Robust Multi-label Active Learning for Missing Labels

J.Rozen[1] , **T.Younesian**[1] , **A.Ghiassi**[1] , **L.Chen**[1]
[1]TU Delft

Abstract

Multi-label classification has gained a lot of attraction in the field of computer vision over the past couple of years. Here, each instance belongs to multiple class labels simultaneously. There are numerous methods for Multi-label classification, however all of them make the assumption that either the training images are completely labelled or that label correlations are given. Since Active Learning is frequently used when not much data is available, it could be used to determine the missing labels by querying an oracle. This paper proposes a novel solution that combines the current state-of-the-art for Multi-label classification with Active Learning to infer the missing labels. This is done with sampling strategies that try to select the most informative sample from the dataset by exploring the amount of missing labels. With these strategies, we try to minimize the relabeling cost for all samples, while maximizing the information gained. The chosen method called Hard sampling with entropy then looks to select those samples that both the model and we find informative. The chosen measure along with the other measure are then explored and evaluated on a subset of the MSCOCO dataset on 20%, 40% and 60% noise. Hard sampling with entropy then outperforms the state-of-the-art by more then 30%, as well as the baseline sampling method by 2% for 60% noise.

1 Introduction

Most real-world images cannot just be labeled by one label, but with multiple labels. Here, an emerging extension of multi-class classification, Multi-label learning (MLL), tries to categorise an instance into multiple relevant labels, as opposed to the usual one class. For this reason Multi-Label learning and its variants are closer to actual real-world applications. In MLL, classification models are trained to predict these labels with high accuracy [1]. However acquiring a fully labeled dataset for a Multi-labeled scenario is a very expensive and time-consuming task.

The main problem this paper will seek to tackle are missing or partial labels [3]. An example of this can be found in **Figure 1**. This might not sound like a big problem, however as can be seen in **Figure 2**, the impact of noise is quite substantial. Missing labels occur more often than one would think, this could happen when the labels are collected via crowd-sourcing to reduce the labeling cost and effort. Furthermore, some applications train on incomplete label sets, and predicting the unavailable labels is preferred to reduce the annotation cost. A way to infer those missing

Figure 1: An illustration of Multi-label Learning with missing labels. On the left are all the correct labels for the image, on the right you can see that some labels are missing

labels would be with a method called Active Learning or AL [2]. This is a special case of machine learning in which a learning algorithm can make use of an oracle (or in the form of a human expert) to label new data points with the desired labels. To reduce the cost and effort that goes into labeling, Active Learning methods have been widely used in this field [6, 14–16].
Active Learning is therefore used to increase the accuracy of the model while keeping the labeling cost and effort to a minimum. By selecting the most informative samples from our dataset with partial labels, we can reduce the amount of queries to our oracle while increasing the performance of our model as much as possible.

The problem with some of the current studies [14–16]. On Multi-label active learning is that they ignore a crucial fact: Not all labels are available [4]. This then gives us the the following research question: How can one determine the missing labels using Active Learning?
This paper therefore proposes to use the current state-of-the-art [7] as a Multi-label classifier and evaluate new informativeness measures which will try to exploit the ratio of missing to given labels. A comparison between these new

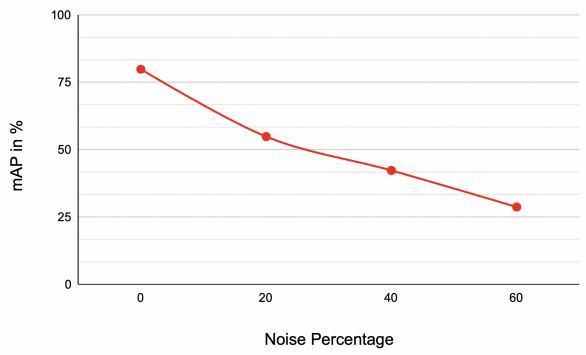


Figure 2: In this graph, the performance of the current state-of-the-art on various degrees of noise is shown. It is clear that missing labels have a significant impact on the performance of MLC.

informativeness measures to Random Sampling will be will be made on the MSCOCO [9] dataset, as this is another commonly used measure. Both the newly introduced measures as well some existing ones will be further elaborated upon in the following sections.

In the next section, we go into other related works. In section 3, we first formally introduce the main topics of this paper, after which we discuss the algorithm and the and present the new informativeness measures. Section 4 will first describe the setup in which experiments will be performed and then evaluate the methods described in section 3 as well as provide a comparison between the proposed approach and the baseline asymmetric loss model on varying amounts of noise. Reproducibility and ethical aspects of this research will be talked about in section 5. Finally we conclude this paper and and talk about future work in section 6.

2 Related Work

In this section we categorize other, related works into three main categories, Multi-label learning, Active learning and missing labels in MLL. Each of these categories briefly introduces relevant works and their proposed methods.

Multi-label learning Y Yu. et al. [13] perform Multi-label classification by exploiting label correlations. They introduced two novel MLC algorithms based on the variable precision neighborhood rough sets, called Multi-label classification using rough sets (MLRS) and MLRS using local correlation (MLRS-LC). These two algorithms consider two important factors that affect the accuracy of prediction, namely the existing uncertainty within the mapping between the feature space, and the label correlation. While Y. Xing et al. [20] introduces an approach called Multi-label Co-Training (MLCT). Their method addresses the class-imbalance challenge by leveraging the information concerning pairwise co-occurrence of labels. MLCT selects samples using a predictive reliability measure, and the labels of the chosen samples are communicated among the co-training classifiers by applying label-wise filtering.

Active learning Some of the existing works for AL include the method proposed by Hsu, W. N. and Lin, H. T. [2]. Here they design a learning algorithm that connects the well-known multi-armed bandit problem with AL. They suggest that it is possible to dynamically estimate the performance of different strategies by giving a appropriate choice for the multi-armed bandit learner. In the paper by Z. Zha et al. [12] they propose a novel active learning approach based on the optimum experimental design criteria in statistics. Their approach exploits the local structure of the sample as well as the samples density, relevance and diversity information. All the while making use of both labelled and unlabelled data.

Missing labels in Multi-label learning A variety of solutions already exist for the missing label problem. K. Ibrahim et al. [4], proposes accounting for the confidence in the missing labels by modifying the loss function. This can be done by adding weighting factors to binary cross entropy loss. This confidence-based weight per sample is applied for each of the negative and positive labels independently.
Two novel MLL strategies are proposed by S. Huang et al. [5], which use Active Learning to infer missing labels: a label cardinality inconsistency strategy and a max-margin prediction uncertainty strategy. These exploit the relative Multi-label classification margin structure on each unlabeled instance and the statistical label cardinality information, respectively, to measure the unified informativeness of unlabeled instances. Moreover, they investigate an adaptive integration framework of these two strategies by applying a novel approximate generalization error to measure the unified informativeness of unlabeled instances. The problem with these solutions is that they ignore correlations the labels might have.

T. Durand et al. [3] uses an iterative strategy based on Curriculum Learning to predict some missing labels. A scalable Generative Model for MLL with missing labels is proposed by V. Jain et al. [10]. The framework used couples an exposure model to account for whether the label is actually missing, together with a latent factor model for the binary label matrix. In this paper a AL solution to infer the missing labels is preferred however.

The approach by J. Wu et al. [6] is based on problem transformation, and makes use of conditional label dependence [17] to explore label correlations on the weak label problem. It uses this to construct a unified sampling strategy to evaluate the informativeness of each example-label pair and determine which unknown label needs to be acquired from oracle.

With this paper we seek to improve the current state-of-the-art rather come up with alternative solution. We will do this by combining the Asymmetric Loss [7] with Active Learning and using novel measures on which to select samples.

3 Methodology

This section is structured as follows, first Multi-label classification will be introduced after which the chosen classifier will be discussed. In the subsequent subsection the Active Learning and the informativeness measures will be introduced and elaborated on. Finally the algorithm will be explained it its entirety in the last sub-section. In the following sections 2 assumptions are made: First, we assume that for a given dataset, it is known how many unique labels are in the dataset and how many labels are missing per sample. Second, when a sample is presented to the oracle, the oracle fully labels this sample with all the correct labels.

3.1 Multi-label classification

Mutli-class classification is formally defined as follows: Assume that a sample x of size S represents an image of S pixels, $Y = y_1, y_2, ..., y_L$ represents a Multi-label set that contains L distinct labels. Given a labeled training set $\{T = (x_i, Y_i)|1 \leq i \leq n\}$ where n is the number of samples, $x \in X^{n \times S}$ and $Y_i \in Y$, MLC is used to learn a Multi-label classifier m $x :\rightarrow Y$.

The state-of-the-art for MLC is described in [7], and was used to design the classifier. Here we use the loss function introduced by [7], which modifies the Binary Cross Entropy by adding Asymmetric Focusing (1) and Asymmetric Probability Shifting (2).

Asymmetric Focusing separates the focusing levels of the positive and negative samples, this is done by introducing a positive and a negative focusing parameter γ^+, γ^- respectively. With these parameters the contribution of positive and negative samples can be better controlled, and despite their infrequency help the network learn meaningful features from positive samples. **Asymmetric Probability Shifting** fully discards negative samples with very low probability, i.e., it performs hard thresholding of very easy negative samples, shifted probability p_m is defined in (2). Let's define the shifted probability, pm, as:

$$\begin{cases} L_+ = (1-p)^{\gamma_+} \log(p) \\ L_- = p^{\gamma_-} \log(1-p) \end{cases} \tag{1}$$

$$p_m = max(p - m, 0) \tag{2}$$

Combined we obtain the following for the Asymmetric Loss Function:

$$\begin{cases} L_+ = (1-p)^{\gamma_+} \log(p) \\ L_- = (p_m)^{\gamma_-} \log(1-p_m) \end{cases} \tag{3}$$

This loss function (3) is then combined with the medium TResNet [18] model that was pre-trained on the ImageNet dataset [19] and will then be used as our classifier.

3.2 Active Learning

Before elaborating on the measures, we distinguish as well as introduce 2 types of sampling strategies, instance based and example-label pair based. The instance based sampling strategy selects a data sample from the dataset based on the informativeness of the data sample as a whole. While the example-label pair strategies take each example-label pair and measure their informativeness. Using an intelligent sampling strategy is therefore key to reducing the labeling cost to the greatest extent possible. The samples are added in batches, and after every batch is added the model is re-trained and validated on the same test set.

The methods proposed by this paper will be introduced below. Using our first assumption, we can then derive the amount of known labels (labels that are given/not missing) and unknown labels (missing labels). The informativeness of a sample based is then on the ratio between these amounts. The proposed methods seek to exploit the fact that the total number of classes and the amount of given are known. Using this we can determine the amount of missing labels per sample. We then define the following, let U be the set of all the unlabelled/noisy instances, let $n_{unknown}$ and n_{known} respectively be the number of missing labels and given labels for a set n, with $n \in U$.

1. Medium Sampling: Select sample where the ratio between, unknown to known labels is the closest to 1 (See E.q. 4). With this we label the most "medium" samples, meaning choose samples where we can obtain a good amount of information from the chosen sample without having to relabel a lot of labels per sample.

$$x^* = argmin_{u \in U} |\frac{|u_{unknown}|}{|u_{known}|} - 1| \tag{4}$$

2. Hard Sampling: Select sample where the ratio between, unknown to known labels is the highest (See E.q. 5). These are the Hardest samples as since very little is known about them. In this case every sample added gives as much information as possible.

$$x^* = argmax_{u \in U} \frac{|u_{unknown}|}{|u_{known}|} \tag{5}$$

3. Hard and Medium: Sample the first 10% to be Hard sampling and the remainder with Medium. With this measure it is expected that it will not outperform the Hard Sampling when it comes to accuracy increase, but the ratio of samples labelled per percentage increase might be lower. A thorough evaluation of various splits will be presented in the Hard + Medium sampling section of the appendix.
4. Uncertainty: Otherwise known as entropy (3), where p_i is simply the frequentist probability of an label i in our data, this computed for all L labels in our dataset. Entropy is a measure of uncertainty or disorder, it represents the uncertainty of a classifier when classifying a sample. It is crucial to incorporate this measure as we can use this to see with which samples the classifier struggles the most with. The higher this value the more classifier can learn by having the true labels of the sample. Then for $x_j \in X$:

$$u(x_j) = \sum_{i=1}^{L} -p_i \ln p_i \tag{6}$$

All the measures mentioned above will be combined with the entropy. This is to combine both what we deem informative as well as to see what the model finds informative.

3.3 Full Algorithm

The full algorithm then looks as follows, first we start by splitting our dataset. A clean set of around 10% of the dataset where all the all the labels are known, A test set of 10% or more, and an unlabelled set where some of the labels are missing. A classifier will then be trained on the clean set, after which it will be tested on the test set. After training and testing, a certain number of samples are selected from the unlabelled set, using one of the measures, fully labelled by the oracle and added into the clean set. We then repeat this until some stopping criterion has been achieved. Pseudocode for the algorithm can be found below in **Algorithm 1**. Implementation for the measures were not included as they a very straightforward.

Algorithm 1: ALASL

Input : Pretrained model M, Clean test set s_{test}, Clean train set s_{train}, Noisy set s_{noise}, Measure_nr m, Oracle O Number of Iterations n Samples per Iteration t

Output: mAP per iteration and Cost for all samples relabelled

```
Initialize i and l to 0
Initialize maps to a List of size n
while i < n do
    train M on s_train
    maps[i] = validate M on s_test
    if m uses entropy then
        e = get entropy from validating M on s_noise
    end
    c = 0
    while c < t do
        x = pick sample from s_noise with or without e
        p-1 = nr. missing labels in x
        c = c + 1
        l = l + p_-1
        Add relabel x and add it to s_train
        Remove x from s_noise
    end
    i = i + 1
end
```

4 Evaluation

To determine which out of the previously discussed measures is optimal, thorough evaluation needs to be performed. This section will start with brief explanation on some of the decisions made followed by describing the setup in which these measures were tested. The results will then be presented and briefly discussed in the Results section. While a more thorough examination of the results will be performed in the Discussion section.

4.1 Design choices and label noise

First the dataset, the results were obtained by training and testing on a subset of the entire MSCOCO [9] dataset, namely the val2017 dataset with 5000 images. A subset of the dataset was chosen due to size and cost of training on the full MSCOCO, given the current hardware constraints. It is very clear that there exist a very significant imbalance between the occurrences of the labels in MSCOCO as can be seen in [9], to counteract this, the measures below only use example based sampling. This is done so that the occurrence of a labels will be independent form the samples selected. Some of the hyper-parameters were scaled down from the ones used in the original ASL code, to reduce the computing time of the algorithm. The changed hyper-parameters include, a batch size of 16 and image size of 224x224.

After obtaining the dataset, noise needs to be injected into the samples to simulate the missing labels. Since clean train and test sets are needed, this noise was only injected in the unlabelled set. This was done by way of collecting all the labels and replacing some percentage of the labels by -1. Since the noise was applied on the all samples the spread is not uniform, meaning some labels have more missing labels, while other may have little to no noise. Various degrees of noise were injected into the dataset, 20%, 40% and 60%. By this we mean that 20% of the **labels** are missing, not 20% of the samples have some missing labels.

Regarding sampling, we evaluate each measure on a budget of 500 samples added in 10 iterations of 50 samples. The decision to opt for adding the samples in batches instead of one batch, was made to better observe the growth of the classifier as more samples were introduced. There is no need to add all samples in the unlabelled set as we are only looking to add the ones that give us the most information.

Before adding a sample the total cost needs to be incremented by the cost of relabelling this sample. This cost is then defined as the amount of labels that need to be relabelled by the oracle. Relabelled labels was chosen instead of samples to clearly differentiate between the cost of picking different samples. Otherwise the cost to relabel a sample with only 1 missing label is equal to the cost of relabelling one with 40 missing labels.

The measures will be compared on 20%, 40% and 60% label noise using the following 3 heuristics:

- Highest mAP (mean Average Precision), meaning the measure with which we obtain the highest mAP after adding all the samples.
- Highest I, where I is defined as the percentage increase per label added (See E.q. 7). Where $a_{x,i}$ and $c_{x,i}$ are the mAP and cost of measure x at iteration i respectively.

$$I(x) = \frac{a_{x,10} - a_{x,0}}{c_{x,10}} \tag{7}$$

- Highest mAP at cost equal to the cost of Random Sampling in iteration 10. With this heuristic we aim to look

at the performance of the different measures at a fixed cost. The cost of Random Sampling at iteration 10 was chosen, since it is our baseline, and its reaches it highest mAP at that iteration. We want to look at how other measure perform against it when the cost is equal.

40% noise

Finally, the final mAP of each iteration is the average of the of the accuracies on the test set in that iteration. Since the code for training and validation were adapted from Asymmetric Loss paper, every epoch after training, the model is validated on test set. To then obtain the final mAP, the mean of the accuracy on the validation set is computed.

4.2 Setup

The setup would then be the following: Train the pre-trained classifier with ASL on the clean dataset for 20 epochs and select 50 new samples with the current informativeness measure from the unlabelled set for relabelling. After relabelling the noisy samples, add them to the clean set, count the amount of labels that were relabelled for the cost and repeat for 10 iterations.

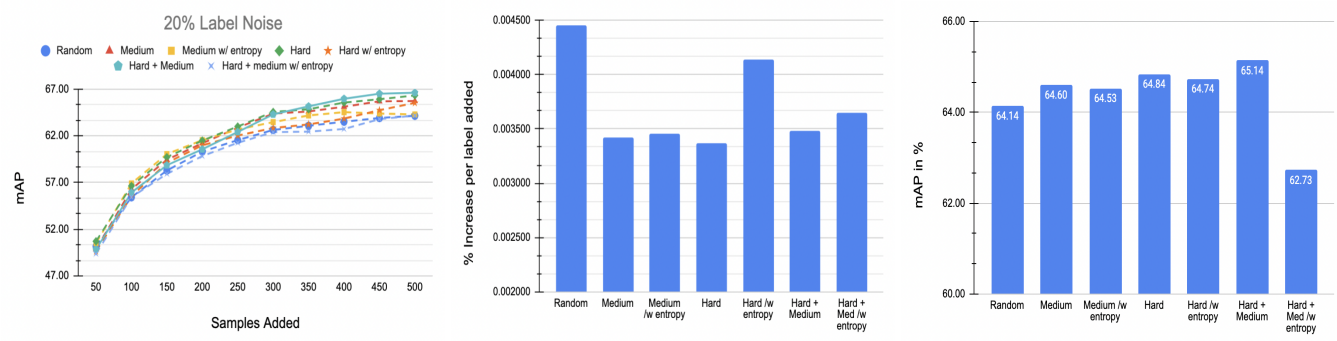
4.3 Results

Here the performance as well as the cost of each of the measures will be shown to 20%, 40% and 60% noise. Besides the measures mentioned in 3, one additional measure is displayed, Random sampling. This method picks one random samples out of the unlabelled set for relabeling. This will be used as a baseline to compare the other measures with.

Although not included in the graphs below, the initial mAP for the classifier without any labels added is approximately 28. This was excluded to make the graphs more readable, as it should be equal among all measures. At 60% noise, some measures had a cost lower than that of Random sampling in their final iteration. Therefore the lowest cost (Medium sampling with a cost of 21000) was used to perform the comparison.

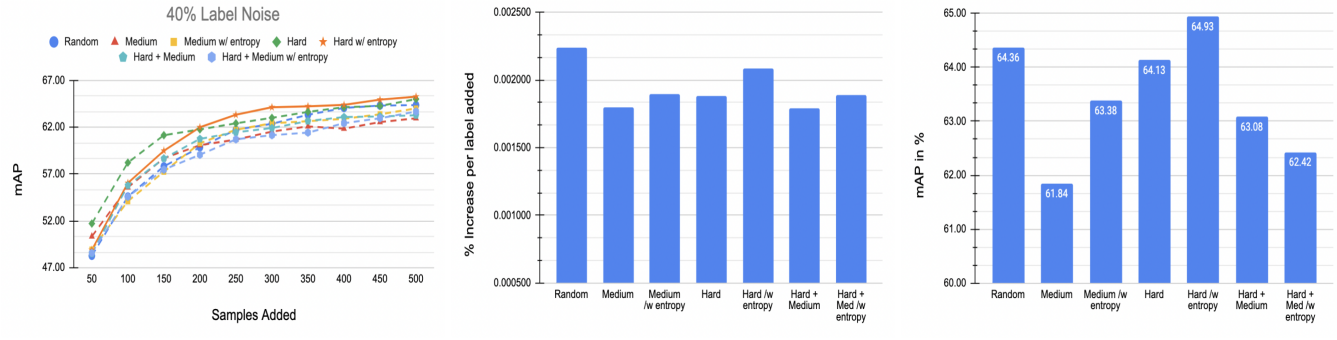
20% noise

The results in **Figure 3** show that although Hard + Medium Sampling has the highest mAP both in graph 3a as well as in 3c , the % increase per label is very low, the opposite is true for Random sampling. Hard Sampling with entropy however, performs quite well on all 3 measures.



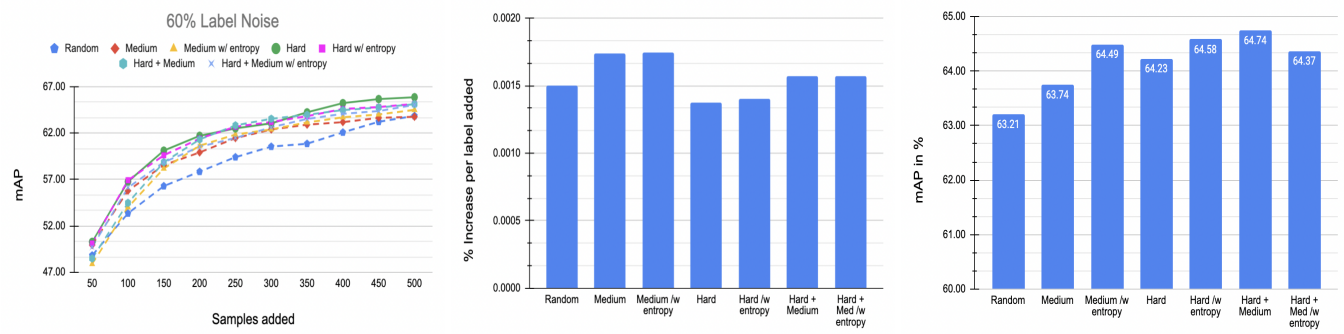
(a) (b) (c)

Figure 3: The evaluation of the measures on 20% noise (a) The mAP per 50 added samples on 20% label noise for all informativeness measures (b) The performances of the measures on 20% noise with labeling cost of approximately 8000 (c) The increase per label of all measures on 20% noise



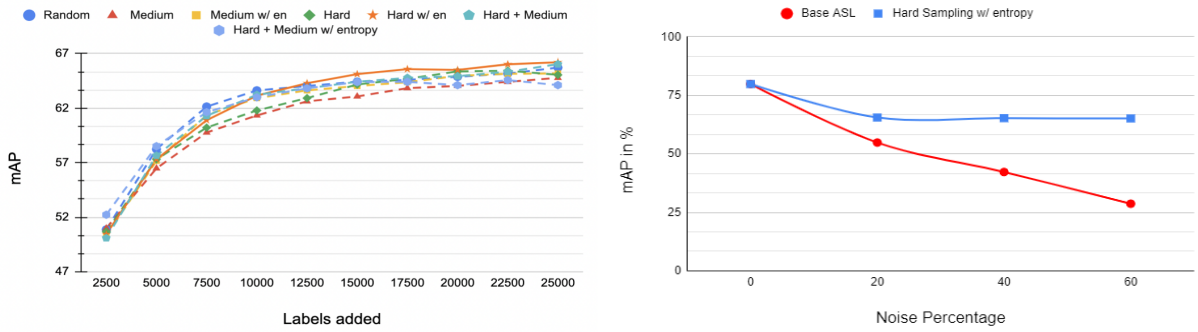
(a) (b) (c)

Figure 4: The evaluation of the measures on 40% noise (a) The mAP per 50 added samples on 40% label noise for all informativeness measures (b) The performances of the measures on 40% noise with labeling cost of approximately 8000 (c) The increase per label of all measures on 40% noise



(a) (b) (c)

Figure 5: The evaluation of the measures on 60% noise (a) The mAP per 50 added samples on 60% label noise for all informativeness measures (b) The performances of the measures on 60% noise with labeling cost of approximately 21000 (c) The increase per label of all measures on 60% noise



(a) (b)

Figure 6: (a) Evaluation of the measures based on their performance after various amounts of labels were cleaned, evaluated with 40% noise. (b) A comparison of the proposed sampling strategy Hard sampling with entropy against the current state-of-the-art, Asymmetric Loss on various degrees of noise.

40% noise

The same applies for 40% noise, as can be seen in **Figure 4**. Random sampling still has the highest increase per label, but Hard Sampling with entropy has very similar increase but, higher overall mAP and better mAP at a fixed cost

60% noise

In **Figure 5** the results vary slightly, Hard sampling has the highest mAP after all samples are added. While the % increase per label is highest for Medium Sampling and Hard + Medium sampling has the best mAP at a fixed cost. Even tough Hard Sampling with entropy does not surpass every other measures, it still performs quite well.
It is worth noting that the Random Sampling has slowly been performing worse across all measures as the % noise increases.

4.4 Discussion

As is visible on the graphs in the previous section, no single measure clearly outperforms the others. Although there is no clear best, we do however want to highlight the measure that combines Hard Sampling with entropy. This measure consistently performs well across all degrees of noise not just when it comes to mAP but regarding cost as well. This is especially visible in **Figure 6a**, were the measures are compared after various amount of labels were cleaned and added. In **Figure 6b**, it is clear by how much the proposed method helps improve the performance of the Asymmetric Loss on a dataset with missing labels.

Despite the fact that the best overall performance will be achieved using Hard Sampling with entropy, there might be scenarios where one might want to have the lowest possible relabelling cost. In which case Random Sampling might be better suited for a dataset with noise percentage lower than 50%, above that Medium sampling has the lower cost. Conversely if one wants to obtain the highest possible mAP, a measure using Hard Sampling might be preferred. Important to note, when a few samples are added there is still a notable difference in mAP. However once enough samples are added all measures will eventually converge to the same value, despite there being a difference in cost.

5 Ethics and Reproducibility

In this section, we elaborate any ethical aspects this paper might have as well as discuss the reproducibility of the experiments performed by this paper. Due to the lack of in-

teraction between a human and a computer any or collection of data, any privacy issues were deemed not relevant. This section will therefore only the discuss the reproducibility of the experiments.

Reproducing the method proposed by this paper should be quite straightforward. The algorithm itself is clearly explained, and in addition pseudo-code is given for it. Despite the fact that no pseudo-code was given for the measures, they are clearly formulated in addition to being simple to implement. The dataset used is explicitly mentioned and publicly available. Although the manner in which noise is injected is also clearly explained, the exact noise might vary, as the noise was injected randomly.
Finally, any choice of hyper-parameters was clearly mentioned and described. The exact results might not able to be achieved, this however natural, since the mAP might vary from run to run.

6 Conclusion and Future work

This paper proposed a novel way to perform Active Learning for missing labels in a multi-label classification setting, i.e., the current state-of-the-art was combined with informativeness measures that exploit the ratio of the missing labels while taking the uncertainty of the model into account. By combining the Asymmetric Loss model with Active Learning and using Hard Sampling with entropy to select samples, we can increase the mAP of ASL by 10%, 20% and 30% for 20%, 40% and 60% respectively, without incurring a high relabelling cost. Our experimental results in section 4.3 show that this measure also outperforms another commonly used sampling strategy as well as improve ASL.

In the future some additional issues can be explored:

- To reduce the relabeling costs even further, methods to propagate the sample to an automatic labeller should be investigated. This way, the oracle is only queried if the algorithm cannot determine the label on its own. One such a labeling methods is described in [6].
- Since this paper only handled a subset of the entire MSCOCO dataset due to hardware restrictions, it should be evaluated on the entire dataset as well as other datasets.
- Alternatively, additional sampling measures could be explored, with higher performance than the methods described by this paper.

References

[1] M. Zhang and Z. Zhou, "A Review on Multi-Label Learning Algorithms," in IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 8, pp. 1819-1837, Aug. 2014, doi: 10.1109/TKDE.2013.39.

[2] Hsu, W. N., Lin, H. T. (2015, February). Active learning by learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 29, No. 1).

[3] T. Durand, N.Mehrasa, and G. Mori. Learning a deep convnet for multi-label classification with partial labels. In CVPR, pages 647–657. IEEE, 2019.

[4] K. Ibrahim, E. Epure, G.Peeters, G. Richard. Confidence-based Weighted Loss for Multi-label Classification with Missing Labels. The 2020 International Conference on Multime- diaRetrieval (ICMR-20), Jun 2020, Dublin, Ireland. ff10.1145/3372278.3390728ff.

[5] Huang, S. J., Chen, S., Zhou, Z. H. (2015, July). Multi-label active learning: query type matters. In IjCAI (pp. 946-952).

[6] J. Wu et al., "Weak-Labeled Active Learning With Conditional Label Dependence for Multilabel Image Classification," in IEEE Transactions on Multimedia, vol. 19, no. 6, pp. 1156-1169, June 2017, doi: 10.1109/TMM.2017.2652065.

[7] E. Ben-Baruch, T. Ridnik, N. Zamir, A. Noy, I. Friedman, M. Protter, L. Zelnik-Manor. (2020). Asymmetric Loss For Multi-Label Classification.

[8] M. Everingham, et al. (2015, p.98-136)The pascal visual object classes challenge: A retrospective . Int. J. Comput. Vis., 111(1).

[9] Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ra- manan, D., Zitnick, C. L. Dollar, P. (2014). Microsoft COCO: Common Objects in Context (cite arxiv:1405.0312 Comment: 1) updated annotation pipeline description and figures; 2) added new section describing datasets splits; 3) updated author list)

[10] Jain, V., Modhe, N., Rai, P. (2017, July). Scalable generative models for multi-label learning with missing labels. In International Conference on Machine Learning (pp. 1636-1644). PMLR.

[11] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS- WIDE: a real-world web image database from national university of singapore. In Stephane Marchand-Maillet and Yiannis Kompatsiaris, editors, CIVR. ACM, 2009.

[12] Zha, Z. J., Wang, M., Zheng, Y. T., Yang, Y., Hong, R., Chua, T. S. (2011). Interactive video indexing with statistical active learning. IEEE Transactions on Multimedia, 14(1), 17-27.

[13] Yu, Y., Pedrycz, W., Miao, D. (2014). Multi-label classification by exploiting label correlations. Expert Systems with Applications, 41(6), 2989-3004.

[14] X. Li, L. Wang, E. Sung. (2004). Multi-label SVM active learning for image classification. International Conference on Image Processing (ICIP), 2004, Vol. 4, pp. 2207-2210.

[15] J. Tang, Z. J. Zha, D. Tao, T. S. Chua. (2012). Semantic-gap-oriented active learning for multi-label image annotation. IEEE Transactions on Image Processing, 2012, 21(4), 2354-2360.

[16] X.Li,Y.Guo.(2013). Active learning with multi-label svm classification. AAAI International Joint Conference on Artificial Intelligence (IJCAI), 2013, pp.1479-1485.

[17] Dembczynski, K., Waegeman, W., Cheng, W., Hüllermeier, E. (2010, June). On label dependence in multi-label classification. In Workshop proceedings of learning from multi-label data (pp. 5-12).

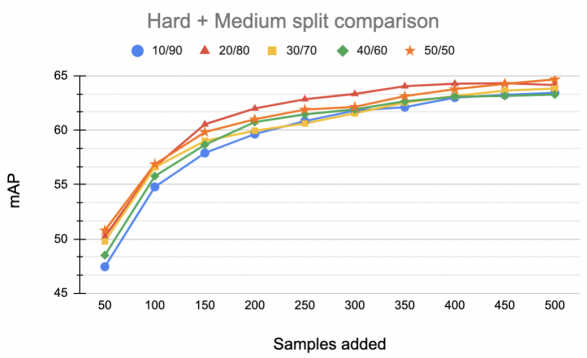
[18] T. Ridnik, H. Lawen, A. Noy, I. Friedman. (2020). TResNet: High Performance GPU-Dedicated Architecture.

[19] Deng, J. et al., 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition.

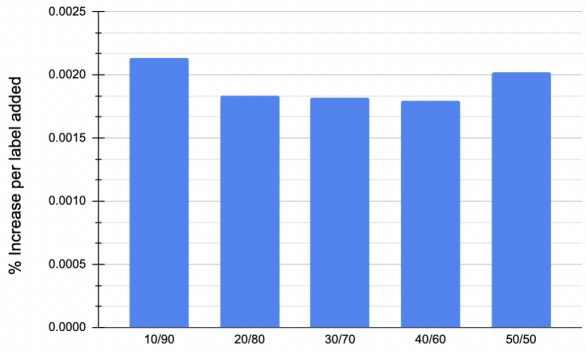
[20] Xing, Y., Yu, G., Domeniconi, C., Wang, J., Zhang, Z. (2018, July). Multi-label co-training. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (pp. 2882-2888).

A Hard + Medium sampling

In this section we evaluate various splits for Hard + Medium sampling to determine which split is optimal. We compare the splits by their final mAPs as well as increase per label. Even tough there is no significant difference in the mAP of the different splits (see **Figure 7a**), in **Figure 7b** it is visible that the in 10/90 splits the increase in mAP highest making it the preferred split.



(a)



(b)

Figure 7: The evaluation of various splits for Hard + Medium Sampling (a) A the mAP for various splits on 40% noise (b) The increase per label for the various splits on 40% noise