Angular Visual Hardness

Beidi Chen 1 Weiyang Liu 2 Zhiding Yu 3 Jan Kautz 3 Anshumali Shrivastava 1 Animesh Garg 3 4 5 Anima Anandkumar 3 6

Abstract

Recent convolutional neural networks (CNNs) have led to impressive performance but often suffer from poor calibration. They tend to be overconfident, with the model confidence not always reflecting the underlying true ambiguity and hardness. In this paper, we propose angular visual hardness (AVH), a score given by the normalized angular distance between the sample feature embedding and the target classifier to measure sample hardness. We validate this score with an in-depth and extensive scientific study, and 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 improve on the classification of harder examples. We observe that the training dynamics of AVH is vastly different compared to the training loss. Specifically, AVH quickly reaches a plateau for all samples even though the training loss keeps improving. This suggests the need for designing better loss functions that can target harder examples more effectively. We also find that AVH has a statistically significant correlation with human visual hardness. Finally, we demonstrate the benefit of AVH to a variety of applications such as self-training for domain adaptation and domain generalization.

1. Introduction

Convolutional neural networks (CNNs) have achieved great progress on many computer vision tasks such as image classification (He et al., 2016; Krizhevsky et al., 2012), face recognition (Sun et al., 2014; Liu et al., 2017b; 2018a), and scene understanding (Zhou et al., 2014; Long et al., 2015a). On certain large-scale benchmarks such as ImageNet, CNNs have even surpassed human-level performance (Deng et al., 2009). Despite the notable progress, CNNs are still far from matching human-level visual recognition in terms of robustness (Goodfellow et al., 2014; Wang et al., 2018c), adaptability (Finn et al., 2017) and few-shot generalizability (Hariharan & Girshick, 2017; Liu et al., 2019), and could suffer from various biases. For example, ImageNet-trained CNNs are reported to be biased towards textures, and these biases may result in CNNs being overconfident, or prone to domain gaps and adversarial attacks (Geirhos et al., 2019).

Softmax score has been widely used as a confidence measure for CNNs but it tends to give over-confident output (Guo et al., 2017; Li & Hoiem, 2018). To fix this issue, one line of work considers confidence calibration from a Bayesian point of view (Springenberg et al., 2016; Lakshminarayanan et al., 2017). Most of these methods tend to focus on the calibration and rescaling of model confidence by matching expected error or ensemble. But how much they are correlated with human confidence is yet to be thoroughly studied. On the other hand, several recent works (Liu et al., 2016; 2017c; 2018b) conjecture that softmax feature embeddings tend to naturally decouple into norms and angular distances that are related to intra-class confidence and inter-class semantic difference, respectively. While inspiring, their conjecture lacks thorough investigation and we make surprising observations partially contradicting to the conjecture on intra-class confidence. This motivates us to conduct rigorous studies for reliable and semantics-related confidence measure.

Human vision is considered much more robust than current CNNs, but this does not mean humans cannot be confused.

Figure 1. Example images that confuse humans. Top row: images with degradation. Bottom row: images with semantic ambiguity.
Many images appear ambiguous or hard for humans due to various image degradation factors such as lighting conditions, occlusions, visual distortions, etc. or due to semantic ambiguity in not understanding the label category, as shown in Figure 1. It is therefore natural to consider such human ambiguity or visual hardness on images as the gold-standard for confidence measures. However, explicitly encoding human visual hardness in a supervised manner is generally not feasible, since hardness scores can be highly subjective and difficult to obtain. Fortunately, a surrogate for human visual hardness was recently made available on the ImageNet validation set (Recht et al., 2019). This is based on human selection frequency (HSF) - the average number of times an image gets picked by a crowd of annotators from a pool belonging to certain specified category. We adopt HSF as a surrogate for human visual hardness in this paper to validate our proposed angular hardness measure in CNNs.

Contribution: angular visual hardness (AVH). Given a CNN, we propose a novel score function for measuring sample hardness. It is the normalized angular distance between the image feature embedding and the weights of the target category (See Figure 2 as a toy example). The normalization takes into account the angular distances to other categories.

We make observations on the dynamic evolution of AVH scores during ImageNet training. We find that AVH plateaus early in training even though the training (cross-entropy) loss keeps decreasing. This is due to the nature of parameterization in softmax loss, of which the minimization goes in two directions: either by aligning the angles between feature embeddings and classifiers or by increasing the norms of feature embeddings. We observe two phases popping up during training: (1) Phase 1, where the softmax improvement is primarily due to angular alignment, and later, (2) Phase 2, where the improvement is primarily due to significant increase in feature-embedding norms.

The above findings suggest that the AVH can be a robust universal measure of hardness since angular scores are mostly frozen early in training. In addition, they suggest the need to design better loss functions over softmax loss that can improve performance on hard examples and focus on optimizing angles, e.g., (Liu et al., 2017b; Deng et al., 2019; Wang et al., 2018b; a). We verify that better models tend to have better average AVH scores, which validates the argument in (Recht et al., 2019) that improving on hard examples is the key to improved generalization. We show that AVH has a statistically significant stronger correlation with human selection frequency than widely used confidence measures such as softmax score and embedding norm across several CNN models. This makes AVH a potential proxy of human perceived hardness when such information is not available.

Finally, we empirically show the superiority of AVH with its application to self-training for unsupervised domain adaptation and domain generalization. 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. Our proposed new loss function based on AVH also shows drastic improvement for the task of domain generalization.

2. Related Work

Example hardness measures. An automatic detection of examples that are hard for human vision has numerous applications. (Recht et al., 2019) showed that state-of-the-art models perform better on hard examples. 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 (Bengio et al., 2009) and self-paced learning (Kumar et al., 2010) where being able to detect hard examples is crucial. Measuring sample confidence is also important in partially-supervised problems such as semi-supervised learning (Zhu; Zhou et al., 2012), unsupervised domain adaptation (Chen et al., 2011) and weakly-supervised learning (Tang et al., 2017) due to their under-constrained nature. Sample hardness can also be used to identify implicit distribution imbalance in datasets to ensure fairness and remove societal biases (Buolamwini & Gebru, 2018).

Angular distance in neural networks. (Zhang et al., 2018) uses 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. (Liu et al., 2017c) 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. (Liu et al., 2018b) proposes to decouple the inner product as norm and angle, arguing that norms correspond to intra-class variation, and angles correspond to inter-class semantic difference. However, this work does not perform in-depth studies to prove this conjecture. Recent research (Liu et al., 2018a; Lin et al.,...
Figure 3. Visualization of two Gaussian classification. The left plot is input data from two classes (generated Gaussian samples) on the unit sphere. The middle one is their AVH heatmap, the samples with lighter color are exactly the visually hard examples. Lighter color represents higher values. The right one is the heatmap for $\ell_2$-norm of the feature embeddings, showing no correlation to hard samples.

2020; Liu et al., 2020) comes up with an angle-based hyperspherical energy to characterize the neuron diversity and improve generalization by minimizing this energy.

**Deep model calibration.** Confidence calibration aims to predict probability estimates representative of the true correctness likelihood (Guo et al., 2017). It is well-known that the deep neural networks tend to be mis-calibrated and there has been a rich set of literature trying to solve this problem (Kumar et al., 2018; Guo et al., 2017). While establishing correlation between model confidence and prediction correctness, the connection to human confidence has not been widely studied from a training dynamics perspective.

**Uncertainty estimation.** In uncertainty estimation, two types of uncertainties are often considered: (1) Aleatoric uncertainty which captures noise inherent in the observations; (2) Epistemic uncertainty which accounts for uncertainty in the model due to limited data (Der Kiureghian & Ditlevsen, 2009). The latter is widely modeled by Bayesian inference (Kendall & Gal, 2017) and its approximation with dropout (Gal & Ghahramani, 2016; Gal et al., 2017), but often at the cost of additional computation. The fact that AVH correlates well with human selection frequency indicates its underlying connection to aleatoric uncertainty. This makes it suitable for tasks such as self-training. Yet unlike Bayesian inference, AVH can be naturally computed during regular softmax training, making it convenient to obtain with only one-time training and a drop-in uncertainty measure for most existing neural networks.

### 3. Discoveries in CNN Training Dynamics

**Notation.** Denote $S^n$ as the unit $n$-sphere, i.e., $S^n = \{ x \in \mathbb{R}^{n+1} \| x \|_2 = 1 \}$. Below by $A(\cdot, \cdot)$, we denote the angular distance between two points on $S^n$, i.e., $A(u, v) = \arccos(\frac{u \cdot v}{\| u \| \| v \|})$. Let $x$ be the feature embedding input for the last layer of the classifiers in the pretrained CNNs (e.g., FC-1000 in VGG-19). 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** (Model Confidence). We define Model Confidence on a single sample as the probability score of the true objective class output by the CNN models, $\frac{e^{w_y^T x}}{\sum_{i=1}^{C} e^{w_i^T x}}$.

**Definition 2** (Human Selection Frequency). We define one way to measure human visual hardness on pictures as Human Selection Frequency (HSF). Quantitatively, given $m$ number of human workers in a labeling process described in (Recht et al., 2019), 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 HSF is defined as $\frac{b}{m}$.

**3.1. Proposal and Intuition**  

**Definition 3** (Angular Visual Hardness). The AVH score, for any $(x, y)$, 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.

**Theoretical Foundations of AVH.** There are theoretical supports of AVH from both machine learning and vision science perspectives. On the machine learning side, we have briefly discussed above that AVH is directly related to the angle between feature embedding and the classifier weight of ground truth class. (Soudry et al., 2018) theoretically shows that the logit of ground truth class must diverge to infinity in order to minimize cross-entropy loss to zero under gradient descent. Assuming input feature embeddings have fixed unit-norm, the norm of classifier weight grows to infinity. Similar result is also shown in (Wei & Ma, 2019) where generalization error of a linear classifier is controlled by the output margins normalized with classifier norm. Although the above analyses make certain assumptions, they indicate that norm is a less calibrated variable towards measuring properties of model/data compared to angle. This conclusion is comprehensively validated by our experiments in section 3. On the vision science side, there have been wide studies showing that human vision is highly adapted for extracting structural information (Zhang et al., 2018; Wang et al., 2004), while the angular distance in AVH is precisely
Angular Visual Hardness

Figure 4. The left four plots show the number of Epochs v.s. Average $\ell_2$ norm across ImageNet validation samples which are split into five bins based on HSF information. The middle three plots represent number of Epochs v.s. Average $\text{AVH}(x)$. The right ones present number of Epochs v.s. Model Accuracy. From top to bottom, we use AlexNet, VGG-19, ResNet-50 and DenseNet-121. The shadow for the figures in the first two columns denotes the corresponding standard deviations.

good at capturing such information (Liu et al., 2018b). This also justifies our angular based design as an inductive bias towards measuring human visual hardness.

The AVH score is inspired by the observation from Figure 2 as well as (Liu et al., 2018b) that samples from each class concentrate in a convex cone in the embedding space along with some interesting theoretical results that are discussed above. Naturally, we conjecture AVH, a measure with angle or margin information, could be the useful component of softmax score indicating the input sample hardness. We also perform a simulation providing visual intuition of how AVH instead of feature embedding norms corresponds to visually hard examples on two Gaussians in Figure 7 (simulation details and analyses in Appendix A).

3.2. Observations and Conjecture

Setup. We aim to observe the complete training dynamics of models that are trained from scratch on ImageNet instead of the pretrained models. Therefore, we follow the
Angular Visual Hardness

standard training process of AlexNet (Krizhevsky et al., 2012), VGG-19 (Simonyan & Zisserman, 2014), ResNet-50 (He et al., 2016) and DenseNet-121 (Huang et al., 2017). For consistency, we train all 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 VGG-19 is 0.01 and for DenseNet-121 and ResNet-50 is 0.1. 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 HSF respectively. In Appendix B, we further provide experimental results on different datasets, such as MNIST, CIFAR10/100, and degraded ImageNet with different contrast or noise level, to better validate our proposal. For all the figures in this section, epoch starts from 1.

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.

**Observation 1:** The norm of feature embeddings keeps increasing during training.

First row of Figure 4 present 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 $||x||_2$. These observations are consistent in all models and compatible with (Soudry et al., 2018) although it is more about the norm of the classifier weights. More interestingly, we find that neural networks with shortcuts (e.g., ResNets and DenseNets) tend to make the norm of the images with different HSF the same, while 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.

Middle row of Figure 4 exhibits the change of average AVH for validation samples in 90 epochs of training on three models. The average AVH for AlexNet and VGG-19 decreases sharply at the beginning and then starts to bounce back a little bit before converging. However, the dynamics of the average AVH for DenseNet-121 and ResNet-50 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, 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 and we show a simple illustration in Appendix C. It shows the necessity of designing loss functions that focus on optimizing angles.

**Observation 3:** AVH’s correlation with human selection frequency consistently holds across models throughout the training process.

In Figure 4, we average over validation samples in five HSF 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 HSF. 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 a model’s generalization ability.

From Figure 4, 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 weakest model, and its overall average AVH and average AVH on each of the five bins are higher than those of the other three models. In addition, we have found that when testing Hypothesis 3 for better models, their AVH correlations with HSF are significantly stronger than correlations of Model Confidence. The above observations are aligned with earlier observations from (Recht et al., 2019) that better models also tend to generalize better on samples across different levels of human visual hardness. In addition, AVH is potentially a better measure for the generalization of a pretrained model as shown in (Liu et al., 2017b), the norms of feature embeddings are often related to training data priors such as data imbalance and class granularity (Krizhevsky et al., 2012).
Table 1 presents the spearman rank correlations between HSF and AVH, Model Confidence and L2 Norm of the Embedding in ResNet-50 for different visual hardness bin of samples. Note that 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 HSF from 0 to 0.2.

<table>
<thead>
<tr>
<th></th>
<th>z-score</th>
<th>Total Coef</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>Number of Samples</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AVH</td>
<td>-</td>
<td>29987</td>
<td>837</td>
<td>2732</td>
<td>6541</td>
<td>11066</td>
<td>8811</td>
</tr>
<tr>
<td>Model Confidence</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>$</td>
<td></td>
<td>x</td>
<td></td>
<td>_2$</td>
<td>0.337</td>
<td>0.325</td>
<td>0.192</td>
</tr>
</tbody>
</table>

However, when extracting features from unseen classes that do not exist in the training set, such training data prior is often undesired. Since AVH does not consider such feature embedding norms, it potentially presents a better measure towards the open set generalization of a deep network.

Conjecture on training dynamics of CNNs. From Figure 4 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).

4. Connections to Human Visual Hardness

From Section 3.2, we conjecture that AVH has a stronger correlation with human selection frequency - a reflection of human visual hardness that is related to aleatoric uncertainty. In order to validate this claim, we design statistical testings for the connections between Model Confidence, AVH, $||x||_2$ and HSF. Studying the precise connection or gap between human visual hardness and model uncertainty is usually prohibitive because it is laborious to collect such highly subjective human annotations. In addition, these annotations are application or dataset specific, which significantly reduces the scalability of uncertainty estimation models that are directly supervised by them. This makes yet another motivation to this work since AVH is naturally obtained for free without any confidence supervision. In our case, we only leverage such human annotated visual hardness measure for correlation testing. 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

We use the pre-trained network model to extract the feature embedding $x$ from each validation sample and also provide the class weights $w$ to compute $AVH(x)$. Note that we linearly scale the range of $AVH(x)$ to $[0, 1]$. Table 1 shows the overall consistent and stronger correlation between $AVH(x)$ and HSF ($p$-value is $< 0.001$ rejecting the null hypothesis). From the coefficients shown in different bins of sample hardness, we can see that the harder the sample, the weaker the correlation. Also Note that we validate the results across different CNN architectures and found that better models tend to have higher coefficients.

The plot on the left in Figure 5 indicates the strong correlation between $AVH(x)$ and HSF on validation images. One intuition behind this correlation is that the class weights $W$ might correspond to human perceived semantics for each category and thereby $AVH(x)$ corresponds to human’s se-
Table 2 presents the Spearman rank, Pearson, and Kendall Tau correlations between human selection frequency and AVH, Model Confidence in ResNet50 along with the significance testing between coefficient pairs. Note that p-value < 0.05 represents the result is significant.

<table>
<thead>
<tr>
<th></th>
<th>Coef with AVH</th>
<th>Coef with Model Confidence</th>
<th>$Z_{avh}$</th>
<th>$Z_{mc}$</th>
<th>$Z$ value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spearman</td>
<td>0.360</td>
<td>0.325</td>
<td>0.377</td>
<td>0.337</td>
<td>4.85</td>
<td>&lt;.00001</td>
</tr>
<tr>
<td>Pearson</td>
<td>0.385</td>
<td>0.341</td>
<td>0.406</td>
<td>0.355</td>
<td>6.2</td>
<td>&lt;.00001</td>
</tr>
<tr>
<td>Kendall Tau</td>
<td>0.257</td>
<td>0.231</td>
<td>0.263</td>
<td>0.255</td>
<td>3.38</td>
<td>.0003</td>
</tr>
</tbody>
</table>

Table 2 presents the Spearman rank, Pearson, and Kendall Tau correlations between human selection frequency and AVH, Model Confidence in ResNet50 along with the significance testing between coefficient pairs. Note that p-value < 0.05 represents the result is significant.

There are three steps for testing if two correlation coefficients are significantly different. First step is applying Fisher Z-Transformation to both coefficients. The Fisher Z-Transformation is a way to transform the sampling distribution of the correlation coefficient so that it becomes normally distributed. Therefore, we apply Fisher transformation for each correlation coefficient: $Z$ score for coefficient of AVH becomes 0.377 and that of Model Confidence becomes 0.337. The second step is to compute the $Z$ value of two $Z$ scores. Then we determined the $Z$ value to be 4.85 from the two above-mentioned $Z$ scores and sample sizes. The last step is that find out the p-value according to the $Z$ table. 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 HSF than Model Confidence. In later section 5.1, we also empirically show that such stronger correlation brings cumulative advantages in some applications.

In Table 2, besides the Spearman correlation coefficient, we also show the coefficients of Pearson and Kendall Tau. In addition, in Appendix D, we run the same tests on four different architectures to test whether the same conclusion holds for different models. Our conclusion is that for all the considered models, AVH correlates significantly stronger than Model Confidence, and the correlation is even stronger for better models. This indicates that besides what we have shown in section 3, AVH is also better aligned with human visual hardness which is related to aleatoric uncertainty.

**Hypothesis 4.** $||x||_2$ has a correlation with Human Selection Frequency.

**Outcome: Failure to Reject Null Hypothesis**

(Liu et al., 2018b) 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 HSF. 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 HSF. The right plot in figure 5 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 HSF.

We also provide a detailed discussion on the difference between AVH and Model Confidence in Appendix E.

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

**Outcome: Null Hypothesis Rejected**

There are three steps for testing if two correlation coefficients are significantly different. First step is applying Fisher Z-Transformation to both coefficients. The Fisher Z-Transformation is a way to transform the sampling distribution of the correlation coefficient so that it becomes normally distributed. Therefore, we apply Fisher transformation for each correlation coefficient: $Z$ score for coefficient...
5. Applications

5.1. AVH for Self-training and Domain Adaptation

Unsupervised domain adaptation (Ben-David et al., 2010) presents an important transfer learning problem and deep self-training (Lee, 2013) recently emerged as a powerful framework to this problem (Saito et al., 2017a; Shu et al., 2018; Zou et al., 2018; 2019). Here we show the application of AVH as an improved confidence measure in self-training that could significantly benefit domain adaptation.

Dataset: We conduct experiments on the VisDA-17 (Peng et al., 2017) 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 (Lin et al., 2014) 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 known source labels and improve over CBST with network prediction/pseudo-label estimation. The complete self-training process involves alternative 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. We propose the following definition of angular visual confidence (AVC) to represent the predicted probability of class c:

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

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

\[ \hat{y}_t^{(k)*} = \begin{cases} 1, & \text{if } k = \arg \max_{k} \text{AVC}(k; x_t; w) \geq \lambda_k \\ 0, & \text{otherwise} \end{cases} \]

where \( \lambda_k \) is differently determined by referring to \( \text{AVC}(k; x_t; w) \) ranked at porton 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 (2).

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

**Improved selection:** \( \text{AVC}(k; x_t; w) > \lambda_k \)

**Improved classification:** \( k = \arg \max_{k} \frac{\text{AVC}(k; x_t; w)}{\lambda_k} \)

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 improved confidence measure that better resembles human visual hardness, both aspects are likely to 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 6. 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.

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 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”
we have no assumption on how the unseen domain looks like, the generalization on the unseen domains will mostly depend on the generalizability of the neural network. We use the challenging PACS dataset (Li et al., 2017) which consists of Art painting, Cartoon, Photo and Sketch domains. For each domain, we leave it out as the test set and train our models on rest of the three domains.

Specifically, we train a 10-layer plain CNN with the following AVH-based loss (additional details in Appendix F):

$$L_{AVH} = \sum_i \exp \left( s (\pi - A(x_i, w_k)) \right) \sum_{k=1}^{n} \exp \left( s (\pi - A(x_i, w_k)) \right),$$

where $s$ is hyperparameter that adjusts the scale of the output logits and implicitly controls the optimization difficulty. This hyperparameter is typically set by cross-validation. Experimental results are reported in Table 5. With the proposed new loss which directly has an AVH-based design, a simple CNN is outperforming baseline and recent methods that are based on more complex models. In fact, similar learning objectives have also been shown useful in image recognition (Liu et al., 2017c) and face recognition (Wang et al., 2017; Ranjan et al., 2017), indicating that AVH is generally effective to improve generalization in various tasks.

6. Concluding Remarks

We propose a novel measure for CNN models known as angular visual hardness. Our comprehensive empirical studies show that AVH can serve as an indicator of generalization abilities of neural networks, and improving SOTA accuracy entails improving accuracy on hard examples. AVH also has a significantly stronger correlation with human selection frequency. We empirically show the huge advantage of AVH over Model Confidence in self-training for domain adaptation task and loss function for domain generalization task. AVH can be useful in other applications such as deep metric learning, fairness, knowledge transfer, etc. and we plan to investigate them in the future (discussions in Appendix G).

---

Table 3. Class-wise and mean classification accuracies on VisDA-17.

<table>
<thead>
<tr>
<th>Method</th>
<th>Source</th>
<th>Painting</th>
<th>Cartoon</th>
<th>Photo</th>
<th>Sketch</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBST + AVH</td>
<td>87.4</td>
<td>89.7</td>
<td>88.9</td>
<td>88.4</td>
<td>89.6</td>
<td></td>
</tr>
<tr>
<td>CBST/CRST</td>
<td>87.0</td>
<td>89.7</td>
<td>89.0</td>
<td>88.5</td>
<td>89.4</td>
<td></td>
</tr>
<tr>
<td>CRST (Zou et al., 2019)</td>
<td>87.2</td>
<td>89.7</td>
<td>89.0</td>
<td>88.5</td>
<td>89.4</td>
<td></td>
</tr>
<tr>
<td>CRST (Zou et al., 2019)</td>
<td>88.0</td>
<td>89.7</td>
<td>89.0</td>
<td>88.5</td>
<td>89.4</td>
<td></td>
</tr>
<tr>
<td>Proposed</td>
<td>93.3</td>
<td>92.5</td>
<td>88.0</td>
<td>88.5</td>
<td>89.5</td>
<td>81.5</td>
</tr>
</tbody>
</table>

Figure 6. Adaptation accuracy vs. epoch for different comparing methods on VisDA-17.

Table 4. 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</th>
<th>Norm $|x|$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBST+AVH</td>
<td>0.844</td>
<td>0.811</td>
<td>0.901</td>
<td>20.84</td>
</tr>
<tr>
<td>CBST/CRST</td>
<td>0.848</td>
<td>0.811</td>
<td>0.976</td>
<td>21.28</td>
</tr>
</tbody>
</table>

Table 5. Domain generalization accuracy (%) on PACS dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Painting</th>
<th>Cartoon</th>
<th>Photo</th>
<th>Sketch</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet (Li et al., 2017)</td>
<td>62.86</td>
<td>66.97</td>
<td>89.50</td>
<td>57.51</td>
<td>69.21</td>
</tr>
<tr>
<td>MLDG (Li et al., 2018)</td>
<td>66.23</td>
<td>66.88</td>
<td>88.00</td>
<td>58.96</td>
<td>70.01</td>
</tr>
<tr>
<td>MetaReg (Balaji et al., 2018)</td>
<td>69.82</td>
<td>70.35</td>
<td>91.07</td>
<td>59.26</td>
<td>72.62</td>
</tr>
<tr>
<td>Feature-rect (Li et al., 2019)</td>
<td>64.89</td>
<td>71.72</td>
<td>89.94</td>
<td>61.85</td>
<td>72.10</td>
</tr>
<tr>
<td>Baseline CNN-10</td>
<td>66.46</td>
<td>67.88</td>
<td>89.70</td>
<td>51.72</td>
<td>68.94</td>
</tr>
<tr>
<td>CNN-10 + AVH</td>
<td>72.02</td>
<td>66.42</td>
<td>90.12</td>
<td>61.26</td>
<td>72.46</td>
</tr>
</tbody>
</table>
Acknowledgements

We would like to thank Shiyu Liang, Yue Zhu and Yang Zou for the valuable discussions that enlighten our research. We are also grateful to the anonymous reviewers for their constructive comments that significantly helped to improve our paper. Weiyang Liu is partially supported by Baidu scholarship and NVIDIA GPU grant. This work was supported by NSF-1652131, NSF-BIGDATA 1838177, AFOSR-YIPFA9550-18-1-0152, Amazon Research Award, and ONR BRC grant for Randomized Numerical Linear Algebra.

References


Dekel, R. Human perception in computer vision, 2017.


Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. Maximum classifier discrepancy for unsupervised domain adaptation. 2017b.


Zhu, X. Semi-supervised learning tutorial.
