Comparing Sample-Wise Learnability across Deep Neural Network Models

Seung-Geon Lee,[1*] **Jaedeok Kim,**[2†] **Hyun-Joo Jung,**[2] **Yoonsuck Choe**[2,3]
[1]Department of Computer Science and Engineering, Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul, Korea, 08826
[2]Machine Learning Lab, Artificial Intelligence Center, Samsung Research, Samsung Electronics Co.
56 Seongchon-gil, Secho-gu, Seoul, Korea, 06765
[3]Department of Computer Science and Engineering, Texas A&M University
College Station, TX, 77843, USA

Abstract

Estimating the relative importance of each sample in a training set has important practical and theoretical value, such as in importance sampling or curriculum learning. This kind of focus on individual samples invokes the concept of sample-wise learnability: How easy is it to correctly learn each sample (cf. PAC learnability)? In this paper, we approach the sample-wise learnability problem within a deep learning context. We propose a measure of the learnability of a sample with a given deep neural network (DNN) model. The basic idea is to train the given model on the training set, and for each sample, aggregate the hits and misses over the entire training epochs. Our experiments show that the sample-wise learnability measure collected this way is highly linearly correlated across different DNN models (ResNet-20, VGG-16, and MobileNet), suggesting that such a measure can provide deep general insights on the data's properties. We expect our method to help develop better curricula for training, and help us better understand the data itself.

Introduction

The performance of DNN models depends heavily on the quantity and quality of data. Furthermore, the order in which the data points are sampled during training makes a big difference in the learning outcome, as shown in latest studies in curriculum learning and self-paced learning (Bengio et al. 2009; Jiang et al. 2015). In this paper, we propose the concept of sample-wise learnability: How easy is it to learn each individual sample, in general, when multiple learning models are considered. Learnability is a well known concept in computational learning theory. However, in the PAC-learning framework for example, learnability is usually defined over a whole concept class, not over individual samples. We show that sample-wise learnability for a fixed data set, measured using different DNN models, are strongly linearly correlated. This way, our approach helps us gain deeper insights into the data itself, and we expect our measure to help automatically generate better curricula for improved performance in DNN training.

Sample-wise Learnability

Let $\mathcal{X}$ be a domain of inputs and $\mathcal{Y} := \{1, \cdots, L\}$ be the set of all possible labels. A DNN model is a prediction function $f\colon \mathcal{X} \to [0,1]^L$ over $\mathcal{X}$, $f(x) := (f_1(x), \cdots, f_L(x))$, such that $\sum_{l=1}^{L} f_l(x) = 1$ for $x \in \mathcal{X}$. During training, the weights of the DNN model f is updated by an optimizer. So we denote by $f^{(t)}$ the DNN model after training step t.

We take a sample (X_c, Y_c), a pair of input and label, from $\mathcal{X} \times \mathcal{Y}$ as our reference. Then $f^{(t)}(X_c)$ is the prediction of X_c by the DNN model after t training steps and $\{f^{(t)}(X_c),\ t \geq 0\}$ can be considered a stochastic process of predictions (by the DNN Model) of the tagged sample (X_c, Y_c) during training. If the tagged sample X_c is easily learnable, in most training steps a model $f^{(t)}$ should correctly predict the true label Y_c of the tagged input X_c.

Based on such an intuition, we define the learnability of an individual sample X_c with respect to a model f as

$$L_f(X_c, Y_c) := \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} f_{Y_c}^{(t)}(X_c)\right] \tag{1}$$

where T denotes the total number of training steps. Although $f_{Y_c}^{(t)}(X_c)$ is the probability that the model predicts the label of X_c as Y_c, it is still a random variable since the model $f^{(t)}$ is evolved randomly due to the randomness in the initialization and optimization. Eq. 1 is the expected value over such a quantity $\{f^{(t)},\ t \geq 0\}$.

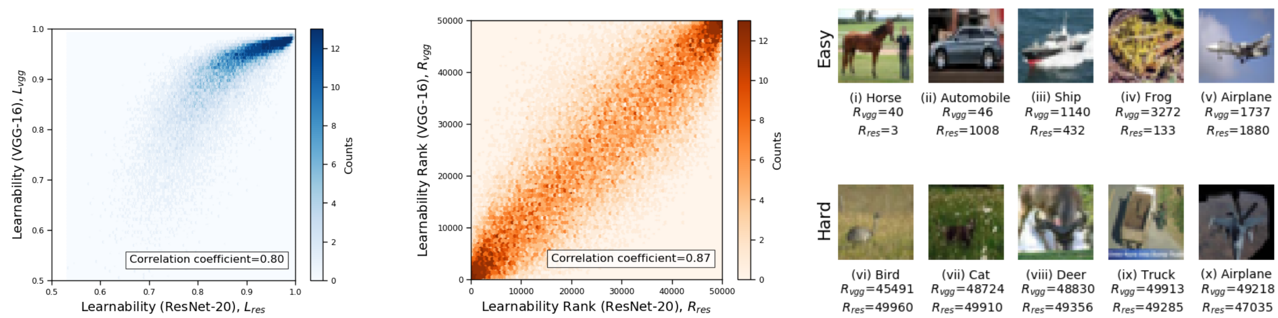
So, if X_c is easily learnable, the value of $f_{Y_c}^{(t)}(X_c)$ increases rapidly to 1 as the training step t increases. Accordingly, the value of $L_f(X_c, Y_c)$ also increases. Otherwise, the value of the probability $f_{Y_c}^{(t)}(X_c)$ remains small and so does the value of $L_f(X_c, Y_c)$. We therefore can say that Eq. 1 faithfully represents the learnability of the sample (X_c, Y_c).

Training of a DNN model is considerably affected by the order in which the samples are drawn and presented to the model, e.g. as shown in curriculum learning (Jiang et al. 2015). So it is also worthy to consider the relative order among training samples in terms of the learnability.

Denoted by $\mathcal{D} := \{(X_1, Y_1), \cdots, (X_N, Y_N)\}$ a training dataset of size N over $\mathcal{X} \times \mathcal{Y}$. Let $R_{f,i}$ be the learnability

[*]This work was performed when Seung-Geon Lee, an undergraduate student, worked as a summer intern at Samsung Research (2018).

[†]Corresponding author (05jaedeok@gmail.com)



(a) Learnability (b) Learnability Rank (c) Example Images

Figure 1: (a) and (b): Distribution (histogram) of samples from the CIFAR-10 data set. The x- and y-axes correspond to ResNet-20 and VGG-16-based learnability/learnability rank, respectively. (c): Example images from the CIFAR-10 dataset with their ground truth label and learnability ranks (top row: easy, bottom row: hard). R_{vgg} and R_{res} represent the learnability rank induced by VGG-16 and ResNet-20 respectively. Learnability rank 1 means the easiest and 50,000 means the hardest to learn.

rank of the ith training sample (X_i, Y_i) in $\mathcal{D}$ with respect to the model f. Formally, we can write

$$R_{f,i} = \sum_{j=1}^{N} \mathbf{1}_{[L_f(X_i,Y_i) \leq L_f(X_j,Y_j)]}. \quad (2)$$

Then $L_f(X_i, Y_i) > L_f(X_j, Y_j)$ if $R_{f,i} < R_{f,j}$, which implies learning the ith sample is easier than learning the jth sample in terms of the learnability.

Experimental Results

We applied the proposed learnability measure to the CIFAR-10 data set, using ResNet-20 (He et al. 2016), VGG-16 (Simonyan and Zisserman 2014), and MobileNet (Howard et al. 2017). To compare the learnability of each sample with respect to different models, we used the same training options for all models. In our experiment we considered a single training epoch as a training step and used $T = 200$.

We plot the learnability of samples with respect to the VGG-16 and ResNet-20 in Figure 1a. As we can see in the figure, the learnability of both models are positively correlated, and the correlation coefficient is 0.80. Figure 1b shows the relation of learnability rank induced by VGG-16 and that induced by ResNet-20. Similar with the case of learnability, the learnability rank of samples are also positively correlated (correlation coefficient = 0.87).

Figure 1c shows actual examples from the CIFAR-10 training set (the set includes a total of 50,000 images). The images in the top row have high rank (small learnability rank value) which means that they are easy to learn. As we can see in the figure, the images in the top row have well defined features and we can easily classify them. In contrast, the images in the bottom row have low rank (large learnability rank value) and hard to classify even for humans. For example, scale is too small (Figure 1c (vi) and (vii)) or viewpoint is atypical (Figure 1c (ix) and (x)).

The full comparison across all tested models in summarized in Table 1. The correlation coefficients in all cases are higher than 0.71. The results suggest that our proposed learnability and rank are consistent across models.

Table 1: Correlation across models. The correlation coefficients of learnability and that of learnability rank (parenthesized) are shown. Note: correlation matrices are symmetric, so redundant information was omitted.

	VGG-16	ResNet-20	MobileNet
VGG-16	-	0.796 (0.867)	0.713 (0.792)
ResNet-20	-	-	0.774 (0.782)
MobileNet	-	-	-

From the above results, we can argue that the proposed sample-wise learnability is an effective measure to estimate the importance of individual samples in a given training set.

Conclusion

In this paper, we introduced the concept of sample-wise learnability (and it's rank-based variant) based on the prediction performance during training. We experimentally showed that the sample-wise learnability (and its rank) for a given data set is linearly correlated across different models. We expect our measure to help develop better curricula for training, and help us better understand the data itself.

References

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *ICML*, 41–48. ACM.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Jiang, L.; Meng, D.; Zhao, Q.; Shan, S.; and Hauptmann, A. G. 2015. Self-paced curriculum learning. In *AAAI*, volume 2, 6.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.