# Appendix: Angular Visual Hardness

## A. Simulation Details

**Gaussian Simulation Plot:** We generate 2000 3-d random vectors from two multivariate normal distribution (1000 for each) and normalize to unit norm, shown in red and green color on the left plot in Figure 7. Then these data points are passed as the inputs to a simple multi layer perceptron classification model with one $3 \times 2$ hidden layer. Upon convergence, we compute the AVH scores for each data point. The middle image shows the visualization of AVH scores for all data points, with lighter color representing higher AVH scores. It is obvious that AVH scores for points lying on the intersection of two clusters are higher, which agrees with the intuition that those are hard examples. We also compute the $\ell_2$ norm of the feature embeddings shown in the right plot. One can see there is less correlation with hard examples.

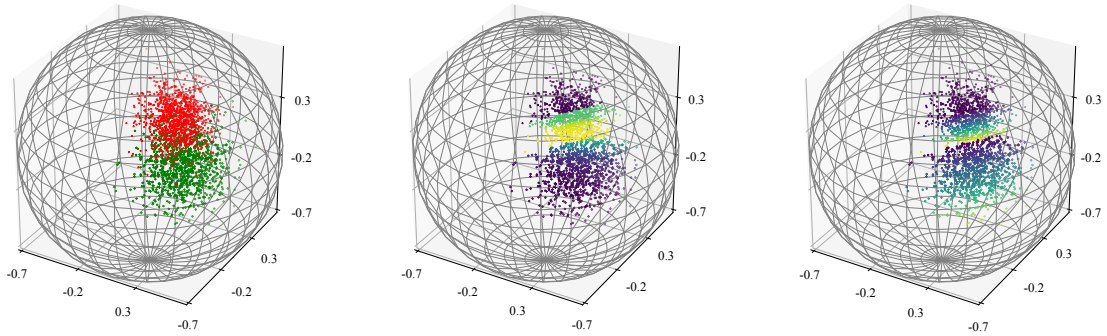


*Figure 7.* Toy example of two overlapping Gaussian distributions (classes) on a unit sphere. Left: samples from the distributions as input to a multi layer perceptron (MLP). Middle: AVH heat map produced by MLP, where samples in lighter colors (higher hardness) are mostly overlapping hard examples. Right: $\ell_2$-norm heat map, where certain non-overlapping samples also have higher values.

**MNIST Simulation Plot:** We train MNIST with a very simple CNN model which the dimension of the embedding (right before the classifier) is 2. Figure 8 shows the visualization of those 2D embeddings.

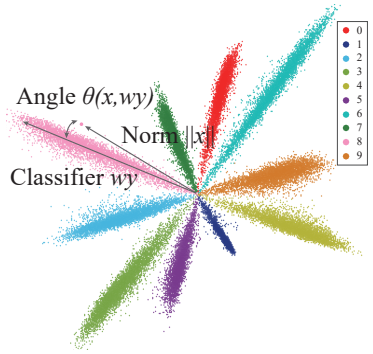


*Figure 8.* Visualization of embeddings on MNIST by setting their dimensions to 2 in a CNN.

# B. Additional Results of Training Dynamics

## B.1. Additional Results on ImageNet

**Model Confidence:** Figure 9 shows the training dynamics of the model confidence corresponding to AlexNet, VGG-19, and ResNet-50.

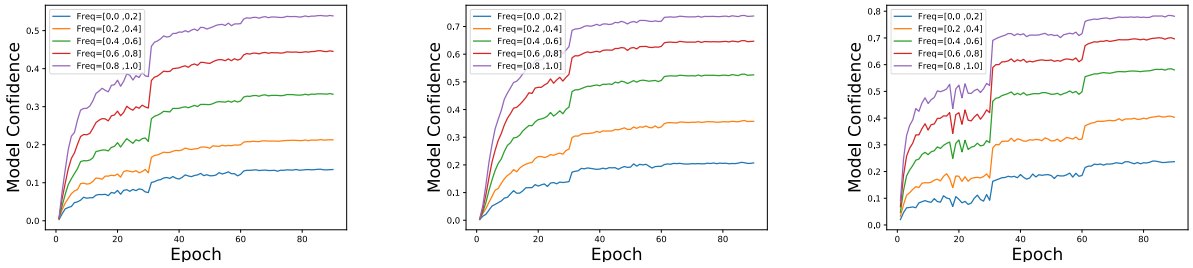


*Figure 9.* Number of epochs vs. Model Confidence. Results from left to right correspond to AlexNet, VGG-19 and ResNet-50.

**Averaged training dynamics:** In Figure 10, we plot the average embedding norm, AVH and model accuracies for AlexNet, VGG-19, ResNet-50 and DenseNet-121 over the validation samples.

**Image degradation:** Because CNNs and humans can achieve similar accuracy on large-scale benchmark dataset such as ImageNet, a number of works have investigated similarities and differences between CNNs and human vision (Martin Cichy et al., 2017; Kheradpisheh et al., 2016; Dodge & Karam, 2017; Dekel, 2017; Pramod & Arun, 2016; Berardino et al., 2017). Since human annotation data is relatively hard to obtain, researchers have proposed an alternative measure of visual hardness on images based on image degradation (Lindsay & Norman, 2013). This involves adding noise or changing image properties such as contrast, blurriness, and brightness. (Geirhos et al., 2018) employed psychological 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 the following additional experiments, we also consider different level of degradation as the surrogate of human visual hardness besides Human Selection Frequency.

**Definition 4** (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.*

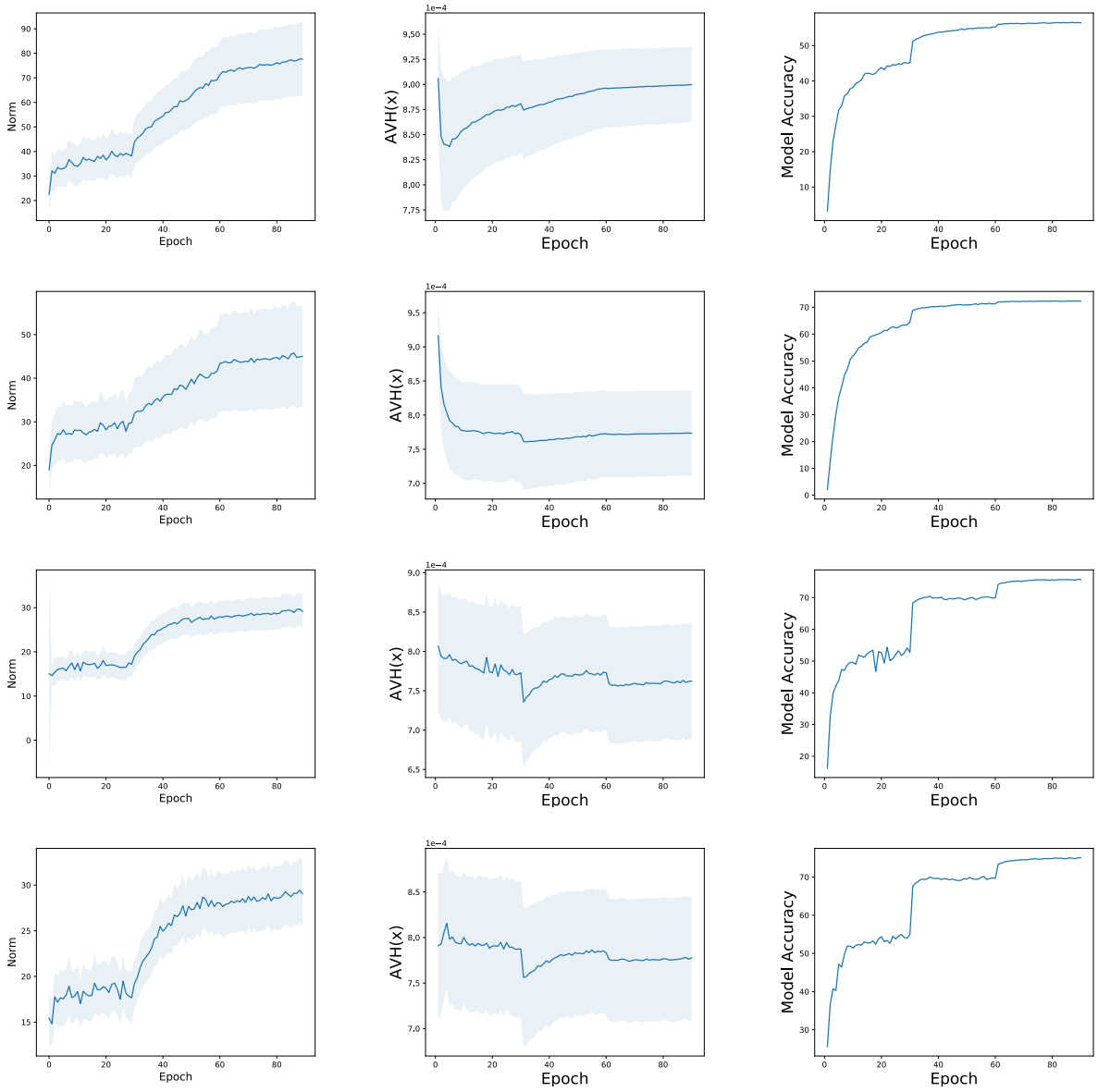


*Figure 10.* Averaged training dynamics on ImageNet validation set. Columns from left to right: number of epochs vs. average $\ell_2$ norm, number of epochs vs. average AVH score, and number of epochs vs. model accuracy. Rows from top to bottom: dynamics corresponding to AlexNet, VGG-19, ResNet-50, and DenseNet-121.

**Dynamics across noise degradation levels:** In Figure 11, we illustrate the averaged training dynamics on the ImageNet validation set across five image noise degradation levels - [0.4, 0.3, 0.2, 0.1, 0.0].

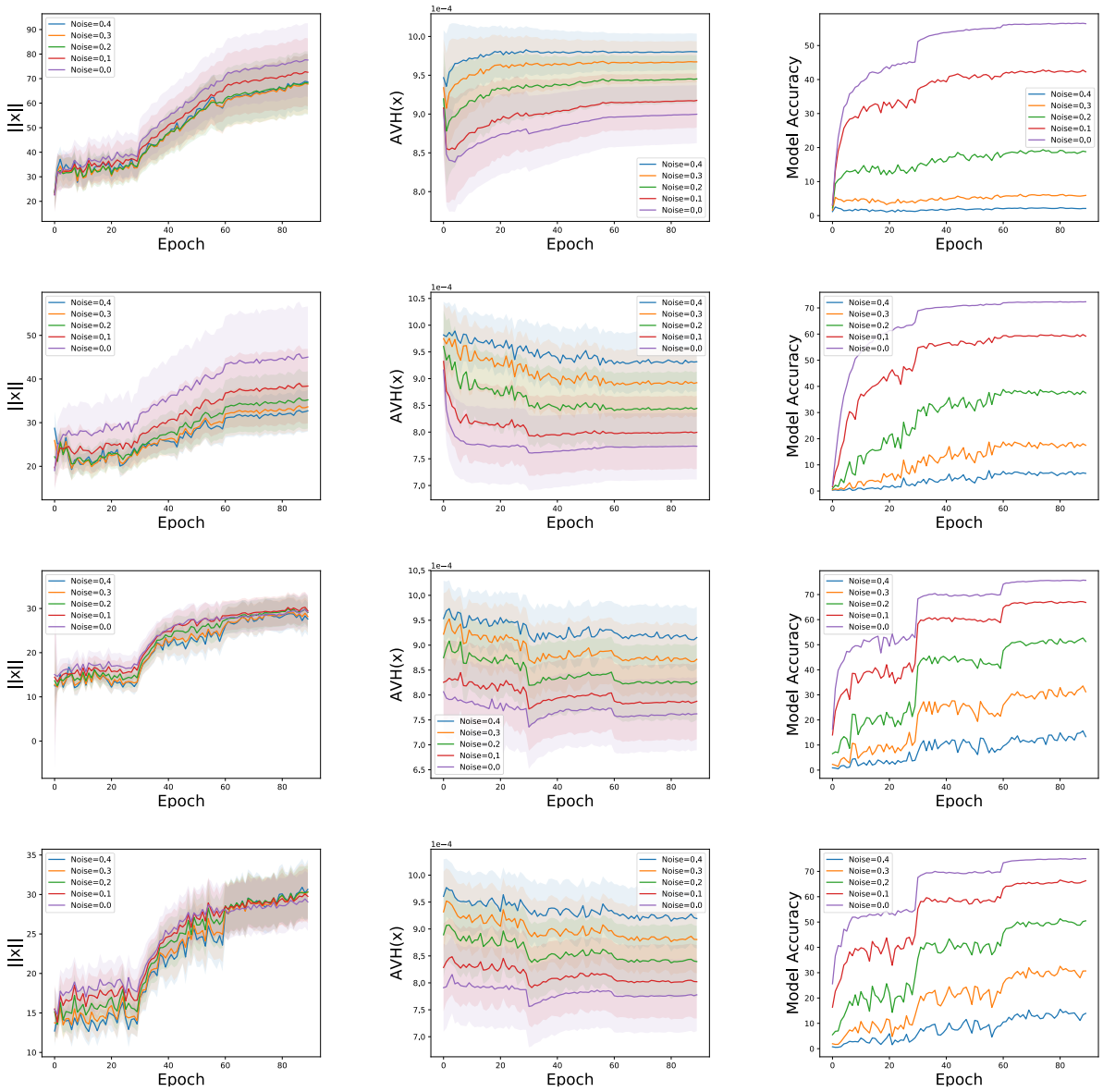


*Figure 11.* Averaged training dynamics across different noise degradation levels. Columns from left to right: number of epochs vs. average $\ell_2$ norm, number of epochs vs. average AVH score, and number of epochs vs. model accuracy. Rows from top to bottom: dynamics corresponding to AlexNet, VGG-19, ResNet-50, and DenseNet-121.

**Dynamics across contrast degradation levels:** In Figure 12, we illustrate the averaged training dynamics on the ImageNet validation set across five image contrast degradation levels - [0.1, 0.2, 0.3, 0.6, 1.0].

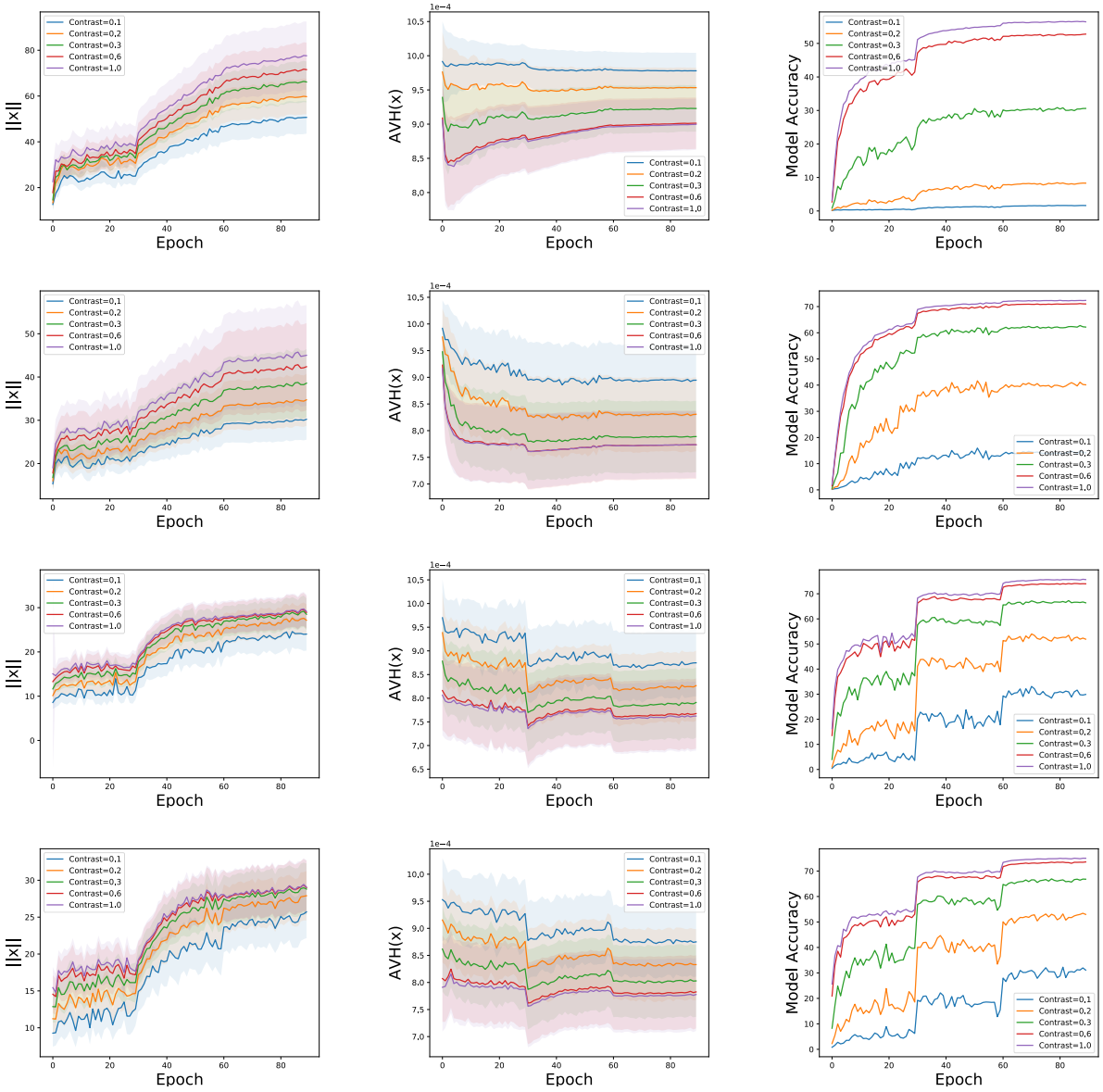


*Figure 12.* Averaged training dynamics across different contrast degradation levels. Columns from left to right: number of epochs vs. average $\ell_2$ norm, number of epochs vs. average AVH score, and number of epochs vs. model accuracy. Rows from top to bottom: dynamics corresponding to AlexNet, VGG-19, ResNet-50, and DenseNet-121.

One can see that the observations from Section 3 in the main paper also hold on this set of experiments.

### B.2. Additional Results on CIFAR-10, CIFAR-100 and MNIST

Figure 13 and 14 show the dynamics of average $\ell_2$ norm of the embeddings and average AVH(x) on CIFAR-10 and CIFAR-100 datasets respectively. We can observe the similar phenomenons we have discussed in section 3. It further supports our theoretical foundation from (Soudry et al., 2018) that gradient descent converges to the same direction as maximum margin solutions irrelevant to the $\ell_2$ norm of classifier weights or feature embeddings.

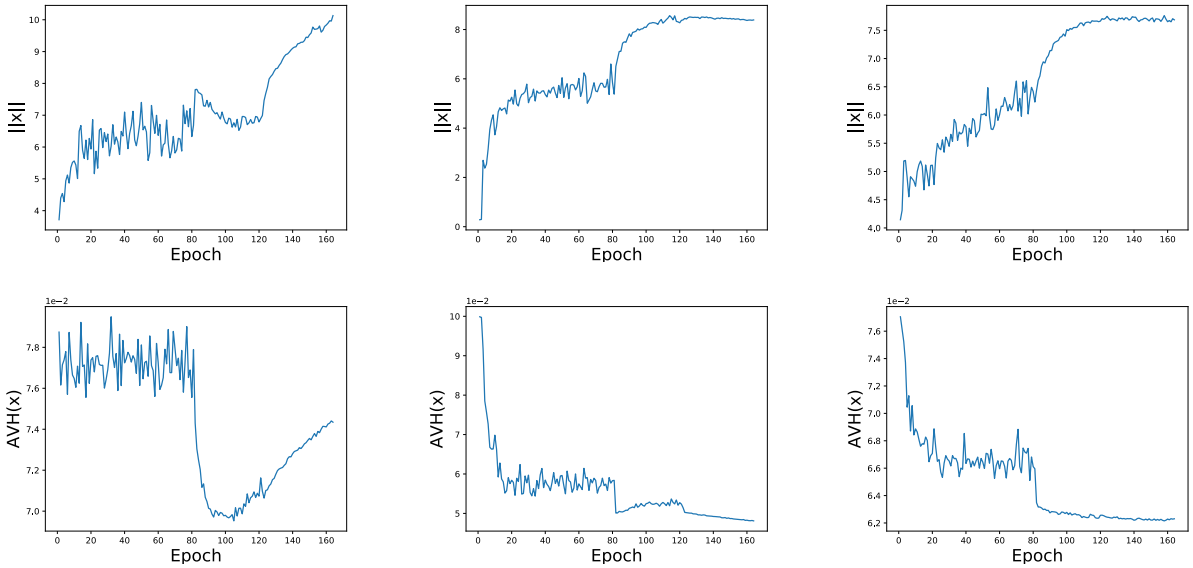


*Figure 13.* The top three plots show the number of Epochs v.s. Average $\ell_2$ norm across CIFAR-10 validation samples. The bottom three plots represent number of Epochs v.s. Average AVH(x). From left to right, we use AlexNet, VGG-19 and ResNet-50.

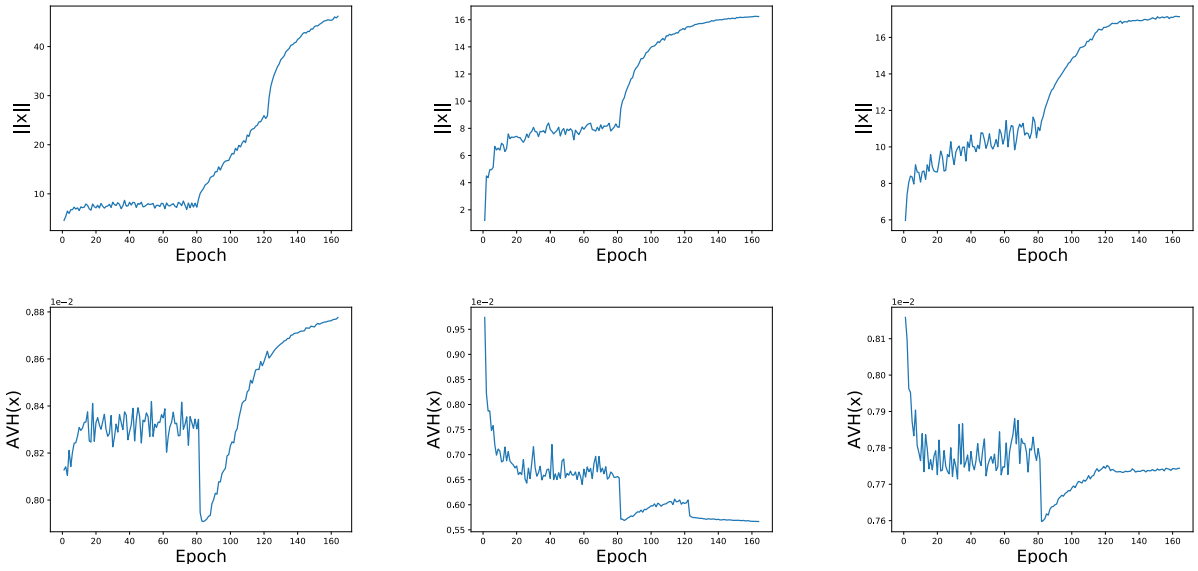


*Figure 14.* The top three plots show the number of Epochs v.s. Average $\ell_2$ norm across CIFAR-100 validation samples. The bottom three plots represent number of Epochs v.s. Average AVH(x). From left to right, we use AlexNet, VGG-19 and ResNet-50.

Figure 15 illustrates how the average norm of the feature embedding and AVH between feature and class embedding for testing samples vary in 60 iterations during the training process on MNIST. 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.

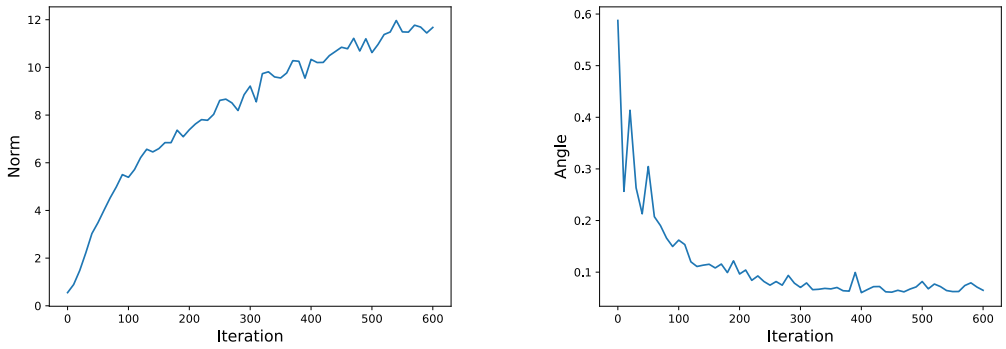


*Figure 15.* Average $\ell_2$ norm and angle of the embedding across all testing samples v.s. iteration number.

## C. Additional Discussions for Observations in Training Dynamics

Observation 2 in section 3 describes that AVH hits a plateau very early even when the accuracy or loss is still improving. AVH is more important than $\|\mathbf{x}\|_2$ 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 for easy examples. Let us consider a simple binary classification case where the softmax score for class 1 is

$$\frac{\exp(\boldsymbol{w}_1\boldsymbol{x})}{\sum_i \exp(\boldsymbol{w}_i\boldsymbol{x})} = \frac{\exp(\|\boldsymbol{w}_1\|\|\boldsymbol{x}\|\cos(\theta_{\boldsymbol{w}_1,\boldsymbol{x}}))}{\sum_i \exp(\|\boldsymbol{w}_i\|\|\boldsymbol{x}\|\cos(\theta_{\boldsymbol{w}_i,\boldsymbol{x}}))} \tag{6}$$

where $\boldsymbol{w}_i$ is the classifier weights of class $i$, $\boldsymbol{x}$ is the input deep feature and $\theta_{\boldsymbol{w}_i,\boldsymbol{x}}$ is the angle between $\boldsymbol{w}_i$ and $\boldsymbol{x}$. To simplify, we assume the norm of $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are the same, and then the classification result is based on the angle now. For easy examples, during early stage of the training, $\theta_{\boldsymbol{w}_1,\boldsymbol{x}}$ quickly becomes smaller than $\theta_{\boldsymbol{w}_2,\boldsymbol{x}}$ and the network will classify the sample $\boldsymbol{x}$ as class 1. However, in order to further minimize the cross-entropy loss after making $\theta_{\boldsymbol{w}_1,\boldsymbol{x}}$ smaller than $\theta_{\boldsymbol{w}_2,\boldsymbol{x}}$, the network has a trivial solution: increasing the feature norm $\|\boldsymbol{x}\|$ instead of further minimizing the $\theta_{\boldsymbol{w}_1,\boldsymbol{x}}$. It is obviously a much more difficult task to minimize $\theta_{\boldsymbol{w}_1,\boldsymbol{x}}$ rather than increasing $\|\boldsymbol{x}\|$. Therefore, the network will tend to increase the feature norm $\|\boldsymbol{x}\|$ to minimize the cross-entropy loss, which is equivalent to maximizing the Model Confidence in class 1. In fact, this also matches our empirical observation that the feature norm keeps increasing during training. Moreover, this also matches our empirical result that AVH easily gets saturated while Model Confidence can keep improving. For hard examples, after some time of training, the feature norms are unavoidable also increasing (although slower than those of easy examples). We can see from equation 6 that when $\|\boldsymbol{x}\|$ is very large and $\cos(\theta_{\boldsymbol{w}_i,\boldsymbol{x}})$ is very small, improving the angle becomes much harder because for a bit improvement on these examples, the model needs to sacrifice a lot for those easy ones. For the case of value of $\cos(\theta_{\boldsymbol{w}_i,\boldsymbol{x}})$ is around the decision boundary, a little change to AVH can cause a lot improvement on loss and accuracy and thereby we can still observe the change of accuracy and loss while AVH plateaus. More details about why $\|\boldsymbol{x}\|$ might be harmful in the training process is in Appendix E.

# D. Additional Experiments for Connections to Human Visual Hardness

## D.1. Additional Results for Correlation Testings

In order to run rigorous correlation testings, besides computing the Spearman coefficient, we provide additional results on Pearson and Kendall Tau correlation coefficients. Moreover, we show results for all four architectures, AlexNet, VGG-19, ResNet-50 and DenseNet-121 in Table 6, 7, 8 and 9 respectively to support our claims in section 4.

*Table 6.* This table presents the Spearman's rank correlation coefficients between Human Selection Frequency and AVH, Model Confidence on AlexNet. Note that we show the absolute value of the coefficient which represents the strength of the correlation. Z value is computed by Z scores of both coefficients. p-value$<$ 0.05 indicates that the result is statistically significant.

| Type | Coef with AVH | Coef with Model Confidence | $Z_{avh}$ | $Z_{mc}$ | Z value | p-value |
|---|---|---|---|---|---|---|
| Spearman's rank | 0.339 | 0.325 | 0.352 | 0.337 | 1.92 | 0.027 |
| Pearson | 0.324 | 0.31 | 0.336 | 0.320 | 1.90 | 0.028 |
| Kendall's Tau | 0.244 | 0.23 | 0.249 | 0.234 | 1.81 | 0.035 |

*Table 7.* This table presents the Spearman's rank correlation coefficients between Human Selection Frequency and AVH, Model Confidence on VGG-19. Note that we show the absolute value of the coefficient which represents the strength of the correlation. Z value is computed by Z scores of both coefficients. p-value$<$ 0.05 indicates that the result is statistically significant.

| | Coef with AVH | Coef with Model Confidence | $Z_{avh}$ | $Z_{mc}$ | Z value | p-value |
|---|---|---|---|---|---|---|
| Spearman's rank | 0.349 | 0.335 | 0.364 | 0.348 | 1.94 | 0.026 |
| Pearson | 0.358 | 0.343 | 0.374 | 0.357 | 2.09 | 0.018 |
| Kendall's Tau | 0.244 | 0.229 | 0.249 | , 0.233 | 1.94 | 0.026 |

*Table 8.* This table presents the Spearman's rank correlation coefficient between Human Selection Frequency and AVH, Model Confidence on ResNet-50. Note that we show the absolute value of the coefficient which represents the strength of the correlation. Z value is computed by Z scores of both coefficients. p-value$<$ 0.05 indicates that the result is statistically significant.

| | Coef with AVH | Coef with Model Confidence | $Z_{avh}$ | $Z_{mc}$ | Z value | p-value |
|---|---|---|---|---|---|---|
| Spearman's rank | 0.360 | 0.325 | 0.377 | 0.337 | 4.85 | $<$ .00001 |
| Pearson | 0.385 | 0.341 | 0.406 | 0.355 | 6.2 | $<$ .00001 |
| Kendall's Tau | 0.257 | 0.231 | 0.263 | 0.235 | 3.38 | .0003 |

*Table 9.* This table presents the Spearman's rank correlation coefficients between Human Selection Frequency and AVH, Model Confidence in DenseNet-121. Note that we show the absolute value of the coefficient which represents the strength of the correlation. Z value is computed by Z scores of both coefficients. p-value $<$ 0.05 indicates that the result is statistically significant.

| | Coef with AVH | Coef with Model Confidence | $Z_{avh}$ | $Z_{mc}$ | Z value | p-value |
|---|---|---|---|---|---|---|
| Spearman's | 0.367 | 0.329 | 0.4059 | 0.355 | 6.2 | $<$ .00001 |
| Pearson | 0.390 | 0.347 | 0.412 | 0.362 | 6.09 | $<$ .00001 |
| Kendall's Tau | 0.262 | 0.234 | 0.268 | 0.238 | 3.65 | .0001 |

### D.2. Additional Plots for Hypothesis Testings

**Additional plots for Section 4:** Figure 16 presents the correlation between Human Selection Frequency and AVH using AlexNet, VGG-19 and DenseNet-121.

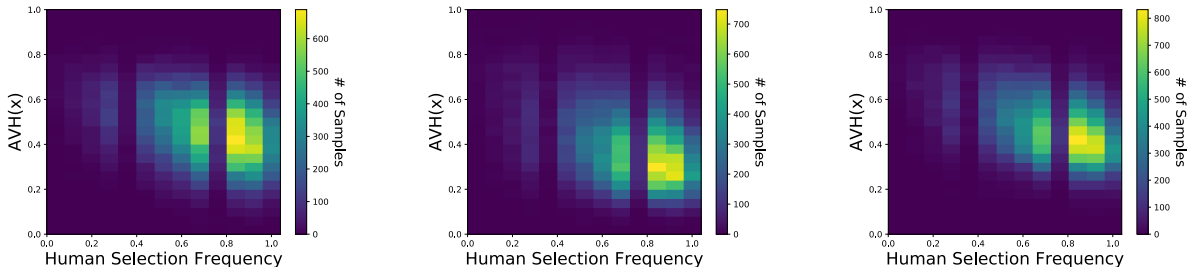


*Figure 16.* The three plots present the correlation between Human Selection Frequency and AVH using AlexNet, VGG-19 and DeseNet121.

**Correlation between AVH and image degradation:** In order to test if the results in Figure 5 from the main paper also hold on proxies other than human visual hardness (image degradation level), we perform the similar experiments but on the augmented ImageNet validation set. Figure 17 shows the correlation between $\mathcal{AVH}(\mathbf{x})$ and different noise degradation levels, while the plots in Figure 18 shows the correlation between $\mathcal{AVH}(\mathbf{x})$ and different contrast degradation levels. Along with Figure 16, these results all indicate that $\mathcal{AVH}(\mathbf{x})$ is a reliable measure of Human Visual Hardness.

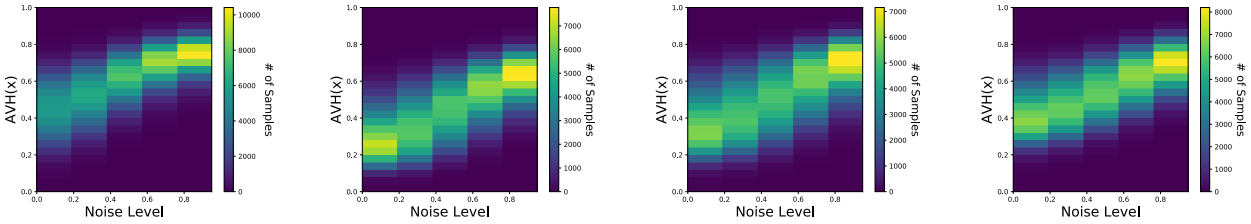


*Figure 17.* Correlation between noise degradation levels and AVH scores on AlexNet, VGG-19, ResNet-50 and DenseNet-121. Note that the larger the noise level is, the harder a human can recognize the image.

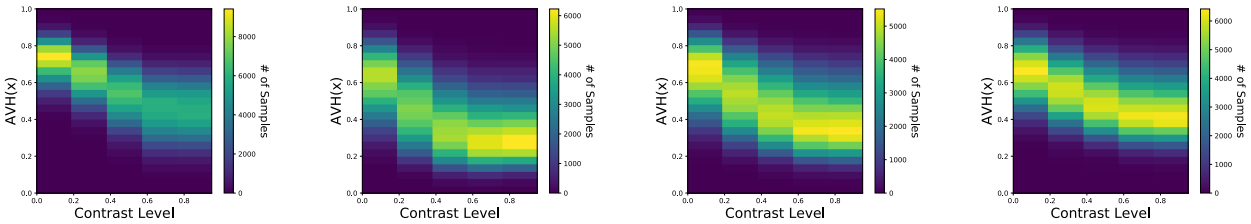


*Figure 18.* Correlation between contrast degradation levels and AVH scores on AlexNet, VGG-19, ResNet-50 and DenseNet-121. Note that the larger the contrast Level is, the easier a human can recognize the image.

**Additional plots for Hypothesis 4:** We further verify if presenting all samples across 1000 different classes affects the visualization of the correlation. According to WordNet (Fellbaum, 2005) hierarchy, we map the original 1000 fine-grained classes to 45 higher hierarchical classes. Figure 19 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.

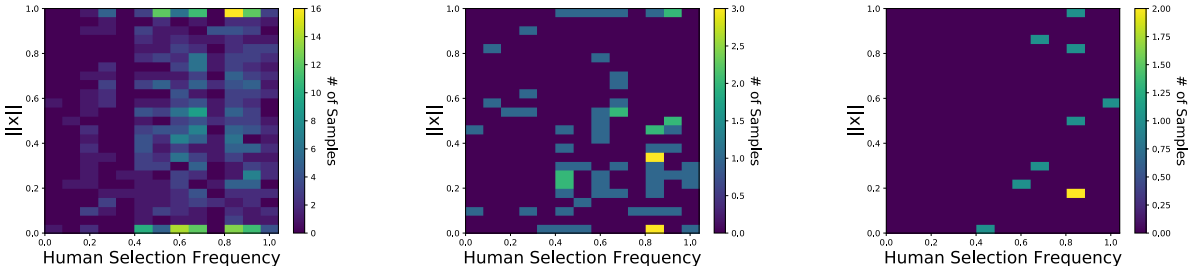


*Figure 19.* $\ell_2$ norm of the embedding vs. Human Selection Frequency under different class granularities (according to WordNet hierarchy). From left to right, there are 58, 7, 1 classes respectively. Human Selection Frequency is therefore computed based on the new class granularity.

## E. Additional discussions on the Difference between AVH and Model Confidence

The difference between AVH and Model Confidence lies in the feature norm and its role during training. To illustrate the difference, we consider a simple binary classification case where the softmax score (i.e., Model Confidence) for class 1 is

$$\frac{\exp(\boldsymbol{w}_1\boldsymbol{x})}{\sum_i \exp(\boldsymbol{w}_i\boldsymbol{x})} = \frac{\exp(\|\boldsymbol{w}_1\|\|\boldsymbol{x}\|\cos(\theta_{\boldsymbol{w}_1,\boldsymbol{x}}))}{\sum_i \exp(\|\boldsymbol{w}_i\|\|\boldsymbol{x}\|\cos(\theta_{\boldsymbol{w}_i,\boldsymbol{x}}))}$$

where $\boldsymbol{w}_i$ is the classifier weights of class $i$, $\boldsymbol{x}$ is the input deep feature and $\theta_{\boldsymbol{w}_i,\boldsymbol{x}}$ is the angle between $\boldsymbol{w}_i$ and $\boldsymbol{x}$. To simplify, we assume the norm of $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are the same, and then the classification result is based on the angle now. Once $\theta_{\boldsymbol{w}_1,\boldsymbol{x}}$ is smaller than $\theta_{\boldsymbol{w}_2,\boldsymbol{x}}$, the network will classify the sample $\boldsymbol{x}$ as class 1. However, in order to further minimize the cross-entropy loss after making $\theta_{\boldsymbol{w}_1,\boldsymbol{x}}$ smaller than $\theta_{\boldsymbol{w}_2,\boldsymbol{x}}$, the network has a trivial solution: increasing the feature norm $\|\boldsymbol{x}\|$ instead of further minimizing the $\theta_{\boldsymbol{w}_1,\boldsymbol{x}}$. It is obviously a much more difficult task to minimize $\theta_{\boldsymbol{w}_1,\boldsymbol{x}}$ rather than increasing $\|\boldsymbol{x}\|$. Therefore, the network will tend to increase the feature norm $\|\boldsymbol{x}\|$ to minimize the cross-entropy loss, which is equivalent to maximizing the Model Confidence in class 1. In fact, this also matches our empirical observation that the feature norm keeps increasing during training. Most importantly, one can notice that AVH will stay unchanged no matter how large the feature norm $\|\boldsymbol{x}\|$ is. Moreover, this also matches our empirical result that AVH easily gets saturated while Model Confidence can keep improving. Therefore, AVH is able to better characterize the visual hardness, since it is trivial for the network to increase feature norm. This is the fundamental difference between Model Confidence and AVH.

To get a more intuitive sense of how feature norm can affect the Model Confidence, we plot the value of the Model Confidence for two scenarios: $\theta_{\boldsymbol{w}_1,\boldsymbol{x}} < \theta_{\boldsymbol{w}_2,\boldsymbol{x}}$ and $\theta_{\boldsymbol{w}_1,\boldsymbol{x}} > \theta_{\boldsymbol{w}_2,\boldsymbol{x}}$. Under the case that the sample $\boldsymbol{x}$ belongs to class 1, once we have $\theta_{\boldsymbol{w}_1,\boldsymbol{x}} < \theta_{\boldsymbol{w}_2,\boldsymbol{x}}$, then we only need to increase the feature norm and can easily get nearly perfect confidence on this sample. In contrast, AVH will stay unchanged during the entire process and therefore is a more robust indicator for visual hardness than Model Confidence.

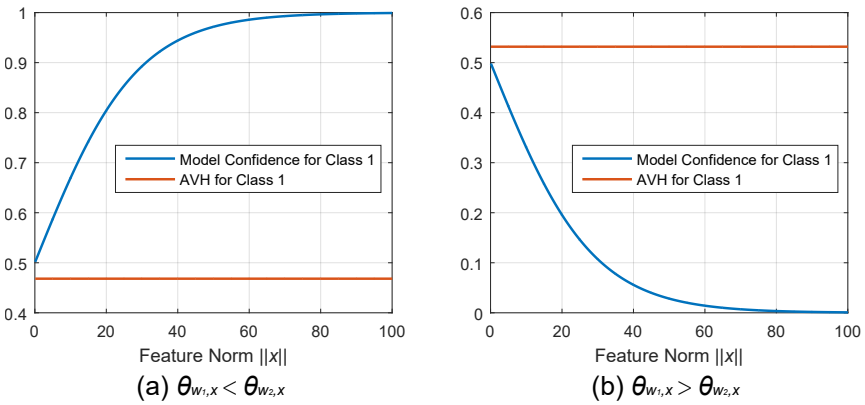


*Figure 20.* The comparison between AVH and Model Confidence when the feature norm keeps increasing. The figure is plotted according to the binary classification example discussed above. We assume $\|\boldsymbol{w}_1\| = \|\boldsymbol{w}_1\|$. When $\theta_{\boldsymbol{w}_1,\boldsymbol{x}} < \theta_{\boldsymbol{w}_2,\boldsymbol{x}}$, we use $\theta_1 = \pi/4 - 0.05$ and $\theta_2 = \pi/4 + 0.05$. When $\theta_{\boldsymbol{w}_1,\boldsymbol{x}} > \theta_{\boldsymbol{w}_2,\boldsymbol{x}}$, we use $\theta_1 = \pi/4 + 0.05$ and $\theta_2 = \pi/4 - 0.05$. Note that, unlike Model Confidence, the smaller AVH is, the more confident the network is (*i.e.*, the easier the sample is).

## F. Experimental Details

**Self-training and domain adaptation:** As mentioned in section 5, a major challenge of self-training is the amplification of error due to misclassified pseudo-labels. Therefore, traditional self-training methods such as CBST often use Model Confidence as the measure to select confidently labeled examples. The hope is that higher confidence potentially implies lower error rate. While this generally proves useful, the model tends to focus on the "less informative" samples, whereas ignoring the "more informative", harder ones near classier boundaries that could be essential for learning a better classifier. Figure 21 shows examples of what AVH selects and labels correctly but the softmax score does not select in CBST. We can see they are all visually confusing examples which can better help with the iterative self-training process when pseudo labeled correctly. The left one has the true label "Truck" but is easy to be confused with the "Car". The right one has the true label "Person" but is easy to be confused with the "Motor".

*Figure 21.* Two example images which AVH selects but softmax score do not. The left one has the true label "Truck" but is easy to be confused with the "Car". The right one has the true label "Person" but is easy to be confused with the "Motor".

**Domain generalization:** For domain generalization, we use the PACS benchmark dataset (Li et al., 2017) which contains consists of art painting, cartoon, photo and sketch domains. Each domain has the same 7 classes. Our experimental settings basically follow (Li et al., 2017). Specifically, we pick one domain as the unseen testing domain and train our model on the remaining three domains. The testing accuracy is evaluated on the unseen testing domain. Therefore, we will have 4 testing accuracies in total and we can use the average accuracy as the final evaluation metric. We use a convolutional neural network similar to (Liu et al., 2017c) with the detailed structure of [7×7, 64] $\Rightarrow$ 2×2 Max Pooling $\Rightarrow$ [3×3, 64]×3 $\Rightarrow$ 2×2 Max Pooling $\Rightarrow$ [3×3, 128]×3 $\Rightarrow$ 2×2 Max Pooling $\Rightarrow$ [3×3, 256]×3 $\Rightarrow$ 2×2 Max Pooling $\Rightarrow$ 512-dim Fully Connected. For example, [3×3, 64]×3 denotes 3 cascaded convolution layers with 64 filters of size 3×3. We use momentum SGD with momentum as $0.9$ and batch size $40$. Batch normalization and ReLU activation are used by default. Following the existing methods (Li et al., 2017; 2018; Balaji et al., 2018; Li et al., 2019), we will first pretrain our network on ImageNet with standard learning rate and decay, and then finetune on the PACS dataset with batch size $40$ and smaller learning rate ($1e-3$).

## G. Extensions and Applications

### Adversarial Example: A Counter Example?

Our claim about the stronger 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.

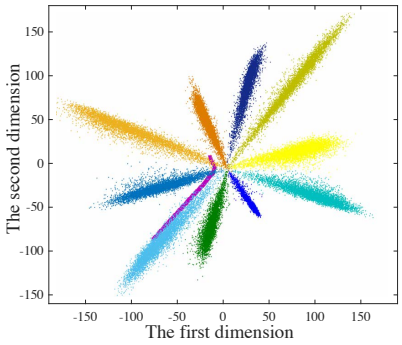


*Figure 22.* Trajectory of an adversarial example switching from one class to another. The purple line denotes the trajectory of the adversarial example.

We show a special case in Figure 22 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 22, 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.

**Measuring Human Visual Hardness is Hard:** Measuring Human Visual Hardness is non-trivial and dependent on many factors such as (i) How much are the annotators penalized for wrong answers and how much time are they given? (ii) What are the cultural and language differences that can cause annotators to be confused about the label categories. Figure 23 shows an example of groom from ImageNet dataset. Since a large contingent of Mturk users are from India, they have high confidence for this image, but the answer would be very different if asked a different population. The proxies we used in this paper, Human Selection Frequency and Image Degradation Level are best efforts.

*Figure 23.* An image of an Indian Groom from ImageNet.

**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). For instance, objective functions in deep metric learning consist of *e.g.*, triplet loss (Schroff et al., 2015) or contrastive loss (Hadsell et al., 2006), 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 (Schroff et al., 2015) and hard negative sample mining. These negative sample mining techniques highly depend on how one defines the hardness of samples. AVH can be potentially useful in this setting.

**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 (Yi et al., 2014) mostly contains white celebrities, so the neural network trained on CASIA-WebFace is highly biased against the other races. (Buolamwini & Gebru, 2018) 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 (Hinton et al., 2015) is partially determined by the sequence of input training data. (Liu et al., 2017a) theoretically shows feeding easy samples first and hard samples later (known as curriculum learning) can improve the convergence of model. (Bengio et al., 2009) 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.