CLIP-based Few-Shot Multi-Label Classification Methods: A Comparative Study

Yağmur Çiğdem Aktaş
Vicomtech Foundation
Basque Research and Technology Alliance (BRTA)
Mikeletegi 57, 20009 Donostia-San Sebastián (Spain)
yaktas@vicomtech.org

Jorge García Castaño
Vicomtech Foundation
Basque Research and Technology Alliance (BRTA)
Mikeletegi 57, 20009 Donostia-San Sebastián (Spain)
jgarciac@vicomtech.org

***Abstract*—Categorizing darkweb 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. While many methods have been proposed to classify this image content, multi-label multi-class classification remains underexplored. The complexity of darkweb imagery, combined with the need for efficient classification systems, demands innovative approaches that can handle both the technical challenges and the sensitive nature of the content. In this paper, we present a comparative study of few-shot multi-label classification methods using the multimodal model CLIP. Our research addresses the growing need for robust classification systems that can effectively categorize diverse and complex image content while maintaining high accuracy and computational efficiency. We particularly focus on the challenges of handling multiple labels simultaneously and the scalability of these systems in real-world applications. We analyze and compare four different approaches: CLIP+Label Empower Adapter, CLIP Sigmoid, SIGLIP, and CLIP+ML-Decoder. Our study evaluates these methods based on their precision, recall, and ability to handle increasing class numbers efficiently. Finally, our research contributes to the field by providing detailed insights into the strengths and limitations of each method.**

***Index Terms*—Multi-label image classification, Multi-class image classification, CLIP, SIGLIP, ML-Decoder**

I. Introduction

The dark web, a specialized subset of the vast global internet, is accessible only through dedicated web browsers. It is widely recognized as a hub for illicit activities due to its core characteristics: anonymity, which provides users with a level of privacy not typically available on the surface web, and untraceability, which makes it extremely challenging to track the origins and destinations of data transfers. As a result, the dark web serves as a critical source of threat intelligence, offering insights into cyber-attacks, stolen assets, illegal trade, arms trafficking, confidential data leaks, child exploitation, and other criminal activities.

Automatically analyzing the visual content of the dark web —such as images and videos- is essential for efficient image categorization, enabling the detection and prevention of potential threats. However, dark web image classification remains an open challenge due to the complex nature of its content.

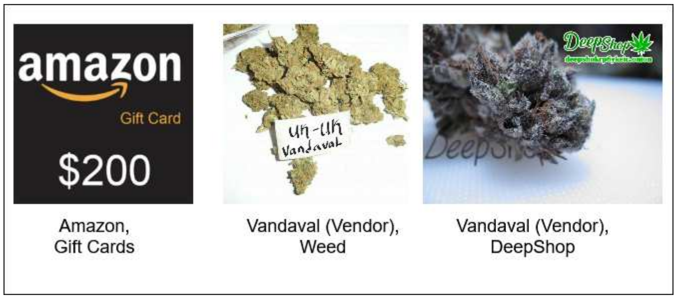
Amazon, Gift Cards
Vandaval (Vendor), Weed
Vandaval (Vendor), DeepShop

Fig. 1: Multi-label image examples from Dark web dataset

Many images suffer from low resolution, blur, and small object sizes, often containing objects that blend into the background due to color similarities. Additionally, the coexistence of multiple overlapping domains and intra-class variations — where instances within the same category exhibit significant visual differences — further complicate classification tasks, as illustrated in Fig. 1.

This complexity highlights the importance of multi-label classification, where an image can be assigned multiple relevant labels rather than a single category.

A range of multi-label classification techniques have been explored to tackle this problem. Some approaches decompose multi-label classification into separate single-label problems[1], [2], while others refine loss functions or activation functions to enhance classification accuracy[3], [4]. Additionally, graph-based networks utilize learned label embeddings to locate key discriminative features [5], [6], while transformer-based architectures have been adapted for multi-label classification tasks [7].

While these methods show promising results, they rely on large, well-balanced datasets—which are difficult to curate in real-world applications. Expanding a dataset with a balanced class distribution is an exponentially complex process [8], making large-scale multi-label learning infeasible in many scenarios.

To address this, recent advances in vision-language models like CLIP [9] offer new possibilities. By aligning visual and textual embeddings, CLIP enables zero-shot and few-shot learning in open-vocabulary settings. However, most CLIP-

979-8-3315-0993-4/25/31.002025$IEEE$

based methods are tailored to single-label classification.

In this study, we address the problem of few-shot multi-label image classification, focusing on the dark web as a real-world, high-stakes application domain. Our hypothesis is that adapting state-of-the-art few-shot multi-label classification techniques to CLIP can improve generalization and performance in low-resource, high-variability environments.

Our contributions can be summarized as follows:

- We analyze four state-of-the-art few-shot multi-label classification techniques and provide a portable pipeline applicable to any real-world dataset requiring few-shot multi-label classification.
- We evaluate these approaches by implementing CLIP-based methodologies for dark web multi-label image classification.
- We compare our previous approach with newer CLIP-based multi-label methodologies, presenting a transparent comparative study which contributes to the field by providing detailed insights into the strengths and limitations of each method.

II. Related Work

A variety of methods have been proposed to solve the multi-label classification problem, which can be broadly categorized into (i) problem transformation methods and (ii) algorithm adaptation techniques.

Early research focused on decomposing multi-label classification into independent binary classification tasks[1]. However, this approach fails to capture inter-label correlations and requires training separate classifiers for each category—introducing a significant computational burden. To address this, some studies propose classifier chaining[10], where each classifier's output serves as an input feature for subsequent classifiers. While this strategy improves inter-label dependency modeling, it still suffers from high computational costs due to the growing number of classifiers required.

Another transformation-based approach, known as Label Powerset (LP) [2], converts multi-label problems into multi-class problems by treating each unique label combination as a distinct class. However, LP struggles with the exponential growth of label combinations, making it impractical for large datasets.

Several methods aim to improve multi-label classification by optimizing the loss function[3] or activation function[4] to mitigate class imbalance issues.

In computer vision, multiple approaches have been developed for multi-label image classification. Hypotheses-CNN-Pooling (HCP) [11] generates a large number of proposals through object detection techniques, treating each proposal as a single-label classification problem.

Beyond CNN-based methods, Graph Convolutional Networks (GCNs) have demonstrated high efficacy across various vision tasks, including multi-label classification [5], [6], [12]. GCN-based methods build classifiers by modeling label relationships within graph networks.

Other approaches employ weakly supervised learning for multi-label classification [13], leveraging knowledge distillation techniques from object detection models.

More recently, transformers[14]—originally developed for modeling long-range dependencies in natural language processing (NLP)[15]–[17]—have demonstrated strong performance across computer vision tasks, including image classification[7], [9] and object detection[18].

Several transformer-based multi-label classification techniques have emerged:

Multi-label Transformer (MLT) [19], which models pixel-wise attention to enhance feature extraction. Transformer-based query learning [20], which uses decoder queries to predict label existence. Graph-enhanced Transformers [21], which combine GCNs with attention mechanisms. Metric Learning Transformers [22], which integrate metric learning to assess label similarities. Transformer-CNN Hybrids [23], which generate attention maps per label, capturing intra-class dependencies through convolutional layers.

Despite these advancements, existing multi-label methods still require large datasets to achieve high accuracy. However, data scarcity is a prevalent challenge in real-world applications, where collecting and labeling extensive datasets is both resource-intensive and time-consuming.

Large-scale open-vocabulary models trained on vast datasets, such as CLIP [9], offer a promising alternative by leveraging zero-shot single-label classification and few-shot multi-label classification. By adapting these models, we aim to explore their potential for enhancing multi-label classification in dark web imagery, where traditional methods face significant constraints.

III. CLIP-based Multi-label Classification Methodologies

There are four recent multi-label classification methods suitable for CLIP multi-modal model adaptation: (i) CLIP+Label Empower(our previous model), (ii) CLIP Sigmoid (iii) SIGLIP (iv) CLIP + ML-Decoder

A. *CLIP+Label Empower*

Label Empower [2] is one of the approaches that convert the multi-label classification problem into a single-label classification, by multi-labels to binary vectors by hot-encoding them and assigning a unique label for each different binary vector. As demonstrated in [24], it has the highest performance among the other approaches to solving the multi-label classification task by transforming the problem into a single-label one. Fig. 2 shows the architecture of our previously proposed method CLIP+Label Empower Adapter.

Although our previously proposed model has the highest recall, as shown in Table I, among even these four state-of-the-art CLIP-adapted multi-label classification approaches, not only between the traditional ones, it preserves a risk of accuracy drop within a large number of classes due to its exponentially growing hot-encoded labels. This possible risk is especially important for such a dataset like darkweb, which

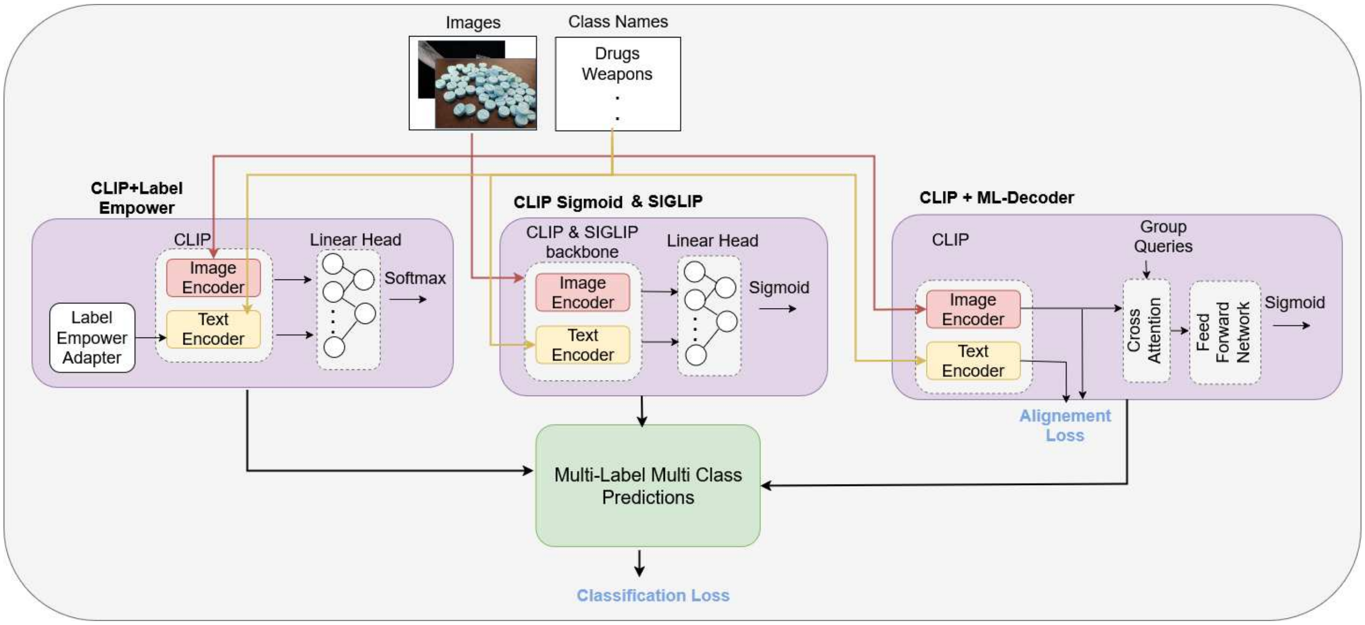

Fig. 2: CLIP based Multi-Label Methodologies

is expected to grow fast in the sense of number of classes, since new type of images and titles are expected to occur.

- **Worst Case:** $O(N!)$ (All class labels exist in all possible permutations in the dataset)
- **Best Case:** $O(N)$ (The dataset contains only single labels)

B. *CLIP Sigmoid*

CLIP has the capability to classify an extremely huge number of different classes, thanks to its open-vocabulary nature. Yet we might need to fine-tune this state-of-the-art model for our real-world datasets like darkweb dataset. In a single-label pipeline, the proper way of fine-tuning such a model is to add a linear layer following the image and text embeddings, having output nodes as much as class numbers in the dataset. In a single-label, multi-class classification scenario, softmax is commonly used to distribute probabilities among different classes, ensuring that the sum of all class probabilities equals 1.

$$P(y_i|x) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \tag{1}$$

where z_i represents the logits for class i and N is the total number of classes.

However, in multi-label classification, where multiple labels can be present simultaneously, softmax is suboptimal as it forces mutual exclusivity among classes. Instead, sigmoid activation is more appropriate, as it independently predicts each label's probability:

$$P(y_i|x) = \frac{1}{1 + e^{-z_i}} \tag{2}$$

This allows the model to assign a probability close to 1.0 for each relevant label in an image, ensuring that all present classes are predicted without competition from other labels.

Thus, fine-tuning CLIP with a linear classification head using Sigmoid activation is another approach. Although this method does not have any risk of exponentially growing with the number of class numbers, our experimental results show its poor accuracy in comparison with our previously proposed method.

C. *SIGLIP*

While SIGLIP [25] is a state-of-the-art, multi-modal model, having a very similar architecture to CLIP, its main difference is using Sigmoid, instead of Softmax while pre-training. This fact provides us with the possibility to see if the poor performance of CLIP + Sigmoid fine-tuning methodology occurred due to using a different activation function in the fine-tuning than pre-training.

Fig. 2 shows the very similar architecture of SIGLIP.

D. *CLIP+ML-Decoder*

ML-Decoder [26] is a state-of-the-art decoder implementation aiming to overcome the exponentially growing input size in classification pipelines built with default transformer decoders (Fig. 3) due to the self-attention layer. The method proposes to remove the self-attention layer and claims its affectless on the classification accuracy. It is also claimed to use ”group queries”, instead of fixed query embeddings per class as in traditional transformer-decoders. Being the number of groups a new hyperparameter in this architecture refers to the amount of input query embeddings the decoder will have, independently from the number of classes in the dataset. Being K is the number of group queries, the decoder learns to create K queries referencing N number of classes in a dataset, instead of having N fixed queries for N classes as in default transformer decoders. CLIP+ML-Decoder approach uses both classification loss coming from the final prediction logits and an alignment loss coming from the similarity of text and image embeddings to obtain a final loss. The fixed text embeddings are used only for this purpose, whereas non-fixed group queries are learnable embeddings and they are updated during the training.

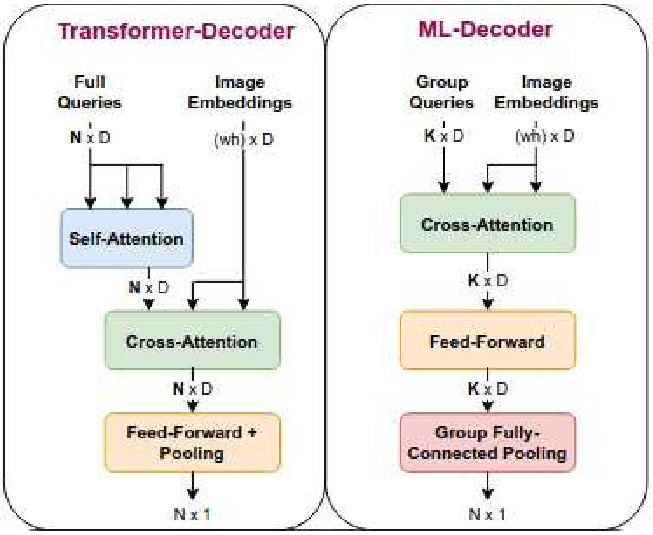

Fig. 3: Traditional vs ML Decoders

Therefore this approach enhances scalability, with two key optimizations: (1) Self-attention removal, which reduces the computational complexity from $O(N^2)$ to $O(N)$, and (2) Group decoding, which further optimizes inference by shifting the complexity from $O(N)$ to $O(K)$, where KK represents the number of meaningful groups instead of processing all classes independently.

Self-attention removal:

$$O(N^2) \rightarrow O(N)$$

Group decoding:

$$O(N) \rightarrow O(K)$$

Fig. 2 shows the architecture of CLIP + ML-Decoder.

IV. Experimental Study

A. Dark web dataset

The Dark web dataset, sourced from CFLW's Dark Web Monitor[1], 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. Fig. 1 shows some image samples along with their ground truth labels from various classes.

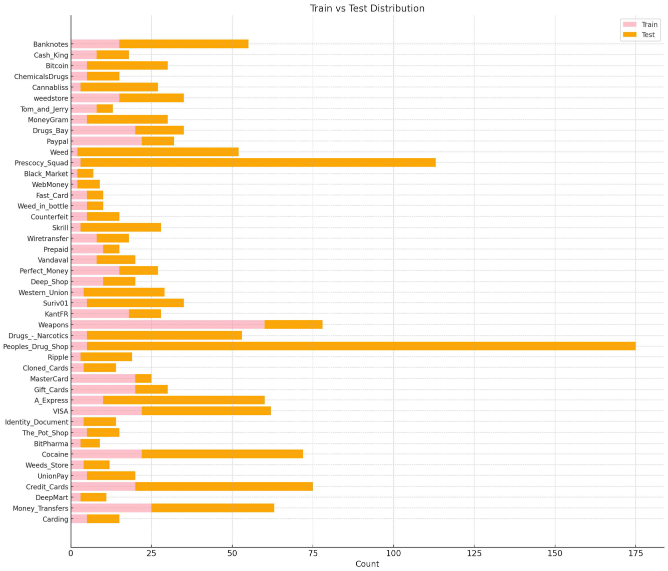

Fig. 4: Data distribution of the dark web dataset for both train and test sets.

The dataset contains 46 classes from various categories 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 existing in the same image. Being the current darkweb dataset is an experimental one, many more classes are expected to join and thus, the scalability performance is important than a regular Fig. 4 shows the data distribution of the dark web dataset for both train and test sets.

The darkweb 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 collecting only a few samples of its subgroups. Fig. 1 shows 3 image samples from the dark web dataset, 1 belonging to a financial organization category, 2 belonging to the drugs and narcotics category containing pictures of weed and the vendor names. While vendor names may vary, each image having a vendor name information also contains weed class, doubling the amount of "weed" class in comparison of class names referring to the individual vendor names. Which is robust example showing the reason of strong class imbalance problem in darkweb dataset.

Lastly, another challenge is that some classes, especially the vendor names like "deep shop", or "vandaval", have textual information rather than visual features which brings the need for considering the combination of vision and language data.

B. Implementation Details

The fine-tuning strategy varies for each of the methods we examine: CLIP + Label Empower was finetuned by freezing the CLIP pre-trained model, and updating the weights of only for between 5 to 10 epochs, using Adam optimizer with default

[1] https://cflw.com/dwm/

learning rate 1×10^{-3}. Softmax final activation function and Cross Entropy Loss.

CLIP Sigmoid model was finetuned by freezing the CLIP pre-trained model, and updating the weights of only the additional linear classification head, for 50 epochs until convergence, using the same optimizer and learning rate as previous approach, with a Binary Cross Entropy Loss over the logits, as a default approach of sigmoid classification.

SIGLIP model was finetuned without freezing any part of the architecture, since our implementation any additional part and the nature of the architecture is already compatible for multi-label classification. It is finetuned 80 epochs until convergence, using Adam optimizer with a learning rate of 5×10^{-5}.

CLIP+ML-Decoder method was fine-tuned by freezing the CLIP model and only focusing on updating the weights of Decoder. A grid search over all the possible hyper-parameters of the decoder block was made and the best results was obtained with: multi-head attention head amount 4, dropout in feedforward network 0.5, number of groups K 8, number of layers (decoder block amount) 2. Similarly to the previous methods, Adam optimizer is used, with a learning rate of 1×10^{-4}.

C. Evaluation Metrics

We evaluate model performance using standard classification metrics: precision, recall, and f1-score. We furthermore analyze the scalability of the model.

- **Precision:** Measures the proportion of correctly predicted labels among all predicted labels. 3
- **Recall:** Measures the proportion of correctly predicted labels among all true labels.4
- **F1-Score:** The harmonic mean of precision and recall. 5
- **Scalability:** Assesses how well the model adapts to increasing class sizes.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1\_score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5}$$

The true positives, false positives, false negatives are calculated similarly to single-label classification. Fig. 5 shows an image sample having two ground truth labels: VISA and Banknotes. In case the predicted classes are ”VISA and 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 ”Banknotes” classes.

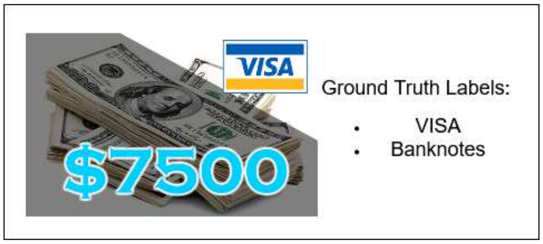

Fig. 5: An example multi-labeled image with 2 classes.

V. Results and Discussion

Table I compares the performance of the four methods in terms of precision, recall, f1-score and scalability and summarizes our findings. CLIP+Label Empwoer method has a low scalability since it has O(N!) as worst case scenario, CLIP Sigmoid and SIGLIP has a moderate scalability since they provide an improvement, but neither bring any growth, meaning the worst and best case scenario is the same and O(N). While CLIP+ML-Decoder improves the scalability from O(N) to O(K), being N is the number of classes and K is the number of groups defined by the end-user.

Table II compares the two models giving the best results for dark web dataset, on the MS-COCO multi-label dataset, causing the CLIP+LE method to have a huge drop of accuracy, due to expanding 80 base classes to 234,581 hot-encoded classes. These results show the aforementioned potential risk of CLIP+LE method on the datasets having high relation between the classes, causing the existence of various permutations of multi-label vectors among the dataset. Even though the results might be impressive for such amount of classes, it is visible thatthe Label Empower method has a huge impact on CLIP multi-modal model’s classification capacity by exploding the class numbers unnecessarily.

TABLE I
COMPARISON OF FEW-SHOT MULTI-LABEL CLASSIFICATION METHODS. LE: LABEL EMPOWER, MLD: ML-DECODER

Method	Precision	Recall	F1_score	Scalability
CLIP+LE	0.944	**0.931**	**0.937**	Low
CLIP Sigmoid	0.493	0.276	0.351	Moderate
SIGLIP	0.880	0.735	0.801	Moderate
CLIP+MLD	**0.958**	0.916	0.936	**High**

Our results show that the CLIP+Label Empower Adapter achieves the best recall, making it still a strong candidate for recall-sensitive applications built for datasets not having huge amount of classes, due to its low scalability and the potential risk of providing poorer results with a huge number of classes. On the other hand, the results show the medium-level scalability models (CLIP+Sigmoid, SIGLIP), that are not improving nor worsen the architecture according to the number of classes, have poor accuracy in comparison with CLIP + adapter solutions. However, CLIP+ML-Decoder performs well in both precision and recall, while also addressing the scalability issue effectively and it should be a definite choice for dataset having huge number of classes, where to not have false positive predictions is more important than not missing any true positive prediction, regarding it's precision beating the CLIP+Label Empower method while providing lower recall.

TABLE II
COMPARISON OF MULTI-LABEL MS-COCO MAP SCORE. LE: LABEL EMPOWER AND MLD: MULTI-LABEL DECODER

Method	Precision	Recall
CLIP+LE	0.592	0.555
CLIP+MLD	**0.839**	**0.809**

VI. Conclusion and Future Works

In this study, we analyzed four few-shot multi-label classification methods based on CLIP. Our results demonstrate that CLIP+Label Empower Adapter excels in the recall, whereas CLIP+ML-Decoder provides a more scalable solution by mitigating the exponential growth problem in input features, providing also a robust performance on accuracy metrics. Future work will explore the efficient deployment pipeline for CLIP+ML-Decoder approach for real-world multi-label classification tasks.

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] M.-L. Zhang, Y.-K. Li, X.-Y. Liu, and X. Geng, "Binary relevance for multi-label learning: An overview," *Frontiers of Computer Science*, vol. 12, no. 2, pp. 191–202, Mar. 2018, ISSN: 2095-2236. DOI: 10.1007/s11704-017-7031-7. [Online]. Available: http://dx.doi.org/10.1007/s11704-017-7031-7.

[2] G. Tsoumakas, A. Dimou, E. Spyromitros-Xioufis, V. Mezaris, I. Kompatsiaris, and I. Vlahavas, "Correlation-based pruning of stacked binary relevance models for multi-label learning," Jan. 2009, pp. 101–116.

[3] E. Ben-Baruch, T. Ridnik, N. Zamir, *et al.*, *Asymmetric loss for multi-label classification*, 2021. arXiv: 2009.14119 [cs.CV].

[4] A. F. T. Martins and R. F. Astudillo, *From softmax to sparsemax: A sparse model of attention and multi-label classification*, 2016. arXiv: 1602.02068 [cs.CL].

[5] Y. Wang, D. He, F. Li, *et al.*, *Multi-label classification with label graph superimposing*, 2019. arXiv: 1911.09243 [cs.CV].

[6] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, *Multi-label image recognition with graph convolutional networks*, 2019. arXiv: 1904.03582 [cs.CV].

[7] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, *Training data-efficient image transformers distillation through attention*, 2021. arXiv: 2012.12877 [cs.CV].

[8] W. Zhang, C. Liu, L. Zeng, B. Ooi, S. Tang, and Y. Zhuang, "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)*, Oct. 2023, pp. 1423–1432.

[9] A. Radford, J. W. Kim, C. Hallacy, *et al.*, *Learning transferable visual models from natural language supervision*, 2021. arXiv: 2103.00020 [cs.CV].

[10] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains: A review and perspectives," *Journal of Artificial Intelligence Research*, vol. 70, pp. 683–718, Feb. 2021, ISSN: 1076-9757. DOI: 10.1613/jair.1.12376. [Online]. Available: http://dx.doi.org/10.1613/jair.1.12376.

[11] Y. Wei, W. Xia, M. Lin, *et al.*, "Hcp: A flexible cnn framework for multi-label image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1901–1907, Sep. 2016, ISSN: 1939-3539. DOI: 10.1109/tpami.2015.2491929. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2015.2491929.

[12] T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin, *Learning semantic-specific graph representation for multi-label image recognition*, 2019. arXiv: 1908.07325 [cs.CV].

[13] Y. Liu, L. Sheng, J. Shao, J. Yan, S. Xiang, and C. Pan, "Multi-label image classification via knowledge distillation from weakly-supervised detection," in *Proceedings of the 26th ACM international conference on Multimedia*, ACM, Oct. 2018. DOI: 10.1145/3240508.3240567. [Online]. Available: http://dx.doi.org/10.1145/3240508.3240567.

[14] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL].

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2019. arXiv: 1810.04805 [cs.CL].

[16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:160025533.

[17] T. B. Brown, B. Mann, N. Ryder, *et al.*, *Language models are few-shot learners*, 2020. arXiv: 2005.14165 [cs.CL].

[18] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, *Deformable detr: Deformable transformers for end-to-end object detection*, 2021. arXiv: 2010.04159 [cs.CV].

[19] X. Cheng, H. Lin, X. Wu, *et al.*, *Mltr: Multi-label classification with transformer*, 2021. arXiv: 2106.06195 [cs.CV].

[20] S. Liu, L. Zhang, X. Yang, H. Su, and J. Zhu, *Query2label: A simple transformer way to multi-label classification*, 2021. arXiv: 2107.10834 [cs.CV].

[21] J. Ye, J. He, X. Peng, W. Wu, and Y. Qiao, *Attention-driven dynamic graph convolutional network for multi-label image recognition*, 2020. arXiv: 2012.02994 [cs.CV].

[22] K. Prokofiev and V. Sovrasov, *Combining metric learning and attention heads for accurate and efficient multilabel image classification*, 2022. arXiv: 2209.06585 [cs.CV].

[23] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, *Learning spatial regularization with image-level supervisions for multi-label image classification*, 2017. arXiv: 1702.05891 [cs.CV].

[24] Y. Ç. Aktaş and J. G. Castaño, "Few-shot multi-label multi-class classification for dark web image categorization," in *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, 2024, pp. 1–6. DOI: 10.1109/ISDFS60797.2024.10527297.

[25] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, *Sigmoid loss for language image pre-training*, 2023. arXiv: 2303.15343 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2303.15343.

[26] T. Ridnik, G. Sharir, A. Ben-Cohen, E. Ben-Baruch, and A. Noy, *Ml-decoder: Scalable and versatile classification head*, 2021. arXiv: 2111.12933 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2111.12933.