ANGULAR VISUAL HARDNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

The mechanisms behind human visual systems and convolutional neural networks (CNNs) are vastly different. Hence, it is expected that they have different notions of ambiguity or hardness. In this paper, we make a surprising discovery: there exists a (nearly) universal score function for CNNs whose correlation with human visual hardness is statistically significant. We term this function as angular visual hardness (AVH) and in a CNN, it is given by the normalized angular distance between a feature embedding and the classifier weights of the corresponding target category. We conduct an in-depth scientific study. We observe that CNN models with the highest accuracy also have the best AVH scores. This agrees with an earlier finding that state-of-art models tend to improve on classification of harder training examples. We find that AVH displays interesting dynamics during training: it quickly reaches a plateau even though the training loss keeps improving. This suggests the need for designing better loss functions that can target harder examples more effectively. Finally, we empirically show significant improvement in performance by using AVH as a measure of hardness in self-training tasks.

1 INTRODUCTION

Convolutional Neural Networks (CNN) have pushed the boundaries on a wide range of computer vision tasks. On certain large-scale benchmarks such as ImageNet [10], CNNs have even surpassed human-level accuracy. Despite such progress, CNNs are still far from matching human vision on many measures such as robustness, adaptability and few shot learning [22] and could suffer from various biases. For example, CNNs pre-trained on Imagenet are biased towards textures [18]. These biases can result in CNNs being overconfident, or prone to domain gaps and adversarial attacks.

Human and CNNs fundamentally differ in terms of uncertainty or ambiguity. CNNs suffer from calibration problems and tend to be overconfident even if they are wrong [20, 33], which is not the case with human vision. Thus, the two systems differ in what they view as hard examples that appear ambiguous or uncertain. In this paper, we bridge the gap by proposing a novel score function on CNNs that correlates closely with human visual hardness. The first piece of this puzzle starts with the question of what is a good measure of human visual hardness. Recently, [48] argued that human selection frequency is a good measure. This is the average number of times an image gets picked by a crowd of annotators, when they are asked to pick an image from a pool that belongs to a certain specified category. Intuitively, human selection frequency depends on various factors like object sizes, poses, special filters applied to images etc. [48] collected human selection frequency scores on ImageNet validation set using the MTurk platform. In this paper, we use this dataset to verify our hypotheses. In addition to human selection frequencies, we also employ certain image degradation models [36] that have been shown to correlate well with human visual hardness.

An automatic detection of examples that are hard for human vision has numerous applications. [48] showed that state-of-art (SOTA) models perform better on hard examples (i.e., hard for humans). This implies that in order to improve generalization, the models need to improve accuracy on hard examples. This can be achieved through various learning algorithms such as curriculum learning [2] and self-paced learning [31] where being able to detect hard examples is crucial. Measuring sample confidence is also important in partially-supervised problems such as semi-supervised learning [62, 63], unsupervised domain adaptation [8] and weakly-supervised learning [57] due to their under-constrained nature. For instance, self-training [64] can easily reach to trivial solutions if one does not select pseudo-labels carefully based on correct measure of hardness. Moreover, by identifying hard
examples, one can detect various biases in current CNN models. Sample hardness can also be used to identify implicit distribution imbalance in datasets to ensure fairness and remove societal biases [4].

Our contributions are summarized as follows:

**Angular visual hardness (AVH):** Given a CNN, we propose a novel score function that has strong correlation with human visual hardness. It is the normalized angular distance between the image feature embedding and the weights of the target category (See Figure 1). 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 and we find strong empirical evidence to support this claim.

The AVH score is inspired by the intuition in [41] 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. [41] used this insight to improve the generalizability and adversarial robustness of the neural network. On the other hand, we use it to study the correspondence with human visual hardness.

We validate that there is a statistically-significant correlation between AVH and 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 serves as its proxy on datasets where such information is not available and is beneficial to a number of downstream tasks.

We observed the evolution of AVH score during training of CNN models. It plateaus early in training even if the training (cross-entropy) loss keeps improving. This suggests the need to design better loss functions that can improve performance on hard examples. We also validate the argument in [48] that improving on hard examples is the key to improve the generalization by verifying that the state-of-art (SOTA) models have the best average AVH scores.

Finally, we empirically show the superiority of AVH with its application to self-training for unsupervised domain adaptation. With AVH being an improved confidence measure, our proposed self-training framework renders considerably improved pseudo-label selection and category estimation, leading to state-of-the-art results with significant performance gain over baselines.

2 RELATED WORK

**Example hardness measures:** Recently, measuring sample hardness for deep learning models has been widely studied with loss value [54], relative Euclidean distance [52, 58] and gradient norm [27]. On the other hand, there is a rich history in cognitive and neuroscience communities to understand human visual perception [6, 7, 15, 43], 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 [36]. They can be affected by the ambiguity of different semantics [26] such as occlusion, distortion, motion blur, and inherent similarity among objects. Due to the expensive human labeling process, such detailed semantic information is typically not present in large-scale image benchmarks used to train the CNN models.

**Angular distance in CNNs:** [61] 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. [39] 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. [41] 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. [34, 40] propose well-performing regularizations based on angular diversity to improve the network generalization.

**Image degradation:** Because CNNs and humans achieve similar accuracy on a wide range of tasks on benchmark datasets, a number of works have investigated similarities and differences between CNNs and human vision [3, 5, 9, 12, 13, 28, 44, 47, 59]. Since human annotation data is hard to come by, researchers have proposed an alternative measure of visual hardness on images based on image degradation [36]. It involves adding noise or changing image properties such as contrast, blurriness, and brightness. [17] 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 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 [20]. It is well-known that the deep neural networks are mis-calibrated and there has been a rich literature trying to solve this problem [20, 30]. 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 from Scientific Testing: 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. We use the ImageNet [11] benchmark in all following experiments. Particularly, we take advantage of the Human Selection Frequency information for validation images provided by the recent paper [48]. 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 [29], VGG19 [55], DenseNet121 [25], ResNet50 [23]. We select ResNet50 as the representative model for some experiments.

Denote $\mathbb{S}^n$ as the unit $n$-sphere, formally, $\mathbb{S}^n = \{ x \in \mathbb{R}^{n+1} \| x \|_2 = 1 \}$. Below by $A(\cdot, \cdot)$, we denote the angular distance between two points on $\mathbb{S}^n$, i.e., $A(u, v) = \arccos \left( \frac{\langle u, v \rangle}{\|u\| \|v\|} \right)$. 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 $W = w_i | 0 < i \leq C$ as the set of weights for all $C$ classes in the final layer of the classifier.

**Definition 1 (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.

**Definition 2 (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, $e^{w_y x} \sum_{i=1}^{C} e^{w_i x}$.

**Definition 3 (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 [48], if $b$ 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 is defined as $\frac{b}{m}$.
Table 1: presents the spearman rank correlations between human selection frequency and AVH, Model Confidence and L2 Norm of the Embedding in ResNet50 for different visual hardness bin of samples. Noted here we show the absolute value of the coefficient which represents the strength of the correlation. For example, [0, 0.2] denotes the samples that have human selection frequency from 0 to 0.2.

<table>
<thead>
<tr>
<th>Number of Samples</th>
<th>z-score</th>
<th>Total</th>
<th>[0, 0.2]</th>
<th>[0.2, 0.4]</th>
<th>[0.4, 0.6]</th>
<th>[0.6, 0.8]</th>
<th>[0.8, 1, 0]</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVH</td>
<td>0.377</td>
<td>0.36</td>
<td>0.228</td>
<td>0.125</td>
<td>0.124</td>
<td>0.103</td>
<td>0.094</td>
</tr>
<tr>
<td>Model Confidence</td>
<td>0.337</td>
<td>0.325</td>
<td>0.192</td>
<td>0.122</td>
<td>0.102</td>
<td>0.078</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>-0.0017</td>
<td>0.0015</td>
<td>0.0007</td>
</tr>
</tbody>
</table>

3.2 Connections and Gaps Between the Human Visual System and CNN

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. In this section, we first provide four hypothesis and test them accordingly.

**Hypothesis 1.** AVH has a correlation with Human Selection Frequency.

**Outcome: Null Hypothesis Rejected**

Correspondingly, after evaluating each validation sample on pre-trained models, we extract feature embeddings $x$ and also the class weights $W$ to compute $AVH(x)$. Noted that we linear scale the range of $AVH(x)$ to $[0, 1]$. Table 1 shows the overall strong correlation of $AVH(x)$ and Human Selection Frequency consistently (p-value is $< 0.001$ rejecting the null hypothesis). From the coefficients represented for different bins of example hardness, we can see that the harder the examples, the weaker the correlation. Noted that we also check the results across four different CNN architectures and we found that better model has higher coefficient.

The plot on the left in Figure 2 help visualize the 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. In order to test if the strong correlation holds for all models, we perform the same experiments on AlexNet, VGG19 and DenseNet121.

**Hypothesis 2.** Model Confidence has a correlation with Human Selection Frequency.

**Outcome: Null Hypothesis Rejected**

An interesting observation in [48] 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.
We first apply Fisher transformation for each correlation coefficient: \( Z \) would be:

After discovering the strong correlation of human visual hardness and A VH score, a natural question we conduct similar experiments like previous section to demonstrate the correlation between the norm is larger, the prediction from the model is also more confident, to some extent. Therefore, from the two above-mentioned Z scores and sample sizes. According to Z table, \( p-value \) is 0.00001. Therefore, we reject the null hypothesis and conclude that AVH has statistically significant stronger correlation with Human Selection Frequency than Model Confidence. In later section 5, we also empirically show that such stronger correlation brings cumulative advantages in some applications.

**Hypothesis 3. AVH has a stronger correlation to Human Selection Frequency than Model Confidence.**

**Outcome: Null Hypothesis Rejected**

We first apply Fisher transformation for each correlation coefficient: \( Z \) score for coefficient of AVH becomes 0.377 and that of Model Confidence becomes 0.337. Then we determined the \( Z \) value to be 4.85 from the two above-mentioned \( Z \) scores and sample sizes. According to \( Z \) table, \( p-value \) is 0.00001. Therefore, we reject the null hypothesis and conclude that AVH has statistically significant stronger correlation with Human Selection Frequency than Model Confidence. In later section 5, we also empirically show that such stronger correlation brings cumulative advantages in some applications.

**Hypothesis 4. \( \|x\|_2 \) has a correlation with Human Selection Frequency.**

**Outcome: Failure to Reject Null Hypothesis**

[41] 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. Table 1 presents the results for the correlation test. We omit the results for \( p-value \) in the table and report here that they are all much higher than 0.05, indicating there is no correlation between \( \|x\|_2 \) and Human Selection Frequency.

The right plot in figure 2 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 [14] 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 \( \|x\|_2 \) 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.

### 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 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 [29], VGG19 [55], ResNet50 [23] and DenseNet121 [25] (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
Figure 3: The top three plots show the number of Epochs v.s. Average $\ell_2$ 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. From left to right, we use AlexNet, Vgg19 and ResNet50. The plots for DenseNet are in Appendix.

DenseNet121 and ResNet50 is 0.1. For human visual hardness based on Human Selection Frequency, we split all the validation images into 5 bins, $[0, 0.2], [0.2, 0.4], [0.4, 0.6], [0.6, 0.8], [0.8, 1.0]$, 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 3, 10 and 11 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 3, 10 and 11 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 [34, 38, 40].

Observation 3: AVH’s correlation with human selection frequency consistently holds across models throughout the training process. In Figure 3, 10 and 11, 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 3, 15, 10 and 11, 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 [48] that better models also generalize better on samples across different human visual hardness. Moreover, 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 [38] and class granularity [29]. But when extracting the features for the classes that do not exist in training set, such training data prior is undesired. Since AVH does not consider the norm of feature embeddings, it may better evaluate the generalization of the deep network.

Conjecture on training dynamics of CNNs. From Figure 3 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 caused by that the CNN is unable to decrease the angle to correctly classify examples and thereby also unable to increase the norms (because it may otherwise increase the loss).

5 Application to Self-training for Domain Adaptation

(Unsupervised) domain adaptation [1] presents an important transfer learning problem with wide applications. Deep self-training [32] recently emerged as a powerful framework towards addressing this problem [49, 53, 64, 65]. Here we show the application of AVH as an improved confidence measure in self-training that could significantly benefit the domain adaptation task.

Dataset: We conduct experiments on the VisDA-17 [46] dataset which is a widely used major benchmark for domain adaptation in image classification. The dataset contains a total number of 152,409 2D synthetic images from 12 categories in the source training set, and 55,400 real images from MS-COCO [35] with the same set of categories as the target domain validation set. We follow the protocol of previous works to train a source model with the synthetic training set, and report the model performance on target validation set upon adaptation.
Baseline: We choose class-balanced self-training (CBST) [64] as our starting self-training baseline considering its good performance on domain adaptation. We also compare our model with confidence regularized self-training (CRST) [65], a more recent state-of-the-art self-training framework improved over CBST with network prediction/pseudo-label regularized with smoothness. Specifically, our work follows the exact implementation of CBST/CRST in [65].

Specifically, given the labeled source domain training set \(x^s_t \in X^S\) and the unlabeled target domain data \(x^t \in X^T\), with known source labels \(y^s_t = (y^s_t(1), ..., y^s_t(K)) \in Y^S\) and unknown target labels \(\hat{y}^t = (\hat{y}^t(1), ..., \hat{y}^t(K)) \in Y^T\) from \(K\) classes, CBST performs joint network learning and pseudo-label estimation by treating pseudo-labels as discrete learnable latent variables with the following loss:

\[
\min_{w, Y^t} \mathcal{L}_{CB}(w, Y^t) = - \sum_{s \in S} \sum_{k=1}^K y^s_t(k) \log p(k|x^s_t; w) - \sum_{t \in T} \sum_{k=1}^K \hat{y}^t(k) \log p(k|x^t; w) \lambda_k
\]

s.t. \(\hat{y}^t \in E^K \cup \{0\}, \forall t\)

where the feasible set of pseudo-labels is the union of \(\{0\}\) and the \(K\) dimensional one-hot vector space \(E^K\), and \(w\) and \(p(k|x; w)\) represent the network weights and the classifier’s softmax probability for class \(k\), respectively. In addition, \(\lambda_k\) serves as a class-balancing parameter controlling the pseudo-label selection of class \(k\), and is determined by the softmax confidence ranked at portion \(p\) (in descending order) among samples predicted to class \(k\). Therefore, only one parameter \(p\) is used to determine all \(\lambda_k\)’s. The optimization problem in (1) can be solved via minimizing with respect to \(w\) and \(Y^t\) alternatively, and the solver of \(Y^t\) can be written as:

\[
y_t^{(k)*} = \begin{cases} 
1, & \text{if } k = \arg \max_k \left\{ \frac{p(k|x^t; w)}{\lambda_k} \right\} \text{ and } p(k|x^t; w) > \lambda_k \\
0, & \text{otherwise}
\end{cases}
\]

The optimization with respect to \(w\) is simply normal network re-training with source labels and estimated pseudo-labels. And the complete self-training process involves alternatively repeat of network re-training and pseudo-label estimation.

CBST+AVH: We seek to improve the pseudo-label solver with better confidence measure from AVH. To this end, we propose the following definition of angular visual confidence (AVC) to represent the predicted probability of class \(c\):

\[
AVC(c|x; w) = \frac{\pi - A(x, w_c)}{\sum_{c=1}^K (\pi - A(x, w_k))},
\]

and the pseudo-label estimation in CBST+AVH is accordingly defined as:

\[
y_t^{(k)*} = \begin{cases} 
1, & \text{if } k = \arg \max_k \left\{ \frac{AVC(k|x^t; w)}{\lambda_k} \right\} \text{ and } AVC(k|x^t; w) > \lambda_k \\
0, & \text{otherwise}
\end{cases}
\]

where \(\lambda_k\) is differently determined by referring to \(AVC(k|x^t; w)\) ranked at portion \(p\) among samples predicted to class \(k\) by AVH, following a similar definition of \(\lambda_k\) in CBST. In addition, network re-training in CBST+AVH follows the softmax self-training loss in (1).

One could see that AVH changes the self-training behavior from two ways with the conditions in (4):

Improved selection: This is determined by \(AVC(k|x^t; w) > \lambda_k\).

Improved classification: This is determined by \(k = \arg \max_k \left\{ \frac{AVC(k|x^t; w)}{\lambda_k} \right\}\).

Specifically, the former determines which samples are not ignored during self-training based on AVC, whereas the latter determines the pseudo-label class by taking argmax over normalized AVC scores. With calibrated confidence that better resembles human visual hardness, both aspects are likely to considerably influence the performance of self-training.

Experimental Results: We present the results of the proposed method in Table 3, and also show its performance with respect to different self-training epochs in Figure 4. One could see that CBST+AVH outperforms both CBST and CRST by a very significant margin. We would like to emphasize that this is a very compelling result under “apples to apples” comparison with the same source model, implementation and hyper-parameters.

---

We consider MRKLD+LRENT which is reported to be the highest one in [65].
Table 2: Statistics of the examples selected by CBST+AVH and CBST/CRST.

<table>
<thead>
<tr>
<th></th>
<th>TP Rate</th>
<th>AVH (avg)</th>
<th>Model Confidence(avg)</th>
<th>Norm $|x|$ (avg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBST+AVH</td>
<td>0.844</td>
<td>0.118</td>
<td>0.961</td>
<td>20.84</td>
</tr>
<tr>
<td>CBST/CRST</td>
<td>0.848</td>
<td>0.117</td>
<td>0.976</td>
<td>21.28</td>
</tr>
</tbody>
</table>

Table 3: Experimental results on VisDA17.

<table>
<thead>
<tr>
<th>Method</th>
<th>Aero</th>
<th>Bike</th>
<th>Bus</th>
<th>Car</th>
<th>Horse</th>
<th>Knife</th>
<th>Motor</th>
<th>Person</th>
<th>Plant</th>
<th>Skateboard</th>
<th>Train</th>
<th>Truck</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source [51]</td>
<td>55.1</td>
<td>53.3</td>
<td>61.9</td>
<td>59.1</td>
<td>80.6</td>
<td>17.9</td>
<td>79.7</td>
<td>31.2</td>
<td>81.0</td>
<td>26.5</td>
<td>73.5</td>
<td>8.5</td>
<td>52.4</td>
</tr>
<tr>
<td>MMD [42]</td>
<td>87.1</td>
<td>63.0</td>
<td>76.5</td>
<td>42.0</td>
<td>90.3</td>
<td>42.9</td>
<td>85.9</td>
<td>53.1</td>
<td>49.7</td>
<td>36.3</td>
<td>85.8</td>
<td>20.7</td>
<td>61.1</td>
</tr>
<tr>
<td>DANN [16]</td>
<td>81.9</td>
<td>77.7</td>
<td>82.8</td>
<td>44.3</td>
<td>81.2</td>
<td>29.5</td>
<td>65.1</td>
<td>28.6</td>
<td>51.9</td>
<td>54.6</td>
<td>82.8</td>
<td>7.8</td>
<td>57.4</td>
</tr>
<tr>
<td>ENT [10]</td>
<td>80.3</td>
<td>75.5</td>
<td>75.8</td>
<td>48.5</td>
<td>77.9</td>
<td>27.3</td>
<td>69.7</td>
<td>40.2</td>
<td>46.5</td>
<td>46.6</td>
<td>79.3</td>
<td>16.0</td>
<td>57.0</td>
</tr>
<tr>
<td>MCD [50]</td>
<td>87.0</td>
<td>60.9</td>
<td>83.7</td>
<td>64.0</td>
<td>88.9</td>
<td>79.6</td>
<td>84.7</td>
<td>76.9</td>
<td>88.6</td>
<td>40.3</td>
<td>83.0</td>
<td>25.8</td>
<td>71.9</td>
</tr>
<tr>
<td>ADR [51]</td>
<td>87.8</td>
<td>79.5</td>
<td>83.7</td>
<td>65.3</td>
<td>92.3</td>
<td>61.8</td>
<td>88.9</td>
<td>73.2</td>
<td>87.8</td>
<td>60.0</td>
<td>85.5</td>
<td>32.3</td>
<td>74.8</td>
</tr>
<tr>
<td>Source [65]</td>
<td>68.7</td>
<td>36.7</td>
<td>61.3</td>
<td>70.4</td>
<td>61.9</td>
<td>5.9</td>
<td>82.6</td>
<td>25.5</td>
<td>75.6</td>
<td>29.4</td>
<td>83.8</td>
<td>10.9</td>
<td>51.6</td>
</tr>
<tr>
<td>CBST [65]</td>
<td>87.2</td>
<td>78.8</td>
<td>56.5</td>
<td>55.4</td>
<td>85.1</td>
<td>79.2</td>
<td>83.8</td>
<td>77.7</td>
<td>82.8</td>
<td>88.8</td>
<td>69.0</td>
<td>72.0</td>
<td>76.4</td>
</tr>
<tr>
<td>CRST [65]</td>
<td>88.0</td>
<td>79.2</td>
<td>61.0</td>
<td>60.0</td>
<td>87.5</td>
<td>81.4</td>
<td>86.3</td>
<td>78.8</td>
<td>85.6</td>
<td>86.6</td>
<td>73.9</td>
<td>68.8</td>
<td>78.1</td>
</tr>
<tr>
<td>Proposed</td>
<td>93.3</td>
<td>80.2</td>
<td>78.9</td>
<td>60.9</td>
<td>88.4</td>
<td>89.7</td>
<td>88.9</td>
<td>79.6</td>
<td>89.5</td>
<td>86.8</td>
<td>81.5</td>
<td>60.0</td>
<td>81.5</td>
</tr>
</tbody>
</table>

Analysis: 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 confidence measure to select confidently labeled examples. The hope is that higher confidence potentially implies lower error rate. While this in generally proves useful, the model tends to focus on the “less informative” samples, whereas ignoring the “more informative”, harder ones near classifier boundaries that could be essential for learning a stronger classifier.

Table 3: Experimental results on VisDA17.

An advantage we observe from AVH is that the improved calibration leads to more frequent sampling of harder samples, whereas the pseudo-label classification on these hard samples generally outperforms softmax results. Table 2 shows some statistics of the examples selected with AVH and Model Confidence respectively at the beginning of the training process. The true positive rate (TP Rate) for of CBST+AVH remains similar to CBST/CRST, indicating AVH is not overall introducing more noisy examples compare to model confidence. On the other hand, it is observed that the average model confidence of AVH selected samples is lower, indicating there are more selected hard samples that are closer to the decision boundary. It is also observed that the average sample norm by AVH is also lower, confirming the influence of sample norm on ultimate model confidence.

Figure 4: Experimental results of comparison among AVH+CBST, CBST and CRST on VisDA2017 dataset.

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 [45, 56, 58]. For instance, objective functions in deep metric learning consist of e.g., triplet loss [52] or contrastive loss [21], 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 [52] 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 [60] mostly
contains white celebrities, so the neural network trained on CASIA-WebFace is highly biased against
the other races. [4] 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 trans-
fer [24] is partially determined by the sequence of input training data. [37] theoretically shows
feeding easy samples first and hard samples later (known as curriculum learning) can improve the
convergence of model. [2] 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.

7 **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. Then we empirically
show the huge advantage of AVH over Model Confidence in self-training for domain adaptation task.
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 other applications such as deep metric learning, fairness,
knowledge transfer, etc. and we plan to investigate them in future.

**REFERENCES**


[36] Peter H Lindsay and Donald A Norman. *Human information processing: An introduction to psychology*. Academic press, 2013. 1, 2, 3


[57] Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In CVPR, 2017. 1


[63] Xiaojin Zhu. Semi-supervised learning tutorial. 1


Appendix

A ADDITIONAL EXPERIMENTS

Figure 5: The three plots present the correlation between Noise Degradation Level and $\|x\|$ using AlexNet, VGG19 and ResNet50 (DenseNet121 in Appendix). Noted here, the larger the Noise Level is, the harder human can visualize the image.

Figure 6: The three plots present the correlation between Contrast Degradation Level and $\|x\|$ using AlexNet, VGG19 and ResNet50 (DenseNet121 in Appendix). Noted here, the larger the Contrast Level is, the easier human can visualize the image.

Figure 7: The three plots present the correlation between Human Selection Frequency and $\|x\|$ using AlexNet, VGG19 and ResNet50 (DenseNet121 in Appendix).

A.1 ADDITIONAL PLOTS FOR THE PREDICTIONS FOR THE HYPOTHESIS

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.

Hypothesis: AVH has a strong correlation with Image Degradation Level

In order to test if the results from Prediction 1 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 5 show the strong correlation between AVH$(x)$ and Noise Degradation Level while the plots in Figure 6 present the strong correlation between AVH$(x)$ and Contrast Degradation.
Figure 8: The left, middle and right plots respectively present the correlation between Human Selection frequency, Noise Degradation Level, Contrast Degradation Level and $\|x\|$ using DenseNet121.

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.

Hypothesis: $\|x\|_2$ has a 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 [14] 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.

Hypothesis: AVH has a correlation with Human Selection Frequency Additional Plots for DenseNet121 is shown in Figure 8.
Figure 10: The top three plots show the number of Epochs v.s. Average $\ell_2$ norm across ImageNet validation samples which are split into five bins based on image contrast degradation level 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. From left to right, we use AlexNet, Vgg19 and ResNet50. The plots for DenseNet are in Appendix.
Figure 11: The top three plots show the number of Epochs v.s. Average $\ell_2$ norm across ImageNet validation samples which are split into five bins based on image noise degradation level 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. From left to right, we use AlexNet, Vgg19 and ResNet50. The plots for DenseNet are in Appendix.
A.2 Additional Experiments for Observing Dynamics on MNIST

Figure 12 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 13 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 12: Average $\ell_2$ norm and angle of the embedding across all testing samples v.s. iteration number.](image1)

![Figure 13: $\ell_2$ norm and angle of the embedding of an easy sample and a hard sample v.s. iteration number.](image2)
A.3 ADDITIONAL EXPERIMENTS FOR TRAINING DYNAMICS ON IMAGE NET

Figure 14 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 15, 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 15, 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 15, 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 14](image1.png)

Figure 14: Average $\ell_2$ norm and angle of the embedding across testing samples with different level of hardness v.s. iteration number on DenseNet121.

![Figure 15](image2.png)

Figure 15: The left plot shows the number of Epochs v.s. Average $\ell_2$ 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.

![Figure 16](image3.png)

Figure 16: The left plot shows the number of Epochs v.s. Average $\ell_2$ 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.
Figure 17: The left plot shows the number of Epochs v.s. Average $\ell_2$ norm across ImageNet validation samples which are split into five bins based on image contrast 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.
B  A Special Case: Adversarial Examples

We show a special case in Figure 18 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 18, 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.

Figure 18: Trajectory of an adversarial example switching from one class to another. The purple line denotes the trajectory of the adversarial example.