On Understanding the Hardness of Samples in Neural Networks

Beidi Chen [1] **Weiyang Liu** [2] **Animesh Garg** [3] **Zhiding Yu** [4] **Anshumali Shrivastava** [1] **Anima Anandkumar** [5]

Abstract

The mechanisms used by the human visual system and artificial convolutional neural networks (CNN) to understand images are vastly different. The two systems have different notions of hardness, meaning the set of images which appear to be ambiguous and hard to classify are different. In this paper, we answer the following question: are there measures we can compute in the trained CNN models that correspond closely to human visual hardness? We employ human selection frequency, the frequency with which human annotators label a given image, as a surrogate for human visual hardness. This information is recently made available on the ImageNet validation set (16). The CNN model confidence does not correlate well with this human visual hardness score, and it is not surprising given that there are calibration issues in the models. We propose a novel measure known as angular visual hardness (AVH). It is the normalized angular distance between the image feature embedding and the weights of the target category. We demonstrate that AVH is strongly correlated with human visual hardness across a broad range of CNN architectures. We conduct an in-depth scientific study and test multiple hypotheses to draw this conclusion. We observe that CNN models with the highest validation accuracy also have the best AVH scores. This agrees with the earlier finding that the state-of-art (SOTA) models are improving classification of harder examples. We also observe that during the training of CNNs, AVH reaches a plateau in early stages even as the training loss keeps improving. We conjecture the different causes for such plateau of easy and hard examples, which suggests the need to design better loss functions that can target harder examples more effectively and improve SOTA accuracy.

[*]Equal contribution [1]Rice University [2]Georgia Institute of Technology [3]Stanford University [4]NVIDIA [5]California Institute of Technology. Correspondence to: Beidi Chen <beidi.chen@rice.edu>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

1. Introduction

Convolutional Neural Networks (CNN) have achieved rapid progress on many computer vision tasks such as image classification (8), face recognition, and scene analysis. On large benchmark datasets such as ImageNet (3) they have even surpassed human-level accuracy. Despite this, these models are no match to the human visual system when it comes to other measures such as robustness and few-shot learning (1; 21; 15). This is not surprising given that they have entirely different processing mechanisms. Due to the black-box nature of CNNs and our limited understanding of the human brain, it is challenging to map out these differences precisely. In this paper, we focus on one key aspect, viz., how different are the hard examples for these two systems? By hard examples, we mean the the set of images that appear ambiguous and are error prone.

Much of current deep learning research focuses on measuring the hardness of an image sample for deep models rather than for a human. For example, hardness for models can be defined using the loss value (18), relative Euclidean distance (17; 20) and gradient norm (11). On the other hand, there is a rich history in cognitive and neuroscience communities to understand human visual perception (6; 2). Many works focus on mechanisms used by the human brain to translate visual information into mental representations. These representations are subject to many correspondence differences and errors, and thereby are not isomorphic to the real world (13). They can be affected by the ambiguity of different semantics (10) such as occlusion, distortion, motion blur, and inherent similarity among objects. However, such detailed semantic information is typically not present in large-scale image benchmarks used to train the CNN models.

A natural approach to measure human visual hardness is the human selection frequency, i.e. the rate with which human annotators select a specific image as belonging to a certain category. Unfortunately, most publicly available benchmarks do not have this information. But thanks to the recent efforts of (16), we now have this information on the

ImageNet test set. We employ this new dataset in this paper and as a result, come up with many novel insights. In addition, we propose a novel measure that closely aligns with the human selection frequency, and hence, can be employed in other datasets where such information is not available.

Our Contributions: We conduct an in-depth exploratory study on the ImageNet testset with newly available human selection frequency information. We employ the scientific method and carefully test multiple hypotheses and present our findings.

- We observe that the CNN model confidence and human selection frequency are not strongly correlated. The CNN model tends to be overconfident and this is a well known calibration issue (7).
- We propose a new measure known as angular visual hardness (AVH) described below.
- We observe that AVH is strongly correlated with human selection frequency across a wide range of CNN models. To the best of our knowledge, this is the first model score that correlates strongly with human visual hardness. Hence, it can serve as its proxy on datasets where such information is not available.
- We observe that the state-of-art (SOTA) models have the highest AVH score. This implies that improving SOTA accuracy will entail improving accuracy of hard examples. AVH serves as a good measure to mine such hard examples in any dataset.
- We observed the evolution of AVH score during training of CNN models. It plateaus early in training even as the training (cross-entropy) loss function keeps improving. This suggests the need to design better loss functions that can improve performance on hard examples.

Angular visual hardness: We propose a new score for a given CNN model based on the normalized angular distance between the image feature embedding and the weights of the target category. The normalization takes into account the angular distances to other categories. We argue that the semantic ambiguity that affects human visual hardness is strongly correlated with this score. This is inspired by the intuition in (14) that the angle between image feature embedding and the weights of the target class accounts for the inter-class semantic differences while the ℓ_2 norm of the feature embedding accounts for intra-class variation. (14) used this insight to try to improve the generalization of the model. On the other hand, we use it to study the correspondence with human visual hardness.

2. A Discovery of the Bridge: Angular Visual Hardness

In order to quantify Human Visual Hardness and Model Predictions for convenience purposes in experiments, we use corresponding surrogates which are formally defined as the following throughout the paper.

Definition 1 (Model Confidence)**.** *We define model confidence on a single sample as the probability score of the true objective class output by the CNN models,* $\frac{e^{W_y x}}{\sum_{i=1}^{C} e^{W_i x}}$.

We employ the standard image benchmark ImageNet(4) in all following experiments. Particularly, we take advantage of the Human Selection Frequency information for validation images provided by the recent paper (16). Recall that such information serves as a proxy for Human Visual Hardness. Besides, in order to verify that the our experimental results hold consistently across models instead of a particular model, we use four popular ImageNet pre-trained models AlexNet (12), VGG19 (19), DenseNet121 (9), ResNet50 (8). We select ResNet50 as the representative model for some experiments.

2.1. Gap between Human Visual Hardness and Model Predictions

Studying the precise connection or gap between human visual hardness and model predictions is not feasible because data collection involving human labelling or annotation requires large amount of work. In addition, usually those human data is application or dataset specific, which makes the scalability of this study even worse. Therefore, all the testing and experiments we design are at best effort given the limited resources. That is exactly another motivation for us to bridge the gap between Human and models because models predictions require minimum costs compared to human efforts.

An interesting observation in (16) shows that Human Selection Frequency has strong influence on the Model Confidence. Specifically, examples with low Human Selection Frequency tends to have relatively low Model Confidence. Naturally we examine if the correlation between Model Confidence and Human Selection Frequency is strong. Specifically, all ImageNet validation images are evaluated by the pre-trained models. The corresponding output is simply the Model Confidence on each image. In addition, because 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, i.e. Human Selection Frequency.

The left plot in figure 1 presents a two-dimensional histogram for the correlation visualization. The x-axis represents Human Selection Frequency, and the y-axis represents Model Confidence. Each bin exhibits the number of images which lie in the corresponding 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

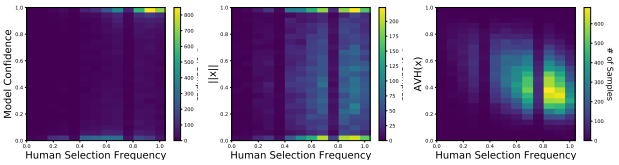

Figure 1. ℓ_2 norm and angle of the embedding of an easy sample and a hard sample v.s. iteration number.

or high model accuracy. Overall, Model Confidence and Human Selection Frequency are not in direct proportion and thereby not strongly correlated.

2.2. Bridging the Gap

Followed by identifying the gap in last section, we naturally propose a hypothesis:

Hypothesis 2. *There exists some characteristic in CNN Models strongly correlates with Human Selection Frequency to bridge the gap?*

In this section, We first provide two predictions and test them accordingly. Denote $\mathbb{S}^n$ as the unit n-sphere, formally, $\mathbb{S}^n = \{\boldsymbol{x} \in \mathbb{R}^{n+1} | \|\boldsymbol{x}\|_2 = 1\}$. Below by $\mathcal{A}(\cdot,\cdot)$, we denote the angular distance between two points on $\mathbb{S}^n$, *i.e.*, $\mathcal{A}(\boldsymbol{u},\boldsymbol{v}) = \arccos(\frac{\langle \boldsymbol{u},\boldsymbol{v}\rangle}{\|\boldsymbol{u}\|\|\boldsymbol{v}\|})$. Let x be the feature embeddings input for the layer before the last one of the classifier of the pretrained CNN models, eg. FC2 for VGG19. Let $\mathcal{C}$ be the number of classes for a classification task. Denote $\mathcal{W} = w_i | 0 < i \leq \mathcal{C}\}$ as the set of weights for all $\mathcal{C}$ classes in the final layer of the classifier.

Definition 3 (Angular Visual Hardness (AVH))**.** *AVH, for any x, is defined as,* ***AVH****(x) = $A(x,w_y)\frac{}{\sum_{i=1}^{C}\mathcal{A}(x,w_i),which}w_y$ represents the weights of the target class.*

Prediction 1: $\|x\|_2$ has a strong correlation with Human Selection Frequency

(14) conjectures that $\|x\|_2$ accounts for intra-class Human/Model Confidence. Particularly, if the norm is larger, the prediction from the model is also more confident, to some extent. Therefore, we conduct similar experiments like previous section to demonstrate the correlation between $\|x\|_2$ and Human Selection Frequency. Initially, we compute the $\|x\|_2$ for every validation sample for all models. Then we normalize $\|x\|_2$ within each class. The middle plot in figure 1 uses a two-dimensional histogram to show the correlation 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 $\|x\|_2$ and Human Selection Frequency.

We further verify if presenting all samples across 1000 different classes affects the visualization of the correlation. According to WordNet (5) hierarchy, we map the original 1000 fine-grained classes to 45 higher hierarchical classes. Figure 7 exhibits the relationship between Human Selection Frequency and $\|x\|_2$ for three representative higher classes containing $58, 7, 1$ fine-grained classes respectively. Noted that there is still not any visible direct proportion between these two variables across all plots.

Prediction 2: AVH(x) has a strong correlation with Human Selection Frequency

We test the correlation between **AVH**(x) and Human Selection Frequency. Correspondingly, after evaluating each validation sample on pre-trained models, we extract feature embeddings x and also the class weights $\mathcal{W}$ to compute **AVH**(x). Noted that we linear scale the range of **AVH**(x) to $[0, 1]$.

The plot on the right in Figure 1 shows strong correlation between **AVH**(x) and Human Selection Frequency for validation images. One intuition behind this correlation is that the class weights W might corresponds to human semantic for each category and thereby **AVH**(x) corresponds to human semantic categorization of an image. Embedding ℓ_2 Norm $\|x\|_2$ is on the other hand irrelevant.

In order to test if the strong correlation holds for all models, we perform the same experiments on AlexNet, VGG19 and DenseNet121. Figure 6 shows the strong correlation of **AVH**(x) and Human Selection Frequency consistently.

3. Dynamics of AVH during Training

After discovering the strong correlation of human visual hardness and AVH score, a natural question would be: What role does AVH play during the training process? Optimization Algorithms are used to update weights and biases i.e. the internal parameters of a model to improve the training loss. Both the angles between feature embedding and classifiers and the L_2 norm of the embedding can influence the loss. While it is well-known that the training loss or accuracy keeps improving but it is not obvious what would be the dynamics of the angles and norms separately during training. we design the experiments to observe the training dynamics of various network measurements like $\|x\|_2$ and AVH(x).

Experiment Settings For datasets and models, we use exactly the same setting as the experiments in **??**. Nevertheless, observing training dynamics involves training models from scratch on ImageNet training set instead of directly using the pre-trained models. Therefore, we follow the standard training process of AlexNet (12), VGG19 (19), DenseNet121 (9), ResNet50 (8). For consistency, we train all four models for 90 epochs and decay the initial learning

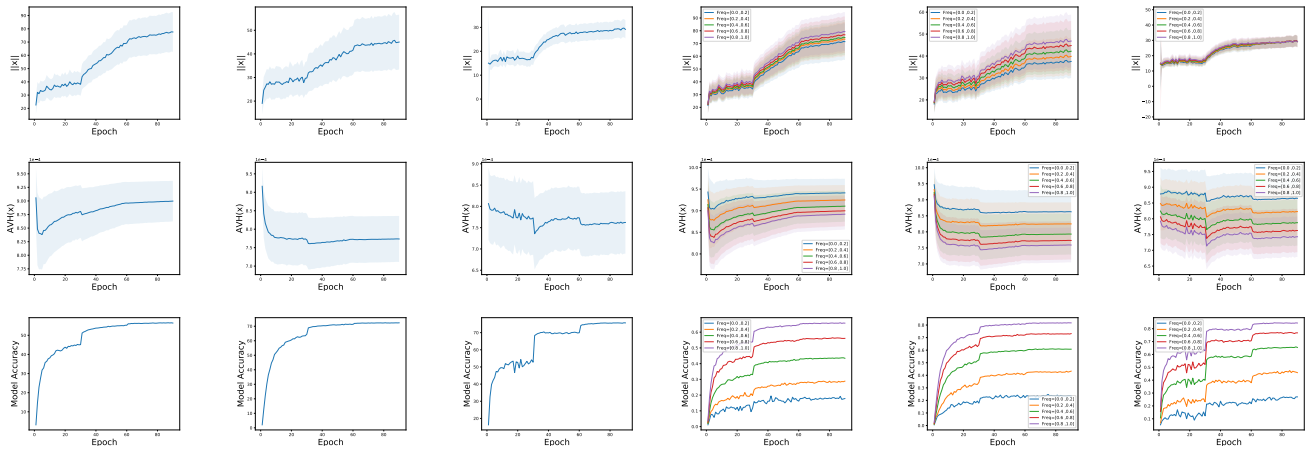

Figure 2. The top three plots show the number of Epochs v.s. Average ℓ_2 norm across all ImageNet validation samples. The middle three plots represent number of Epochs v.s. Average $AVH(x)$. The bottom ones present number of Epochs v.s. Model Accuracy. From left to right, we use AlexNet, Vgg19 and ResNet50. The plots for DenseNet are in Appendix.

Figure 3. The top three plots show the number of Epochs v.s. Average ℓ_2 norm across ImageNet validation samples which are split into five bins based on human selection frequency information. The middle three plots represent number of Epochs v.s. Average $AVH(x)$. The bottom ones present number of Epochs v.s. Model Accuracy. We use AlexNet, Vgg19 and ResNet50.

rate by a factor of 10 every 30 epochs. The initial learning rate for AlexNet and VGG19 is 0.01 and for DensetNet121 and ResNet50 is 0.1. We split all the validation images into 5 bins based on their human selection frequency, respectively $[0.0, 0.2], [0.2, 0.4], [0.4, 0.6], [0.6, 0.8], [0.8, 1.0]$ for experiments on dynamics under different human selection frequency. Note that for all the figures in this section, Epoch starts from 1.

Observation 1: Dynamics of $\|x\|_2$ and model accuracy are similarly increasing Figure 2 presents the dynamics of the average $\|x\|_2$ and the dynamics of the accuracy for testing samples vary in 90 epochs during the training on four architectures. Note that we are using the validation data for dynamics observation and thereby have never fit them into the model. The average $\|x\|_2$ increases with a small initial slope but it suddenly climbs after 30 epochs when the first learning rate decay happens. The accuracy curve is very similar to that of the average $\|x\|_2$. The above observations are consistent in all models.

Observation 2: AVH(x) hits a plateau very early even when the accuracy or loss is still improving Figure 2 exhibits the change of average AVH(x) for testing samples in 90 epochs of training on four models. The average AVH(x) for AlexNet and VGG19 decreases sharply at the beginning and then starts to bounce back a little bit before converging. However, the dynamics of the average AVH(x) for DenseNet121 and ResNet50 are different. They both decrease slightly and then quickly hits a plateau in all three learning rate decay stages. But the common observation is that they all stop improving even when $\|x\|_2$ and model accuracy are increasing. However, AVH(x) is more important than $\|x\|_2$ because during inference it is the key factor deciding which class the input sample is classified to The dynamics for $\|x\|_2$ should be monotonically decreasing so we raise the question that is the cross-entropy loss the best for CNNs?

Observation 3: AVH(x)'s correlation with human selection frequency holds across models and throughout the training process. Figure 3 demonstrates the change of $\|x\|_2$ and AVH(x) similar to Figure 2, but average over testing samples in five human selection frequency bins separately. We can observe that for $\|x\|_2$, the gaps between the samples with different human visual hardness are not obvious in ResNet50 and DenseNet121. Besides and closed near convergence. However, for AVH(x), such gaps are significant and consistent across every single model during the whole training process.

Observation 4: AVH(x) is an indicator of model's generalization ability From Figure 2 and Figure 3, we observe that better models have better AVH(x) throughout the training process and also across samples under different human selection frequency. For instance, Alexnet is the worst model and its overall AVH(x) or average AVH(x) on each of five bins are worse than those of the other three models. This observation aligned with the earlier observations of (16) that better models also generalize better on samples across different human selection frequencies.

References

[1] Charles F. Cadieu, Ha Hong, Daniel L. K. Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Na-

jib J. Majaj, and James J. DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Computational Biology*, 10(12):e1003963, Dec 2014.

[2] Lichao Chen, Sudhir Singh, Thomas Kailath, and Vwani Roychowdhury. Brain-inspired automated visual object discovery and detection. *Proceedings of the National Academy of Sciences*, 116(1):96–105, 2019.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[5] Christiane Fellbaum. Wordnet and wordnets. 2005.

[6] Daniel J Felleman and DC Essen Van. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.

[7] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org, 2017.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[10] Martina Jakesch, Helmut Leder, and Michael Forster. Image ambiguity and fluency. *PLoS One*, 8(9):e74084, 2013.

[11] Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling. *arXiv preprint arXiv:1803.00942*, 2018.

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

[13] Peter H Lindsay and Donald A Norman. *Human information processing: An introduction to psychology*. Academic press, 2013.

[14] Weiyang Liu, Zhen Liu, Zhiding Yu, Bo Dai, Rongmei Lin, Yisen Wang, James M Rehg, and Le Song. Decoupled networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2771–2779, 2018.

[15] Radoslaw Martin Cichy, Aditya Khosla, Dimitrios Pantazis, and Aude Oliva. Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, 153:346358, Jun 2017.

[16] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? *arXiv preprint arXiv:1902.10811*, 2019.

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[18] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *ICCV*, 2015.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *ICCV*, 2017.

[21] D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):86198624, May 2014.

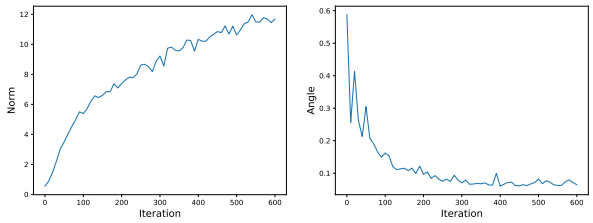

Figure 4. Average ℓ_2 norm and angle of the embedding across all testing samples v.s. iteration number.

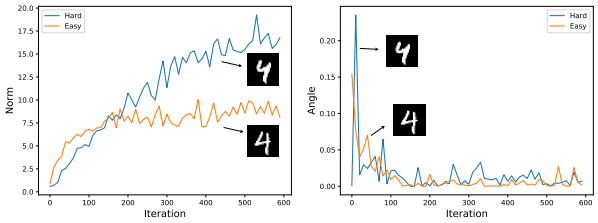

Figure 5. ℓ_2 norm and angle of the embedding of an easy sample and a hard sample v.s. iteration number.

A. Additional Experiments

Figure 4 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 5 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.

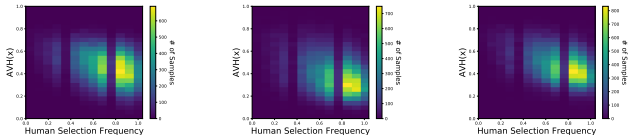

Figure 6. The three plots present the correlation between Human Selection Frequency and $\|x\|$ using AlexNet, VGG19 and DenseNet121.

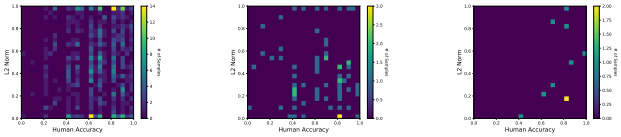

Figure 7. ℓ_2 norm of the embedding v.s. human selection frequency under different class granularity (according to WordNet hierarchy). From left to right, there are 58, 7, 1 classes respectively. The human selection frequency is therefore computed based on the new class granularity.