Experiments Around Training Data Selection Methods for Image Classification

David Crandall[*]
John Koo[†]
Michael Trosset[‡]

December 16, 2019

Abstract

Convolutional neural networks (CNNs) are now widely utilized to fit highly accurate image classification models. However, in order to achieve these results, CNNs require vast amounts of training data, especially as the size of these networks grows in an effort to achieve increasingly better performance. In real-world applications, large amounts of training data are often difficult to obtain due to data collection and labelling limitations or difficult to work with due to computational limitations. Expanding upon on previous work by Bambach, Crandall, Smith, and Yu [1], our work explores various methods for subsampling training images under a data budget for fitting an image classification model and compares the results against uniform random sampling. Our methods make use of image embeddings to determine image diversity and outlyingness.

Introduction

LeNet-5, 1998 [8], was the first widely recognized CNN architecture for image classification. Consisting of only seven layers,three of which are convolutional, training this network on 32×32 greyscale images of handwritten digits involved fitting 60,000 model parameters. Over time, as larger datasets and more powerful computer hardware became available, CNN architectures grew deeper and more complicated: AlexNet, 2012 [7], consists of 8 layers and 60M parameters, and VGG16, 2015 [11], consists of 41 layers and 138M parameters). Due to the massive number of training parameters, these deep models require large amounts of data to prevent overfitting. In fact, it has been observed that performance gains can continue to scale with the training set size, even into the billions [9].

However, obtaining large datasets for deep models is not always feasible. Manually labelling thousands or even millions of images can be tedious, time-consuming, and expensive [2]. Some proposed and empirically verified solutions to the limited data problem include using a smaller network with fewer parameters, starting with a pre-trained model, and image augmentation [10]. In this paper, we propose methods for selectively choosing training images under a set data "budget" and discuss how they compare against uniform random sampling.

Previous Work

Bambach et al. [1] demonstrated that given a large pool of images and a fixed training set size, it is possible to tailor the training set for fitting a VGG16 network that results in better or worse performance on a separate test set. They then described two characteristics of the datasets that seemed to correspond to model performance: object size (how much of the image the object takes up) and diversity (after embedding the images in $\mathbb{R}^d$, how much space the training point cloud takes up). Their study showed that model performance correlated positively with object size in the training set, and training sets consisting of "diverse" images tended to outperform those consisting of "similar" images.

[*]Department of Computer Science, Indiana University Bloomington
[†]Department of Statistics, Indiana University Bloomington
[‡]Department of Statistics, Indiana University Bloomington

Replication Study

The data in the above study is as follows:

- Training images were sampled from the frames of first person video feeds, from the point of view of toddlers and parents playing with one of 24 toys. The video was taken with a 70° lens. Bounding boxes of the toys were drawn for each image to determine the size and location of the toy. The images were blurred around the object using the bounding box information to simulate visual acuity.
- Validation and testing sets consisted of artificially generated images of the 24 toys.

Two experiments were then performed on these data. Both experiments involved fitting VGG16 networks on a particular training set.

The size experiment can be described as follows: Frames were randomly sampled from the video feeds and ranked according to object size (median of around 10%). These were then split into a training set of “big” objects and a training set of “small” objects. It was shown that the model fit on the big objects outperformed the model fit on the small objects when comparing test accuracies. The images were also cropped into the object to simulate varying focal lengths from the original 70° down to 30° in increments of 10°, and the croppedimages outperformed the original images, further supporting this result.

The diversity experiment can be described as follows: Again, frames were randomly sampled from video feeds. These frames were then embedded into Euclidean space using GIST features [12]. Three training subsets were sampled based on the GIST features: a “diverse” subset that maximizes pairwise distances, a “similar” subset that minimizes pairwise distances, and a “random” subset. Models fit on the random subset outperformed the models fit on the diverse subset which outperformed the models fit on the similar subset, using test accuracy to compare models. Images were again cropped to simulate various focal lengths, and lower focal lengths again resulted in better model performance.

We attempted to replicate this study using the Stanford Dogs dataset [5], which consists of around 20,000 images of 120 dog breeds. Most images contain one dog per image, and images that contain multiple dogs were discarded. For each breed, 100 images were randomly selected for the training set (which were further divided into 50-50 training subsets based on the experiment), 25 images were randomly selected for the validation set, and the rest were set aside for testing. No blurring was applied to these images. It is assumed that all images were taken with a 70° lens. Each experiment was replicated 10 times.

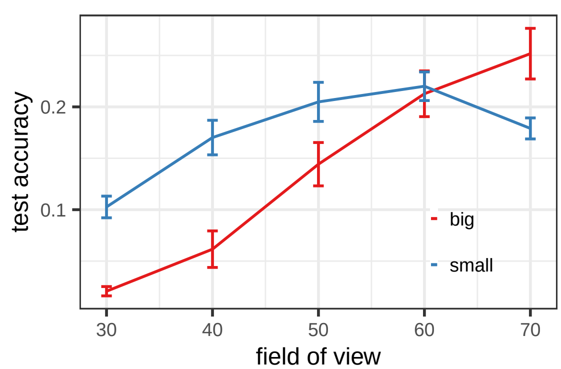

Figure 1: "Size" experiment on the Stanford Dogs dataset. Error bars indicate ±1 standard deviation from the mean of 10 repetitions.

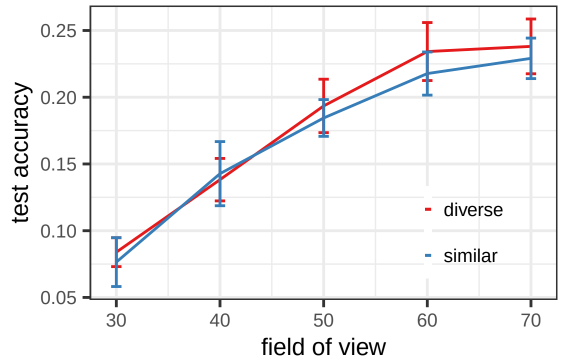

Figure 2: "Diversity" experiment on the Stanford Dogs dataset. Error bars indicate ±1 standard deviation from the mean of 10 repetitions

Fig. 1 shows that for the original, uncropped images (FoV = 70°), models fit on the “big” subset tend to outperform models fit on the “small” subset. However, instead of increased performance with reduced field of view, we observed the opposite: as the images were cropped into the objects of interest, the resulting models performed worse, and this was especially true of the “big” subset. Closer inspection of the training images reveals that the median bounding box coverage was around 50%, compared to 10% of the toys dataset, and cropping often resulted in cutting off parts of the object of interest (Fig. 3).

Results of the diversity experiment (Fig. 2) suggest that there may be some difference between the models fitted on “diverse” vs. “similar” training sets at the original focal lengths, but it is not clear if this is a significant result.

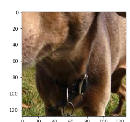

Figure 3: An image from the Stanford Dogs dataset cropped from an assumed 70° FoV to 30°.

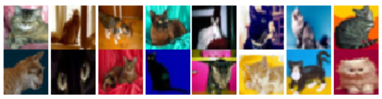

Figure 4: Images of cats selected by the diverse sampling method.

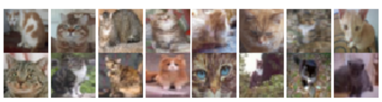

Figure 5: Images of cats selected by the similar sampling method.

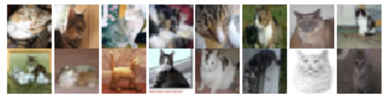

Figure 6: Images of cats selected by the random sampling method.

Focusing just on diversity, we tried a similar experiment using the CIFAR-10 [6] dataset, which consists of 10 object classes, each class consisting of 5,000 training images and 1,000 test images. Instead of varying focal lengths (which is not possible with CIFAR-10 images), we varied the training set sizes. For a given training set size n, we sampled $2n$ images from each class and set aside half for validation. The sampling was done according to the "diverse", "similar", and "random" sampling methods (see Figs. 4, 5, 6).

Other Related Work

Wang et al. [13] proposed a two-round training approach to better fit CNNs on a subset of the training set. In their approach, A CNN was fit on a large training set, then for each image in the training set, an influence measure was computed over all images in the validation set. If the influence was negative, then that training image was discarded. The CNN was then fit again using the reduced training set, and the resulting model was observed to achieve a higher test accuracy than the original model trained prior to subsetting the training data.

Kaushal et al. [4] also proposed an active learning approach to limit the size of training data. Their method is as follows: First, a small subset of a large training set is selected, and a model is fit on the subset. Predictions are then made on the unselected images in the training set based on this model. Images are chosen based on the uncertainty of the model prediction and added to the training subset, and the model is refitted using the larger training subset. This is then repeated over multiple rounds, increasing the size of the training subset each round. The study demonstrates that this method outperforms training on randomly sampled subsets of the same size.

Ferreira [3] proposed a maximum entropy based subset selection method for selecting training data, given that the inputs are of the form $x_i \in \mathbb{R}^d$. Their method starts with a large training set to estimate the density of the feature space.

Wilson [14] demonstrated that edited k-nearest neighbors classifiers outperform regular "unedited" k-nearest neighbors classifiers. The Wilson editing method is described as follows: Given a training set $X_1, ..., X_n \in \mathbb{R}^d$ with corresponding discrete labels $Y_1, ..., Y_n \in \{1, 2, ..., q\}$, use leave-one-out cross-validated k-nearest neighbors to determine $\hat{Y}_1, ..., \hat{Y}_n$. Discard $i \in \{1, ..., n\}$ where $Y_i \neq \hat{Y}_i$ to construct a reduced, edited training set. Finally, fit a new k-nearest neighbors model on the reduced training set. Wilson's results show that the model fit on the edited data tend to outperform models fit on the entire training set, suggesting that outliers in the training set are detrimental to the resulting model performance.

Based on the literature, training set selection methods can be classified as denoising/filtering methods or as diversification methods. Denoising and filtering methods remove outliers or atypical observations, while diversification methods aim to make training observations as different from each other as possible. These two ideas appear to be at odds with one another. There also doesn't appear to be as much literature on how to apply such methods to image data, as they assume that the data can be naturally represented in Euclidean space (i.e., as a data matrix). However, this does provide some sense of how we can "construct" a good training sample: Given a "good" embedding of images such that the embedding space can be separated by some set of manifolds into regions that correspond to each class, for each class, we should choose a training sample that fills up that class' region without crossing over into any other regions.

One thing that is not clear is how this relates to the data collection process. Most previous studies sample from a pool of preexisting images, but in a more practical scenario, the data collection process would involve creating new images (e.g., by taking photographs).

Experiments and Results

Conclusions and Future Work

References

[1] Sven Bambach, David Crandall, Linda Smith, and Chen Yu. Toddler-inspired visual object learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1201–1210. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7396-toddler-inspired-visual-object-learning.pdf.

[2] Per H. Christensen and Wojciech Jarosz. The path to path-traced movies. *Foundations and Trends® in Computer Graphics and Vision*, 10(2):103–175, 2016. ISSN 1572-2759. doi: 10.1561/0600000073. URL http://dx.doi.org/10.1561/0600000073.

[3] Pedro M. Ferreira. Unsupervised entropy-based selection of data sets for improved model fitting. *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3330–3337, 2016.

[4] Vishal Kaushal, Rishabh Iyer, Suraj Kothawade, Rohan Mahadev, Khoshrav Doctor, and Ganesh Ramakrishnan. Learning from less data: A unified data subset selection and active learning framework for computer vision, 2019.

[5] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

[6] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999134.2999257.

[8] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[9] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining, 2018.

[10] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning, 2017.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[12] Antonio Torralba. Contextual priming for object detection. *Int. J. Comput. Vision*, 53(2):169–191, July 2003. ISSN 0920-5691. doi: 10.1023/A:1023052124951. URL https://doi.org/10.1023/A:1023052124951.

[13] Tianyang Wang, Jun Huan, and Bo Li. Data dropout: Optimizing training data for convolutional neural networks, 2018.

[14] Dennis L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans. Systems, Man, and Cybernetics*, 2:408–421, 1972.