Identifying Measures of Training Data Quality for Classifying Images

David Crandall
John Koo
Michael Trosset

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 and measures of dissimilarity and distance.

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 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.

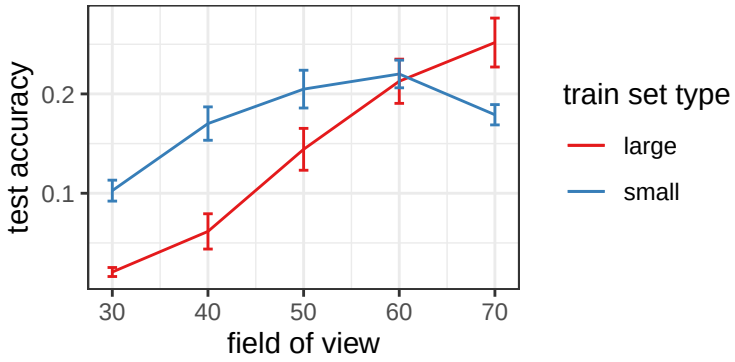

Figure 1: "Size" experiment on the Stanford Dogs dataset.

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).

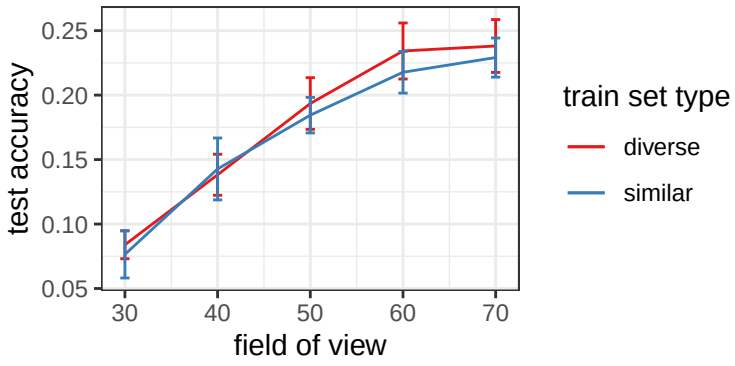

Figure 2: "Diversity" experiment on the Stanford Dogs dataset.

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.

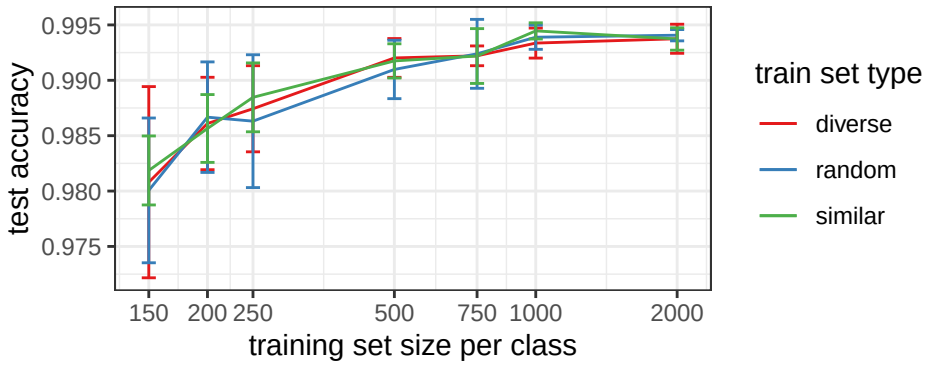

Figure 3: Diversity experiment results on MNIST data.

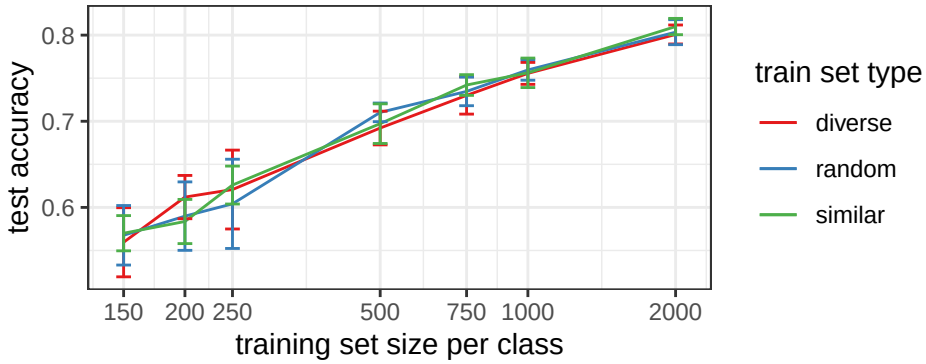

Figure 4: Diversity experiment results on CIFAR-10 data.

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 training samples (Fig. 5, 6). However, again, we see no significant difference in the three training data subsets.

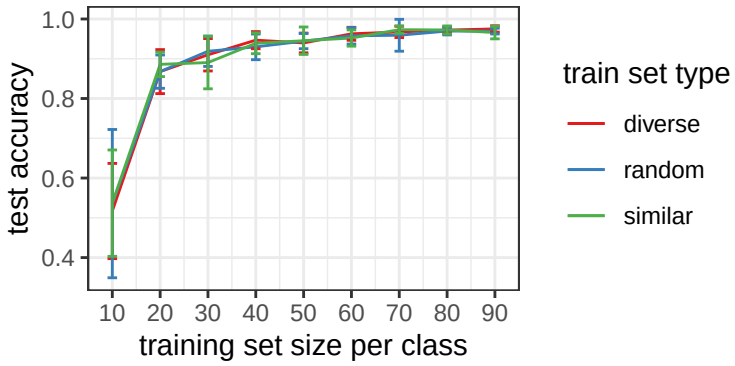

Figure 5: Diversity experiment results on small MNIST data.

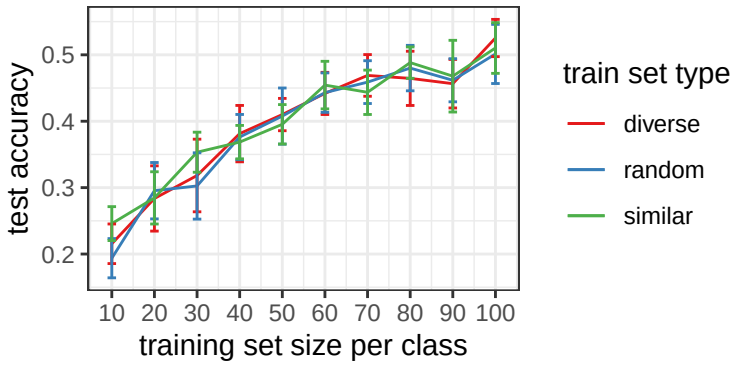

Figure 6: Diversity experiment results on small CIFAR-10 data.

Alternative Methods for Training Set Selection

In the previous experiments, the training and validation sets were selected as follows:

1. A validation set was randomly sampled from the training set such that it contains 1,000 images of each class.
2. Embed the remaining training images in $\mathbb{R}^{960}$ using GIST features.
3. From the reamining training images, sample n_{sub} images from each class such that their GIST-embedded points are as far from each other as possible (in practice, we use a greedy algorithm for this since it is not possible to try all possible combinations of points, and this induces randomness in the sampling). The corresponding images comprise the "diverse" training set.
4. Sample n_{sub} images from each class such that their GIST-embedded points are as close to each other as possible. This is done by selecting a random point and then selecting the $n_{sub} - 1$ nearest points. The corresponding images comprise the "similar" training set.
5. Sample n_{sub} images from each class uniformly at random without replacement.

Note that this method does not necessarily force the diverse and similar training sets to be disjoint (see Fig. 7). In Fig. 8, 9, 10, 10 cat images were sampled from a pool of 5,000 using the diverse, similar, and random sampling methods respectively.

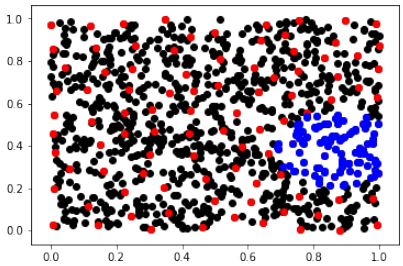

Figure 7: Diverse (red) vs similar (blue) samples taken from a uniform point cloud. The two samples are not disjoint, and some of the blue points may be masking some of the red points.

Figure 8: Images of cats selected by the diverse sampling method.

Figure 9: Images of cats selected by the similar sampling method.

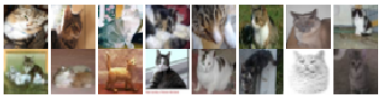

Figure 10: Images of cats selected by the random sampling method.

We believe that this method is as close as possible to the diversity experiment described by Bambach et al. [2018] using our datasets. However, this method raises some problems. First, if we were to think of this method as intelligently selecting a small training sample, then the large and randomly selected validation set seems out of place. Second, if we draw points from a $\mathbb{R}^d$ such that the points are as far from each other as possible and our sample size is $n \leq 2d$, and we think of the original sample as a point cloud, we would expect all of our points to lie on the edges of the point cloud with no interior points. This may manifest itself in selecting "outlier" images, or images that are unlike the typical image of its class. We can address the first issue by selecting our validation set as follows: First, draw $2n_{sub}$ images for both the diverse and similar training sets, then for each training set, set aside half for validation. This will be called the "validation from training" sampling method. We can address the second issue by first taking a random sample of size $2n_{sub}$ and then dividing that sample into diverse and similar subsets by choosing the diverse subset according to maximal distance and setting the remainder as the similar subset. We will call this the "subset before sampling training" sampling method.

Another possible way to address the second issue is by projecting the embedded points into a lower dimensional space before sampling.

[insert table of different sampling methods here]

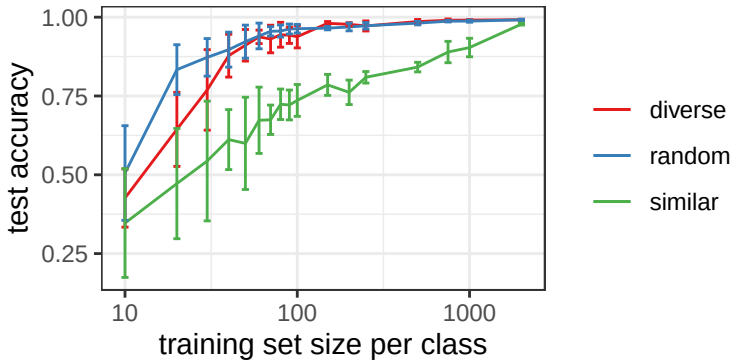

Figure 11: Diversity experiment on MNIST data, using the "validation from training" sampling method.

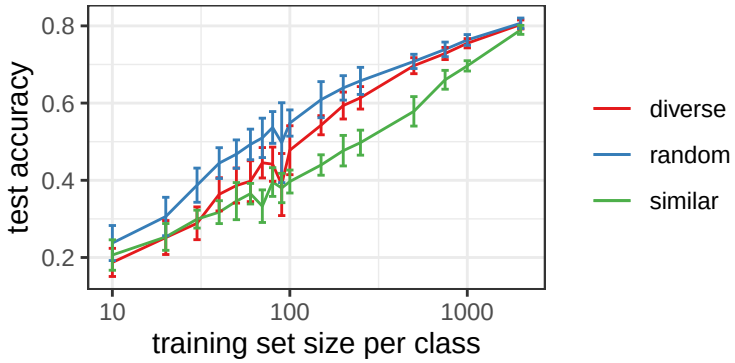

Figure 12: Diversity experiment on CIFAR-10 data, using the "validation from training" sampling method.

Figures 11 and 12 suggest that the diverse sampling method results in higher test accuracy than the similar sampling method when the validation sets are sampled from the training sets. However, for either dataset, the randomly sampled training sets perform as well or outperform the diversely sampled training sets. This may be because when the validation sets are large random samples, the model is able to generalize despite sub-optimal training sets. We can also see that for very small training set sizes, models trained on diverse samples tend to perform worse than models trained on random samples, and as training set sizes increase, the performance of models trained on samples selected with either method begin to coincide. One explanation of this might be that the diverse sampling method tends to choose outliers or atypical images, and as the sample size increases, the diverse sampling method exhausts the set of outliers and begins to sample more typical images. Previous studies show that classification models tend to perform better when outliers or atypical observations are removed from the training set.

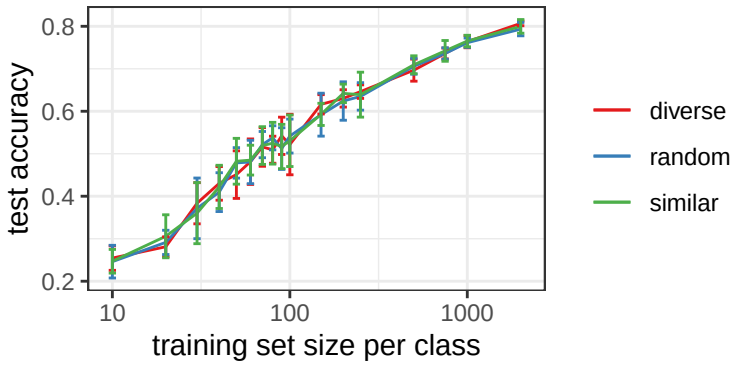

Figure 13: Diversity experiment on CIFAR-10 data using the "subset before sampling training" sampling method.

Fig. 13 shows that the diverse, similar, and random training samples using the "subset before sampling training" sampling method results in no significant differences in model performance.

Other Related Work

Wang et al. [2018] 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. However, this still requies training on a large dataset, and their goal is to fine-tune a large training set by dropping "bad" observations rather than building an optimal small training set.

Kaushal et al. [2019] 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 sampled, and a model is fit on the subset. Predictions are then made on the hold-out images in the training set based on this model. Finally, additional images are chosen based on the uncertainty of the model predictions, and these images are added to the training subset. This is then repeated over multiple rounds, increasing the size of the training data over each round. The study demonstrates that this method outperforms training on randomly sampled subsets of the same size.

Ferreira [2016] proposed a maximum entropy based sampling method for selecting training data, given that the inputs are of the form $x_i \in \mathbb{R}^d$.

Wilson [1972] demonstrated that edited k-nearest neighbors classifiers outperform regular "unedited" k-nearest neighbors classifiers. This suggests that outliers in the training sample are detrimental to model training.

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 much literature on how to apply such methods on 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).

New Proposed Methods

Training Set Measures

One question we would like to answer is whether we can define a measure or statistic on a training sample that correlates with model performance.

Edited Training Data

The data editing method described by Wilson [1972] is as follows:

1. Given a training set $\mathcal{X}$ such that $\mathcal{X} \subset \mathbb{R}^d$, $0 < i \leq n$, for each i, construct a kNN model from $\mathcal{X} \setminus \{x_i\}$ to obtain $\hat{y}_i$.
2. For each i, if $\hat{y}_i \neq y_i$, set $\mathcal{X} \leftarrow \mathcal{X} \setminus \{x_i\}$, to obtain the edited training set $\tilde{\mathcal{X}}$.
3. Construct a kNN model from $\tilde{\mathcal{X}}$.

The results of this paper demonstrate that models constructed in this way tend to outperform models constructed using the entire training set. The intuition behind this result is that the training set editing method removes per-class outliers to construct smoother decision boundaries.

The three training samples in this section are constructed as follows:

1. Draw a random sample from the training data.
2. Draw a "similar" sample from the training data.
3. Edit the training data using Wilson's data editing method, then draw a "diverse" sample from the edited training data.

The validation sets are subsetted from the three training samples after they are drawn (instead of drawing randomly from the original training set).

The intuition here is that the diverse sampling method may favor drawing "outliers", but we still want to impose diversity on the training sample, so the hope is that the data editing method will remove those outliers and the diverse sampling method can draw from a "clean" pool of images.

In our data editing method, we first embedded the images using GIST features, then used principal component analysis to reduce the dimensionality (using enough components to obtain 80% explained variance). Then we constructed a 5-NN model on the projected embedding to determine which images to remove.

Fig. 14 suggests that diverse sampling from an edited set results in worse model performance than just drawing a random sample. Fig. 15 shows the impact of editing before drawing a diverse sample.

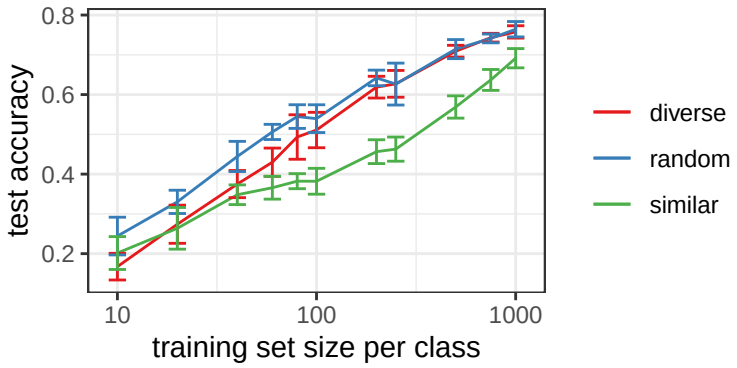

Figure 14: Diversity experiment on CIFAR-10 data, drawing from an edited dataset.

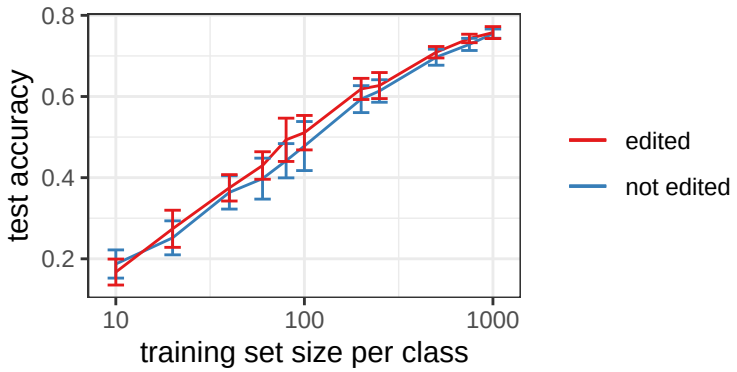

Figure 15: Impact of editing the training data prior to drawing a "diverse" sample

[**also try random sampling from the edited data**]

Clustering

We might be able to think of each class as having "subclasses". For example, one of the classes in the CIFAR-10 dataset is "dog", but we can further subdivide that into different breeds of dogs (or groups of breeds of dogs). It might be beneficial to have at least a couple representative images from each subclass. The sampling procedures tested in this section are described as follows:

1. Start with an overall training set. Optionally, perform Wilson editing to remove outliers.
2. Draw a random sample ("random" set).
3. For each class, perform k-means clustering. Then draw a per-cluster random sample ("random by cluster" set) and a per-cluster diverse sample ("diverse by cluster" set).

Again, we split validation sets from each of the training sets sampled by this procedure. Here, we chose $k = 10$.

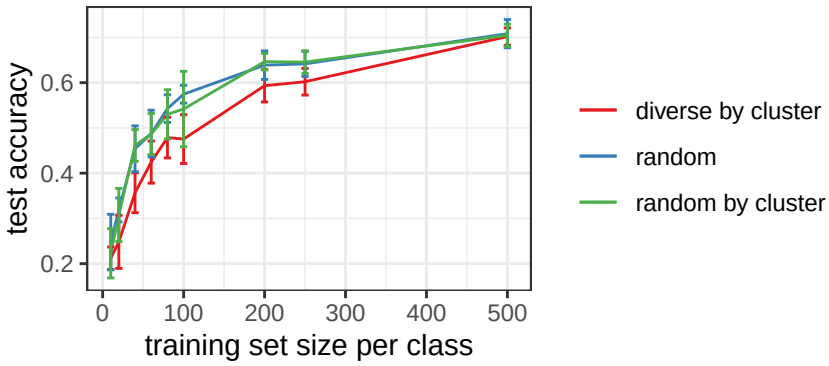

Figure 16: Diversity experiment on CIFAR-10 data, drawing from clusters.

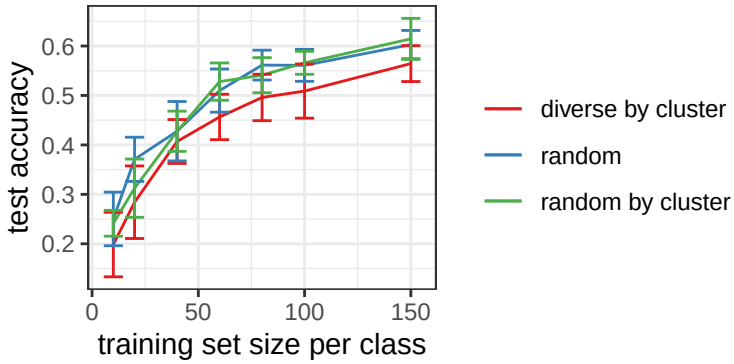

Figure 17: Diversity experiment on CIFAR-10 data, drawing from clusters. The overall training set was edited prior to sampling.

- One cluster per image
- Using clustering to remove outliers

Transfer Learning-Based Methods

- Using an embedding from the last layer of another CNN model

Active Learning-Based Methods

- Pre-training on a small subset to obtain an embedding

Methods Leveraging Extra Information

- CIFAR-100
- COIL-20 and COIL-100

Variational Autoencoders

- How to map from embedding to image

Conclusions and Next Steps

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.

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.

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.

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.

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

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