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

Karim M Ibrahim, Elena Epure, Geoffroy Peeters, Gael Richard

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

Karim M. Ibrahim
LTCI, Télécom Paris, Institut Polytechnique de Paris

Elena V. Epure
Deezer Research

Geoffroy Peeters
Gaël Richard
LTCI, Télécom Paris, Institut Polytechnique de Paris

ABSTRACT

The problem of multi-label classification with missing labels (MLML) is a common challenge that is prevalent in several domains, e.g. image annotation and auto-tagging. In multi-label classification, each instance may belong to multiple class labels simultaneously. Due to the nature of the dataset collection and labelling procedure, it is common to have incomplete annotations in the dataset, i.e. not all samples are labelled with all the corresponding labels. However, the incomplete data labelling hinders the training of classification models. MLML has received much attention from the research community. However, in cases where a pre-trained model is fine-tuned on an MLML dataset, there has been no straightforward approach to tackle the missing labels, specifically when there is no information about which are the missing ones. In this paper, we propose a weighted loss function to account for the confidence in each label/sample pair that can easily be incorporated to fine-tune a pre-trained model on an incomplete dataset. Our experiment results show that using the proposed loss function improves the performance of the model as the ratio of missing labels increases.

CCS CONCEPTS

• **Computing methodologies** → *Neural networks.*

KEYWORDS

Multi-label classification; missing labels; neural networks

ACM Reference Format:
Karim M. Ibrahim, Elena V. Epure, Geoffroy Peeters, and Gaël Richard. 2020. Confidence-based Weighted Loss for Multi-label Classification with Missing Labels. In *Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR '20), June 8–11, 2020, Dublin, Ireland.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3372278.3390728

1 INTRODUCTION

Multi-label classification [14, 30] is a common task in various research fields, such as audio auto-tagging [1], image annotation [3], text categorization [23], and video annotation [25]. Multi-label classification is concerned with the problem of predicting multiple correct labels for each input instance. A relevant problem to multi-label classification is missing labels. Collecting a multi-label dataset is a challenging and demanding task that is less scalable than collecting a single-label dataset [10]. This is because collecting a consistent and complete list of labels for every sample requires significant effort. It is shown in [40] that learning with corrupted labels can lead to very poor generalization performances. Various strategies in dataset collection, such as crowdsourcing platforms like Amazon Mechanical Turk[1] or web services like reCAPTCHA[2], lead to datasets with a set of well-labelled positive samples and a set of missing negative labels. The set of missing labels is often not known. Hence, this problem of MLML is different from the problem of partial labels [11], where the position of the missing labels is known but its value is unknown, and noisy labels [31] where a set of both positive and negative labels are corrupted.

The problem of MLML is a common challenge in previous research that received much attention [2, 4, 6, 12, 24, 34, 38, 39]. Most of the previous approaches relied on exploiting the correlation between labels to predict the missing negative labels [2, 6, 34, 37]. However, the state-of-the-art approaches in MLML [15, 17] are not easily usable in cases where a pre-trained model is used. They either rely on jointly learning the correlations between the labels along with the model parameters, require prior extraction of manually engineered features for the task [15], or assume the location of the missing labels is known but the value is missing [17]. These methods do not allow to fine-tune a pre-trained model on a dataset with missing labels. This is limiting because it has been shown that fine-tuning a pre-trained architecture is useful and, in most cases, gives superior results to models trained from scratch [18, 28]. Multiple domains exploit existing pre-trained models especially when access to large annotated data is challenging, such as medical image classification [13, 22, 42], or when access to resources and computation power to fully train a complex model are scarce.

In this work, we propose a solution to address MLML with pre-trained models. In this direction, our main contributions are:

(1) Present a new weighted loss function that accounts for the confidence in the labels while training in a way that is easily usable to fine-tune pre-trained models.
(2) A weighting schema per sample per label to estimate the presence of the missing labels, unlike previous work where the location of missing labels is assumed to be known [11, 17]
(3) We demonstrate the benefit of our approach on synthetic experimental setup by comparing the performance of the

ICMR '20, June 8–11, 2020, Dublin, Ireland

ACM ISBN 978-1-4503-7087-5/20/06…$15.00
https://doi.org/10.1145/3372278.3390728

[1] https://www.mturk.com/
[2] https://www.google.com/recaptcha/

fine-tuned models with the weighted loss compared to non-weighted loss on two different datasets. The improvement is consistent across different ratios of missing labels[3].

2 PREVIOUS WORK

Label correlations has been frequently used in several approaches on multi-label classification to predict the missing labels in the dataset. For example, [5, 36] proposed using a matrix completion approach to predict the missing labels. Similarly, [7, 10] learned correlation between the categories to predict some missing labels. Additionally, [35] used a mixed graph approach to discover label dependencies to recover missing labels. Recently, [15] proposed an approach to jointly learn independent binary classifiers, while also learning label correlation for a multi-label classifier. However, most of the recent approaches are not usable to train a deep neural network with unknown missing labels. They either assume the location of the missing labels is known, or require solving an optimization problem with the training set in memory, which is not practically usable to fine-tune a pre-trained model.

Unlike these methods, we propose an approach that is scalable and usable to fine-tune a pre-trained model. To train our model, we introduce a new loss function that accounts for the confidence in the labels. Weighted loss functions is a common approach for different problems, e.g. to solve class imbalance [32], to focus on samples that are harder to predict [20], or to solve a similar problem of partial labels [11]. However, to our knowledge, this is the first attempt to use a per sample per label weighted loss for missing labels where the missing labels are unknown.

3 PROPOSED APPROACH

We propose to modify the binary cross entropy loss to account for the confidence in the missing labels. This can be done by adding weighting factors to our loss function. We apply confidence-based weight per sample for each of the positive and negative labels independently. We hypothesise that using these weights can improve our model performance in predicting the correct label by giving less weight to samples with low confidence in their label.

Formally, let $\mathbb{X} = \mathbb{R}^d$ denote the d-dimensional space for the input vector, $\mathbb{Y} = \{0, 1\}^m$ denote the label space marking the absence or presence of each of the m labels for each instance. The task of multi-label classification is to estimate a classifier $f : \mathbb{X} \mapsto \mathbb{Y}$ using the labelled dataset $D = \{(\mathbf{x_i}, \mathbf{y_i})|1 < i \leq n\}$.

We can describe our classifier as $f(\mathbf{y_i}|\mathbf{x_i}, \theta)$, which estimates the labels $\mathbf{y_i}$ for the given sample $\mathbf{x_i}$, while θ represents the trainable parameters of the model. The model parameters are trained by minimizing a loss function $J(D, \theta)$ that describes how the model is performing over the training examples. In multi-label classification, it is common to use the binary cross entropy loss:

$$CE(\mathbf{x_i}, \mathbf{y_i}) = -\sum_{c=1}^{m} y_{i,c} log(f_c(\mathbf{x_i})) + (1 - y_{i,c}) log(1 - f_c(\mathbf{x_i})) \quad (1)$$

where $y_{i,c}$ is the c^{th} label in $\mathbf{y_i}$ and $f_c(\mathbf{x_i})$ is the output of the classifier f corresponding to the c^{th} label.

[3]https://github.com/KarimMibrahim/Sample-level-weighted-loss.git

The cross entropy is made of two terms, one is "active" when the label is positive while the second is zero, and vice versa. We propose to modify each term to add a weighting factor, one relative to the confidence in the positive label and a second one relative to the confidence of a negative label for each sample.

$$CE_{proposed}(\mathbf{x_i}, \mathbf{y_i}) = -\sum_{c=1}^{m} \omega_{i,c} y_{i,c} log(f_c(\mathbf{x_i})) + \bar{\omega}_{i,c}(1 - y_{i,c}) log(1 - f_c(\mathbf{x_i})) \quad (2)$$

where $\omega_{i,c}$ represents the confidence in the positive label, while $\bar{\omega}_{i,c}$ represents the confidence in the negative label.

Estimating the weights

The weights used in the proposed loss function depend vastly on the problem in hand. In most cases, the missing labels exist in the negative labels while the positive labels are complete, i.e. $\omega_{i,c} = 1$. $\bar{\omega}_{i,c}$ depends on the information we have about the collection of the dataset. However, a common way to estimate them is by using the labels correlation [34]. Even if missing, a label could still be correctly inferred, signalled by frequently co-occurring labels associated with the same sample. The weights can be estimated as:

$$\bar{\omega}_{i,c} = P(y_{i,c} = 0|\mathbf{y_i}) \quad (3)$$

which corresponds to the probability of having a negative label for the c^{th} label given the vector of labels $\mathbf{y_i}$ for the point $\mathbf{x_i}$. This probability can be estimated from the ground-truth label matrix based on label co-occurrences. However, we propose using positive label co-occurrences when computing $\mathbf{y_i}$, and ignore the negative labels since there is lower confidence in them, i.e. we would not rely on missing labels to estimate the missing labels. Regarding the positive weights $\omega_{i,c}$, while less common to have low confidence in the positive labels, a potential approach is using inter-raters agreement in cases where the labels are crowd-sourced to put emphasize on samples with higher agreement.

4 EXPERIMENTS

To validate the advantage of using the weighted cross entropy, we compare between the performance of the same model trained with original cross entropy and with the weighted cross entropy across different ratios of artificially created missing labels. Specifically, we use the proposed loss on the problem of image classification with multiple labels. Image classification is a popular problem with multiple approaches proposed to it and a vast repertoire of pre-trained models on large datasets. We apply one of the commonly used pre-trained models, which is inception-resnet v2 [27], on two different datasets: MSCOCO [21] and NUS-WIDE [8]. We use two different schemes for computing the weights:

(1) Setting the weights for the missing labels to zero and one otherwise (by using our knowledge of which labels are missing, which is not the case in most real-world datasets) referred to as `ignore missing weighted cross entropy (IM-WCE)`;
(2) Estimating them using label co-occurrences, referred to as `correlation-based weighted cross entropy (CB-WCE)`.

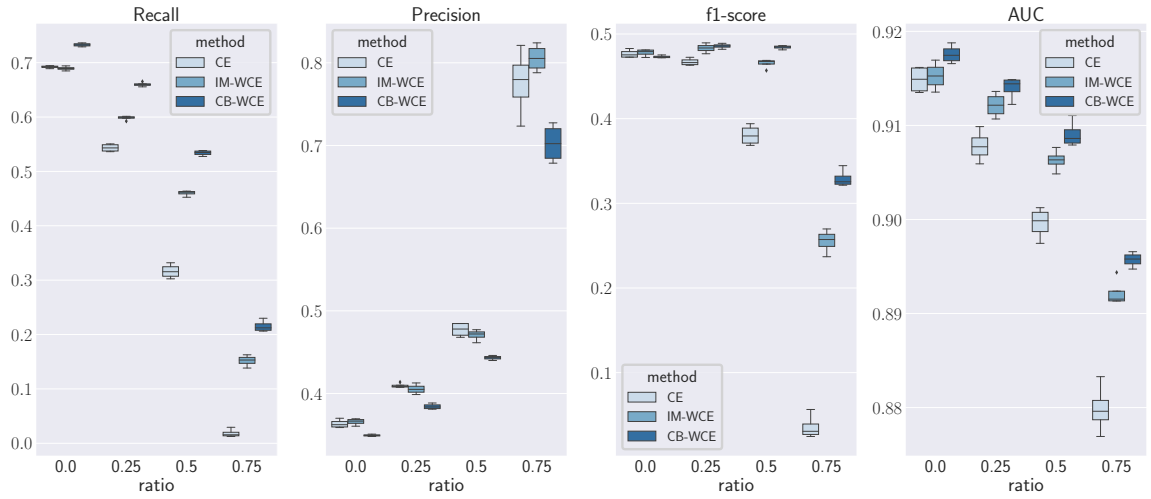

Figure 1: Results of the weighted cross entropy loss and original cross entropy on the MSCOCO dataset with different ratios of missing labels

Datasets

The experiment requires a strongly labelled multi-label dataset with no missing labels at the start. Hence, we decided to work with MSCOCO[4] [21] which is commonly used in multi-label classification with and without missing label [11, 16, 19, 33]. The dataset is originally intended for image segmentation, but is also usable for image classification. We use the 2017 version of the dataset. The dataset contains ~122k images and 80 classes. However, after filtering out the samples with less than 4 labels, the total number of images drops to ~33k images.

The second dataset is NUS-WIDE[5] [8], which is another image classification dataset that is suitable for this problem and also commonly used in the multi-label classification studies along with MSCOCO. The dataset contain ~270k images and 81 classes. However, the number of images drops to ~24k images after filtering out samples with less than 3 labels per sample. We reduced the threshold to 3 for this dataset because it has less labels co-occurrences compared to MSCOCO.

Creating artificial missing labels

An important part of the experiment is creating missing labels in the training dataset. We propose to create missing labels with different ratios. We follow a similar procedure to [11]. We hide the labels randomly as a ratio of the complete labels per image, i.e. we hide $x_i = r * n_i$ labels for each image, where r is the ratio of labels to hide, n_i is the total number of positive labels of the image i, and x_i is the corresponding number of labels to hide in this image. We use ratios of 0.0, 0.25, 0.5, and 0.75 missing labels to complete labels.

Classification model

We propose to use a pre-trained classification model for the task of image classification that needs to be fine-tuned to a different dataset with missing labels. Previous papers on multi-label classification used models as VGG16 [19] and resnset-101 [11]. The exact architecture of the model is not the focus of this work and would not have a significant effect on the comparison between the two losses. Hence, we used the inception-resnet v2 [27], which is one of the best performing models in image classification, pretrained on the ImageNet dataset [9] through the TensorFlow pre-trained models library[6].

Evaluation Procedure

We perform a 4-fold cross-validation, each with the aforementioned ratios of missing labels in the training sets and no missing labels in the test sets. We evaluate the model performance using standard multi-label classification metrics: Precision, Recall, f1-score and AUC [41], all computed with 'micro' averaging to account for the large number of classes with few samples [26]. The effect of the missing labels is specifically prominent in predicting the positive labels correctly. As the ratio of missing labels increases, the models learn to predict all zeros. Hence, the selected metrics are useful in evaluating the model's performance particularly in these cases.

4.1 Evaluation Results

Figure 1 shows the results obtained when training with the 3 different losses on the MSCOCO dataset on different ratios of missing labels. It shows that using the weighted loss clearly improves the performance of the model in terms of recall, f1-score and ROC-AUC, with an expected decrease in the precision. The decrease in the precision is explained by the fact that the model trained with the unweighted loss is learning to predict mostly zeros and the few samples that are predicted as positive are more likely to be correct. This is evident when observing the recall and the f1-score results alongside the precision. Additionally, the improvement is larger as the ratio of missing labels increases. We can also notice the effect

[4]http://cocodataset.org
[5]https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDE.html
[6]https://github.com/tensorflow/models/tree/master/research/slim

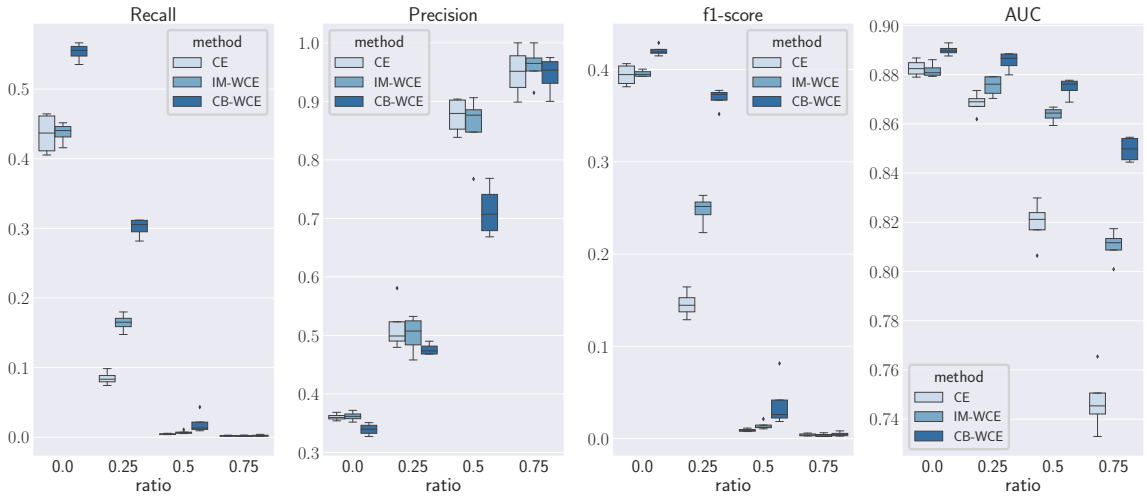

Figure 2: Results of the weighted cross entropy loss and original cross entropy on the NUS-WIDE dataset with different ratios of missing labels

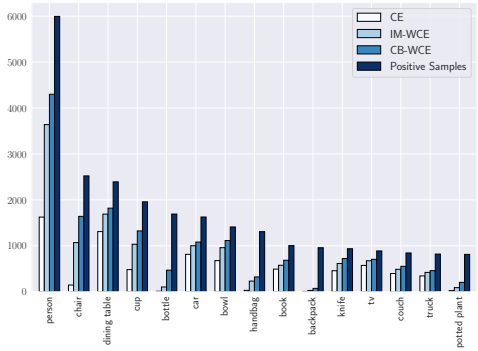

Figure 3: Comparison of the correctly predicted positive samples between the different methods

of missing labels on the performance of the model. The higher the ratio of the missing labels is the worse the model performs. Moreover, we find that using the correlation as weights generally gives better results in all cases even in the case where there are no missing labels. This is understandable since using the correlations in training a multi-label classifier leads the model to learn the underlying relationship between labels [29].

Similarly in Figure 2, we find the results of applying the different loss function on the NUS-WIDE dataset. We find a similar pattern in the performance of the model across different values of missing labels which shows the advantage of using the weighted loss function. However, as NUS-WIDE shows less correlation and co-occurrences between the labels on average compared to MSCOCO, the improvement is less impactful, yet evident.

Additionally, Figure 3 shows a comparison of the true positives of each of the methods for the 15 most frequent classes in the MSCOCO dataset with a ratio of 0.5 missing labels. It is evident that the correlation based (CB-WCE) gives superior results in all classes even compared to (IM-WCE). However, the improvement is particularly noticeable in certain classes, such as "bottle" and "chair", which we interpret such that some classes that are harder to learn becomes easier when emphasizing their co-occurrences with other classes.

Considering the evaluation results on these two different datasets, we can advise towards using the weighted loss for multi-label classification when missing labels are present. We experimented with using the correlations to weight the loss function and concluded an evident improvement across different evaluation metrics. While there are various proposed solutions for missing labels, our proposal is particularly more suitable to be used in the cases of fine-tuning a pre-trained model, or even in cases where a specific deep learning architecture is preferred to be used and a simple modification in the loss is needed to account for the missing labels.

5 CONCLUSION

In this paper, we presented a weighted loss function for multi-label classification with missing labels. The weighted loss depends on estimating the confidence in the labels when used in training followed by adjusting the loss accordingly. Hence, sample/label pairs with higher confidence contribute more in the learning process. The proposed approach show a clear improvement compared to original unweighted loss. The proposed approach is simple to integrate in pre-trained models, which is a relevant solution that has not been previously explored in the literature. Future work includes studying additional weighting schemes for estimating the missing labels for both the positive and negative cases and in other domains such as text and audio classification and other scenarios for creating the artificial missing labels.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068.

REFERENCES

[1] Thierry Bertin-Mahieux, Douglas Eck, and Michael Mandel. 2011. Automatic tagging of audio: The state-of-the-art. In *Machine audition: Principles, algorithms and systems*. IGI Global, 334–352.

[2] Wei Bi and James T Kwok. 2014. Multilabel classification with label correlations and missing labels. In *Proceedings of 28th AAAI Conference on Artificial Intelligence*.

[3] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. 2004. Learning multi-label scene classification. *Pattern recognition* 37, 9 (2004), 1757–1771.

[4] Serhat Selcuk Bucak, Rong Jin, and Anil K. Jain. 2011. Multi-label learning with incomplete class assignments. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society.

[5] Ricardo S Cabral, Fernando Torre, João P Costeira, and Alexandre Bernardino. 2011. Matrix completion for multi-label image classification. In *Proceedings of the Advances in neural information processing systems*.

[6] Gang Chen, Yangqiu Song, Fei Wang, and Changshui Zhang. 2008. Semi-supervised multi-label learning by solving a sylvester equation. In *Proceedings of the 2008 SIAM International Conference on Data Mining*.

[7] Minmin Chen, Alice Zheng, and Kilian Weinberger. 2013. Fast image tagging. In *Proceedings of the International conference on machine learning*.

[8] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: a real-world web image database from National University of Singapore. In *Proceedings of the ACM international conference on image and video retrieval*.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[10] Jia Deng, Olga Russakovsky, Jonathan Krause, Michael S Bernstein, Alex Berg, and Li Fei-Fei. 2014. Scalable multi-label annotation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

[11] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. 2019. Learning a Deep ConvNet for Multi-label Classification with Partial Labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[12] Charles Elkan and Keith Noto. 2008. *Learning Classifiers from Only Positive and Unlabeled Data*.

[13] Mingchen Gao, Ulas Bagci, Le Lu, Aaron Wu, Mario Buty, Hoo-Chang Shin, Holger Roth, Georgios Z Papadakis, Adrien Depeursinge, Ronald M Summers, et al. 2018. Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 6, 1 (2018), 1–6.

[14] Eva Gibaja and Sebastián Ventura. 2015. A tutorial on multilabel learning. *ACM Computing Surveys (CSUR)* 47, 3 (2015), 52.

[15] Zhi-Fen He, Ming Yang, Yang Gao, Hui-Dong Liu, and Yilong Yin. 2019. Joint multi-label classification and label correlations with missing labels and feature selection. *Knowledge-Based Systems* 163 (2019), 145–158.

[16] Mengying Hu, Hu Han, Shiguang Shan, and Xilin Chen. 2018. Multi-label Learning from Noisy Labels with Non-linear Feature Transformation. In *Proceedings of the Asian Conference on Computer Vision*.

[17] Jun Huang, Feng Qin, Xiao Zheng, Zekai Cheng, Zhixiang Yuan, Weigang Zhang, and Qingming Huang. 2019. Improving multi-label classification with missing labels by learning label-specific features. *Information Sciences* 492 (2019), 124–146.

[18] Simon Kornblith, Jonathon Shlens, and Quoc V Le. 2019. Do better imagenet models transfer better?. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[19] Yuncheng Li, Yale Song, and Jiebo Luo. 2017. Improving pairwise ranking for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European conference on computer vision*.

[22] Jan Margeta, Antonio Criminisi, R Cabrera Lozoya, Daniel C Lee, and Nicholas Ayache. 2017. Fine-tuned convolutional neural nets for cardiac MRI acquisition plane recognition. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 5, 5 (2017), 339–349.

[23] Andrew McCallum. 1999. Multi-label text classification with a mixture model trained by EM. In *Proceedings of the AAAI workshop on Text Learning*.

[24] Olivier Petit, Nicolas Thome, Arnaud Charnoz, Alexandre Hostettler, and Luc Soler. [n.d.]. *Handling Missing Annotations for Semantic Segmentation with Deep ConvNets*. Technical Report.

[25] Guo-Jun Qi, Xian-Sheng Hua, Yong Rui, Jinhui Tang, Tao Mei, and Hong-Jiang Zhang. 2007. Correlative multi-label video annotation. In *Proceedings of the 15th ACM international conference on Multimedia*.

[26] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management* 45, 4 (2009), 427–437.

[27] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*.

[28] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging* 35, 5 (2016), 1299–1312.

[29] Grigorios Tsoumakas, Anastasios Dimou, Eleftherios Spyromitros, Vasileios Mezaris, Ioannis Kompatsiaris, and Ioannis Vlahavas. 2009. Correlation-based pruning of stacked binary relevance models for multi-label learning. In *Proceedings of the 1st international workshop on learning from multi-label data*.

[30] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. 2009. Mining multi-label data. In *Data mining and knowledge discovery handbook*. Springer, 667–685.

[31] Arash Vahdat. 2017. Toward robustness against label noise in training deep discriminative neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*.

[32] Byron C Wallace, Kevin Small, Carla E Brodley, and Thomas A Trikalinos. 2011. Class imbalance, redux. In *Proceedings of the IEEE 11th international conference on data mining*.

[33] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. 2016. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

[34] Baoyuan Wu, Zhilei Liu, Shangfei Wang, Bao-Gang Hu, and Qiang Ji. 2014. Multi-label learning with missing labels. In *Proceedings of 22nd International Conference on Pattern Recognition*.

[35] Baoyuan Wu, Siwei Lyu, and Bernard Ghanem. 2015. Ml-mg: Multi-label learning with missing labels using a mixed graph. In *Proceedings of the IEEE international conference on computer vision*.

[36] Miao Xu, Rong Jin, and Zhi-Hua Zhou. 2013. Speedup matrix completion with side information: Application to multi-label learning. In *Proceedings of the Advances in neural information processing systems*.

[37] Miao Xu, Gang Niu, Bo Han, Ivor W Tsang, Zhi-Hua Zhou, and Masashi Sugiyama. 2018. Matrix Co-completion for Multi-label Classification with Missing Features and Labels. *arXiv preprint arXiv:1805.09156* (2018).

[38] Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon. 2014. Large-scale multi-label learning with missing labels. In *International conference on machine learning*.

[39] Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon. 2014. Large-scale multi-label learning with missing labels. In *Proceedings of the International conference on machine learning*.

[40] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530* (2016).

[41] Min-Ling Zhang and Zhi-Hua Zhou. 2013. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering* 26, 8 (2013), 1819–1837.

[42] Zongwei Zhou, Jae Shin, Lei Zhang, Suryakanth Gurudu, Michael Gotway, and Jianming Liang. 2017. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.