Identifying Measures of Training Data Quality for Classifying Images

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. Our work aims to address this problem by setting a training data "budget" by limiting the number of images allowed for model fitting and developing a method of choosing training images for the best model performance under this constraint. We explore measures of image quality and diversity to define a training data quality metric, inspired by image embedding methods as well as various measures of performance for generative adversarial networks (GANs), and show how these correlate with fitted model performance.

Introduction

LeNet-5 [Lecun et al., 1998] 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 [Krizhevsky et al., 2012], consists of 8 layers and 60M parameters, and VGG16 [Simonyan and Zisserman, 2014], 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 [Mahajan et al., 2018].

However, obtaining large datasets for deep models is not always feasible. Manually labelling thousands or even millions of images can be a tedious, time-consuming, and expensive [Christensen and Jarosz, 2016]. 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 [Perez and Wang, 2017]. In this paper, we propose methods for selectively choosing training images under a set data "budget" and discuss various measures and statistics on training sets that correlate to model performance on a separate test set.

Previous Work

[Bambach et al., 2018] 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.

Replication Study

The data in the above study can be described as follows:

- Training images were sampled from the frames of first person video feeds. Each image contained one of 24 toys that the toddlers were playing with. The video was taken with a 70° lens. Bounding boxes of the toys were drawn for each image to determine how much of the image each toy took up.
- The training images were blurred around the object to simulate acuity.
- Validation and testing sets consisted of artificially generated images of the 24 toys.

The size experiment was as follows:

- A training set was selected of the “largest” objects (images in which the object we wish to classify took up more of the image). Another training set was selected of the “smallest” objects (images in which the object took up less of the image).
- The images were cropped to simulate different fields of view between 30° and the original 70°.
- For each object size training subset and field of view, a VGG16 network pretrained on the ImageNet dataset was fit to these images using a fixed set of hyperparameters. Then performance was measured on the test set. This was repeated 10 times to obtain interval estimates for performance.

The results of this experiment suggested that using a training set of “larger” images results in better model performance than using a training set of “smaller” images. In addition, decreasing the field of view (i.e., cropping into the object of interest) of the training images resulted in progressively better performance. This suggests that in order to obtain the best training set, the objects we wish to classify must be prominently displayed in the images.

The diversity experiment was as follows:

- The training images were embedded into high dimensional Euclidean space using GIST features [Torralba, 2003].
- Points were sampled from this embedding using a greedy algorithm to maximize the distance between the points. The corresponding images became the “diverse” training set while the reamining images became the “similar” training set.
- VGG16 networks pretrained on the ImageNet dataset were fit to the two training sets, using the same fixed set of hyperparameters as before. This was repeated 10 times. Performance was measured on the test set. Field of view was also adjusted incrementally as in the previous experiment.

The results of the diversity experiment suggest that the “diverse” training set results in a better performing model than the “similar” training set. As before, decreasing the field of view improved performance.

Stanford Dogs Dataset

In order to replicate the study as closely as possible with a different dataset, we needed an image classification dataset consisting of only one object/class per image and yet also containing either bounding box or segmentation mask information for each image, in order to both measure the object size in the image as well as zoom into the object to simulate different fields of view. One dataset that contains this information is the Stanford Dogs dataset [Khosla et al., 2011], consisting of around 20,000 images of 120 different dog breeds. We assumed that the original images were taken with a 70° field of view. Some of the images were of multiple dogs, and these images were discarded. For each breed, 100 images were randomly selected as the training set, 25 images were randomly selected as the validation set, and the rest were set aside for testing. We did not blur these images.

Size Experiment

Using the bounding box information, we calculated the proportion of the image that the dogs took up. Then for each breed, we split the 100 training images so that the “large” dataset contains images in which the dogs take up more of the image and the “small” dataset contains images in which the dogs take up less of the image. We also incrementally zoomed into the center of the bounding boxes to simulate lower fields of view.

The results (Figure 1) suggest that when using the unscaled images (70° field of view), we obtain better model performance when training on images in which the dogs take up more of the image. This is consistent with the results of the previous study. However, in our results, we do not get progressively better performance as we decrease the field of view. Visual inspection of the images shows that the dogs already take up most of the image in many cases, compared to the toys taking up relatively little of the images in the toys dataset (median bounding box proportion of around 50% for the dogs dataset vs. around 10% for the toys dataset). As we crop the images to simulate lower fields of view, in some cases, we end up cropping so much that we are left with just a patch of fur in our image. We also suspect that since in the toys dataset, the validation and testing images prominently display the toys, compared to the training dataset where the toys take up relatively little of the images, cropping in makes the training set look more like the testing set. No such disparity exists in the training vs. validation vs. testing sets for the dogs datasets. However, the fact that the "large" dataset outperformed the "small" dataset using the uncropped data suggests that there may be some relationship between model performance and object size.

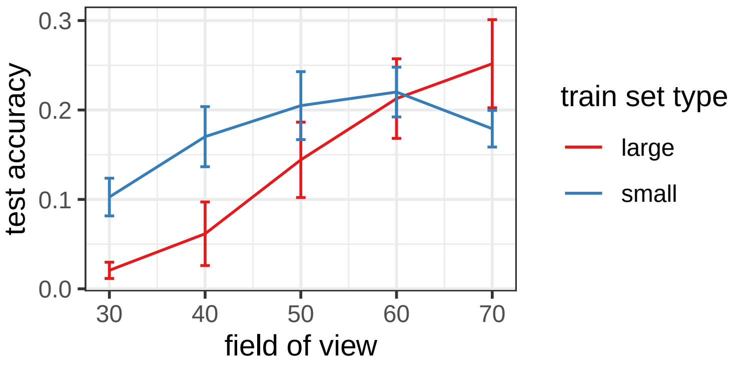

Figure 1: "Size" experiment on the Stanford Dogs dataset

Diversity Experiment

As in the previous study, we embedded the images using GIST features and for each breed, sampled 50 points to maximize Euclidean distance, and set aside the corresponding images as the "diverse" training set. The reamining 50 images for each breed were set aside as the "similar" training set.

As in the size experiment, decreasing the field of view reduced model performance. There was no significant difference in model performance between the "diverse" and "similar" training sets, although perhaps we would've observed a more pronounced difference if we increased the number of repetitions (Fig 2).

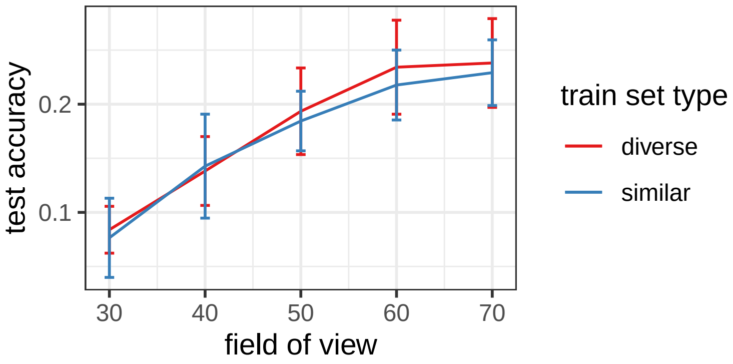

Figure 2: "Diversity" experiment on the Stanford Dogs dataset

CIFAR-10 and MNIST Datasets

Focusing just on training set diversity and ignoring object size and field of view allows us to expand this study to more datasets, as we no longer require bounding boxes or segmentation masks. Two datasets commonly used for image classification experiments are the CIFAR-10 [Krizhevsky, 2009] and MNIST [LeCun et al., 2010] datasets. Each consist of 10 object classes. We repeated the diversity experiment on each dataset, this time varying the training set size instead of the field of view.

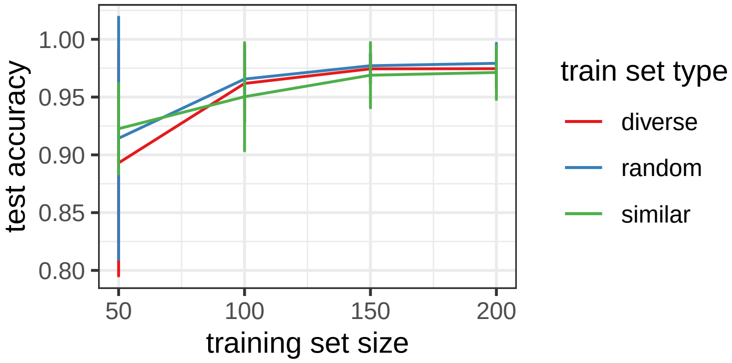

Figure 3: Diversity experiment results on MNIST data

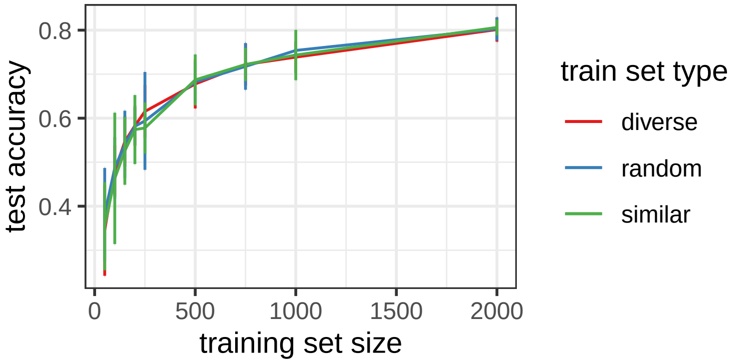

Figure 4: Diversity experiment results on CIFAR-10 data

Here, our results suggest no significant relationship between image diversity and model performance (Fig. 3, 4). However, this may perhaps be due to the fact that this classification task is easier. In particular, once we reach ~100 images per class in the MNIST data, all training sets attain ~95% accuracy. We tried repeating this experiment with very small datasets (Fig. 5, 6).

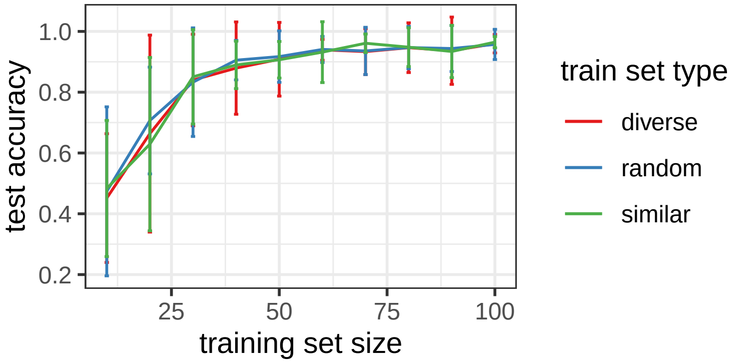

Figure 5: Diversity experiment results on small MNIST data

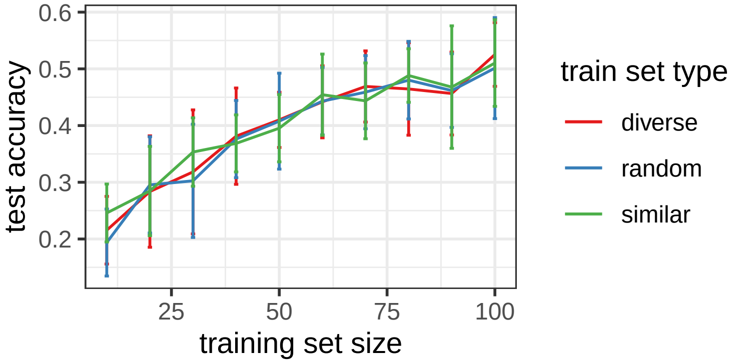

Figure 6: Diversity experiment results on small CIFAR-10 data

Next steps and ideas/concerns

- A previous study used qualitative measures of diversity to obtain better model performance. Perhaps we can try something like this on a dataset with sub-classes (e.g., CIFAR-100)—for each class, we sample images from subclasses using a multinomial distribution, the parameters of which we adjust depending on how diverse we want to make the training images (equal probabilities for a "diverse" set, skewed probabilities for a "similar" set).
- To more closely match the original study, a (relatively large) random subset of the original training data was set aside for validation. Perhaps if we match both the size and "diversity" of the validation set to the training sets, we would see a more pronounced difference (and this would be a more realistic scenario).
- Tweak the hyperparameters
- See how the diverse/similar image sampling method corresponds to other commonly used measures of image diversity, such as Inception score
- Look at performance consistency rather than just average performance (perhaps training on a diverse set lowers the probability of finding a bad local minimum)
- Consider different methods for selecting "diverse" or "similar" subsets from an embedding (although we'd want to avoid having to construct a full distance matrix)
- Consider different image embeddings
- More lit review on GAN performance measures (currently the only ones I can find require a pre-trained image classification model, which somewhat defeats the purpose)

- If we sample from a elliptical point cloud in $\mathbb{R}^d$ in such a way that we want to maximize pairwise distances, and if we draw fewer than $2d$ points, we will tend to pick points on the edges of the ellipse. In our case, GIST features are in $\mathbb{R}^{960}$, and we are drawing fewer than 2×960 points, so we are very likely only drawing exterior points. This might still be desirable, but I'm not sure what the implications of this would be.

References

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.

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.

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.

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

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.

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.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2, 2010.

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.

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

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

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.