Project Title

Optional Subtitle

Hermanni Hälvä [1]

MSc. Computational Statistics and Machine Learning

Supervisor: Prof. Bradley Love

Submission date: Day Month Year

[1]**Disclaimer:** This report is submitted as part requirement for the MSc CSML degree at UCL. It is substantially the result of my own work except where explicitly indicated in the text. The report may be freely copied and distributed provided the source is explicitly acknowledged

Abstract

Summarise your report concisely.

Contents

Chapter 1

Literature Review

PAPERS THAT COMBINE DL AND LINGUISTICS

Learning Hierarchical Visual Representations in DNNs Using Hierarchical Linguistic Labels - Peterson et al. 2018.

Overview

The study most closely related to ours is Peterson *et al.* [1] who investigate whether the use of hierarchical labels helps DNNs to learn better visual representations. As DNNs are typically trained against single labels, the features learned are biased by the arbitrary level of these labels; for instance, different breeds of dogs each have their own label in the widely used ImageNet data-set, but there is no information in the labels that informs the models that all the different breeds are part of a larger category of 'dogs' [1]. In order to control for these hierarchical relationships in categories, the authors define two levels of labels for each image, called baseline (top-level, e.g. 'dog') and subordinate level (original low-level label e.g. 'golden retriever'). Peterson *et al.* argue that this is much closer to how humans represent and learn object categories. To test this, they train four versions of the InceptionV3 [**?**] DNN architecture on the ImageNet Large Scale Visual Recognition Challenge 2012 data (ImageNet hereon): original pre-trained model that only uses subordinate-level labels, model pre-trained on the subordinate labels and fine-tuned on basic-level labels, model pre-trained on basic-level and fine tuned on subordinate lables, and finally a model trained purely on basic-level labels. Several interesting results were attained. First, the authors found that including the basic-level labels led to a much more clustered representation of the feature spaces (e.g. the features vectors of different breeds of dogs were now bundeled up together whilst if trained on just subordinate labels, then there was no clustering of similar categories). Additionally, dendrogram of hierarchical clustering of thelearned feature space with basic labels showed a clear separation between between nature related objectes 'natural images' and images of man-made objects 'artificial images', eventhough no such information about the two groups was given to the model *a priori*. The authors note that this is close to humans' mental representations. To further compare the model's learn representation to humans, the authors investigate how well the model is able to capture human similarity judgements. For the models, similarity between two images is measured as the inner product of their feature representations, and this was compared to human ratings (on scale 1 to 10) of similarity of the same images. Whilst the authors found that on aggregate including basic labels improved the explanatory power of the DNN, the R^2 was not very

high (0.57) and as noted, similar results have in the past been attained using only the subordinate labels. Finally, the paper also presents a generalization experiment in which the model is given just a few examples of either sub or basic level images and then told to find other images from the data set which it would predict to have the same label. The results of this few-shot generalization experiment show similar results as corresponding human studies [**?**] in that introducing basic-level in model training leads to basic-level bias (even if the example given is of subordinate level, the model will generalize by seeking matches in basic level).

Thoughts:

- This is probably the most relevant papers to ours, though I dont actually think it's too similar because they use only two levels of labels so focus on just basic/subordinate, so I dont really see this as hierarchical. It's quite different from how we are going to do with word2vec or some other embedding that allows a much richer relationship between the words rather than just a 'vertical' link between two categories.
- They approach quite strongly from psychology point of view and dont even talk about the accuracy of the model. I think I will have more ML focus than this and will definitely look at the models' performance
- I do like the psychology approach in that if we think of AI more generally then I suppose we wish to achieve human-like semantic understanding and one could argue that this paper perhaps has some of that going on
- I am tempted to also use InceptionV3 in case we do wish to repeat any of their experiments, and for the reason it was used here which is that it's near state-of-the-art and pretty quick to train
- Find it bit weird that they define the model's feature space to be just the final layer: '..we pose multi-level labeling problem simply as learning a set of independent softmax classifiers that are unconnected to each other and fully connected to the final representation layer of deep CNN while other alternative approaches exist for defining the network architecture and loss function, this approach provides a single embedding space for all images, which allows us to inspect the representations with classic psychological methods such as hierarchical clustering.' Not the biggest fan of this approach as would expect also hierarchy of representations through out the network (c.f. human visual cortex)
- further, with above in mind the authors only fine-tune the final layer: 'For fine-tuning models, we freeze all but the weights in the last block of the model to speed up training'. If we wish to look a representation across hierarchy of layers, we cant do this. Hopefully this wont be too much computation...
- the t-sne and dendrograms of representations i do like so might consider something similar
- their human similarity judgement experiment is nice but am not very convinced by their results. Would be cool to top them but not sure how realistic it is to get hands on that data since they collected using Amazon MTurk which costs some moneys

- Their generalization experiment with basic level bias is also unconvincing to me. They only have two levels of labels with one more general than the other, and a model where features are hierarchical so it doesnt surprise me that the higher level subsumes the lower. If they would have three levels and would generalize to the middle level, then that would be pretty cool. Need to think if this experiment is going to be relevant to us

MULTI-LABEL CLASSIFICATION PAPERS

According to current plan, we are concerned with typical single-label classification. Nevertheless, multi-label image classification is relevant to us since those papers have to explicitly try to learn relationships between many related categories in objects. Different papers below approach this in different ways and have somewhat differing aims. They do share our goal of trying enforce some way of semantic understanding into DL papers. I do sometimes feel that multi-label classification is 'under rated'; it's much closer to how I imagine that humans process visual data.

YOLO 9000 - Redmon and Farhadi 2016

Overview

This is a object detection paper so in a way quite different from ours i.e. it tries to predict locations of all objects in the images and then label those. This paper is relevant to us due to its use of hierarchical labelling, which the authors use to combine distinct datasets of unequal hierarchy into one dataset, so their purpose of hierarchies is though very different from ours. More specifically, in this paper the authors combine imaages from classification data sets and object detection (includes localization and possible multiple objects). Problem is that imagenet has very specific labels (c.f. subordinate in above paper) whilst detection data set has only very generic labels e.g. dog (c.f. basic labels above paper). Need a coherent way to marge these different level labels, to create the joint data set. Also, say 'norfolk terrier' and 'dog' may not be mutually exclusive in photos thus cant merge using just a single softmax like in DL typically. Thus the authors instead create a hierarchical tree of the imagenet labels (which are based on WordNet). To classify with their hierarchical tree, predict conditional probabilities at each node i.e. probability of each hyponym of that synset given that synset. e.g. P(norfolk terrier—terrier), P(Yorkshire terrier— terrier). Can then calculate marginal probabilities by traversing upwards through the graph. This allows the authors to combine COCO and ImageNet data sets and efficiently do detection on over 9000 labels with only marginal performance drop. The authors also show that with the hierarchical model, the performance of the model degrades gracefully when it sees for instance a dog but isnt certain about the breed of the dog

Thoughts:

- the idea of hierarchies is more similar to us here than in the Peterson paper
- would be interesting thus compare our models performance to this one, though it may be hard given the weird training done here
- the hierarchical approach is more of a 'means' here and in itself is not much analysed. e.g. there is no discussion of reprsentations
- Whilst how theyve built the tree from wordnet is impressive, it only captures hierarchies whilst a word emebedding space may be able to capture more complex relationships
- I find their point about graceful degradation fascinating and will look to explore that as well. Very curious if a hierarchically trained version of our model turns out to be more robust to adversarial attacks.

CNN-RNN: A Unified Framework for Multi-label Image Classification - Wang et al.

Overview

In reality, visual views rarely ever contain just a single object with a unique label; rather, we perceive rich semantic information in even the simplest views. Single label classification fails to capture this, hence multi-label classification. Usual approach to multi-label classification treats it as multiple single label classification problems. This fails to capture the dependency between multiple labels e.g. sky and clouds usually appear together but cars and water shouldnt. Machine vision has in the past captured these type of dependencies using markov random fields but such grid models only really control for pairwise dependencies. Cannot handle complex higher order relationships in images. To do this the authors of this paper use RNN on multi-label data and show that this significantly improves classification accuracy.

Usually CNNs share features across different classifications, but the problem here is that small objects in images are hard to classify because the features are built on classifying the whole image in the best way possible. RNN helps in this because, as the paper shows, it implicitly creates an 'attention' model where the classifier focuses on different areas of the image based on the RNN memory state. More specifically, this is done by learning join image-label embedding to model semantic relevance, where the image embedding is the lower layer of the CNN. These are projected into same sub-space as label embeddings, and the LSTM memory thus captures higher order dependecies in this embedding space in particular. Model is best understood from figure 4. In prediction stage, beam search algorithm is used because markov property is not satisfied. And to be clear, training is done on multi-label data sets such as MS-COCO.

Results: state-of-the-art on multi-label since can weed out labels that can't possibly co-occur. On 1000 label data set the performance is poor because the DL model is trained on image net, which doesnt have concepts such as actor/actress that occur in the large multi-label data-set On MS-COCO data set poor performance with few tiny objects such as toaster/hair dryer that have little dependence on other categories. The authors also show nearest neighbours in label-image joint space to illustrate that the model indeed captures semantic similarity. They also show that (using de-convolution), there is implicit attentional mechanism where first the model focuses on entire image and then moves on to smaller parts.

Thoughts:

- A big difference between all these multi-label approaches and ours would be that here semantics are really controlled only by what's visible in images, whilst we could use entire corpus of text / ImageNet hierarchy even for a single image.
- On the other hand, our model wouldn't be able to do multi-label classification as it's hierarchical 'understanding' would derive entirely from word embeddings..
- A little like that attention here, I wonder if we could also visualize if using hierarchical model changes the regions of the image that the model focuses on vs. non-hierarchical model
- the observation about toaster/hair dryer being difficult to capture semantically is interesting. I wonder if our word embedding approach would suffer from similar?

- bit like the nearest neighbour label prediction here, we should attempt also to predict into the embedding space after training the model. I really wonder if we could thus train on single label data but actually somehow accomplish multi-label predictions (based on predicted embedding space). I guess this would be kind of what the zero-shot learning papers (see the section on those papers) attempt to do using word embeddings.

Other relevant papers and intro stuff

Computer vision researchers have already long before the advent of deep learning, been interested in multi-label classification as it would greatly enhanc our ability, for example to search and reterieve large quantities of image data [**?**]. Further, multilabel classification much closer corresponds to how humans perceive complex visual scenes and thus developing models with rich semantic ability would be a valuable step for development of artificial intelligence technology such as autonomous vehicles. Multi-label classification already of interest before deep learning: [**?**].

How is multi-label classification different from semantic segmentation / object detection? Not trying to locate/bound any object but rather concerned with labeling. Why would we need something other than YOLO9000 (Redmon and Farhadi 2016). Further, in our approach we are not trying to perform multi-label classification, rather we still perform single label classification but trying to benefit from semantic dependencies in training..

Bibliography

[1] J. C. Peterson, P. Soulos, A. Nematzadeh, and T. L. Griffiths, "Learning Hierarchical Visual Representations in Deep Neural Networks Using Hierarchical Linguistic Labels," 2018.