Angular Visual Hardness

Beidi Chen¹, Weiyang Liu², Animesh Garg³⁴, Zhiding Yu¹, Anshumali Shrivastava¹, Anima Anandkumar³⁵
¹Rice University  ²Georgia Institute of Technology    ³NVIDIA Research
⁴University of Toronto    ⁵California Institute of Technology

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 two scores as surrogates for human visual hardness. One is human selection frequency, the frequency with which human annotators label a given image. This information is recently made available on the ImageNet validation set [38]. The other measure is based on an image degradation method [14] that adds noise and changes various properties such as contrast to make the image more ambiguous to humans. 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 called 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-the-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.

1 Introduction

Convolutional Neural Networks (CNN) have achieved unprecedented performance on many computer vision tasks, such as image classification [18, 24, 41], face recognition [31, 44, 43], and scene analysis [49, 33, 15]. On large-scale benchmark datasets such as ImageNet [8], CNNs have already surpassed human-level accuracy. Despite such progress, CNNs are still no match to the human visual system when it comes to other measures such as robustness and few-shot recognition [4, 46, 35, 23, 11, 10, 7, 37, 2]. 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 refer to the set of images that appear semantically ambiguous and are error-prone.

Most 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 [40], relative Euclidean distance [39, 45] and gradient norm [22]. On the other
hand, there is a rich history in cognitive and neuroscience communities to understand human visual perception [13, 5, 34, 6], where many of them focus on mechanisms used by the human brain to translate visual information to mental representations. These representations are subject to many correspondence differences and errors and thereby are not isomorphic to the real world [27]. They can be affected by the ambiguity of different semantics [21] 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.

There have been many attempts made to design human perceptual scores that act as a surrogate for human visual hardness, i.e. whether a typical human would perceive the image to be ambiguous [27]. For instance, since human annotation data is hard to come by, researchers have proposed other measures of human visual hardness based on image degradation [27]. It involves adding noise and changing image properties such as contrast. [14] employed psychology studies to validate the degradation method as a way to measure human visual hardness. Another 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 [38], we now have this information on the ImageNet validation set. We adopt this new information of the dataset in the paper and come up with many novel insights as a result. In addition, we propose a novel measure that closely aligns with the visual hardness based on human selection frequency as well as image degradation, 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 validation set with newly available human selection frequency information. We employ the scientific method, carefully test multiple hypotheses and present our findings. Furthermore, we test our findings on ImageNet validation set with various degradations.

- We observe that the CNN model confidence and human selection frequency are not strongly correlated. The CNN model tends to be overconfident, which is a well known calibration issue [16].
- We propose a new measure known as angular visual hardness (AVH) described in Section 3.3.
- We observe that AVH is strongly correlated with human selection frequency across a wide range of CNN models. Moreover, AVH is also strongly correlated with image degradation level, which is another proxy to human visual hardness. 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 observed the evolution of AVH score during training of CNN models. It plateaus early in training even as the training (cross-entropy) 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 improve performance on hard examples. It also validates the argument in [38] that improving the hard examples is the key to improve the generalization.
- We observe that the state-of-art (SOTA) models have the best average AVH score over all the validation images. Therefore, an AVH score can serve as a good measure to mine such hard examples in any datasets to facilitate the study of better generalization.
- Finally, we discuss some potential applications where AVH can be useful.

Angular visual hardness: We propose a new score for a specific image and 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 [30] that the angle between image feature embedding and the weights of the target class accounts for the inter-class semantic differences while the $\ell_2$ norm of the feature embedding accounts for intra-class variation. [30] 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.

A Preview of our positive experimental findings:

- AVH is linearly correlated with human selection frequency while other measures such as model confidence and embedding norm have low correlation with human selection frequency. We observe similar results for image degradation level.
• Images with different human selection frequencies or degradation levels have a consistent AVH gap across different network architectures. Thus, AVH is a reliable universal measure to detect hard examples.
• We measure the dynamics of AVH evolution during the training of CNN models. We observed that images with different human selection frequencies or degradation levels have a consistent gap on average AVH across different training iterations. Thus, AVH is a robust measure to detect hard examples even when model training is underway.
• Models with the highest accuracies also have the best average AVH scores. [38] and the above observations both imply that in SOTA models, the “low-hanging fruits” i.e. the easy examples are all correctly classified and the gains in accuracy will come from improving on hard examples.
• During training, AVH score plateaus early even as the training (cross-entropy) loss is improving and $L_2$ embedding norm keeps increasing. Hence, the improvement in the training loss is mostly coming from the improvement of embedding norm, rather than AVH.

A Preview of our negative experimental findings:
• Fine-tuning on AVH score does not improve SOTA accuracy. We believe that this is because the optimization landscape for AVH score is challenging and this has been observed before [31]. It is an open problem how we can leverage AVH and design better loss functions to improve SOTA performance.
• We evaluate AVH with adversarial examples. Our experiments show that adversarial example may change AVH without being more difficult to recognize, which makes adversarial examples a counterexample for AVH. Hence, AVH will work for natural distribution of images, but not on adversarial perturbations to the distribution. However, on a positive note, we find that for the adversarial example to switch to another class, it tends to first shrink its norm and then change its angle, indicating that AVH is in general more difficult to change compared to the norm. This suggests AVH is more robust to adversarial perturbations.

2 Related Work

Angular Distance in CNNs: [48] uses the deep features to quantify the semantic difference between images, indicating that deep features contain the most crucial semantic information. It empirically shows that the angular distance between feature maps in deep neural networks is very consistent with the human in distinguishing the semantic difference. However, because of the different goal mentioned above, they have not studied or shown any strong correlation of human visual hardness and the angular distance on natural images. [32] proposes a hyperspherical neural network that constrains the parameters of neurons on a unit hypersphere and uses angular similarity to replace the inner product similarity. [30] decouples the inner product as the norm and the angle and argues that the norm corresponds to intra-class variation, and the angle corresponds to inter-class semantic difference. However, this work does not consider any human factors, while our goal is to bridge the gap between CNNs and human perception. [29] proposes a well-performing regularization based on angular diversity to improve the network generalization.

Image Degradation: Because CNNs and humans achieve similar accuracy on a wide range of (narrow) tasks on benchmark datasets, a number of works have investigated similarities and differences between CNNs and human vision [4, 46, 35, 23, 11, 10, 7, 37, 2]. Since human annotation data is hard to come by, researchers have proposed an alternative measure of visual hardness on images based on image degradation [27]. It involves adding noise or changing image properties such as contrast, blurriness, and brightness. [14] employed psychology studies to validate the degradation method as a way to measure human visual hardness. It should be noted that the artificial visual hardness introduced by degradation is a different concept from the natural visual hardness. The hardness based on degradation only reflects the hardness of a single original image with various of transformations, while natural visual hardness based on the ambiguity of human perception across a distribution of natural images. In this paper, we consider both as the surrogates of human visual hardness.

Deep model calibration. Confidence calibration is the problem of predicting probability estimates representative of the true correctness likelihood [16]. It is well-known that the deep neural networks are mis-calibrated and there has been a rich literature trying to solve this problem [25, 16]. However, this is a somewhat different issue because the confidence calibration is a problem introduced by two measurements of the model, which does not have any involvement of human visual hardness.
3 A Discovery of the Bridge: Angular Visual Hardness

3.1 Notations and Setup

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, formally, \( \frac{e^{W_y x}}{\sum_{i=1}^{C} e^{W_i x}} \).

**Definition 2 (Human Selection Frequency).** We define one way to measure human visual hardness on pictures as Human Selection Frequency. Quantitatively, given \( m \) number of human workers in a labeling process described in [38], if \( a \) out of \( m \) label a picture as a particular class and that class is the target class of that picture in the final dataset, then Human Selection Frequency = \( \frac{a}{m} \).

**Definition 3 (Image Degradation Level).** We define another way to measure human visual hardness on pictures as Image Degradation Level. We consider two degradation methods in this paper; decreasing contrast and adding noise. Quantitatively, Image Degradation Level for decreasing contrast is directly the contrast level. Image Degradation Level for adding noise is the amount of pixel-wise additive uniform noise.

We use the ImageNet [9] benchmark in all following experiments. Particularly, we take advantage of the Human Selection Frequency information for validation images provided by the recent paper [38]. Recall that such information can serve as one of the proxy for Human Visual Hardness. To test if our findings with Human Selection Frequency hold on another proxy, image degradation, we create an augmented validation set based on two image degradation methods, decreasing contrast and adding noise. We label them with corresponding degradation level. 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 [24], VGG19 [41], DenseNet121 [20], ResNet50 [18]. We select ResNet50 as the representative model for some experiments.

3.2 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 [38] 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 or high model accuracy. Overall, Model Confidence and Human Selection Frequency are not in direct proportion and thereby not strongly correlated.

3.3 Bridging the Gap

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

**Hypothesis 4.** 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 $S^n$ as the unit $n$-sphere, formally, $S^n = \{ x \in \mathbb{R}^{n+1} \parallel x \parallel_2 = 1 \}$. Below by $A(\cdot, \cdot)$, we denote the angular distance between two points on $S^n$, i.e., $A(u, v) = \arccos(\frac{\langle u, v \rangle}{\parallel u \parallel \parallel v \parallel})$. Let $x$ be the feature embeddings input for the layer before the last one of the classifier of the pretrained CNN models, e.g., FC2 for VGG19. Let $C$ be the number of classes for a classification task. Denote $\mathcal{W} = w_i \{ 0 < i \leq C \}$ as the set of weights for all $C$ classes in the final layer of the classifier.

**Definition 5** (Angular Visual Hardness (AVH)). AVH, for any $x$, is defined as,

$$AVH(x) = \frac{A(x, w_y)}{\sum_{i=1}^{C} A(x, w_i)},$$

which $w_y$ represents the weights of the target class.

**Prediction 1:** $\parallel x \parallel_2$ has a strong correlation with Human Selection Frequency

[30] conjectures that $\parallel x \parallel_2$ 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 $\parallel x \parallel_2$ and Human Selection Frequency. Initially, we compute the $\parallel x \parallel_2$ for every validation sample for all models. Then we normalize $\parallel x \parallel_2$ 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 $\parallel x \parallel_2$ and Human Selection Frequency.

We further verify if presenting all samples across 1000 different classes affects the visualization of the correlation. According to WordNet [12] hierarchy, we map the original 1000 fine-grained classes to 45 higher hierarchical classes. A figure in appendix exhibits the relationship between Human Selection Frequency and $\parallel x \parallel_2$ 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 has a strong correlation with Human Selection Frequency
We test the correlation between $\text{AVH}(x)$ and Human Selection Frequency. Correspondingly, after evaluating each validation sample on pre-trained models, we extract feature embeddings $x$ and also the class weights $W$ to compute $\text{AVH}(x)$. Noted that we linear scale the range of $\text{AVH}(x)$ to $[0, 1]$.

The plot on the right in Figure 1 shows strong correlation between $\text{AVH}(x)$ 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 $\text{AVH}(x)$ corresponds to human semantic categorization of an image. Embedding $\ell^2$ 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 2 shows the strong correlation of $\text{AVH}(x)$ and Human Selection Frequency consistently.

**Prediction 3: AVH has a strong correlation with Image Degradation Level**

In order to test if the results from Prediction 2 hold on another proxy to human visual hardness, Image Degradation Level, we perform the similar experiments but on the augmented ImageNet validation set. The plots in Figure 3 show the strong correlation between $\text{AVH}(x)$ and Noise Degradation Level while the plots in Figure 4 present the strong correlation between $\text{AVH}(x)$ and Contrast Degradation Level. They, along with Figure 2, demonstrate that $\text{AVH}(x)$ strongly correlates with Human Visual Hardness.

## 4 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 $\text{AVH}$ 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 the 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 architectures.

**Experiment Settings.** For datasets and models, we use exactly the same setting as the experiments in 3.1. 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 [24], VGG19 [41], ResNet50 [18] and DenseNet121 [20] (DenseNet results are...
put in Appendix). For consistency, we train all four models for 90 epochs and decay the initial learning rate by a factor of 10 every 30 epochs. The initial learning rate for AlexNet and VGG19 is 0.01 and for DenseNet121 and ResNet50 is 0.1. For human visual hardness based on Human Selection Frequency, we split all the validation images into 5 bins, [0.0, 0.2], [0.2, 0.4], [0.4, 0.6], [0.6, 0.8], [0.8, 1.0], based on their human selection frequency respectively. For human visual hardness based on Image Degradation Level, we create an augmented validation set based on two image degradation methods, decreasing contrast and adding noise. We label them with corresponding degradation level as well. Note that for all the figures in this section, Epoch starts from 1.

**Observation 1:** The norm of feature embeddings keeps increasing during training. Figure 6, 7 and 8 presents the dynamics of the average $\|x\|_2$ and the dynamics of the accuracy for validation samples vary in 90 epochs during the training on three architectures. Note that we are using the validation data for dynamics observation and therefore 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. More interestingly, we find that neural networks with shortcut connections (e.g., ResNets and DenseNets) tend to make the norm of the images with different human selection frequency become the same, while the neural networks without shortcuts (e.g., AlexNet and VGG) tend to keep the gap of norm among the images with different human visual hardness.

**Observation 2:** AVH hits a plateau very early even when the accuracy or loss is still improving. Figure 6, 7 and 8 exhibits the change of average AVH for validation samples in 90 epochs of training on three models. The average AVH 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 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. AVH is more important than $\|x\|_2$.
in the sense that it is the key factor deciding which class the input sample is classified to. However, optimizing the norm under the current softmax cross-entropy loss would be easier so, which cause the plateau of angles for easy examples. However, the plateau for the hard examples can be caused by the limitation of the model itself. As a result, it shows the necessity and importance of designing loss functions that focus on optimizing angles, such as [31, 29, 26].

**Observation 3:** AVH’s correlation with human selection frequency consistently holds across models throughout the training process. In Figure 6, 7 and 8, we average over validation samples in five human selection frequency bins or five degradation level bins separately, and then compute the average embedding norm, AVH and model accuracies. We can observe that for $\|x\|_2$, the gaps between the samples with different human visual hardness are not obvious in ResNet and DenseNet, while they are quite obvious in AlexNet and VGG. However, for AVH, such AVH gaps are very significant and consistent across every network architecture during the entire training process. Interestingly, even if the network is far from being converged, such AVH gaps are still consistent across different human selection frequency. Also the norm gaps are also consistent. The intuition behind this could be that the angles for hard examples are much harder to decrease and probably never in the region for correct classification. Therefore the corresponding norms would not increase otherwise hurting the loss. It validates that AVH is a consistent and robust measure for visual hardness (and even generalization).

**Observation 4:** AVH is an indicator of model’s generalization ability. From Figure 6, 14, 7 and 8, we observe that better models (*i.e.*, higher accuracy) have lower average AVH throughout the training process and also across samples under different human visual hardness. For instance, Alexnet is the worst model, and its overall average AVH and average AVH on each of five bins are worse than those of the other three models. This observation is aligned with the earlier observations of [38] that better models also generalize better on samples across different human visual hardness. Moreover, we AVH is potentially a better measure for generalization as a pretrained model. The norm of feature embeddings is often embedded with training data prior such as data imbalance [31] and class granularity [24]. But when extracting the features for the classes that do not exist in training set,
Conjecture on training dynamics of CNNs. From Figure 6 and observations above, we conjecture that the training of CNN has two phases. 1) At the beginning of the training, the softmax cross-entropy loss will first optimize the angles among different classes while the norm will fluctuate and increase very slowly. We argue that it is because changing the norm will not decrease the loss when the angles are not separated enough for correct classification. As a result, the angles get optimized firstly. 2) As the training continues, the angles become more stable and change very slowly while the norm increases rapidly. On the one hand, for easy examples, it is because when the angles get decreased enough for correct classification, the softmax cross-entropy loss can be well minimized by purely increasing the norm. On the other hand, for hard examples, the plateau is cause by unable to decrease the angle to correctly classify examples and thereby also unable to increase the norms otherwise hurting loss.

5 Extensions and Applications

Adversarial Example: A Counter Example? Our claim about the strong correlation between AVH score and human visual hardness does not apply on non-natural images such as adversarial examples. For such examples, the human can not tell the difference visually, but the adversarial example has a worse AVH than the original image, which runs counter to our claim that AVH has strong correlation with human visual hardness. So this claim is limited to distribution of natural images. However, on a positive note, we do find that AVH is slower to change compared to the embedding norm during the dynamics of adversarial training. See Appendix for details.

Connection to deep metric learning: Measuring the hardness of samples is also of great importance in the field of deep metric learning [36, 42, 45]. For instance, objective functions in deep metric learning consist of e.g., triplet loss [39] or contrastive loss [17], which 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 [39] and hard negative sample mining. These negative sample
Connections to fairness in machine learning: Easy and hard samples can implicitly reflect imbalances in latent attributes in the dataset. For example, the CASIA-WebFace dataset [47] mostly contains white celebrities, so the neural network trained on CASIA-WebFace is highly biased against the other races. [3] demonstrates a performance drop of faces of darker people due to the biases in the training dataset. In order to ensure fairness and remove dataset biases, the ability to identify hard samples automatically can be very useful. We would like to test if AVH is effective in these settings.

Connections to knowledge transfer and curriculum learning: The efficiency of knowledge transfer [19] is partially determined by the sequence of input training data. [28] theoretically shows feeding easy samples first and hard samples later (known as curriculum learning) can improve the convergence of model. [1] also show that the curriculum of feeding training samples matters in terms of both accuracy and convergence. We plan to investigate the use of AVH metric in such settings.

6 Concluding Remarks

Human perception and deep neural networks in general have different notions of visual hardness. Our paper studies the gap between them, and attempts to bridge this gap by proposing a novel measure for CNN models known as angular visual hardness. Our comprehensive empirical studies show that AVH has many nice properties. First, AVH has a strong correlation with human selection frequency and image degradation level. Second, this holds across different network architectures and throughout the training process. Third, AVH can serve as an indicator of generalization abilities of neural networks, and improving SOTA accuracy entails improving accuracy on hard examples. It is still an open problem of designing an appropriate loss function that can focus on improving AVH during training. AVH can be very useful in diverse applications such as deep metric learning, fairness, knowledge transfer, etc. and we plan to investigate them in future.
References


[27] Peter H Lindsay and Donald A Norman. Human information processing: An introduction to psychology. Academic press, 2013. 2, 3


Appendix

A Additional Experiments

A.1 Additional Plots for the Predictions for the Hypothesis

**Prediction 1: \( \|x\|_2 \) has a strong correlation with Human Selection Frequency**

We further verify if presenting all samples across 1000 different classes affects the visualization of the correlation. According to WordNet [12] hierarchy, we map the original 1000 fine-grained classes to 45 higher hierarchical classes. Figure 9 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.

![Figure 9](image)

*Figure 9: \( \ell_2 \) 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.*

**Prediction 2: AVH has a strong correlation with Human Selection Frequency** Additional Plots for DenseNet121 is shown in Figure 10.

**Prediction 3: AVH has a strong correlation with Image Degradation Level** Additional Plots for DenseNet121 is shown in Figure 10.

![Figure 10](image)

*Figure 10: The left, middle and right plots respectively present the correlation between Human Selection frequency, Noise Degradation Level, Contrast Degradation Level and \( \|x\| \) using DenseNet121.*
A.2 Additional Experiments for Observing Dynamics on MNIST

Figure 11 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 12 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 11: Average $\ell_2$ norm and angle of the embedding across all testing samples v.s. iteration number.

Figure 12: $\ell_2$ norm and angle of the embedding of an easy sample and a hard sample v.s. iteration number.
A.3 Additional Experiments for Training Dynamics on ImageNet

Figure 13 presents the dynamics of the average $\|x\|_2$ and the dynamics of the accuracy for validation samples vary in 90 epochs during the training on AlexNet, VGG19, DenseNet121 and ResNet50. In figure 14, we average over validation samples in five human selection frequency bins separately, and then compute the average embedding norm, AVH and model accuracies. In figure 14, we average over validation samples in five image noise degradation level bins separately, and then compute the average embedding norm, AVH and model accuracies. In figure 14, we average over validation samples in five image contrast degradation level bins separately, and then compute the average embedding norm, AVH and model accuracies.

![Figure 13: Average ℓ₂ norm and angle of the embedding across testing samples with different level of hardness v.s. iteration number on DenseNet121.](image)

![Figure 14: The left plot shows the number of Epochs v.s. Average ℓ₂ norm across ImageNet validation samples which are split into five bins based on human selection frequency information. The middle plot represent number of Epochs v.s. Average AVH(x). The bottom ones present number of Epochs v.s. Model Accuracy on DenseNet121.](image)

![Figure 15: The left plot shows the number of Epochs v.s. Average ℓ₂ norm across ImageNet validation samples which are split into five bins based on image noise degradation level information. The middle plot represent number of Epochs v.s. Average AVH(x). The bottom ones present number of Epochs v.s. Model Accuracy on DenseNet121.](image)
B A Special Case: Adversarial Examples

We show a special case in Figure 17 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 17, 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 adversarial sample will first approach to the origin in order to become a hard sample for this class. Then 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.