# Pre-trained Convolutional Networks and generative statistical models: a comparative study in large datasets

John Michael and Luís F. Teixeira

INESC TEC
FEUP - MIEEC
Michael.Galveira@gmail.com
luisft@fe.up.pt

**Abstract.** This study explored the viability of out-the-box, pre-trained ConvNet models as a tool to generate features for large-scale classification tasks. A juxtaposition with generative methods for vocabulary generation was drawn. Both methods were chosen in an attempt to integrate other datasets (transfer learning) and unlabelled data, respectively. Both methods were used together, studying the viability of a ConvNet model to estimate category labels of unlabelled images. All experiments pertaining to this study were carried out over a two-class set, later expanded into a 5-category dataset. The pre-trained models used were obtained from the Caffe Model Zoo.
The study showed that the pre-trained model achieved best results for the binary dataset, with an accuracy of 0.945. However, for the 5-class dataset, generative vocabularies outperformed the ConvNet (0.91 vs. 0.861). Furthermore, when replacing labelled images with unlabelled ones during training, acceptable accuracy scores were obtained (as high as 0.903). Additionally, it was observed that linear kernels perform particularly well when utilized with generative models. This was especially relevant when compared to ConvNets, which require days of training even when utilizing multiple GPUs for computations.

**Keywords:** computational geometry, graph theory, Hamilton cycles

## 1 Introduction

Image classification is a central problem of Computer Vision and Machine Learning. One of the greatest obstacles faced when handling large volumes of images and video is tied to the fact that, more often than not, the visual data is unlabelled. Whilst unlabelled data is readily available and easy to extract, labelled data is scarce and quite costly to obtain. It's therefore important to find reliable methods which minimize the need for labelled data. A similar argument could be drawn towards using labelled data from categories or applications which are sufficiently similar to the desired one, a form of transfer learning. This can be done indirectly by utilizing pre-trained models as estimators for the class of images

from the target dataset. Exploring these various options can help understand how to overcome the scarcity of labelled data. The pre-trained ConvNets used to carry out this study (trained on the full Imagenet dataset) follow a slightly modified AlexNet architecture [2] and an architecture based on the one proposed for the ZFNet ConvNet [1]. The models were obtained from the Caffe Model Zoo page. A SPM+SVM classification scheme [5] was used to study the performance of methods for generation of a visual vocabulary, namely Sparse Coding (SC, [4]), Latent Dirichilet Alocation (LDA, [7]), probabilistic Latent Semantic Analysis (pLSA, [3]) and Ficher's Vectors (these built on top of a GMM model instead, as seen in [6]). These generative methods can be seen as a form of feature reduction and description, and are interesting due to being fairly orthogonal to lower-level feature manipulation (in analogy to what's achieve in preprocessing steps for ConvNets). In an attempt to bring both approaches together, the ConvNet was also utilized together with these generative methods to explore the viability of utilizing these pre-trained models to estimate the class of unlabelled images.

## 2 Methodology

All methods were tested on two distinct datasets: an initial dataset with 5 thousand images of the dog and cat categories, varied in resolution, background and race (obtained from Kaggle competitions), which was then supplemented with three additional categories, whale, fish and galaxy (obtained from other Kaggle competitions and the Imagenet dataset). These classes were picked for their very high variability between each individual image (especially true for the cat and dog classes), their similar backgrounds (fish and whale classes frequently have marine backgrounds with predominant shades of blue, in an attempt to understand if the classifiers learn interesting features from the class and not the backgrounds themselves) and due to their uniqueness and sheer oddity (the galaxy class being highly regular and featuring a low amount of colour information, as well as low resolution). The datasets for each class were randomly split between test and training sets for each trial. Features were initially obtained from images utilizing both PCA-SIFT (64 components) and colour information in the form of a Bag-of-Colours (BoC) descriptor with 32-bins. Lastly, as previously explained, an attempt to enhance their performance through assistance of the pre-trained ConvNet was studied, by altering the typical SVM cost function:

$$\begin{aligned} \gamma := \min_{\bar{\gamma},w} \|w^2\| - C\sum_{i=1}^{n} \epsilon_i \\ \text{Subject to:} \\ \epsilon_i > 1 \\ y_i(w^T x_i) \geq 1 - \epsilon_i \end{aligned} \tag{1}$$

So that the scalar term $C$ becomes a diagonal matrix, where each non-zero entry $c_{(i,i)}$ represents the weight of the $i$th sample, estimated by class score as-

signed by the ConvNet. In the training scheme, the sample scores were fixed to $0.5 \times \alpha \times c_{CONV,i} \times \lambda$, where $\alpha$ represents the accuracy score of the pre-trained ConvNet on all the available training labelled data and $c_{CONV,i}$ the class score attributed to unlabelled image $i$ by the ConvNet. This weighting reflects both an estimation of the confidence in the ConvNet's tentative performance (through the parameter $\alpha$) and an estimation of the tentative classification of the image through $c_{CONV,i}$. After class scores were generated for all unlabelled data, each image was assigned a tentative label through $\max(c_{cat,i}, c_{dog,i}, ...)$, where $c_{cat,i}, c_{dog,i}, ...$ are the normalized scores for the various categories (such as "cat" or "dog"). Values for the average score and maximum score for each class, $c_{max,j}, c_{avg,j}; j = cat, dog, ...$, were also were computed. These were used to discard images which were exceptionally noisy or in which the ConvNet struggled to attribute a class with some certainty. In practice, unlabelled images are drawn if their score is greater than $c_{avg,i}$ (and therefore not necessarily the highest scoring images for either class in the list). If there aren't enough images scoring over $c_{avg,i}$, images with lower scores are drawn until enough unlabelled samples are obtained. Furthermore, the parameter $\lambda$ is a SVM empirical parameter, optimized through model cross-validation. This parameter controls the penalty of misclassifying a data point to maximize the margin for the remaining points.

**Generated topic vectors** The LDA/pLSA models are generated utilizing labelled images from both categories and unlabelled images without distinction. The interpretation of the labelled corpus of either category is tied to the overall dataset- that is, the topical vectors generated for an image of a certain category depends upon the corpus of images utilized when creating the LDA/pLSA model (and thus, its topic representation varies with the corpus of other categories and the unlabelled corpus used to create the model in the first place). Once the model is obtained, one can now generate a topic representation for each image belonging to each category. These topic vectors can be used to generate an overall frequency representation for topics belonging to each category. If this is done for each class, it can subsequently be normalized to obtained a new, category-specific multinomial model that represents each individual probability $p(z_i|C_j)$, with $z_i$ the ith topic and $C_j$ the jth category. In order to keep some form of sparsity-constraint, mirroring what the original LDA/pLSA formulation aims to achieve (it's desirable that topics are associated with a small portion of the most salient words, as to represent the variance between different classes correctly), the median number of topics represented in each category, $\hat{z}_j$,is also computed. After computing the new multinomial model, uniformly drawn random numbers can be used to create new, "fictitious" topic vectors. This is useful when supplementing classes for which the amount of data is smaller, and was tested upon as an alternative to oversampling (as it involves some variance when creating new samples). It also bypasses a number of feature detection, extraction and reduction steps, saving some computational resources. A topic is added to this new vector if the random number, $r$, is such that $r \geq p(z_i|C_j)$ with $0 \geq r \geq 1$. If the

number of topics for this vector is greater than $1.5\hat{z}_j$, topics with smaller probabilities (smaller $p(z_i|C_j)$ factors) are set to zero until the sparsity constraint is verified. While this method has the advantage of promoting more salient topics, it also has the adverse effect of being much more prone to overfitting, so the number of generated topic vectors has to be kept conservatively small. The factor of 1.5 utilized in the sparsity constraint is purely empirical and was set after some brief experimentation with other factors in the 1˜2 range.

### 2.1 Cross-validation scheme and testing

A stratified 4-fold cross validation scheme was used for all the experiments. The final model is obtained by aggregating the results of all the folds through model averaging. Testing each model for performance was done on 12500 images on both the binary and the 5-category datasets. In both cases, the number of images belonging to each category was the same, allowing for isotropic priors which simplify calculations without any meaningful loss of generality. Training and testing was carried out on an Intel i7 4720 processor and 16 GB RAM for the majority of the methods. The exception was training and testing with the pre-trained ConvNet, which required two Nvidia GTX970 GPUs to yield results in a timely manner (training took roughly two days on this hardware).

## 3 Results

A summary of the main results achieved during the course of this study is presented in table 1. The classical SPM+SVM approached applied to a visual vocabulary obtained through the Sparse Coding or Latent Semantic Analysis used one thousand labelled images for training. The hybrid ConvNet+generative approach used 500 labelled and 500 unlabelled images. Furthermore, optimization regarding the generative methods used for visual vocabulary generation is present in figure **??**, where the effect of overfitting and underfitting can be observed, as expected.

As is customary, ConvNets remain unparalleled in terms of performance for similar image categories. Furthermore, if sufficient computational resources and labelled examples are available, they'll always outperform any other of the presented methods. However, when labelled data is limited, simpler SPM formulations paired up with generative methods offer a viable and computationally lighter solution. These can be used ether in alternative to pre-trained ConvNet models or in conjunction with these, in an ensemble that utilizes the strong points of either approach. One thing to note is that the regularity of classes (that is, how similar elements of each class are) is correlated with accuracy in methods which use SVMs. This can be seen by observing the confusion matrices 2 and 3, noting that the galaxy, fish and whale category have higher regularity. This is an obvious result, as variance within the same class results in more features being captured in each individual image and also in the possibility of some of those features being similar to those present in other classes (as is the

case for the cat and dog categories). If a class is very regular, it's also very easily predicted as some very salient, distinct features can be found across all elements of such a class. Considerations about the nature of dataset, the idiosyncrasies of the classification task and limiting factors related to computational resources and time available can weight in favor of some methods and detriment of others. A trade-of was ultimately shown to be present when choosing which method better fits a specific problem.

| Method \ Dataset | Bin | 5-class |
|---|---|---|
| **SPM+SVM** | 0.82696 | 0.8848 |
| **Pre-trained ConvNet** | 0.945 | 0.861 |
| **Both methods** | 0.80928 | 0.87648 |

Table 1: Summary table of the best accuracy in each dataset (results presented-SC for the binary case, pLSA for the 5-category case)

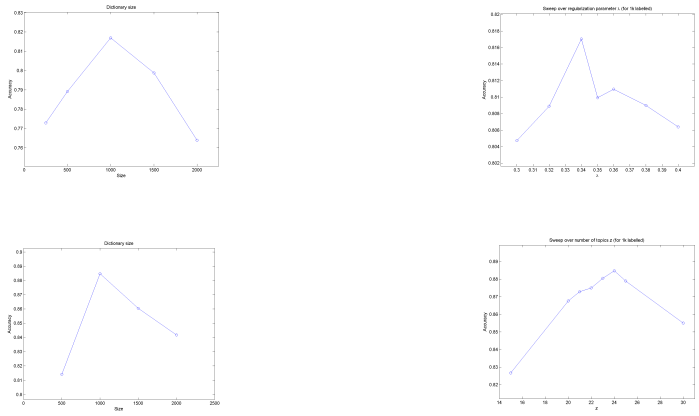

Fig. 1: Example images of the 5-category dataset. From top to bottom, a pair of images with the galaxy, cat, dog, fish and whale labels are presented, respectively

## 4 Conclusion

Despite the limitations on available computational resources and the modest time frame in which the study was carried, it was shown that the features ex-

| Labelled \ Predicted | Cat | Dog | Fish | Whale | Galaxy |
|---|---|---|---|---|---|
| **Cat** | 2013 | 429 | - | - | - |
| **Dog** | 440 | 2004 | - | - | - |
| **Fish** | - | - | 2249 | 201 | - |
| **Whale** | - | - | 144 | 2309 | - |
| **Galaxy** | - | - | - | - | 2481 |

Table 2: SPM+pLSA confusion matrix for training with 1000 labelled examples for the 5-category dataset, accuracy of 0.8848

| Labelled \ Predicted | Cat | Dog | Fish | Whale | Galaxy |
|---|---|---|---|---|---|
| **Cat** | 1998 | 466 | - | - | - |
| **Dog** | 473 | 2000 | - | - | - |
| **Fish** | - | - | 2231 | 202 | - |
| **Whale** | - | - | 189 | 2261 | - |
| **Galaxy** | - | - | - | - | 2466 |

Table 3: SPM+pLSA confusion matrix for training with 500 labelled and 500 unlabelled examples for the 5-category dataset, accuracy of 0.87648

tracted by the pre-trained ConvNet model were useful for image classification tasks, yielding comparable accuracy to more traditional methods. Furthermore, when used in conjunction with more typical generative methods for visual vocabulary generation, these allowed to replace labelled data with unlabelled data through the class estimation scheme described without incurring a significant accuracy loss. The results validate the hypothesis that pre-trained ConvNet models can be quite useful in providing an earlier estimation of the class to which an unlabelled image belongs, for posterior use in other models. Furthermore, it was shown that, through some empirical tuning of various weight parameters, the class scores generated by these pre-trained models can offer a satisfactory estimation of the confidence for the tentative labelling provided by the ConvNet. Through the various experiments in both datasets, the performance of multiple SPM models with generative methods creating a visual vocabulary on a dataset with a mixture of labelled and unlabelled data was either kept at a competitive accuracy level. Particularly, the pLSA formulation showed slight performance increases in the labelled and unlabelled mixed sets (of slight over 1%) compared to utilizing only the ConvNet, whilst displaying only very minimal accuracy losses (less than 2% in all cases) compared to the usage of solely labelled data for training. The study showed the benefit of allying the statistical formulation from generative models, which captures richer information in smaller, sparser feature vectors, with out-the-box ConvNet models trained in large, generic datasets. Further, this yielded a decrease in the training time, due to a shorter vocabulary

creation step. The fictitious topic vectors created from the multinomial model built on top of the LDA/pLSA models were also validated as another tool to combat insufficient data for some categories, if used with care.

In the future, studying other ways to further integrate these generative methods with ConvNets or explore more elaborate schemes for utilizing the class scores from pre-trained networks might yield even better results.

## References


1. Jia, Y. and Shelhamer, E. and Donahue, J. and Karayev, S. and Long, J. and Girshick, R. and Guadarrama, S. and Darrell, T.: Caffe: Convolutional Architecture for Fast Feature Embedding (2014)
2. Krizhevsky, A. and Sutskever, I. and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks Advances In Neural Information Processing Systems, 1–9 (2012)
3. Bosch, A. and Zisserman, A. and Munoz, X.: Scene classification using a hybrid generative/ discriminative approach IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, 712–727 (2008)
4. J. Yang and K. Yu and Y. Gong and T. Huang: Linear spatial pyramid matching using sparse coding for image classification Cvpr'09 (2009)
5. Lazebnik, S. and Schmid, C. and Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2169–2178 (2006)
6. Sanchez, Jorge and Perronnin, Florent and Mensink: Image Classification with the Fisher Vector : Theory and Practice, Cvpr'13 (2013)
7. Lee, Chu-hui and Chiang, Kun-cheng: Latent Semantic Analysis for Classifying Scene Images Proceedings of the International MultiConference of Engineers and Computer Scientists 2010, 17–20 (2010)