

Technology Review:

Confidence-based Weighted Loss for Multi-label Classification with Missing Labels

Brendan Lynch¹

Department of Computer Science, University of Illinois at Urbana-Champaign, 201 S. Goodwin Ave, Urbana IL 61801, USA

Overview

The purpose of this paper is to review the methods described by Ibrahim et al.(2020) with respect to their applicability our project of classifying biology paper text into various biomimicry functions. This review will attempt evaluate the method, with a score of 1-10, against the following three criteria:

- **Understandability:** Are the proposed methods easy to comprehend within the timeframe of the semester/project?
- **Availability:** Is the source code available and well-documented? Or will the methods need to be implemented from scratch?
- **Applicability:** Does the method generalize to our project task and dataset?

Introduction

Multi-label text classification with missing labels (MLML) has been widely researched to overcome the challenge and costs of collecting high-quality and complete label sets for all samples in a classification dataset. Most research in this area has focused on predicting missing labels based on correlations between labels, however, there is currently no straight-forward manner in which a pre-trained model can be fine-tuned to an MLML dataset, especially since there is little to no information on which specific labels are missing or their locations. In their paper, "Confidence-based Weighted Loss for Multi-label Classification with Missing Labels" (2020), Ibrahim et al. propose to address this challenge with a weighted loss function that considers label confidence while training and a weighting schema that estimates the presence of missing labels. They attempt to demonstrate that their weighted loss method improves the performance of a pre-trained model while the ratio of missing labels increases.

Understandability: The described approach is analogous to the weighted loss functions used for solving class imbalances or partial labels. Score 9/10.

Availability: The authors provide all source code (with detailed comments) and evaluation datasets on their github repository. Score 10/10.

Applicability: Given our preference in using pre-trained transformer models for our project, this MLML method is particularly applicable. Score 9/10.

Conclusion: Overall the methods described by Ibrahim et al. should be useful for our tasks of classifying biology paper text into biomimicry functions.

References

Ibrahim, K., Epure, E., Peeters, G., and Richard, G. (2016). Confidence-based Weighted Loss for Multi-label Classification with Missing Labels. ICMR '20: Proceedings of the 2020 International Conference on Multimedia Retrieval. June 2020. Pages 291-295.
<https://doi.org/10.1145/3372278.3390728>
<https://github.com/KarimMibrahim/Sample-level-weighted-loss.git>