On Understanding the Hardness of Samples in Neural Networks

Anonymous Authors[1]

Abstract

Deep neural network (DNN) has been an important tool for nonlinear function approximation. While the effectiveness is widely validated in practice, current DNN models somehow remain black-box when making predictions. In this paper, we aim to provide a novel perspective for understanding deep neural networks. Specifically, we hypothesize that the hardness of a sample largely depends on the angle between its embeddings and the corresponding ground truth classifier. To validate our hypothesis, we compare the hardness evaluated by human and our proposed characterization, which reveals their intrinsic correlations. Inspired by these empirical observations, we propose to employ novel forms of hashing for hard sample mining at scale. In addition to improving generalization, we show our proposal has a substantial impact on a broad range of applications such as detecting the intrinsic long-tail intra-class variations, correcting "bias to ensure fairness in learning, and principled knowledge distillation.

1. Introduction

In recent years, neural networks have achieved unprecedented successes in a wide variety of tasks. Despite such successes, how deep neural networks understand the semantics of an input and make prediction accordingly remains not fully explored. There has been an active research line (Zhang et al., 2018; Zeiler & Fergus, 2014; Mahendran & Vedaldi, 2015; Simonyan et al., 2013; Fong & Vedaldi, 2017; Koh & Liang, 2017; Bau et al., 2017) that focuses on the visualization and diagnosis of deeply learned representations. Different from these works, we cast a novel view on understanding neural networks by identifying the sample hardness.

Inspired by the observation in Figure 1 (also in (Liu et al., 2018b)) that the deeply learned features are naturally decoupled to the norm and the angle, we study how the hardness of samples affects the norm and angle of the embeddings separately. Specifically, we empirically evaluate the hardness of samples using human annotation accuracy and draw the connections between the intermediate network outputs and human-defined hardness.

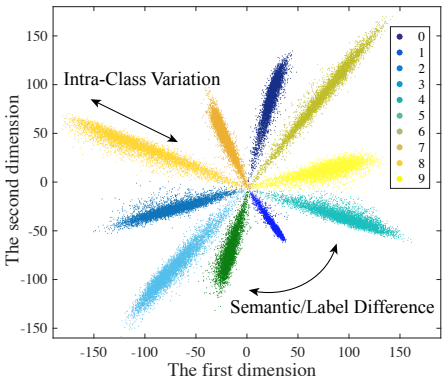

Figure 1. 2D CNN Features on MNIST.

(Liu et al., 2017b; 2018a) show that the angular information of the embeddings is crucial to visual semantics. It motivates us to hypothesize that the hardness of a sample is mostly determined by the angle between its feature embedding and the corresponding ground truth classifier. To validate our argument, we conduct experiments on the re-annotated ImageNet testing set from (Recht et al., 2019). More specifically, we use the human annotation accuracy as a reflection of the ground truth hardness of samples. We find that the angular characterization of sample hardness is highly correlated with human annotation accuracy.

Connection to deep metric learning. Measuring the hardness of samples is also of great importance in the field of deep metric learning (Oh Song et al., 2016; Sohn, 2016; Wu et al., 2017), because most objective functions in deep metric learning, , triplet loss (Schroff et al., 2015) and contrastive loss (Hadsell et al., 2006), requires data pair/triplet mining in order to perform well in practice. One of the most widely used data sampling strategies is semi-hard negative sample mining (Schroff et al., 2015) and hard negative sample mining. These negative sample mining techniques highly depend on how one defines the hardness of samples, which motivates us to study how the hardness of samples is

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

represented in deep neural networks.

Connections to fairness in machine learning. Easy and hard samples can implicitly reflect the inductive bias of the dataset, and the hardness of samples is partially affected by the data imbalance problem. For example, the CASIA-WebFace dataset (Yi et al., 2014) mostly contains western celebrities, so the neural network trained on CASIA-WebFace is highly biased towards westerners. (Buolamwini & Gebru, 2018) also demonstrates the performance drop of faces of darker people due to the biases from the training dataset. In order to ensure fairness and remove dataset biases, the ability to identify hard samples and improve the model accuracy on these hard samples is of great significance.

Connections to knowledge transfer. The efficiency of knowledge transfer (Hinton et al., 2015) is also partially determined by the sequence of input training data. (Liu et al., 2017a) theoretically shows that feeding easy samples first and hard samples later can improve the convergence of the model. (Bengio et al., 2009) also show that the curriculum of feeding training samples matters in terms of both accuracy and convergence.

2. Identifying the Hardness of Samples

There are various ways to define hardness of samples (Wu et al., 2017), but we focus on how the norm and angle of the embedding reflect the sample hardness. To be more specific, we first define the norm and angle of the embedding. The embedding of the i-th sample x_i (belonging to class j) is extracted from the input of the classification layer (, classifiers). The classifier for the i-th class is denoted as $_i$. The norm of the embedding is defined as $\|x_i\|$, while the angle of x_i is defined as the angle between x_i and the classifier of the ground truth class W_j. According to Figure 1, the angle stands for the semantic hardness of a sample. The norm corresponds to the intra-class variation, and specifically when the angle is fixed, smaller norm indicates harder sample. We will validate these characterizations with some preliminary empirical observations.

3. Empirical Observations

Our goal is to answer the following questions empirically: 1) How does human accuracy correlate with model accuracy? 2) How does human accuracy correlate with ℓ_2 norm of the embedding? 3) How does human accuracy correlate with the angle between the feature embedding and the class embedding? 3) What is the evolution of the norm and the angle of the embedding throughout the training process?

Datasets and Settings. We employ two standard image datasets MNIST (LeCun et al., 1998) and ImageNet(Deng et al., 2009) in the following experiments. Since a recent paper (Recht et al., 2019) demonstrates the findings when building new ImageNet validation set, we also take advantage of the new dataset. The additional information about human labeling frequency which is a surrogate for the human confidence level is recorded for each image. For experiments on ImageNet, we use ResNet50 (He et al., 2016) as the pre-trained model.

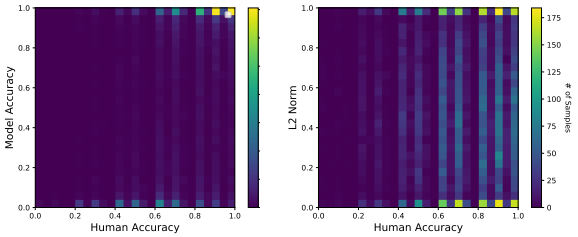

Figure 2. The left plot presents Model accuracy v.s. human accuracy. The right one presents ℓ_2 norm of the embedding v.s. human accuracy. Different color indicates different number of samples.

Baseline Results. We first benchmark the correlation between human accuracy and model accuracy as our observational baseline. Specifically, after being passed through the pre-trained model, each of the 10, 000 validation images obtains a softmax score for their labeled class. We define the softmax score as the model accuracy. In addition, each such image is provided with the frequency of being identified as the labeled class out of 50 workers who manually perform the labeling task. We also define the ratio of the frequency and the total number of 50 as human accuracy. The left plot in figure 5 presents a two-dimensional histogram for the correlation visualization. The x-axis represents human accuracy, and the y-axis represents model accuracy. Each bin exhibits the number of images which lie in the corresponding human and model accuracy range. We can observe the high density at the right corner, which means the majority of the images have both high human and model accuracy. However, there is a considerable amount of density on the range of medium human accuracy but either extremely low or high model accuracy. Overall, human and model accuracy is not in direct proportion.

3.1. Embedding ℓ_2 Norm

Both Figure 1 and (Liu et al., 2018b) demonstrate the observation that the embedding ℓ_2 norm accounts for intra-class variation. If the norm is larger, the prediction from the model is also more confident, to some extent. In this section, we try to understand the relationship between human accuracy and embedding ℓ_2 norm (a measurement of model confidence). Initially, we compute the ℓ_2 norm of the embedding layer for every validation sample. Then we normalize the value of the norm within each class. The right plot in figure 5

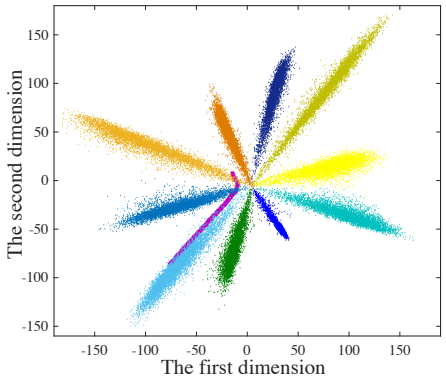

Figure 3. Trajectory of an adversarial example switching from one class to another. The purple line denotes the trajectory of the adversarial example.

uses a two-dimensional histogram to show the relationship between human accuracy and embedding ℓ_2 norm or model confidence for all the validation images. Given that the norm has been normalized with each class, naturally, there is notable density when the norm is 0 or 1. Except for that, there is no obvious correlation between human accuracy and norm.

We further verify if presenting all samples across 1000 different classes affects the visualization of the relationship between human accuracy and norm. According to WordNet (Fellbaum, 2005) hierarchy, we map the original 1000 fine-grained classes to 45 higher hierarchical classes. Figure 4 exhibits the relationship between human accuracy and norm for four representative higher classes containing $90, 58, 7, 1$ fine-grained classes respectively. Noted that from all four plots, there is still not any visible direct proportion between these two properties.

3.2. ℓ_1 Normalized Logits and Cosine Similarity

In this section, we explore the correlation between human accuracy and ℓ_1 normalized logits. We extract the logits of each validation sample before the softmax layer and process them with ℓ_1 normalization, formally,

$$\|W_y^T x\|_1 = \frac{\|W_y\|_2 \|x\|_2 cos\theta_y}{\sum_{i=1}^{C} \|W_i\|_2 \|x\|_2 cos\theta_i} = \frac{\|W_y\|_2 cos\theta_y}{\sum_{i=1}^{C} \|W_i\|_2 cos\theta_i}, \tag{1}$$

where W_y represents the class embedding for the labeled class. Noted that we linear scale the range of the normalized logits to $[0, 1]$. The plot on the left in Figure 5 shows positive correlation between human accuracy correlates with ℓ_1 normalized logits for validation images. According to equation 1, embedding ℓ_2 norm has been canceled in the final form and thereby does not have any influence to the correlation. Moreover, in order to explore if ℓ_2 norm of the class embedding has any effect to the correlation, we normalize it, formally,

$$\frac{\|W_y\|_2 cos\theta_y}{\sum_{i=1}^{C} \|W_i\|_2 cos\theta_i} \rightarrow \frac{cos\theta_y}{\sum_{i=1}^{C} cos\theta_i}. \tag{2}$$

This represents purely the normalized cosine similarity between class embedding and feature embedding. The right plot in Figure 5 shows almost exactly the same correlation. Our direct intuition from those observations is that ℓ_1 normalized logits (or the cosine similarity) corresponds to human semantic categorization of an image. Embedding ℓ_2 Norm is on the other hand irrelevant. Recall the definition of hard example described in Section 2, we can derive the above intuition to: *Hard examples for human correspond to the ones with lower cosine similarity between example and labeled class embedding in the model*.

3.3. Dynamic Experiments over ℓ_2 Norm and Cosine Similarity

In previous sections, we demonstrate some insights into the correlation between human and model hard examples. In this section, we aim to observe how the norm and the angle change throughout the training process on MNIST and ImageNet dataset. Figure 6 presents how the average norm of the feature embedding and the cosine of the angles ($cos\theta$) between feature and class embedding for testing samples vary in 90 epochs of training on ResNet50 (He et al., 2016). Note that both examples are testing data and thereby have never fit into the model. The average norm increases with a small initial slope but it suddenly climbs after 30 epochs when the first learning rate decay happens. On the other hand, the average $cos\theta$ increases sharply at the beginning and then suddenly peaks also at 30 epoch. Then it gradually drops until the second-time learning rate decay on epoch 60. After this second peak, the curve remains stable.

Moreover, we explore the difference between norm and angle change for easy and hard human examples in more details. With the human accuracy information on ImageNet testing dataset provided by (Recht et al., 2019), we split the samples into five bins corresponding to a different level of human hardness. Figure 7 demonstrates the change of the norm and $cos\theta$ similar to Figure 6, but average over testing samples in five bins separately. We can observe that for the norm, the gaps between the samples with different level of hardness are not obvious. Besides, the gaps are closed when the training converges. However, for the angle, the gaps are significant and consistent across the training process.

With all critical observations above, we propose a hypothesis that human harder testing examples correspond to relatively larger angles throughout the training process. However, the model would adjust the norm to be much higher just for optimization purpose, eg. reducing the loss.

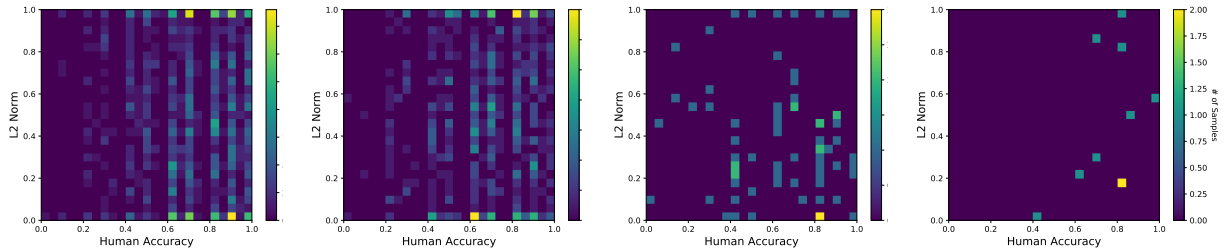

Figure 4. ℓ_2 norm of the embedding v.s. human accuracy under different class granularity (according to WordNet hierarchy). From left to right, there are 90, 50, 7, 1 classes respectively. The human accuracy is therefore computed based on the new class granularity.

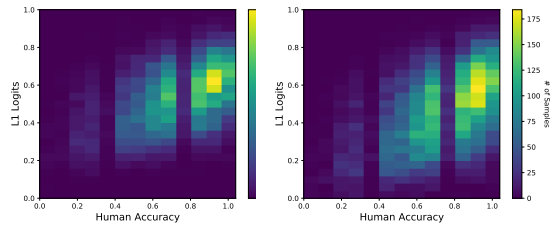

Figure 5. ℓ_1-normalized logits v.s. human accuracy. Different color indicates different number of samples

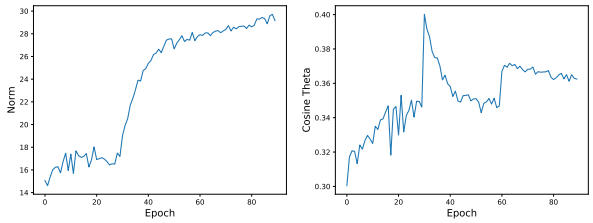

Figure 6. Average ℓ_2 norm and angle of the embedding across all testing samples v.s. iteration number.

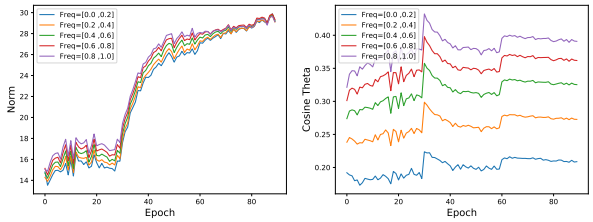

Figure 7. ℓ_2 norm and angle of the embedding of an easy sample and a hard sample v.s. iteration number.

3.4. Special Case: Adversarial Example

We show a special case in Figure 3 to illustrate how the norm and the angle change when one sample switches from one class to another. Specifically, we change the sample from one class to another using adversarial perturbation. It is essentially performing gradient ascent to the ground truth class. In Figure 3, the purple line denotes the trajectory of an adversarial sample switching from one class to another. We can see that the sample will first shrink its norm towards origin and then push its angle away from the ground truth class. Such a trajectory indicates that the norm corresponds to the intra-class hardness, and the adversarial sample will first approach to the origin in order to become a hard sample for this class. The angle corresponds to the inter-class semantic hardness, and the sample will change the angle in order to switch its label. This special example fully justifies the importance of both norm and angle in terms of the hardness of samples.

4. Future Work: Adaptive Sampling via LSH

The key ingredient of our proposal is to use Locality-Sensitive Hashing (Indyk & Motwani, 1998) as an adaptive sampler to efficiently sample the hard or easy examples based on designed metrics. LSH is a popular, sub-linear time algorithm for approximate nearest-neighbor search. The high-level idea is to place similar items into the same bucket of a hash table with high probability. An item returned as a candidate from a (K, L)-parameterized LSH algorithm is sampled with probability $1-(1-p^K)^L$, where p is the collision probability of LSH function. The LSH family defines the precise form of p that used to build the hash tables.

Recent advances in maximum inner product search (MIPS) using asymmetric locality sensitive hashing has made it possible to sample large inner products. Assume that given a collection $\mathcal{C}$ of vectors and query vector Q, using (K, L)-parameterized LSH algorithm with MIPS hashing (Shrivastava & Li, 2014), we get a candidate set S that every element $x_i \in \mathcal{C}$ is sampled with probability $p_i \leq 1$, where p_i is a monotonically increasing function of $Q \cdot x_i$. Thus, we can pay a one-time linear cost of preprocessing $\mathcal{C}$ into hash tables, and any further adaptive sampling for query Q only requires few hash lookups. We can also compute the probability of getting x.

References

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, 2017.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *ICML*, 2009.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pp. 77–91, 2018.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Fellbaum, C. Wordnet and wordnets. 2005.

Fong, R. C. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, 2017.

Hadsell, R., Chopra, S., and LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Indyk, P. and Motwani, R. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pp. 604–613. ACM, 1998.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *ICML*, 2017.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Liu, W., Dai, B., Humayun, A., Tay, C., Yu, C., Smith, L. B., Rehg, J. M., and Song, L. Iterative machine teaching. In *ICML*, 2017a.

Liu, W., Zhang, Y.-M., Li, X., Yu, Z., Dai, B., Zhao, T., and Song, L. Deep hyperspherical learning. In *NeurIPS*, 2017b.

Liu, W., Lin, R., Liu, Z., Liu, L., Yu, Z., Dai, B., and Song, L. Learning towards minimum hyperspherical energy. In *NeurIPS*, 2018a.

Liu, W., Liu, Z., Yu, Z., Dai, B., Lin, R., Wang, Y., Rehg, J. M., and Song, L. Decoupled networks. In *CVPR*, 2018b.

Mahendran, A. and Vedaldi, A. Understanding deep image representations by inverting them. In *CVPR*, 2015.

Oh Song, H., Xiang, Y., Jegelka, S., and Savarese, S. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? *arXiv preprint arXiv:1902.10811*, 2019.

Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

Shrivastava, A. and Li, P. Asymmetric lsh (alsh) for sublinear time maximum inner product search (mips). In *Advances in Neural Information Processing Systems*, pp. 2321–2329, 2014.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. In *NeurIPS*, 2016.

Wu, C.-Y., Manmatha, R., Smola, A. J., and Krahenbuhl, P. Sampling matters in deep embedding learning. In *ICCV*, 2017.

Yi, D., Lei, Z., Liao, S., and Li, S. Z. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

Zhang, Q., Nian Wu, Y., and Zhu, S.-C. Interpretable convolutional neural networks. In *CVPR*, 2018.

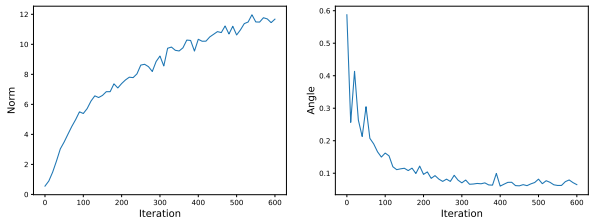

Figure 8. Average ℓ_2 norm and angle of the embedding across all testing samples v.s. iteration number.

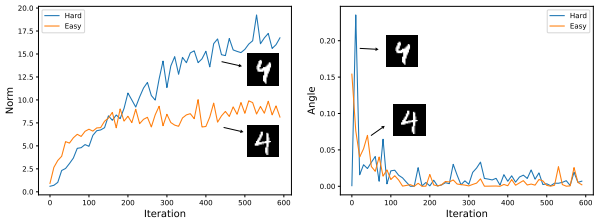

Figure 9. ℓ_2 norm and angle of the embedding of an easy sample and a hard sample v.s. iteration number.

A. Additional Experiments

Figure 8 illustrates how the average norm of the feature embedding and angles between feature and class embedding for testing samples vary in 60 iterations during the training process. The average norm increases with a large initial slope but it flattens slightly after 10 iterations. On the other hand, the average angle decreases sharply at the beginning and then becomes almost flat after 10 iterations.

Moreover, we explore the difference between norm and angle change for easy and hard human examples in more details. Figure 9 also plots the angle and norm changes for two examples, which are hard and easy for human visualization, in the training phase. Note that both examples are testing data and thereby have never fit into the model. We can see that for the angle, both of them drop largely initially and then the angle for the easy one converges to a much lower value. For the norm, both of them are increasing drastically at an early stage but that for the harder example keeps climbing even when that for the easy one saturates.