Few-Shot Multi-Label Multi-Class Continuous Learning for Dark Web Image Categorization

Yağmur Çiğdem Aktaş[1], Mikel Aramburu[2], and Jorge García Castaño[3]

[1]Vicomtech Foundation, Mikeletegi 57, 20009 Donostia-San Sebastián (Spain)

ABSTRACT

Categorizing dark web image content is critical for identifying and averting potential threats. However, this remains a challenge due to the nature of the data, which includes multiple co-existing domains and intra-class variations, as well as continuously having newer classes due to the rapidly augmenting amount of criminals in Darkweb. While many methods have been proposed to classify this image content, multi-label multi-class continuous learning classification remains under explored. In this paper, we propose a novel and efficient strategy for transforming a zero-shot single-label classifier into a few-shot multi-label classifier. This approach combines a label empowering methodology with few-shot data. We use CLIP, a conservative learning model that uses image-text pairs, to demonstrate the effectiveness of our strategy. Furthermore, we demonstrate the most appropriate continuous learning methodology to overcome with the challenges of accessing old data and training over and over again for each newly added class. Finally, we compare the performance with multi-label methodologies applied to CLIP, leading multi-label methods and the continuous learning approaches.

Keywords: Multi-label image classification, Multi-class image classification, CLIP, Label empowering, Continuous learning

1. INTRODUCTION

The dark web is a specific subset of the vast global network, accessible only through specialized web browsers, which is often a hotbed for illegal or illicit activities. The principal traits are the provision of anonymity, offering users a level of privacy not often found on the traditional internet, and untraceability, meaning that the origins and destinations of data are extremely difficult to track and identify. Thus, the dark web is a valuable source of threat intelligence that investigators can use to identify cyber-attacks, stolen goods, illegal substances, arms trafficking, confidential information, child exploitation, or violent acts.

Understanding automatically the visual content of the dark web, i.e. images or videos, is crucial for efficient image categorization in different domains enabling the identification and prevention of potential threats. Dark web image categorization remains still an open challenge due to the nature of the visual content. Most of the images present blur and low resolution including small object sizes or objects with similar colors to the background. However, the most challenging issues are the multiple co-existing domains, which can be overlapped or semi-occluded, and intra-class variations, where a single domain can have significant visual differences as can be seen in Fig. 1. This fact highlights the importance of multi-label classification, which aims to categorize an image not only with a single label but into more than one label.

Most recent multi-classification approaches rely on dividing the multi-label problem into separate single-label problems,[1,2] optimizing the loss or activation function,[3,4] using the embeddings learned from label graph-based networks to discover the locations of discriminative features,[5,6] or adapting transformers to create a brand-new architecture specifically for multi-label classification.[7] While these approaches provide a suitable performance, they require a large well-balanced dataset, which is a complex situation that typically arises in real-world applications. The cost of expanding the dataset to a larger class vocabulary with balanced data is not linear but exponential.[8]

Recently, the combination of computer vision and natural language processing has witnessed significant advancements in image classification.[9] Although prompt-based zero-shot transfer learning showed promising

yaktas@vicomtech.org, maramburu@vicomtech.org, jgarciac@vicomtech.org

Figure 1: Multi-label image examples from Dark web dataset

performances, it is not enough for specific use cases such as dark web visual data. To avoid this issue, different works have proposed to utilize few-shot data to improve the classification capability.[10] However, these approaches have been specifically designed for single-label classification. We extend these classification capabilities to address multi-label multi-class classification. Our approach is light and simple, yet effective. It is based on the adaption of a label-empowering multi-label methodology by incorporating the use of embeddings for images and text.

Another challenge that comes with dark web dataset is that new classes, i.e. new vendors, new products to categorize and track occurs way more often than a usual classification problem. This brings us the importance of a continuous learning pipeline along with this multi-label few shot classification problem to no train the model all the time from scratch, also to not be obligated to preserve the old data which quickly becomes bigger and bigger .

In this work, we propose a novel efficient methodology to convert a zero-shot single-label classifier into a few-shot multi-label classifier using Label Empowering methodology in combination with few-shot data. For this purpose, our contributions are as follows;

- We adapt the label-empowering multi-label methodology to a zero-shot classifier providing a portable pipeline for any real-world dataset that needs few-shot multi-label classification.
- We demonstrate the effectiveness of our approach by implementing the proposed methodology using CLIP[9] on a dark web multi-label image classification dataset.
- We compare the performance of our approach with other multi-label methodologies applied on CLIP, as well as other state-of-the-art multi-label architectures.
- We adapt a continuous learning solution to our portable few-shot multi-label classification and demonstrate the efficient results.

2. RELATED WORK

Numerous methods have been proposed to solve the multi-label classification problem, which can be divided into (i) problem transformation and (ii) algorithm adaptation methods.

Early works focus on decomposing the multi-label task into several independent binary class tasks[1] which suffer from the correlation information between labels. Also, training an individual classifier for each class brings a heavy computational cost. Some approaches involve linking these binary classifiers together in a chain structure,[11] such that the prediction of a previous binary classifier becomes a feature for other classifiers. Even though this method better covers the correlation between classes, it still has the same computational cost problem due to having a classifier for each class. Another method[2] converts the labels into binary variables and transforms the multi-label multi-class problem into a multi-class single-label problem by considering each unique label combination as a separate class.

Some other works contribute to multi-label classification by optimizing the loss function[3] or the activation function[4] to tackle the challenges of multi-label classification.

More focused on computer vision, multiple approaches have been proposed for multi-label image classification as well.[12] propose a Hypotheses-CNN-Pooling (HCP) framework that generates a large number of proposals by objectness detection methods without requiring bounding box annotations and treats each proposal as a single-label image recognition problem. As a variation of CNN-based methods, Graph Convolutional Networks (GCNs) have been demonstrated as having remarkable capacity in multiple vision tasks. In the range of multi-label classification tasks, GCN[5,6,13] trains classifiers by mapping the correlation between the labels into graph networks. Some approaches follow the objectness detection method but instead of hypothesis extraction, they adapt the object detection model approaches like,[14] where the proposed method is training a multi-label classifier with knowledge distillation by an object detector.

Lastly, transformers[15] was created to examine long-range dependencies among word embeddings in NLP[16–18] which later was discovered to have strong capacity in various computer vision tasks like image classification[7,9] and object detection.[19] Following these improvements,[20] proposes a Multi-label Transformer architecture constructed pixel and cross-window attention whereas[21] leverages transformer decoders to query the existence of a class label. With the increasing interest in transformers, various works are combining the mentioned approaches to create a multi-label classifier.[22] combines graph convolution networks with attention, while[23] applies metric learning with attention to investigate the similarity of the labels.[24] proposes a model producing attention maps for each label and capturing the underlying connections between them using convolutions.

Even though there are many inspiring works for multi-label classification, they still require a huge amount of data to obtain high accuracy. However, it is a common issue to not have a big enough custom dataset, as well as it is time-consuming. Open-vocabulary large language models trained on the enormous size of datasets to learn how to extract and match visual and text embeddings like,[9] provides promising results with their strong capacity to be zero-shot single-label image classifiers and few-shot multi-label classifiers via some improvements.

Continuous learning is divided into three categories according to the problem: domain, task and class incremental. In domain incremental, the output nodes remain unchanged, whereas the input type differs for each incremental train. In task incremental, the output node amount stays unchanged, whereas in each incremental train, another type of classes are added along with task id. In contrary of these types, class incremental refers to add new class(es) to the architecture for each incremental step. Fig. 2 shows the different types of continous learning.

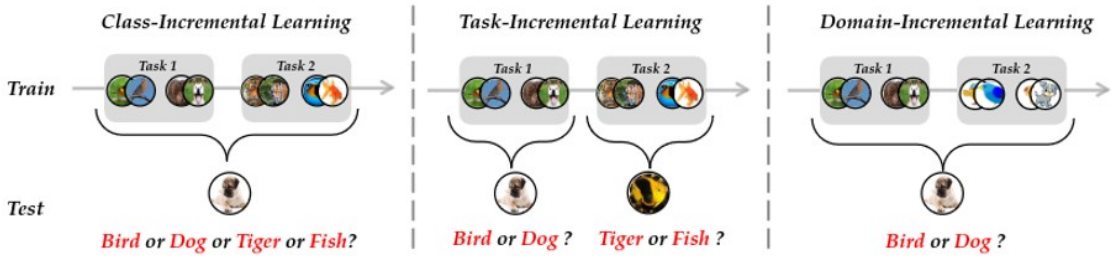


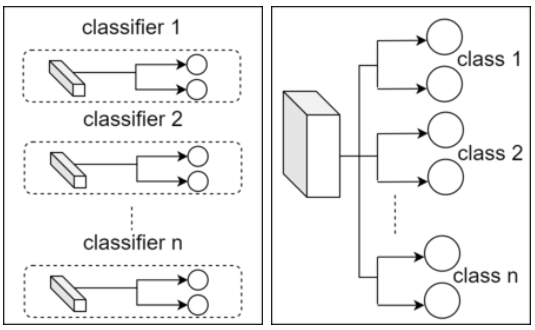
Figure 2: Continuous learning types: class, task and domain incremental

Continuous learning methods can be divided in 4 category: Distillation based, adaptive-plasticity-based, data-centric based, architecture based. Different methodologies belonging any of these categories can serve for one of the mentioned 3 different types of continuous learning tasks. [25–27] applies distillation between the latest trained model and the new one containing the additional task. [28,29] are the ones using replay data, to remind the features of the old tasks during the new train using only a few sample. [30,31] are using techniques like Fisher matrix to analyse the importance of the parameters of the models and decide which ones should be preserved during the new training, to prevent forgetting old task. [32] adapts the architecture for each new task by expanding the backbone, while [33] applies also distillation.

Being class-incremental is the most challenge full one between three mentioned categories of continuous learning, we provide knowledge distillation based, class-incremental multi-label few-shot classification pipeline, without the need of replaying old data, or expanding the backbone for each train.

3. MULTI-LABEL CLASSIFICATION METHODOLOGIES

There are three multi-label classification methods that apply to image classification: (i) binary relevance, (ii) multi-task binary heads and (iii) label empowering.



Label Powerset					
	class 1	class 2	class 3	class n	Powered Label
Sample 1	1	0	0	0	a
Sample 2	0	0	1	0	b
Sample 3	1	1	0	0	c
Sample 4	1	0	0	0	a
Smaple 5	1	0	0	1	d

Figure 3: Multi-label methodologies: (a) Binary relevance; (b) Multi-task Binary Heads; (c) Label empowering.

3.1 Binary Relevance

Binary relevance method,[1] also named One-vs-One, includes as many binary classifiers as the number of classes. Each binary classifier is composed of the backbone and the classifier head and individually trained for each class, in which one node corresponds to the existence of the class and the second one is non-existence. Therefore images containing the related class become positive samples and the rest of the images become negative samples. At inference time, the image is passed through all classifiers one by one. The multi-label result is obtained by combining the predicted classes. The classes whose classifier gives a higher probability than a given threshold. Figure ?? shows the architecture of a binary relevance model created for n classes. As it can be noted, binary relevance is not an efficient solution for both training and inference time, especially for datasets having a large number of classes. It requires individual training for each classifier and all of them should be deployed for inference, which causes time and memory inefficiency.

3.2 Multi-task Binary Heads

In contrast to binary relevance, the multi-task binary heads method[34] presents a common backbone for all binary classifiers. Therefore, each binary head is connected to the same backbone having a unified architecture. The configuration of positive and negative samples remains equal. Figure ?? shows the architecture of multi-task binary heads created for n classes. Although this method improves both time and memory efficiency in comparison to binary relevance, our experimental results show that it can not preserve its accuracy while the number of classes increases.

3.3 Label Empowering

While both binary relevance and multi-task binary heads methods convert the multi-class problem into separate binary class tasks, the label empowering method[2] translates multi-label classification into traditional single-label but still a multi-class classification task. This method converts the multi-labels to binary vectors by hot-encoding them and assigning a unique label for each different binary vector. Figure ?? shows a dataset example having 5 input images and 3 classes, where each image's multi-label ground-truths are hot encoded, and for every unique combination, a unique label is assigned to create the final label powerset having 4 single-label. While the first two methodologies have various shortcomings, label empowering hinges on increasing the number of classes to be learned. This transforms the problem into a single-label classification approach, which is easily affordable.

4. APPROACH

Fig. 4 shows the proposed multi-label methodology which is divided into three parts: (i) label curation by implementing the label empowering multi-label classification method; (ii) model training by implementing a few-shot fine-tunning technique; and (iii) the deployment of the model by defining a multi-label inference process.

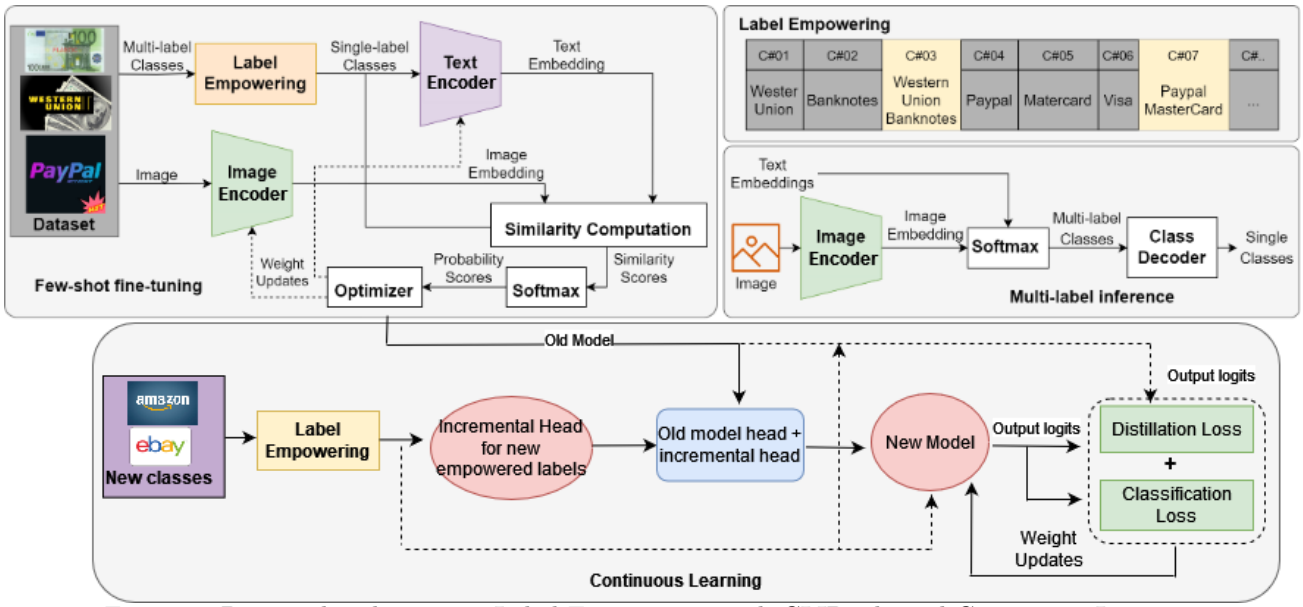


Figure 4: Proposed architecture: Label Empowering with CLIP adapted Continuous Learning

4.1 Label empowering implementation

We formalize the multi-label classification task as follows. $L = y_j : 1...l$ represents a finite set of labels, $D = \{(X_i, Y_i), i = 1...n\}$ represents a set of examples, where X_i is a feature vector and $Y_i \subseteq L$ the set of labels assigned to the ith example. The label empowering method transforms each Y_i into a single and unique label. If $Y1 = \{y_1, y_3\}$ and $Y2 = \{y_1, y_2, y_3\}$, single-labels representing these subsets can be denoted as $y_{1,3}$ and $y_{1,2,3}$, respectively. After the label empowering transformation, the transformed set of examples is represented as $D = \{(X_1, y_{LP1}), (X_2, y_{LP2}), ..., (X_n, y_{LPn})\}$, with $L_{LP} = y_{LP1}, y_{LP2}, ..., y_{LPm}$ the new set of target labels and m is the number of unique subsets of labels present in the D.

4.2 Few-shot fine-tunning

We fine-tune CLIP image and text encoders appending linear layers as the classifier head on top of CLIP, containing as many nodes as our empowered label amount. Using the Softmax function, we calculate the probability of similarities between image and text embeddings and accept the head node having the highest probability as the model prediction while using Cross Entropy Loss for loss calculation and backpropagation.

Therefore we convert the open-vocabulary CLIP classifier into a specific multi-label classifier for our dark web dataset using only a few amount of samples per class.

4.3 Multi-label inference

For inference, we use the text embedding generated for each class during the training phase. This, along with the image embedding extracted from the input image, is used to compute the similarity score and select the empowered single class. Lastly, a class decoder function is used to convert this single class into multiple labels.

4.4 Knowledge distillation for incremental tasks

For each new class we need to add to our classifier, we distill the output logits from the latest fine-tuned model and add a distillation loss additionally to the classification loss, representing the difference between the logits of old and new model. To achieve this goal, we consider the predictions of the old model as truth label and send them to cross-entropy loss as in classification loss. Furthermore we merge both losses to obtain the final loss for back propagation.

$$Loss_{classification} = CE(logits, labels \tag{1}$$

$$Loss_{distillation} = CE(logits, old\_model\_logits) \tag{2}$$

To prevent forgetting old classes in a more efficient way, we merge the weights of the old and new model at the end of each incremental task with a coefficient $\alpha = 0.5$,

$$incremented\_model = \alpha \times new\_model + (1 - \alpha) \times old\_model \tag{3}$$

5. EXPERIMENTAL RESULTS

5.1 Dark web dataset

The Dark web dataset, sourced from CFLW's Dark Web Monitor*, includes images collected from dark web domains about various crime categories like financial crime and organizations, drugs and narcotics, weapons as well as their vendors as individual classes. Similar to many real-world datasets, the dark web dataset contains images with multiple labels, meaning an image can belong to more than one class. It also includes images with a single label. Figure 1 shows some image samples along with their ground truth labels from various classes.

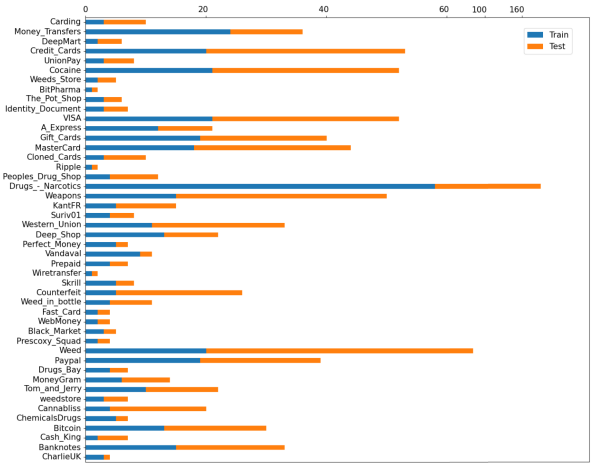


Figure 5: Data distribution of the large dark web dataset for both train and test sets.

The dataset contains two different subsets: (i) tiny dark web financial dataset; and (ii) large dark web dataset. The former contains only 5 highly correlated classes. These classes are Paypal, VISA, American Express, Credit Cards, and MasterCard. The latter contains 46 classes from various categories, not only financial, like Drugs, Narcotics, Weapons, and Financial Organizations which do not have a strong correlation between them but it includes subcategories having a strong relationship and the possibility of being exist in the same image. Figure 5 shows the data distribution of the large dark web dataset for both train and test sets.

The large dark web dataset presents an imbalanced data problem. Some dominant classes have many samples, whereas some other classes suffer from a lack of data. Some classes like drugs and narcotics, coming along with any type of drug type or any individual vendor that has to be categorized, become a very dominant class by

*https://cflw.com/dwm/

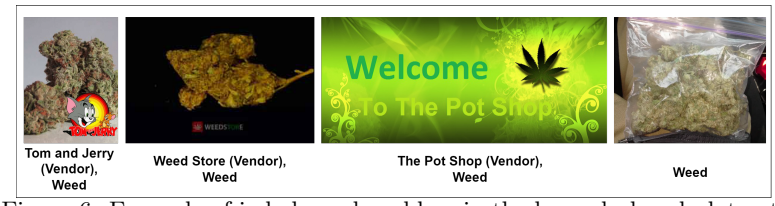


Figure 6: Example of imbalanced problem in the large dark web dataset.

collecting only a few samples of its subgroups. Figure 6 shows 4 image samples from the large dark web dataset related to the "weed" class. It can be noted that this class can occur as a single instance. Furthermore, it becomes a dominant parent class for other image samples having vendor information. This issue creates an imbalance problem with one label for each vendor and 4 labels for the "weed" class. Lastly, another challenge is that some classes, especially the vendor names like people drug stores, or vandaval shop, have textual information rather than visual features which brings the need for considering the combination of vision and language data.

5.2 Metrics

We use standard classification metrics: precision, recall and f1-score; to compare the performance of multi-label multi-class classification as shown in equations 4, 5 and 6, respectively.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1\_score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

Since these metrics cannot be directly applied to multi-label classification, we convert the predicted powered labels back to single labels. Then, we use these single labels to perform the calculation in the usual way. Figure 7 shows an image sample having three ground truth labels: VISA, AmericanExpress, and MasterCard. In case the predicted class is "VISA, Paypal" class, after extracting the single labels as "VISA" and "Paypal", this prediction would contribute as a true positive for "VISA" class, false positive for "Paypal" class and a false negative for "MasterCard" class and "AmericanExpress" classes.

Figure 7: An example multi-labeled image with 3 classes.

5.3 Implementation Details

The tiny dark web financial dataset includes 20 image samples for training and 85 samples for testing, where both splits are balanced. The large dark web dataset includes 272 samples for training and 593 samples for testing without balancing. We follow slightly different training strategies for each multi-label classification methodology, according to their underfitting and overfitting ranges. In this essence, each model belonging to the binary relevance method is trained 3 epochs taking around 0.76 seconds for both datasets, multi-task binary heads are trained 11 epochs for the tiny dataset and 15 epochs for the large dataset, each taking around 1 min and 21 seconds and 4 min and 10 seconds. Finally, the tiny dataset has been trained for 5 epochs for the label

empowering method, taking around 1 minute and 34 seconds for 14 labels constructed from 5 base classes and the large dataset has been trained for 10 epochs taking around 2 min and 26 seconds for 98 labels constructed from 42 base classes.

5.4 Ablation Study

We compare the results of our few-shot multi-label classifier using label empowering methodology against the binary relevance and multi-task binary heads methodologies. To investigate the effect of the asymmetric loss, being the most relevant state-of-the-art loss function for imbalanced datasets, we conducted an additional experiment by adapting this loss to the label empowering method. Table 1 shows the comparison of the multi-label classification methodologies on the tiny dark web financial dataset. The label-empowering methods outperform the binary relevance and multi-task binary heads methodologies. In particular, the label empowering method in combination with the asymmetric loss achieves the best metric results.

Table 1: Comparison of binary relevance (BR), multi-task binary heads (MTH), and label empowering methods (LE and LE + ASL) on the tiny dark web financial dataset, extracting 14 labels as the final powerset from 5 classes.

Method	Precision	Recall	F1_score	Model Size
BR	0.986	0.648	0.742	453.8 x 5
MTH	0.907	0.906	0.904	453.9
LE	1.0	0.908	0.947	453.9
LE + ASL	1.0	0.912	0.953	453.9

Table 2: Comparison of binary relevance (BR), multi-task binary heads (MTH), and label empowering methods (LE and LE+ASL) on the large dark web dataset, extracting 98 labels as the final powerset from 42 classes.

Method	Precision	Recall	F1_score	Model Size
BR	0.772	0.621	0.659	453.8 x 42
MTH	0.67	0.54	0.6	453.9
LE	0.936	0.936	0.935	454.1
LE + ASL	0.936	0.940	0.937	454.1

Similarly, Table 2 shows the comparison of the multi-label classification methodologies on the large dark web dataset. The metrics results reveal a similar behavior even with the large dark web dataset where the labels are not balanced. This demonstrates that performance does not change regardless of whether the data is balanced or not. Additionally, we observe that the number of classes also has no effect.

5.5 Comparison with State-of-the-art

We compare our approach in Table 3 with two state-of-the-art methods: (i) a zero-shot image classification approach[9] and (ii) a few-shot image classification[10] on the large dark web dataset. Results demonstrate the effectiveness of our approach since it outperforms both state-of-the-art methods. In particular, our method achieves 93.6% and 94.0% for precision and recall, respectively.

5.6 Continuous Learning Adaptation and Results

To create and evaluate our continuous learning pipeline, we start with 44 initial classes and add 1 more class at each incremental step, reaching to the 52 classes at the end.

Table 3: Comparison with state-of-the-art methods

Method	Precision	Recall	F1_score
CLIP	0.847	0.862	0.854
Tip-Adapter	0.901	0.916	0.909
LE + ASL (Ours)	**0.936**	**0.940**	**0.937**

We compare our pipeline, including knowledge distillation and weight merging at the end of the training, with other methods like weight ensembling, means to apply the weight merge not at the end of the training but after each iteration I, distillating from the zero-shot CLIP model instead of the latest fine-tuned one, adding weight consolation loss, refering to add the difference of the teacher and student model's parameters as the third loss.

Table 4 shows us the comparison between the mentioned methods and our approach. First refers to the first fine-tuned multi-label model accuracy, Average refers to the average accuracy of incremental training from 44 to 52 classes, Last means the final model's accuracy with 52 classes and δ shows the forgetting amount of old classes at the end of continuous learning to reach 52 classes.

Table 4: Comparison of continuous learning approaches adapted into our multi-label few shot pipeline. Incremental: default incremental learning without any additional method, DIST Incremental: Knowledge distillation based Incremental learning, WF: weight merge at the end of training using $\alpha = 0.5$, WE: weight merge after each iteration I, WC: weight consolation loss

Method	First	Average	Last	δ
Incremental	0.912	0.901	0.881	0.031
DIST Incremental from zero-shot	0.820	0.803	0.792	0.028
DIST Incremental	0.925	0.905	0.897	0.028
DIST Incremental+WF (Ours)	0.925	0.912	0.910	**0.015**
DIST Incremental+WE	0.925	0.907	0.895	0.030
DIST Incremental+WC	0.930	0.890	0.880	0.040

6. CONCLUSION AND FUTURE WORKS

We have proposed a novel efficient methodology to convert a zero-shot single-label classifier into a few-shot multi-label classifier using label-empowering methodology in combination with few-shot data. We have specifically adapted distillation based continuous learning to this portable pipeline for any real-world dataset that needs a few-shot multi-label classification. We have demonstrated the effectiveness of our approach by implementing the proposed methodology on a dark web multi-label image classification dataset. Finally, we have compared the performance of our approach with other multi-label methodologies, other state-of-the-art multi-label architectures as well as continuous learning approaches. The present study has provided valuable insights into the multi-label, continuous few-shot classification research area, yet there are several avenues for future research warrant exploration. Our main focus on future work would be discovering new continuous learning methodologies and adapting to our pipeline to make the retraining process even more efficient.

Acknowledgements

The work described in this paper is performed in the H2020 project STARLIGHT ("Sustainable Autonomy and Resilience for LEAs using AI against High Priority Threats"). This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101021797.

REFERENCES

[1] Zhang, M.-L., Li, Y.-K., Liu, X.-Y., and Geng, X., "Binary relevance for multi-label learning: an overview," *Frontiers of Computer Science* **12**, 191–202 (Mar. 2018).

[2] Tsoumakas, G., Dimou, A., Spyromitros-Xioufis, E., Mezaris, V., Kompatsiaris, I., and Vlahavas, I., "Correlation-based pruning of stacked binary relevance models for multi-label learning," *Proceedings of the 1st International Workshop on Learning from Multi-Label Data* , 101–116 (01 2009).

[3] Ben-Baruch, E., Ridnik, T., Zamir, N., Noy, A., Friedman, I., Protter, M., and Zelnik-Manor, L., "Asymmetric loss for multi-label classification," (2021).

[4] Martins, A. F. T. and Astudillo, R. F., "From softmax to sparsemax: A sparse model of attention and multi-label classification," (2016).

[5] Wang, Y., He, D., Li, F., Long, X., Zhou, Z., Ma, J., and Wen, S., "Multi-label classification with label graph superimposing," (2019).

[6] Chen, Z.-M., Wei, X.-S., Wang, P., and Guo, Y., "Multi-label image recognition with graph convolutional networks," (2019).

[7] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H., "Training data-efficient image transformers distillation through attention," (2021).

[8] Zhang, W., Liu, C., Zeng, L., Ooi, B., Tang, S., and Zhuang, Y., "Learning in imperfect environment: Multi-label classification with long-tailed distribution and partial labels," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*], 1423–1432 (October 2023).

[9] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I., "Learning transferable visual models from natural language supervision," (2021).

[10] Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., and Li, H., "Tip-adapter: Training-free adaption of clip for few-shot classification," in [*Computer Vision – ECCV 2022*], Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T., eds., 493–510, Springer Nature Switzerland, Cham (2022).

[11] Read, J., Pfahringer, B., Holmes, G., and Frank, E., "Classifier chains: A review and perspectives," *Journal of Artificial Intelligence Research* **70**, 683–718 (Feb. 2021).

[12] Wei, Y., Xia, W., Lin, M., Huang, J., Ni, B., Dong, J., Zhao, Y., and Yan, S., "Hcp: A flexible cnn framework for multi-label image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38**, 1901–1907 (Sept. 2016).

[13] Chen, T., Xu, M., Hui, X., Wu, H., and Lin, L., "Learning semantic-specific graph representation for multi-label image recognition," (2019).

[14] Liu, Y., Sheng, L., Shao, J., Yan, J., Xiang, S., and Pan, C., "Multi-label image classification via knowledge distillation from weakly-supervised detection," in [*Proceedings of the 26th ACM international conference on Multimedia*], ACM (Oct. 2018).

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I., "Attention is all you need," (2023).

[16] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., "Bert: Pre-training of deep bidirectional transformers for language understanding," (2019).

[17] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I., "Language models are unsupervised multitask learners," (2019).

[18] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D., "Language models are few-shot learners," (2020).

[19] Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J., "Deformable detr: Deformable transformers for end-to-end object detection," (2021).

[20] Cheng, X., Lin, H., Wu, X., Yang, F., Shen, D., Wang, Z., Shi, N., and Liu, H., "Mltr: Multi-label classification with transformer," (2021).

[21] Liu, S., Zhang, L., Yang, X., Su, H., and Zhu, J., "Query2label: A simple transformer way to multi-label classification," (2021).

[22] Ye, J., He, J., Peng, X., Wu, W., and Qiao, Y., "Attention-driven dynamic graph convolutional network for multi-label image recognition," (2020).

[23] Prokofiev, K. and Sovrasov, V., "Combining metric learning and attention heads for accurate and efficient multilabel image classification," (2022).

[24] Zhu, F., Li, H., Ouyang, W., Yu, N., and Wang, X., "Learning spatial regularization with image-level supervisions for multi-label image classification," (2017).

[25] Li, Z. and Hoiem, D., "Learning without forgetting," (2017).

[26] Dhar, P., Singh, R. V., Peng, K.-C., Wu, Z., and Chellappa, R., "Learning without memorizing," (2019).

[27] Douillard, A., Cord, M., Ollion, C., Robert, T., and Valle, E., "Podnet: Pooled outputs distillation for small-tasks incremental learning," (2020).

[28] Rebuffi, S.-A., Kolesnikov, A., Sperl, G., and Lampert, C. H., "icarl: Incremental classifier and representation learning," (2017).

[29] Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., and Fu, Y., "Large scale incremental learning," (2019).

[30] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R., "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences* **114**, 3521–3526 (Mar. 2017).

[31] Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., and Tuytelaars, T., "Memory aware synapses: Learning what (not) to forget," (2018).

[32] Yan, S., Xie, J., and He, X., "Der: Dynamically expandable representation for class incremental learning," (2021).

[33] Wang, F.-Y., Zhou, D.-W., Ye, H.-J., and Zhan, D.-C., "Foster: Feature boosting and compression for class-incremental learning," (2022).

[34] Zhang, Y. and Yang, Q., "A survey on multi-task learning," (2021).