Loss framework for entity-wide 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 entity-wide 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, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As neural network models are able to learn more and more abstract 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 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 understandings, 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 1 to disambiguate the terminology used in this research. There seems to exist a concensus over the terms multiclass and multilabel, meaning respectively mutually exclusive and mutually inclusive labels. 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 sub-entity unilabel and entity-wide multilabel. The former refers

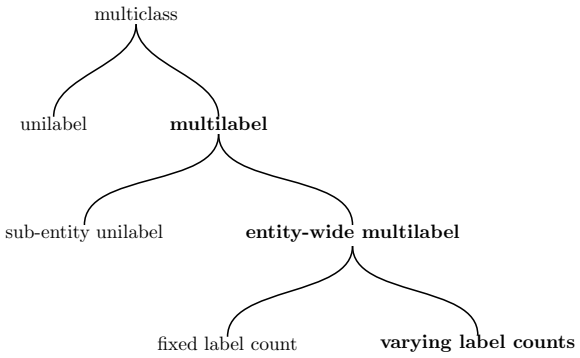


Figure 1: "multiclass" nomenclature

to tasks where elements within each example can be singled-out (objects in an image or expressions in a text) and assigned a single label. In the countrary, this paper focusses on entity-wide multilabel training with varying label counts. These disctinctions will prove useful troughout the text.

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. The singled-out elements can be subsets of the original feature space (typically in object detection like with the COCO dataset [16] and TODO : others). Similarly, it is possible that abstract representations of the feature set would allow for singling out of tasks that predict a single true label in the future (like GPT-3 TODO : source) TODO : more examples. This might also carry prospects of generalizability of the model [26].

But for now, in certain retrieval tasks such as scientific journal tagging, the predictive power of each feature for each single label remains opaque. In that context, the problem is conveniently framed as multilabel binary prediction with varying amount of positive labels in the groundtruth. The reason for distanciating singling-out from entity-wide labels, is that it has been shown that as soon as

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

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

singling-out is possible, models that work on instances are more accurate TODO : sources.

To allow the use of existing diffentiable loss fonctions, previous research papers tend to reframe the problem into either (I) a unilabel top-1 for each feature (as described above, with the COCO dataset as an example of isolation of features [16]), (II) a top-1 prediction (III) a top-k prediction with fixed k (IV) a top-k prediction with varying k after training (V) redefine backpropagation for multilabel prediction [27] (VI) multitask learning [8]. This order reflects how close modelling is to the groundtruth, which remains multilabel with varying amounts of labels. TODO : group them

TODO : delta with hierarchical labels

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.

A dataset of multilabel images for visual representation learning was recently released along with its pretrained model. In contrast to most modles trained on image-net TODO : source, the implementation is trained on a multi-label prediction task from scratch [24].

immediately point out the delta

first identifying elements is the solution to the problem.

Sometimes it seems useful to optimize a neural network directly on the evaluation metric [9]. In the case of 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 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.

what is multilabel

RQ: does a contrastive method to deal with varying amount of labels help with predictions (measured by multilabel F1)

contrastive

what is the problem if we don't take into account the number of wrongs and rights. Find who suggested that this should be done.

top Kappa is more realistic in cases where ranking differs.

Similarily to the focal loss, sigmoidF1 loss deals with class imbalance (see [15]), robustness to outliers [15].

While these recent studies are focussing on top-k prediction with fixed k, or often top 1. Among different kinds of animals, identify the specie of this animal. Note that there can be several animals (or more generally, objects) on the picture, but the task is still top-1 prediction, couples with an object detection task. i.e. contrasting it from object detection (typically COCO dataset or text entity recognition (typically this dataset).

We propose a framework that explicitly deals with varying numbers of positive groundtruth labels per example.

it is not a tag problem, because it is user generated and limitless number of classes

1.1 our contribution

We propose a general mathematical formulation of 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.

2 BUILDING UP ON LOSSES

Multi-label learning can be divided into two major fields: *problem transformation* and *algorithm adaptation* [29]. 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 [15]) 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 [3, Section 4.3.4]. More generally, the *problem transformation* formulation amounts to minimizing the loss on a class of neural networks, such that

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

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:

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

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

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

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 [12, 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

Early representatives of *algorithm adaptation* stem from heterogenous domains of machine learning. Multi-Label k-Nearest Neighbors [28], Multi-Label Decision Tree [7], Ranking Support Vector Machine [10] and Backpropagation for Multi-Label Learning [27]. More recently, two papers introduced the idea of multitask learning for *label prediction* and *label count prediction* for text (ML_{NET}) [8] and image [14] 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.).

3.2 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 CNN from scratch with millions of images specifically for multilabel tasks [24]. Similarly to [14] mentioned in the previous paragraph, this task is loosely related to object detection.

3.3 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 [15], 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 [17], or per-class-thresholding after training [6] and eventually abstracting the threshold away via a *soft-F1* measure [4] TODO : say more about this method. In the latter two cases, the task is to predict genre from movie posters.

TODO : nicer plot on the right dataset

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

4 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.

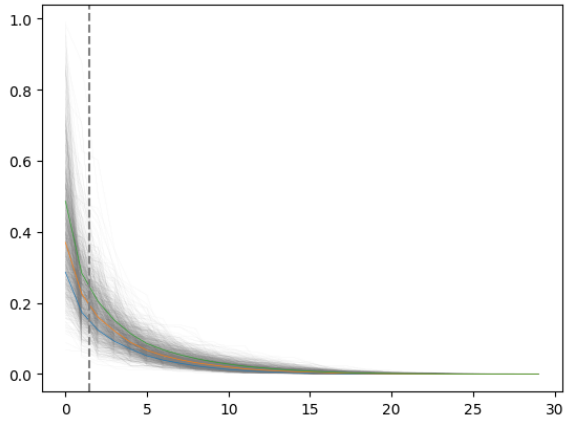


Figure 2: ordered per-label cross-entropy predictions for each example (each grey line) with the median (orange) and IQR (green & blue) over all examples. Determining a global threshold can be related to visually finding the "knee" in that median curve (dotted line)

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

TODO : check this formula

$$\mathbf{p}_i = \begin{cases} \hat{\mathbf{y}} & \text{if } y = 1 \\ 1 - \hat{\mathbf{y}} & \text{otherwise} \end{cases} \tag{4}$$

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:

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

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

We also define precision and recall

$$\begin{aligned} P &= \frac{tp}{tp + fp} \\ R &= \frac{tp}{tp + fn} = \frac{tp}{|Y^+|} \end{aligned} \tag{5}$$

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 [21].

DOUBT : maybe ignore F_β and only mention F_1

$$F_\beta = \left(1 + \beta^2\right) \frac{P \cdot R}{\beta^2 P + R} \tag{6}$$

Or equivalently:

$$\begin{aligned} F_\beta &= \left(1 + \beta^2\right) \frac{tp}{\left(1 + \beta^2\right) tp + \beta^2 fn + fp} \\ &= \left(1 + \beta^2\right) \frac{tp}{\beta^2 |Y + | + tp + fp} \end{aligned} \tag{7}$$

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.

4.1 soft F1 score

It is possible define a *soft F1* score [4] DOUBT : can we cite a Medium post? with smooth confusion matrix entries (i.e. tp, fp and fn are not natural numbers anymore):

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

$$\mathcal{L}_{\text{softF1}} = \frac{\tilde{tp}}{2\tilde{tp} + \tilde{fn} + \tilde{fp}} \tag{8}$$

softF1 is

4.2 sigmoidF1 score

We define *sigmoidF1*, inspired by the *Maximum F1-score criterion* for automatic mispronounciation detection [13]. Whereas A sigmoid function $S(u)$

$$S(u) = \frac{1}{1 + \exp(-\beta u)} \tag{9}$$

Confusion matrix entries then become

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

DOUBT : mention smooth hinge loss [20]

4.3 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 [17] 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 [SOURCE HERE] is similar but includes weighing to account for class imbalance, i.e. weighing each class by the number of groundtruth positives.

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

Accuracy is the overall fraction of correctly predicted labels [17]:

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

TODO compare to [25]

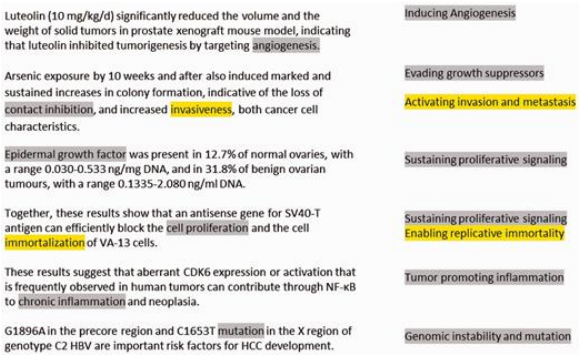
5 DATASETS

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

DOUBT : optional paragraph In light of the problem definition leading to the sigmoidF1 framework in the introduction and in order to clearly delimit the proposed method, following are a few datasets that are not suitable for the task.

Among the three datacets used for benchmarking ML-NET [8], a cancer hallmark dataset is of sub-entity unilabel nature [11]: the research clearly describe a process of annotating several expressions within paper abstracts. The remaining two, seem to fit to the entity wide multilabel definition and have a strong hierarchical nature.

Cacncer can be described according to its complexity with different principles, named hallmarks [11]. A corpus of 1580 PubMed abstracts are manually annotated for 10 hallmarks. This is a sub-entity labelling task and will therefore not be used here.



, a dataset for movie posters. Music genre

music genre, Arxiv publications, medical publications.

In order to test sigmoidF1 on different settings, image, sound and text

The datasets are namely a movie poster dataset, a toxic comments dataset and a medical publications dataset.

- Multilabel classification for text [22]
- Scenery dataset for images [1].
- movie Posters dataset:

https://www.kaggle.com/neha1703/movie-genre-from-its-poster

pre-scraped: https://www.kaggle.com/neha1703/movie-genre-from-its-poster/discussion/35485 (I removed all jpg's that are empty.)

TODO download posters myself, to see if I get more (see utils in here)

- https://www.kaggle.com/c/imaterialist-challenge-fashion-2018/data
- https://archive.ics.uci.edu/ml/datasets/DeliciousMIL\%3A+A+Data+Set+for+Multi-Label+Multi-Instance+Learning+with+Instance+Labels#

Not what we are looking for:

some datasets have spacially differing labels such as Amazon rainforest.

citing Kaggle datasets https://www.kaggle.com/data/46091

6 EXPERIMENTAL RESULTS

6.1 implementation

varying b in the sigmoid function as if it is an adaptive learning rate.

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.c. 256), in order for meaningful F1 surrogates to be calculated.

VanillaResnet

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**

TencentResnet

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**

DenseNet

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**

some TextNet

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**

7 FUTURE WORK

Apply the loss function to more sophisticated neural network architectures that use F1 score as an evaluation metric such as ACSUM-GAN [2].

This model can be adapted for hiarchical multilabel classification or active learning (for both see [19]).

Combine the proposed loss functions with representation learning [18, 23] or self-supervised learning, in order to model abstract relationships between the labels.

adapt to *extreme* multilabel prediction [5]

8 DRAWBACKS

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.

REFERENCES

[1] 2007. *Advances in Neural Information Processing Systems 19* (2007). https://doi.org/10.7551/mitpress/7503.003.0206

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

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

[4] 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

[5] 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

[6] 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

[7] 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

[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] 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

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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

[17] 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

[18] 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

[19] 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

[20] Jason DM Rennie. 2005. Smooth hinge classification. *Proceeding of Massachusetts Institute of Technology* (2005).
[21] C.J. van Rijsbergen. [n.d.]. *Information retrieval* (2nd [rev.] ed. ed.). Butterworths, London [etc.
[22] Nithum Thain, Lucas Dixon, and Ellery Wulczyn. 2017. Wikipedia Talk Labels: Toxicity. https://doi.org/10.6084/m9.figshare.4563973.v2
[23] 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
[24] 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
[25] 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]
[26] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. https://arxiv.org/abs/1611.03530
[27] 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
[28] 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
[29] 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