Computer Vision

Multi-label Image Classification

Group 25
Riajul Islam, Andreas Calonius Kreth, Christine Midtgaard

***Abstract*—Here is a summary of the project and the conclusions.**

***Index Terms*—Multi-label learning, deep learning, computer vision, multi-label classification, deep learning for MLC.**

I. Introduction

Multi-label classification is the supervised learning problem where an instance may be associated with multiple labels. Image classification is a computer vision task that requires assigning a label or multiple labels to an image. Single-label classification, or multi-class classification, refers to the problem where an image contains only one object to be identified. However, natural images usually contain multiple objects or concepts, highlighting the importance of multi-label classification [1]. In this project we investigate two methods aimed to solve two different problems within multi-label learning: Query2Label targets the challenges of imbalance and object localization, whereas MLSPL addresses the challenge of training an effective multi-label classifier from minimal supervision.

A. Multi-Label Classification Problem

II. Related Work

Multi-label learning is a well studied problem within computer vision [2].

a) Loss Functions:

b) Convolutional Neural Networks:

c) Vision Transformer:

d) PU learning: Learning from positive and unlabeled data: a survey by Bekker and Davis [3].

Learning to Classify Texts Using Positive and Unlabeled Data by Li and Liu [4].

e) Partially Observed Labels:

f) Heuristics: Heuristics can be used to reduce the required annotation effort [34, 18], but this runs the risk of increasing error in the labels [2].

III. Theoretical Background

Contrary to multi-class or binary classification where each instance is associated with only one label from a set of categories (e.g. person, dog, car, etc.), multi-label classification may assign an instance multiple labels [2]. Assuming K categories, an input image $x \in \mathcal{X}$ is associated with a vector of labels $y = [y_1, ..., y_K]$ from the label space $\mathcal{Y} = \{0,1\}^K$, where $y_k = 1$ represents that k is present in x and $y_k = 0$ otherwise. The goal is to create a model that outputs the probability of the presence of a category $p = [p_1, ..., p_K]$ [2], [5].

This project focuses on finding solutions to some of the problems that arise when training a network that assigns multiple labels to an input image: (i) the general multi-label learning issue accurately identifying multiple objects in an image under class imbalance, representing the complexity of real-world images, and (ii) the scenario where the training data underwent sparse supervision, which is often the case for data.

A. Feature Extraction

In the field of computer vision there exists different deep learning methods that extract features from input images, and the two main neural networks are convolutional neural networks and vision transformers.

B. Query2Label: A Simple Transformer Way to Multi-Label Classification

Query2Label: A Simple Transformer Way to Multi-Label Classification (Query2Label) by Liu et al. [5] is a two-stage framework for multi-label classification. It uses transformer decoders to extract features with multi-head attentions focusing on different parts of an object category and learn label embeddings from data automatically.

a) Backbone architecture: The CvT by Wu et al. [6].

C. Multi-Label Learning from Single Positive Labels

Multi-Label Learning from Single Positive Labels (MLSPL) by Cole et al. [2]. How it uses weak supervision and contrastive learning.

a) Online Estimation of Unobserved Labels:

$$\mathcal{L}_{ROLE}(\boldsymbol{F}_B, \tilde{\boldsymbol{Y}}_B) = \frac{\mathcal{L}'(\boldsymbol{F}_B)|\tilde{\boldsymbol{Y}} + \mathcal{L}'(\tilde{\boldsymbol{Y}}|\boldsymbol{F}_B)}{2} \tag{1}$$

for a batch B, where $\tilde{\boldsymbol{Y}} \in [0,1]^{N \times L}$ represent the estimated labels, and $\boldsymbol{F} \in [0,1]^{N \times L}$ is the matrix of classifier predictions.

The goal is to train the label estimator $g(\cdot;\phi)$ and the image classifier $f(\cdot;\theta)$

IV. Methodology

In real-world datasets, obtaining full label annotations is practiaclly impossible. This project focuses on finding solutions to the multi-label learning problems:

- The challenge of multi-label learning in scenarios where only a single positive label per image is available durin training (MLSPL).
- Label imbalance (Query2Label).
- Feature localization (Query2Label).

A. Overview of Approach

A brief description of what we did to solve the multi-label classification problem.

B. Dataset

The MS-COCO 2014 [7] dataset is used as a benchmark for evaluation both Query2Label and MLSPL. MS-COCO (Microsoft Common Objects in Context) is a large-scale dataset commonly used for object detection, segmentation, and multi-label image classification. COCO consists of 82,081 training images and 80 classes, and a validation set of 40,137 images.

C. Query2Label

Write about why we chose Query2Label for solving the multi-label classification problems.

D. MLSPL

Write about why we chose MLSPL for solving the multi-label classification problem.

E. Experiments

A description of our setup versus the author's setups.

All experiments were run on a single NVIDIA GeForce RTX 3060 GPU (12 GB VRAM), with Python 3.11.8, NVIDIA driver 550.90.07 and CUDA 12.4.

1) Query2Label: All Query2Label experiments were conducted on environment versions cuda==12.4, torch==2.1.0+cu121, torchvision==0.16.0+cu121, python==3.11.8. We evaluated the Query2Label model using the best performing backbone model, the CvT-w24 backbone, with a 384 input resolution, pretrained on the ImageNet-22k dataset, as described by Liu et al. [5]. For our experiments, we used a pretrained model checkpoint released by the authors, which was trained on the MS-COCO 2024 dataset. The model was tested on the MS-COCO 2024 validation set with a batchsize of 16 and otherwise no fine-tuning or modification.

2) Multi-Label Learning from Single Positive Labels: All MLSPL experiments were conducted on environment versions cuda==12.4, torch==2.2.1+cu121, torchvision==0.17.1+cu121, python==3.11.8.

a) Data preparation: The dataset preparation for the MS-COCO dataset in MLSPL relies on converting the standard multi-label MS-COCO dataset to a single positive label format, simulating a weakly supervised setting. Cole et al. does this beginning with the fully annotated multi-label image dataset and corrupt it by discarding annotations, simulating single positive training data by randomly selecting one positive label for each training example [2].

To convert into a single positive label format, the instructions provided by Cole et al. are followed: firsly, both images and annotations for training and validation are downloaded. Secondly, pre-extracted features for COCO, provided by the authors are downloaded. Lastly, the script `format_coco.py` is used to produce uniformly formatted image lists and labels.

b) Hyperparameters:

F. Evaluation Metrics

To asses model performance, we adopt mean Average Precision (mAP), a standard metric widely reported in multi-label classification tasks as it is used to analyze the performance object detection and segmentation. Both Query2Label and MLSPL report results in terms of mAP.

V. Results and Discussion

This section compares results from the experiments to the those of the papers. Discuss why they might not be the same, or describe the similarities. Discuss whether the methods are able to solve the problems.

A. Query2Label

B. Multi-Label Learning from Single Positive Labels

VI. Conclusion

References

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

[2] E. Cole, O. M. Aodha, T. Lorieul, P. Perona, D. Morris, and N. Jojic, "Multi-label learning from single positive labels," 2021. [Online]. Available: https://arxiv.org/abs/2106.09708

[3] J. Bekker and J. Davis, "Learning from positive and unlabeled data: a survey," *Machine Learning*, vol. 109, no. 4, p. 719–760, Apr. 2020. [Online]. Available: http://dx.doi.org/10.1007/s10994-020-05877-5

[4] X. Li and B. Liu, "Learning to classify texts using positive and unlabeled data." 01 2003, pp. 587–594.

[5] S. Liu, L. Zhang, X. Yang, H. Su, and J. Zhu, "Query2label: A simple transformer way to multi-label classification," 2021. [Online]. Available: https://arxiv.org/abs/2107.10834

[6] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," 2021. [Online]. Available: https://arxiv.org/abs/2103.15808

[7] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015. [Online]. Available: https://arxiv.org/abs/1405.0312

TABLE I
COMPARISON OF MAP RESULTS BETWEEN OUR EXPERIMENTS AND REPORTED MAP RESULTS ON THE MS-COCO 2014 DATASET.

Method	**Backbone**	**Resolution**	**mAP(Ours)**	**mAP (Paper)**
Q2L-R101-448	ResNet-101	448×448	84.9	84.9
Q2L-R101-576	ResNet-101	576×576	86.5	86.5
Q2L-SwinL	Swin-L(22k)	384×384	90.5	90.5
Q2L-CvT	CvT-w24(22k)	384×384	91.3	91.3