Learning Latent Parameters without Human Response Patterns: Item Response Theory with Artificial Crowds

John P. Lalor[1]*, **Hao Wu**[2], **Hong Yu**[1,3,4]
[1]College of Information and Computer Sciences, UMass Amherst
[2]Department of Psychology and Human Development, Vanderbilt University
[3]Department of Computer Science, UMass Lowell
[4]Center for Healthcare Organization & Implementation Research, Bedford VA
john.lalor@nd.edu, hao.wu.1@vanderbilt.edu, hong.yu@umassmed.edu

Abstract

Supplemental material for the EMNLP-IJCNLP submission "Learning Latent Parameters without Human Response Patterns: Item Response Theory with Artificial Crowds."

A Applications in other domains

To further demonstrate the usefulness of learned latent parameters, we provide additional results here for two vision data sets: MNIST and CIFAR.

A.1 MNIST

The MNIST data set (LeCun et al.) is a data set of handwritten digits from 0 to 9. It includes 60,000 training examples and 10,000 test examples, and is regularly used to benchmark new machine learning models. With MNIST, we use a straightforward convolutional neural network (CNN) architecture (LeCun et al., 1995) with two convolutional layers and two fully connected layers with ReLU activations (Nair and Hinton, 2010). Max-pooling layers (Krizhevsky et al., 2012) are included after both convolutional layers, and there is a dropout layer between the first and second fully connected layers (Srivastava et al., 2014). Models were trained for 100 epochs using stochastic gradient descent (SGD) with a learning rate of 0.01 and momentum of 0.5.

A.2 CIFAR

The CIFAR data set (Krizhevsky and Hinton, 2009) is another popular image recognition data set where each image is associated with 1 of 10 classes. Class labels include "dog," "automobile," and "truck." CIFAR consists of 50,000 training examples and 10,000 test examples. For the CIFAR experiments we use the VGG-16 CNN model (Simonyan and Zisserman, 2014), a deep CNN model that has shown impressive performance on image recognition tasks, including CIFAR. CIFAR models were trained for 1000 epochs using SGD with a learning rate of 0.01, momentum of 0.9, and weight decay of 0.0005. MNIST and CIFAR models were implemented in Pytorch (Paszke et al., 2017).

A.3 Qualitative Analysis of Learned Parameters

We were also able to show that the learned difficulty parameters are interpretable for image tasks such as MNIST and CIFAR. Figure 1 shows the easiest and hardest items in the test data sets. For a certain class, there are items that one may consider more difficult than others, due to noise or irregular lines (in the case of MNIST), and this is reflected in the learned difficulty parameters. As we can see, there is interpretability in the learned difficulty parameter b_i. The difference between the easiest and hardest items in the MNIST test set for each digit is clear. The easiest items are very much prototypical examples of their specific digit, while the hardest items for each digit are outliers and in some cases (e.g. 3 and 8) are hard to distinguish from certain other digits. For CIFAR the differences are present but more subtle because the variation in the images is greater. For the hardest examples it seems that the difficulty arises mostly from the subject of the image being non-typical for the class, either according to color or orientation. For example, the hardest "car" is a car in a rotary lift, which is not common for cars, and the hardest "ship" is sitting on land instead of water. The hardest "frog" is blue, and the hardest "dog" is wearing an orange sweater. These are not consistent with the typical cases for each class, which may be the reason that the DNN models do not perform well with regards to labeling them.

*Current affiliation: Mendoza College of Business, University of Notre Dame

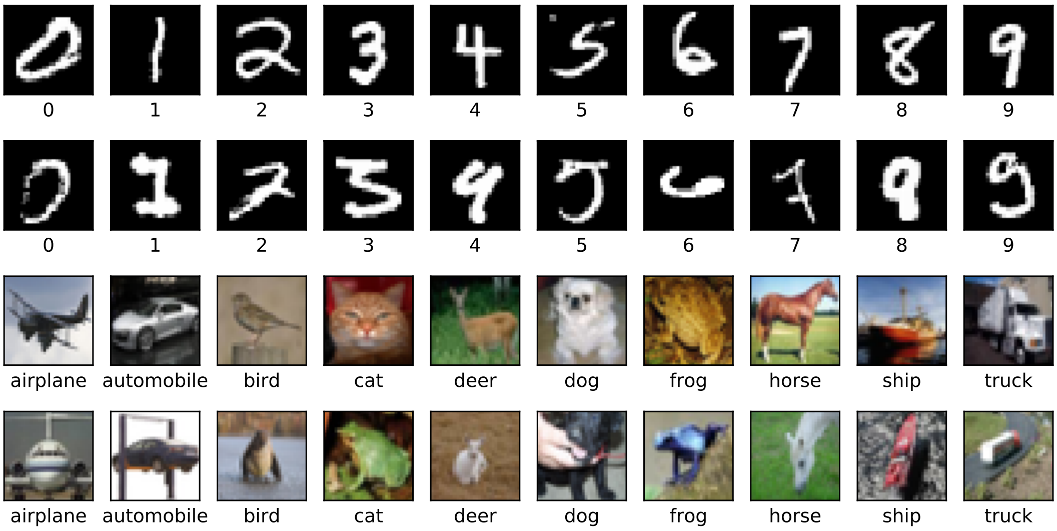

Figure 1: The easiest (first and third rows) and hardest (second and fourth rows) items in the MNIST and CIFAR test sets.

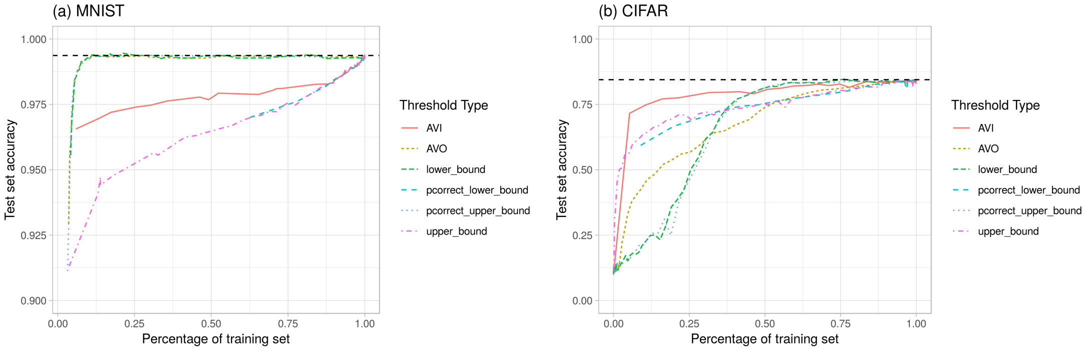

Figure 2: Test set accuracy for each filtering strategy plotted as a function of the percentage of training data retained.

A.4 Data set filtering

Figure 2 shows results of the training data filtering experiments for MNIST and CIFAR, respectively. Note that for MNIST, test set accuracy was above 90% even for very small percentages of the training set, and therefore the MNIST plot y axis is truncated to show variations more clearly (Fig. 2). For both data sets, removing up to 50% of the training data according to learned difficulty maintains test set accuracy within a few percentage points of the baseline. For MNIST, baseline accuracy is maintained with as little as 15% of the training data.

For MNIST, AVO and LB are more effective filtering strategies than AVI and UB, while AVI is the most effective for CIFAR. For MNIST, relative variance within the class is small. That is, even the hardest "3" still looks like a single numerical digit. Therefore it is unnecessary to include a large number of items of average difficulty in order to learn a particular class, making AVO an effective strategy. Similarly, the easiest items in a class can be ignored in favor of more challenging ones (LB).

For CIFAR, on the other hand, there is much more variance within each class. In these cases the easiest and hardest examples may truly be outliers in terms of the class. Therefore the DNN models would require more items from the middle of the difficulty distribution to learn a representation of the class (AVI). That said, LB is the least effective strategy in both cases, indicating that it is not enough to only include the most difficult examples.

As with the SNLI experiments, the filtering strategy we used did not take class labels into consideration. More advanced sampling strategies that maintain training set distribution or sample data using a Bayesian approach are left for future work.

A.5 Additional Examples of Difficulty Parameters: SNLI

Table 1 shows examples of the easiest and hardest sentence pairs for each class from the SNLI training data set. Table 2 shows examples of the easiest and hardest sentence pairs for each class from the SNLI testing data set.

Premise	Hypothesis	Label	Difficulty
A spherical shaped sea aquarium building.	A sphere shaped building.	entailment	-1.3775
Workers in front of thatched homes breaking up dirt with hoes.	People are working on the landscape.	entailment	-1.360
Two postal workers handle canned goods with a smile.	Two people are touching cans.	entailment	-1.356
A lady wearing a white hat sitting on bench sleeping	The hat is white	entailment	2.622
A bunch of people walking down sidewalk.	A group of friends are walking to a restaurant together	entailment	2.643
The boy is swinging outside.	The boy is having fun.	entailment	2.670
A man wearing ear coverings is cutting wood with a power saw.	A man is burning a bawnfire.	contradiction	-1.528
A man dressed in business attire stands next to an orange and green taxi cab.	There are no taxis on the street.	contradiction	-1.414
Three little kids on tricycles race downhill.	Four grown men race go-karts down an empty highway.	contradiction	-1.397
A white dog holds a stick in its mouth while it runs through snow.	The dog is outside.	contradiction	2.694
A shirtless man in a white cap relaxes in a deck chair, close to three parked bicycles.	A man is wearing clothes.	contradiction	2.729
A black dog runs through the snow.	A brown dog is outside	contradiction	2.763
Two men competing at a jujitsu tournament in a gymnasium.	Two men are sparring in a jiujitsu tournament being judged by the Olympic committee.	neutral	-1.421
Five ladies shopping in a shopping district.	5 moms are shopping for their cruise	neutral	-1.416
2 guys practice fighting ones foot is in the other face with what appears to be a referee butting in.	The men want to learn how to become professional boxers.	neutral	-1.406
A brown dog is on the green grass.	A dog sits in the grass.	neutral	2.817
A white dog standing on leaves on the ground.	A white dog is outside.	neutral	2.857
A man is riding in a boat on the water.	A male is outdoors.	neutral	2.890

Table 1: The easiest and hardest items for each class in the SNLI training data set.

Table 2: The easiest and hardest items for each class in the SNLI test data set.

Premise	Hypothesis	Label	Difficulty
Two men and a woman are inspecting the front tire of a bicycle.	There are a group of people near a bike.	entailment	-3.675
A street vendor selling cupcakes.	There is a person outside in this picture	entailment	-3.506
A young boy in a red shirt plays on a mini-trampoline in a grassy field	Someone is outside.	entailment	-3.483
This is nice place to relax and chat.	the place is nice	entailment	2.235
Neck and neck to the finish line, every competitor has been training for this race.	The competitors have trained very hard and are all very close to the finish line.	entailment	2.759
A girl in a newspaper hat with a bow is unwrapping an item.	The girl is going to find out what is under the wrapping paper.	entailment	3.144
Two dogs playing in snow.	a cat sleeps on floor	contradiction	-4.014
Girls playing soccer competitively in the grass.	Nobody is playing soccer.	contradiction	-3.558
The backside of a woman leaning against the guard rail of a passenger boat looking out at the open ocean.	The woman is driving her car on the highway.	contradiction	-3.407
A rider mid-jump on a snowmobile during a race.	A snowboarder in mid-air during a race.	contradiction	3.639
A man and woman walking away from a crowded street fair.	There are a group of men walking together.	contradiction	3.658
Man sweeping trash outside a large statue.	A man is on vacation.	contradiction	3.766
People sitting in chairs with a row flags hanging over them.	a family reunion for fourth of july	neutral	-3.603
Two men together, one watching, one resting.	Two men are together because they are friends.	neutral	-3.446
two girls on a bridge dancing with the city skyline in the background	The girls are sisters.	neutral	-3.385
A wielder works on wielding a beam into place while other workers set beams.	The wielder is working on a building.	neutral	2.864
Two soccer players on the field running into each other.	There are two people colliding and falling.	neutral	3.422
A group of dancers are performing.	The audience is silent.	neutral	3.798

References

Alex Krizhevsky and Geoffrey Hinton. 2009. Learning multiple layers of features from tiny images. Technical report, Citeseer.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Yann LeCun, Yoshua Bengio, et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

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

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.