Multi-label Deepfake Classification

Inder Pal Singh†, Nesryne Mejri†, Van Dat Nguyen†, Enjie Ghorbel†, Djamila Aouada†
†Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg
{inder.singh, nesryne.mejri, dat.nguyen, enjie.ghorbel, djamila.aouada}@uni.lu

***Abstract*—In this paper, we investigate the suitability of current multi-label classification approaches for deepfake detection. With the recent advances in generative modeling, new deepfake detection methods have been proposed. Nevertheless, they mostly formulate this topic as a binary classification problem, resulting in poor explainability capabilities. Indeed, a forged image might be induced by multi-step manipulations with different properties. For a better interpretability of the results, recognizing the nature of these stacked manipulations is highly relevant. For that reason, we propose to model deepfake detection as a multi-label classification task, where each label corresponds to a specific kind of manipulation. In this context, state-of-the-art multi-label image classification methods are considered. Extensive experiments are performed to assess the practical use case of deepfake detection.**

***Index Terms*—Deepfake detection, Multi-Label Classification, Stacked Manipulations**

I. Introduction

The recent advances in Deep Learning (DL) techniques have led to the emergence of highly realistic facial manipulations, known as deepfakes. The subtlety of these forgeries makes their distinction from authentic images increasingly challenging. Given this threat, many efforts have been dedicated to developing deepfake detection techniques [1]–[3]. Typically, these approaches formalize the problem of deepfake detection as a binary classification [2]–[8]. Given an input image or video, they predict whether it has been forged or not; therefore classifying it as 'real' or 'fake'. However, binary predictions are opaque and are difficult to interpret, while in real-world applications, explainable predictions in deepfake detectors are of utmost importance. In fact, an image predicted as fake can be produced by one or multiple manipulations. In existing face editing software, such as FaceTune[1], it is common for the same image to undergo several edits, which we refer to as *stacked manipulations* or *multi-step operations*, as illustrated in Fig. 1(a).

As an alternative, we propose in this paper to reformulate the task of deepfake detection as a multi-label classification problem, where each label corresponds to a specific manipulation. Such a formulation is supported by the fact that multiple forgeries can be present in the same image.

Recently, Shao et al. [9] highlighted the necessity of detecting multi-step manipulations. For that purpose, they have introduced a novel deepfake dataset incorporating sequences of facial forgeries, along with their annotations. However, instead of considering multi-label classification, they framed the problem of deepfake detection as an image-to-sequence task. This means that their goal was not only to recognize the different manipulations applied to a given image, but also to retrieve their chronological order. Nevertheless, predicting the temporal structure of a forgery sequence adds complexity to the problem without having a clear benefit in a practical scenario.

In this paper, we argue that for detecting stacked manipulations, it is sufficient to formulate deepfake detection as a multi-label image classification task. As we are the first to explicitly rethink deepfake detection as such, we propose to show the suitability of existing multi-label image classification methods for the practical scenario of detecting multi-step manipulations. Our main finding is that current deepfake multi-label image datasets might be too simplistic since they were created under controlled conditions. This emphasizes the need for more realistic deepfake datasets, as the existing ones may not accurately reflect the performance of state-of-the-art multi-label classification methods.

In summary, our contributions are twofold: (1) we reformulate deepfake detection as a multi-label classification problem and show that more explainable predictions can be achieved regardless of the forgery order; (2) we compare multiple state-of-the-art multi-label classification techniques in the context of deepfake detection and present an extensive analysis of the obtained results.

In the remainder of this work, Section II formulates the problem of multi-label deepfake classification. Section III presents an overview of the considered multi-label image classification techniques. In Section IV, we detail the experimental setup and present our results. Finally, Section V concludes this work and offers interesting perspectives.

II. Formulating Deepfake Detection as a Multi-label image Classification problem

Let $\mathcal{I}$ be a dataset formed by a set of real and fake images. Given an image $\mathbf{I} \in \mathcal{I}$, traditional deepfake detection methods consider that the label of $\mathbf{I}$ belongs to $[\![0, 1]\!]$. In other words, they classify an image as real or fake, formulating the problem as a simple binary classification. Nevertheless, a deepfake image might result from multi-step manipulations that enclose different properties. As detecting the nature of these manipulations is highly relevant for obtaining a more explainable output, we propose to define the problem of deepfake detection as a multi-label classification. Let $\mathbf{I} \in \mathcal{I}$ be a given image, we aim at estimating a function f that predicts

This work is supported by the Luxembourg National Research Fund, under the BRIDGES2021/IS/16353350/FaKeDeTeR and UNFAKE, ref.16763798 projects, and by Post Luxembourg.

[1]https://www.facetuneapp.com/

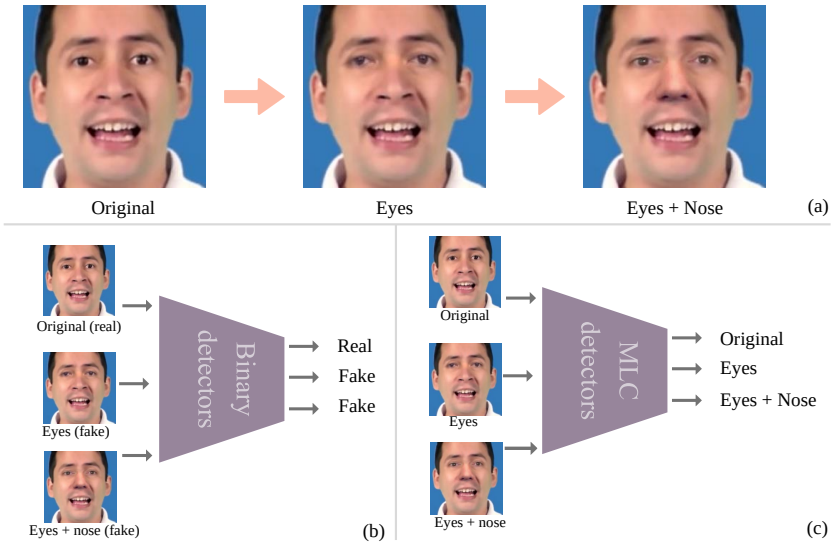

Fig. 1. **(a)** Examples of single-step manipulation affecting only the eyes and multi-step manipulations affecting both the eyes and the nose. **(b)** Binary deepfake detectors treat single-step and multi-step forged images equally, which implicitly assumes that only one manipulation took place in the image. **(c)** Whereas Multi-Label deepfake Classifiers (MLC) predict more informative outputs by indicating the labels of the applied manipulations.

the presence or not of N different manipulations. This can be written as follows,

$$f : \mathbb{R}^{w \times h} \rightarrow [\![0, 1]\!]^N$$
$$\mathbf{I} \mapsto \mathbf{y} = (y_i)_{i \in [\![1, N]\!]},$$

where w and h are respectively the pixel-wise width and height of the image. It is to note that $y_i = 1$ if the manipulation i is present in $\mathbf{I}$, otherwise, $y_i = 0$.

III. Comparison of Multi-Label Image Classification for Deepfake Detection

The multi-label image classification problem has received a lot of attention from the computer vision research community in recent years. Many methods have demonstrated outstanding performances in light of current developments in deep learning techniques. We propose in this paper to evaluate the performance and assess the current state of existing multi-label image classification methods in the context of deepfake detection, as formulated in Section II. For that purpose, two main categories of methods are considered, namely, direct and indirect methods. We describe these methods in the subsections that follow.

A. Direct methods

In order to determine if multiple objects are present in an image or not, direct methods employ a single stream deep neural network f that directly maps a given image to a binary vector. In other words, f is usually learned in an end-to-end manner in this case. Generally, these single-stream architectures are constituted of two main components, namely: (1) a block of Convolutional Neural Networks (CNN) which seeks to extract discriminative image features; and (2) a classification head that employs a Multi-Layer Perceptron (MLP) to directly translate these features into the probability of occurrence of each considered label.

Among direct methods, the ResNet architecture is probably one of the most successful [10]–[13]. Recently, TResNet [11]- an improved version of ResNet [10] that takes advantage of GPU capabilities, has also been proposed for multi-label classification. Moreover, by combining the recently introduced Asymmetric Loss (ASL) [13] with TResNet, improved results have been achieved. Note that the ASL loss acts differently on positive and negative labels.

Herein, we compare the effectiveness of some popular direct techniques in the context of deepfake detection, namely: (1) ResNet50 [10]; (2) ResNet101 [10]; and (3) TResNetM [11]. Additionally, we couple these methods with ASL [13]. Section IV provides more details on the quantitative performance of these methods.

B. Indirect methods

While direct approaches have shown great performance, they tend to require a large number of layers to work effectively. To avoid using very deep networks, a second research line has attempted to model label dependencies. In fact, label correlations are important cues since some labels are more likely to appear together in the same image. For example, we have a higher chance to observe a “sheep” and some “grass” in one image than a “sheep” and a “bicycle”. We refer to these approaches as indirect methods.

Graphs have been particularly useful for modeling label correlations. Graph-based approaches are typically formed by two streams. They usually combine a CNN denoted by f_1 that learns discriminative image features with a Graph Convolutional Network (GCN) for generating interdependent label-wise classifier denoted by f_2 [16], [18], [19]. These generated

classifiers are directly applied to the features resulting from f_1. In other words, images are mapped to a binary vector using the function $f = f_2 \circ f_1$. The pioneering work on graph-based multi-label classification [16] made use of word embeddings [17] to represent graph nodes. More recent techniques [18], [19] have generated image-based embeddings to improve the performance. Additionally, earlier graph methods are mainly based on a pre-computed fixed adjacency matrix where weak edges are ignored using an empirically fixed threshold. This may lead to a significant loss of information. To overcome this issue, ML-AGCN [19] attempts to adaptively learn the adjacency matrix by computing an attention weight for each node pair.

Herein, we compare the effectiveness of some recent indirect graph-based techniques in the context of deepfake detection, namely: (1) ML-GCN [16]; (2) IML-GCN [18]; and (3) ML-AGCN [19]. We use both word [17] and image-based [18] node embeddings to assess the performance of the aforementioned indirect methods. We generate the label graph using the co-occurrences of each manipulation pair in an image over the entire dataset, as in [16]. Section IV gives more details on the quantitative performance of these methods.

IV. Experiments

A. Datasets

For our experiments, we use the dataset referred to as *Deep-Seq* proposed in [9]. Initially, this dataset was proposed for image-to-sequence tasks. Nevertheless, its annotations are compatible with multi-label classification. As compared to [9], the order constraint is not considered. More specifically, the dataset consists of two subsets depicting various manipulations. The first subset, called *Sequential facial components manipulations (Seq-Com-Deepfake)*, shows forgeries that alter the appearance of facial attributes such as hair bangs or the beard. In the second subset, the manipulations are applied by swapping facial regions, such as the eyes, the mouth, etc., between an original and a reference image, respectively. This subset is termed *Sequential facial attributes manipulations (Seq-Att-Deepfake)*. For both sub-collections, one to five manipulations are applied to the same image. Hence, the label vector is formed by five elements ($N = 5$).

B. Evaluation metrics

We provide the mean Average Precision (mAP) as well as the number of model parameters (# Params) in order to assess the effectiveness of current state-of-the-art multi-label classification approaches in the context of deepfake detection. In addition, as in [13], we report the following evaluation metrics on both subsets of the Deep-Seq dataset: average per-Class Precision (CP), average per-Class Recall (CR), average per-Class F1-score (CF1), the average Overall Precision (OP), average overall recall (OR) and average Overall F1-score (OF1).

C. Implementation details

In the context of deepfake detection, the effectiveness of both direct and indirect approaches is assessed. For that purpose, we employ ResNet [10] and TResNet [11] as direct approaches, in addition to ML-GCN [16], IML-GCN [18] and ML-AGCN [19] as indirect methods. More specifically, we utilize both Resnet50 and Resnet101 variations. For TResNet, we adapt a smaller version known as TResNet-M.

We use the original train and test split that was initially provided in the dataset [9] to train our models. For the subset of facial attribute manipulations (Seq-Attr-Deepfake), we use 41600 samples for training and 4160 samples for testing, and for the subset of facial component manipulations (Seq-Comp-Deepfake), we use 29408 and 2860 samples for training and testing, respectively. Using conventional image augmentation techniques, the image samples are reshaped to 224x224 as suggested in the original methods [10], [11], [16], [19]. We train the models on an NVIDIA TITAN V GPU with a total memory of 12GB using PyTorch in Python with a batch size of 128 for a total of 40 epochs or until convergence.

D. Experimental Results

We report in Table I and Table II the results obtained for both Seq-Att-Deepfake and Seq-Com-Deepfake subsets, respectively.

1) Comparison of direct methods: In general, all the results obtained for direct methods are comparable. However, it is interesting to note that, in our experiments, ResNet50 outperforms TResNetM in terms of mAP regardless of ResNet's depth, with an improvement of approximately 2%. It should be noted, though, that TResNetM allows for a larger batch size than ResNet50 while still utilizing the same GPU memory. In addition, surprisingly, the use of the ASL loss does not seem to influence the results importantly on the Seq-Attr-Deepfake subset, only inducing a variation of 0.1% in mAP. On the other hand, a marginal performance improvement on the other subset i.e., Seq-Comp-Deepfake can be noticed when comparing direct methods to their counterpart ASL-based ones. However, the recall (CR, OR) and F-1 score (CF, OF), both per-class and overall, increase significantly when these methods are combined with ASL. This might be explained by two points: 1) since ASL aims to focus more on positive labels than negative ones, the model tends to predict more false positives; and 2) the models may overfit the distribution of manipulations in Seq-Attr-Deepfake.

Finally, it can be noted that the best performance is achieved when using a deeper architecture, i.e ResNet101, with an enhancement between 2% and 4%, in terms of mAP, on Seq-Attr-Deepfake and Seq-Com-Deepfake, respectively. Nevertheless, this slight improvement comes at the cost of an important increase in terms of number of parameters (almost multiplied by a factor of 2).

2) Comparison of indirect methods: The largest architecture corresponding to ML-GCN outperforms other graph-based methods. More specifically, an improvement of 0.5-12% and 1.6-8% can be observed in terms of mAP for Seq-

TABLE I
COMPARISON OF EXISTING MULTI-LABEL IMAGE CLASSIFICATION METHODS ON DEEPFAKE ATTRIBUTE MANIPULATIONS SUBSET (**SEQ-ATTR-DEEPFAKE**). BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Category	Method	# Params (↓)	mAP (↑)	CP (↑)	CR (↑)	CF1 (↑)	OP (↑)	OR (↑)	OF1 (↑)
Direct methods	ResNet50 [10]	**23.8**	96.0	**93.5**	80.7	86.5	93.7	80.9	86.9
	ResNet 50 (with ASL) [10]	**23.8**	95.9	89.2	91.6	**90.4**	89.3	91.7	**90.5**
	ResNet101 [10]	42.8	**96.1**	92.9	83.6	87.8	93.1	83.7	88.2
	ResNet101 (with ASL) [10]	42.8	96.0	88.1	**92.3**	90.1	88.1	**92.4**	90.2
	TResNetM [11]	29.4	93.9	93.4	69.2	79.1	**93.8**	69.4	79.8
	TResNetM (with ASL) [13]	29.4	94.0	86.1	89.5	87.7	86.2	89.5	87.8
Indirect methods	ML-GCN† [16]	44.9	**95.1**	93.3	74.9	82.9	93.6	75.0	83.3
	IML-GCN [18]	**31.6**	82.5	76.7	72.4	74.3	76.9	72.6	74.7
	IML-GCN† [18]	31.7	94.0	84.8	90.6	87.6	84.9	90.6	87.7
	ML-AGCN [19]	36.3	82.7	74.8	77.1	75.9	74.9	77.1	76.0
	ML-AGCN† [19]	36.6	94.3	85.3	**90.9**	**88.0**	85.3	**90.9**	**88.0**
	ML-AGCN† w/o ASL [19]	36.6	94.5	**93.8**	70.0	79.9	**94.1**	70.2	80.4

†Graph-based indirect approaches with word embeddings [17]

TABLE II
COMPARISON OF EXISTING MULTI-LABEL IMAGE CLASSIFICATION METHODS ON DEEPFAKE COMPONENT MANIPULATIONS SUBSET (**SEQ-COMP-DEEPFAKE**). BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Category	Method	# Params (↓)	mAP (↑)	CP (↑)	CR (↑)	CF1 (↑)	OP (↑)	OR (↑)	OF1 (↑)
Direct methods	ResNet50 [10]	**23.8**	89.8	89.0	68.5	77.0	89.6	68.5	77.6
	ResNet 50 (with ASL) [10]	**23.8**	90.5	80.3	87.7	83.8	80.3	87.9	84.0
	ResNet101 [10]	42.8	91.7	**89.3**	74.4	80.7	**89.7**	74.5	81.4
	ResNet101 (with ASL) [10]	42.8	**92.7**	82.7	**90.0**	**86.2**	82.7	**90.0**	**86.2**
	TResNetM [11]	29.4	87.1	88.3	58.2	68.6	89.4	58.7	70.8
	TResNetM (with ASL) [13]	29.4	87.2	79.9	82.1	80.9	80.1	82.2	81.1
Indirect methods	ML-GCN† [16]	44.9	**89.6**	87.5	69.4	76.9	87.8	69.8	77.8
	IML-GCN [18]	**31.6**	81.7	**94.9**	18.8	27.9	92.3	20.0	32.9
	IML-GCN† [18]	31.7	87.7	79.8	82.1	80.9	80.0	82.2	81.1
	ML-AGCN [19]	36.3	81.7	80.1	65.5	71.8	80.2	65.3	72.0
	ML-AGCN† [19]	36.6	87.1	78.1	**84.9**	**81.3**	78.2	**85.1**	**81.5**
	ML-AGCN† w/o ASL [19]	36.6	88.0	90.7	54.4	65.5	**91.7**	54.7	68.5

†Graph-based indirect approaches with word embeddings [17]

Attr-Deepfake and Seq-Comp-Deepfake subsets, respectively. This is consistent with the results obtained for direct methods, as the feature extraction branch of ML-GCN is based on ResNet101. Moreover, while image embeddings have significantly improved the performance on standard multi-label image classification datasets, word embeddings give a higher mAP when tested on both deepfake subsets. In fact, as reported in Table I and Table II, for both IML-GCN and ML-AGCN, the mAP decreases by more than 12% when paired with image-based embeddings. This is counter-intuitive since, unlike the word embeddings that were initially proposed for the task of Natural Language Processing (NLP), image-based embeddings are semantically more meaningful for image classification as discussed in [18]. This might be caused by the fact that the discrepancy between the image embeddings produced by two different manipulations is not significant. In contrast to generic objects, image embeddings may fail to describe the manipulation semantics. Last but not least, the attention mechanism proposed in [19] does not improve the performance of the standard ML-GCN.

3) Direct methods versus indirect methods: In Fig. 2 and Fig. 3, we visualize the average performance of both direct and indirect methods on Seq-Attr-Deepfake subset and Seq-Comp-Deepfake subset, respectively. Given Fig. 2 and Fig. 3, two observations can be made. First, direct methods seem to be more suitable for multi-label deepfake classification. This

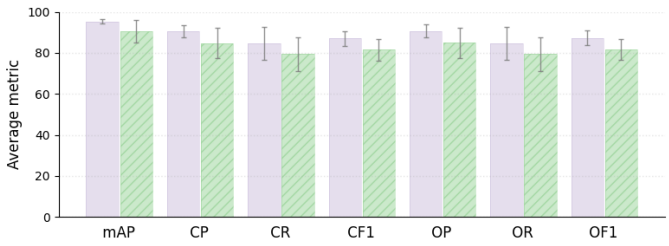

Fig. 2. Comparison of the average performance of numerous direct (non-hatched) and indirect (hatched) approaches on the Seq-Attr-Deepfake subset.

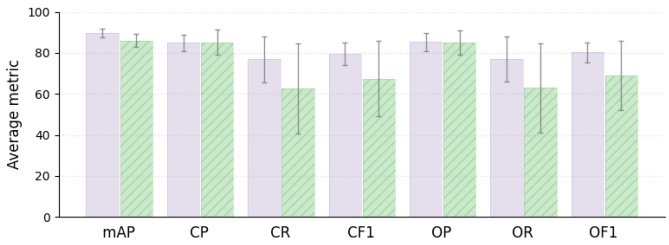

Fig. 3. Comparison of the average performance of numerous direct (non-hatched) and indirect (hatched) approaches on the Seq-Comp-Deepfake subset.

comes again in contradiction with the results obtained in the generic field of multi-label image classification, showing that modelling the label correlations is highly beneficial. This can be explained by the fact that the present deepfake dataset has not been spontaneously generated, but has been produced in a controlled environment. The distribution of the generated manipulations is assumed to be uniform, which does not necessarily reflect a realistic scenario. Second, we can observe that the facial components manipulations subset is relatively more challenging than the facial attribute subset, especially for indirect methods that enclose a high standard deviation in terms of performance metrics.

V. Conclusion

Existing deepfake detection techniques model the problem as a simple binary classification task, with the aim to determine whether or not a particular image is fake. However, this makes the classification task hardly interpretable. For obtaining more explainable outputs, this work proposes to tackle deepfake detection problem as a multi-label classification problem, with the objective of simultaneously identifying several categories of image manipulations. To this end, state-of-the-art multi-label classification methods are benchmarked on a recently proposed deepfake dataset incorporating multi-label annotations. This allows assessing the effectiveness of current multi-label classification methods, including both direct and indirect, in the practical use case of deepfake detection. Multiple results are against intuition, showing the need to investigate further multi-label deefake classification. These future investigations might be supported by the introduction of more complex and realistically generated multi-label deepfake datasets.

References

[1] Mejri, N., Ghorbel, E., and Aouada, D. (2023). UNTAG: Learning Generic Features for Unsupervised Type-Agnostic Deepfake Detection. *In IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1-5).

[2] Rössler A., Cozzolino D., Verdoliva L., Riess C., Thies J. and Nießner M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. *In Proceedings of IEEE/CVF International Conference on Computer Vision* (pp. 1-11)

[3] Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., and Guo, B. (2020). Face x-ray for more general face forgery detection. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5001-5010).

[4] Mejri, N., Papadopoulos, K., and Aouada, D. (2021). Leveraging High-Frequency Components for Deepfake Detection. *In 2021 IEEE 23rd International Workshop on Multimedia Signal Processing* (pp. 1-6).

[5] Shiohara K., and Yamasaki T. (2022). Detecting Deepfakes with Self-Blended Images. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (pp. 18720–18729).

[6] Chen L., Zhang Y., Song Y. Liu L. and Wang J. (2022). Self-supervised Learning of Adversarial Examples: Towards Good Generalizations for DeepFake Detections. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*

[7] Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., and Xia, W. (2021). Learning self-consistency for deepfake detection. *In Proceedings of the IEEE/CVF international conference on computer vision* (pp. 15023-15033).

[8] Wang, Y., Yu, K., Chen, C., Hu, X., and Peng, S. (2023). Dynamic Graph Learning With Content-Guided Spatial-Frequency Relation Reasoning for Deepfake Detection. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7278-7287).

[9] Rui S., Tianxing W. and Ziwei L. (2022). Detecting and Recovering Sequential DeepFake Manipulation. *In Proceedings of European Conference on Computer Vision* (pp. 712-728).

[10] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[11] Ridnik, T., Lawen, H., Noy, A., Ben Baruch, E., Sharir, G., and Friedman, I. (2021). "Tresnet: High performance gpu-dedicated architecture." *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1400-1409).

[12] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017. Densely connected convolutional networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).

[13] Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M. and Zelnik-Manor, L., 2021. Asymmetric loss for multi-label classification. *In Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 82-91).

[14] Zhu, F., Li, H., Ouyang, W., Yu, N., and Wang, X. (2017). Learning spatial regularization with image-level supervisions for multi-label image classification. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5513-5522).

[15] Qu, X., Che, H., Huang, J., Xu, L., & Zheng, X. (2021). Multi-layered semantic representation network for multi-label image classification. *arXiv preprint* arXiv:2106.11596.

[16] Chen, Z.M., Wei, X.S., Wang, P. and Guo, Y., 2019. Multi-label image recognition with graph convolutional networks. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5177-5186).

[17] Pennington, J., Socher, R. and Manning, C.D., 2014, October. Glove: Global vectors for word representation. *In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

[18] Singh, I.P., Oyedotun, O., Ghorbel, E. and Aouada, D., 2022. IML-GCN: Improved Multi-Label Graph Convolutional Network for Efficient yet Precise Image Classification. *In AAAI-22 Workshop Program-Deep Learning on Graphs: Methods and Applications.*

[19] Singh, I. P., Ghorbel, E., Oyedotun, O., and Aouada, D. (2022). Multi label image classification using adaptive graph convolutional networks (ML-AGCN). *In IEEE International Conference on Image Processing.*

[20] MultiLabelSoftMarginLoss — PyTorch 2.0 documentation. (n.d.). Pytorch.org. Retrieved May 28, 2023, from https://pytorch.org/docs/stable/generated/torch.nn.MultiLabelSoftMarginLoss.html