nature portfolio

https://doi.org/10.1038/s42256-022-00516-1

Supplementary information

Advances, challenges and opportunities in creating data for trustworthy AI

In the format provided by the authors and unedited

Supplementary Information

Details of the image classification experiments shown in Figure 1c. The image classification task we used is the "Cat vs. Dog" task from the MetaShift dataset[48]. The training set contains 1500 images, where 750 are cat images and 750 are dog images. Following the original MetaShift setting, this dataset is challenging because the class labels correlate spuriously with the image contexts: for the cat class, 80% of the cat images are indoor and 20% outdoor. For the dog class, 80% of the dog images are outdoor and 20% indoor. The test set contains 238 cat images and 240 dog images.

Furthermore, 8% of the training images have incorrect annotations which are more prevalent in less common contexts (e.g. some outdoor cats are mislabeled as dogs). We trained three popular deep learning architectures: ResNet-101, DenseNet-121, and VGG. The classifiers are trained for 15 epochs with a batch size of 32 on the training set and evaluated on the test set. The training objective is standard cross-entropy loss and the optimizer is SGD with a learning rate of 0.001, and a momentum of 0.9. The classifiers are initialized with their ImageNet pre-trained weights. We also randomly down-sampled the training set to 250, 500, 750, 1000, 1250 images and reported the experiment results in Figure 1c.

To demonstrate the effectiveness of data-centric approaches, the data Shapley[49] method is applied to the noisy training set. The Shapley value of each training point is estimated using the TMC-Shapley method[49]. We dropped the training data points with negative Shapley value, and trained the classifier on the remaining training data points. For example, on the full noisy training set, 45 out of the 1500 training images have a negative Shapley value and thus are removed. Training the ResNet classifier on the remaining 1255 training images led to substantial improvements. We repeated the aforementioned data Shapley value computation and filtering procedure when down-sampling the training set to generate the plot. All results are aggregated over 5 random seeds. The data and code for this experiment is available at https://github.com/Weixin-Liang/data-centric-AI-perspective.