Loss framework for uni-instance multiclass multilabel prediction with varying number of labels

Anonymous Author(s)
Submission Id: xx

ABSTRACT

Multilabel classification is a common task in text, image or video (scene) prediction.

KEYWORDS

Keyword; Keyword; Keyword

ACM Reference Format:
Anonymous Author(s). 2021. Loss framework for uni-instance multiclass multilabel prediction with varying number of labels. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11-15, 2021, Montréal, Canada.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As neural network models are able to learn increasingly abstracter representations via deeper networks, representation learning and self-supervision, it might be reasonable to expect that, thanks to their conferred broader understanding of the world, they get better at predicting more abstract labels. Beyond objects types, face recognition, expressions, neural networks might be able to predict genres/categories TODO : other things as well? of text, image and sound. While researchers are working hard at building neural networks with very high level of understanding in the embedding space, there seems to be few research on developing loss functions that are adapted for these higher level concepts in the output space.

Although multilabel binary prediction (commonly referring to mutually inclusive labels) is a task thoroughly covered in existing literature, there does not seem to exist a framework that deals with different amounts of positive labels in the groundtruth. For example, a scientific journal can be tagged as *machine learning* and *economics*, or a movie can be tagged as *romance* and *comedy*. These instances might as well be assigned only one tag in the groundtruth, or many more within the possible tags (classes).

Before, exploring the subject further, we will use Figure 5 to disambiguate the terminology used in this research. There seems to exist a consensus over the terms multiclass and multilabel learning, meaning respectively mutually exclusive and mutually inclusive labels TODO : source. Multilabel can therefore be seen as a sub domain of multiclass learning, where more than one class can be true for the same example. Within multilabel training, we introduce the distinction between multi-instance multilabel (e.g. [26]) and

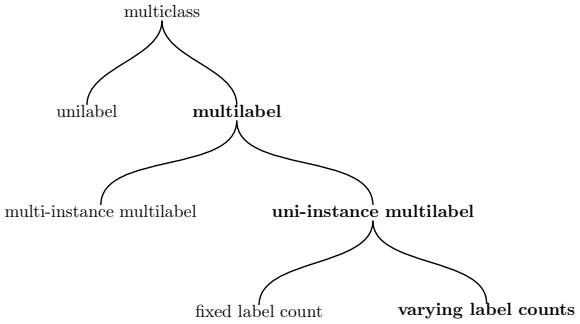

Figure 1: "multiclass" nomenclature

uni-instance multilabel. The former refers to tasks where elements within each example can be singled-out (objects in an image or expressions in a text) and assigned one or more labels. In the contrary, this paper focuses on uni-instance multilabel training (sparse occurrences of the term holistic can be found in the literature to describe this phenomenon for image [11, 15] and a recent video dataset [7] TODO : read these), more specifically with varying label counts. To the best of our knowledge, there are few existing representatives of that type of labeling task in the literature. TODO : cite more milestone examples for each category. TODO : delta with hierarchical label learning

The particularity of tasks like scientific paper tagging or movie genre classification is that it remains unclear what elements in an image/video or text can be singled out as predictive of a particular tag/genre. Rather, a complex interaction between these elements in the feature space steer the predictions. For example, the sole mention of the term "machine learning" in a paper should not be a sufficient condition to tag it as such. Instead, one could expect from the publisher to get acquainted with the paper enough to determine whether the research is a worthwhile contribution or application of *machine learning* to deserve the tag. This involves thorough understanding of the proposed method and background knowledge on state-of-the-art methods. An analogous argument can be made for movie genre classification for movie posters.

However, if elements in an image/text can be singled out as predictive of a single tag, the problem reverts back to predicting with the a priori knowledge of the existence of only one true label (i.e. multi-instance multilabel learning). The reason for distanciating singling-out from uni-instance labels, is that it has been shown that as soon as singling-out is possible, models that work on instances are more accurate TODO : rewrite this paragraph and sources. The singled-out elements can be subsets of the original feature space (typically in object detection like with the COCO dataset [19] or

SIGIR '21, July 11-15, 2021, Montréal, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

the Amazon Rainforest Dataset[1] TODO : others). Similarly, recent research has shown that the singled-out elements can be located in the abstract representations (embeddings) of the feature set and might individually predict a single true label (like GPT-3 TODO : source) TODO : more examples. This might also carry prospects of generalizability of the model [30] TODO : elaborate.

But for now, in certain retrieval tasks such as scientific journal tagging, the effect of sub-entities (either expressions in the text or single features in the embedding space) on the prediction of each label remains hard to assess. Instead we propose uni-instance (sometimes referred to as holistic) multilabel learning for varying amount of labels, with a focus on custom loss functions.

To allow the use of existing diffentiable loss functions, previous research papers tend to re-frame the problem into either (I) a multi-instance uni-label (as described above, with the COCO dataset as an example of isolation of features [19]), (II) uni-instance unilabel prediction (III) uni-instance multilabel prediction with fixed label count (IV) uni-instance multilabel prediction with varying label count with post-training thresholding (V) redefine backpropagation for multilabel prediction [31] (VI) multitask learning [8]. This order reflects in ascending order how close modelling seem to fit the original task, which remains uni-instance multilabel learning with varying amounts of labels. TODO : group them

Sometimes it seems useful to optimize a neural network directly on the evaluation metric [9]. In the case of uni-instance multilabel classification with varying amount of labels, common loss functions such as cross-entropy loss or multinomial logit loss deliver predictions on the unit interval. In the case of mutually exclusive labels, this is a viable solution. Sometimes the groundtruth consists of a collection of positive and negative valued classes, in varying amounts across observations. In other terms, one is looking for a top-k prediction with varying k accross observations.

For the problem definitions (III) and (IV)

In a number of retrieval tasks, a model's out of sample accuracy is measured on metrics such as AUROC, F1 score, etc. These reflect an objective catered towards evaluating the model over an entire ranking. Due to to lack of differentiability, these metrics cannot be directly used as loss functions at training time (in-sample). A seminal study [9] derived a general framework for doing so.

our contribution

We propose a general mathematical formulation of uni-instance multilabel learning for varying amount of groundtruth labels. The generalization encompasses different levels of complexity, from the classical cross-entropy loss up to the proposed loss function. *sigmoidF1* is a F1 score surrogate which allows to optimize for label prediction and count simultaneously in a single task and is robust to outliers. It delivers more precise predictions than the current state-of-the-art on several different metrics, accross text and image related tasks.

2 BUILDING UP ON LOSSES

Multi-label learning can be divided into two major fields: *problem transformation* and *algorithm adaptation* [33]. In the former case, multilabel classification is re-framed as a binary, multiclass classification or label ranking problem. In the latter, one tries to adapt multiclass algorithms to the problem. The current endeavor focusses on *algorithm adaptation*.

For the purpose of *problem transformation*, we define $\mathcal{L}_{\text{multiclass}}$, a class of loss functions that minimize predictions in relative terms. Binary cross-entropy, logit and their variants such as focal loss or hinge loss (deemed unstable [18]) are common choices when it comes to multiclass prediction. Cross-entropy loss can be formulated as $\mathcal{L}_{\text{CE}} = -\sum \log(p_i)$. Note that minimizing binary cross-entropy is equivalent to maximizing for log-likelihood [2, Section 4.3.4]. More generally, the *problem transformation* formulation amounts to minimizing the loss on a class of neural networks, such that

$$\min_{\mathcal{L}_{\text{multiclass}}} \mathcal{F}(\cdot;\Theta;\mathcal{L}_{\text{multiclass}}(\mathbf{y},\hat{\mathbf{y}})), \tag{1}$$

In the context of *algorithm adaptation*, where the number of positive labels in the groundtruth is unknown a priori, we aim to both obtain a propensity of each label being true and a prediction of the number of true labels:

$$\min_{\mathcal{L}_{\text{multiclass}},\mathcal{L}_{\text{count}}} \mathcal{F}(\cdot;\Theta;\mathcal{L}_{\text{multiclass}}(\mathbf{y},\hat{\mathbf{y}}) + \lambda\mathcal{L}_{\text{count}}(\mathbf{n},\hat{\mathbf{n}})), \tag{2}$$

where $n_i = \sum_j \mathbb{1}_{y_i^j=1}$ is the count of positive labels per example. We thus impose a constraint for the retrieval of label counts. For example, a cross-entropy loss surrogate would penalize for the number of wrongly predicted labels $\mathcal{L}_{\text{CE+N}} = \mathcal{L}_{\text{CE}} + \lambda(\sum tp/\sum p)$, with $tp = \sum_{i\in Y^+} \mathbb{1}_{\mathbf{p_i}\geq b}$ and b a threshold to be defined. TODO : tencent loss.

This formulation is most straightforward but suffers from higher parameterization and the lack of modelling of the interactions between label counts and label prediction. To mitigate these issues, we propose a unified loss formulation, namely

$$\min_{\mathcal{L}_{\text{multitag}}} \mathcal{F}\left(\cdot;\Theta;\mathcal{L}_{\text{multitag}}(\mathbf{y},\hat{\mathbf{y}},\mathbf{n},\hat{\mathbf{n}})\right), \tag{3}$$

Although predictions and counts explicitly appear in that formulation, $\mathcal{L}_{\text{multitag}}$ can optimize for both metrics implicitely (see proposed *sigmoidF1* below).

TODO : look at YOU ONLY TRAIN ONCE: LOSS-CONDITIONAL TRAINING OF DEEP NETWORKS

TODO : cite stat learning [13, p. 308-310]

3 RELATED WORK

TODO : look at [[https://www.sciencedirect.com/topics/computer-science/extractive-summarization][extractive summarization]]

This section will be guided by the previous section's formulation of the multitags problem, we will therefore focus on *algorithm adaptation*, *metrics as losses* and *dynamic thresholding*.

3.1 algorithm adaptation

[1] Available at https://www.kaggle.com/c/planet-understanding-the-amazon-from-space

Loss framework for uni-instance multiclass multilabel prediction with varying number of labels

Anonymous Author(s)
Submission Id: xx

ABSTRACT

Multilabel classification is a common task in text, image or video (scene) prediction.

KEYWORDS

Keyword; Keyword; Keyword

ACM Reference Format:
Anonymous Author(s). 2021. Loss framework for uni-instance multiclass multilabel prediction with varying number of labels. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11-15, 2021, Montréal, Canada.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

4 INTRODUCTION

As neural network models are able to learn increasingly abstracter representations via deeper networks, representation learning and self-supervision, it might be reasonable to expect that, thanks to their conferred broader understanding of the world, they get better at predicting more abstract labels. Beyond objects types, face recognition, expressions, neural networks might be able to predict genres/categories TODO : other things as well? of text, image and sound. While researchers are working hard at building neural networks with very high level of understanding in the embedding space, there seems to be few research on developing loss functions that are adapted for these higher level concepts in the output space.

Although multilabel binary prediction (commonly referring to mutually inclusive labels) is a task thoroughly covered in existing litterature, there does not seem to exist a framework that deals with different amounts of positive labels in the groundtruth. For example, a scientific journal can be tagged as *machine learning* and *economics*, or a movie can be tagged as *romance* and *comedy*. These instances might as well be assigned only one tag in the groundtruth, or many more within the possible tags (classes).

Before, exploring the subject further, we will use Figure 5 to disambiguate the terminology used in this research. There seems to exist a concensus over the terms multiclass and multilabel learning, meaning respectively mutually exclusive and mutually inclusive labels TODO : source. Multilabel can therefore be seen as a subdomain of multiclass learning, where more than one class can be true for the same example. Within multilabel training, we introduce the distinction between multi-instance multilabel (e.g. [26]) and

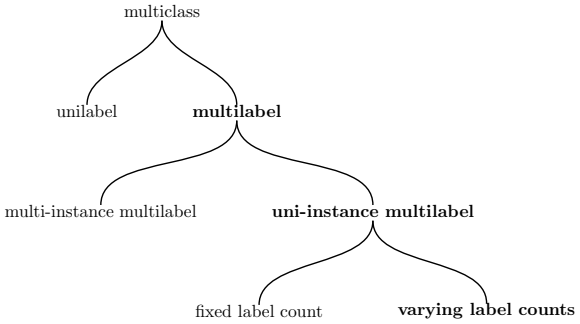

Figure 2: "multiclass" nomenclature

uni-instance multilabel. The former refers to tasks where elements within each example can be singled-out (objects in an image or expressions in a text) and assigned one or more labels. In the countrary, this paper focusses on uni-instance multilabel training (sparse occurences of the term holistic can be found in the litterature to describe this phenomenon for image [11, 15] and a recent video dataset [7] TODO : read these), more specifically with varying label counts. To the best of our knowledge, there are few existing representatives of that type of labelling task in the literature. TODO : cite more milestone examples for each category. TODO : delta with hierarchical label learning

The particularity of tasks like scientific paper tagging or movie genre classification is that it remains unclear what elements in an image/video or text can be singled out as predictive of a particular tag/genre. Rather, a complex interaction between these elements in the feature space steer the predictions. For example, the sole mention of the term "machine learning" in a paper should not be a sufficient condition to tag it as such. Instead, one could expect from the publisher to get acquainted with the paper enough to determine wether the research is a worthwhile contribution or application of *machine learning* to deserve the tag. This involves thorough understanding of the proposed method and background knowledge on state-of-the-art methods. An analogous argument can be made for movie genre classification for movie posters.

However, if elements in an image/text can be singled out as predictive of a single tag, the problem reverts back to predicting with the a priori knowledge of the existence of only one true label (i.e. multi-instance multilabel learning). The reason for distanciating singling-out from uni-instance labels, is that it has been shown that as soon as singling-out is possible, models that work on instances are more accurate TODO : rewrite this paragraph and sources. The singled-out elements can be subsets of the original feature space (typically in object detection like with the COCO dataset [19] or

SIGIR '21, July 11-15, 2021, Montréal, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

the Amazon Rainforest Dataset[1] TODO : others). Similarly, recent research has shown that the singled-out elements can be located in the abstract representations (embeddings) of the feature set and might individually predict a single true label (like GPT-3 TODO : source) TODO : more examples. This might also carry prospects of generalizability of the model [30] TODO : elaborate.

But for now, in certain retrieval tasks such as scientific journal tagging, the effect of sub-entities (either expressions in the text or single features in the embedding space) on the prediction of each label remains hard to assess. Instead we propose uni-instance (sometimes referred to as holistic) multilabel learning for varying amount of labels, with a focus on custom loss functions.

To allow the use of existing diffentiable loss fonctions, previous research papers tend to reframe the problem into either (I) a multi-instance uni-label (as described above, with the COCO dataset as an example of isolation of features [19]), (II) uni-instance unilabel prediction (III) uni-instance multilabel prediction with fixed label count (IV) uni-instance multilabel prediction with varying label count with post-training thresholding (V) redefine backpropagation for multilabel prediction [31] (VI) multitask learning [8]. This order reflects in ascending order how close modelling seem to fit the original task, which remains uni-instance multilabel learning with varying amounts of labels. TODO : group them

Sometimes it seems useful to optimize a neural network directly on the evaluation metric [9]. In the case of uni-instance multilabel classification with varying amount of labels, common loss functions such as cross-entropy loss or multinomial logit loss deliver predictions on the unit interval. In the case of mutually exclusive labels, this is a viable solution. Sometimes the groundtruth consists of a collection of positive and negative valued classes, in varying amounts across observations. In other terms, one is looking for a top-k prediction with varying k accross observations.

For the problem definitions (III) and (IV)

In a number of retrieval tasks, a model's out of sample accuracy is measured on metrics such as AUROC, F1 score, etc. These reflect an objective catered towards evaluating the model over an entire ranking. Due to to lack of differentiability, these metrics cannot be directly used as loss functions at training time (in-sample). A seminal study [9] derived a general framework for doing so.

our contribution

We propose a general mathematical formulation of uni-instance multilabel learning for varying amount of groundtruth labels. The generalization encompasses different levels of complexity, from the classical cross-entropy loss up to the proposed loss function. *sigmoidF1* is a F1 score surrogate which allows to optimize for label prediction and count simultanuously in a single task and is robust to outliers. It delivers more precise predictions than the current state-of-the-art on several different metrics, accross text and image related tasks.

5 BUILDING UP ON LOSSES

Multi-label learning can be divided into two major fields: *problem transformation* and *algorithm adaptation* [33]. In the former case, multilabel classification is reframed as a binary, multiclass classification or label ranking problem. In the latter, one tries to adapt multiclass algorithms to the problem. The current endeavour focusses on *algorithm adaptation*.

For the purpose of *problem transformation*, we define $\mathcal{L}_{\text{multiclass}}$, a class of loss functions that minimize predictions in relative terms. Binary cross-entropy, logit and their variants such as focal loss or hinge loss (deemed unstable [18]) are common choices when it comes to multiclass prediction. Cross-entropy loss can be formulated as $\mathcal{L}_{\text{CE}} = -\sum \log(p_i)$. Note that minimizing binary cross-entropy is equivalent to maximizing for log-likelihood [2, Section 4.3.4]. More generally, the *problem transformation* formulation amounts to minimizing the loss on a class of neural networks, such that

$$\min_{\mathcal{L}_{\text{multiclass}}} \mathcal{F}\left(\cdot;\Theta;\mathcal{L}_{\text{multiclass}}(\mathbf{y},\hat{\mathbf{y}})\right), \tag{4}$$

In the context of *algorithm adaptation*, where the number of positive labels in the groundtruth is unknown a priori, we aim to both obtain a propensity of each label being true and a prediction of the number of true labels:

$$\min_{\mathcal{L}_{\text{multiclass}},\mathcal{L}_{\text{count}}} \mathcal{F}\left(\cdot;\Theta;\mathcal{L}_{\text{multiclass}}(\mathbf{y},\hat{\mathbf{y}}) + \lambda\mathcal{L}_{\text{count}}(\mathbf{n},\hat{\mathbf{n}})\right), \tag{5}$$

where $n_i = \sum_j \mathbb{1}_{y_i^j=1}$ is the count of positive labels per example. We thus impose a constraint for the retrieval of label counts. For example, a cross-entropy loss surrogate would penalize for the number of wrongly predicted labels $\mathcal{L}_{\text{CE+N}} = \mathcal{L}_{\text{CE}} + \lambda(\sum tp/\sum p)$, with $tp = \sum_{i\in Y^+} \mathbb{1}_{\mathbf{p_i}\geq b}$ and b a threshold to be defined. TODO : tencent loss.

This formulation is most straightfoward but suffers from higher parametrization and the lack of modelling of the interactions between label counts and label prediction. To mitigate these issues, we propose a unified loss formulation, namely

$$\min_{\mathcal{L}_{\text{multitag}}} \mathcal{F}\left(\cdot;\Theta;\mathcal{L}_{\text{multitag}}(\mathbf{y},\hat{\mathbf{y}},\mathbf{n},\hat{\mathbf{n}})\right), \tag{6}$$

Although predictions and counts explicitely appear in that formulation, $\mathcal{L}_{\text{multitag}}$ can optimize for both metrics implicitely (see proposed *sigmoidF1* below).

TODO : look at YOU ONLY TRAIN ONCE: LOSS-CONDITIONAL TRAINING OF DEEP NETWORKS

TODO : cite stat learning [13, p. 308-310]

6 RELATED WORK

TODO : look at [[https://www.sciencedirect.com/topics/computer-science/extractive-summarization][extractive summarization]]

This section will be guided by the previous section's formulation of the multitags problem, we will therefore focus on *algorithm adaptation*, *metrics as losses* and *dynamic thresholding*.

6.1 algorithm adaptation

Early representatives of *algorithm adaptation* stem from heterogenous domains of machine learning. Multi-Label k-Nearest Neighbors [32], Multi-Label Decision Tree [6], Ranking Support Vector Machine [10] and Backpropagation for Multi-Label Learning [31]. More recently, two papers introduced the idea of multitask learning

[1] Available at https://www.kaggle.com/c/planet-understanding-the-amazon-from-space

for *label prediction* and *label count prediction* for text (ML_{NET}) [8] and image [17] data. The latter research is loosely catered towards object detection (although not formally presented as such) and is thus out-of-scope: elements in a picture are predicted that tend to be unilabel as defined by the groundtruth (e.g. cat, flower, vase, person, bottle etc.).

6.2 metrics as losses

Often, machine learning post-training evaluation metrics (e.g. AUROC, F1) are not differentiable. There are motivations TODO : which motivations for optimizing a model directly on a metric at training time. A general framework for AUC, AUROC and F1 is presented in [9], but the proposed F1 surrogate remains short of being explicitly derived for stochastic gradient descent. TODO : check again with the authors if I can't get inspired from their work. Recently, a similar work has been proposed to train a Convolutional Neural Network (CNN) from scratch with a few millions of images and hundreds of labels specifically for multilabel tasks [28]. This task is loosely related to object detection, similarly to [17] mentioned in the previous paragraph.

6.3 dynamic thresholding

dynamic thresholding accross classes or examples is an issue as soon as the number of labels to predict is unknown. Certain variants of cross-entropy loss accomodate imbalanced label data [18], but remain agnostic towards the number of labels to predict. Solutions have been tailored to that end, starting with determining an ideal global *threshold* depending on use-cases [20], or per-class-thresholding after training [5] and eventually abstracting the threshold away via a *soft-F1* measure [3] TODO : say more about this method. In the latter two cases, the task is to predict genre from movie posters.

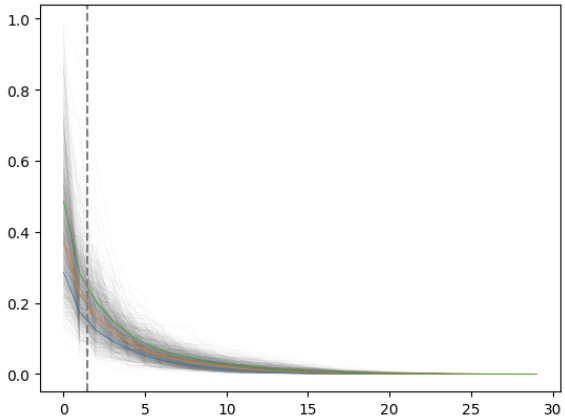

Figure 3: ordered per-label cross-entropy predictions for each example (each grey line) with the median (orange) and IQR (green & blue) over all examples. Determining a global threshold can be related to visually finding the "knee" in that median curve (dotted line)

TODO : nicer plot on another dataset (this is from RTL)

The proposed method is positioned in the lineage of *algorithm adaptation*, using *metric as losses* and allowing for *dynamic thresholdig*.

7 SIGMOID F1 LOSS

For a class of multilayer perceptron $\mathcal{F}(\cdot;\Theta) : \mathcal{X} \rightarrow \mathcal{Y}$, we consider a special case, where $\mathbf{x} = \{x_1, ..., x_n\}$. Each observation is attributed one or more classes out of a label set $\mathbf{l} = \{l_1, ..., l_c\}$. Labels y_i^j are available for each observation i and class j.

For each observation i, label class probabilities can be defined based on predictions as

TODO : check this formula

$$\mathbf{p}_i = \begin{cases} \hat{\mathbf{y}} & \text{if } y = 1 \\ 1 - \hat{\mathbf{y}} & \text{otherwise} \end{cases} \tag{7}$$

Let tp and fp be number of true and false positives respectively. It is necessary to define a bound b, at which a prediction is dichotomized:

$$tp = \sum_{i \in Y^+} \mathbb{1}_{\mathbf{p}_i \geq b} \quad fp = \sum_{i \in Y^-} \mathbb{1}_{\mathbf{p}_i \geq b} \quad fn = \sum_{i \in Y^+} \mathbb{1}_{\mathbf{p}_i < b} \tag{8}$$

$\mathbb{1}_{\mathbf{p}_i \geq b}$, $\mathbb{1}_{\mathbf{p}_i < b}$ are thus the count of positive and negative predictions at threshold b,

We also define precision and recall

$$\begin{aligned} P &= \frac{tp}{tp + fp} \\ R &= \frac{tp}{tp + fn} = \frac{tp}{|Y^+|} \end{aligned} \tag{9}$$

We can then define F_β, which can be expressed as the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall than precision [25].

DOUBT : maybe ignore F_β and only mention F_1

$$F_\beta = \left(1 + \beta^2\right) \frac{P \cdot R}{\beta^2 P + R} \tag{10}$$

Or equivalently:

$$\begin{aligned} F_\beta &= \left(1 + \beta^2\right) \frac{tp}{\left(1 + \beta^2\right) tp + \beta^2 fn + fp} \\ &= \left(1 + \beta^2\right) \frac{tp}{\beta^2 |Y + | + tp + fp} \end{aligned} \tag{11}$$

Given the presence of the step indicator function $\sum \mathbb{1}_{\mathbf{p}_i \geq b}$, F_β is not differentiable for gradient based methods. One way of surpassing that problem is to use a smooth surrogate.

7.1 soft F1 score

It is possible define a *soft F1* score [3] DOUBT : can we cite a Medium post? with smooth confusion matrix entries (i.e. tp, fp and fn are not natural numbers anymore):

$$\overline{tp} = \sum \hat{\mathbf{y}} \odot \mathbf{y} \quad \overline{fp} = \sum \hat{\mathbf{y}} \odot (\mathbf{1} - \mathbf{y}) \quad \overline{fn} = \sum (\mathbf{1} - \hat{\mathbf{y}}) \odot \mathbf{y}$$

$$\mathcal{L}_{\text{softF1}} = \frac{\overline{tp}}{2\overline{tp} + \overline{fn} + \overline{fp}} \tag{12}$$

tp, fp and fn are now replaced by rough surrogates, this method has the advantage of

7.2 sigmoidF1 score

We define *sigmoidF1*, inspired by the *Maximum F1-score criterion* for automatic mispronounciation detection [14]. Whereas a sigmoid function $S(u)$

$$S(u;\beta,\eta) = \frac{1}{1+\exp(-\beta(u+\eta))}, \tag{13}$$

with β and η tunable parameters for slope and offset respectively. Higher β results in steeper slope at the center of the sigmoid and thus more stringent thresholding. At its extreme, $lim_{\beta\to\infty}S(u;\beta,\eta)$ corresponds to the step function used in Equation 29. with $S(u)$, the confusion matrix entries then become

$$\widetilde{tp} = \sum S(\hat{\mathbf{y}})\odot\mathbf{y} \quad \widetilde{fp} = \sum S(\hat{\mathbf{y}})\odot(\mathbf{1}-\mathbf{y}) \quad \widetilde{fn} = \sum(\mathbf{1}-S(\hat{\mathbf{y}}))\odot\mathbf{y}$$

And thus

$$\mathcal{L}_{\text{softF1}} = \frac{\widetilde{tp}}{2\widetilde{tp}+\widetilde{fn}+\widetilde{fp}} \tag{14}$$

DOUBT : mention smooth hinge loss [24]

For *sigmoidF1* β and η are tuned globally as hyperparameters. *SAdF1* (Sigmoid Adaptive F1), is an alternative where β is first set to a relatively low value and increased after each epoch. This way, a loose threshold first allows Stochastic Gradient Descent (SGD) to broadly scan the parameter space accross several local minima, before narrowing parameter search down to a promissing region (similarily to adaptive learning rates).

SBayesF1 (sigmoid Bayes F1) replaces point estimates for β and η with posterior distribution estimates.

$$S(u_i) = \frac{1}{1+\exp(-\beta_i(u_i+\eta_i))} \tag{15}$$

β_i and η_i are estimated with MCMC at training time of the neural network. They are therefore implicitely allowed to vary across examples.

TODO : try SadF1 and SBayesF1 in practice

7.3 Robustness

Similarly to the focal loss [18], sigmoidF1 loss deals with class imbalance, robustness to outliers.

TODO : statistical robustness assessment

7.4 Evaluation Metrics

The metrics described below are a result of a survey of different common practices for measuring accuracy of multilabel prediction. When true positives and false positives are used, recall that $tp = \sum_{i\in Y^+}\mathbb{1}_{\mathbf{p_i}\geq b}$ and $fp = \sum_{i\in Y^-}\mathbb{1}_{\mathbf{p_i}\geq b}$, and thus a threshold b must be set. When $b = 0.5$, as is commonly done [SOURCE HERE], a risk remains that a lot of examples remain without predictions.

Extending F_1 to multi-class binary classification amounts to deciding wether to un/pool classes. In a first pooled iteration, micro F_1 [SOURCE HERE] equates to creating a single 2x2 confusion matrix for all classes:

$$F_1^{micro} = \frac{\sum tp_c}{2\sum tp_c + \sum fn_c + \sum fp_c} \quad for \quad c\in C$$

Macro F_1 [20] amounts to creating one confusion matrix per class:

$$F_1^{macro} = \frac{1}{c}\sum_{j=1}^{c} F_1$$

DOUBT : Do we need to justify optimizing for an F1 surrogate at training time and to then use F1 itself as a metric?

Weighted macro F_1 TODO : find source is similar but includes weighing to account for class imbalance, i.e. weighing each class by the number of groundtruth positives.

$$F_1^{weighted} = \frac{1}{c}\sum_{j=1}^{c} n_j F_1 \quad where \quad n_j = \sum_i \mathbb{1}_{y_i^j=1}$$

Accuracy is the overall fraction of correctly predicted labels [20]:

$$Acc = \frac{tp+tn}{tp+tn+fp+fn}$$

TODO compare to [29]

8 DATASETS

sigmoidF1 is tested across different modalities, namely image, video, sound and text, with a focus on text: the most comparable research was on text data.

Among the three datasets used for benchmarking ML-NET [8], a cancer hallmark dataset is of multi-instance multilabel nature [12] [2]: the research clearly describe a process of annotating several expressions within paper abstracts. The remaining two datasets for chemical exposure [16] [3] and diagnosis codes assigment [23] [4], seem to fit to the entity wide multilabel definition but have a strong hierarchical nature. Although slightly out-of-scope, the three datasets above will be used for benchmarking, since they were used to test ML-NET, which is the state-of-the-art in *algorithm adaptation* for text to the best of our knowledge.

For a broader scope in learning for text data, we also use the newly created *Arxiv dataset* [5] with data on abstracts of 1.7 million open source articles and their categories (suitably mutually inclusive and of varying count per example).

In the vision domain, a dataset of movie posters [6] and their genre is used. Similarly, labels are mutually inclusive and of varying count per example. It is arguable that is hard to single out elements in the image of a poster that define the genre of a movie. Rather it might be a combination of the title font, the background image, the presence of actors and specific objects such as cars, weapons etc.

TODO : I removed all jpg's that are empty in the prescraped data. I could try to scrape the posters myself to see if I get more

Another recently created dataset was made available for *Large Scale Holistic Video Understanding* [7] [7], as defined in the introduction.

TODO : this is an ambitious number of datasets. Add longer description of each dataset, depending on which ones I keep: sample size, number of classes etc. see utils here: https://github.com/ashrefm/multi-label-soft-f1

DOUBT : cite Kaggle datasets formally instead of using links: https://www.kaggle.com/data/46091

DOUBT : add a music genre classification dataset, for which Vincent Koops at RTL could help train

[2] Available at https://www.cl.cam.ac.uk/sim;sb895/HoC.html
[3] Available at https://figshare.com/articles/Corpus$_a nd_S oftware$/4668229
[4] Available at https://physionet.org/works/ICD9CodingofDischargeSummaries
[5] Available at https://www.kaggle.com/Cornell-University/arxiv
[6] Labels available at https://tinyurl.com/y7ydyedu and prescraped images from IMDB at https://tinyurl.com/y7lfpvlx
[7] Available at https://github.com/holistic-video-understanding/HVU-Dataset

9 EXPERIMENTAL RESULTS

varying b in the sigmoid function as if it is an adaptive learning rate TODO : actually try it out.

one b per class

if we consider b and c to be probabilistic, we can then use tensorflow probability to assess their distribution

the batch size has to be relatively large (i.e. 256), in order for meaningful F1 surrogates to be calculated.

movie posters (CNN)

Loss	macroF @ 0.5	microF1 @ 0.5	weightedF1 @ 0.5	Precision @ 0.5	Recall @ 0.5
$\mathcal{L}_{\text{CE}}$	0.057	0.200	0.159	0.106	0.106
$\mathcal{L}_{\text{FL}}$	0.055	0.192	0.154	0.115	0.115
$\mathcal{L}_{\text{CE+N}}$	0	0	0	0	0
$\mathcal{L}_{\text{CE+T}}$	0	0	0	0	0
$\mathcal{L}_{\text{macroSoftF1}}$	0.132	0.323	0.280	0.105	0.105
$\mathcal{L}_{\text{sigmoidF1}}$	**0.117**	**0.240**	**0.263**	**0.103**	**0.103**

Arxiv (distillBERT)

Metric	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{FL}}$	$\mathcal{L}_{\text{CE+N}}$	$\mathcal{L}_{\text{CE+T}}$
$P(\%)$	0	0	0	0
$R(\%)$	0	0	0	0
$F_1(\%)$	0	0	0	**0**

Cancer hallmark (distillBERT)

Metric	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{FL}}$	$\mathcal{L}_{\text{CE+N}}$	$\mathcal{L}_{\text{CE+T}}$
$P(\%)$	0	0	0	0
$R(\%)$	0	0	0	0
$F_1(\%)$	0	0	0	**0**

Chemical exposure (distillBERT)

Metric	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{FL}}$	$\mathcal{L}_{\text{CE+N}}$	$\mathcal{L}_{\text{CE+T}}$
$P(\%)$	0	0	0	0
$R(\%)$	0	0	0	0
$F_1(\%)$	0	0	0	**0**

Chemical exposure (distillBERT)

Metric	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{FL}}$	$\mathcal{L}_{\text{CE+N}}$	$\mathcal{L}_{\text{CE+T}}$
$P(\%)$	0	0	0	0
$R(\%)$	0	0	0	0
$F_1(\%)$	0	0	0	**0**

simulated data

Metric	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{FL}}$	$\mathcal{L}_{\text{CE+N}}$	$\mathcal{L}_{\text{CE+T}}$
$P(\%)$	0	0	0	0
$R(\%)$	0	0	0	0
$F_1(\%)$	0	0	0	**0**

10 CONCLUSION

Shortcomings

it is debatable wether any task is intrinsincly multilabel and wether the image / text cannot be decomposed in parts that are single labelled.

not long training and small models, but aibility to demonstrate the statement anyways.

Results

In this paper we defined a new problem in deep learning for mulitple modalities that harness the current advances in abstract representation of the input space. A general loss framework is proposed to locate that solution within the existing multiclass multilabel losses and a specific loss function is formulated. *sigmoidF1* achieves significantly results for different F1 values on all datasets.

Future work

Apply the loss function to more sophisticated neural network architectures that use F1 score as an evaluation metric such as AC-SUM-GAN [1].

This model can be adapted for hiarchical multilabel classification or active learning (for both see [22]).

Combine the proposed loss functions with representation learning [21, 27] or self-supervised learning, in order to model abstract relationships between the labels.

adapt to *extreme* multilabel prediction [4]

REFERENCES

[1] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras. 2020. AC-SUM-GAN: Connecting Actor-Critic and Generative Adversarial Networks for Unsupervised Video Summarization,. *IEEE Transactions on Circuits and Systems for Video Technology* (2020).

[2] Christopher M. Bishop. 2007. *Pattern recognition and machine learning, 5th Edition.* Springer. https://www.worldcat.org/oclc/71008143

[3] Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S. Dhillon. 2019. The Unknown Benefits of using a Soft-F1 Loss in Classification Systems. *Towards Data Science* (Dec 2019). https://towardsdatascience.com/the-unknown-benefits-of-using-a-soft-f1-loss-in-classification-systems-753902c0105d

[4] Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S. Dhillon. 2020. Taming Pretrained Transformers for Extreme Multi-label Text Classification. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining* (Jul 2020). https://doi.org/10.1145/3394486.3403368

[5] Wei-Ta Chu and Hung-Jui Guo. 2017. Movie Genre Classification based on Poster Images with Deep Neural Networks. *Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes* (Oct 2017). https://doi.org/10.1145/3132515.3132516

[6] Amanda Clare and Ross D. King. 2001. Knowledge Discovery in Multi-label Phenotype Data. *Lecture Notes in Computer Science* (2001), 42–53. https://doi.org/10.1007/3-540-44794-6_4

[7] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jurgen Gall, Rainer Stiefelhagen, and Luc Van Gool. 2019. *Large Scale Holistic Video Understanding.* arXiv:1904.11451v3 [cs.CV]

[8] Jingcheng Du, Qingyu Chen, Yifan Peng, Yang Xiang, Cui Tao, and Zhiyong Lu. 2019. ML-Net: multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association* 26, 11 (Jun 2019), 1279–1285. https://doi.org/10.1093/jamia/ocz085

[9] Elad ET. Eban, Mariano Schain, Alan Mackey, Ariel Gordon, Rif A. Saurous, and Gal Elidan. 2016. *Scalable Learning of Non-Decomposable Objectives.* arXiv:1608.04802v2 [stat.ML]

[10] André Elisseeff and Jason Weston. 2001. A Kernel Method for Multi-Labelled Classification. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic* (Vancouver, British Columbia, Canada) *(NIPS'01)*. MIT Press, Cambridge, MA, USA, 681–687.

[11] Mingchen Gao, Ulas Bagci, Le Lu, Aaron Wu, Mario Buty, Hoo-Chang Shin, Holger Roth, Georgios Z. Papadakis, Adrien Depeursinge, Ronald M. Summers, and et al. 2016. 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 (Jun 2016), 1–6. https://doi.org/10.1080/21681163.2015.1124249

[12] Douglas Hanahan and Robert A. Weinberg. 2011. Hallmarks of Cancer: The Next Generation. *Cell* 144, 5 (Mar 2011), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

[13] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. The Elements of Statistical Learning. *Springer Series in Statistics* (2009). https://doi.org/10.1007/978-0-387-84858-7

[14] Hao Huang, Haihua Xu, Xianhui Wang, and Wushour Silamu. 2015. Maximum F1-Score Discriminative Training Criterion for Automatic Mispronunciation Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 4 (Apr 2015), 787–797. https://doi.org/10.1109/taslp.2015.2409733

[15] Alberto Jaenal, Francisco-Angel Moreno, and Javier Gonzalez-Jimenez. 2019. Experimental study of the suitability of CNN-based holistic descriptors for accurate visual localization. *Proceedings of the 2nd International Conference on Applications of Intelligent Systems - APPIS '19* (2019). https://doi.org/10.1145/3309772.3309800

[16] Kristin Larsson, Ilona Silins, Yufan Guo, Anna Korhonen, Ulla Stenius, and Marika Berglund. 2014. Text mining for improved human exposure assessment. *Toxicology Letters* 229 (Sep 2014), S119. https://doi.org/10.1016/j.toxlet.2014.06.427

[17] Yuncheng Li, Yale Song, and Jiebo Luo. 2017. Improving Pairwise Ranking for Multi-label Image Classification. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Jul 2017). https://doi.org/10.1109/cvpr.2017.199

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal Loss for Dense Object Detection. *2017 IEEE International Conference on Computer Vision (ICCV)* (Oct 2017). https://doi.org/10.1109/iccv.2017.324

[19] 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. *Lecture Notes in Computer Science* (2014), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[20] Zachary C. Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. 2014. Optimal Thresholding of Classifiers to Maximize F1 Measure. *Lecture Notes in Computer Science* (2014), 225–239. https://doi.org/10.1007/978-3-662-44851-9_15

[21] Timo Milbich, Omair Ghori, Ferran Diego, and Björn Ommer. 2020. Unsupervised representation learning by discovering reliable image relations. *Pattern Recognition* 102 (Jun 2020), 107107. https://doi.org/10.1016/j.patcog.2019.107107

[22] Felipe Kenji Nakano, Ricardo Cerri, and Celine Vens. 2020. Active learning for hierarchical multi-label classification. *Data Mining and Knowledge Discovery* 34, 5 (2020), 1496–1530. https://doi.org/10.1007/s10618-020-00704-w

[23] Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2014. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association* 21, 2 (Mar 2014), 231–237. https://doi.org/10.1136/amiajnl-2013-002159

[24] Jason DM Rennie. 2005. Smooth hinge classification. *Proceeding of Massachusetts Institute of Technology* (2005).

[25] C.J. van Rijsbergen. [n.d.]. *Information retrieval* (2nd [rev.] ed. ed.). Butterworths, London [etc.

[26] H. Soleimani and D. J. Miller. 2017. Semisupervised, Multilabel, Multi-Instance Learning for Structured Data. *Neural Computation* 29, 4 (2017), 1053–1102. https://doi.org/10.1162/NECO_a_00939

[27] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, and et al. 2020. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), 1–1. https://doi.org/10.1109/tpami.2020.2983686

[28] Baoyuan Wu, Weidong Chen, Yanbo Fan, Yong Zhang, Jinlong Hou, Jie Liu, and Tong Zhang. 2019. Tencent ML-Images: A Large-Scale Multi-Label Image Database for Visual Representation Learning. *IEEE Access* 7 (2019), 172683–172693. https://doi.org/10.1109/access.2019.2956775

[29] Hichame Yessou, Gencer Sumbul, and Begüm Demir. 2020. *A Comparative Study of Deep Learning Loss Functions for Multi-Label Remote Sensing Image Classification.* arXiv:2009.13935v1 [cs.CV]

[30] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. https://arxiv.org/abs/1611.03530

[31] Min-Ling Zhang and Zhi-Hua Zhou. 2006. Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization. *IEEE Transactions on Knowledge and Data Engineering* 18, 10 (Oct 2006), 1338–1351. https://doi.org/10.1109/tkde.2006.162

[32] Min-Ling Zhang and Zhi-Hua Zhou. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition* 40, 7 (Jul 2007), 2038–2048. https://doi.org/10.1016/j.patcog.2006.12.019

[33] Min-Ling Zhang and Zhi-Hua Zhou. 2014. A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26, 8 (Aug 2014), 1819–1837. https://doi.org/10.1109/tkde.2013.39

Early representatives of *algorithm adaptation* stem from heterogeneous domains of machine learning. Multi-Label k-Nearest Neighbors [32], Multi-Label Decision Tree [6], Ranking Support Vector Machine [10] and Backpropagation for Multi-Label Learning [31]. More recently, two papers introduced the idea of multitask learning for *label prediction* and *label count prediction* for text (ML_{NET}) [8] and image [17] data. The latter research is loosely catered towards object detection (although not formally presented as such) and is thus out-of-scope: elements in a picture are predicted that tend to be unilabel as defined by the groundtruth (e.g. cat, flower, vase, person, bottle etc.).

10.1 metrics as losses

Often, machine learning post-training evaluation metrics (e.g. AUROC, F1) are not differentiable. There are motivations TODO : which motivations for optimizing a model directly on a metric at training time. A general framework for AUC, AUROC and F1 is presented in [9], but the proposed F1 surrogate remains short of being explicitly derived for stochastic gradient descent. TODO : check again with the authors if I can't get inspired from their work. Recently, a similar work has been proposed to train a Convolutional Neural Network (CNN) from scratch with a few millions of images and hundreds of labels specifically for multilabel tasks [28]. This task is loosely related to object detection, similarly to [17] mentioned in the previous paragraph.

10.2 dynamic thresholding

dynamic thresholding across classes or examples is an issue as soon as the number of labels to predict is unknown. Certain variants of cross-entropy loss accommodate imbalanced label data [18], but remain agnostic towards the number of labels to predict. Solutions have been tailored to that end, starting with determining an ideal global *threshold* depending on use-cases [20], or per-class-thresholding after training [5] and eventually abstracting the threshold away via a *soft-F1* measure [3] TODO : say more about this method. In the latter two cases, the task is to predict genre from movie posters.

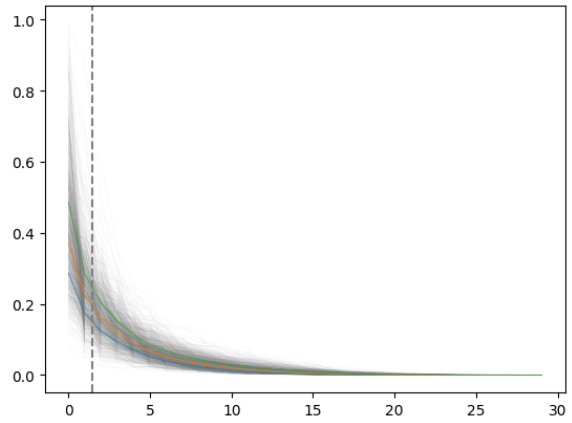

Figure 4: ordered per-label cross-entropy predictions for each example (each grey line) with the median (orange) and IQR (green & blue) over all examples. Determining a global threshold can be related to visually finding the "knee" in that median curve (dotted line)

TODO : nicer plot on another dataset (this is from RTL)

The proposed method is positioned in the lineage of *algorithm adaptation*, using *metric as losses* and allowing for *dynamic thresholding*.

11 SIGMOID F1 LOSS

For a class of multilayer perceptron $\mathcal{F}(\cdot;\Theta) : \mathcal{X} \to \mathcal{Y}$, we consider a special case, where $\mathbf{x} = \{x_1, ..., x_n\}$. Each observation is attributed one or more classes out of a label set $\mathbf{l} = \{l_1, ..., l_c\}$. Labels y_i^j are available for each observation i and class j.

For each observation i, label class probabilities can be defined based on predictions as

TODO : check this formula

$$\mathbf{p}_i = \begin{cases} \hat{\mathbf{y}} & \text{if } y = 1 \\ 1 - \hat{\mathbf{y}} & \text{otherwise} \end{cases} \tag{16}$$

Let tp and fp be number of true and false positives respectively. It is necessary to define a bound b, at which a prediction is dichotomized:

$$tp = \sum_{i \in Y^+} \mathbb{1}_{\mathbf{p}_i \geq b} \quad fp = \sum_{i \in Y^-} \mathbb{1}_{\mathbf{p}_i \geq b} \quad fn = \sum_{i \in Y^+} \mathbb{1}_{\mathbf{p}_i < b} \tag{17}$$

$\mathbb{1}_{\mathbf{p}_i \geq b}$, $\mathbb{1}_{\mathbf{p}_i < b}$ are thus the count of positive and negative predictions at threshold b,

We also define precision and recall

$$\begin{aligned} P &= \frac{tp}{tp + fp} \\ R &= \frac{tp}{tp + fn} = \frac{tp}{|Y^+|} \end{aligned} \tag{18}$$

We can then define F_β, which can be expressed as the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall than precision [25].

DOUBT : maybe ignore F_β and only mention F_1

$$F_\beta = \left(1 + \beta^2\right) \frac{P \cdot R}{\beta^2 P + R} \tag{19}$$

Or equivalently:

$$\begin{aligned} F_\beta &= \left(1 + \beta^2\right) \frac{tp}{\left(1 + \beta^2\right) tp + \beta^2 fn + fp} \\ &= \left(1 + \beta^2\right) \frac{tp}{\beta^2 |Y+| + tp + fp} \end{aligned} \tag{20}$$

Given the presence of the step indicator function $\sum \mathbb{1}_{\mathbf{p}_i \geq b}$, F_β is not differentiable for gradient based methods. One way of surpassing that problem is to use a smooth surrogate.

11.1 soft F1 score

It is possible define a *soft F1* score [3] DOUBT : can we cite a Medium post? with smooth confusion matrix entries (i.e. tp, fp and fn are not natural numbers anymore):

$$\overline{tp} = \sum \hat{\mathbf{y}} \odot \mathbf{y} \quad \overline{fp} = \sum \hat{\mathbf{y}} \odot (\mathbf{1} - \mathbf{y}) \quad \overline{fn} = \sum (\mathbf{1} - \hat{\mathbf{y}}) \odot \mathbf{y}$$

$$\mathcal{L}_{\text{softF1}} = \frac{\overline{tp}}{2\overline{tp} + \overline{fn} + \overline{fp}} \tag{21}$$

tp, fp and fn are now replaced by rough surrogates, this method has the advantage of

11.2 sigmoidF1 score

We define *sigmoidF1*, inspired by the *Maximum F1-score criterion* for automatic mispronunciation detection [14]. Whereas a sigmoid function $S(u)$

$$S(u;\beta,\eta) = \frac{1}{1+\exp(-\beta(u+\eta))}, \tag{22}$$

with β and η tunable parameters for slope and offset respectively. Higher β results in steeper slope at the center of the sigmoid and thus more stringent thresholding. At its extreme, $lim_{\beta\to\infty}S(u;\beta,\eta)$ corresponds to the step function used in Equation 29. with $S(u)$, the confusion matrix entries then become

$$\widetilde{tp} = \sum S(\hat{\mathbf{y}}) \odot \mathbf{y} \quad \widetilde{fp} = \sum S(\hat{\mathbf{y}}) \odot (\mathbf{1}-\mathbf{y}) \quad \widetilde{fn} = \sum (\mathbf{1}-S(\hat{\mathbf{y}})) \odot \mathbf{y}$$

And thus

$$\mathcal{L}_{\text{softF1}} = \frac{\widetilde{tp}}{2\widetilde{tp}+\widetilde{fn}+\widetilde{fp}} \tag{23}$$

DOUBT : mention smooth hinge loss [24]

For *sigmoidF1* β and η are tuned globally as hyperparameters. *SAdF1* (Sigmoid Adaptive F1), is an alternative where β is first set to a relatively low value and increased after each epoch. This way, a loose threshold first allows Stochastic Gradient Descent (SGD) to broadly scan the parameter space accross several local minima, before narrowing parameter search down to a promising region (similarly to adaptive learning rates).

SBayesF1 (sigmoid Bayes F1) replaces point estimates for β and η with posterior distribution estimates.

$$S(u_i) = \frac{1}{1+\exp(-\beta_i(u_i+\eta_i))} \tag{24}$$

β_i and η_i are estimated with MCMC at training time of the neural network. They are therefore implicitely allowed to vary across examples.

TODO : try SadF1 and SBayesF1 in practice

11.3 Robustness

Similarly to the focal loss [18], sigmoidF1 loss deals with class imbalance, robustness to outliers.

TODO : statistical robustness assessment

11.4 Evaluation Metrics

Loss framework for uni-instance multiclass multilabel prediction with varying number of labels

Anonymous Author(s)
Submission Id: xx

ABSTRACT

Multilabel classification is a common task in text, image or video (scene) prediction.

KEYWORDS

Keyword; Keyword; Keyword

ACM Reference Format:
Anonymous Author(s). 2021. Loss framework for uni-instance multiclass multilabel prediction with varying number of labels. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11-15, 2021, Montréal, Canada.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

12 INTRODUCTION

As neural network models are able to learn increasingly abstracter representations via deeper networks, representation learning and self-supervision, it might be reasonable to expect that, thanks to their conferred broader understanding of the world, they get better at predicting more abstract labels. Beyond objects types, face recognition, expressions, neural networks might be able to predict genres/categories TODO : other things as well? of text, image and sound. While researchers are working hard at building neural networks with very high level of understanding in the embedding space, there seems to be few research on developing loss functions that are adapted for these higher level concepts in the output space.

Although multilabel binary prediction (commonly referring to mutually inclusive labels) is a task thoroughly covered in existing litterature, there does not seem to exist a framework that deals with different amounts of positive labels in the groundtruth. For example, a scientific journal can be tagged as *machine learning* and *economics*, or a movie can be tagged as *romance* and *comedy*. These instances might as well be assigned only one tag in the groundtruth, or many more within the possible tags (classes).

Before, exploring the subject further, we will use Figure 5 to disambiguate the terminology used in this research. There seems to exist a concensus over the terms multiclass and multilabel learning, meaning respectively mutually exclusive and mutually inclusive labels TODO : source. Multilabel can therefore be seen as a subdomain of multiclass learning, where more than one class can be true for the same example. Within multilabel training, we introduce the distinction between multi-instance multilabel (e.g. [26]) and

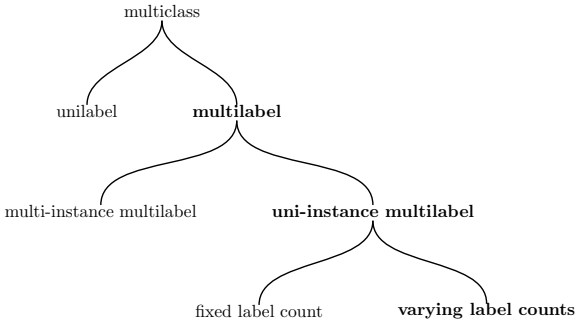

Figure 5: "multiclass" nomenclature

uni-instance multilabel. The former refers to tasks where elements within each example can be singled-out (objects in an image or expressions in a text) and assigned one or more labels. In the countrary, this paper focusses on uni-instance multilabel training (sparse occurences of the term holistic can be found in the litterature to describe this phenomenon for image [11, 15] and a recent video dataset [7] TODO : read these), more specifically with varying label counts. To the best of our knowledge, there are few existing representatives of that type of labelling task in the literature. TODO : cite more milestone examples for each category. TODO : delta with hierarchical label learning

The particularity of tasks like scientific paper tagging or movie genre classification is that it remains unclear what elements in an image/video or text can be singled out as predictive of a particular tag/genre. Rather, a complex interaction between these elements in the feature space steer the predictions. For example, the sole mention of the term "machine learning" in a paper should not be a sufficient condition to tag it as such. Instead, one could expect from the publisher to get acquainted with the paper enough to determine wether the research is a worthwhile contribution or application of *machine learning* to deserve the tag. This involves thorough understanding of the proposed method and background knowledge on state-of-the-art methods. An analogous argument can be made for movie genre classification for movie posters.

However, if elements in an image/text can be singled out as predictive of a single tag, the problem reverts back to predicting with the a priori knowledge of the existence of only one true label (i.e. multi-instance multilabel learning). The reason for distanciating singling-out from uni-instance labels, is that it has been shown that as soon as singling-out is possible, models that work on instances are more accurate TODO : rewrite this paragraph and sources. The singled-out elements can be subsets of the original feature space (typically in object detection like with the COCO dataset [19] or

SIGIR '21, July 11-15, 2021, Montréal, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

the Amazon Rainforest Dataset[1] TODO : others). Similarly, recent research has shown that the singled-out elements can be located in the abstract representations (embeddings) of the feature set and might individually predict a single true label (like GPT-3 TODO : source) TODO : more examples. This might also carry prospects of generalizability of the model [30] TODO : elaborate.

But for now, in certain retrieval tasks such as scientific journal tagging, the effect of sub-entities (either expressions in the text or single features in the embedding space) on the prediction of each label remains hard to assess. Instead we propose uni-instance (sometimes referred to as holistic) multilabel learning for varying amount of labels, with a focus on custom loss functions.

To allow the use of existing diffentiable loss fonctions, previous research papers tend to reframe the problem into either (I) a multi-instance multiclass (as described above, with the COCO dataset as an example of isolation of features [19]), (II) uni-instance multiclass prediction (III) uni-instance multilabel prediction with fixed label count (IV) uni-instance multilabel prediction with varying label count with post-training thresholding (V) redefine backpropagation for multilabel prediction [31] (VI) multitask learning [8] (VII) custom loss function [28]. This order reflects in ascending order how close modelling seem to fit the original task, which remains uni-instance multilabel learning with varying amounts of labels. DOUBT : group them

Common loss functions such as cross-entropy loss (for mutually inclusive labels) or multinomial logit loss (for mutually exclusive labels) deliver predictions on the unit interval. Thresholding the output to assess the performance of the model against the groundtruth can be done after training for (I), (II), (III) and (IV). TODO : give a very sound reason as to why we'd rather not do things post-training and rather at training-time. Problem formulations (V), (VI) and (VII) suggest a solution at training time. We think that a custom loss function (VII) is the best alternative. TODO : explain why

In a number of retrieval tasks, a model's out of sample accuracy is measured on metrics such as AUROC, F1 score, etc. These reflect an objective catered towards evaluating the model over an entire ranking. Due to to lack of differentiability, these metrics cannot be directly used as loss functions at training time (in-sample). A seminal study [9] derived a general framework for deriving decomposable surrogates to some of these metrics. We propose our own decomposable F1 surrogate tailored for the problem at hand.

We first propose a general mathematical formulation of uni-instance multilabel learning for varying amount of groundtruth labels. The generalization encompasses different levels of complexity, from the classical cross-entropy loss up to the proposed loss function. *sigmoidF1* is a F1 score surrogate which allows to optimize for label prediction and count simultanuously in a single task and is robust to outliers. It delivers more precise predictions than the current state-of-the-art on several different metrics, accross text and image related tasks. *sigmoidF1* and its adaptive *SadF1* and Bayesian *SBayesF1* counterparts are benchmarked against loss functions commonly used in multilabel learning and others tailored specifically to the uni-instance multilabel with varying number of labels setting.

[1]Available at https://www.kaggle.com/c/planet-understanding-the-amazon-from-space

13 BUILDING UP ON LOSSES

Multi-label learning can be divided into two major fields: *problem transformation* and *algorithm adaptation* [33]. In the former case, multilabel classification is reframed as a binary, multiclass classification or label ranking problem. In the latter, one tries to adapt multiclass algorithms to the problem. The current endeavour focusses on *algorithm adaptation*.

For the purpose of *problem transformation*, we define $\mathcal{L}_{\text{multiclass}}$, a class of loss functions that minimize predictions in relative terms. Binary cross-entropy, logit and their variants such as focal loss or hinge loss (deemed unstable [18]) are common choices when it comes to multiclass prediction. Cross-entropy loss can be formulated as $\mathcal{L}_{\text{CE}} = -\sum \log(p_i)$. Note that minimizing binary cross-entropy is equivalent to maximizing for log-likelihood [2, Section 4.3.4]. More generally, the *problem transformation* formulation amounts to minimizing the loss on a class of neural networks, such that

$$\min_{\mathcal{L}_{\text{multiclass}}} \mathcal{F}(\cdot;\Theta;\mathcal{L}_{\text{multiclass}}(\mathbf{y},\hat{\mathbf{y}})), \tag{25}$$

In the context of *algorithm adaptation*, where the number of positive labels in the groundtruth is unknown a priori, we aim to both obtain a propensity of each label being true and a prediction of the number of true labels:

$$\min_{\mathcal{L}_{\text{multiclass}},\mathcal{L}_{\text{count}}} \mathcal{F}(\cdot;\Theta;\mathcal{L}_{\text{multiclass}}(\mathbf{y},\hat{\mathbf{y}}) + \lambda\mathcal{L}_{\text{count}}(\mathbf{n},\hat{\mathbf{n}})), \tag{26}$$

where $n_i = \sum_j \mathbb{1}_{\mathbf{y}_i^j=1}$ is the count of positive labels per example. We thus impose a constraint for the retrieval of label counts. For example, a cross-entropy loss surrogate would penalize for the number of wrongly predicted labels $\mathcal{L}_{\text{CE+N}} = \mathcal{L}_{\text{CE}} + \lambda(\sum tp/\sum p)$, with $tp = \sum_{i\in Y^+} \mathbb{1}_{\mathbf{p}_i\geq b}$ and b a threshold to be defined. TODO : tencent loss.

This formulation is most straightfoward but suffers from higher parametrization and the lack of modelling of the interactions between label counts and label prediction. To mitigate these issues, we propose a unified loss formulation, namely

$$\min_{\mathcal{L}_{\text{multitag}}} \mathcal{F}\left(\cdot;\Theta;\mathcal{L}_{\text{multitag}}(\mathbf{y},\hat{\mathbf{y}},\mathbf{n},\hat{\mathbf{n}})\right), \tag{27}$$

Although predictions and counts explicitly appear in that formulation, $\mathcal{L}_{\text{multitag}}$ can optimize for both metrics implicitely (see proposed *sigmoidF1* below).

TODO : look at YOU ONLY TRAIN ONCE: LOSS-CONDITIONAL TRAINING OF DEEP NETWORKS

TODO : cite stat learning [13, p. 308-310]

14 RELATED WORK

TODO : look at [[https://www.sciencedirect.com/topics/computer-science/extractive-summarization][extractive summarization]]

This section will be guided by the previous section's formulation of the multitags problem, we will therefore focus on *algorithm adaptation*, *metrics as losses* and *dynamic thresholding*.

14.1 algorithm adaptation

Early representatives of *algorithm adaptation* stem from heterogenous domains of machine learning. Multi-Label k-Nearest Neighbors [32], Multi-Label Decision Tree [6], Ranking Support Vector Machine [10] and Backpropagation for Multi-Label Learning [31]. More recently, two papers introduced the idea of multitask learning for *label prediction* and *label count prediction* for text (ML_{NET}) [8] and image [17] data. The latter research is loosely catered towards object detection (although not formally presented as such) and is thus out-of-scope: elements in a picture are predicted that tend to be unilabel as defined by the groundtruth (e.g. cat, flower, vase, person, bottle etc.).

14.2 metrics as losses

Often, machine learning post-training evaluation metrics (e.g. AUROC, F1) are not differentiable. There are motivations TODO : which motivations for optimizing a model directly on a metric at training time. A general framework for AUC, AUROC and F1 is presented in [9], but the proposed F1 surrogate remains short of being explicitly derived for stochastic gradient descent. TODO : check again with the authors if I can't get inspired from their work. Recently, a similar work has been proposed to train a Convolutional Neural Network (CNN) from scratch with a few millions of images and hundreds of labels specifically for multilabel tasks [28]. This task is loosely related to object detection, similarly to [17] mentioned in the previous paragraph.

14.3 dynamic thresholding

dynamic thresholding accross classes or examples is an issue as soon as the number of labels to predict is unknown. Certain variants of cross-entropy loss accomodate imbalanced label data [18], but remain agnostic towards the number of labels to predict. Solutions have been tailored to that end, starting with determining an ideal global *threshold* depending on use-cases [20], or per-class-thresholding after training [5] and eventually abstracting the threshold away via a *soft-F1* measure [3] TODO : say more about this method. In the latter two cases, the task is to predict genre from movie posters.

TODO : nicer plot on another dataset (this is from RTL)

The proposed method is positioned in the lineage of *algorithm adaptation*, using *metric as losses* and allowing for *dynamic thresholdig*.

15 SIGMOID F1 LOSS

For a class of multilayer perceptron $\mathcal{F}(\cdot;\Theta) : \mathcal{X} \rightarrow \mathcal{Y}$, we consider a special case, where $\mathbf{x} = \{x_1, ..., x_n\}$. Each observation is attributed one or more classes out of a label set $\mathbf{l} = \{l_1, ..., l_c\}$. Labels y_i^j are available for each observation i and class j.

For each observation i, label class probabilities can be defined based on predictions as

TODO : check this formula

$$\mathbf{p}_i = \begin{cases} \hat{\mathbf{y}} & \text{if } y = 1 \\ 1 - \hat{\mathbf{y}} & \text{otherwise} \end{cases} \tag{28}$$

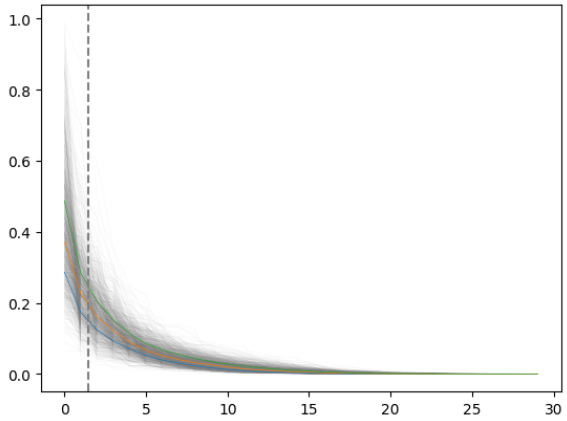

Figure 6: ordered per-label cross-entropy predictions for each example (each grey line) with the median (orange) and IQR (green & blue) over all examples. Determining a global threshold can be related to visually finding the "knee" in that median curve (dotted line)

Let tp and fp be number of true and false positives respectively. It is necessary to define a bound b, at which a prediction is dichotomized:

$$tp = \sum_{i \in Y^+} \mathbb{1}_{\mathbf{p}_i \geq b} \quad fp = \sum_{i \in Y^-} \mathbb{1}_{\mathbf{p}_i \geq b} \quad fn = \sum_{i \in Y^+} \mathbb{1}_{\mathbf{p}_i < b} \tag{29}$$

$\mathbb{1}_{\mathbf{p}_i \geq b}$, $\mathbb{1}_{\mathbf{p}_i < b}$ are thus the count of positive and negative predictions at threshold b,

We also define precision and recall

$$\begin{aligned} P &= \frac{tp}{tp + fp} \\ R &= \frac{tp}{tp + fn} = \frac{tp}{|Y^+|} \end{aligned} \tag{30}$$

We can then define F_β, which can be expressed as the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall than precision [25].

DOUBT : maybe ignore F_β and only mention F_1

$$F_\beta = \left(1 + \beta^2\right) \frac{P \cdot R}{\beta^2 P + R} \tag{31}$$

Or equivalently:

$$\begin{aligned} F_\beta &= \left(1 + \beta^2\right) \frac{tp}{\left(1 + \beta^2\right) tp + \beta^2 fn + fp} \\ &= \left(1 + \beta^2\right) \frac{tp}{\beta^2 |Y + | + tp + fp} \end{aligned} \tag{32}$$

Given the presence of the step indicator function $\sum \mathbb{1}_{\mathbf{p}_i \geq b}$, F_β is not differentiable for gradient based methods. One way of surpassing that problem is to use a smooth surrogate.

15.1 soft F1 score

It is possible define a *soft F1* score [3] DOUBT : can we cite a Medium post? with smooth confusion matrix entries (i.e. tp, fp and fn are not natural numbers anymore):

$$\overline{tp} = \sum \hat{\mathbf{y}} \odot \mathbf{y} \quad \overline{fp} = \sum \hat{\mathbf{y}} \odot (\mathbf{1} - \mathbf{y}) \quad \overline{fn} = \sum (\mathbf{1} - \hat{\mathbf{y}}) \odot \mathbf{y}$$

$$\mathcal{L}_{\text{softF1}} = \frac{\overline{tp}}{2\overline{tp} + \overline{fn} + \overline{fp}} \tag{33}$$

tp, fp and fn are now replaced by rough surrogates, this method has the advantage of

15.2 sigmoidF1 score

We define *sigmoidF1*, inspired by the *Maximum F1-score criterion* for automatic mispronounciation detection [14]. Whereas a sigmoid function $S(u)$

$$S(u; \beta, \eta) = \frac{1}{1 + \exp(-\beta(u + \eta))}, \tag{34}$$

with β and η tunable parameters for slope and offset respectively. Higher β results in steeper slope at the center of the sigmoid and thus more stringent thresholding. At its extreme, $lim_{\beta \to \infty} S(u; \beta, \eta)$ corresponds to the step function used in Equation 29. with $S(u)$, the confusion matrix entries then become

$$\widetilde{tp} = \sum S(\hat{\mathbf{y}}) \odot \mathbf{y} \quad \widetilde{fp} = \sum S(\hat{\mathbf{y}}) \odot (\mathbf{1} - \mathbf{y}) \quad \widetilde{fn} = \sum (\mathbf{1} - S(\hat{\mathbf{y}})) \odot \mathbf{y}$$

And thus

$$\mathcal{L}_{\text{softF1}} = \frac{\widetilde{tp}}{2\widetilde{tp} + \widetilde{fn} + \widetilde{fp}} \tag{35}$$

DOUBT : mention smooth hinge loss [24]

For *sigmoidF1* β and η are tuned globally as hyperparameters. *SAdF1* (Sigmoid Adaptive F1), is an alternative where β is first set to a relatively low value and increased after each epoch. This way, a loose threshold first allows Stochastic Gradient Descent (SGD) to broadly scan the parameter space accross several local minima, before narrowing parameter search down to a promissing region (similarily to adaptive learning rates).

SBayesF1 (sigmoid Bayes F1) replaces point estimates for β and η with posterior distribution estimates.

$$S(u_i) = \frac{1}{1 + \exp(-\beta_i(u_i + \eta_i))} \tag{36}$$

β_i and η_i are estimated with MCMC at training time of the neural network. They are therefore implicitely allowed to vary across examples.

TODO : try SadF1 and SBayesF1 in practice

15.3 Robustness

Similarly to the focal loss [18], sigmoidF1 loss deals with class imbalance, robustness to outliers.

TODO : statistical robustness assessment

15.4 Evaluation Metrics

The metrics described below are a result of a survey of different common practices for measuring accuracy of multilabel prediction. When true positives and false positives are used, recall that $tp = \sum_{i \in Y^+} \mathbb{1}_{\mathbf{p}_i \geq b}$ and $fp = \sum_{i \in Y^-} \mathbb{1}_{\mathbf{p}_i \geq b}$, and thus a threshold b must be set. When $b = 0.5$, as is commonly done [SOURCE HERE], a risk remains that a lot of examples remain without predictions.

Extending F_1 to multi-class binary classification amounts to deciding wether to un/pool classes. In a first pooled iteration, micro F_1 [SOURCE HERE] equates to creating a single 2x2 confusion matrix for all classes:

$$F_1^{micro} = \frac{\sum tp_c}{2\sum tp_c + \sum fn_c + \sum fp_c} \quad for \quad c \in C$$

Macro F_1 [20] amounts to creating one confusion matrix per class:

$$F_1^{macro} = \frac{1}{c} \sum_{j=1}^{c} F_1$$

DOUBT : Do we need to justify optimizing for an F1 surrogate at training time and to then use F1 itself as a metric?

Weighted macro F_1 TODO : find source is similar but includes weighing to account for class imbalance, i.e. weighing each class by the number of groundtruth positives.

$$F_1^{weighted} = \frac{1}{c} \sum_{j=1}^{c} n_j F_1 \quad where \quad n_j = \sum_i \mathbb{1}_{y_i^j = 1}$$

Accuracy is the overall fraction of correctly predicted labels [20]:

$$Acc = \frac{tp + tn}{tp + tn + fp + fn}$$

TODO compare to [29]

16 DATASETS

sigmoidF1 is tested across different modalities, namely image, video, sound and text, with a focus on text: the most comparable research was on text data.

Among the three datasets used for benchmarking ML-NET [8], a cancer hallmark dataset is of multi-instance multilabel nature [12] [2]: the research clearly describe a process of annotating several expressions within paper abstracts. The remaining two datasets for chemical exposure [16] [3] and diagnosis codes assigment [23] [4], seem to fit to the entity wide multilabel definition but have a strong hierarchical nature. Although slightly out-of-scope, the three datasets above will be used for benchmarking, since they were used to test ML-NET, which is the state-of-the-art in *algorithm adaptation* for text to the best of our knowledge.

For a broader scope in learning for text data, we also use the newly created *Arxiv dataset* [5] with data on abstracts of 1.7 million open source articles and their categories (suitably mutually inclusive and of varying count per example).

In the vision domain, a dataset of movie posters [6] and their genre is used. Similarly, labels are mutually inclusive and of varying count per example. It is arguable that is hard to single out elements in the image of a poster that define the genre of a movie. Rather it might be a combination of the title font, the background image, the presence of actors and specific objects such as cars, weapons etc.

TODO : I removed all jpg's that are empty in the prescraped data. I could try to scrape the posters myself to see if I get more

Another recently created dataset was made available for *Large Scale Holistic Video Understanding* [7] [7], as defined in the introduction.

TODO : this is an ambitious number of datasets. Add longer description of each dataset, depending on which ones I keep: sample size, number of classes etc. see utils here: https://github.com/ashrefm/multi-label-soft-f1

DOUBT : cite Kaggle datasets formally instead of using links: https://www.kaggle.com/data/46091

DOUBT : add a music genre classification dataset, for which Vincent Koops at RTL could help train

[2] Available at https://www.cl.cam.ac.uk/sim;sb895/HoC.html
[3] Available at https://figshare.com/articles/Corpus$_a nd_S of t ware$/4668229
[4] Available at https://physionet.org/works/ICD9CodingofDischargeSummaries
[5] Available at https://www.kaggle.com/Cornell-University/arxiv
[6] Labels available at https://tinyurl.com/y7ydyedu and prescraped images from IMDB at https://tinyurl.com/y7lfpvlx
[7] Available at https://github.com/holistic-video-understanding/HVU-Dataset

17 EXPERIMENTAL RESULTS

varying b in the sigmoid function as if it is an adaptive learning rate TODO : actually try it out.

one b per class

if we consider b and c to be probabilistic, we can then use tensorflow probability to assess their distribution

the batch size has to be relatively large (i.e. 256), in order for meaningful F1 surrogates to be calculated.

movie posters (CNN)

Loss	macroF @ 0.5	microF1 @ 0.5	weightedF1 @ 0.5	Precision @ 0.5	Recall @ 0.5
$\mathcal{L}_{\text{CE}}$	0.057	0.200	0.159	0.106	0.106
$\mathcal{L}_{\text{FL}}$	0.055	0.192	0.154	0.115	0.115
$\mathcal{L}_{\text{CE+N}}$	0	0	0	0	0
$\mathcal{L}_{\text{CE+T}}$	0	0	0	0	0
$\mathcal{L}_{\text{macroSoftF1}}$	0.132	0.323	0.280	0.105	0.105
$\mathcal{L}_{\text{sigmoidF1}}$	**0.117**	**0.240**	**0.263**	**0.103**	**0.103**

Arxiv (distillBERT)

Metric	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{FL}}$	$\mathcal{L}_{\text{CE+N}}$	$\mathcal{L}_{\text{CE+T}}$
$P(\%)$	0	0	0	0
$R(\%)$	0	0	0	0
$F_1(\%)$	0	0	0	**0**

Cancer hallmark (distillBERT)

Metric	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{FL}}$	$\mathcal{L}_{\text{CE+N}}$	$\mathcal{L}_{\text{CE+T}}$
$P(\%)$	0	0	0	0
$R(\%)$	0	0	0	0
$F_1(\%)$	0	0	0	**0**

Chemical exposure (distillBERT)

Metric	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{FL}}$	$\mathcal{L}_{\text{CE+N}}$	$\mathcal{L}_{\text{CE+T}}$
$P(\%)$	0	0	0	0
$R(\%)$	0	0	0	0
$F_1(\%)$	0	0	0	**0**

Chemical exposure (distillBERT)

Metric	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{FL}}$	$\mathcal{L}_{\text{CE+N}}$	$\mathcal{L}_{\text{CE+T}}$
$P(\%)$	0	0	0	0
$R(\%)$	0	0	0	0
$F_1(\%)$	0	0	0	**0**

simulated data

Metric	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{FL}}$	$\mathcal{L}_{\text{CE+N}}$	$\mathcal{L}_{\text{CE+T}}$
$P(\%)$	0	0	0	0
$R(\%)$	0	0	0	0
$F_1(\%)$	0	0	0	**0**

18 CONCLUSION

Shortcomings

it is debatable wether any task is intrinsincly multilabel and wether the image / text cannot be decomposed in parts that are single labelled.

not long training and small models, but aibility to demonstrate the statement anyways.

Results

In this paper we defined a new problem in deep learning for mulitple modalities that harness the current advances in abstract representation of the input space. A general loss framework is proposed to locate that solution within the existing multiclass multilabel losses and a specific loss function is formulated. *sigmoidF1* achieves significantly results for different F1 values on all datasets.

Future work

Apply the loss function to more sophisticated neural network architectures that use F1 score as an evaluation metric such as AC-SUM-GAN [1].

This model can be adapted for hiarchical multilabel classification or active learning (for both see [22]).

Combine the proposed loss functions with representation learning [21, 27] or self-supervised learning, in order to model abstract relationships between the labels.

adapt to *extreme* multilabel prediction [4]

REFERENCES

[1] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras. 2020. AC-SUM-GAN: Connecting Actor-Critic and Generative Adversarial Networks for Unsupervised Video Summarization,. *IEEE Transactions on Circuits and Systems for Video Technology* (2020).

[2] Christopher M. Bishop. 2007. *Pattern recognition and machine learning, 5th Edition.* Springer. https://www.worldcat.org/oclc/71008143

[3] Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S. Dhillon. 2019. The Unknown Benefits of using a Soft-F1 Loss in Classification Systems. *Towards Data Science* (Dec 2019). https://towardsdatascience.com/the-unknown-benefits-of-using-a-soft-f1-loss-in-classification-systems-753902c0105d

[4] Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S. Dhillon. 2020. Taming Pretrained Transformers for Extreme Multi-label Text Classification. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining* (Jul 2020). https://doi.org/10.1145/3394486.3403368

[5] Wei-Ta Chu and Hung-Jui Guo. 2017. Movie Genre Classification based on Poster Images with Deep Neural Networks. *Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes* (Oct 2017). https://doi.org/10.1145/3132515.3132516

[6] Amanda Clare and Ross D. King. 2001. Knowledge Discovery in Multi-label Phenotype Data. *Lecture Notes in Computer Science* (2001), 42–53. https://doi.org/10.1007/3-540-44794-6_4

[7] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jurgen Gall, Rainer Stiefelhagen, and Luc Van Gool. 2019. *Large Scale Holistic Video Understanding.* arXiv:1904.11451v3 [cs.CV]

[8] Jingcheng Du, Qingyu Chen, Yifan Peng, Yang Xiang, Cui Tao, and Zhiyong Lu. 2019. ML-Net: multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association* 26, 11 (Jun 2019), 1279–1285. https://doi.org/10.1093/jamia/ocz085

[9] Elad ET. Eban, Mariano Schain, Alan Mackey, Ariel Gordon, Rif A. Saurous, and Gal Elidan. 2016. *Scalable Learning of Non-Decomposable Objectives.* arXiv:1608.04802v2 [stat.ML]

[10] André Elisseeff and Jason Weston. 2001. A Kernel Method for Multi-Labelled Classification. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic* (Vancouver, British Columbia, Canada) *(NIPS'01)*. MIT Press, Cambridge, MA, USA, 681–687.

[11] Mingchen Gao, Ulas Bagci, Le Lu, Aaron Wu, Mario Buty, Hoo-Chang Shin, Holger Roth, Georgios Z. Papadakis, Adrien Depeursinge, Ronald M. Summers, and et al. 2016. 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 (Jun 2016), 1–6. https://doi.org/10.1080/21681163.2015.1124249

[12] Douglas Hanahan and Robert A. Weinberg. 2011. Hallmarks of Cancer: The Next Generation. *Cell* 144, 5 (Mar 2011), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

[13] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. The Elements of Statistical Learning. *Springer Series in Statistics* (2009). https://doi.org/10.1007/978-0-387-84858-7

[14] Hao Huang, Haihua Xu, Xianhui Wang, and Wushour Silamu. 2015. Maximum F1-Score Discriminative Training Criterion for Automatic Mispronunciation Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 4 (Apr 2015), 787–797. https://doi.org/10.1109/taslp.2015.2409733

[15] Alberto Jaenal, Francisco-Angel Moreno, and Javier Gonzalez-Jimenez. 2019. Experimental study of the suitability of CNN-based holistic descriptors for accurate visual localization. *Proceedings of the 2nd International Conference on Applications of Intelligent Systems - APPIS '19* (2019). https://doi.org/10.1145/3309772.3309800

[16] Kristin Larsson, Ilona Silins, Yufan Guo, Anna Korhonen, Ulla Stenius, and Marika Berglund. 2014. Text mining for improved human exposure assessment. *Toxicology Letters* 229 (Sep 2014), S119. https://doi.org/10.1016/j.toxlet.2014.06.427

[17] Yuncheng Li, Yale Song, and Jiebo Luo. 2017. Improving Pairwise Ranking for Multi-label Image Classification. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Jul 2017). https://doi.org/10.1109/cvpr.2017.199

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal Loss for Dense Object Detection. *2017 IEEE International Conference on Computer Vision (ICCV)* (Oct 2017). https://doi.org/10.1109/iccv.2017.324

[19] 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. *Lecture Notes in Computer Science* (2014), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[20] Zachary C. Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. 2014. Optimal Thresholding of Classifiers to Maximize F1 Measure. *Lecture Notes in Computer Science* (2014), 225–239. https://doi.org/10.1007/978-3-662-44851-9_15

[21] Timo Milbich, Omair Ghori, Ferran Diego, and Björn Ommer. 2020. Unsupervised representation learning by discovering reliable image relations. *Pattern Recognition* 102 (Jun 2020), 107107. https://doi.org/10.1016/j.patcog.2019.107107

[22] Felipe Kenji Nakano, Ricardo Cerri, and Celine Vens. 2020. Active learning for hierarchical multi-label classification. *Data Mining and Knowledge Discovery* 34, 5 (2020), 1496–1530. https://doi.org/10.1007/s10618-020-00704-w

[23] Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2014. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association* 21, 2 (Mar 2014), 231–237. https://doi.org/10.1136/amiajnl-2013-002159

[24] Jason DM Rennie. 2005. Smooth hinge classification. *Proceeding of Massachusetts Institute of Technology* (2005).

[25] C.J. van Rijsbergen. [n.d.]. *Information retrieval* (2nd [rev.] ed. ed.). Butterworths, London [etc.

[26] H. Soleimani and D. J. Miller. 2017. Semisupervised, Multilabel, Multi-Instance Learning for Structured Data. *Neural Computation* 29, 4 (2017), 1053–1102. https://doi.org/10.1162/NECO_a_00939

[27] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, and et al. 2020. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), 1–1. https://doi.org/10.1109/tpami.2020.2983686

[28] Baoyuan Wu, Weidong Chen, Yanbo Fan, Yong Zhang, Jinlong Hou, Jie Liu, and Tong Zhang. 2019. Tencent ML-Images: A Large-Scale Multi-Label Image Database for Visual Representation Learning. *IEEE Access* 7 (2019), 172683–172693. https://doi.org/10.1109/access.2019.2956775

[29] Hichame Yessou, Gencer Sumbul, and Begüm Demir. 2020. *A Comparative Study of Deep Learning Loss Functions for Multi-Label Remote Sensing Image Classification.* arXiv:2009.13935v1 [cs.CV]

[30] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. https://arxiv.org/abs/1611.03530

[31] Min-Ling Zhang and Zhi-Hua Zhou. 2006. Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization. *IEEE Transactions on Knowledge and Data Engineering* 18, 10 (Oct 2006), 1338–1351. https://doi.org/10.1109/tkde.2006.162

[32] Min-Ling Zhang and Zhi-Hua Zhou. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition* 40, 7 (Jul 2007), 2038–2048. https://doi.org/10.1016/j.patcog.2006.12.019

[33] Min-Ling Zhang and Zhi-Hua Zhou. 2014. A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26, 8 (Aug 2014), 1819–1837. https://doi.org/10.1109/tkde.2013.39

The metrics described below are a result of a survey of different common practices for measuring accuracy of multilabel prediction. When true positives and false positives are used, recall that $tp = \sum_{i \in Y^+} \mathbb{1}_{\mathbf{p_i} \geq b}$ and $fp = \sum_{i \in Y^-} \mathbb{1}_{\mathbf{p_i} \geq b}$, and thus a threshold b must be set. When $b = 0.5$, as is commonly done [SOURCE HERE], a risk remains that a lot of examples remain without predictions.

Extending F_1 to multi-class binary classification amounts to deciding wether to un/pool classes. In a first pooled iteration, micro F_1 [SOURCE HERE] equates to creating a single 2x2 confusion matrix for all classes:

$$F_1^{micro} = \frac{\sum tp_c}{2 \sum tp_c + \sum fn_c + \sum fp_c} \quad for \quad c \in C$$

Macro F_1 [20] amounts to creating one confusion matrix per class:

$$F_1^{macro} = \frac{1}{c}\sum_{j=1}^{c} F_1$$

DOUBT : Do we need to justify optimizing for an F1 surrogate at training time and to then use F1 itself as a metric?

Weighted macro F_1 TODO : find source is similar but includes weighing to account for class imbalance, i.e. weighing each class by the number of groundtruth positives.

$$F_1^{weighted} = \frac{1}{c}\sum_{j=1}^{c} n_j F_1 \quad where \quad n_j = \sum_i \mathbb{1}_{y_i^j=1}$$

Accuracy is the overall fraction of correctly predicted labels [20]:

$$Acc = \frac{tp + tn}{tp + tn + fp + fn}$$

TODO compare to [29]

19 DATASETS

sigmoidF1 is tested across different modalities, namely image, video, sound and text, with a focus on text: the most comparable research was on text data.

Among the three datasets used for benchmarking ML-NET [8], a cancer hallmark dataset is of multi-instance multilabel nature [12] [8]: the research clearly describe a process of annotating several expressions within paper abstracts. The remaining two datasets for chemical exposure [16] [9] and diagnosis codes assignment [23] [10], seem to fit to the entity wide multilabel definition but have a strong hierarchical nature. Although slightly out-of-scope, the three datasets above will be used for benchmarking, since they were used to test ML-NET, which is the state-of-the-art in *algorithm adaptation* for text to the best of our knowledge.

For a broader scope in learning for text data, we also use the newly created *Arxiv dataset* [11] with data on abstracts of 1.7 million open source articles and their categories (suitably mutually inclusive and of varying count per example).

In the vision domain, a dataset of movie posters [12] and their genre is used. Similarly, labels are mutually inclusive and of varying count per example. It is arguable that is hard to single out elements in the image of a poster that define the genre of a movie. Rather it might be a combination of the title font, the background image, the presence of actors and specific objects such as cars, weapons etc.

TODO : I removed all jpg's that are empty in the prescraped data. I could try to scrape the posters myself to see if I get more

Another recently created dataset was made available for *Large Scale Holistic Video Understanding* [7] [13], as defined in the introduction.

TODO : this is an ambitious number of datasets. Add longer description of each dataset, depending on which ones I keep: sample size, number of classes etc. see utils here: https://github.com/ashrefm/multi-label-soft-f1

DOUBT : cite Kaggle datasets formally instead of using links: https://www.kaggle.com/data/46091

DOUBT : add a music genre classification dataset, for which Vincent Koops at RTL could help train

[8] Available at https://www.cl.cam.ac.uk/sim;sb895/HoC.html
[9] Available at https://figshare.com/articles/Corpus$_a nd_S oftware$/4668229
[10] Available at https://physionet.org/works/ICD9CodingofDischargeSummaries
[11] Available at https://www.kaggle.com/Cornell-University/arxiv
[12] Labels available at https://tinyurl.com/y7ydyedu and prescraped images from IMDB at https://tinyurl.com/y7lfpvlx
[13] Available at https://github.com/holistic-video-understanding/HVU-Dataset

20 EXPERIMENTAL RESULTS

varying b in the sigmoid function as if it is an adaptive learning rate TODO : actually try it out.

one b per class

if we consider b and c to be probabilistic, we can then use tensorflow probability to assess their distribution

the batch size has to be relatively large (i.e. 256), in order for meaningful F1 surrogates to be calculated.

movie posters (CNN)

Loss	macroF @ 0.5	microF1 @ 0.5	weightedF1 @ 0.5	Precision @ 0.5	Recall @ 0.5
$\mathcal{L}_{CE}$	0.057	0.200	0.159	0.106	0.106
$\mathcal{L}_{FL}$	0.055	0.192	0.154	0.115	0.115
$\mathcal{L}_{CE+N}$	0	0	0	0	0
$\mathcal{L}_{CE+T}$	0	0	0	0	0
$\mathcal{L}_{macroSoftF1}$	0.132	0.323	0.280	0.105	0.105
$\mathcal{L}_{sigmoidF1}$	**0.117**	**0.240**	**0.263**	**0.103**	**0.103**

Arxiv (distillBERT)

Metric	$\mathcal{L}_{CE}$	$\mathcal{L}_{FL}$	$\mathcal{L}_{CE+N}$	$\mathcal{L}_{CE+T}$
$P(\%)$	0	0	0	0
$R(\%)$	0	0	0	0
$F_1(\%)$	0	0	0	**0**

Cancer hallmark (distillBERT)

Metric	$\mathcal{L}_{CE}$	$\mathcal{L}_{FL}$	$\mathcal{L}_{CE+N}$	$\mathcal{L}_{CE+T}$
$P(\%)$	0	0	0	0
$R(\%)$	0	0	0	0
$F_1(\%)$	0	0	0	**0**

Chemical exposure (distillBERT)

Metric	$\mathcal{L}_{CE}$	$\mathcal{L}_{FL}$	$\mathcal{L}_{CE+N}$	$\mathcal{L}_{CE+T}$
$P(\%)$	0	0	0	0
$R(\%)$	0	0	0	0
$F_1(\%)$	0	0	0	**0**

Chemical exposure (distillBERT)

Metric	$\mathcal{L}_{CE}$	$\mathcal{L}_{FL}$	$\mathcal{L}_{CE+N}$	$\mathcal{L}_{CE+T}$
$P(\%)$	0	0	0	0
$R(\%)$	0	0	0	0
$F_1(\%)$	0	0	0	**0**

simulated data

Metric	$\mathcal{L}_{CE}$	$\mathcal{L}_{FL}$	$\mathcal{L}_{CE+N}$	$\mathcal{L}_{CE+T}$
$P(\%)$	0	0	0	0
$R(\%)$	0	0	0	0
$F_1(\%)$	0	0	0	**0**

21 CONCLUSION

Shortcomings

it is debatable whether any task is intrinsically multilabel and wether the image / text cannot be decomposed in parts that are single labeled.

not long training and small models, but ability to demonstrate the statement anyways.

Results

In this paper we defined a new problem in deep learning for multiple modalities that harness the current advances in abstract representation of the input space. A general loss framework is proposed to locate that solution within the existing multiclass multilabel losses and a specific loss function is formulated. *sigmoidF1* achieves significantly results for different F1 values on all datasets.

Future work

Apply the loss function to more sophisticated neural network architectures that use F1 score as an evaluation metric such as AC-SUM-GAN [1].

This model can be adapted for hierarchical multilabel classification or active learning (for both see [22]).

Combine the proposed loss functions with representation learning [21, 27] or self-supervised learning, in order to model abstract relationships between the labels.

adapt to *extreme* multilabel prediction [4]

REFERENCES

[1] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras. 2020. AC-SUM-GAN: Connecting Actor-Critic and Generative Adversarial Networks for Unsupervised Video Summarization,. *IEEE Transactions on Circuits and Systems for Video Technology* (2020).

[2] Christopher M. Bishop. 2007. *Pattern recognition and machine learning, 5th Edition.* Springer. https://www.worldcat.org/oclc/71008143

[3] Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S. Dhillon. 2019. The Unknown Benefits of using a Soft-F1 Loss in Classification Systems. *Towards Data Science* (Dec 2019). https://towardsdatascience.com/the-unknown-benefits-of-using-a-soft-f1-loss-in-classification-systems-753902c0105d

[4] Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S. Dhillon. 2020. Taming Pretrained Transformers for Extreme Multi-label Text Classification. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining* (Jul 2020). https://doi.org/10.1145/3394486.3403368

[5] Wei-Ta Chu and Hung-Jui Guo. 2017. Movie Genre Classification based on Poster Images with Deep Neural Networks. *Proceedings of the Workshop on Multimodal Understanding of Social, Affective and Subjective Attributes* (Oct 2017). https://doi.org/10.1145/3132515.3132516

[6] Amanda Clare and Ross D. King. 2001. Knowledge Discovery in Multi-label Phenotype Data. *Lecture Notes in Computer Science* (2001), 42–53. https://doi.org/10.1007/3-540-44794-6_4

[7] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jurgen Gall, Rainer Stiefelhagen, and Luc Van Gool. 2019. *Large Scale Holistic Video Understanding.* arXiv:1904.11451v3 [cs.CV]

[8] Jingcheng Du, Qingyu Chen, Yifan Peng, Yang Xiang, Cui Tao, and Zhiyong Lu. 2019. ML-Net: multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association* 26, 11 (Jun 2019), 1279–1285. https://doi.org/10.1093/jamia/ocz085

[9] Elad ET. Eban, Mariano Schain, Alan Mackey, Ariel Gordon, Rif A. Saurous, and Gal Elidan. 2016. *Scalable Learning of Non-Decomposable Objectives.* arXiv:1608.04802v2 [stat.ML]

[10] André Elisseeff and Jason Weston. 2001. A Kernel Method for Multi-Labelled Classification. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic* (Vancouver, British Columbia, Canada) *(NIPS'01)*. MIT Press, Cambridge, MA, USA, 681–687.

[11] Mingchen Gao, Ulas Bagci, Le Lu, Aaron Wu, Mario Buty, Hoo-Chang Shin, Holger Roth, Georgios Z. Papadakis, Adrien Depeursinge, Ronald M. Summers, and et al. 2016. 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 (Jun 2016), 1–6. https://doi.org/10.1080/21681163.2015.1124249

[12] Douglas Hanahan and Robert A. Weinberg. 2011. Hallmarks of Cancer: The Next Generation. *Cell* 144, 5 (Mar 2011), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

[13] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. The Elements of Statistical Learning. *Springer Series in Statistics* (2009). https://doi.org/10.1007/978-0-387-84858-7

[14] Hao Huang, Haihua Xu, Xianhui Wang, and Wushour Silamu. 2015. Maximum F1-Score Discriminative Training Criterion for Automatic Mispronunciation Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 4 (Apr 2015), 787–797. https://doi.org/10.1109/taslp.2015.2409733

[15] Alberto Jaenal, Francisco-Angel Moreno, and Javier Gonzalez-Jimenez. 2019. Experimental study of the suitability of CNN-based holistic descriptors for accurate visual localization. *Proceedings of the 2nd International Conference on Applications of Intelligent Systems - APPIS '19* (2019). https://doi.org/10.1145/3309772.3309800

[16] Kristin Larsson, Ilona Silins, Yufan Guo, Anna Korhonen, Ulla Stenius, and Marika Berglund. 2014. Text mining for improved human exposure assessment. *Toxicology Letters* 229 (Sep 2014), S119. https://doi.org/10.1016/j.toxlet.2014.06.427

[17] Yuncheng Li, Yale Song, and Jiebo Luo. 2017. Improving Pairwise Ranking for Multi-label Image Classification. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Jul 2017). https://doi.org/10.1109/cvpr.2017.199

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal Loss for Dense Object Detection. *2017 IEEE International Conference on Computer Vision (ICCV)* (Oct 2017). https://doi.org/10.1109/iccv.2017.324

[19] 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. *Lecture Notes in Computer Science* (2014), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[20] Zachary C. Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. 2014. Optimal Thresholding of Classifiers to Maximize F1 Measure. *Lecture Notes in Computer Science* (2014), 225–239. https://doi.org/10.1007/978-3-662-44851-9_15

[21] Timo Milbich, Omair Ghori, Ferran Diego, and Björn Ommer. 2020. Unsupervised representation learning by discovering reliable image relations. *Pattern Recognition* 102 (Jun 2020), 107107. https://doi.org/10.1016/j.patcog.2019.107107

[22] Felipe Kenji Nakano, Ricardo Cerri, and Celine Vens. 2020. Active learning for hierarchical multi-label classification. *Data Mining and Knowledge Discovery* 34, 5 (2020), 1496–1530. https://doi.org/10.1007/s10618-020-00704-w

[23] Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2014. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association* 21, 2 (Mar 2014), 231–237. https://doi.org/10.1136/amiajnl-2013-002159

[24] Jason DM Rennie. 2005. Smooth hinge classification. *Proceeding of Massachusetts Institute of Technology* (2005).

[25] C.J. van Rijsbergen. [n.d.]. *Information retrieval* (2nd [rev.] ed. ed.). Butterworths, London [etc.

[26] H. Soleimani and D. J. Miller. 2017. Semisupervised, Multilabel, Multi-Instance Learning for Structured Data. *Neural Computation* 29, 4 (2017), 1053–1102. https://doi.org/10.1162/NECO_a_00939

[27] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, and et al. 2020. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), 1–1. https://doi.org/10.1109/tpami.2020.2983686

[28] Baoyuan Wu, Weidong Chen, Yanbo Fan, Yong Zhang, Jinlong Hou, Jie Liu, and Tong Zhang. 2019. Tencent ML-Images: A Large-Scale Multi-Label Image Database for Visual Representation Learning. *IEEE Access* 7 (2019), 172683–172693. https://doi.org/10.1109/access.2019.2956775

[29] Hichame Yessou, Gencer Sumbul, and Begüm Demir. 2020. *A Comparative Study of Deep Learning Loss Functions for Multi-Label Remote Sensing Image Classification.* arXiv:2009.13935v1 [cs.CV]

[30] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. https://arxiv.org/abs/1611.03530

[31] Min-Ling Zhang and Zhi-Hua Zhou. 2006. Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization. *IEEE Transactions on Knowledge and Data Engineering* 18, 10 (Oct 2006), 1338–1351. https://doi.org/10.1109/tkde.2006.162

[32] Min-Ling Zhang and Zhi-Hua Zhou. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition* 40, 7 (Jul 2007), 2038–2048. https://doi.org/10.1016/j.patcog.2006.12.019

[33] Min-Ling Zhang and Zhi-Hua Zhou. 2014. A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26, 8 (Aug 2014), 1819–1837. https://doi.org/10.1109/tkde.2013.39