Project Title

Optional Subtitle

Hermanni Hälvä [1]

MSc. Computational Statistics and Machine Learning

Supervisor: Prof. Bradley Love

Submission date: Day Month Year

Abstract

Summarise your report concisely.

Contents

Chapter 1

Intro/Literature Review Notes

Humans are able to extract rich semantic information from visual scenes. For instance, upon viewing a picture of a dog, we may also be able to identify it as a specific breed such as a golden retriever. Further, our understanding of the image benefits from semantic knowledge that is not captured explicitly in the visual features of a specific image. For example, we also know that the dog belongs in the category of mammals which, in turn, are animals and thus living entities. While this type of hierarchical semantic visual understanding comes to us effortlessly, it remains a challenging task for computer vision. In fact, most image classification models are trained on data sets with single, mutually exlusive labels and thus the learnt feature representations do not account, for example, for both cat and dog being four-legged animals. Why is this a problem? Generalization? An exception to this is the area of knowledge transfer and related tasks such as zero-shot learning in which the aim is to predict labels of previously unseen classes of images; a popular approach for zero-shot learning is to borrown strength from, say text data, to create a semantic space that embed all the possible labels of images including those not seen previously. Mapping between images and the semantic space is then learned using the 'seen' images. Once this mapping is learnt, it can be used to transform previously unseen images into the semantic space and then apply some distance measure to label it as the category thats cloest in the embedding space. We postulate that the same idea could also be employed in the simple image classification tasks to achieve better generalization performance. In particular, we will train a deep learning model for image classification against a hierarchical semantic embedding derived from the Word-Net database which, as we will show, will correspond to regularizing the model weights to account for these semantic relationships. We will use the recently developed poincare embeddings as the resulting embedding space can capture both the similarity between the possible labels as well as the hierarchical semantic relationships encoded by WordNet graphical model. The learnt model is then transferred to perform standard image classification by adding a final softmax classification layer and fine tuning this agains the ImageNet database. Our results show ... hopefully some improvements over training the model only against ImageNet which illustrates the benefit of incorporatig semantic knowledge..

PAPERS ON PSYCHOLOGY OF VISUAL PERCEPTION & SEMANTICS

Basic Objects in Natural Categories [1]

Overview

The visual world that we experience consists of an infinite number of different stimuli, yet we are able to classify this environment into distinct objects. This is because there is an inherent correlative structure in real world objects. For instance, dogs never have wings and birds will not in general have a furry tail; humans are able to exploit this non-independence by grouping objects into categories based on their visual attributes. Categories in turn can be organised into hierarchical taxonomies. For instance, dogs and birds can be distinct categories but in a taxonomy can also be grouped into a higher level category of animals which are visually distinct from some other higher level category such as furnitures. In this paper the authors show that there is in general one level of abstraction in the taxonomy of visual objects at which humans naturally make the most fundamental distinction between objects i.e. they are the most differentiated at this level; this is called the basic level. The reason for this is that given infinite visual stimuli, it is more efficient to only differentiate objects that are relevant to any task at hand - the basic level represents the level of abstraction that is the most useful in most situations. Categories one level of abstraction above the basic level are called superordinate categories and they are less fundamental for separating objects as they have fewer common attributes with each other. Categories one level below the basic level are subordinate categories, and they are in turn less fundamental for separating objects because there is more sharing of visual features across contrasting subordinate categories.

Thoughts:

- DL for classification usually doesn't take account of taxonomies though one could argue that most datasets are at the basic level. Nevertheless it would be of benefit to account for entire hierarchy

- several limitations in this paper (see [2])

Pictures and Names: Making the Connection [2]

Overview

PAPERS USING WORD HIERARCHIES IN IMAGE CLASSIFICATION

Learning and Using Taxonomies For Fast Visual Categorisation - Griffin and Perona 2008

Overview

Pre DL-era paper in which the authors attempt to move from linear computational cost in number of categories of one-vs-other classifier to logarithmic cost by considering a tree-like hierarchical classification. This is particularly importan for large number of labels, say 10^4. The authors conjecture that humans are able to perform fast visual categorization by moving down a hierarchical taxonomy an hence avoid considering irrelevant classes e.g. to classify an 'object' as a dog first decide if something is an animal or not and after that if the animal is a dog or a cat. To do

this need to create hierarchy of categories as a binary tree. This is done using Spatial Pyramid Matchin on the images and their labels.

Thoughts:

- hierarchy here is in the classification model though the idea is similar to ours that humans hold visual categories

- the model here is very complex in that it requires several pre-processing steps e.g with SIFT so very different from DL approach

- rather than using wordnet hierarchy they calculate hierarchy from images and their labels. Using wordnet embedding would gives us the ability to draw strength from categories not necessarily in our images (though if we use imagenet, they should mostly be there?)

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.* [3] 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 dataset, 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' [3]. 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

NLP WORD EMBEDDING PAPERS

These papers provide important background on word embeddings. Main idea is to represent words in vector space such that distance between any two vectors reflects the similarity of those two words (semantically).

Improving Word Representations via Global Context and Multiple Word Prototypes - Huang et al. 2012

Overview

Main point is that in order to cluster words vectors appropriately, need to consider both syntax and semantics. Most previous approaches have only done this in the context of the local problem, but employing a global data can provide more accurate resuls, as it would better capture semantics whilst still contolling also for syntax. A joint model is hence used. Indeed, the authors show that this joint model provides better correlation with human similarity judgement of words.

More specifically, 'the model jointly learns word representation while learning to discriminate the next word, given a short word sequence (local contexti, syntactic information) and a document of text the sequence appears in (global context, semantic information)'. Two standard neural networks are used to predict scores that reflect the likelihood of the most recent word occuring given the local and global context. Sum of the two scores creates a total score for the most recent word, which is then used to construct a hinge-loss type of function to discriminate against any other possible word that could have occured, which is then minimized.

Thoughts:

- This seems like a pre-cursor to GloVe

ZERO-SHOT LEARNING AND RELATED PAPERS

These zero-shot learning papers are very much related to ours in the sense that they try to achieve generalization via word embeddings into unseen image classes. The DeViSE paper is probably the most similar though others relevant too. In fact, some of them already used this idea before the deep learning revolution.

Large Scale Image Annotation: Learning to Rank with Joint Word-image Embeddings - Weston et al. 2010

Overview

This is a pre-deep learning paper, on how to handle efficiently annotation of 'new' large scale data-sets such as ImageNet. Approach is to learn to represent images and annotations jontly in a low dimensional embedding space, idea being that low-dimension at test time translated into faster model. Indeed, a pre-deep learning methods such as performing knn in image feature space suffer from the curse of dimensionality. To do this they use a loss function that allow the model to learn-to-rank i.e. to optimize precision of k-top annotations. In practice, their model performs linear mapping into joint embedding space (i.e. annotations and images are multiplied by their own matrices of equal dimensions).

Thoughts:

- while impressive at the time, no longer state-of-the-art, but their point about curse of dimensionality may still be important even for DL. Say if a DL model has a softmax of too high N, it becomes harder and harder to separate out different classes so word embedding may help here

- The model used by the authors is just a linear matrix transformation. Thus DeViSE paper can be seen as a big improvment over this.

Zero-shot Learning Through Cross-modal Transfer - Socher et al. 2013

Overview

Motivated by the idea that humans have the ability to identify unseen objects even if they know about that object from having read about it. The authords introduce a model that can predict both seen and unseen classes: 'without having ever seen a cat the model can say whether it is indeed a cat or another category it has seen during training'. Large unsupervised text data is used to create word embeddings and then images are mapped into this space using neural networks. Prediction of unknown classes works by determining whether the test image is on the same manifold as known examples - this is based on outlier detection. One advantage over previous works is that unsupervised text corpus can be used rather than a manually constructed word embedding space.

The model used to create unsupervised word embedding is based on [**?**] (see above), which can be seen as earlier version of GloVe. They perform this on Wikipedia text data to create word embeddings that capture local and global context. Image features are then computed in unsupervised fashion using orthogonal matching pursuit. With the word and image feature vectors available, a two layer neural network is trained on the images with known classes into the word embedding space. The authors then use t-sne to visually illustrate that if unseen classes are fed

into the model, the predicted word vectors will not be close to the vectors of known classes (which are clustered in turn). The closest known classes to the zero-shot classes however give idea of their semantics. In practice, to accomplish above we need to first detect whether an image is of a known class or of unknown class, and for this purpose a binary novelty random variable is employed (this is becuase otherwise would never predict any of the unseen classes if only used training data of seen classes). An outlier detection (based on Gaussian distribution) to predict whether an image is of unseen class given its predicted word embedding vector. If image is predicted as seen, a softmax classifier is used to predict its label, whereas, if image is predicted as new, an isometric Gaussian is assumed around its word vector and its class is assigned based on likelihood i.e. distance betwen predicted word embedding and full-word embedding space that contains also the unseen classes (this is the classic zero-shot trick using word data). The equations in the paper make this much clearer so will add those. In training, CIFAR-10 image data is used with 2 classes omitted for testing. The results depend strongly on the threshold used to determine whether image is in an unseen or seen class.

Other interesting points: The authors show how the zero-shot learning can be framed in a fully bayesian way. They also compare the novelty detection in word embedding space as above, to doing novelty detection in image feature space and find the former superior as it adds the additional information from semantics.

Thoughts:

- the zero-shot approach here may not be fully relevant to us but I still wrote it up so have a general idea of why word embeddings are used in these papers. Further, I do find this an interesting application and if we have time, may want to consider zero-shot learning

- no deep models used here really though they do call their two layer neural network a deep model, different times ... no convolutions though

- hierarchical information is not used

- a major problem is that the threshold used to discriminate between known and unknown class creates an inherent trade-off between the ability to predict these two. The DeViSE model improves on this (see below)

- a lot of useful references to go through from this paper, and some other intresting ideas not covered above may be worth exploring later

DeViSE: A Deep Visual-Semantic Embedding Model
Overview
Typical object classification models treat all the categories unrelated, via N-way softmax. This leads to models that 'cannot transfer semantic information about learned labels to unseen words or phrases. Solution this paper proposes is to use both standard image data and then an unrelated large unannotated text data to learn semantic information. In particular, the model maps image inputs into this rich semantic embedding space. The results show similar performance to standard

state-of-the-art DL models, but with a significant improvement in that much less 'semantically unreasonable' mistakes are made. Perhaps more importantly, they show this joint training allows the model to generalize to 20000 visual categories, despite having trained the model on just 1000 categories.

DeViSE extends work of Weston et al. (above) as it allows non-linearities and also capture semantics from text that's not contained in the image lables, hence allowing for zero-shot generalization. It also improves on the Socher et al. paper by using a deep model and avoid the trade-off which that paper has in predicting seen and unseen classes. That model also had several different models whilst this one uses a unified model that only uses embeddings.

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.

<u>YOLO 9000 - Redmon and Farhadi 2016</u>

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.

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

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: [?].

Similarly, word embeddings have been used extensively to capture image semantics much before deep learning. As an example [?] perform unsupervised image auto-annotation with a probabilistic model in which a joint latent space is used to represent co-occurence of image features and words. More specifically, the authors constraint the latent space to be mainly defined by word-features as they argue that the semantic relationships between words is richer than that of image features, and second, co-occurence of words in text is semantically more meaningful than that of visual features. The authors' first point may be less relevant for the state-of-the-art today as convolutional neural networks have shown the ability to build complex fvisual features. Nevertheless, if we train model that captures full semantic relationships in large sample text, this should be richer still; for instance, ImageNet is represented by a Directed Graph rather than just a tree hierarchy. The second point too remains apt, image feature co-occurence indeed may not mean much (though deeper layer ones do probably more so in CNNs). The presence of varying background and other objects, makes it difficult to maintain semantic similarity on the basis of visual features alone.

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...

Is it really a reasonable assumption that semantically similar words should be close to each other in image space. Perhaps true for e.g. cat and a dog. But what about sheep, clouds or snakes and cables. Should they be close to each other in an optimal embedding space?

Bibliography

[1] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem, "Basic objects in natural categories," *Cognitive Psychology*, vol. 8, no. 3, pp. 382–439, 1976.

[2] P. Joliceur, M. A. Gluck, and S. M. Kosslyn, "Pictures and Naming: Making the Connection," *Cognitive Psychology*, vol. 16, pp. 243–275, 1984.

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