Elsevier Editorial System(tm) for Computers & Graphics
Manuscript Draft

Manuscript Number:

Title: Deep Learning on Small Datasets using Online Image Search

Article Type: Research Paper

Keywords: convolutional neural network; deep learning; image classification; reinforcement learning

Corresponding Author: Mr. Martin Kolář, M.Sc.

Corresponding Author's Institution: Brno University of Technology

First Author: Martin Kolář, M.Sc.

Order of Authors: Martin Kolář, M.Sc.; Pavel Zemčík

Abstract: This paper tackles a very hard and important problem of training deep models with small amounts of data. We propose a semi-supervised self-training bootstrap to deep learning on small datasets by retrieving and utilizing additional images from internet image search.

We adopt the Pseudo-Label method proposed by Dong-Hyun Lee in 2013, previously used on the elementary MNIST handwritten digit classification task. We show that by suitable changes to its example weighting and selection mechanisms it can be adapted to general image classification tasks supported by online image search.

This approach does not require any human supervision, it is practical and efficient, and it actively avoids overtraining. The usefulness of the proposed method is demonstrated on the SUN 397 dataset with only 50 training images per category. When exploiting results of Google's Image Search, we achieve a significant improvement over current state-of-the-art, with a classification accuracy of 51%.

Cover Letter

Deep Learning on Small Datasets using Online Image Search

February 14, 2016

The main contribution of this work is the ability to learn classifiers on small datasets, where this was previously impossible. Datasets with as few as 5 images per class are shown to be perform well. The approach is also demonstrated on the SUN 397 dataset, where the highest accuracy ever published is achieved.

This method adapts the work of Lee, Dong-Hyun: *"Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks." (Workshop on Challenges in Representation Learning, ICML. Vol. 3. 2013.)* (**published**). It is closely related to CNN bootstraps for noisy labels, such as Reed, Scott, et al: *"Training deep neural networks on noisy labels with bootstrapping." arXiv preprint arXiv:1412.6596 (2014)* (**published**).

The original Pseudo-label approach was incapable of handling large datasets. In this work, we improved the example selection method and weighting. Previous bootstrapping methods for Deep Learning worked on nosily-labelled data, but our method optimally utilises correctly labelled images as well as noisy data, in order to achieve state-of-art results.

Deep Learning on Small Datasets using Online Image Search

Abstract

This paper tackles a very hard and important problem of training deep models with small amounts of data. We propose a semi-supervised self-training bootstrap to deep learning on small datasets by retrieving and utilizing additional images from internet image search.

We adopt the Pseudo-Label method proposed by Dong-Hyun Lee in 2013, previously used on the elementary MNIST handwritten digit classification task. We show that by suitable changes to its example weighting and selection mechanisms it can be adapted to general image classification tasks supported by online image search.

This approach does not require any human supervision, it is practical and efficient, and it actively avoids overtraining. The usefulness of the proposed method is demonstrated on the SUN 397 dataset with only 50 training images per category. When exploiting results of Google's Image Search, we achieve a significant improvement over current state-of-the-art, with a classification accuracy of 51%.

Keywords: convolutional neural network, deep learning, image classification, reinforcement learning

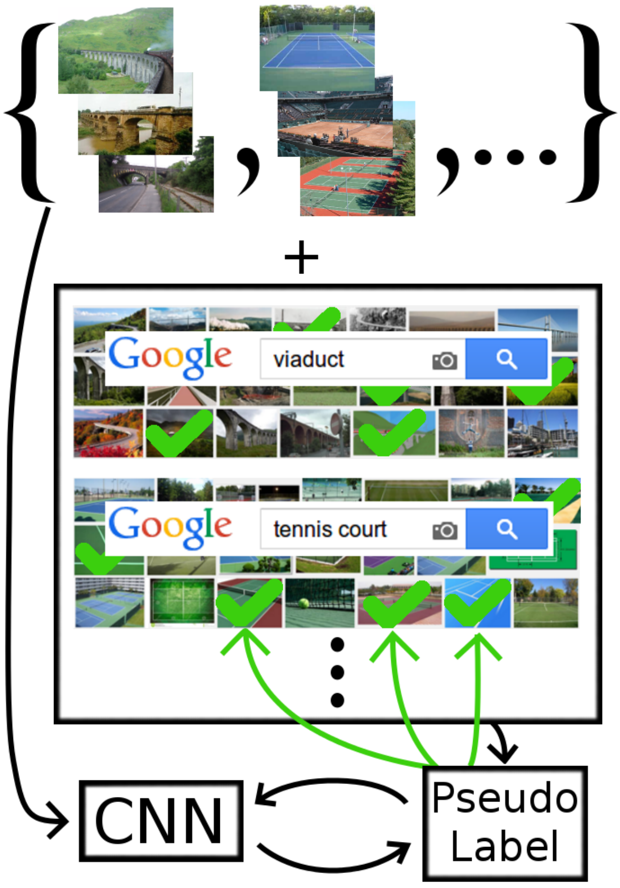

Figure 1: Pseudolabel selects useful additional images from an unreliable source, to help train a Deep Learning classifier

1. Introduction

Image classification is an important and challenging problem of Computer Vision. Traditionally, visual categories could be learned by Support Vector Machines on histograms of local features [34]. Current approaches have shifted towards Convolutional Neural Networks [12, 27, 5], which require vast amounts of data and computational power to learn millions of parameters. Such approaches have achieved near-human performance in face recognition [32], and have beaten previous approaches in classification of both very broad and very specific categories [23]. The purpose of our approach is to be able to use datasets with few examples, rather than to fine-tune training on a small dataset where CNNs already achieve high accuracy.

Deep Learning relies on large labelled datasets, with several hundred images for each category, but the creation of such datasets is demanding. Imperfect datasets can be created cheaply in an automated fashion, but near-perfect labelling, required by current approaches, relies on manual selection.

Creating datasets autonomously from the web, so that classifiers can be trained, has been demonstrated to work well when the requested data is in the form of text labels [25, 26]. In this paper, we focus on the problem where a few images are already known, and a label can also be retrieved, so that we can also learn classifiers when the label is ambiguous (such as "crane") or machine-generated.

The contribution of this paper is the ability to learn visual categories from very few images. One way to do Deep Learning on small datasets is to initialize network parameters from an existing network. Networks have been show to produce excellent embeddings, which generalize well to new categories [21, 5]. However, this approach is limited, and a larger dataset will further improve results.

Therefore, there is a natural need for an approach which couples cheap automatically retrieved weakly labelled images with pre-initialized Convolutional Neural Networks. Pseudolabel [15] is such an approach. However, it remains to be adopted to a large, challenging classification dataset. Pseudolabel exploits the iterative nature of Neural Network training, and expands a small set of correctly labelled training data with some of the imperfectly labelled training data. This approach selects samples from the imperfect dataset to best supplement the correct data.

Figure 1 shows the presented method. It relies on an adapted Pseudolabel approach, allowing for use on web-scale datasets of millions of images. The results are demonstrated on a toy problem devised from the SUN 397 dataset, and on the full SUN 397 dataset expanded with images gathered from Google's image search without human intervention. The toy problem allows us to analyse the properties of the data selection progress during training. Using these findings, state-of-art accuracy is achieved on the full dataset.

2. Previous Work

As discussed in section 2.1, Convolutional Neural Networks produce state-of-art results, but deal poorly with small datasets. The Pseudo-Label method, section 2.2, uses an unlabelled dataset to mitigate this. Any such approach needs to fully consider Dataset Bias and Limitations, section 2.3. Semi-Supervised Learning offers a structured approach to utilize labelled data in conjunction with an unlabelled dataset, and this work is discussed in section 2.4.

2.1. Convolutional Neural Networks

Convolutional Neural Networks [13] are the state-of-art approach for image classification, achieving the best accuracy for classification and detection [23]. These methods require large datasets [33], and this is handled by dataset augmentation with rotation, distortion, and other changes to the used images [12].

While much excellent work has been done to enhance the abilities of CNNs on large datasets [38, 31, 27, 28, 30, 20], it has generally been accepted that small datasets cannot be directly trained upon with random weight initialization. In this work, we focus on using the CNN structure to improve accuracy, rather than explicitly attempting to improve features, because features can be transferred from classifiers trained on other datasets.

Other approaches to train on small datasets without Neural Networks have been published, with limited success, such as a generative models [7] and a V1-like model [4].

2.2. Pseudo-Label

Pseudo-Label [15] introduced Semi-Supervised Learning to Convolutional Neural Networks. The CNN is trained in the usual way, but training images are supplemeneted from an unlabeled dataset. Low-density separation between classes justifies the use of entropy regularization on additional data.

In addition, at each iteration, the mixed set is classified with the current network, and these predictions are used as labels for the next iteration. Random selection from the mixed set, and increasing weights for the selected subset, are meant to help convergence to a classifier principally influenced by the training set.

This approach is justified by the cluster assumption, which states that the decision boundary should lie in low-density regions to improve generalization performance [1]. Rather than explicitly searching for low-density regions, the Pseudo-Label approach implicitly finds these, because changes in classification are more likely to occur in regions where the consensus among examples can be perturbed by few label changes. The Pseudo-Label approach helps with the MNIST dataset, divided artificially into a training set and a mixed set for which labels are unknown. An accuracy comparable to that achieved by using the entire set was reached. However, this dataset is long considered solved [36], and similar results remain to be demonstrated on a challenging problem.

2.3. Dataset Bias and Limitations

Datasets can have a variety of biases, which will affect the trained classifier [33]. Since object classification should perform well across a broad spectrum of variances, such as lighting or deformation, datasets should exhibit these as well. Most datasets used are created semi-automatically: images are retrieved from a good automated source, and manually sifted through. Depending on the source, this leads to different forms of bias: ImageNet is known to contain mainly centered images, SUN 397 is mostly composed of canonical ('archetypal') scenes.

By augmenting a biased dataset with additional data, the bias can be reduced and the resulting classifier may demonstrate less unwanted specificity. This can be accomplished by extending the datasets manually, and image classifiers have greatly benefited from new, larger datasets (see Table 1). Similarly, human level performance on the Labelled Faces in the Wild dataset[1] [11] was achieved by pretraining on a private dataset of 800 to 1 200 faces for 4 030 people [32].

Table 1 lists the most popular image classification datasets. While a larger number of categories makes a classification increasingly difficult, the top published classification accuracy is more correlated with the number of example images per category.

A kind of database bias can even be seen in raw images from Google's image search: low accuracy, constructive error, and canonicity. For instance, a search for the SUN 397 category "marsh" will yield many images of people with the surname "Marsh", and a search for "mountain" will yield a disproportional number of images of the Matterhorn and visually pleasing photographs. Google's image search accuracy decreases as further images are retrieved, see Figure 2.3. However, the reasoning behind using such data is that the sheer number of images guarantees that there will nevertheless be many representative ones.

2.4. Semi-Supervised Learning

Weakly Supervised Multiple Instance Learning (WSMIL) is a subproblem of Semi-Supervised Learning. By making the

[1] 13 323 web photos of 5 749 celebrities

dataset	# categories	# images containing instance	top published classification accuracy
MNIST [14]	10	5 421 - 6 745 (mean 6 000)	99.79 [36]
ImageNet [23]	1 000	732 - 1 300 (mean 1 281)	68.4 (top-1), 92.3 (top-5) [39]
PASCAL VOC 2012 [6]	20	303 - 4 087 (mean 834)	90.3 mAP [37]
SUN 397 [40]	397	100 - 2 361 (mean 274)	47.2 ± 0.2 [24]
Caltech 256 [8]	256	80 - 827 (mean 119)	82.2 [41]
Caltech 101 [7]	101	31 - 800 (mean 90)	93.42 ± 0.5 [9]
MS COCO [18]	91	~300 - ~600 000 (mean 7 849)	59.0 mAP [10]

Table 1: Comparison of image classification datasets. Note that the top-1 metric is inherently inappropriate for ImageNet

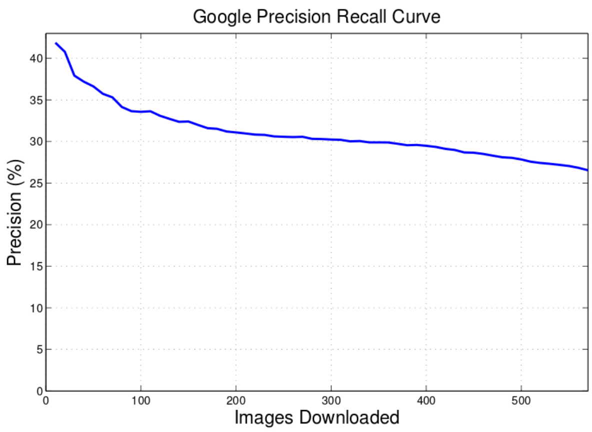

Figure 2: The portion of images returned by Google in 2007 rated *good* while constructing the Caltech 256 dataset [8].

assumption that at least one of the retrieved images for each class is correctly labelled, training with online image search data becomes WSMIL [35]. This approach has been coupled with the traditional image classification approach of a dividing hyperplane in a feature histogram hyperspace [17, 16].

CNNs have also been coupled with WSMIL [20, 3], but in the setting of searching through an image for the object instance, rather than searching through weakly labelled images. The CNN training process is sensitive to noisy labels, and semi-supervised learning approaches have been proposed to handle this issue [22, 29].

3. Method

This section describes the data, the method, and the implementation. The description of the method is divided into how a CNN is trained without Pseudolabels, how it is trained with Pseudolabels, Pseudolabel selection, and Pseudolabel weighting.

Pseudolabels are labels assigned during each epoch to any unlabelled images based on classifier responses. In our setting, these are the images retrieved from any online image search.

3.1. Notation

Throughout this paper, the following conventions are adopted: $\mathbf{X}$ is a set of images $\{X_1, X_2, X_3, \dots\}$, $\mathbf{y}$ is a set of labels $\{y_1, y_2, y_3, \dots\}$ where $y_n \in [1, C]$. C denotes the number of categories. Data has the form $(\mathbf{X}, \mathbf{y})$. Every i model update iterations is referred to as one epoch, and a set of images and labels during the duration of epoch e is denoted $(\mathbf{X}_e, \mathbf{y}_e)$.

Correctly labeled images are divided into a train set and test set: $(\mathbf{X}^{\text{train}}, \mathbf{y}^{\text{train}})$, $(\mathbf{X}^{\text{test}}, \mathbf{y}^{\text{test}})$, where

$$\forall y^{\text{train}} \in \mathbf{y}^{\text{train}}, \quad y^{\text{train}} \in [1, C]$$
$$\forall y^{\text{test}} \in \mathbf{y}^{\text{test}}, \quad y^{\text{test}} \in [1, C]$$

In addition to the train and test sets, query images are retrieved from an online image search engine separately for each category. The retrieved images are denoted $(\mathbf{X}^{\text{query}}, \mathbf{y}^{\text{query}})$, where

$$\forall y^{\text{query}} \in \mathbf{y}^{\text{query}}, \quad y^{\text{query}} \in [1, C]$$

3.2. Training CNN

In order to train a CNN without pseudolabels, training images are propagated forward through the network in batches to produce outputs, for which error gradients are calculated. To complete an iteration, these are backpropagated, in batches. This process is repeated until convergence. Throughout training, the accuracy of the network is typically evaluated on the test set for monitoring.

All images are resized so that the smaller dimension is p pixels, and a central crop of $p \times p$ pixels is extracted. This has been shown to work better than other cropping methods [2].

3.3. Pseudolabels with Retrieved Images

The method published here relies on a different Pseudo-Label selection mechanism and a different Pseudo-Label weighting to the original approach [15]. When training with pseudolabel data, the CNN is trained as described in section 3.2. However, $\mathbf{X}^{\text{query}}$ images are repeatedly evaluated with the current CNN, and some are selected with pseudolabels $\mathbf{X}^{\text{pl}}$, for training.

At the beginning of training, $\mathbf{X}^{\text{pl}}_0$ is empty.

$$\mathbf{X}^{\text{pl}}_0 = \emptyset$$

For the first i iterations (during epoch 0), the CNN is trained only with $(\mathbf{X}^{\text{train}}, \mathbf{y}^{\text{train}})$. Then, $\mathbf{X}^{\text{query}}_0$ is propagated forward through the CNN, to produce a set of vectors of beliefs for all labels $\mathbf{b}_0$ for every image. These beliefs correspond to the normalized outputs of the last fully connected layer, before applying the last softmax layer.

Then, a randomized selection process chooses which predicted labels $\mathbf{y}^{\text{query}}$ will be trusted. $\mathbf{X}_e^{\text{pl}}$ from the previous epoch are not included.

$$(\mathbf{X}_{e+1}^{\text{pl}}, \mathbf{y}_{e+1}^{\text{pl}}) = selected(\mathbf{X}^{\text{query}} \setminus \mathbf{X}_e^{\text{pl}}, \mathbf{y}^{\text{query}}, \mathbf{b}_e)$$

The selection method proposed in this paper is explained in section 3.4. The rest of $\mathbf{X}^{\text{query}} \setminus \mathbf{X}_e^{\text{pl}}$ is unused in this epoch.

This is the end of epoch 0. In each following epoch e, the CNN is trained with $\{(\mathbf{X}_e^{\text{pl}}, \mathbf{y}_e^{\text{pl}}), (\mathbf{X}^{\text{train}}, \mathbf{y}^{\text{train}})\}$. Section 3.5 discusses how $\mathbf{y}_e^{\text{pl}}$ can be weighted against $\mathbf{y}^{\text{train}}$ for better convergence stability.

3.4. *Pseudolabel Selection*

A number of factors affect the quantitative benefit of using pseudolabeled images: dataset belief, accuracy of the selection method, difference between datasets, selection variability over epochs, and randomization. Our selection method balances these by selecting images in a randomized way, depending on class accuracies and classifier belief for the correct class.

The accuracy λ_c for each class c on unlabelled data is the ratio of images classified as class c to the number of queried images in class c. By making the weak assumption that retrieval class accuracies across queried data are similar, class accuracies λ_c for the classifier are an indicator of training data and class complexity for each category.

Classes with higher accuracy on the query dataset are given lower pseudolabel priority. This is accomplished with the $(1 - \lambda_c)$ factor.

Another factor in selection is classifier belief. By using the normalised the belief in the $\mathbf{y}^{\text{query}}$ class, the selection favours images the classifier is more confident about, thus removing incorrect query images. This belief is normalized across network responses.

The last step is randomisation. A portion of query images is randomly removed during selection. In our experiments, we chose to remove 50%, and found this beneficial. This is justified by a need to regularize across data when the CNN is trained.

Hence, each example image is chosen with probability:

$$\frac{(1 - \lambda_c) * b_e}{2}$$

3.5. *Pseudolabel Weighting*

Pseudolabels are likely to affect the classifier adversely when it hasn't yet reached a sufficient accuracy, just as the classifier would train badly on raw query data. Self-training is prone to quickly converge to suboptimal solutions, because the classifier assigns high confidence to wrong examples. How this is mitigated is explained below.

In the original pseudolabel paper, images from the training set had constant weights, and the pseudolabel losses were weighted by α, where α increases with time according to two hyperparameters.

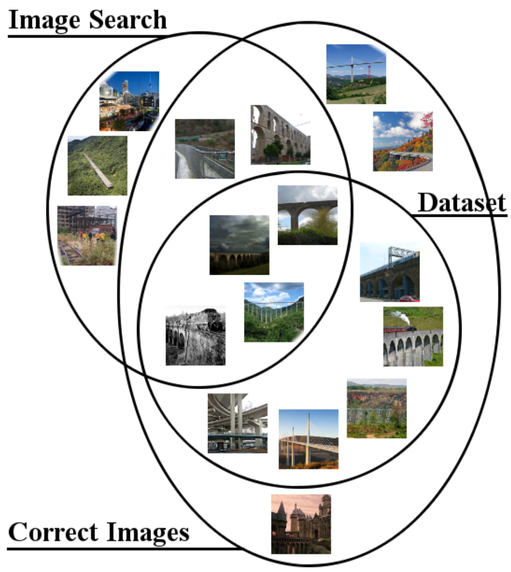

Figure 3: Example images of the viaduct class

Our experiments showed that this method is not more effective than setting $\alpha = 0$ until the network reaches near-top accuracy, and then setting $\alpha = 1$. This method crucially relies on the network's ability to create a weak classifier from the training data alone, and we found that this is the case with the previously published α tuning method as well.

This weighting method, albeit crude, simplifies hyperparameter tuning, and at the cost of a few epochs, achieves the same accuracy.

3.6. *Dataset Belief*

For an automatically retrieved set of images, a crucial information for deciding whether to use Pseudo-Labels is the query accuracy of the retrieved data. The unknown proportion of images which belong to the queried category is B, or dataset belief.

Query images can be wrong, misleading, and contain correctly and incorrectly labelled images from the training dataset, see Figure 3.

An imperfect selection must vary over epochs, in order to mitigate convergence to a non-median representation of the category.

3.7. *Difference Between Datasets*

If the training dataset and the images retrieved from Image Search are the same, the method will not be of benefit. It is important that they are complementary, albeit with an overlap, and that they disagree to a degree. The disagreement creates jitter between images where the classifier should not be divisive, and supports convergence to a decision boundary elsewhere.

We found that selecting (X^{query}, y^{query}) which fully agrees the current classifier does not boost classifier accuracy over not using pseudolabels at all. This is because the data don't create disagreement, and therefore no novelty. In our experiments, we found that a degree of wrong and randomly labelled images

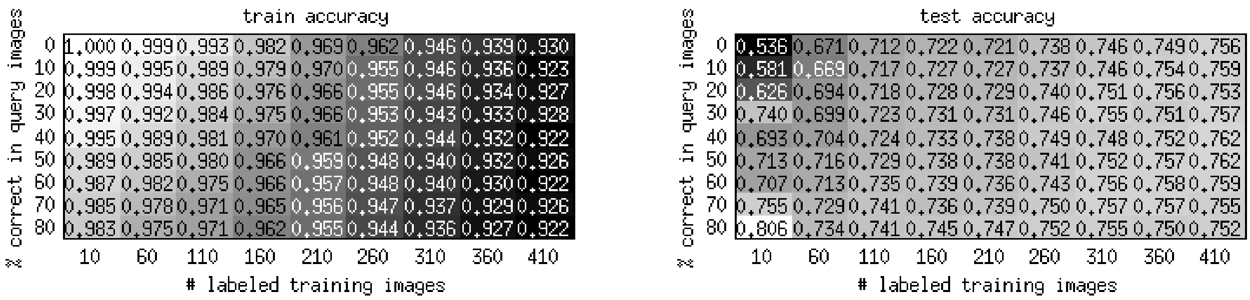

Figure 4: Train and test accuracies with varying correct query images, and varying train set sizes for each class

helped the classifier to converge to higher accuracy over the test set. Adding this form of noise achieves regularisation.

3.8. Implementation

All images $\mathbf{X}^{\text{train}}, \mathbf{X}^{\text{test}}, \mathbf{X}^{\text{query}}$ were resized so that the smaller dimension is 224 pixels, and a central crop of 224×224 pixels is extracted. This has been shown to work better than other cropping methods [2].

The AlexNet architecture was used, and initialized with weights trained on the ImageNet dataset. The network was retrained by keeping all but the last fully connected layer locked, and updating weights on the last layer. This was shown to achieve the best results.

The network was trained over 100 epochs of 500 iterations each with each combination of parameters.

The ratio of testing to retrieved accuracies is an indicator of the retrieved datasets accuracy or similarity. Assuming no constructive errors, such as those CNNs have been demonstrated to fall to when synthesizing examples [19], the number of correctly classified images is a lower bound on how many really belong into the category. A large difference between this number and the actual number (B), directly indicates how much further benefit the new data can have for training.

4. Results

We performed experiments in two setups: the 6 most numerous SUN 397 classes, artificially divided into "labeled" and "query" subsets, and the full SUN 397 dataset with images retrieved from Google's Image Search. For each set of train, test, and pseudolabel accuracies in figures 4 and 5, the network was trained independently.

4.1. Artificial Dataset

By varying the percentage of correct images in the "query" subset, it was possible to analyse the tolerance of the algorithm. The 6 classes with most images in the SUN 397 dataset contain 1126 to 2439 images, and these were divided into training, testing, and query subsets. The query subset was then diluted with images from all other SUN 397 classes to varying degrees. Experimental results are in Figure 4.

Training accuracy, which increases beyond testing accuracy when overtraining, goes down with more training images, as well as with a higher proportion of correct query images. This demonstrates that by applying our method, overtraining is being mitigated. Test accuracy benefits most from pseudolabels with 60 to 160 training images per class, and only when there are at least 20% correct images in the query dataset.

Interestingly, with only 10 training images per class and highly accurate query data, classifier accuracy fluctuates, and sometimes reaches better results than by using the same amount of correct images by training without pseudolabels. This may be because the classifier is able to ignore outliers among training images, which correspond unhelpful examples.

4.2. Full SUN 397 dataset

Train and test images are retrieved from the SUN 397 dataset. These are divided into a train set and test set randomly, by using n for training, and the rest for testing. We performed experiments with $n = [5, 20, 50]$

The query set were retrieved from Google's image search separately for each category, by searching the full name of the SUN 397 category (ex.: "swimming pool indoor"), and retrieving all full scale original images. Only images which produced an erroneous http query were ignored, and the number of images found was between 230 and 1 359, with mean 796. There is a total of 316 024 images, see Figure 6 for the distribution of counts. An automated image similarity search was applying to remove duplicate images, to avoid overcounting problems.

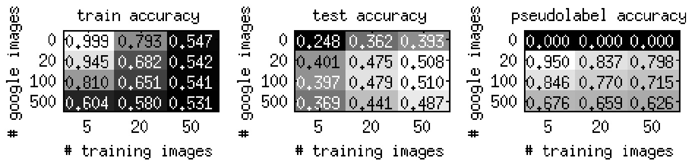

Figure 5: train, test, and Pseudo-Label accuracy sets with SUN 397 supplemented by online query images, for various numbers of training images and top google image search queries. Top rows are without Pseudo-Labels.

Figure 5 shows the accuracy distribution across classes with and without pseudolabels. Note that the quality of retrieved images decreases with additional images, offsetting the benefit from Pseudo-Labels on larger queries. We can see that the pseudolabel approach reaches higher accuracy than classifiers trained without it, and that too many additional Google images

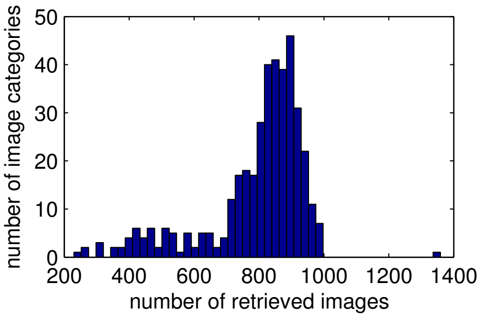

Figure 6: Image counts for categories retrieved through Google

are detrimental. This may be because they vastly outnumber true images, and true labels are drowned out in noise.

5. Conclusion and Future Work

The goal of this paper was to demonstrate that CNN training in the semi-supervised setting can be beneficial with small datasets supplemented by images retrieved from online image search. Experimental results demonstrate that this method is of significant benefit especially if the number of training samples is small (60 - 160), or the images in the training sample are not as representative as the query data.

By adapting pseudo-labels to real-world datasets, groundbreaking results have been accomplished, facilitating progress in classification and localization where image data is sparse. The method was justified, experimentally analysed, and validated. Finally, state-of-art results were presented on the SUN 397 dataset with few images in each category.

Future work includes searching for a way to work with online image search data only, without the need for a labelled dataset, or with a labelled dataset of a few images only. The existing method does not work with useful data from other categories, and may benefit from a model of interclass relationships. In addition, retrieving images from a given category is a non-trivial problem, and the method may be expanded to handle further information on where query images came from, and map intra-class disparities. It will also be beneficial to experiment on classifiers created with subsets of the ImageNet dataset, so that accuracy can be compared for various class sizes even for hierarchical classes of varying complexity.

6. References

[1] O. Chapelle and A. Zien. Semi-supervised classification by low density separation. *AI STATS 2005*, 2004.

[2] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014.

[3] R. G. Cinbis, J. Verbeek, and C. Schmid. Multi-fold mil training for weakly supervised object localization. In *CVPR*, pages 2409–2416. IEEE, 2014.

[4] J. J. DiCarlo, N. Pinto, and D. D. Cox. Why is real-world visual object recognition hard? 2008.

[5] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

[6] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2014.

[7] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[8] A. P. P. Griffin, G. Holub. The caltech 256. 2007.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *arXiv preprint arXiv:1406.4729*, 2014.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[11] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[13] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[15] D. Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. *Workshop on Challenges in Representation Learning, ...*, 2013.

[16] W. Li, L. Duan, I. W. Tsang, and D. Xu. Batch mode adaptive multiple instance learning for computer vision tasks. In *CVPR*, pages 2368–2375. IEEE, 2012.

[17] W. Li, L. Duan, D. Xu, and I. W. Tsang. Text-based image retrieval using progressive multi-instance learning. In *ICCV*, pages 2049–2055. IEEE, 2011.

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer, 2014.

[19] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *CVPR*, 2015.

[20] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Weakly supervised object recognition with convolutional neural networks. Technical Report HAL-01015140, INRIA, 2014.

[21] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *CVPR*, pages 512–519. IEEE, 2014.

[22] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 2015.

[24] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Compressed fisher vectors for large-scale image classification. *Rapport de recherche RR-8209, INRIA*, 2013.

[25] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(4):754–766, 2011.

[26] F. Schroff, A. Zisserman, and A. Criminisi. *Semantic image segmentation and web-supervised visual learning*. PhD thesis, University of Oxford, 2009.

[27] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. Dec. 2013.

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[29] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. *CoRR*, abs/1409.4, 2014.

[31] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6, 2013.

[32] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, pages 1701–1708. IEEE, 2014.

[33] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *CVPR*, pages 1521–1528. IEEE, 2011.

[34] K. E. Van De Sande, T. Gevers, and C. G. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[35] S. Vijayanarasimhan and K. Grauman. Keywords to visual categories: Multiple-instance learning forweakly supervised object categorization. In *CVPR*, pages 1–8. IEEE, 2008.

[36] L. Wan, M. Zeiler, S. Zhang, Y. L. Cun, and R. Fergus. Regularization of neural networks using dropconnect. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1058–1066, 2013.

[37] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. Cnn: Single-label to multi-label. *arXiv preprint arXiv:1406.5726*, 2014.

[38] R. Wu, S. Yan, Y. Shan, Q. Dang, and G. Sun. Deep Image: Scaling up Image Recognition. *arXiv preprint arXiv:1501.02876*, 2015.

[39] Z. Wu, Y. Zhang, F. Yu, and J. Xiao. A gpu implementation of googlenet.

[40] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492. IEEE, 2010.

[41] S. Yan, J. Dong, Q. Chen, Z. Song, Y. Pan, W. Xia, Z. Huang, Y. Hua, and S. Shen. Generalized hierarchical matching for sub-category aware object classification. In *Visual Recognition Challange workshop, ECCV*, volume 5, 2012.

***Research Highlights (for review)**

-The problem of training deep models with small amounts of data is tackled
-A mechanism for augmenting labeled data with noisy data is proposed
-The weighting and selection processes of the pseudolabel approach are improved
-The proposed mechanism reaches state-of-art accuracy on the SUN 397 dataset