Deep Multi-Label Learning

by

Jan André Marais

Thesis presented in partial fulfilment of the requirements for the degree of Master of Commerce (Mathematical Statistics) in the Faculty of Economic and Management Sciences at Stellenbosch University

Supervisor: Dr. S. Bierman

December 2017

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: .....

Abstract

Deep Multi-Label Learning

J. A. Marais

Thesis: MCom (Mathematical Statistics)

December 2017

English abstract

Uittreksel

Diep Multi-Etiket Leer

("Deep Multi-Label Learning")

J. A. Marais

Tesis: MCom (Wiskundige Statistiek)

Desember 2017

Afrikaans abstract

Acknowledgements

I would like to express my sincere gratitude to the following people and organisations ...

Contents

List of Figures

List of Tables

Nomenclature

Constants

g =	$9.81\,\mathrm{m/s^2}$	

Variables

Re_D	Reynolds number (diameter)	[ ]
x	Coordinate	[m]
$\ddot{x}$	Acceleration	[$\mathrm{m/s^2}$]
θ	Rotation angle	[rad]
τ	Moment	[N·m]

Vectors and Tensors

$\overrightarrow{v}$	Physical vector, see equation ...

Subscripts

a	Adiabatic
a	Coordinate

Chapter 1

Introduction

1.1 Motivation

The motivation for this thesis is two-fold:

1. Image classification is a highly relevant topic in Computer Vision, Machine Learning and Statistical Learning. It is a thoroughly researched domain and already by many regarded as a 'solved' problem. This progress is mainly attributed to the yearly large-scale image classification competition, *ImageNet*[1], and the development of *Convolutional Neural Networks* (CNNs). The last five winners of ImageNet all used a variant of CNNs in their solution. However, the main focus up until recently was on problems of single label classification. Therefore, the field of multi-label image classification is nowhere near the maturity level of its single-label counterpart. Multi-label classification has a wide range of applications, not only in image classification. It has been applied to problems in text categorisation, multimedia, biology, chemical data analysis, social network mining and e-learning among others. This is most likely the reason why it has seen such a rapid increase of academic publications (see **??**). However, researchers have not yet reached consensus on how to deal with many of the aspects when learning from multi-labelled data, *e.g.* dependency between labels. There are a very limited number of publications specifically dealing with multi-label classification of images, even more so while using CNNs for this task. The field can gain from an up-to-date review of the literature, more statistical perspectives on some of the challenges, additional benchmark datasets and quality empirical evaluations of the theory.

2. Deforestation is a massive global problem. It contributes to reduced biodiversity, habitat loss, climate change and other devastating effects.

[1] http://www.image-net.org/

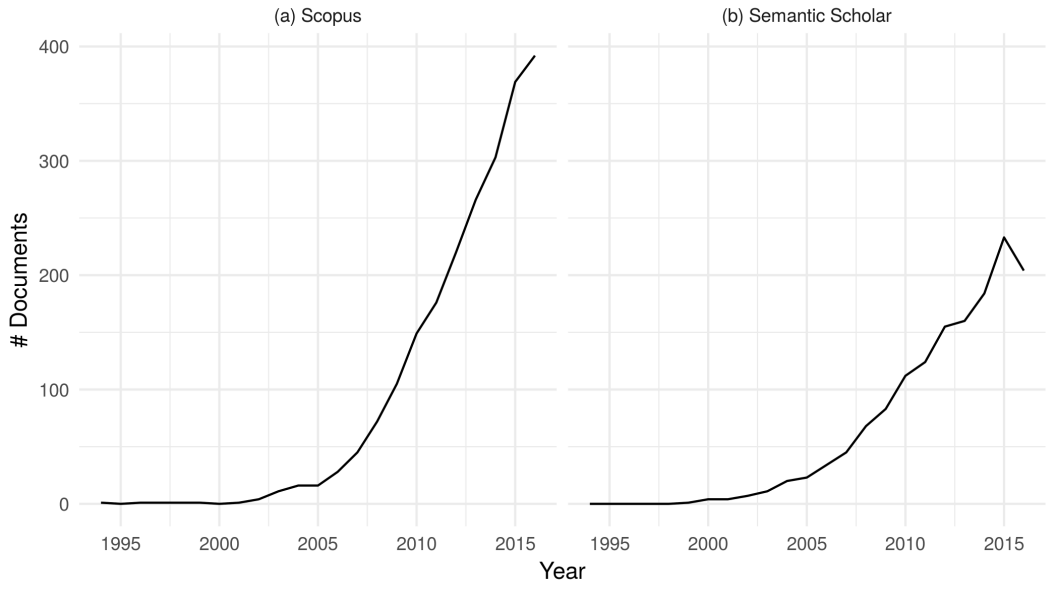

Figure 1.1: Line graphs illustrating the rise in multi-label learning publications per year for two databases. The database searches were done on 24-03-2017. The searches were not identical since they were limited to the search features of the databases. (a) The search on Scopus (cite) was for all documents (conference papers, articles, conference, articles in press, reviews, book chapters and books) in any subject area with either the words *multi-label* or *multilabel* and either the words *learning* or *classification* found in either their titles, abstracts or keywords. (b) The search on Semantic Scholar was based on machine learning principles and thus automatically decides which research documents are relevant to a specific search query. The query used was *multilabel multi-label learning classification*. The search only returns research in the computer science and neuroscience fields of study. More technical details can be found on the respective engine's websites.

It is said that the world loses an area of forest the size of 48 football fields per minute and the area most affected is in the Amazon basin (cite Kaggle). This problem can be fought more effectively by governments and local stakeholders if better data about the location of deforestation and human invasion on forests are continuously available to them - an ideal task for machine learning! Planet[2] and SCCON[3] constructed a dataset of labelled satellite images taken of the Amazon basin and released it as part of a competition on Kaggle[4], challenging competitors to build algorithms that can automatically label these images with atmospheric

[2]Designer and builder of the world's largest constellation of Earth-imaging satellites - www.planet.com

[3]Remote sensing experts - www.sccon.com.br/eng

[4]Runs programming contests to crowd source machine learning solutions - www.kaggle.com

conditions and various classes of land use/cover[5]. Resulting algorithms will help the global community better understand where, how, and why deforestation happens all over the world - and ultimately how to respond.

1.2 Thesis Objectives

This thesis works towards building a multi-label classifier that can label satellite images of the Amazon as accurately as possible. The method thought best to achieve this goal is to:

1. Identify the most important and latest developments in the literature for: multi-label classification, image classification and *remote sensing* (analysis of satellite images).
2. Provide an extensive review and discussion of these methods and how they compare to each other.
3. Empirically evaluate and compare them on the satellite image data in order to find the best strategies for our labeling task.

Since practically every state-of-the-art solution to an image classification problem is a CNN, it is reasonable to restrict the space of possible classifiers to CNNs. This thesis should provide the reader with a clear understanding of CNNs and how to effectively apply them to a multi-label classification problem, and especially in the domain of remote sensing. The main contribution of this thesis is a review of multi-label CNNs.

update this section as progress is made with thesis.

1.3 Data

This section covers an initial introduction to the data available for the problem at hand. The elements of the data important to know before moving on will be discussed here and the rest will be addressed throughout the thesis, as it becomes relevant to the discussion. This is done here to get a better understanding of the problem before exploring the literature.

1.3.1 Image Format

The data for this task comes from a set of images (also referred to as chips). Each chip is a small excerpt from a larger image of a specific scene in the Amazon taken by satellites. The chip size in pixels is 256×256, representing roughly 90 hectares of land, and is taken from a larger scene of 6600×2200

[5]Land cover indicates the physical land type such as forest or open water whereas land use documents how people are using the land.

pixels. All of the satellite images were taken between January 1, 2016 and February 1, 2017. The format of these images differ from the standard image format. Each image contains four spectral bands: red (R), green (G), blue (B) and near infrared (NIR), where the standard format images usually only contain R, G and B. The additional NIR colour channel is common in remote sensing[6] applications and supposedly allows for clear distinction between water and vegetation in satellite images, for example.

Another difference between these images and the usual format is that these have pixel intensities in 16-bit digital number format as opposed to the usual 8-bit of standard RGB images. This allows the colours in the images to have a much higher range since 16-bit pixel intensities have 65536 (2^{16}) levels, compared to 256 levels of 8-bit images. This becomes useful, for example, to distinguish between very dark or very bright areas in an image. If the pixel values of a chip gets flattened out into a vector, it will be of size 262144 ($256 \times 256 \times 4$). However, CNNs take the images in their array form as input.

1.3.2 Collection and Labelling of the Images

The image collection was created by first specifying a "wish list" of scenes containing the phenomena the creators wanted to be included, in addition to a rough estimate of the number of such scenes that are necessary for a sufficient representation in the final collection. This set of scenes was then searched for manually on Planet Explorer[7]. From these scenes the 4-band chips were created. A schematic of this process can be seen in **??**. The chips were labelled manually by crowd sourcing. The utmost care was taken to get a large and well-labelled dataset, but that does not mean the labels all correspond to the ground-truth, *i.e.* the data will contain some inherent error. The creators believe that the data has a reasonable high signal to noise ratio.

Note, the training and test splits were determined by the Kaggle competition creators. The training chips are labeled but at the time of writing this, the test chips are not yet made available to competitors. Predicted labels for the test chips can be submitted to Kaggle to evaluate in terms of the F_2-score, a metric which will be discussed in Chapter ??. This setup prevents competitors from using the test chips for training a classifier. There are 40479 training chips and 61191 test chips.

1.3.3 Class Labels

The class labels for the images can be divided into three groups: atmospheric conditions, common land cover/use phenomena and rare land cover/use phenom-

[6]The use of satellite- or aircraft-based sensor technologies to detect and classify objects on Earth [https://en.wikipedia.org/wiki/Remote_sensing].

[7]A web based interactive map of Earth consisting of satellite images, similar to Google Earth - www.planet.com/explorer

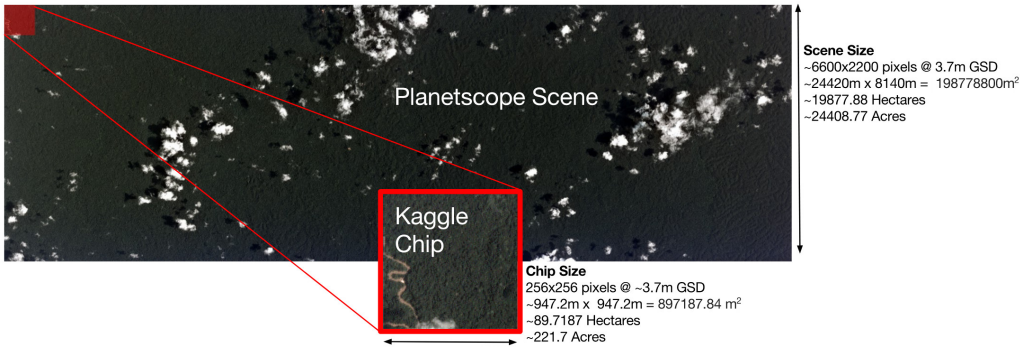

Figure 1.2: Schematic of the image collection process.

ena. In total there are 17 posssible labels. Each chip will have one atmospheric label and zero or more common and rare labels. Chips that are labeled as cloudy should have no other labels.

The atmospheric condition labels are: *clear*, *haze*, *partly cloudy* and *cloudy*. They are relevant to a chip when:

- **clear**: there are no evidence of clouds.
- **haze**: clouds are visible but they are not so opaque as to obscure the ground.
- **partly cloudy**: scenes show opaque cloud cover over any portion of the image but the land cover/use phenomena are still visible.
- **cloudy**: 90% of the image is obscured with opaque cloud cover.

Examples of chips with atmospheric labels can be found in **??**. Each chip should only have one atmospheric label and therefore this classifying task simplifies to a multiclass problem. This allows for the option to break up the labeling task of all the labels into two tasks: a multiclass classification problem for the atmospheric labels and a multi-label classification problem for the land cover/use labels. This approach might save some computational time and give extra information to the multi-label learners for classifying the land cover/use labels. We will experiment with these approaches in Chapter ??.

The common land cover/use labels are: *primary*, *agriculture*, *water*, *habitation*, *road*, *cultivation* and *bare ground*. They are relevant to a chip when:

- **primary**: it is primarily consisting of rain forest (virgin forest), *i.e.* dense tree cover.
- **agriculture**: it contains any land cleared of trees that is being used for agriculture or range land.
- **water**: it contains any one of the following: rivers, reservoirs, or oxbow lakes.
- **habitation**: it contains human homes or buildings.

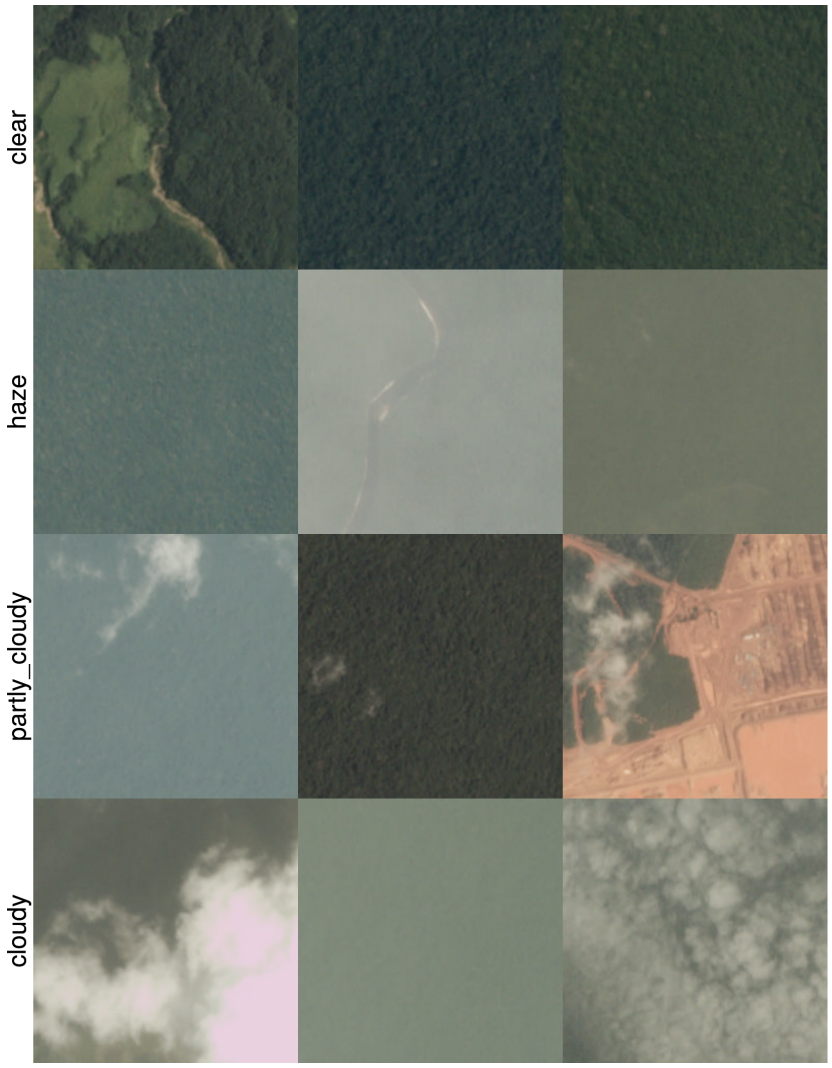

Figure 1.3: Examples of chips with atmospheric labels. These (along with all the other chips plotted throughout the thesis) are the JPEG conversions of the original 4-band, 16-bit images.

- **road**: it contains any type of road.
- **cultivation**: it shows signs of smaller-scale/informally cleared land for farming.
- **bare ground**: it contains naturally (not the caused by humans) occurring tree-free areas.

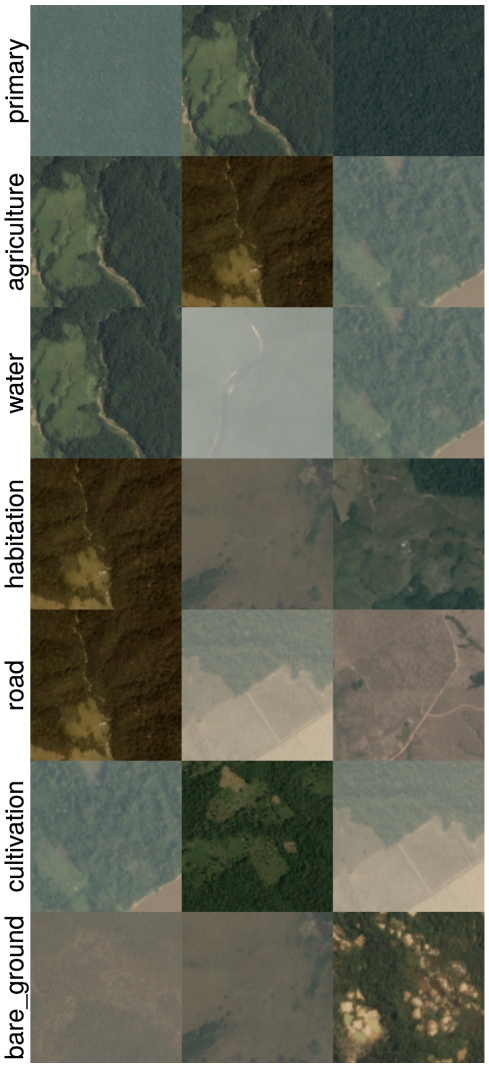

Figure 1.4: Examples of chips with common land cover/use labels.

Examples of chips with common land cover/use labels are found in **??**. According to the competition page on Kaggle, small, single-dwelling habitations are often difficult to spot but usually appear as clumps of a few pixels that are bright white. Roads sometimes look very similar to rivers and therefore these two labels might be noisy. The NIR band might give a classifier additional information to help distinguish between the two. Cultivation is a subset of

agriculture and is normally found near smaller villages, along major rivers or at the outskirts of agricultural areas. It typically covers very small areas.

The less common land cover/use labels are: *slash and burn*, *selective logging*, *blooming*, *conventional mine*, *artisinal mine* and *blow down*. Chips are tagged with these labels when:

- **slash and burn**: there are signs of the farming method that involves the cutting and burning of the forest to create a field. These look like cultivation patches with black or dark brown areas.
- **selective logging**: winding dirt roads are present adjacent to bare brown patches in otherwise primary rain forest. Selective logging is the practice of selectively removing high values tree species from the rainforest.
- **blooming**: there are signs of trees flowering. Blooming is a natural phenomena where particular species of flowering trees bloom, fruit and flower at the same time. These trees are quite big and the phenomena can be seen in the chips. They usually appear as white dots.
- **conventional mine**: it contains signs of large-scale legal mining operations.
- **artisinal mine**: it contains signs of small-scale (sometimes illegal) mining operations.
- **blow down**: there are signs of trees uprooted or broken by wind. High speed winds (~160km/h) in the Amazon are generated when the cold dry air from the Andes settles on top of the warm moist air in the rainforest and then sinks down with incredible force, toppling larger rainforest trees. These open areas are visible from space.

Examples of chips with these less common land cover/use labels are given in **??**. These labels are more challenging to identify in the chips and since they also appear less frequently, it might be difficult for the classifier to learn these labels. The imbalance in the class distribution is apparent in **??**.

1.4 Code and Reproducibility

All of the code for this project, including the source docoments, is made available at https://github.com/jandremarais/Thesis. The data is hosted on Kaggle at https://www.kaggle.com/c/planet-understanding-the-amazon-from-space/data. More instructions on how to implement the code is contained in the file named, `README.md`, in the GitHub repository.

1.5 Important Concepts and Terminology

Not suprisingly, a convolutional neural network is a type of neural network. Neural networks will be discussed in **??**, but there are some preliminary concepts

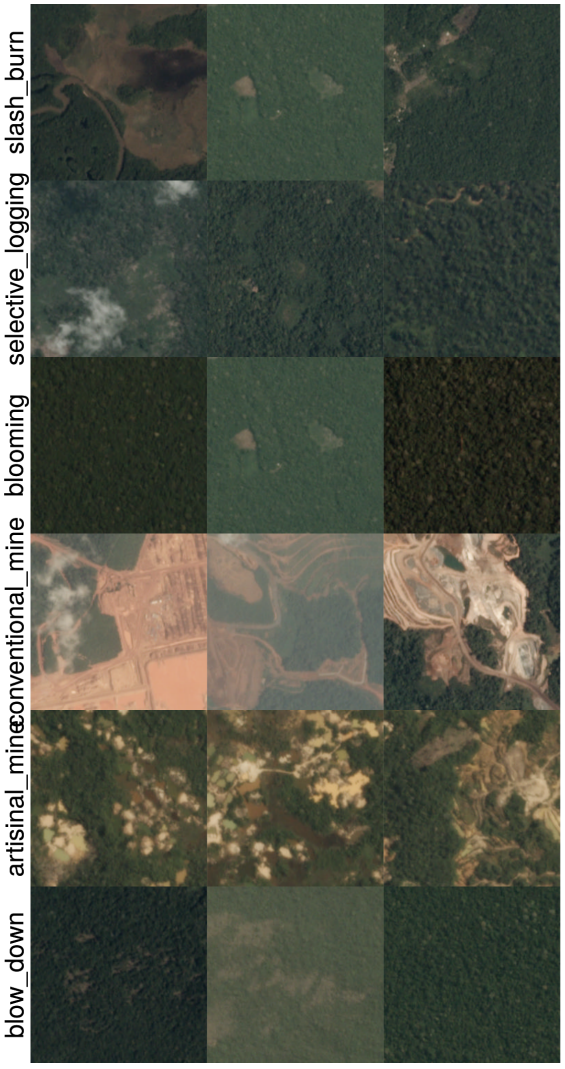

Figure 1.5: Examples of chips with less common land cover/use labels.

to introduce here to ensure a better understanding of neural networks and ultimately CNNs. First, a brief introduction to the general problem of image classification is given.

1.5.1 Image Classification

There are three main tasks in computer vision (CV), namely: image classification, object detection and image segmentation. Traditional image classification is the task of assigning one label from a fixed set of categories to an input image. More recently the task has been generalised to assigning multiple labels to an input image, *i.e.* multi-label classification (MLC). First, we will only look at the single label case.

Image classification is the core of computer vision tasks and probably the

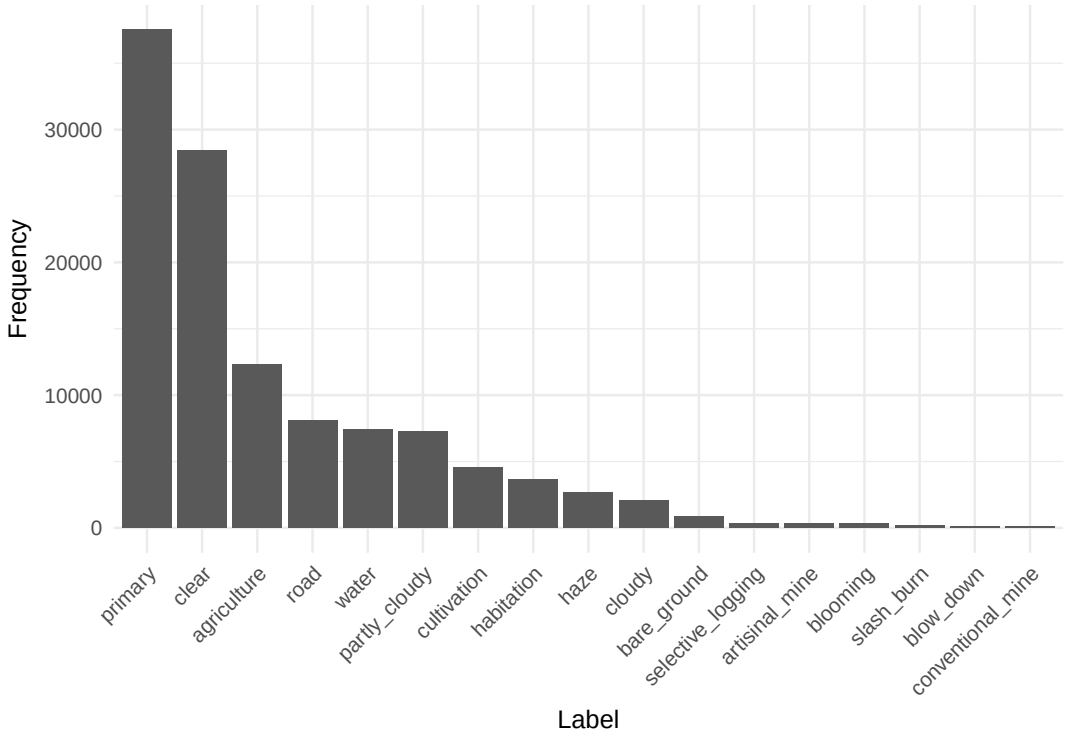

Figure 1.6: Class distribution of the labels in the training set.

most explored since it has a large variety of practical applications. It can be shown that the other two CV tasks, detection and segmentation, can be reduced to classification. Classification will be the main theme of this thesis but we will have a look at segmentation and detection later on.

> show visual difference between classification, segmentation and detection.

Instead of hard coding rules on how to classify images into an image classification model, it can learn to classify images by seeing many examples of images and its corresponding labels. In this way it learns the visual appearance of each class. This is sometimes referred to as a data-driven approach. A very intuitive approach to image classification (and supervised learning in general) is called the nearest neighbour approach.

Although this approach is rarely used in practice, this description helps with the understanding of the image classification problem. The nearest neighbour classifier will take a test image, compare it to every single one of the training images, and predict its label to be the label of the closest training image. This leaves the question of how to measure the similarity between images.

An image is a grid of many small, square cells of different colors. These cells are known as pixels and one pixel represents one color. A grayscale image, 32 pixels wide and 32 pixels long, can be represented by a 32×32 matrix of

integers, where each integer represents the 'brightness' (intensity) of each pixel. These integers are usually in $[0, 255]$, such that the greater the integer the brighter the pixel, *i.e.* a pixel with intensity 0 is totally black and a pixel with intensity 255 is totally white. Note that a color image consists of 3 spectral bands, red, green and blue (RGB), *i.e.* the color of one pixel is determined by 3 integers each representing the intensity of the color red, green and blue, respectively.

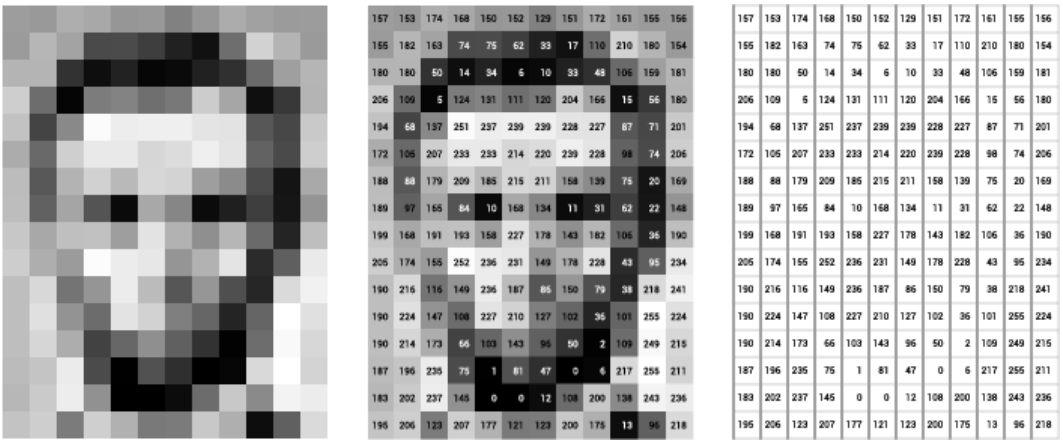

Figure 1.7: Greyscale intensities. http://ai.stanford.edu/~syyeung/cvweb/tutorial1.html

The (dis)similarity between two images can now be measured pixel by pixel. It is possible to represent the grayscale image mentioned above in a vector of length 32×32. Suppose a grayscale Image 1 is flattened out to be represented by the vector $\boldsymbol{I}_1 = \{I_{11}, I_{12}, \ldots, I_{1p}\}$ and similary, Image 2 by $\boldsymbol{I}_2$, where $p = 32 \times 32$. Then the dissimilarity between Image 1 and Image 2 can be calculated by the L_1-distance:

$$d_1(\boldsymbol{I}_1, \boldsymbol{I}_2) = \sum_{j=1}^{p} |I_{1j} - I_{2j}|.$$

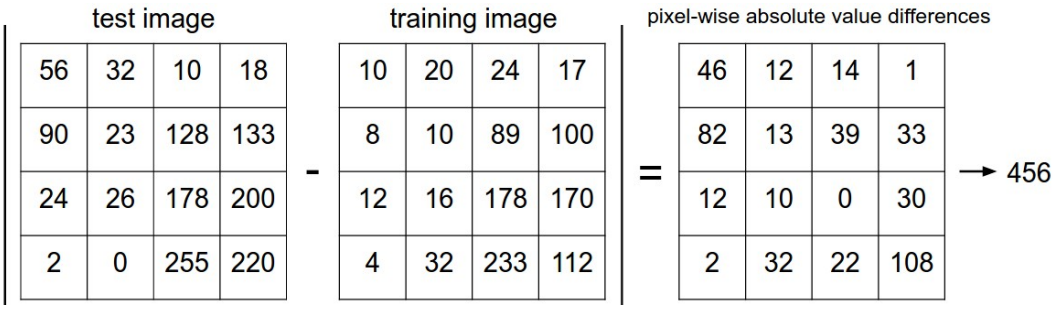

Figure 1.8: Pixelwise difference

Now, suppose we want to predict the label of an test image a, then the nearest neighbour approach would assign the label of train image b^* to test image a if:

$$b^* = \arg\min_b d_1(\boldsymbol{I}_a, \boldsymbol{I}_b),$$

for $b = 1, 2, /dots, N$, where N is the number of training images. Of course there are other ways of measuring the dissimilarity between images. Another example would be to use the L_2-distance:

$$d_2(\boldsymbol{I}_1, \boldsymbol{I}_2) = \sqrt{\sum_{j=1}^{p}(I_{1j} - I_{2j})^2}.$$

The chosen metric depends on the use case.

The nearest neighbour approach can be generalised to use more than 1 nearest neighbour when predicting the label of a test image. This approach is called the k-Nearest Neighbours (k-NN). The only difference is that, you now search for the k (instead of just 1) images with the smallest dissimilarity with the test image and then combine the labels of these k images, either through averaging or majority voting, to predict the label of the test image. Choosing the right value of k is important and is usually done by cross-validation. See Hastie ref.

The advantage of using k-NN is that it is simple and requires no time to train. Unfortunately, when it comes to test time, the algorithm needs to calculcate the distance between the test image and all the other images in the training set, which is computationally very expensive. Also in [Haste ref], they show that k-NN suffers severely from the *curse of dimensionality* and that it is mostly only useful to classify lower dimensional objects. Images are very high-dimensional objects.

The dissimilarity measures discussed above are actually proven to be very poor in discriminating between images in an image classification problem. Images that are nearby in terms of the L_1 and L_2 distances are much more of a function of the general color distribution of the images, or the type of background rather than their semantic identity. Refer to the t-SNE figure.

1.5.2 Score Function

The following simple approach to image classification naturally extends to neural networks and convolutional neural networks and is therefore very important to comprehend. This approach has to major components: a score function and a loss function. The score function maps raw data (*e.g.* an image) to a set of class scores, and a loss function quantifies the agreement between the predicted class scores and the actual ground truth labels associated with the raw data. This approach can then be described as an optimization problem in which the

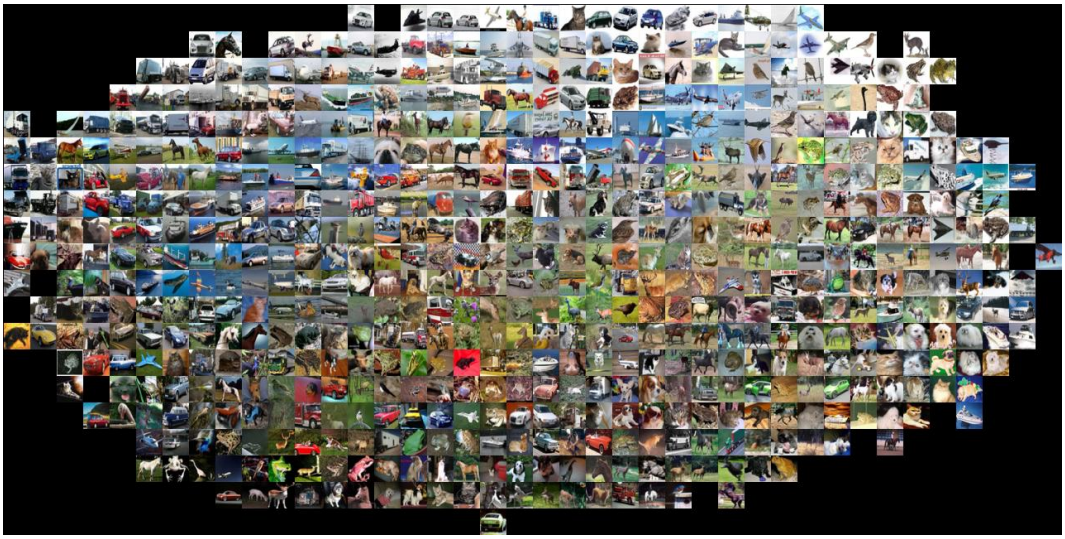

Figure 1.9: Caption

minimisation of the loss function with respect to the parameters of the score function is the main goal.

Some notation is needed to formally define this approach. Suppose we have N training images $\boldsymbol{x}_i \in \mathbb{R}^p$ each associated with a label $y_i \in \{1, 2, \ldots, K\}$, where $i = 1, 2, \ldots, N$ and K is the number of possible categories an image can belong to and p the number of pixels of each image. The score function is then defined as the function f that maps the raw image pixels to class scores:

$$f : \mathbb{R}^p \to \mathbb{R}^K.$$

The simplest possible score function is a linear mapping:

$$f(\boldsymbol{x}_i, W, b) = W\boldsymbol{x}_i + \boldsymbol{b}.$$

In the above equation, Image i is flattened out to be represented by a p-dimensional vector. The paramters of f are the matrix $W : K \times p$ and the vector $\boldsymbol{b}$, often called the weights and biases, respectively. These terms are comparable to the coefficient and constant terms in a statistical linear model and thus should not be confused with bias in the statistical sense.

We assume the pairs $(\boldsymbol{x}_i, y_i)$ to be fixed, but we do have control over the W and $\boldsymbol{b}$ terms. Our goal will be to set these in such a way so that the computed class scores for each image in the training set match the associated ground truth label as close as possible. What we have described thus far is very similar to the approach taken by convolutional neural networks, but instead the function, f, which maps the raw pixels to class scores, is much more complicated with plenty more parameters to tune.

Notice that this score function determines the score for each class as a weighted sum of the pixel values across all 3 of its spectral bands. We would

imagine that a linear classifier trained to classify, say, ships would have a weight matrix that assigns heavier weights to blue pixels on the sides of an image, which loosely corresponds to water.

If we picture the images as points in a high-dimensional space, f is a hyperplane, W determines the angle of the hyperplane and $\boldsymbol{b}$ translates the hyperplane through the space. Another interpretation of this linear classifier is that each row of the weight matrix is a so-called template for the corresponding class. The linear classifier matches the input image with each of the class templates in W by calculating a dot product. A high class score would translate to a higher similarity between the input image and the class template. This interpretation is closely related to the nearest neighbour approach, but here only the test image's distance (here the negative of the inner product) to each of the K class templates are calculated instead of its distance to each of the N images in the training set.

Later on it becomes too cumbersome to keep track of two sets of parameters, W and $\boldsymbol{b}$, and therefore, for the rest of the thesis we will write the linear classifier as:

$$f(\boldsymbol{x}_i, W) = W\boldsymbol{x}_i,$$

where $\boldsymbol{b}$ is now contained in the last(/first?) column of W and the last element of $\boldsymbol{x}_i$ is now the constant, 1. This is the so-called bias trick.

Note that thus far we have used raw pixel values in the range of $[0, 255]$ as input. However, in practice, it is more common to subject the input images to some preprocessing before inputting them into the score function. The benefits of this will be made clear in the optimisation section. Common preprocessing techniques are the centering and scaling of the pixels so that their values lie in the range of $[-1, 1]$. To center the input image, is to calculate a *mean image* from the training images and subtract each of its pixel values to from the corresponding pixel values of each image in the training set. This is identical to zero mean centering for standard statistical learning tasks - each pixel is seen as in input feature. Scaling is done by dividing each pixel by a function of its variance accross the whole training set.

1.5.3 Loss Function

To evaluate the agreement between the score function and the ground truth labels, we need a loss function. A loss function, also known as the cost function or the objective, is high when the score function does a poor job of mapping the input images to the class scores, and low when it does so accurately. There are multiple ways of defining such a loss function.

1.5.3.1 Multiclass Support Vector Machine Loss

A commonly used loss is the Multiclass Support Vector Machine (SVM) loss. In statistical learning this is more commonly known as the Hinge Loss. The SVM loss is designed in such a way that it wants the correct class for each image to have a score higher than the incorrect classes by some fixed margin Δ. More precisely, the multiclass SVM loss for the i-th example with label y_i can be given by:

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + \Delta),$$

where $s_j = f(x_i, W)_j$ is the score for the j-th class computed for image i. Here L_i consists of $K - 1$ components, each representing an incorrect class. A component will make no contribution to the loss if the calculated class score for the corresponding incorrect class is less than the correct class score by a margin of Δ, *i.e.* $s_{y_i} - s_j > \Delta$. It will make a positive contribution otherwise. As an example, suppose we have three predicted class scores for an image $s = [4, 5, -3]$ and that the second class is the true label. Let $\Delta = 2$. The loss computed for this image will then consist of 2 components:

$$\begin{aligned} L_i &= \max(0, 4 - 5 + 2) + \max(0, -3 - 5 + 2) \\ &= 1 + 0 \end{aligned}$$

We see that although the predicted class score for class 1 was smaller that the predicted class score for the true label, class 2, it was still within a margin of $\Delta = 2$ and there had a positive contribution to the loss. The predicted class score for class 3 was far lower than predicted class score for the true label and therefore did not make any contibution to the loss. In summary, the SVM loss function wants the score of the correct class to be larger than the incorrect class scores by at least Δ, if not, we will accumulate a loss.

Note that the loss is typically evaluated on a set of images and not just one, as we have describe thus far. The average loss of a set with N images can be written as $L = \frac{1}{N} \sum_{i=1}^{N} L_i$. Another variation of the SVM loss is to replace the $\max(0, \cdot)$ term with the term, $\max(0, \cdot)^2$, which results in the squared hinge loss or the L_2-SVM loss. This penalises violated margins more heavily and may work better in some cases. [https://arxiv.org/abs/1306.0239]

There is still one problem with the SVM loss described thus far. Suppose we have found a weight matrix W that correctly classifies all input images and by the correct margins, *i.e.* $L_i = 0$, $\forall i$, then setting the weight matrix to λW, for $\lambda > 1$ will have the same solution. This means the solution to the optimisation problem is not unique. It would make the optimisation task easier if we could remove this ambiguity. This can be done by adding a penalty term to the loss function, also know as regularisation. The mosty common regularisation penalty, $R(W)$, is the L_2-norm:

$$R(W) = \sum_k \sum_l W_{k,l}^2,$$

which is simply the sum of the squared elements of the weight matrix. The full SVM loss can now be defined as:

$$L = \frac{1}{N} \sum_i L_i + \lambda R(W).$$

The two components of the loss can be called the *data loss* and the *regularisation loss*. λ determines how much regularisation should be done. If λ is large, more regularisation will take place. The value of λ is typically determined through cross-validation.

The regularisation penalty ensures a unique (or less solutions?) solution to the optimisation problem by restricting the weight parameters in size. Greater weight parameters will result in bigger loss, if everything else remain constant. Another appealing property is that penalising large weights tends to improve generalisation, because it means that no input dimension can have a very large inlfuence on the scores all by itself.

Typically, only the weight parameters are regularised, since the bias terms do not control the strength of influence of an input dimension. Howerver, in practice the often turns out to have a negligible effect.

To return to the value of Δ - it turns out that Δ and λ control the same trade-off and therefore we can safely set $\Delta = 1$ and only use cross-validation for determining λ. This might not seem obvious, but the key to understanding this is to realise that the weights in W have a direct influence on the class scores and therefore also on the differences between them. If all the elements in W are shrinked, all the differences in class scores will shrink and if all the elements are scaled up, the opposite will happen. Therefore, the margin Δ becomes meaningless in the sense that the weights can shrink or stretch to match Δ. Thus the only real trade-off is how large we allow the weights to be and this we specify through λ.

1.5.3.2 Softmax Classifier

The linear classifier combined with the SVM loss we call the SVM classifier. We will now look at the Softmax Classifier, which is the linear classifier combined with a different loss function. In statistics, the softmax classifier is better known as the multiclass logistic regressor. The biggest difference between the SVM classifier and the softmax classifier is that the latter gives a slightly more intuitive ouput in the form of normalised class probabilities, instead of the uncalibrated and less interpretable output of the SVM classifier. The loss function used for the softmax classifier is the *cross-entropy loss*:

$$\begin{aligned} L_i &= -\log \frac{e^{f_{y_i}}}{\sum_j e^{f_j}} \\ &= -f_{y_i} + \log \sum_j e^{f_j}. \end{aligned}$$

As before, the full loss is the mean of L_i over the whole dataset with an additional regularisation penalty.

To see where this loss function comes from, first consider the softmax function:

$$h_j(\boldsymbol{z}) = \frac{e^{z_j}}{\sum_k e^{z_k}}.$$

$h_j(\boldsymbol{z})$ squeezes the elements of the real-valued vector, $\boldsymbol{z}$, to fit in the range of $[0, 1]$ and that their sum always add to 1. Now, in information theory, the cross-entropy between a 'true' distribution p and an estimated distribution q is defined as:

$$H(p, q) = -\sum_x p(x) \log q(x).$$

Consider the case where the 'true' distribution, p, is a vector of zeros except at the y_i-th position, where the value is 1, and the estimated distribution, q, is the estimated class probabilities, $q = \frac{e^{f_{y_i}}}{\sum_j e^{f_j}}$. Clearly, $H(p, q)$ then simplifies to L_i. Thus the softmax classifier minimises the cross-entropy between the estimate class probabilities and the true distribution.

In the probabilistic interpretation of this classifier, we are minimising the negative log likelihood of the correct class, which can be interpreted as performing *maximum likelihood estimation* (MLE). From this view, the term $R(W)$ can be interpreted as coming from a Gaussian prior over the weight matrix, W, where instead of MLE we are performing *maximum a posteriori*.

To be clear, the softmax classifier interprets the scores computed by f to be the unnormalised log probabilities. Therefore, it undergoes the exponentiating and division (to become the normalized probabilities) before being used as input the cross-entropy loss.

Note that although we used the term 'probabilities' to describe the output the softmax classifier, these are not probabilities in the statistical sense. They do sum to 1 and are in the range of $[0, 1]$, but they are still technically confidence scores rather than probabilities, *i.e.* their order is interpretable but not their absolute values. The reason for this is that they depend heavily on the regularisation strength determined by λ. The higher λ is, the more uniform the probabilities become.

SVM and Softmax comparable. See http://cs231n.github.io/linear-classify/

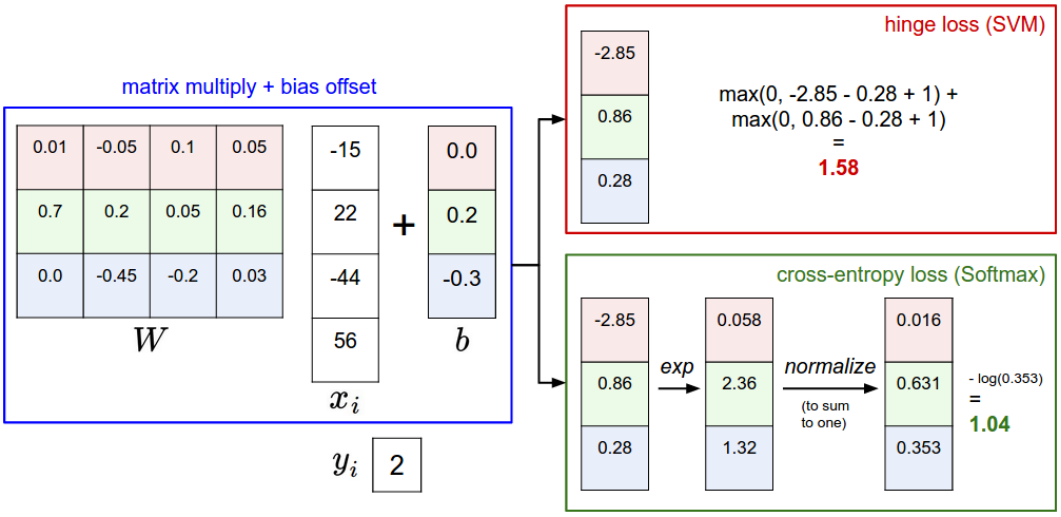

Figure 1.10: Figure to help with interp. Explain.

remember, only single label multiclass classification has been considered thus far and that some of these do not hold for multilabel classification.

1.5.4 Optimisation

From the previous sections we learned that the key components for the image classification task is the score function and the loss function. We looked at the linear mapping of raw pixel values to class scores and various loss functions, such as the hinge loss and cross-entropy loss, to evaluate the mapping against the ground truth labels. Putting all of this together, the SVM classifier can be reduced to the problem of minimising the loss:

$$L = \frac{1}{N} \sum_{i} \sum_{j \neq y_i} [\max(0, f(\boldsymbol{x}_i; W)_j - f(\boldsymbol{x}_i; W)_{y_i} + 1)] + \alpha R(W),$$

where $f(\boldsymbol{x}_i; W) = W\boldsymbol{x}_i$. This process of minimising the loss is also known as optimisation, which is the third key component. Optimisation is the process of finding the set of parameters W that minimise the loss function.

Once we get to convolutional neural networks, the only major difference is the use of a more complicated score function. The loss and optimisation components remain mostly unchanged.

Visualise a loss function in 2-dimensions to give idea of how it looks. [http://cs231n.github.io/optimization-1/]

SVM classifier has a convex loss function. Whole research field in convex optimisation. When we get to more complex neural netorks, the loss becomes non-convex.

> The loss functions we use are technnically non-differentiable, since there are 'kinks' in the loss function (gradients not define everywhere). However, the subgradient still exists and is commonly used instead. [https://en.wikipedia.org/wiki/Subderivative]

For this discussion on how to minimise the loss function with respect to W, we will use the SVM loss. The methods discussed may seem odd, since it is a convex optimisation problem. We only use this example for simplicity, since when we get to complex neural networks, the optimisation will not be a convex problem.

The core idea of this approach to minimise the loss with respect to W is that of iterative refinement - start with a random W and then iteratively refining it to get a lower loss. Finding the best set of weights, W is hard, but the problem of refining a specific set of weights to only be slightly better, is much easier.

A helpful analogy is that of the blindfolded hiker, who is on a hilly terrain, trying to reach the bottom. The height of the terrain represents the loss achieved. A possible strategy for the hiker to reach the bottom would be to test a step into a random direction and the only take the step if it leads downhill. In optimisation terms, we can start with a random initialisation of W, generate random perturbations δW to it and if the loss ath the perturbed $W + \delta W$ is lower, we will perform an update. This approach is better than a random search of W but still inefficient and computationally expensive.

It turns out that it is actually not necessary to randomly search for a good direction to move towards. The best direction can be determined mathematically. This best direction along which the weights should change corresponds to the direction of steepest descend and is related to the gradient of the loss function. In the hiking analogy, this approach roughly corresponds to feeling the slope of the hill below our feet and stepping down the direction that feels the steepest.

In one-dimensional functions, the slope is the instantaneous rate of change of the function at any specified point. The gradient is a generalisation of slope for multi-dimensional functions and is simply a vector of slopes, better known as derivatives, for each dimension in the search space. Mathematically, the expression for the derivative of a 1-dimensional function with respect to its input is:

$$\frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

When the function of interest take a vector of numbers instead of a single number, we call the derivatives partial derivatives. The gradient is simply the vector of partial derivatives in each dimension.

There are two approaches to computing the gradient: the **numerical gradient** and the **analytic gradient**. Their pro's and con's are discussed in the following section.

1.5.4.1 Computing the Gradient Numerically

Iterate over all dimensions one by one, make a small change, h, along that dimension and calculate the partial derivative of the loss function along that dimension by seeing how much the function changed. Ideally, we want h to be as small as possible, since the mathematical formulation requires $h \to 0$. In practice it often works better to compute the numeric gradient using the centered difference formula: $\frac{f(x+h)-f(x-h)}{2h}$.

Note that the update of W should be made in the negative direction of the gradient, since we wish to decrease the loss function.

The gradient tells us the direction in which the function has the steepest rate of increase, but it does not tell us how far along this direction we should step, *i.e.* what is the value of the step size? This value is also known as the *learning rate* and we will soon learn that it is one of the most important hyperparameters of a neural network. Choosing a small step size in the direction of steepest descent will ensure consistent but slow progress. A large step in this direction may lead to a quicker descent but also has the risk of overshooting the optimal point.

The obvious downfall of this approach (in addition to that is only an approximation) is that we need to calculate the gradient in each direction/dimension. Neural networks have millions of parameters and therefore optimising them in this manner is clearly not feasible.

1.5.4.2 Computing the Gradient Analytically

The second way to compute the gradient is analytically using Calculus. A direct formula for the gradient can be derived and it also very fast to compute. This approach is more error prone to implement which is why in practice it is very common to perform a *gradient check*, which is the comparision of the analytic gradient to the numeric gradient to chech the correctness of the implementation.

By using the SVM loss for a single data point as an example:

$$L_i = \sum_{j \neq y_i} \left[\max(0, \boldsymbol{w}_j^T \boldsymbol{x}_i - \boldsymbol{w}_{y_i}^T \boldsymbol{x}_i + \Delta) \right].$$

Now, we want to differentiate the function with respect to the weights. Taking the gradient *w.r.t.* $\boldsymbol{w}_{y_i}$, gives:

$$\nabla_{\boldsymbol{w}_{y_i}} L_i = -\left(\sum_{j \neq y_i} \mathbb{I}(\boldsymbol{w}_j^T \boldsymbol{x}_i - \boldsymbol{w}_{y_i}^T \boldsymbol{x}_i + \Delta > 0) \right) \boldsymbol{x}_i,$$

where $\mathbb{I}$ is the indicator function. This is simply the data vector scaled by the negative of the number of classes scores that did not meet the desired margin. The gradient with respect to the other rows of W where $j \neq y_i$ is:

$$\nabla_{\boldsymbol{w}_j} L_i = \mathbb{I}(\boldsymbol{w}_j^T \boldsymbol{x}_i - \boldsymbol{w}_{y_i}^T \boldsymbol{x}_i + \Delta > 0)\boldsymbol{x}_i.$$

Determining these equations are the tricky part. Once this is done, it is easy to implement the expressions and use them to perform gradient updates.

1.5.4.3 Gradient Descent

The procedure of repeatedly evaluating the gradient and then performing a parameter update is called *gradient descent*. This is by far the most common and established way of optimising neural network loss functions. Although there are some 'bells and whistles' to add to this algorithm, the core ideas remains the same when optimising neural networks.

One of the advantages of gradient descent is that a weight update can be made by only evaluating the gradient over a subset of the data, called *mini-batch gradient descent*. This is extremely helpful for large-scale applications, which are almost the norm for Deep Learning, since it is not necessary to compute the full loss function over the entire dataset. This leads to faster convergence and allows for the processing of large datasets that are too big to fit into a computer's memory. A typical batch consists of 64/128/256 data points, but it depends on the computational power at hand. The gradient computed using a mini-batch is only an approximation of the gradient of the full loss. This seems to be sufficient in practice since the data points/images are correlated.

The specification of the mini-batch size is not very important and is usually determined based on memory constraints. Usually they are in powers of two, because in practice many vectorised operation implementations work faster when ther inputs are sized in powers of 2. The extreme case of mini-batch gradient descent is when the batch size is selected to be 1. This is called *Stochastic Gradient Descent* (SGD). Recently, this is much less common, since it is more efficient to calculate the gradient in larger batches compared to only using one example. However, it is still widely acceptable to use the term SGD even though you are referring mini-batch gradient descent. This is actually the norm.

1.5.4.4 Backpropagation

Way of computing gradients of expressions through recursive application of the chain rule. Critical to understanding the optimisation of neural networks.

The core problem for this section is: We are given some function $f(\boldsymbol{x})$, where $\boldsymbol{x}$ is a vector of inputs, and we are interested in computing the gradient of f at $\boldsymbol{x}$, *i.e.* $\nabla f(\boldsymbol{x})$. In our case, f corresponds to the loss function (*e.g.*

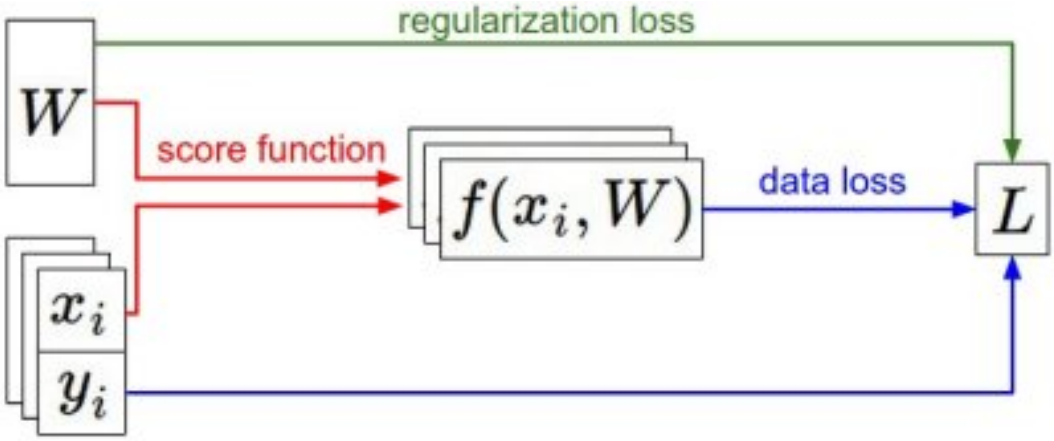

Figure 1.11: Good visual summary of data flow.

SVM loss) and the inputs $\boldsymbol{x}$ will consists of the training data and the neural network weights.

Consider this simple example to introduce some of the conventions. Suppose we have the following function $f(x, y) = xy$. The partial derivative for either input is then:

$$\frac{\partial f}{\partial x} = y,$$
$$\frac{\partial f}{\partial y} = x$$

These indicate the rate of change of f with respect to x and y respectively surrounding an infinitelly samll region near a particular point. For example, if $x = 2$ and $y = -5$, then $f(x, y) = -10$. The derivative on x is -5, which tells us that if we were to increase the value of x by a tiny amount, the effect on the whole expression would be to decrease by 5 times that amount.

As used before, the vector of partial derivatives is called the gradiemt, ∇f. So for the previous simple example we have $\nabla f = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right] = [y, x]$. What follows are another two simple examples that will prove to be useful in later discussions.

For $f(x, y) = x + y$, $\nabla f = [1, 1]$, and if $f(x, y) = \max(x, y)$, then $\nabla f = [\mathbb{I}(x \geq y), \mathbb{I}(y \geq x)]$. Technically, *nablaf* for the latter function is called a subgradient, since the derivative for $\max(x, y)$ is not defined everywhere (?).

1.5.4.5 Compound Expressions with the Chain Rule

Now for the calculating of a more complicated expression, we will use the chain rule. Consider the expression $f(x, y, z) = (x + y)z$. Note that this expression can be decomposed into two expressions: $q = x + y$ and $f = qz$. From the previous simpler examples, we saw how to calculalte the gradient for these

simple expression of addition and multiplication separately. But what we are really interested in is how to calculate the gradient of f *w.r.t.* its inputs, x, y, z. This can be done using the *chain rule*. According to the chain rule, $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q}\frac{\partial q}{\partial x}$, and similarly for $\frac{\partial f}{\partial y}$ and $\frac{\partial f}{\partial z}$. This can be viewed as the simplest form of backpropogation.

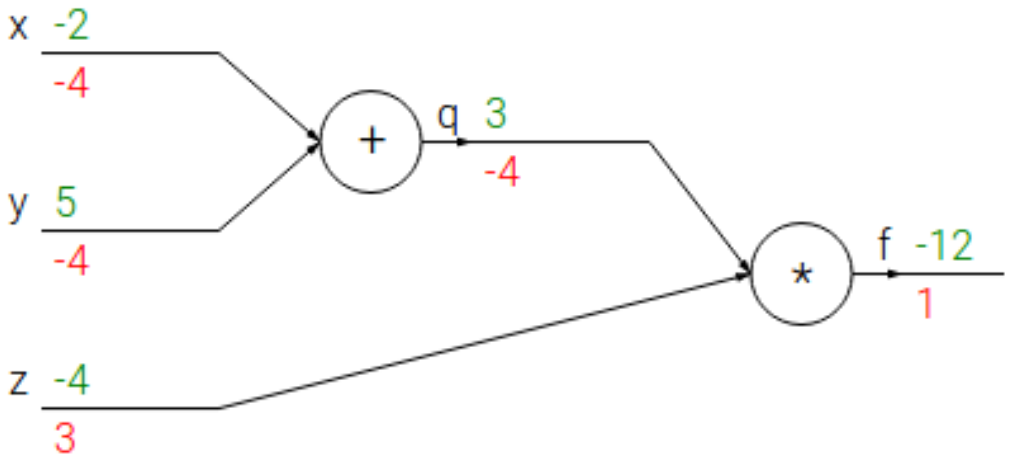

Figure 1.12: Simple circuit diagram to visualise backpropogation.

Suppose we want to compute the gradient at inputs $x = -2$, $y = 5$ and $z = -4$. First, we make a *foreward pass* to compute the outputs from the given inputs, *i.e.* $q = 3$ and then $f = -12$. These values are shown in green in the circuit diagram. The following step is to make a *backward pass* (backpropogation), which is to start at the end and recursively apply the chain rule to copmute the gradients, shown in red in the circuit diagram, all the way to the inputs of the circuit. In the example, $\frac{\partial f}{\partial f} = 1$, $\frac{\partial f}{\partial z} = 3$, $\frac{\partial f}{\partial q} = -4$, $\frac{\partial f}{\partial x} = -4$ and $\frac{\partial f}{\partial y} = -4$. The gradients can be thought of as flowing backwards through the circuit.

Each circle in the diagram can be referred to as a gate. Notice that every gate (the addition gate ($+$) and the multiplication gate ($*$)) gets some inputs and can right away compute its output value and the local gradient of its inputs with respect to its output value. This is done completely indpendently without being aware of any of the details of the full circuit that they are embedded in. However, during backpropogation the gate will eventually learn about the gradient of its ouput value on the final output of the entire circuit. According to the chain rul, the gate should take that gradient and multiply it into every gradient it normally computes for all of its inputs. Let us look at the example again to make this clear.

The (+) gate received inputs $[2, -5]$ and computed output 3. It also computed its local gradient with respect to both of its inputs, which is 1, since it is an addition operation. The rest of the circuit copmuted the final value to be -12. During the backward pass, the (+) gate learns that the gradient for its output was -4. It then takes that gradient and multiplies it to all of the local gradients for its inputs, which results in -4 and -4. This implies that if x, y were to decrease (responding to their negative gradients) then the (+) gate's output would decrease, which in turn makes the $(*)$ gate's output increase. Thus backpropogation can be thought of as gates communicating to each other through the gradient signal whether they want their outputs to increase or decrease, so as to make the final output higher.

1.5.4.6 Modularity

We introduced addition gates and multiplication gates, but any kind of differentiable function can act as a gate. We can also group multiple gates into a single gate or decompose a function into multiple gates whenever it is convenient. Consider the following expression to illustrate this:

$$f(\boldsymbol{w}, \boldsymbol{x}) = \frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2)}}.$$

This function is actualy a common piece in a neural network, but for now we can view it as mapping from inputs $\boldsymbol{x}, \boldsymbol{w}$ to a single number. The function is made up of multiple gates, and aside from the ones already discussed (addition, multiplication and max), they are:

$$\begin{aligned}
f(x) = \frac{1}{x} \quad &\implies \frac{df}{dx} = -\frac{1}{x^2} \\
f_c(x) = c + x \quad &\implies \frac{df}{dx} = 1 \\
f(x) = e^x \quad &\implies \frac{df}{dx} = e^x \\
f_a(x) = ax \quad &\implies \frac{df}{dx} = a
\end{aligned}$$

where c, a are constants. The full circuit for this expression is then:

The long chain of functions (gates) on the dot product of $\boldsymbol{x}$ and $\boldsymbol{w}$ is the decomposition of the *sigmoid function*:

$$\sigma(x) = \frac{1}{1 + e^{-x}}.$$

The derivative of the sigmoid function simplifies to a very convenient expression:

$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1 + e^{-x})^2} = \left(\frac{1 + e^{-x} - 1}{1 + e^{-x}}\right)\left(\frac{1}{1 + e^{-x}}\right) = (1 - \sigma(x))\sigma(x).$$

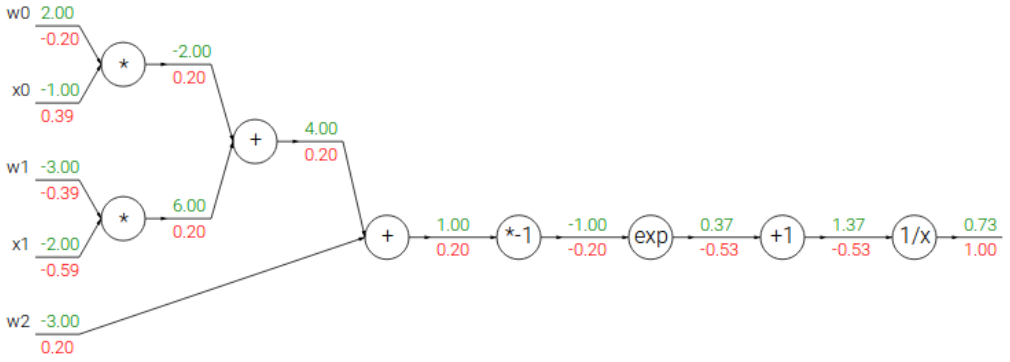

Figure 1.13: Sigmoid circuit.

Therefore in any real practical application it would be very useful to group the operations of the tail chain into a single gate.

1.5.4.7 Patterns in Backward Flow

It is interesting to note that in many cases the backward-flowing gradient can be interpreted on an intuitive level. Take the three most commonly used gates in neural networks, (add, mul, max), as an example. All of them have very simple interpretations in terms of how they act during backpropogation. Consider the following example circuit:

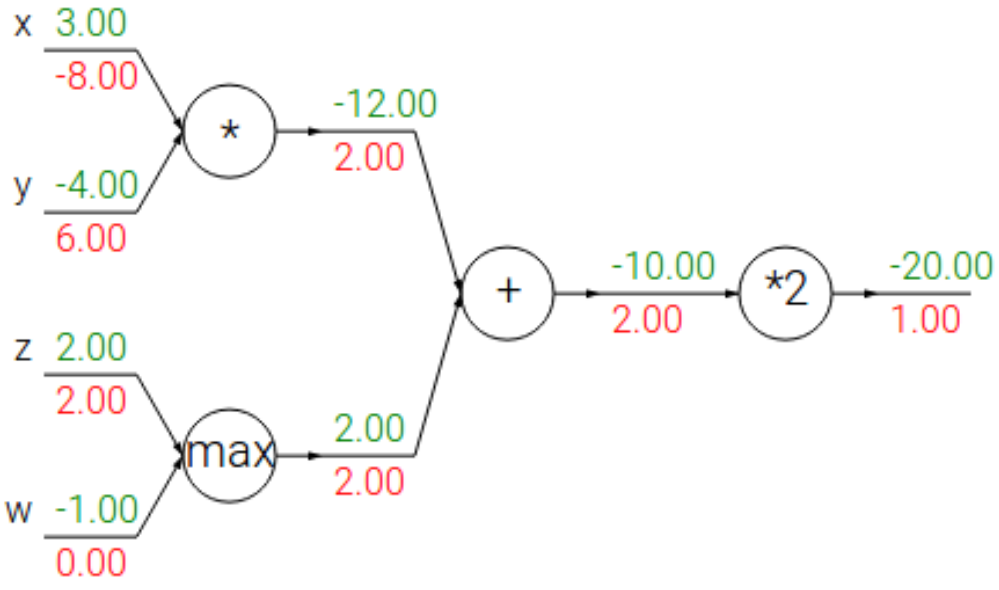

Figure 1.14: example cicuit for interpretation.

From the diagram above, the following patterns should be clear:

- The add gate always takes the gradient on its output and distributes it equally to all of its inputs, regardless of what their values were during the forward pass. This is because the local gradient for the add operation is always 1 for all its inputs.
- The max gate routes the gradient to exactly one of its inputs, the input that had the highest value during the forward pass. This because the local gradient for a max gate is 1 for the highest value and 0 for all other values.
- The multiply gate switches the gradients of its inputs and then multiply it by its output gradient.

Notice that if one of the inputs to the multiply gate is very small and the other is very big, then the multiply gate will do something slightly unintuitve: it will assign a relatively huge gradient to the small input and a tiny gradient to the large input. This is good to know, since in linear classifiers where the weights are multiplied by the inputs, it means that if the inputs are multiplied by a 1000, then the gradient on the weights will be 1000 times larger and you would have to lower the learning rate by that factor to compensate. This show how important preprocessing is for the optimisation of a classifier.

The above sections were concerned with single variables, but all concepts extend in a straight-forward manner to matrix and vector operations. However, one must pay closer attention to dimensions and transpose operations.

- http://cs231n.github.io/

1.6 Outline

We have now discussed the problem to be solved in the thesis, the basics of the data to be analysed and the basic components necessary for understanding Neural Network. The outline for the rest of the thesis will be as follows: In Chapter **??**, Neural Networks are described in detail, along with some of the recent developments in the field. Chapter **??** looks deeper into a variant of Neural Networks which is especially good in Image Classification problems, known as Convolutional Neural Networks. Again, the basics and the recent developments will be discussed. Chapter **??** first introduces the problem of Multi-Label Classification and then explores methods of extending CNNs to fit into this framework. Since we are working with satellite images, Chapter **??** will look at the recent approaches to image classification of satellite and remote sensing images. Then in Chapter **??** the most promising approaches highlighted in the previous chapters will be evaluated and compared to eachother in order

to find the approach with the best performance on our dataset. The results will also be compared to the findings in the literature. The thesis is concluded in Chapter **??** with a summary of the work done in this project, general discussion of the results and literature and what directions can be followed for future research.

Chapter 2

Neural Networks

2.1 Introduction

In **??** we have introduced all of the basic components for building a Neural Network. Recall the linear classifier that mapped the inputs, $\boldsymbol{x}$, to a vector of class scores, $\boldsymbol{s}$, $\boldsymbol{s} = W\boldsymbol{x}$, where W is a matrx of weights. Here W has K rows and p columns corresponding to the number of classes and size of the inputs respectively. As mentioned in **??**, a Neural Network has a more complicated mapping from the inputs to the class scores. An example Neural Network would instead have a mapping like, $\boldsymbol{s} = W_2 \max(0, W_1\boldsymbol{x})$. This time, for example, W_1 could be a matrix transforming the inputs to a 100-dimensional vector, thus of size $100 \times p$. The function $\max(0, \cdot)$ introduces a non-linearity that is applied element-wise. It simply thresholds all values below zero to zero. There are several types of non-linearities that can be applied, but this is a common choice. Finally, W_2, is a matrix of size $K \times 100$, mapping the intermediate vector to the final class scores. The key difference here is the non-linearity. If it is left out, the two matrices could be collapsed into one and therefore the predicted class scores would again be a linear function of the input. W_1, W_2 is learned through stochastic gradient descent (SGD), their gradients are derived with the chain rule and computed with backpropogation.

The above mapping is an example of a two-layer network. A three-layer neural network may look something like this:

$$\boldsymbol{s} = W_3 \max(0, W_2 \max(0, W_1\boldsymbol{x})).$$

Now there are two non-linearities and W_1, W_2, W_3 are all parameters to be learned. Their sizes, which determine the size of the intermediate layers (vectors), are seen as hyperparameters and how they can be determined will be discussed shortly. In the next section we will show where the name, Neural Networks, come from and ...

2.2 Biological Motivation and Connections

Originally primarily inspired by the goal of modelling biological neural systems, but has since diverged and become a matter of engineering and achieving good results in Machine Learning tasks.

Coarse model of biological neural systems and how they can be modelled.

2.3 Common Activation Functions

2.3.1 Sigmoid

2.3.2 Tanh

2.3.3 ReLu

- Leaky ReLu

2.3.4 Maxout

- mention where we will discuss ELU and SELU

2.4 Architectures

Layerwise organisation, naming conventions, size

Representational power -> universal approximators.

More on size, overfitting and generalisation, regularisation

- – https://arxiv.org/abs/1706.01350

2.5 Setup

2.5.1 Data Preprocessing

- mean subtraction
- normalisation
- pca and whitening
- leakage

2.5.2 Weight Initialisation

- all zero
- small random
- calibrating the variances

- sparse initialisation
- initialising biases

2.5.3 Batch Normalisation

- https://arxiv.org/abs/1502.03167

2.5.4 Regularisation

- L1
- L2
- maxnorm
- Dropout: http://www.cs.toronto.edu/~rsalakhu/papers/srivastava14a.pdf, http://papers.nips.cc/paper/4882-dropout-training-as-adaptive-regularization.pdf
- noise in forward pass
- bias regularisation
- per layerregularisation

2.6 Loss Functions

2.6.1 Classification

2.6.2 Attrubute Classification

2.6.3 Regression

2.6.4 Structered Prediction

2.7 Learning

Process of learning the parameters and finding good hyperparameters.

- practical tips for learning

2.7.1 Monitoring

- loss function + learning rate
- train/val acc
- ration of weights

- activation per layer
- first layer viz

2.7.2 Parameter Updates

- momentum
- Nesterov
- decay
- adaptive: adagrad, rmsprop, adam
- something on cyclical?

2.7.3 Freezing Layers

- https://arxiv.org/abs/1706.04983

2.8 Hyperparameter Optimisation

2.9 Evaluation

2.9.1 Ensembles

- same model diff initialisations
- top models through cv
- different checkpoints
- running average of parameters
- the paper on ensembling from cyclical minima
- maybe tta
- Pseudo-Labelling and Knowledge-Distillation

Chapter 3

Convolutional Neural Networks

3.1 Introduction

3.2 ConvNet Layers

3.2.1 Convolutional Layer

3.2.2 Pooling Layer

- mixed pool: https://pdfs.semanticscholar.org/de66/4f22dd4c7b4c15ac4a52513004aee55765ff.pdf and https://arxiv.org/pdf/1509.08985.pdf, can try implementation from https://github.com/fchollet/keras/issues/2816
- maxout?

3.2.3 Normalisation Layer

3.2.4 Fully Connected Layer

- also optiion to replace with conv layer

3.3 ConvNet Architectures

3.3.1 Layer Patterns

3.3.2 Layer Sizing Patterns

3.3.3 Famous Architectures

- AlexNet
- VGG
- ResNet
- DenseNet

- Inception (?)

3.4 Visualizing CNN's

- https://github.com/raghakot/keras-vis
- http://yosinski.com/deepvis

3.4.1 Activations and First Layer Weights

3.4.2 Images with Maximum Activation

3.4.3 t-SNE Embedding

3.4.4 Occluding

- more resources: http://cs231n.github.io/understanding-cnn/

3.5 Transfer Learning

3.5.1 Feature Extractor

3.5.2 Fine-Tuning

3.5.3 Pretrained Models

Chapter 4

Multi-Label Convolutional Neural Networks

4.1 Introduction

First discuss the domain of MLC in general (only the aspects transferable to Deep Learning) and then move on to look at approaches of ML with ConvNets

Multi-label (ML) learning belongs to the supervised learning paradigm and can be viewed as a generalisation of the traditional single-label learning problem. Suppose the data set to be analysed consists of a set of observations each representing a real-world object such as an image or a text document. In the single-label context each object is restricted to belonging to a single, mutually exclusive class, *i.e.* each observation is associated with a single label. One can quite effortlessly come up with tasks that will not fit into this framework: an image annotation problem where each image contains more than one semantic object, a text classification task where each document has multiple topics or an acoustic classification task where the recordings contain the sounds of multiple bird species. Therefore the need for a ML learner that can assign a set of labels to an observation. Let $\mathcal{L} = \{l_1, l_2, \ldots, l_K\}$ denote the complete set of possible labels that can be assigned to an observation. Whereas a single-label learner aims to find which single label l_k, $k = 1, 2, \ldots, K$, belongs to a given observation, a ML learner is capable of assigning a set of labels $L \subseteq \mathcal{L}$ to the observation.

According to (**?**), ML learning can be considered a sub-problem of a wider framework, called multi-target learning, covering all problems where an observation is associated with multiple outputs. When the output variables are binary, it is a ML learning problem. But problems also exist where the output variables are multi-class or numerical and in these settings the problems are respectively known as multi-dimensional learning and multi-output regression. It is also possible that the output variables are combinations of the aforementioned types.

As should be expected, the ML framework has a few concepts novel to the single-label case which should be reviewed before looking at the algorithms for ML learning. In this chapter, the core notation for the thesis will be introduced and a clear definition of the task of ML learning will be given. Then, a deep look is taken into the unique properties of ML data and how these might affect the performance of classifiers. The concepts of label correlation and class imbalance will also be introduced, however, how to deal with these will be discussed in the next chapter (for now). Finally, we will get to the evaluation metrics of ML algorithms. This is an important topic in ML learning, often neglected in the literature [cite]. After completing this chapter, the reader will have a good basis to be able to move on to the discussion of ML learning algorithms.

4.2 Notation

The following notation will be used throughout the thesis. Define the input matrix as

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^\intercal & \boldsymbol{x}_2^\intercal & \dots & \boldsymbol{x}_n^\intercal \end{bmatrix},$$

where n is the number of observations and p is the number of features. $\boldsymbol{x}_i^\intercal$ represents the p-dimensional vector that forms the i-th row of X. For a text classification problem, x_{ij} might indicate the number of times a word j appeared in document i. Define the label or output matrix as

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1K} \\ y_{21} & y_{22} & \cdots & y_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nK} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1^\intercal & \boldsymbol{y}_2^\intercal & \dots & \boldsymbol{y}_K^\intercal \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_{(1)} & \boldsymbol{Y}_{(2)} & \dots & \boldsymbol{Y}_{(K)} \end{bmatrix},$$

where K is the size of the label set $\mathcal{L}$. Y only contains zeros and ones, *i.e.* $y_{ik} = 1$ if label l_k, $k = 1, \dots, K$, is present for observation i and $y_{ik} = 0$ if it is absent. Thus $\boldsymbol{Y}_{(k)}$ is a n-dimensional binary vector indicating which observations are associated with label l_k. A ML data set will be defined as $D = \begin{bmatrix} X & Y \end{bmatrix}$, which contains the n input-output pairs, $\{(\boldsymbol{x}_i, \boldsymbol{y}_i) \,|i = 1, \dots, n\}$. Note that, $\boldsymbol{y}_i = (y_1, y_2, \dots, y_K)$, $y_k \in \{0, 1\}$, used here is the label vector, however, it is also common to use the label set notation, *i.e.* $L_i \subseteq \mathcal{L}$, where $\mathcal{L}$ is the complete label set and L_i is the set of relevant labels for observation i.

4.3 The Task of Multi-Label Learning

A more formal definition of the ML learning task will be given in the following chapter. However, it is important to note that we will define the utlimate task of ML learning as the assigning of multiple labels to an observation. ML learning covers to very similar approaches, namely, ML classification and ML ranking. ML classification algorithms output whether or not labels are relevant to an observation (binary) and ML ranking algorithms outputs a real-valued score assigned to each label indicating its relative importance to an observation. Thus with ML ranking, for each observation we seek a list of labels ordered by their scores representing the confidence in how relevant they are to the specific observation. Many classifiers base their final (categorical) prediction on the thresholding of the real-valued output of the algorithm and thus can also be used for ranking. Similarly, ranking algorithms can also be used for classification if a thresholding function is applied to the real-valued output. (see (**?**) for a more brief description)

- mathematical definition with notation of the task of ML learning.
- real-valued output + thresholding function (ranking vs classification)

The task of ML classification is to find a function h that accurately maps the observations contained in X to the label matrix Y, i.e., $h : X \to Y$, so that given a new observation, h can determine which labels belong to it. The accuracy aforementioned is a topic that will be discussed shortly. The measurement thereof is another unique problem for ML classification.

On the other hand, the goal of ML ranking is to find a function $f : X \to G$, where G is a similar matrix to Y, but with the g_{ij} a real value representing the relative confidence score that label j is relevant to observation i. f is found by optimising a ranking metric, also discussed shortly. From the confidence scores of observation i, $f(\boldsymbol{x}_i)$, a ranking $\boldsymbol{r}_i$ can be obtained, giving the rank of labels in descending order of $f(\boldsymbol{x}_i)$.

mention the calibration factor of (**?**). Finding z_i from r_i

h will be referred to as the ML classifier and f as the ML ranker. When ML learner will be a collective term covering both h and f. Before different ML learners can be discussed, an understanding of how the output of these algorithms are evaluated is necessary, since fitting f of h envolves optimising an evaluation metric. (always?)

Let $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$ define a multi-label dataset. $\boldsymbol{x}$, is the feature/input/instance vector of an observation and is given by a p-dimensional real-valued vector, $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$, *i.e.* $\boldsymbol{x} \in \mathcal{R}^p$. Each instance, $\boldsymbol{x}$ is associated with a subset of labels $L \in 2^{\mathcal{L}}$, where $2^{\mathcal{L}}$ represents the powerset of the full set of labels, $\mathcal{L} = \{l_1, l_2, \ldots, l_K\}$. The subset L is represented as an indicator vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_K)$, where $y_k = 1$ if $l_k \in L$ or else $y_k = 0$, for

$k = 1, 2, \ldots, K$. We assume examples in $\mathcal{D}$ to be independently and identically distributed (*i.i.d.*) from $P(\boldsymbol{X}, \boldsymbol{Y})$. Let h define a multi-label classifier, which is a mapping,

$$h : \boldsymbol{X} \to \boldsymbol{Y}$$

(not sure about this notation). The risk of h is defined as the expected loss over the joint distribution $P(\boldsymbol{X}, \boldsymbol{Y})$:

$$R_L(h) = E_{\boldsymbol{XY}} \left[L\left(\boldsymbol{Y}, h(\boldsymbol{X})\right) \right],$$

where $L(.)$ is a multi-label loss function. The MLC task boils down to given training data, $\mathcal{D}$, drawn independently from $P(\boldsymbol{X}, \boldsymbol{Y})$, learn a classifier h that minimizes the risk with respect to a specific loss function, *i.e.*

$$h^* = \arg\min_h E_{\boldsymbol{XY}} \left[L\left(\boldsymbol{Y}, h(\boldsymbol{X})\right) \right] = \arg\min_h E_{\boldsymbol{X}} \left[E_{\boldsymbol{Y}|\boldsymbol{X}} \left[L\left(\boldsymbol{Y}, h(\boldsymbol{X})\right) \right] \right],$$

where h^* is the so-called risk-minimizing model and can be determined in a pointwise way by the risk minimizer,

$$h^*(\boldsymbol{x}) = \arg\min_{\boldsymbol{y}} E_{\boldsymbol{Y}|\boldsymbol{X}} \left[L(\boldsymbol{Y}, \boldsymbol{y}) \right].$$

Note, here we allow $h(\boldsymbol{x})$ to take on real values, *i.e.* $h(\boldsymbol{x}) \in \mathcal{R}^K$, for the sake of generality. This is to cover multi-label ranking functions and multi-label classifiers that output real real values before thresholding.

4.3.1 Theoretical Results

- evaluate perfomance on many metrics for fairness
- minimisation of surrogate loss functions and consistency
- consistency (**?**):

They were the first to do a theoretical study on the consistency of multi-label learning algorithms, focusing on the ranking loss and the hamming loss. A learning algorithm is said to be consistent if its expected risk converges to the Bayes risk as the size of the training data increases. They found that any convex surrogate loss is inconsistent with the ranking loss and therefore proposed a partial ranking loss (which is consistent with some surrogate loss functions) as an alternative. They also show how some recent multi-label algorithms are inconsistent in terms of the hamming loss and provides a discussion on the consistency of approaches which transforms the multi-label problem into a set of binary classification tasks.

- more theoretical work at (**?**). Mentions: Finding theoretically correct algorithms for other non label-wise decomposable loss functions is still a great challenge.

- more theory: Optimizing the F-Measure in Multi-Label Classification: Plug-in Rule Approach versus Structured Loss Minimization

Other solutions: exploit correlation of labels from both types conditional and unconditional dependencies, features selection methods that are designed especially to handle multi label datasets, and having new stratification methods that are suitable to the nature of multi label datasets (copied from (**?**))

- most of these algorithms suffer from high complexity in the learning process [10]. Based on that, the true challenge is to exploit high order labels correlations locally and maintain a linear complexity at the same time [2].(copied from (**?**))

4.4 Multi-Label Indicators

As with all supervised learning problems, no one ML algorithm performs optimally on all problems. It is common practice in classical single output supervised learning to first consider, for example, the number of features (p) and the number of observations (n) in a data set before deciding on which model(s) to fit to the data. The same naturally holds for a ML problem but with added complexity. The multiple outputs of the data introduces many more factors to consider before continuing to the modelling phase. Some ML data sets have only a few labels per observation, while others have plenty. In some ML data sets the number of label combinations is small, whereas in others it can be very large. Some labels appear more frequently than others. Moreover, the labels can be correlated or not. These characteristics can have a serious impact on the performance of a ML classifier. This is the reason why several specific indicators have been designed to assess ML data set properties.

The two standard measures for the multi-labeledness of a data set are *label cardinality* and *label density*, introduced by (**?**). The label cardinality of a ML data, D, set is the average number of labels per observation:

$$LCard(D) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}.$$

This measure can be normalised to be independent of the label set size, which results in the label density indicator:

$$LDens(D) = \frac{1}{K}LCard(D) = \frac{1}{nK}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}.$$

According to (**?**) it is important to distinguish between these two measures, since two data sets with the same label cardinality but with a great difference in the number of labels might not exhibit the same properties and cause different behaviour to the ML classification methods. These two measures give a good

indication of the label frequency of a data set, but we are also interested in the uniformity and regularity of the labeling scheme. The authors of (**?**) suggested measuring the proportion of distinct label sets and the proportion of label sets with the maximum frequency. Consider the number of distinct label sets, also referred to as the label diversity (**?**), which can be defined as:

> there are multiple ways this is defined in the literature - still need to decide on which one I want to use

$$LDiv(D) = |\{Y|\exists \boldsymbol{x} : (\boldsymbol{x}, Y) \in D\}|,$$

by (**?**). ((**?**) uses $\exists!$ instead of $\exists$ and Y as a vector $\boldsymbol{y}$. I want to consider a way of defining it in matrix notation. Maybe with an indicator function. Some papers define it as DL instead of $LDiv$.) The proportion of distinct label sets in D is then

$$PLDiv\{/PUniq/PDL\}(D) = \frac{1}{n} LDiv(D).$$

The proportion of label sets with the maximum frequency is defined by (**?**) as:

$$PMax(D) = \max_{\boldsymbol{y}} \frac{\text{count}(\boldsymbol{y}, D)}{n},$$

where $\text{count}(\boldsymbol{y}, D)$ is the frequency that label combination $\boldsymbol{y}$ is found in data set D. This represents the proportion of observations associated with the most frequently occurring label sets. High values of $PLDiv$ and $PMax$ indicate an irregular and skewed labeling scheme, respectively, *i.e.* a relatively high number of observations are associated with infrequent label sets and a relatively high number of observations are associated with the most common label sets. (*think about this again*) When this is the case, and the labels are modelled separately, the classifiers will suffer from the class imbalance problem, a common problem in supervised classification tasks. More detail about this will be adressed shortly.

Very little research has been done on how all these ML indicators affect the performance of a ML classifier. (**?**) made a worthy attempt. Their goal was to find a way of determining which ML algorithm to use given a data set with specific properties and with a specific evaluation metric to optimise. They approached this problem by training a so called meta-learner on a meta-data set containing the performance of multiple ML algorithms on benchmark data sets with different properties. This trained meta-learner is then able to predict which ML algorithm is most likely to give the best results in terms of a specific evaluation metric, given the properties of the data set to be analysed. Although we will not use their meta-learner for this thesis, we will consider some of the additional findings in their research. They found that the

following properties (among others) of a ML data set was important to their trained meta-model (which was based on classification trees) in predicting which ML algorithm is most appropriate: K; $LDiv(D)$; $LCard(D)$; the standard deviation, skewness and kurtosis of the number of labels per observation in D; number of unconditionally dependent label pairs (based on what?); average of χ^2 -scores of all dependent label pairs; number of classes with less than 2, 5 and 10 observations; ratio of classes with less than 2, 5, 10 and 50 observations; average, minimal and maximal entropy of labels (def of entropy?); average observations per class. This strengthens the argument that it is important to take ML indicators into account before the training process.

Some rules that they found that I might refer to later:

- for micro-AUC target evaluation measure if label cardinality of training data is above 3.028 then the 2BR method (among the single-classifiers) should be used.
- Another example for an extracted rule is for ranking loss evaluation measure: if minimum of label entropies is zero (i.e. there is at least one certain label in the training set), number of labels is less than 53 and skewness of label cardinality is below or equal to 2.49 then the EPS method (among ensembles) should be used.

4.5 Evaluation Metrics

(**?**) first to categorise into label-based and example-based.

The evaluation of the performance of ML algorithms is another distinct problem to this setting. Compared to the single-label case, many more evaluation metrics exist, with subtle or obvious differences in their measurement. According to (**?**) it is essential to evaluate a ML algorithm on multiple and contrasting measures because of the additional degrees of freedom introduced by the ML setting. In addition, care should be taken when reporting multiple measures and with their interpretation. Since some of the measures are contrasting it is dangerous to report multiple metrics and conclude that on average one learner is better than the other. This was highlighted in (**?**), where the authors suggested that when evaluating the performance of a ML learner, it should be made clear which metric(s) it is aiming to optimise, otherwise the results can be misleading. It is impossible (?) for a learner to have superior performance over others in terms of all the multi-label evaluation metrics simultaneously.

The evaluation measures of predictive performance of multi-label learnerss can be divided into two groups: example-based and label-based measures. Example-based measures compares the actual versus the predicted labels for each observation and then computes the average across all the observations in the dataset. Where label-based measures computes the predictive performance

on each label separately and then averages across all labels (**?**). For both groups the measures can further be partitioned into metrics from a classification persepective and measures from a ranking perspective, *i.e.* metrics for h and metrics for f respectively. The most commonly used metrics in each of the groups will be introduced here.

4.5.1 Brief Taxonomy

- more complicated than single label metrics
- introduce example based vs label based
- for classification and ranking
- diagram / table + where they are used

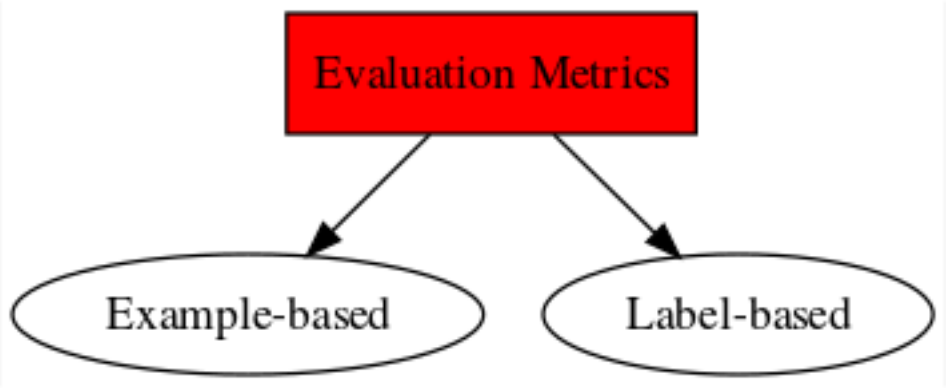

Figure 4.1: Categorisation of the taxonomy of MLL evaluation metrics

- **??** is just an example. The image quality is lacking.

4.5.2 Example-based Metrics

- subset accuracy; hamming loss; accuracy; precision; recall; one-error; coverage; ranking loss; average precision
- definition + brief interpretaion where it is unclear

For the following definitions, let y_i be the set of true labels for observation $\boldsymbol{x}_i$ and z_i the set of predicted labels for the same observation, obtained from the predicted indicator vector of $\hat{h}(\boldsymbol{x}_i)$. The Hamming loss is then defined as

$$\text{hloss}(h) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{K}|z_i \triangle y_i|,$$

where $\triangle$ stands for the symmetric difference and $|.|$, the size of the set. For example, $|\{1,2,3\}\triangle\{3,4\}| = |\{1,2,4\}| = 3$. Thus the Hamming loss counts

the number of labels not in the intersection of the predicted subset of labels and the true subset of labels, as a fraction of the total size of the labelset, averaged across each observation in the dataset. When h returns perfect predictions for each observation in the dataset, $\text{hloss}(h) = 0$, and if h predicts for each observation that it belongs to all the labels except for its the true labels, $\text{hloss}(h) = 1$.

Accuracy is defined as

$$\text{accuracy}(h) = \frac{1}{n}\sum_{i=1}^{n}\frac{|z_i \cap y_i|}{|z_i \cup y_i|}.$$

Thus for each observation the number of correctly predicted labels is calculated as a proportion of the sum of the correctly and incorrectly predicted labels. These quantities are then averaged over each observation in the dataset. If the h perfectly predicts the relevant subset of labels for each observations, $\text{accuracy}(h) = 1$. If h does not manage to predict a single correct label for any observation, $\text{accuracy}(h) = 0$.

The precision and recall are respectively defined as

$$\text{precision}(h) = \frac{1}{n}\sum_{i=1}^{n}\frac{|z_i \cap y_i|}{|z_i|},$$

and

$$\text{recall}(h) = \frac{1}{n}\sum_{i=1}^{n}\frac{|z_i \cap y_i|}{|y_i|}.$$

Precision calculates the average proportion of correctly predicted labels in terms of the number of labels predicted, across all the observations in the dataset. Recall calculates a similar average, with the only difference that the proportion is calculated in terms of the number of true labels per observation. Both these metrics lie in the range $[0, 1]$ with larger values desirable.

The harmonic mean between the precision and the recall is called the F_1-score and is defined as

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2|z_i \cap y_i|}{|z_i| + |y_i|}.$$

The perfect classifier will result in a F_1-score of 1 and the worst possible score is zero.

The subset accuracy or classification accuracy is defined as

$$\text{subsetacc}(h) = \frac{1}{n}\sum_{i=1}^{n} I(z_i = y_i),$$

where $I(.)$ is the indicator function. This the subset accuracy is the proportion of observations that were perfectly predicted by h.

The above are all performance measures of ML classifiers. If the ML learner outputs real-valued confidence scores, these ranking metrics can be used to evaluate the learner's performance:

One-error:

Coverage:

Ranking Loss:

Average Precision:

4.5.3 Label-based Metrics

- micro vs macro ito tp, tn, fp, fn
- auc example

The idea with label-based measures is to compute a single-lable metric for each label based on the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) made by the classifier on a dataset and then obtaining an average of the values (**?**). Note, TN_k, TP_k, FN_k and FP_k denote the quantities for label l_k, $k = 1, 2, \ldots, K$. Thus $TP_k + TN_k + FP_k + FN_k = n$. Let B be any binary classification metric, i.e. $B \in \{\text{accuracy}, \text{precision}, \text{recall}, F_1\}$. B can be written in terms of TN_k, TP_k, FN_k and FP_k, for example

$$\text{accuracy}(TN_k, TP_k, FN_k, FP_k) = \frac{TP_k + TN_k}{TP_k + TN_k + FP_k + FN_k}.$$

B is then calculated for each label and then an average is calulated. The averaging can be done either by the micro or the macro approach. The micro approach considers predictions of all observations together and then calculates the measure across all labels, i.e.

$$B_{micro} = B\left(\sum_{k=1}^{K} TP_k, \sum_{k=1}^{K} TN_k, \sum_{k=1}^{K} FP_k, \sum_{k=1}^{K} FN_k\right).$$

Whereas the macro approach computes one metric for each label and then the values are averaged over all the labels, i.e.

$$B_{macro} = \frac{1}{K}\sum_{k=1}^{K} B(TP_k, TN_k, FP_k, FN_k).$$

Note, also that $\text{accuracy}_{micro}(h) = \text{accuracy}_{macro}(h)$ and that $\text{accuracy}_{micro}(h) + \text{hloss}(h) = 1$, since Hamming loss is the average binary classification error.

Again, all of the above mentioned metrics are from a classification perspective. An example of a label-based metric from a ranking persepective is the macro- and micro-averaged AUC:

Most multi-label classifiers learn from the training observations by explicity or implicitly optimising one specific metric (**?**). That is why in (**?**) the authors

reccomended specifying which of the metrics a new proposed algorithm aims to optimise in order to show if it is succesful. But at the same time it is important to test the algorithm on numerous metrics for fair comparisons against other algorithms (**?**), (**?**). It might be that a algorithm does very well in terms of the Hamming loss, but performs poorly according to the subset accuracy, or vice versa, as shown in (**?**). In (**?**) they claim that the Hamming loss reported together with the micro-average F-measure gives a good indication of the performance of a multi-label classifier.

These multi-label metrics are usually non-convex and discontinuous (**?**). Therefore multi-label classifiers resort to considering surrogate metrics which are easier to optimise.

probably should add an example or maybe later

Other than predictive performance, are there other aspects on which multi-label classifiers can be evaluated, such as efficiency and consistency. Multi-label algorithms should be efficient in the sense that it takes the least amount of computational power for a given level of predictive performance (**?**). These classifiers can take a considerable amount of time to train when complicated ensembles are being implemented on datasets with huge labelsets. In cases where live updating and predictions are needed, this may be a problem [reference]. The other desirable attribute of multi-label classifiers are that they are consistent. This means that the expected loss of the classifier converges to the Bayes loss when the number of observations in the training set tends to infinity. Actually only a very few number of multi-label classifiers satisfy this property (**?**), (**?**).

4.6 Label Dependence

With this chapter I want to investigate the need for approaches in multi-label classification which model the dependence structure between labels. For this we need a sound theoretical definition and analysis of label dependence and then we might want to investigate it empirically with synthetic datasets (or real world). The main papers inspiring this chapter are (**?**) and (**?**), and some content will be taken from (**?**), (**?**) (for empirical evidence maybe), (**?**), (**?**), (**?**). My main hypothesis is that modelling the input-output pairs individually should have just as good, if not better performance compared to approaches trying to model label dependence, since all the available information of the labels should be contained in X and by the assumption that label y_i can be determined with the help of the knowledge of label y_j, it should also be possible to find y_i from X since y_j is found from X. This argument probably only holds for approaches trying to "correct" binary relevance (BR) with regards to its lack of modelling label dependence, such as classifier chains (CC), stacking like MBR/2BR/BR+, etc. Reformulate hypothesis later.

It is essentially a given in multi-label classification literature that in order to obtain competitive results, a learner should be able to model the dependence structure between labels in some way. Whenever a new MLC algorithm is proposed, it will be compared to independent label learning (BR) and if it has superior empirical performance, it is usually ascribed to its ability of modelling label dependence in some ad-hoc way (examples?). The authors of (**?**), (**?**) and (**?**) were the first to point out this lack of understanding of the term *label dependence* in the literature (later on a comprehensive and extended discussion of the topics covered in the aforementioned papers was given in (**?**)). They argued that *label dependence* is only understood and used by most in the literature in a purely intuitive manner, and that in order to build a better understanding of multi-label classifiers, theoretical backing is essential.

Modelling each label independently, *i.e.* using the binary relevance (BR) approach, is one of the simplest and most intuitive approaches to tackling the multi-label problem. But it has been criticized and overlooked by the majority because it does not take into account the possible dependence between labels. However, BR has many advantages. (**?**) shows that BR is the risk minimizer of the Hamming Loss and (**?**) pointed out that it is very rare for 'improved' methods to achieve significantly better results than BR in terms of this measure (also visible in (**?**) (make sure)). In addition, BR is highly resistant to overfitting label combinations, since it does not expect samples to be associated with previously-observed combinations of labels [Read2011a]. It can naturally handle data streaming or other dynamic scenarios where the addition and removal of labels are quite common. BR's biggest strength is its low computational complexity compared to other multi-label classification methods. It scales linearly with increasing number of labels and it is easily parallelizable - desirable properties, especially working with large label sets.

Recently, (**?**) has gone so far as to claim that BR can perform just as well as methods supposedly modelling label dependence, and if it does not, it is usually because of the inadequacy of the base learners used. In other words, if the base learner can extract the right features, BR will be as good as any other multi-label classifier, without the need to model label dependence. Some theoretical justifications were given but the empirical evidence was not convincing. This is what motivated the writing of this chapter - to answer the question, "is it essential for a multi-label classifier to take label correlations into account in order to be optimal?". To investigate this one needs a thorough, theoretical understanding of *label dependence*, how to possibly exploit it and how to evaluate it. This is what this chapter aims to do. Most of the work is based on the papers (**?**) and (**?**). We will also attempt to back up the theory with empirical results.

4.6.1 Two types of label dependence

As mentioned, most mutli-label learning papers display merely an intuitive understanding of *label dependence*, in the sense that in predicting a specific label, the information on the rest of the labels may be helpful. For example in an image recognition problem, if a picture is labelled with *beach* and *ocean*, *sand* will most likely be a relevant label. Clearly, this understanding is insufficient to gain advances in the multi-label learning literature (later on it will also be pointed out why this may indeed not make intuitive sense). In this section, a formal statistical definition of the two types of label dependence will be given. First, we briefly revisit the task of multi-label classification (MLC), in mathematical(?) terms.

4.6.1.1 Marginal vs. conditional dependence

First note that we denote the conditional distribution of $\boldsymbol{Y} = \boldsymbol{y}$ given $\boldsymbol{X} = \boldsymbol{x}$ as

$$P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x}) = P(\boldsymbol{y}|\boldsymbol{x})$$

and the corresponding conditional marginal distribution of Y_k (conditioned on $\boldsymbol{x}$) as

$$P(Y_k = b|\boldsymbol{x}) = \sum_{y_i=b} P(\boldsymbol{y}|\boldsymbol{x}).$$

(can probably also write as $P(Y_k|\boldsymbol{x})$ since b is either 0 or 1?)

(**?**) defines two types of dependence among lables, namely, conditional dependence and marginal dependence. Their definitions follow:

Definition 1 *A random vector of labels $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_K)$ is called marginally independent if*

$$P(\boldsymbol{Y}) = \prod_{k=1}^{K} P(Y_k). \tag{4.6.1}$$

Marginal dependence is also known as unconditional dependence and can be thought of as a measure of the frequency of co-occurrence among labels. Conditional dependence captures the dependence of the labels given a specific observation $\boldsymbol{x}$.

Definition 2 *A random vector of labels is called conditionally independent, given $\boldsymbol{x}$ if*

$$P(\boldsymbol{Y}|\boldsymbol{x}) = \prod_{k=1}^{K} P(Y_k|\boldsymbol{x}). \tag{4.6.2}$$

The conditional joint distribution of a random vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_K)$ can be expressed by the product rule of probability ($P(AB) = P(A|B)P(B)$):

$$P(\boldsymbol{Y}|\boldsymbol{x}) = P(Y_1|\boldsymbol{x}) \prod_{k=2}^{K} P(Y_k|Y_1, \ldots, Y_{k-1}, \boldsymbol{x}). \tag{4.6.3}$$

A similar expression can be given for $P(\boldsymbol{Y})$. If $Y_1, Y_2, \ldots, Y_K$ are conditionally independent, then **??** will simplify to **??**.

Marginal and conditional dependence are closely related - it can be written as:

$$P(\boldsymbol{Y}) = \int_{\mathcal{X}} P(\boldsymbol{Y}|\boldsymbol{x}) d\mu(\boldsymbol{x}), \tag{4.6.4}$$

where μ is the probability measure on the input space $\mathcal{X}$ induced by the joint probability distribution P on $\mathcal{X} \times \mathcal{Y}$. Marginal dependence can roughly be viewed as an 'expected dependence' over all instances. Nevertheless, marginal dependence does not imply conditional independence, or *vice versa*. Two examples from (**?**) are given to illustrate this.

Example 1 *Suppose two labels, Y_1 and Y_2, are independently generated from $P(Y_k|\boldsymbol{x}) = (1 + \exp(-\phi f(\boldsymbol{x}))^{-1}$, where ϕ controls the Bayes error rate. Thus, by definition, the two labels are conditionally independent with conditional joint distribution, $P(\boldsymbol{Y}|\boldsymbol{x}) = P(Y_1|\boldsymbol{x}) \times P(Y_2|\boldsymbol{x})$. However, as $\phi \to \infty$, the Bayes error tends to zero and the marginal dependence increases to an almost deterministic case of $y_1 = y_2$. Showing, conditional independence does not imply marginal independence.*

Example 2 *Suppose two labels, Y_1 and Y_2, are to be predicted by using a single binary feature, x_1. Let the joint distribution $P(X_1, Y_1, Y_2)$ be given by the following table:*

x_1	y_1	y_2	P
0	0	0	0.25
0	0	1	0.00
0	1	0	0.00
0	1	1	0.25
1	0	0	0.00
1	0	1	0.25
1	1	0	0.25
1	1	1	0.00

Thus, the labels are not conditionally independent,

$$P(Y_1 = 0, Y_2 = 0|x_1 = 1) = 0 \neq P(Y_1 = 0|x_1 = 1) \times P(Y_2 = 0|x_1 = 1) = 0.25 \times 0.25,$$

but it can be shown that they are indeed marginally independent. For example,

$$P(Y_1 = 0, Y_2 = 0) = 0.25 = P(Y_1 = 0) \times P(Y_2 = 0) = 0.5 \times 0.5.$$

This holds for all the combination of labels, showing that marginal independence does not imply conditional independence.

This distinction between marginal and conditional dependence is crucial in the attempt to model label dependence in multi-label classification. We describe a multi-output model with the following notation, similar to (**?**):

$$Y_k = h_k(\boldsymbol{X}) + \epsilon_k(\boldsymbol{X}), \tag{4.6.5}$$

for all $k = 1, 2, \ldots, K$. $h_k : \boldsymbol{X} \to \{0, 1\}$ will be referred to as the structural part and $\epsilon_k(\boldsymbol{x})$ as the stochastic part of the model. Note that a common assumption in multi-variate regression (real-outputs) is that

$$E[\epsilon_k(\boldsymbol{x})] = 0. \tag{4.6.6}$$

for all $\boldsymbol{x} \in \boldsymbol{X}$ and $k = 1, 2, \ldots, K$. This is not a reasonable assumption in mutli-label classification (**?**) - the distribution of the noise terms can depend on $\boldsymbol{x}$ and two or more noise terms can depend on each other. Classifier h_k might also be very similar to h_l, $l \neq k; l = 1, 2, \ldots, K$. Thus there are two possible sources of label dependence: the structural part and the stochastic part of the model.

It seems that marginal dependence between labels is caused by the similarity between the structural parts. This assumption is made since it is reasonable to assume that the structural part will dominate the stochastic part. Suppose there exists a function $f(.)$ such that $h_k \approx f \circ h_l$, *i.e.*

$$h_k(\boldsymbol{x}) = f(h_l(\boldsymbol{x})) + g(\boldsymbol{x}), \tag{4.6.7}$$

with $g(.)$ being negligible in the sense that $g(\boldsymbol{x}) = 0$ with high probability. Then this $f(.)$-*dependence* between the classifiers is likely to dominate the averaging process in **??**, compared to $g(.)$ and the stochastic parts. This is what happens in Example 1 when $\phi \to \infty$. Thus we see that even if the dependence between h_k and h_l is only probable, it can still induce a dependence between the labels Y_k and Y_l (verstaan nie presies wat hier bedoel word nie). Another example illustrating idea is given from (**?**).

Example 3 *Consider a problem with a 2-dimensional input $\boldsymbol{x} = (x_1, x_2)$, where x_i is uniformly distributed in $[-1, 1]$ for $i = 1, 2$, and two labels, Y_1, Y_2, determined as follows. Y_1 is set to 1 for all positve values of x_1, i.e. $Y_1 = I(x_1 > 0)$. The second label is generated similarly but with the decision boundary of Y_1 ($x_1 = 0$) rotated by an angle of $\alpha \in [0, \pi]$ (give illustration). In addition, let the two error terms of the model be independent and both flip the*

label with a probability of 0.1. *If* α *is close to zero, the labels will almost be identical and a high correlation will be observed between them. But if* $\alpha = \pi$, *the decision boundaries of the labels are orthogonal and a low correlation will be observed.*

With regards to **??**, in Example 3, $f(.)$ is the identity function and $g(.)$ given by the ± 1 in the regions between the decision boundaries. From this point of view, marginal dependence can be seen as a kind of soft constraint that a learning algorithm can exploit for the purpose of regularization (**?**). (verstaan nie wat dit beteken nie)

For the conditional dependence, it seems that the stochastic part of the model is the cause. In Example 3, Y_1 and Y_2 is conditionally independent because the error terms are assumed to be independent. However, if there is a close relationship between ϵ_1 and ϵ_2, this conditional independence will be lost. (**?**) proves the proposition that a vector of labels is conditionally dependent given $\boldsymbol{x}$ if and only if the error terms in **??** are conditionally dependent given $\boldsymbol{x}$, *i.e.*

$$E\left[\epsilon_1(\boldsymbol{x}) \times \cdots \times \epsilon_K(\boldsymbol{x})\right] \neq E\left[\epsilon_1(\boldsymbol{x})\right] \times \cdots \times E\left[\epsilon_K(\boldsymbol{x})\right].$$

(Include proof?) It should also be noted that conditional independence can also cause marginal dependence because of **??**. Thus the similarity between models is not the only source of of marginal dependence.

What we have learned thus far is that there is a difference between marginal and conditional label dependence. The presence of marginal dependence does not imply conditional label dependence and *vice versa*. If label correlations are observed it can only be assumed that marginal dependence between the labels exist. It does not necessarily imply that there are any dependencies among the error terms (although it could be the cause). On the other hand, if conditional dependence is observed, one can safely assume that there are dependencies among the error terms. Next, we see how to exploit both types of label dependence to improve predictive accuracy.

4.6.2 Link between label dependence and loss minimization

One can view the MLC task from different persepectives in terms of loss minimizations. (**?**) describes three such views, determined by the type of loss function to be minimized, the type of dependence taken into account and the distinction between marginal and joint distribution estimation. The three views and the main questions to consider for each of them are:

1. The individual label view: How can we improve the predictive accuracy of a single label by using information about other labels?

2. The joint label view: What type of non-decomposable MLC loss functions is suitable for evaluating a multi-label prediction as a whole and how to minimize such loss functions?
3. The joint distribution view: Under what conditions is it reasonable to estimate the joint conditional probability distribution over all label combinations?

4.6.2.1 The individual label view

With this view, the goal is to minimize a loss function that is label-wise decomposable and we want to determine whether or not it will help taking label relationships into account. The most common and intuitive label-wise decomposable loss function is the Hamming loss, which is defined as the fraction of labels whose relevance is incorrectly predicted:

$$L_H\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right) = \frac{1}{K} \sum_{k=1}^{K} I\left(y_k \neq \hat{y}_k\right). \tag{4.6.8}$$

?? is only the Hamming loss for one observation. To compute the Hamming loss over an entire dataset, **??** is averaged over all the observations.

It is easy to see that the Hamming loss is minimized when

$$\hat{\boldsymbol{y}} = (\hat{y}_1, \ldots, \hat{y}_K),$$

where

$$\hat{y}_k = \arg \max_{y_k \in \{0,1\}} p(y_k|\boldsymbol{x}),$$

for $k = 1, 2, \ldots, K$. This shows that it is enough to take only the conditional marginal distribution $P(Y_k|\boldsymbol{x})$ into account to solve the problem, at least on a population level. Thus the Hamming loss is minimized by BR. (**?**) also gives a similar result for label-wise decomposable loss functions in general (thus also relevant for F-measure, AUC, *etc.*). This result implies that the multiple single label predictions problem can be solved on the basis of $P(Y_k|\boldsymbol{x})$ alone. Hence, with a proper choice of base classifiers and parameters for estimating the conditional marginal probabilities, there is in principle no need for modelling conditional dependence between the labels. However, in cases where the base classifiers are inadequate, dependence between the errors will exist and BR will give a suboptimal solution (make sure this statement is used correctly). Methods exist to improve BR in these situations and will be discussed shortly.

4.6.2.2 The joint label view

Here we are interested in non-decomposable (label-wise) MLC loss functions such as rank loss and the subset 0/1 loss. We discuss when they are appropriate and how to minimize them. First, consider the rank loss. Suppose the true labels constitute a ranking in which all relevant labels ideally precede all

irrelevant ones and $\boldsymbol{h}(\boldsymbol{x}) = (h_1(\boldsymbol{x}), \ldots, h_K(\boldsymbol{x}))$ is seen as a ranking function representing a degree of label relevance sorted in a decreasing order. The rank loss simply counts the number of label pairs that disagree in these two rankings:

$$L_r\left(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x})\right) = \sum_{(k,l);y_k>y_l} \left(I\left(h_k(\boldsymbol{x}) < h_l(\boldsymbol{x})\right) + \frac{1}{2} I\left(h_k(\boldsymbol{x}) = h_j(\boldsymbol{x})\right) \right). \quad (4.6.9)$$

This function is not convex nor differentiable, thus an alternative would be to minimize a convex surrogate like the hinge or exponentional function. However, (**?**) proves that it is enough to minimize **??** by sorting the labels by their probability of relevance:

Theorem 1 *A ranking function that sorts the labels according to their probability of relevance, i.e. using the scoring function* $\boldsymbol{h}(.)$ *with* $h_k(\boldsymbol{x}) = P(Y_k = 1|\boldsymbol{x})$*, minimizes the expected rank loss.*

(include proof?) This implies again (just like in the case for the label-wise decomposable loss functions) that, in principle, it is not necessary to know the joint label distribution $P(\boldsymbol{Y}|\boldsymbol{x})$ when training a multi-label classifier, *i.e.* risk-minimizing predictions can be made without any knowledge about the conditional dependency between labels. Thus, to minimize the rank loss, one can simply use any approach minimizing the single label losses. Note this results does not hold for the normalized version of rank loss.

Next, we look at the extremely stringent multi-label loss function, the subset 0/1 loss:

$$L_S\left(\boldsymbol{y}, \hat{\boldsymbol{y}}\right) = I\left(\boldsymbol{y} \neq \hat{\boldsymbol{y}}\right). \quad (4.6.10)$$

Although most would agree that this is not a fair measure for MLC performance, since it does not disinguish between almost correct and completely wrong, it is still interesting to study with regards to exploiting label dependence. The risk-minimizing prediction for **??** is given by the mode of the distribution:

$$h_s^*(\boldsymbol{x}) = \arg\max_{\boldsymbol{y}} P(\boldsymbol{Y}|\boldsymbol{x}). \quad (4.6.11)$$

This implies that the entire distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$ is needed to minimize the subset 0/1 loss. Thus a risk minimizing prediction requires the modelling of the joint distribution and hence the modelling of the conditional dependence between labels. Later on we will show an important results that under independent outputs, minimizing the Hamming loss and the subset 0/1 loss is equivalent, implying that BR will indeed also minimize the subset 0/1 loss (consider to show it here).

The cases for F-measure loss and the Jaccard distance is a bit more complicated and will not be discussed here. (give citation of where this can be found)

4.6.2.3 The joint distribution view

not sure if I want to mention the joint distribution view. Maybe only distinguish between single label and joint label prediction approach.

We just saw that minimzing the subset 0/1 loss requires the estimation of the entire conditional joint distribution, $P(\boldsymbol{Y}|\boldsymbol{X})$. Generally, if the joint distribution is known, a risk-minimizing prediction can be derived for any loss function in an explicit way:

$$h^*(\boldsymbol{x}) = \arg\min_{\boldsymbol{y}} E_{\boldsymbol{Y}|\boldsymbol{x}} \left[L(\boldsymbol{Y}, \boldsymbol{y}) \right].$$

In some applications modelling the joint distribution may result in using simpler classifiers, potentially leading to a lower cost and a better performance compared to directly estimating marginal probabilities by means of more complex classifiers. Nevertheless, it remains a difficult task. One has to estimate 2ˆ{K} values to estimate for a given $\boldsymbol{x}$.

4.6.3 Previous attempts to 'exploit' label dependence

4.6.4 Improved attempts to 'exploit' label dependence

Theoretical insights into MLC

- when new MLC algorithm is introduced, it should be specified which loss functions it intends to minimize. Otherwise it may give misleading results (like that it is optimal for many loss functions).
- a classifier supposed to be good in solving one problem may perform poorly on a different problem and vice versa
- restricts attention to hamming loss and subset 0/1 loss.
- hamming is representative of single label scenario and subset 0/1 for the mutli-label loss.
- assumes unconstrained hypothesis space.
- proposition (with proof in paper): The hamming loss and subset 0/1 loss have the same risk-minimizer, *i.e.* $\boldsymbol{h}^*_H(\boldsymbol{x}) = \boldsymbol{h}^*_s(\boldsymbol{x})$, if one of the following conditions holds: (1) Labels $Y_1, \ldots, Y_K$ are conditionally independent, *i.e.* $P(\boldsymbol{Y}|\boldsymbol{x}) = \prod_{k=1}^{K} P(Y_k|\boldsymbol{x})$. (2) The probability of the mode of the joint probability is greater than or equal to 0.5, *i.e.* $P(\boldsymbol{h}^*_S(\boldsymbol{x})|\boldsymbol{x}) \geq 0.5$.
- corollary (with proof in paper): In the separable case (*i.e.* the joint conditional distribution is deterministic, $P(\boldsymbol{Y}|\boldsymbol{x}) = I(\boldsymbol{Y} = \boldsymbol{y})$), the risk minimizers of the hamming loss and subset 0/1 loss coincide.
- Then 3 propositions on upper bounds of these losses. Ponder its relevance.

MLC algorithms for exploiting label dependence

- proposed algorithms improve predictive performance by supposedly modelling label dependence.
- type of dependence and loss to be optimized is omitted
- leads to poor designs and misleading results.
- focus on PT methods
- discussion on BR:
- simplest. does not take marginal or conditional dependence into account.
- in general not able to yield risk-mininizing predictions for multi-label losses but is well suited for loss functions whose risk-minimizer can solely be expressed in terms of marginal (conditional) distributions.
- may be sufficient, but exploiting marginal dependencies may still be beneficial especially for small-sized problems.
- moves discussion to single label predictions
- several methods that exploit similarities between structural parts of the label models.
- general scheme:

$$\boldsymbol{y} = \boldsymbol{b}(\boldsymbol{h}(\boldsymbol{x}), \boldsymbol{x}), \tag{4.6.12}$$

where $\boldsymbol{h}(\boldsymbol{x})$ is the binary relevance learner and $\boldsymbol{b}(.)$ is an additional classifier that shrinks or regularizes the solution of BR. Or

$$\boldsymbol{b}^{-1}(\boldsymbol{y}, \boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x}), \tag{4.6.13}$$

where the output space is first transformed and then the BR classifiers are trained and then transformed back to original. + Stacking follows first scheme. Form of regularization or feature expansion. Not clear which inputs should all be use for second level. + multivariate regression + kernel dependency estimation + compressive sensing + next section on methods that seek to estimate the joint distribution $P(\boldsymbol{Y}|\boldsymbol{x})$. + LP. Largest drawback the number of label combinations + the literature usually claims LP is generally the rigth approach. FALSE. LP takes conditional dependence into account but usually fails for losses like Hamming. + can improve with RAKEL, but it is still not well understood from a theoretical point of view. + PCC. computationally more manageable. ECC to reduce importance of label chain order.

Experimental evidence

- real and synthetic data
- BR, SBR, CC, LP
- hamming loss and subset 0/1
- MULAN
- logistic regression for base classifier.
- marginal independence: stacking does improve on BR, CC similar to SBR, LP also bad. Error increases with number of labels. hamming and subset 0/1 coincide.

- conditional independence: again loss functions coincide. SBR improves over BR, even higher when structural parts are more similar.Supports theoretical claim that the higher the structural similarties the more prominent effect of stacking. Study rest of results.
- conditional dependence:
- xor problem

Conclusions

- study

Nou opsomming van (**?**) - sodra klaar, probeer in hoofstuk inkorporeer.
Introduction

- n-th feature vector $\boldsymbol{x}^{(n)} = [x_1^{(n)}, \ldots, x_p^{(n)}]$, where $x_j \in \mathcal{R}$, $j = 1, \ldots, p$.
- in the traditional binary classification task we are intersted in having a model h to provide a prediction for test instances $\tilde{\boldsymbol{x}}$, *i.e.* $\hat{y} = h(\tilde{\boldsymbol{x}})$. In MLC there are K binary output class variables (labels) and thus $\hat{\boldsymbol{y}} = [\hat{y}_1, \ldots, \hat{y}_K] = h(\boldsymbol{x})$.
- probabilistic speaking h seeks the expecctation $E[\boldsymbol{y}|\boldsymbol{x}]$ of unknown $p(\boldsymbol{y}|\boldsymbol{x})$. This task is typically posed as a MAP estimate of the joint posterior mode
$$\hat{\boldsymbol{y}} = [\hat{y}_1, \ldots, \hat{y}_K] = h(\tilde{\boldsymbol{x}}) = \arg \max_{\boldsymbol{y} \in \{0,1\}^p} p(\boldsymbol{y}|\tilde{\boldsymbol{x}})$$
This corresponds to minimizing the susbet 0/1 loss.
- $h_{BR}(\tilde{\boldsymbol{x}}) := [h_1(\tilde{\boldsymbol{x}}), \ldots, h_K(\tilde{\boldsymbol{x}})]$
- entirety of ML literature point out that BR obtain suboptimal performance because it assumes labels are independent.
- several approaches attempt to correct/regularize BR, SBR.
- others attempt to learn the labels together, LP. $\hat{\boldsymbol{y}} = h_{LP}(\tilde{\boldsymbol{x}})$
- another example is CC done using a greedy search:
$$h_{CC}(\tilde{\boldsymbol{x}}) := [h_1(\tilde{\boldsymbol{x}}), h_2(\tilde{\boldsymbol{x}}, h_1(\tilde{\boldsymbol{x}})), \ldots, h_K(\tilde{\boldsymbol{x}}, \ldots, h_{K-1}(\tilde{\boldsymbol{x}}))]$$
- PCC formulates CC as the joint distribution using the chain rule,
$$h_{CC}(\boldsymbol{x}) := \arg \max_{\boldsymbol{y}} p(y_1|\boldsymbol{x}) \prod_{k=2}^{K} p(y_k|\boldsymbol{x}, y_1, \ldots, y_{K-1})$$
and show that it is indeed possible to make a Bayes-optimal search with guarantees to the optimal solution for 0/1 loss. Several search techniques exist to make the seach optimal, but greedy is still popular.

- order and structure of chains in cc is the main focus point.
- although in theory the chain rule holds regardless of the order of variables, each $p(y_k|\boldsymbol{x}, y_1, \ldots, y_{K-1})$ is only an approximation of the true probability because it is modelled from finite data under a constrainded class of model, and consequently a different indexing of labels can lead to different results in practice.
- many approaches try to find the best order and show better empirical results, but the reason why is not quite clear
- LP can be viewed as modelling the joint probability directly,

$$h_{LP}(\boldsymbol{x}) := \arg\max_{\boldsymbol{y}} p(\boldsymbol{y}, \boldsymbol{x})$$

- two main points from previous papers: (1) the best label order is impossible to obtain from observational data only. (2) the high performance of classifier chains is due to leveraging earlier labels in the chain as additional feature attributes.

The role of label dependence in multi-label classification

- marginal dependence: frequency of co-occurence among labels
- conditional dependence: after conditioning on the input
- modelling complete dependence is intractable
- rather attempt pairwise marginal dependence or use of ensemble.
- many new methods do not outperform each other over a reasonable amount of datasets.
- improvements of prediction on standard multi-label datasets reached a plateau (maybe investigate).
- question the logic, if the ground truth label dependence could be known and modelled, multi-label predictive performance would be optimal and therefore as more technique and computational effort is invested into modelling label dependence, the lead of the new methods over BR and other predecessors will widen.
- BR might be underrated
- modelling label dependence is a compensation of lack of training data and one could only assume that given infinite data two separate binary models on labels y_k and y_l could achieve as good performance as one that models them together.
- the 'intuitive' understanding actually seems quite flawed: if we take two labels and wish to tag images with them, the assumption that label dependence is key to optimal multi-label accuracy is analogous to assuming that an expert trained for visually recognising one label will make optimum classifications only if having viewed the classiication of an expert trained on the other label.

- in reality, modelling label dependence only helps when a base classifier behind one or more labels is inadequate.
- depends on the base classifier
- there is no guarantee that an ideal structure based on label dependence can be found at all given any amount of training data.
- see XOR problem
- take the view that BR can perform as well as any other method when there is no dependence among the outputs given the inputs.
- not to say that BR should perform as well as other methods if there is no dependence *detected*. Due to noisy data or insufficient model dependence may be missed or even introduced.
- if a ML method outperforms BR under the same base classifier then we can say that it using label dependence to compensate for the inadequacy in its base classifiers.
- attempt to remove the dependence among the labels
- dependence generated by inadequate base classifiers

Binary relevance as a state-of-the-art classifier

- CC and LP are representative of PT problems. Succesful on many fronts and can be builded on. Still has some drawbacks. Discusses them.
- BR less parameters to tune.
- multi-label classifiers can be comprised of individual binary models that perform equally as well as models explicitly linked together based on label dependence or even a single model that learns labels together (intrinsic label dependence modelling).
- claim this is the case for example and label based metrics. (not what the previous paper found)
- proposition with proof: given $X = x$, there exists a classifier $h'_2(x) \approx \arg\max_{y_2 \in \{0,1\}} p(Y_2|X)$ that achieves at least as small error as classifier $h_2(x) \approx \arg\max_{y_2 \in \{0,1\}} p(Y_2|Y_1, X)$, under loss $L(y_2, \hat{y}_2) = I(y_2 \neq \hat{y}_2) = I(y_2 \neq h_2(x))$. Instances of X, Y_1, Y_2 are given in the training data but only $\tilde{x}$ is given at test time. (see proof in paper)
- This means that if we are interested in a model for any particular label, best accuracy can be obtained in ignorance of other labels.
- proposition and proof: under observations $X = x$, there exists two individually constructed classifiers $h'_1 \approx \arg\max_{y_1} p(Y_1|X)$ and $h'_2 \approx \arg\max_{y_2} p(Y_2|X)$ such that under 0/1 loss, $[h_1(x), h_2(x)] \equiv \hat{\boldsymbol{y}} \equiv \boldsymbol{h}(x)$ are equivalent, where $\boldsymbol{h} \approx \arg\max_{[y_1,y_2]} p(Y_1, Y_2|X)$ models labels together. Instances of X, Y_1, Y_2 are given in the training data but only x (tilde) is given at test time. (see proof in paper)
- following examples, X represents some document and Y_1, Y_2 represent the relevance of two subject categories for it. Latent variable Z represents

the unobservable current events which may affect both the observation X and the decisions for labelling it. (illustration of all of the scenarios)

- ignore case where input and all labels are independent.
- case of conditional independence - a text document is given independently to two human labelers who each independently identify if the document is relevant to their expert domain.

$$\begin{aligned} p(\boldsymbol{y}, x) &= p(y_1, y_2) \\ &= p(y_1|x)p(y_2|y_1, x) \\ &= p(y_1|x)p(y_2|x) \end{aligned}$$

which obviously can be solved with BR, where $h_k(\tilde{x}) := \arg\max_{y_k} p(y_k|\tilde{x})$.

- a text document is labelled by the first labeller and afterwards by the second expert - potentially biasing the decision to label relevance or not with this second label. If we do not impose any restriction on any $h_k(x)$, it is straightforward to make some latent $z \equiv h_1(x)$ such that $h_2(x, z) \equiv h_2(x, h_1(x))$. We speak of equivalence in the sense that given Z we can recover Y_2 to the same degree of accuracy (probably compared to case without Z). In this analogy the second labeller must learn also the first labeller's knowledge and thus makes the first labeller redundant. If we drop Y_1 we return to the original structure.
- two experts label a document X but both are biased by each other and - possibly to alternate degrees - by an external source of information Z. Can also introduce latent variables Z_1, Z_2 to break the dependence between the labels.
- note the dependence between any variable can be broken by introducing hidden variables not just the label variables. Hence we can further break dependence between X and Y_1 in the same way - if we desire.
- universal approximation: with a finite number of neurons, even with even with a linear output layer, a network can approximate any continuous function. Implies for ML - given a large enough but finite feature representation in the form of a middle layer, any of the labels can be learned independently of the others, *i.e.* a linear BR layer can suffice for optimal classification performance.
- to summarise: if we find dependence between labels it can be seen as a result of marginalizing out hidden variables that generated them. Also, we can add hidden variables to remove the dependence between labels.
- this does not mean we have a method to learn this structure. Which is learning latent variables powerful enough.
- EM and MCMC sampling under energy models to learn latent variables by minimizing the energy and thus maxmimizing the joint probability with observed variables. (iterative procedures).
- unsupervised part more difficult than supervised

- **existing methods to obtain conditional independence among labels.**
- task: making outputs independent of each other by using a different input space to the original such that a simpler classifier can be employed to predict outputs.
- deep learning to learn a powerful higher-level feature representations of the data. (uses multiple hidden layers)
- in MLC the labels can be seen as high-level feature representations.
- **the equivalence of loss metrics under independent outputs**
- if outputs are independent of each other given the input, then minimizing Hamming loss and 0/1 loss is equivalent.
- the risk of Hamming loss is minimized by BR

  $$\hat{y}_k = \arg \max_{y_k \in \{0,1\}} p(y_k|\boldsymbol{x})$$

  for each label. The 0/1 loss on the other hand, is minimized by taking the mode of the distribution,

  $$\hat{\boldsymbol{y}} = \arg \max_{\boldsymbol{y} \in \{0,1\}^K} p(\boldsymbol{y}|\boldsymbol{x})$$

  equivalently written as

  $$\hat{\boldsymbol{y}} = \arg \max_{\boldsymbol{y} \in \{0,1\}^K} p(y_1|\boldsymbol{x}) \prod_{k=2}^{K} p(y_k|\boldsymbol{x}, y_1, \ldots, y_{K-1}).$$

- Noting that when all outputs are independent of each other given the input ($p(y_k|\boldsymbol{x}, y_l) \equiv p(y_k|\boldsymbol{x})$), then for all k, l it becomes

  $$\begin{aligned}\hat{\boldsymbol{y}} &= \arg \max_{\boldsymbol{y} \in \{0,1\}^K} \prod_{k=1}^{K} p(y_k|\boldsymbol{x}) \\ &= \left[\arg \max_{y_1 \in \{0,1\}} p(y_1|\boldsymbol{x}), \ldots, \arg \max_{y_K \in \{0,1\}} p(y_k|\boldsymbol{x}) \right].\end{aligned}$$

- here input refers to the input into the model and not the original features.
- we can replace the input with hidden variables derived from the original feature space in order to make them independent. If this is successful, the above holds, and using BR will achieve the same result as CC on either measure.
- suppose only the third of three outputs is successfully made independent, then prediction of independent models is optimizing

  $$\hat{\boldsymbol{y}} = \left[\arg \max_{y_1, y_2 \in \{0,1\}^2} p(y_1, y_2|\boldsymbol{x}), \arg \max_{y_3 \in \{0,1\}} p(y_3|\boldsymbol{x}) \right].$$

- if this is the case it could be handled elegantly by RAkELd - disjoint labelset segmentations RAkEL. But detecting these mixed dependence sets is difficult.

- RAkEL and ECC benefit from the ensemble effect of reducing variance of estimates but it is not clear what loss measure is being optimized.

Classifier chains augmented with synthetic labels (CCASL)

- difficult to search for good order in CC
- if 'difficult' label is at start of chain, all other labels may suffer.
- present a method that adds synthetic labels to the beginning of the chain and builds up a non-linear representation, which can be leveraged by other classifiers further down the chain. CCASL
- create H synthetic labels.
- many options - they used threshold linear unit (TLU) to make binary, can also try others like ReLU with continuous output. or sigmoid and radial basis.
- the synthetic labels can be interpreted as random cascaed basis functions, except that at prediction time the values are predicted and thus we refer to them as synthetic labels.
- synthetic label $z_k = I(a_k > t_k)$ with activation values

$$a_k = \left([B * W]^T_{k,1:(p+(k-1))} \cdot \boldsymbol{x}'_k \right)$$

  where W is a random weight matrix (sampled from multivariate normal) with identically sized masking matrix B where $B_{i,j} \sim Bernoulli(0.9)$, input $\boldsymbol{x}'_k = [x_1, \ldots, x_p, z_1, \ldots, z_{k-1}]$ (not the same k as label index), and threshold $t_k \sim \mathcal{N}(\mu_k, \sigma_k \cdot 0.1)$
- want to use synthetic labels at beginning of chain to improve prediction of the real labels.
- $\boldsymbol{y}' = [z_1, \ldots, z_H, y_1, \ldots, y_K]$ and from the predictions $\hat{\boldsymbol{y}}'$ we extract the real labels $\hat{\boldsymbol{y}} = [\hat{y}'_{H+1}, \ldots, \hat{y}_{H+K}] = [\hat{y}_1, \ldots, \hat{y}_K]$.
- $\hat{y}_j = \arg\max_{y_j \in \{0,1\}} p(y_j | x_1, \ldots, x_p, z_1, \ldots, z_H, y_1, \ldots, y_{j-1})$
- use LR as base classifier
- label order less of an issue.
- does well on complex non linear synthetic data - overfits on simple linear synthetic data.
- lots of tunable parameters
- few hidden labels are necessary for CCASL, empirical suggests $H = K$.
- **CCASL + BR**
- guards against overfitting, removes connections among the output
- advantages of BR, stacking and CC
- no back prop necessary.
- **CCASL+AML**
- CCASL strucutre is powerful for modeling non-linearities. CCASL+BR regularizes but otherwise does not offer a more powerful classifier.
- whereas we created synthetic labels from feature space, we can do the same from the label space.

- layer of binary nodes which are feature functions created from the label space for each subset
- see rest in paper.
- section on other network based literature
- back prop bad
- simply using a powerful non-linear base classifier may remove the need for transformations of the feature space altogether.

Experiments

- done in python and sklearn
- synthetic dataset and music, scene, yeast, medical, enron, reuters (max K = 103)
- 10 iterations for each datset 60/40 split
- report parameters
- all out-perform BR and CC
- BR_{RF} does best under hamming loss! RF are adequately powerful to model each layer
- CCASL are quite expensive
- the main advantage brought by modelling label dependence via connections among outputs is that of creating a stronger learner.
- did not investigate ensembles

It has been shown repeatedly in the literature that in order to achieve acceptable empirical results, the multi-label algorithm used must in some way or another exploit the dependence/correlation amongst the labels. Unfortunately very little theoretical evidence exists for this suggestion. To delve deeper into this topic an understanding of the evaluation metrics of multi-label classifiers is a fundamental step.

> not sure if this should go here and to what extent. This is only an introduction and the rest will be continued after algorithms are introduced.

It has been mentioned here and many times in literature that the exploitation of label structures is essential to an effective multi-label algorithm. The problem is that the correlation/dependence/relationship between labels is not yet well defined in the literature (**?**). In (**?**) the authors comment that researchers often use the term label dependence in an intuitive sense and not as a formally defined concept. Naturally this makes it a hard problem to solve, if it is not well defined. Some valiatnt attempts were made in (**?**), (**?**) (and others). The following is an overview of them.

In (**?**) the existing strategies for multi-label classification are divided into categories based on the order of label correlations being considered by the algorithms. So-called first-order approaches are those that do not take

label correlations into account. Second-order approaches consider the pairwise relationships between labels and high-order approaches allows for all interactions between labels and/or combinations of labels. First-order strategies simply ignore label correlations, but they are usally simpler. The latter two strategies are far more complex but also limited in some cases. Second-order strategies will not generalise well when higher-order dependencies exist amongst the labels and the the high-order strategies may 'overfit' if only subgroups of the labels are correlated (**?**).

From the Bayesian point of view, the problem of multi-label learning can be reduced to modeling the conditional joint distribution of $P(\boldsymbol{y}|\boldsymbol{x})$. This can be done in various ways. First-order approaches solve the problem by decomposing it into a number of independent task through modelling $P(y_k|\boldsymbol{x})$, $k = 1, \ldots, K$. Second-order approaches solve the problem by considering interactions between a pair of labels through modelling $P((y_k, y_{k'})|\boldsymbol{x})$, $k \neq k'$. High-order approaches solve the problem by adressing correlations between a subset of labels through modelling $P((y_{k_1}, y_{k_2}, \ldots, y_{k_{K'}})|\boldsymbol{x})$, $K' \leq K$. Our goal is to find a simple and efficient way to improve the performance of multi-label learning by exploiting the label dependencies (**?**). Propose LEAD approach.

- (**?**) use the ϕ coefficient to estimate label correlations.
- (**?**)
- mention the holy grail comment
- comment on what 'exploitation' means. Since many authors claim that exploiting label dependence structures is the only way to effectively handle multiple labels, I would assume this means that we can make use of label correlations to spare time and increase accuracy.
- we need to think about how observations are labelled, when will it be useful to take label dependence into account and how.
- Such a solution, however, ne- glects the fact that information of one label may be help- ful for the learning of another related label; especially when some labels have insufficient training examples, the label correlations may provide helpful extra information (**?**)

4.6.5 Symmetry

- (**?**) claims that most of the time the label dependencies are asymmetric and suggest the MAHR algorithm. Also most of the existing methods exploit label correlations globally, which is not necessarily a good assumption if these correlations only exist for some instances (**?**). They suggest a ML-LOC algorithm (which seems to do very well).

4.6.6 Locality

- is local the same as conditional? and global unconditional?
- (?)

Existing approaches to exploiting label correlations either assume the the label correlations are global and shared by all instances, or that the label correlations are local and shared only by a subset of the data. It may be that some label correlations are globally applicable and some share only in a local group of observations.

- give example
- mention GLOCAL (?)
- (?)

Existing approaches typically exploit label correlations globally by assuming that the label correlations are shared by all observations. In the real-world, however, different observations may share different label correlations and few correlations are globally applicable.

- propose ML-LOC approach
- mentions that by assuming global correlations may be hurtful to the performance (?) in empirical discussion
- maybe meta analysis on how others claimed to improved label correlation modelling

4.7 Problem Transformation Approaches

Problem transformation methods consist of first transforming the multi-label problem into one or more single-label problem(s) and then fitting any standard supervised learning algorithm(s) to the single-label data. For that reason, problem transformation methods are called algorithm independent, i.e. once the data is transformed, any single-label classifier can be used (?).

The two main problem transformation algorithms are the binray relevance and label powerset transformations. Both methods suffer from several limitations but they form the basis of arguably any problem transformation method. The state-of-the-art problem transformations algorithms are most of the times extensions of either the standard binary relevance or label powerset algorithms (?). Therefore the understanding of these two basic methods are crucial in dealing with the more complex, modern problem transformation methods.

4.7.1 Binary Relevance

- basic idea
- notation
- cross-training
- T-criterion for avoiding empty prediction
- psuedo-code
- remarks: first-order; parallel; straightforward; building block of state-of-the-art; ignores potential label correlations; may suffer from class-imbalance; computational complexity

The most common transformation method is binary relevance (BR). BR transforms the mutli-label into K single-label problems by modelling the presence of the labels separately. Typically K single-label binary data sets, $D_k = (X, \boldsymbol{Y}_k)$ for $k = 1, ..., K$, would be constructed from the multi-label data set, $D = (X, Y)$. To each D_k any single-label classifier can be applied. In the end, predictions $\hat{\boldsymbol{Y}}_1, ..., \hat{\boldsymbol{Y}}_K$ are obtained separately which can then be combined to allocate all the predicted relevant variables to each instance. Note, that it may occur that all of the single-label learners produces zeroes, which would imply that the instance belongs to an empty set. To avoid this (**?**) suggests following the T-criterion rule. The rule states, briefly, that in such a case the labels associated with the greatest output should be assigned to the instance. Clearly, this will only work if the base learners used gives continuous outputs and it will only make sense if all the base learners are of the same type. I suppose these rules are ad-hoc and I can think of alternatives.

> With this approach the standard single label feature selection procedures can be applied. The relevant subset of features can be identified for each label. This is convenient since it is not unlikely that the optimal subset of features will differ from label to label.

The biggest drawback for this approach is that it models each label separately and ignores the possible correlations between labels. Thus BR assumes that there are no correlations between the labels. However, these correlations can be very helpful in predicting the labels present. This is a first-order strategy. Also it can be time consuming since data sets with hundreds of labels is not rare. This would mean more than a hundred models should be fit and tuned separately. But this complexity scales linearly with increasing K, which is actually not so bad when comparing to other multi-label algorithms. Grouping the labels in a hierachical tree fashion may become useful when K is very large (**?**) (see also Incorporating label dependency into the binary relevance framework for multi-label classification by the same authors).

Another argument against BR from (**?**): The argument is that, due to this information loss, BR's predicted label sets are likely to contain either too many or too few labels, or labels that would never co-occur in practice.

Advantage of BR by (**?**): Its assumption of label independence makes it suited to contexts where new examples may not necessarily be relevant to any known labels or where label relationships may change over the test data; even the label set L may be altered dynamically - making BR ideal for active learning and data stream scenarios.

Nevertheless, BR remains a competitive ML algorithm in terms of efficiency and efficacy, especially when minimising a macro-average loss function is the goal (**?**). The most important advantage of BR is that it is able to optimise several loss functions (**?**) also see small proof. They also show empirically that BR tends to outperform ECC when there are many labels, high label dependency and high cardinality, i.e. when the multi-label data becomes more complicated.

Compared to label powerset (LP) which will be discussed later, BR is able to predict arbitrary combinations of labels (**?**) not restricted only to those in the training set.

(**?**) also proposes a variation of BR called BR+. Its aim is to keep the simplicity of BR but also to consider the possible label correlations. It does so by also creating K binary data sets but this time each of these data sets treat all the label columns not to be predicted by the current single-label classifier as features to the classifier. Thus each sinlge-label classifier will have $p + K - 1$ inputs. So now when predicting label l, all of the original features in X and the remaining variables $\boldsymbol{Y}_k$, $k \neq l$, are used as inputs for classifier l. (second order strategy?)

The problem arises when predicting unseen instances for which the labels are unknown. Thus the input needed for each binary classifier is not available. One workaround is to obtain an initial prediction of the labels using an ordinary BR approach and then using these predictions as inputs to the BR+ algorithm. The BR+ algortihm will most likely produce different predictions to the initial predicitons or BR which can then also be used in a next round of BR+. These steps can be continued until convergence but this seems like the classifier chains approach. (to be investigated).

(**?**) mentions the 2BR strategy that seems very similar/identical to BR+. They describe the 2BR method as follows: first train a binary classifier on each of the K binary data sets and then use their predictions (and or probabilities) as so called meta-features for a second round of BR. They mention that it might be better to train the base and meta learners on separate parts of the training data to avoid biased predictions. They suggest using a cross-validation approach for both learners to also avoid size constraints of the training data. They describe this approach as a stacked generalisation, also mentioned in (**?**), (**?**), (**?**) calls it classifier fusion.

The adding of all the base learner predicitions as meta-feature to the meta-learners is not necessarily desirable. Some label pairs might have no correlation and adding predictions for those labels as inputs to the meta-learner will add noise to the model and waste computation time. (**?**) suggests a solution called

corerlation-based pruning. They calculate the pairwise correlations between labels, ϕ, and only add base learner prediction of label i as a meta-feature to meta-learner j if ϕ_{ij} is greater than some threshold. In this way only label-pairs that are highly correlated will be used in the final prediction of each other.

- BR performs well for Hamming loss, but fails for subset 0/1 loss.
- It is not clear, in general, whether the meta-classifier b should be trained on the BR predictions h(x) alone or use the original features x as additional inputs. Another question concerns the type of information provided by the BR predictions. One can use binary predictions, but also values of scoring functions or probabilities, if such outputs are delivered by the classifier @(**?**).

4.7.2 Label Powerset

4.7.3 Classifier Chains

- basic idea
- notation
- importance of ordering
- ECC brief explanation
- psuedo-code
- remarks: high-order; considers label correlations in a random manner; not parallel; computational complexity

Another extension of BR, similar to 2BR and BR+, is the classifier chains (CC) approach introduced by (**?**). It also consists of transforming the mutli-label data set D to K single-label data sets but the transformations are done sequentially in the sense that the label previously treated as a response will be added as a feature for predicting the next label. This will give data sets similar to $D_1 = (X, \boldsymbol{Y}_1), D_2 = (X, \boldsymbol{Y}_1, \boldsymbol{Y}_2), ...D_K = (X, \boldsymbol{Y}_1, \boldsymbol{Y}_2, ..., \boldsymbol{Y}_K)$, where the last column of each is the response that needs to be predicted. To each of these single-label data sets a classifier can be trained and then their predictions are combined in the same fashion as BR. CC keeps the simplicity of BR but has that additional capacity to model label dependencies by passing label information between classifiers. This should raise the question of what order of labels should the chain consist of and should it stop after one cycle?

In a response to this, the ensembles of classifier chains (ECC) was suggested by (**?**). Here the term ensemble refers to an ensemble of multi-label classifiers instead of an ensemble of binary classifiers already mentioned before. ECC trains m classifier chains, each with a random chain ordering and a random subset of instances. These parameters of ECC contributes to the uniqueness of each classifier chain which helps with variance reduction when their predictions are combined. These predictions are summed by label so that each label receives

a number of votes. A threshold is used to select the most popular labels which form the final predicted multi-label set (**?**) (copied from). More details still to cover in article.

CC and ECC has an advantage over the ensemble methods of BR, that it is not necessary for an initial step of training to obtain predictions of labels that can later be used as features, it does this simultaneously.

Paper still need to look at for CC (**?**).

4.8 Algorithm Adaption Approaches

These are methods tackling the multi-label learning task by adapting, extending and/or customising an existing supervised learning algorithm (**?**).

The main weakness of algorithm adaption methods is that they are mosty tailored to suit a specific model, whereas problem transformation methods are more general and allows for the use of many well-known and effective single-label models (**?**) (algorithm independent).

4.9 Ensemble Approaches

- Ensembles are well known for their effect of increasing overall accuracy and overcoming over-fitting, as well as allowing parallelism. The main idea behind ensembles is to exploit the fact that different classifiers may do well in different aspects of the learning task so combining them could improve overall performance. Ensembles have been extensively used in literature [13] with stacking [14], bagging [15] and boosting [16] being the main methods employed. In the context of multi-label problems, [17] proposes a fusion method where the probabilistic outputs of heterogeneous classifiers are averaged and the labels above a threshold are chosen. Copied from (**?**) (can maybe use to explain why these methods perform better and not because of label dependence)
- evidence of stacking working (**?**). Read conclusions chapter. Ensembling effective. Linear models good for text classification. Thresholding important.

4.9.1 Ensemble of Classifier Chains

4.9.2 Random k-Labelsets

As mentioned before, the LP method has the advantage of taking label correlations into account but typically suffers from a huge class imbalance problem. (**?**) suggested the Random k-labelsets (RAKEL) algorithm to overcome the drawbacks of LP while still being able to model label dependencies. RAKEL is simply an ensemble of LP classifiers, but the LP classifiers are trained on

different subsets of the labelset. The author defined a k-labelset as a set $Y \subseteq L$ with $k = |Y|$, where L is the complete labelset and $|Y|$ the size of the set, Y. Let L^k denote the set of all distinct k-labelsets on L. The size of L^k can thus be given by $|L^k| = \binom{|L|}{k}$.

First, the RAKEL algorithm iteratively constructs m LP classifiers. At each iteration, $j = 1, 2, \ldots, m$, it randomly selects a k-labelset, Y_j, from L^k without replacement, and then learns the classifier $h_j : X \to P(Y_j)$ (review notation). For classifying an instance, x, each model, h_j, provides binary decisions, $h_j(x, \lambda_l)$ for each label λ_l in k-labelset Y_j. The average of these binary decisions are then computed and a final prediction for a label is given if its corresponding average is bigger than some threshold t. Note, the average for label λ_l is not calculated by the sum of $h_j(x, \lambda_l)$ divided by m, but by instead dividing by the number of times λ_l was in Y_j for $j = 1, \ldots, m$.

The values m, k and t, are all parameters to be specified by the user. Clearly, k can only lie between 1 and $|L|$, where if $k = 1$, the algorithm is equivalent to the BR approach, and if $k = |Y|$, the algorithm is equivalent to the LP approach. In the original paper, the author showed empirically that by using small labelsets and an adequate number of iterations, RAKEL will manage to model label correlations effectively. An intuitive value for t would be 0.5, however, in the same paper, it is shown that RAKEL performs well over a wide range of values for t.

A concern might be the number of classes, 2^k that each LP classifiers must deal with. In practice, each LP classifier deals with a much smaller subset of label combinations, since it can only model combinations that exist in the training set. Also, RAKEL is preferred to LP when there are a large number of labels. In this case, RAKEL would only need to model a subset of 2^k possible label combinations compared to LP that needs to model a much larger subset of $2^{|Y|}$ possible label combinations.

In (**?**) it is shown that RAKEL outperforms LP and BR on 3 benchmark datasets with numerous configurations. The author concluded that the randomness of the RAKEL algorithm might not be the best ensemble selection approach since it may lead to the inclusion of models that affect the ensemble's performance in a negative way. Continue with papers that improve on this idea.

- something on (re)sampling

4.10 Spatial Regularization Networks

- https://arxiv.org/pdf/1702.05891.pdf

4.11 From Single to Multi Output Paper ()

4.12 RNN-CNN paper ()

4.13 Direct binary Embedding

- https://arxiv.org/pdf/1703.04960.pdf

4.14 Another ML architecture

- https://arxiv.org/pdf/1609.07982.pdf
- replace last softmax with sigmoid
- replace last FC with maxout
- replace last pooling with spatial pyramid pooling
- other option use FCN with global max pool at end before sigmoid
- dropout only on first layers after representation (fixed VGG)
- note they also used dropout at testing and then combined for mean prediction

4.15 Is object localization for free? – Weakly-supervised learning with convolutional neural networks

- http://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Oquab_Is_Object_Localization_2015_CVPR_paper.pdf
- section on label correlations for MLC

4.16 Near Perfect Protein Multi-Label Classification with Deep Neural Networks

- https://arxiv.org/pdf/1703.10663.pdf
- spp layer after convolutions - not sure if it has to do with MLC

4.17 BP-MLL

- http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.130.7318&rep=rep1&type=pdf
- description on NN and some references

4.18 ML NN for text

- https://arxiv.org/pdf/1312.5419.pdf

4.19 ML MT

- http://www.cripac.ia.ac.cn/irds/People/lwang/M-MCG/Publications/2013/YH2013ICIP.pdf
- but not much detail

4.20 ML Attention:

- https://arxiv.org/pdf/1412.7755.pdf

4.21 Sparsemax ML loss

- https://arxiv.org/pdf/1602.02068.pdf
- Other object detection
- winner of yt8m challenge: https://arxiv.org/pdf/1706.06905.pdf (**?**)

Chapter 5

Remote Sensing

Wiki: In current usage, the term "remote sensing" generally refers to the use of satellite- or aircraft-based sensor technologies to detect and classify objects on Earth, including on the surface and in the atmosphere and oceans, based on propagated signals (e.g. electromagnetic radiation).

The idea is not to review the domain in detail, but rather to discuss the state-of-the-art Deep Learning approaches to Remote Sensing or related problems.

- (?)
- remote sensing images are easily accumulated but the labeling process is most of the infeasible.
- most applications in single-label framework.
- graph-based multi-label classification approach called Multi-Label Classification based on Low Rank Representation for image annotation (MLC-LRR) has been proposed.
- can effectively capture global label correlation.
- take advantage of limited labeled images and abundant unlabeled images
- does very well

5.1 https://arxiv.org/pdf/1706.01171.pdf

- recent years DL have made a breakthrough for satellite image analysis
- land cover classification
- hyperspectral image analysis
- synthetic aperture radar
- large datasets of satellite images with high quality labels are not easily available
- most works use pre trained
- use early and late fusion of rgb images and local binary patterns

- We design deep models by constructing a two-stream deep architecture where texture coded mapped images are used as a second stream and fuse it with the standard RGB stream
- only single label

5.2 Do Deep Features Generalize from Everyday Objects to Remote Sensing and Aerial Scenes Domains?

- http://www.cv-foundation.org/openaccess/content_cvpr_workshops_2015/W13/papers/Penatti_Do_Deep_Features_2015_CVPR_paper.pdf
- the use of deep learning is rapidly growing
- still no evaluation of pretrained convnets in aerial and remote sensing domains
- contribution: evaluation of the generalisation power of convnets from everyday objects to aeriaal and remote sensing domain, comparitive evaluation of global descriptors, correlation analysis among different convnets
- also used fusing
- convnet features the best for aerial
- ACC and BIC best for Coffee Scene
- depends on intrinsic properties of data
- fusion can improve
- looks like fine tuned conv layers aswell
- only output of penultimate layer of pretrained network is used

5.3 Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery

- http://www.mdpi.com/2072-4292/7/11/14680/htm
- can we transfer the successfully pre-trained CNNs to address HRRS scene classification, which is also a typical recognition task with limited amount of training data?
- test two feature extraction from CNN: fc layer before classification, last conv layer with different scaled inputs
- also contribute comparison of pretrained convnets (vgg's, alex, caffe, places)

- lack of investigations on using CNNs for HRRS
- did not pre train convnet feature extractors
- center + corners with horizontal flips augmentaion for testing. Did this to input images and averaged outputs for ultimate feature extraction (shows empirically that it helped)
- for fc feautes, performed better after relu
- Although the features of FC layers capture global spatial layout information, they are still fairly sensitive to global rotation and scaling, making them less suitable for HRRS scenes that greatly differ in orientation and scales.
- not sure how they used BOW on multi scale conv features
- show multi scale does better (wonder if this will be true if skip connections)
- their methods not better than googlenet + finetune
- conv features more suitable than fc, both better than low level

5.4 Land Use Classification in Remote Sensing Images by Convolutional Neural Networks

- https://arxiv.org/pdf/1508.00092.pdf
- pretrain nets (caffe and google) + fine tune
- compare training from scratch, fine tuning and feautre vector
- for UC-Merced: fine tuning got best results and then feature vector
- better than handcrafted features
- for brazillian coffee: from scratch and fine tune were very similar

5.5 Learning Low Dimensional Convolutional Neural Networks for High-Resolution Remote Sensing Image Retrieval

- https://arxiv.org/pdf/1610.03023.pdf
- Investigate how to extract deep feature representations based on convolutional neural networks (CNN) for high-resolution remote sensing image retrieval (HRRSIR)
- compare feature extraction from fc and conv from pre trained, vs training novel cnn on large scale remote sensing and then features are used.
- Nothing newer than vgg
- ambiguous results with fc layer with + without relu
- both better than handcrafted featuers
- feature aggregation methods used (figure out)

- the custom pretrained scheme performed better when the datasets were similar

5.6 What do We Learn by Semantic Scene Understanding for Remote Sensing imagery in CNN framework?

- https://arxiv.org/pdf/1705.07077.pdf
- Compared with object recognition, scene understanding not only needs to identify the targets, it should also understand the distribution of targets in a scene
- model depth acts differently on on different classes
- more complex scenes require greater depth
- multi scale works better for complex scenes

5.7 Unsupervised Geometric Learning of Hyperspectral Images

- https://arxiv.org/pdf/1704.07961.pdf
- way of segmenting hyperspectral images withoud supervision

5.8 Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale

- https://arxiv.org/pdf/1704.02965.pdf
- good survey and resources on other land classification projects
- pretrained model on image net, fine tune on urban atlas and deepsat
- mostly urban environments
- interesting figure on how the area trained on makes a difference in accuracy on other areas
- shows bigger patch size leads to better accuracy

5.9 Using Satellite Imagery for Good: Detecting Communities in Desert and Mapping Vaccination Activities

- https://arxiv.org/pdf/1705.04451.pdf

- interesting insight under implementation section
- better to train on own data than pretrain
- interesting method of adding more data: after n epochs test model on a set, those that were wrong are added to training (unclear on specifics)
- FCN worked better for them, batchnorm didnt
- relu did not work

5.10 Solar Power Plant Detection on Multi-Spectral Satellite Imagery using Weakly-Supervised CNN with Feedback Features and m-PCNN Fusion

- https://arxiv.org/pdf/1704.06410.pdf
- novel architecture

5.11 Hard Mixtures of Experts for Large Scale Weakly Supervised Vision

- https://arxiv.org/pdf/1704.06363.pdf
- MOE for MLC

5.12 To read:

- Vehicle detection in satellite images by hybrid deep convolutional neural networks,
- Unsupervised deep feature extraction for remote sensing image classification,
- Detecting man-made structures and changes in satellite imagery with a content-based information retrieval system built on self-organizing maps
- Deep learning-based classification of hyperspectral data
- Spectralspatial classification of hyperspectral data based on deep belief network,
- Multiclass feature learning for hyperspectral image classification: Sparse and hierarchical solutions
- High-resolution sar image classification via deep convolutional autoencoders

- Deep learning earth observation classification using imagenet pretrained networks
- Representation learning for contextual object and region detection in remote sensing
- Feature learning based approach for weed classification using high resolution aerial images from a digital camera mounted on a uav
- Saliency-guided unsupervised feature learning for scene classification
- Unsupervised deep feature extraction of hyperspectral image
- Deep model for classification of hyperspectral image using restricted boltzmann machine
- Unsupervised feature learning for aerial scene classification
- High-resolution remote-sensing imagery retrieval using sparse features by auto-encoder
- Visual descriptors for content-based retrieval of remote sensing images
- Classification and Segmentation of Satellite Orthoimagery Using Convolutional Neural Networks
- Convolutional Neural Network Based Automatic Object Detection on Aerial Images
- Land Use and Land Cover Classification Using Deep Learning Techniques
- Benchmarking Deep Learning Frameworks for the Classification of Very High Resolution Satellite Multispectral Data. !!!
- Learning Multi-Scale Deep Features for High-Resolution Satellite Image Classification.
- DeepSat - A Learning framework for Satellite Imagery
- https://www.cs.toronto.edu/%7Evmnih/docs/Mnih_Volodymyr_PhD_Thesis.pdf (hinton student thesis), implementation: https://github.com/mitmul/ssai-cnn
- Remote Sensing Image Scene Classification: Benchmark and State of the Art
- [Karalas2015]
- [Karalas]

- Semantic annotation of high-resolution remote sensing images via Gaussian process multi-instance multilabel learning
- deepsat for data: https://github.com/trailbehind/DeepOSM, guide at: https://github.com/developmentseed/skynet-data

Chapter 6

Results (/Application)

In this chapter, some of the theoretical reccommendations will be evaluated empirically, with the goal to find the best approaches for this dataset. The full end-to-end workflow will be discussed and intermediate results also reported.

(Preliminary) The network will consist of two parts: a feature extraction part and a classification part. Basically a fine-tuning framework. The feauture extractor will be a popular Convolutional Neural Network (CNN) pretrained on ImageNet, *e.g.* VGG, ResNet, DenseNet, Inception. The classification part is the mapping from the extracted features to the class scores. This can be done in various ways and will be experimented with, for example: a combination of BR output and LP output, classification head for each label or Rakel output. The weights of the whole classification part will be trained on the Amazon dataset and, in some cases, the last layers of the feature extraction part.

6.1 Comparison of Pretrained Networks for Transfer Learning

- VGG16
- VGG19
- ResNet50
- DenseNet
- Inception

6.2 Experimentation with Classification Heads

- separate vs combined
- BR vs LP vs CC?
- Rakel?
- combo of BR and LP

- FCN
- spp: https://github.com/yhenon/keras-spp

6.3 Sampling and Resampling

- Simulating (**?**) (also gives citations to other papers)
- partitioning mentioned in (**?**) - referred to (**?**)
- (**?**) Therefore they created a ML data generator to simulate ML data on which algorithms can be evaluated.

6.4 Class Imbalance

- https://www.reddit.com/r/MachineLearning/comments/6iq5i8/d_what_are_your_favorite_ways_for_dealing_with/
- (**?**)

6.5 Data Augmentation

- elastice transform: https://pdfs.semanticscholar.org/7b1c/c19dec9289c66e7ab45e80e8c422735
pdf

6.6 Pseudo-Labelling

6.7 Ensembling

Chapter 7

Things that need a place:

- challenges for image classification: (maybe in CNNs in practive)
  - http://cs231n.github.io/classification/
- Feature learning
- one-shot learning:
  - https://github.com/sorenbouma/keras-oneshot
  - https://github.com/fchollet/keras/blob/master/examples/mnist_siamese_graph.py
  - https://sorenbouma.github.io/blog/oneshot/
- multi-task learning:
  - https://arxiv.org/abs/1706.05137
- relational learning:
  - https://arxiv.org/pdf/1706.01427.pdf
- AutoML:
  - https://research.googleblog.com/2017/05/using-machine-learning-to-explore.html

Chapter 8

Conclusion

- summary
- contributions
- reccomendations
- limitations
- future work

Appendices

Appendix A

Benchmark Datasets

The progress of areas in machine/statistical learning is highly dependent on the availability of quality and diverse benchmark data sets. This enables researchers to compare their methods in a wide variety of environments. Recently, a decent amount of ML data sets has been published, but not without critique. (**?**) argues that the MULAN[1] ML data set repository does not have data sets that are truely ML and that most of the data sets are very similar to each other. Most of the data sets have low cardinality and low label dependence. The problem with this is that these data sets may not show the true performance of ML algorithms. In (**?**) the authors also comments on the lack of thorough, comparative empirical studies on these benchmark sets.

Some of the most popular and recent ML benchmark data sets will be introduced here along with their properties. This will give us some form of a reference to compare our data set of satellite images against.

- (**?**) defines a complexity measure as $n \times p \times K$
- (**?**) long list of datasets. Other than MULAN: Plant and Human, Slashdot, LangLog, IMDB
- (**?**)
- https://manikvarma.github.io/downloads/XC/XMLRepository.html
- yelp dataset: http://www.ics.uci.edu/~vpsaini/
- also new yt8m

[1] A Java library for ML learning - http://mulan.sourceforge.net/datasets-mlc.html.

Appendix B

Software

Bibliography

Alazaidah, R. and Ahmad, F.K. (2016). Trending Challenges in Multi Label Classification. *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 7, no. 10.
Available at: www.ijacsa.thesai.org

Charte, F., Rivera, A.J., del Jesus, M.J. and Herrera, F. (2015). Addressing imbalance in multilabel classification: Measures and random resampling algorithms. *Neurocomputing*, vol. 163, pp. 1–14. ISSN 09252312.
Available at: http://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/1790{_}2015-Neuro-Charte-MultiLabel{_}Imbalanced.pdfhttp://linkinghub.elsevier.com/retrieve/pii/S0925231215004269

Chekina, L., Rokach, L. and Shapira, B. (2011). Meta-learning for selecting a multi-label classification algorithm. In: *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 220–227. ISBN 9780769544090. ISSN 15504786.

Dembcz, K., Waegeman, W., Cheng, W., Hüllermeier, E., Tsoumakas, G., Zhang, M.-L., Zhou, Z.-H., Dembczyski, K., Waegeman, W., Cheng, W. and Hüllermeier, E. (2012). On label dependence and loss minimization in multi-label classification. *Mach Learn*, vol. 88, pp. 5–45.

Dembczy, K. (2010). Bayes Optimal Multilabel Classification via Probabilistic Classifier Chains. *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 279–286.
Available at: http://machinelearning.wustl.edu/mlpapers/paper{_}files/icml2010{_}DembczynskiCH10.pdfhttp://www.uni-marburg.de/fb12/kebi/people/cheng/cheng-icml10c.pdf

Dembczynski, K., Waegeman, W., Cheng, W. and Hüllermeier, E. (). On Label Dependence in Multi-Label Classification.

Dembczyński, K., Waegeman, W., Cheng, W. and Hüllermeier, E. (2010). Regret analysis for performance metrics in multi-label classification: The case of hamming and subset zero-one loss. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6321 LNAI, pp. 280–295. ISBN 364215879X. ISSN 03029743.
Available at: https://biblio.ugent.be/publication/1155381/file/1210780.pdf

Dembczynski, K., Waegeman, W. and Hüllermeier, E. (2012). An analysis of chaining in multi-label classification. In: *Frontiers in Artificial Intelligence and Applications*, vol. 242, pp. 294–299. ISBN 9781614990970. ISSN 09226389.
Available at: https://biblio.ugent.be/publication/3132158/file/3132170

Gasse, M., Aussem, A. and Elghazel, H. (2015). On the Optimality of Multi-Label Classification under Subset Zero-One Loss for Distributions Satisfying the Composition Property. *ICML*, vol. 37.

Gibaja, E. and Ventura, S. (2014). Multi-label learning: A review of the state of the art and ongoing research. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 6, pp. 411–444. ISSN 19424795.

Gibaja, E. and Ventura, S. (2015). A Tutorial on Multilabel Learning. *ACM Computing Surveys (CSUR)*, vol. 47, no. 3, pp. 52:1—-52:38. ISSN 0360-0300.
Available at: https://www.researchgate.net/profile/Sebastian{_}Ventura/publication/270337594{_}A{_}Tutorial{_}on{_}Multi-Label{_}Learning/links/54bcd8460cf253b50e2d697b.pdfhttp://doi.acm.org/10.1145/2716262

Godbole, S. and Sarawagi, S. (2004). Discriminative Methods for Multi-labeled Classification. *Lecture Notes in Computer Science*, vol. 3056, pp. 22–30. ISSN 03029743. 978-3-540-24775-3{_}5.
Available at: http://link.springer.com/10.1007/978-3-540-24775-3{_}5

Hastie, T., Tibshiranie, R. and Friedman, J.H. (2009). *No Title*. 2nd edn. New York: Springer.

Huang, S.-j., Yu, Y. and Zhou, Z.-h. (2012). Multi-label hypothesis reuse. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, p. 525.
Available at: http://dl.acm.org/citation.cfm?id=2339530.2339615

Koyejo, O.O., Natarajan, N., Ravikumar, P.K. and Dhillon, I.S. (2015). Consistent Multilabel Classification. *Advances in Neural Information Processing Systems*, pp. 3303–3311. ISSN 10495258.
Available at: http://papers.nips.cc/paper/5883-consistent-multilabel-classification

Lee, J. and Kim, D.-W. (2017). SCLS: Multi-label feature selection based on scalable criterion for large label set. *Pattern Recognition*, vol. 66, no. August 2016, pp. 342–352. ISSN 00313203.
Available at: http://ac.els-cdn.com.ez.sun.ac.za/S003132031730016X/1-s2.0-S003132031730016X-main.pdf?{_}tid=72e6d1d6-1573-11e7-a49b-00000aacb361{&}acdnat=1490897305{_}4ff82d83a1296a46c537536304a7e929http://linkinghub.elsevier.com/retrieve/pii/S003132031730016X

Luaces, O., Díez, J., Barranquero, J., Del Coz, J.J. and Bahamonde, A. (). Binary Relevance Efficacy for Multilabel Classification.

Madjarov, G., Kocev, D., Gjorgjevikj, D. and Džeroski, S. (2012). Author's personal copy An extensive experimental comparison of methods for multi-label learning. Available at: http://www.elsevier.com/copyright

Pachet, F. and Roy, P. (2009). Improving Multilabel Analysis of Music Titles: A Large-Scale Validation of the Correction Approach. *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, vol. 17, no. 2.

Read, J. and Hollmen, J. (2014). A Deep Interpretation of Classifier Chains. *Advances in Intelligent Data Analysis Xiii*, vol. 8819, pp. 251–262. ISSN 0302-9743.
Available at: http://jmread.github.io/papers/Read,Holmen-ADeepInterpretationofClassifierChains.pdf

Read, J. and Hollmén, J. (2015). Multi-label Classification using Labels as Hidden Nodes. pp. 1–23. 1503.09022.
Available at: https://arxiv.org/pdf/1503.09022.pdfhttp://arxiv.org/abs/1503.09022

Read, J., Pfahringer, B., Holmes, G. and Frank, E. (2011*a*). Classifier chains for multi-label classification. *Machine Learning*, vol. 85, no. 3, pp. 333–359. ISSN 08856125. arXiv:1207.6324.

Read, J., Pfahringer, B., Holmes, G., Frank, E., Brodley Read, C.J., Pfahringer, B., Holmes, G. and Frank, E. (2011*b*). Classifier chains for multi-label classification. *Mach Learn*, vol. 85, no. 85, pp. 333–359. ISSN 08856125. arXiv:1207.6324.
Available at: http://download.springer.com/static/pdf/44/art{%}253A10.1007{%}252Fs10994-011-5256-5.pdf?originUrl=http{%}3A{%}2F{%}2Flink.springer.com{%}2Farticle{%}2F10.1007{%}2Fs10994-011-5256-5{&}token2=exp=1490608886{~}acl={%}2Fstatic{%}2Fpdf{%}2F44{%}2Fart{%}25253A10.1007{%}25252Fs10994-011-5256-

Sechidis, K., Tsoumakas, G. and Vlahavas, I. (2011). On the Stratification of Multi-label Data.
Available at: http://download.springer.com/static/pdf/229/chp{%}253A10.1007{%}252F978-3-642-23808-6{_}10.pdf?originUrl=http{%}3A{%}2F{%}2Flink.springer.com{%}2Fchapter{%}2F10.1007{%}2F978-3-642-23808-6{_}10{&}token2=exp=1489589222{~}acl={%}2Fstatic{%}2Fpdf{%}2F229{%}2Fchp{%}25253A10.1007{%}25252F978-3-64

Sorower, M.S. (). A Literature Survey on Algorithms for Multi-label Learning.

Sucar, L.E., Bielza, C., Morales, E.F., Hernandez-Leal, P., Zaragoza, J.H. and Larrañaga, P. (2013). Author's personal copy Multi-label classification with Bayesian network-based chain classifiers.

Tan, Q., Liu, Y., Chen, X. and Yu, G. (2017). Multi-Label Classification Based on Low Rank Representation for Image Annotation. *Remote Sensing*, vol. 9, no. 2, p.

109. ISSN 2072-4292.
Available at: http://www.mdpi.com/2072-4292/9/2/109

Tomás, J.T., Spolaôr, N., Cherman, E.A. and Monard, M.C. (2014). A framework to generate synthetic multi-label datasets. *Electronic Notes in Theoretical Computer Science*, vol. 302, pp. 155–176. ISSN 15710661.
Available at: http://ac.els-cdn.com/S1571066114000267/1-s2.0-S1571066114000267-main.pdf?{_}tid=207a475a-25c4-11e7-9d47-00000aacb35d{&}acdnat=1492691174{_}037266698571d8a927f3feb0eb432995

Tsoumakas, G., Dimou, A., Spyromitros, E., Mezaris, V., Kompatsiaris, I. and Vlahavas, I. (2009). Correlation-based pruning of stacked binary relevance models for multi-label learning. *Proceedings of the Workshop on Learning from Multi-Label Data (MLD'09)*, pp. 101–116. ISSN 1475-925X.
Available at: http://www.ecmlpkdd2009.net/wp-content/uploads/2008/09/learning-from-multi-label-data.pdf{#}page=102

Tsoumakas, G. and Katakis, I. (2007). Multi-Label Classification : An Overview. ISSN 1548-3924.

Tsoumakas, G., Katakis, I. and Vlahavas, I. (). A Review of Multi-Label Classification Methods.

Tsoumakas, G. and Vlahavas, I. (). Random k-Labelsets: An Ensemble Method for Multilabel Classification.

Zhang, M.L. and Zhou, Z.H. (2014). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837. ISSN 10414347.

Zhu, Y., Kwok, J.T. and Zhou, Z.-H. (2017). Multi-Label Learning with Global and Local Label Correlation. 1704.01415.
Available at: https://arxiv.org/pdf/1704.01415.pdfhttp://arxiv.org/abs/1704.01415