	Full	Full	55.47	21.53	20.33
HierC [2]	Few-shot	Full	61.72	19.21	6.24
Add-Net [4]	Full	Full	63.36	19.86	16.28
Add-Net [4]	Few-shot	Full	63.37	15.72	9.82
Concat-Net [4]	Full	Full	63.35	19.55	15.61
Concat-Net [4]	Few-shot	Full	63.37	18.31	5.51
H-CNN [10]	Full	Full	60.25	18.61	15.61
H-CNN [10]	Few-shot	Full	60.48	14.52	8.89
HiMulConE [18]	Full	Full	55.90	19.28	17.59
HiMulConE [18]	Few-shot	Full	57.11	14.20	9.01
BA-CNN [8]	Full	Full	63.36	21.17	10.43
BA-CNN [8]	Few-shot	Full	60.36	21.05	8.55
PMTL	**Full**	**Full**	**98.63**	**39.28**	**46.03**
PMTL	**Few-shot**	**Full**	**63.06**	**38.33**	**50.21**

TABLE III: Hierarchical Taxonomy Classification Performance of Fashion-MNIST Dataset [16] and Comparison with the state-of-the-art.

Method	Train data	Test data	L1 Acc.	L2 Acc.	L3 Acc.
HierC [2]	Full	Full	93.17	88.31	83.74
HierC [2]	Few-shot	Full	71.33	81.12	79.95
Add-Net [4]	Full	Full	95.38	89.38	85.08
Add-Net [4]	Few-shot	Full	61.31	45.04	33.64
Concat-Net [4]	Full	Full	95.43	88.82	84.85
Concat-Net [4]	Few-shot	Full	53.00	43.70	41.26
H-CNN [10]	Full	Full	98.89	95.22	92.49
H-CNN [10]	Few-shot	Full	73.99	57.00	48.36
HiMulConE [18]	Full	Full	99.46	86.05	88.32
HiMulConE [18]	Few-shot	Full	60.00	52.14	40.56
BA-CNN [8]	Full	Full	99.84	96.28	92.56
BA-CNN [8]	Few-shot	Full	95.46	66.79	47.92
PMTL	**Full**	**Full**	**99.88**	**96.44**	**96.53**
PMTL	**Few-shot**	**Full**	**99.09**	**87.26**	**86.38**

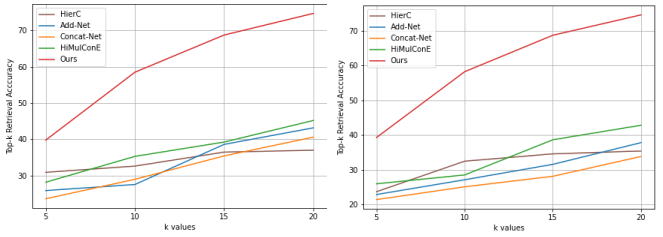

Fig. 3: Comparison for top-k retrieval accuracy for DeepFashion dataset [7]. From left to right: (a). Retrieval for Full dataset, (b). Retrieval for Part dataset.

taxonomy classification, we have considered accuracy as the metric while Top-k retrieval accuracy and Normalized Discounted Cumulative Gain (NDCG@k) [5] are used for similar item retrieval.

B. Fashion Taxonomy Classification

Using our proposed method, we perform fashion taxonomy classification for full-shot and few-shot datasets using all three datasets. Since Shopping100k [1] does not contain train-test split information, we have considered 60,000 images for training and 40,000 images for testing keeping similar image ratio in every class for partition. For few-shot training, we have taken 15 images per class in level 3 and have used 6 images in support set and 4 in query set for each task. For fair comparison, we have retrained all baseline models for all datasets using the dataset split same as the proposed method. To ensure consistency in backbone, ImageNet-pretrained Resnet-18 [3] is used as backbone for all baselines and proposed method for full-shot data. For few-shot experiments, we have used four CNN layer backbone, each having 64 filters of size (3,3) followed by batch normalization, ReLU activation and max pooling. For Fashion-MNIST [16], last two max pooling layers are removed to prevent the network reducing the receptive field below kernel dimension.

The results of the proposed method and the comparison with the state-of-the-art for DeepFashion dataset [7] are given in Table I. Here, we observe that our method significantly outperforms state-of-the-art methods for most of the cases, except for L1 accuracy for few-shot learning which gives comparable performance to other few-shot models. Also, we observe that the improvement in performance of our model is more significant in finer labels (*e.g.,* L2 and L3), since these levels require supervision from previous layers and explicit control on final labels. This shows the efficacy of the proposed training protocol and loss computation in a hierarchical setup irrespective of dataset size.

Similar trend is demonstrated in Table II, where we compare performance of our proposed method with state-of-the-art for Shopping100k dataset [1]. Here, L1 accuracy of proposed model in few-shot training is comparable to the state-of-the-art, however, improvements in performance for L2 and L3 are significant. Further, we perform experiments with Fashion-MNIST dataset [16] to observe the change in performance in datasets with less variations and low resolution. From Table III, we observe that performance improvement using our proposed method is substantial, especially for levels 2 and 3. Even, the few-shot framework using our method gives comparable performance to the full-shot results of the baseline methods and significant improvement can be seen from other few-shot methods.

These results for three datasets quantify the benefit of using our proposed method for hierarchical taxonomy classification in few-shot and full-shot scenario using datasets with various constraints, such as variations in human pose, object style, high inter-class similarity, *etc.*

TABLE IV: Comparison of similar item retrieval of PMTL with state-of-the-art using DeepFashion dataset [7].

Method	Full Data		Few-shot Data	
	Top-10 Acc	NDCG@10	Top-10 Acc	NDCG@10
HierC [2]	32.57	**73.72**	**32.46**	**72.70**
Add-Net [4]	27.49	68.24	27.12	66.89
Concat-Net [4]	28.92	70.53	25.09	65.16
HiMulConE [18]	**35.25**	71.86	28.51	69.45
PMTL	**58.48**	**75.03**	**58.25**	**74.57**

TABLE V: Comparison of class-specific classification accuracy for all levels using models trained with full DeepFashion dataset [7]. Here, top-2 results for each class are highlighted in bold.

Class	HierC	Add-Net	Concat-Net	HiMulConE	PMTL
Level 1 Classes					
Male	16.76	6.51	0.0	**48.37**	**93.58**
Female	94.57	99.58	**99.68**	99.36	**99.64**
Level 2 Classes					
Male Upper	4.18	11.93	0.0	**53.32**	**96.37**
Male Lower	5.29	**27.06**	1.03	0.0	**99.41**
Male Full-body	0.0	0.0	0.0	0.0	**46.67**
Male Outer	0.0	0.0	0.0	0.0	**21.43**
Female Upper	84.00	91.76	**97.38**	**98.63**	96.04
Female Lower	**91.36**	21.36	14.75	36.02	**99.20**
Female Full-body	0.0	0.0	0.0	0.0	**40.55**
Female Outer	0.25	**7.16**	0.0	0.0	**89.07**
Level 3 Classes					
Polos	0.0	**1.83**	0.0	0.0	**75.69**
Men's Hoodies	0.0	0.0	0.0	0.0	**79.70**
Men's Sweaters	0.0	**1.52**	0.0	0.0	**35.50**
Men's tees	0.36	15.68	0.0	**55.23**	**88.48**
Men's Denim	0.0	**18.35**	0.0	0.0	**34.86**
Men's Pants	1.29	**13.55**	0.0	0.0	**86.45**
Men's Shorts	6.13	**26.05**	0.0	0.0	**96.93**
Men's Jacket	**1.11**	0.0	0.0	0.0	**99.00**
Suit	0.0	0.0	0.0	0.0	**98.45**
Blouses	5.70	11.27	0.24	3.96	**79.42**
Cardigans	0.0	0.0	0.0	0.0	**40.70**
Women's Tees	0.30	40.71	**96.21**	**96.68**	66.57
Women's Sweater	0.0	0.0	0.0	**1.36**	**56.05**
Women's Hoodies	0.0	0.0	0.0	0.0	**41.00**
Graphic Tees	**5.6**	2.7	0.0	0.0	**71.51**
Women's Denim	**3.66**	0.0	0.0	0.0	**62.20**
Leggings	**82.35**	0.0	0.0	0.0	**45.86**
Women's Pants	2.00	2.00	0.0	**21.00**	**89.90**
Women's Shorts	3.64	**38.56**	29.76	21.76	**94.03**
Skirts	**5.21**	2.44	0.0	0.0	**90.23**
Women's Jacket	**7.30**	0.0	0.0	0.0	**98.00**
Dresses	11.00	**49.92**	0.21	0.0	**98.00**
Jumpsuit	**4.1**	0.0	0.0	0.0	**67.70**

C. Similar Item Retrieval

Similar item retrieval is an important application in the context of retail physical store and e-commerce setup, where the automatic self-help kiosks in physical stores and e-commerce websites retrieve similar items based on user's search. To facilitate this, traditional deep learning methods create embeddings of each image and retrieve products having minimum difference in embedding space. However, this involves finding similarity between millions of embeddings and hence is a time-consuming and resource-exhaustive process. To circumvent this, we aid fashion taxonomy for retrieval and then use embeddings for re-ranking after reducing the search space. The process happens in two phases, as given below:

Retrieval. To reduce the search space for visual similarity check, we retrieve all products from the retrieval gallery following same taxonomy as the query product. This results in a set of products having same taxonomy but a fraction of retrieval gallery in number.

Re-ranking. Although all the retrieved products follow same taxonomy and hence are "similar", some of them should be visually more similar than others and should be shown to the user before other products which are visually less similar. To facilitate this, we re-rank the order of the products based on their visual similarity computed using their embeddings. For our implementation, we use L_2 distance and retrieve k products with minimum distance.

In Table IV, we tabulate the retrieval results for DeepFashion using the proposed method and compare it with state-of-the-art. From the results, we observe that the proposed PMTL framework significantly outperformed all the existing methods for both full data and few-shot data training. Also, the performance is similar for full data and few-shot data, reinstating the generalizability of our approach. The trend is similar with the variations of number of retrieved products for few-shot and full dataset, as shown in Figure 3. Also, we observe that the large performance difference between the proposed method and other baselines. From this, we can conclude that our method consistently outperformed existing methods by a large margin irrespective of number of retrieved items and dataset size.

V. Analysis and Discussions

A. Class-specific Performance

Hierarchical fashion taxonomy classification deals with large number of classes and correct prediction of each of them is necessary for creating correct taxonomy for fashion products. Hence, class-specific performance is crucial to analyze rather than only checking overall performance. For this, we have compared class-specific performance of our method for DeepFashion dataset [7] for all levels with the state-of-the-art methods [2], [4], [18]. The comparison for full-shot and few-shot setups are given in Tables V and VI respectively. From these results, we observe that all baselines [2], [4], [18] mostly perform inferior to our proposed PMTL framework. Also, it should be noted that all these methods sometimes perform well for a class and perform very poorly for other classes in that level, *e.g.,* Concat-Net [4] gave wrong prediction for all 'Male' classes, however, giving 99.68% for 'Female' class in Table V. On contrary, the proposed method gives consistent performance across all categories. Therefore, it can be considered to be more reliable compared to other state-of-the-art methods.

B. Ablation Study Experiments

To analyze the impact of our training protocol and logit masking strategy, we have performed an extensive ablation study experiments. To showcase the improvement by our training protocol, we have considered four existing training protocols. For fair comparison, all experiments were conducted using same CNN backbone *i.e.,* ImageNet-pretrained Resnet-18 [3] for full-shot dataset and four *Conv* layer network for few-shot dataset. Each training protocol is further used with and without logit masking strategy. This entire set of

TABLE VI: Comparison of class-specific classification accuracy for all levels using models trained with few-shot DeepFashion dataset [7]. Here, top-2 results for each class are highlighted in bold.

Class	HierC	Add-Net	Concat-Net	HiMulConE	PMTL
Level 1 Classes					
Male	25.19	0.00	14.00	27.06	**77.75**
Female	**94.46**	**98.51**	92.92	89.91	84.54
Level 2 Classes					
Male Upper	8.36	0.0	**67.03**	0.48	**70.39**
Male Lower	**30.29**	0.0	0.0	0.0	**76.76**
Male Full-body	**10.00**	0.0	0.0	0.0	**17.78**
Male Outer	0.0	0.0	0.0	0.0	**35.71**
Female Upper	66.17	**99.80**	0.0	46.98	**75.45**
Female Lower	12.72	0.0	8.24	**37.48**	**72.44**
Female Full-body	4.40	0.0	**13.94**	0.0	**10.09**
Female Outer	**25.26**	0.0	0.0	5.49	**37.20**
Level 3 Classes					
Polos	**0.48**	0.46	0.0	0.0	**18.35**
Men's Hoodies	0.0	0.0	0.0	0.0	**50.00**
Men's Sweaters	**11.68**	0.0	0.0	0.0	**27.41**
Men's tees	0.12	0.0	0.0	**5.58**	**43.35**
Men's Denim	15.60	**33.94**	0.0	0.0	**36.70**
Men's Pants	0.32	0.0	**74.52**	0.0	**23.87**
Men's Shorts	**24.14**	0.0	0.0	0.0	**62.45**
Men's Jacket	**6.67**	4.44	0.0	0.0	**98.00**
Suit	0.0	0.0	0.0	0.0	**97.85**
Blouses	**28.99**	0.0	0.0	0.0	**24.26**
Cardigans	1.51	2.01	0.0	**29.4**	**22.86**
Women's Tees	0.0	0.0	0.0	0.0	**36.75**
Women's Sweater	**19.46**	0.0	0.0	0.0	**28.71**
Women's Hoodies	**10.88**	0.0	0.0	0.0	**33.89**
Graphic Tees	10.96	**81.10**	24.38	0.27	**37.81**
Women's Denim	0.0	0.0	0.0	0.0	**46.34**
Leggings	0.0	0.0	**1.47**	0.0	**33.82**
Women's Pants	0.0	0.0	0.0	0.0	**38.81**
Women's Shorts	**8.91**	0.0	0.0	0.0	**52.63**
Skirts	**4.40**	0.0	0.0	0.0	**28.18**
Women's Jacket	**6.06**	0.37	0.73	0.0	**98.00**
Dresses	23.62	0.50	0.0	**92.06**	**36.72**
Jumpsuit	**3.29**	0.0	0.0	0.0	**72.02**

Fig. 4: Visual examples of taxonomy classification and comparison with state-of-the-art. Here, first two examples correspond to models trained on full dataset and next two on few-shot dataset. In predicted result, 'red' color signifies wrong prediction and 'black' as correct prediction.

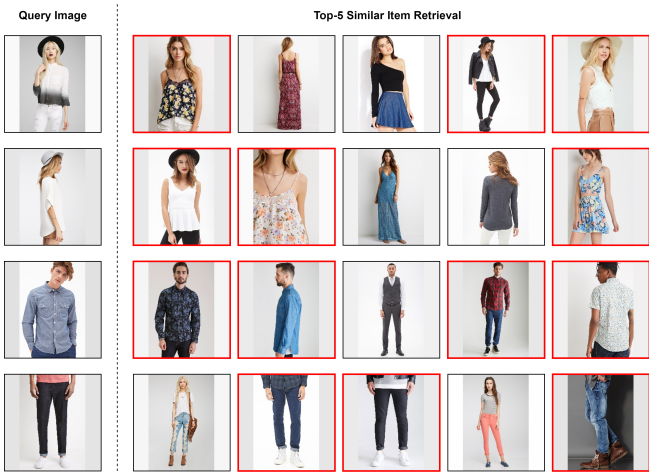

Fig. 5: Visual examples of Top-5 retrieval performance of proposed PMTL. Here, first column corresponds to query image and next 5 columns are retrieved items. Here, items with red boxes correspond to correct retrieval.

experiments are conducted for both full-shot and few-shot datasets. The training protocol used here are given below:

TABLE VII: Ablation Study Experiments for DeepFashion full- and few-shot datasets [7].

Training Protocol	Logit Mask	Full-shot Dataset			Few-shot Dataset		
		L1	L2	L3	L1	L2	L3
TP 1	No	98.68	92.17	70.62	83.47	60.51	32.09
TP 1	Yes	98.68	92.67	76.49	83.47	63.84	36.83
TP 2	No	98.68	91.56	68.79	83.47	59.25	32.16
TP 2	Yes	98.68	92.52	75.71	83.47	62.18	35.45
TP 3	No	98.50	90.96	67.73	72.86	53.86	24.15
TP 3	Yes	97.48	91.41	74.50	78.45	55.79	26.89
TP 4	No	98.68	91.60	68.51	83.47	60.89	31.25
TP 4	Yes	98.68	92.54	74.29	83.47	62.54	35.95
PMTL	**No**	**98.68**	**91.79**	**70.12**	**83.47**	**61.56**	**33.84**
PMTL	**Yes**	**98.68**	**93.05**	**77.30**	**83.47**	**65.18**	**39.37**

Training Protocol (TP) 1: In TP 1, we train one model for each level in a parallel manner, where the response of the model in one level does not depend on its predecessor. This mimics the level-wise training strategy of hierarchical classification.

Training Protocol (TP) 2: In TP 2, we train one model for each node for each level, but all models in a level are trained simultaneously. Here, the response of the model in one level does not depend on its predecessor. This mimics the parent node-wise training strategy of hierarchical classification, however, training models for all nodes in a level together reduces training time without sacrificing performance.

Training Protocol (TP) 3: Here, we train all levels together, where the features are extracted using a single backbone. The *Conv* features are then sent to three parallel classification branches with dense layers for three levels and losses are computed from them. This mimics the global classifier training strategy of hierarchical classification and thus is the simplest of all.

Training Protocol (TP) 4: Here, we train the models similar to TP 2, however, the feature embedding of parent node is added to the feature embedding of child node before classifi-

cation branch. The training of all models corresponding to all parent nodes in a level happen simultaneously. This merges the node-wise and level-wise training strategy of hierarchical classification, but it is computationally expensive.

The results of the ablation study experiments are given in Table VII. From the results, several observations can be made:
1) Training strategy involving one model per parent node is performing inferior to its counterparts having one model per level (*e.g.,* TP 1 performs better than TP 2, proposed method works better than TP 4). This is due to the overfitting problem which fastens the training loss convergence due to dedicated model for each node but the testing performance do not increase due to complex models for each node.
2) performance of the model improves by using logit masking strategy irrespective of training protocol and size of training dataset. Hence, logit masking strategy can be used for any hierarchical multi-label classification problem to alleviate the consistency issue in subsequent levels.
3) Proposed training strategy and logit masking gives improved performance for both full-shot and few-shot dataset and hence can be adopted as a go-to strategy for any hierarchical multi-label classification problem.

C. Qualitative Analysis

To analyze the ability of our method to obtain hierarchical labels, we examine classification performance using a set of images and compare the results with the state-of-the-art methods [2], [4], [18]. These visual results are given in Figure 4. Here, we have obtained results using models trained on full dataset for first two sample images and next two from models trained on few-shot dataset. From the results, we observe that existing models are not able to get all hierarchical label correctly due to label inconsistency and error propagation problems. For example, for third image, Concat-Net gives the taxonomy as *Women* $\rightarrow$ *Male top-wear* $\rightarrow$ *Pants*, which is an impossible set. Contrary to existing methods, our proposed PMTL performs better by giving correct prediction for all labels. Similar to this, we observe the results of retrieval performance of our proposed PMTL using DeepFashion dataset [7]. The results are given in Figure 5. Here, we have considered two male and two female products and we can observe that our approach is able to retrieve relevant products from the gallery.

VI. Conclusion and Future scope

In this paper, we have proposed PMTL, a generalized framework for progressive multi-level training with logit masking strategy for hierarchical taxonomy classification. The proposed approach merges facets of different training schemes by addressing three problems in hierarchical classification: error propagation, label inconsistency and generalized pipeline for full-shot and few-shot scenario. We have experimented with three datasets and our method has shown significant improvement from several state-of-the-art methods. In future, this work can be extended for more granularity, involving attributes and compatibility information. These information enables the model to identify multi-shelf taxonomy of an item which can be part of several true hierarchies using image data. Also, it can be augmented with customer behaviour to provide personalized display-level taxonomy.

References

[1] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Efficient multi-attribute similarity learning towards attribute-based fashion search. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1671–1679. IEEE, 2018.

[2] Hyunsoo Cho, Chaemin Ahn, Kang Min Yoo, Jinseok Seol, and Sanggoo Lee. Leveraging class hierarchy in fashion classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Matheus Inoue, Carlos Henrique Forster, and Antonio Carlos dos Santos. Semantic hierarchy-based convolutional neural networks for image classification. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[5] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[6] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[7] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.

[8] Iván Pizarro, Ricardo Nanculef, and Carlos Valle. An attention-based architecture for hierarchical classification with cnns. *IEEE Access*, 11:32972–32995, 2023.

[9] Nidhi Ranjan, Pranav Vinod Machingal, Sunil Sri Datta Jammalmadka, Veena Thenaknidiyoor, and AD Dileep. Hierarchical approach for breast cancer histopathology images classification. 2022.

[10] Yian Seo and Kyung-shik Shin. Hierarchical convolutional neural networks for fashion image classification. *Expert systems with applications*, 116:328–339, 2019.

[11] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[12] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

[13] Anne-Marie Tousch, Stéphane Herbin, and Jean-Yves Audibert. Semantic hierarchies for image annotation: A survey. *Pattern Recognition*, 45(1):333–345, 2012.

[14] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[15] Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. Hierarchical multi-label classification networks. In *International conference on machine learning*, pages 5075–5084. PMLR, 2018.

[16] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[17] Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. Hd-cnn: hierarchical deep convolutional neural networks for large scale visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2740–2748, 2015.

[18] Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. Use all the labels: A hierarchical multi-label contrastive learning framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16660–16669, 2022.

[19] Xinqi Zhu and Michael Bain. B-cnn: branch convolutional neural network for hierarchical classification. *arXiv preprint arXiv:1709.09890*, 2017.