Convolutional Neural Networks for Multi-Label Image Classification

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

Convolutional Neural Networks for Multi-Label Image Classification

J. A. Marais

Thesis: MCom (Mathematical Statistics)

December 2017

English abstract.

Uittreksel

Konvolusionele Neurale Netwerke vir Multi-Etikel Beeldklassifikasie

("Convolutional Neural Networks for Multi-Label Image Classification")

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

List of Abbreviations and/or Acronyms

AA	Algorithm Adaptation
ANN	Artificial Neural Network
BR	Binary Relevance
CAD	Computer Aided Diagnosis
CC	Classifier Chains
CNN	Convolutional Neural Network
CV	Computer Vision
ECC	Ensemble Classifier Chains
kNN	k-Nearest Neighbour
LP	Label Powerset
mAP	Mean Average Precision
ML-kNN	Multi-Label k-Nearest Neighbour
MLC	Multi-Label Classification
MLIC	Multi-Label Image Classification
PT	Problem Transformation
RAkEL	Random k-Labelsets
SGD	Stochastic Gradient Descent

SotA State-of-the-Art

Nomenclature

N	number of observations in a dataset
p	input dimension or the number of features for an observation
K	number of labels in a dataset
$\boldsymbol{x}$	p-dimensional input vector $(x_1, x_2, \ldots, x_p)^\intercal$
λ	label
$\mathcal{L}$	complete set of labels in a dataset $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_K\}$
Y	labelset associated with $\boldsymbol{x}$, $Y \subseteq \mathcal{L}$
$\hat{Y}$	predicted labelset associated with $\boldsymbol{x}$, $\hat{Y} \subseteq \mathcal{L}$, produced by $h(\cdot)$
$\boldsymbol{y}$	K-dimensional label indicator vector, $(y_1, y_2, \ldots, y_K)^\intercal$, associated with observation $\boldsymbol{x}$
$(\boldsymbol{x}_i, Y_i)_{i=1}^N$	multi-label dataset with N observations
D	dataset
$h(\cdot)$	multi-label classifier $h : \mathbb{R}^p \to 2^{\mathcal{L}}$, where $h(\boldsymbol{x})$ returns the set of labels for $\boldsymbol{x}$
θ	set of parameters for $h(\cdot)$
$\hat{\theta}$	set of parameters for $h(\cdot)$ that optimise the loss function
$L(\cdot, \cdot)$	loss function between predicted and true labels
$f(\cdot)$	label prediction module, $f : \mathbb{R}^p \to \mathbb{R}^K$
$t(\cdot)$	thresholding function, $t : \mathbb{R}^K \to \{0, 1\}^K$
$\mathcal{N}(\boldsymbol{x})$	points in the input space neighbourhood of $\boldsymbol{x}$

Chapter 1

Experiments and Results

"For us, the most important part of rigor is better empiricism, not more mathematical theories."

— Ali Rahimi and Ben Recht, *NIPS 2017*

1.1 Introduction

Write introduction here.

1.2 Evaluation of Approaches?

1.2.1 Evaluatution Metrics

We chose the following metrics to measure the performance of the model on the data:

- Label-based macro F_1-score (F_1^{macro}),
- Label-based micro F_1-score (F_1^{micro}),
- example-based average precision (AP), and
- Label-based macro ROC-AUC

By using these four metrics we will get an all-round estimate of the performance the models. This is a diverse set of metrics. Includes label-based and example-based metrics, F-score, AP and ROC-AUC metrics and micro- and macro-average metrics. The F-score metric variants are popular choices for evaluating MLC models. The AP metric is common in the Computer Vison domain. The ROC-AUC is chosen mainly to be able to compare the models to other work reported on this dataset. ROC-AUC is also a convenient option since it is independent of the classification threshold chosen. When applicable, we will inspect the performance of the models on a per label basis.

When possible, the chosen set of metrics will be reported after each epoch in the form of line graphs. We do it this way because the point of convergence for the loss function being trained on might not be the same as the metric reported. Thus, if we only report the performance of the final (converged) model, we might not see the best possible performance for each of the metrics. The performance of the best (and/or final - I must still choose) models for each training phase will be reported in tabular form.

The final model evaluations will be reported on both the validation and testing sets. No model selection will be done on the test set evaluations.

Where possible we will include the time taken to train until convergence. Also time taken to make a prediciton for a single image.

1.2.2 Validation Approach

The data is split into a training, validation and test set. Since our dataset is large and our computing resources limited, we are comfortable not to use cross-validation. We will use the exact same split as in (paper) for fairer comparisons. The split was made randomly by patient in the following ratios: 70% training, 10% validation, 20% testing. There is no overlap between the patients in different splits to ensure uniqueness of the validation and testing examples.

1.3 Training Procedure

1Cycle policy with learning rate finder.

1.3.1 How will the policy parameters be chosen?

See paper. Parameters:

- ratio between minimum and maximum learning rate,
- decay rate
- momentum decay

- weight decay?

1.3.2 Fine-Tuning or Global Tuning

Prefer fine-tuning where possible to save time. Will not be as accurate as global tuning. Can precompute the activations before the classification head to be tuned, which saves a lot of repetitive computing. Will do complete global tuning for specific models where appropriate. Will use appropriate data augmentation techniques when precomputed activations are not used.

1.4 Model Architecture

Will have an initial experiment to compare different architectures. The majority of the experiments we will run using the smallest version of the chosen architecture type to reduce computational demand. We assume the conclusions are applicable to the larger models unless specified otherwise.

1.5 Base Experiments

1.5.1 Validation Split Experiment

Will it bias results to split data randomly?

1.5.2 Architecture Experiment

Which of the following architectures perform the best on our data ResNets, DenseNets, SEResNet, DarkNet?

1.5.3 Transfer Learning Experiment

Does it help to do transfer learning vs training from scratch?

1.6 Multi-Label Experiments

1.6.1 Loss Function Experiment

By training on which loss function will result in the best metrics?

1.6.2 Classification Head Experiment

Which classification head architecture obtains the best results? Does it learn label dependence?

1.7 Summary

Write summary here.

Appendices