論文の要旨

題目　Treatment of Multi-label and Multi-object Images in Deep Learning: A Focus on Image Annotation and Retrieval
(ディープラーニングにおける複数ラベルおよび複数オブジェクト画像の処理：画像アノテーションと検索に焦点を当てた研究)

JONATHAN MOJOO

INTRODUCTION

Deep learning has seen an unprecedented boost in recent years and has advanced artificial intelligence and its applications in many fields. One of the areas where research in deep learning has yielded amazing results is computer vision, which deals with the understanding of image data by computer-based algorithms. This is largely due to the emergence of convolutional neural networks (CNN), specialized networks that can learn and extract important features from images for various tasks. This is very significant, as digital images have become a part of our everyday life and continue to take up a large part of people's day-to-day activities, from social media to medicine and engineering applications.

Our work is predicated on the fact that most real-world images tend to have metadata that includes multiple labels as opposed to a single label per image. This is because a single image database can be used by different individuals, in different usage scenarios, each possibly requiring different types of information and classification schemes. It is therefore important for deep learning algorithms to learn how to effectively learn various tasks effectively on multi-label image datasets.

We argue that multi-label learning requires special treatment and poses unique challenges that are not encountered in single label learning. We propose improvements to standard multi-label deep learning algorithms in two computer vision tasks: multi-label image annotation and multi-label image retrieval.

MULTI-LABEL ANNOTATION WITH MISSING LABELS

In multi-label annotation, we focus on the recurring problem of incomplete or incorrect labels in multi-label image datasets. This affects the ultimate performance of deep learning algorithms, since these models are greatly affected by the accuracy of the data they are trained on. We propose two separate solutions to solve this problem. We propose a regularization-based approach, that extends the loss function in-order to leverage label co-occurrence patterns inside the training data, and label similarity outside the training data. Additionally, we propose using a separate graphical model (called a restricted Boltzmann machine) to learn label dependencies and reconstruct label vectors to supplement training images with extra labels. We introduce the following graph Laplacian regularizer:

$$G = \sum_{l=1}^{M} \boldsymbol{y}_l^{\mathrm{T}} (\propto L_h + (1-\propto) L_w) \boldsymbol{y}_l$$

Where $\boldsymbol{y}_l$ is a vector of predictions for the l-th example, M is the number of training samples, L_h and L_w are the Laplacian vectors from internal and external label similarity respectively, and $\propto$ is a trade-off parameter between them. We evaluate our methods on 3 benchmark datasets and show that they lead to a more accurate multi-label annotation model that is also more robust to incomplete training labels than the baseline CNN model. We also demonstrate qualitatively how our method adds more relevant labels to test images than the baseline.

MULTI-LABEL AND MULTI-OBJECT DEEP METRIC LEARNING FOR IMAGE RETRIEVAL

On the image retrieval task, we propose a simple and intuitive function for calculating pairwise similarity of multi-label images, based on the intersection-over-union of their label vectors. This allows for a clear distinction between different similarity levels of image pairs in the training data, and leads to more accurate ranking of retrieval results at test time. We also tackle the idea of multi-object images, which are multi-label images with several instances of a particular concept appearing in a single image. We introduce corresponding changes in the loss function to reflect the varying levels of similarity (by introducing an adaptive margin) and in the neural network architecture to aid in the processing of multi-label images (by introducing a relation network module). We use the following loss function to train our model:

$$L = \sum_{i=1}^{N} \max\left(\left\| \boldsymbol{x}_i^a - \boldsymbol{x}_i^p \right\|_2^2 - \| \boldsymbol{x}_i^a - \boldsymbol{x}_i^n \|_2^2 + s_{a,p} m,\ 0 \right)$$

This is similar to the standard triplet ranking loss used in deep metric learning, where $\boldsymbol{x}_i^a$, $\boldsymbol{x}_i^p$, $\boldsymbol{x}_i^n$ are network embeddings for the anchor, positive and negative examples, and m is a tunable margin parameter. However, we introduce the quantity $s_{a,p}$ that adapts the margin parameter based on the level of similarity between the anchor and positive examples.

We further improve upon this method by introducing label reconstruction using a variational auto-encoder, to correct errors in the similarity function introduced by missing labels. We show that our method achieves better results on both multi-label and multi-object retrieval tasks than the baseline, and several other recently proposed methods. We also present retrieval examples demonstrating that our method performs better at multi-label and multi-object image retrieval than competing state-of-the-art methods.