FUNDAMENTAL LIMITS ON THE ROBUSTNESS OF IMAGE CLASSIFIERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We prove that image classifiers are fundamentally sensitive to small perturbations in their inputs. Specifically, we show that given some image space of n-by-n images, all but a tiny fraction of images in any image class induced over that space can be moved outside that class by adding some perturbation whose p-norm is $O(n^{1/\max(p,1)})$, as long as that image class takes up at most half of the image space. We then show that $O(n^{1/\max(p,1)})$ is asymptotically optimal. Finally, we show that an increase in the bit depth of the image space leads to a loss in robustness. We supplement our results with a discussion of their implications for vision systems.

1 INTRODUCTION

Image classification, the task of partitioning images into various classes, is a classical problem in computer science with countless practical applications. Progress on this problem has advanced with leaps and bounds since the advent of deep learning, with modern image classifiers attaining some incredible results(Beyer et al., 2020). However, it has been observed that image classes tend to be brittle - classifiers like to partition images in a way such that most images lie very close to images of different classes(Szegedy et al., 2013). Although usually studied in computer vision systems, such phenomena also appear to manifest in natural vision systems(Elsayed et al., 2018; Zhou & Firestone, 2019).

Given these observations, it is natural to ask the following question: is the brittleness of image classes a result of their construction, or does it arise from some fundamental property of image spaces? Previous work demonstrate that classifiers can be made more robust as a function of how they are constructed, and attempts to improve the robustness of existing computer vision systems through such means is an active area of research(Moosavi-Dezfooli et al., 2016; Madry et al., 2017; Ma et al., 2018; Tramer et al., 2020; Machado et al., 2021). However, there is also a fundamental limit to the robustness achievable by *any* classifier that arises as a consequence of the geometry of image spaces.

In this work we show that this fundamental limit of achievable robustness is surprisingly low. Roughly speaking, in most cases it suffices to change the contents of only a few columns of pixels in an image to change its class. Even smaller changes are sufficient when measured using other metrics, such as the Euclidean metric. Our results are a consequence of the geometry of image spaces, and so they apply regardless of the architecture of the classifier. This suggests that there is an inherent brittleness in the semantic content of images, and that robustness as an objective is only desirable with respect to distributions that are concentrated over small subsets of the image space.

1.1 OUR CONTRIBUTIONS AND RELATED WORK

The observation that image classes tend to be brittle was popularized by Szegedy et al. (2013), where it was observed that tiny perturbations suffice to change the image class of many images. This has since opened up a rich field of research on how the brittleness of image classes arise from specific classifier formulations or training distributions (Goodfellow et al., 2014; Gilmer et al., 2018; Tsipras et al., 2018). While these advances offer insights into the deficiencies of our current methodologies, their analyses ultimately depend on some aspect of the architecture or training distribution, so do not rule out the possible existence of ideal classifiers that do not induce brittle image classes.

By contrast, our work provides a non-trivial upper bound to the robustness of *any* image class[1]. Specifically:

- We prove that most images in any image class consisting of at most half the images in an image space of n-by-n images can be moved into a different class by adding a perturbation whose p-norm is $\mathcal{O}(n^{1/\max(p,1)})$. This is a vanishingly small quantity in relation to the average distances in the image space, which is $O(n^{2/\max(p,1)})$, and therefore provides a non-trivial upper bound for the robustness attainable by even an ideal classifier.
- We show that there exist image classes where most images cannot be moved into a different class with any perturbation whose p-norm is $o(n^{1/\max(p,1)})$ (note the small-o notation[2]). Therefore, the bound we derive is asymptotically optimal in n, so proving stronger robustness bounds will *require* examining classifier-specific properties.
- We show that the discretization through lowering the bit depth of the image space permits the existence of more robust image classes. This lends theoretical backing to the idea of using discretization as a method of defending against adversarial attacks(Panda et al., 2019).
- We demonstrate that brittle features in images can deliver semantic content. We argue that a deeper understanding of such features can lead to advances in aligning human and computer vision systems.

To our knowledge, there are two previous works that investigate upper bounds of robustness that arise from the geometry of image spaces. One is from Fawzi et al. (2018), which provides an upper bound for the probability that an image drawn from a given distribution is far from images of a different class. However, our analysis differs and improves on theirs in a few key aspects. Firstly, they only analyze the case where distance is measured using the 2-norm, while we provide bounds for p-norms for any p. Secondly, they do not account for the discrete nature of image spaces with finite bit depth, which allows for classifiers that are more robust than their bounds imply. Finally, their bound is parametrized by a modulus of continuity which differs depending on the image distribution, potentially resulting in trivial bounds for certain distributions. Furthermore, this parameter cannot be computed exactly, so in application their bound is inexact. By contrast, we formulate our results independently of specific image distributions[3]. Our bounds can therefore be computed exactly and unconditionally, and we are able to show the asymptotic optimality of our result.

The other work is from Diochnos et al. (2018), which investigates partitions of bit vectors. They show that given a finite probability of misclassification, most vectors can be turned into a misclassified vector through a small number of bit flips. This result has since been extended to more general spaces in follow-up work(Mahloujifar et al., 2019). However, these results are still dependent on the existence of a finite fraction of misclassified datapoints, and therefore do not preclude the existence of asymptotically infinitesimal image classes that are robust, something which our analysis does preclude. Furthermore our bounds, due to our focus on image spaces, are much stronger than the ones they derive and are asymptotically optimal.

2 Results

2.1 Preliminaries

Images consist of pixels on a two dimensional grid, with each pixel consisting of a set of channels (for example R, G, and B) of varying intensity. We therefore define the *image space of h-channel images of height n and aspect ratio q*, denoted $\mathcal{I}_{n,q,h,\infty}$, as the set of all real valued tensors with shape (qn, n, h) with entries lying in the interval $[0, 1]$. We require that qn be an integer. The first two dimensions index the x and y coordinate of the pixel, while the third indexes the channel.

Only a finite subset of these images can be represented with finite bit strings. Therefore, we use $\mathcal{I}_{n,q,h,b}$ to denote the subset of $\mathcal{I}_{n,h,q,\infty}$ where each entry in the tensor lies in $[0, 1] \cap \{i/(2^b - 1) | i \in$

[1] Consisting of at most half the images in the image space.

[2] $f(n) \in \mathcal{O}(g(n)) \iff \limsup_{n\to\infty} \frac{f(n)}{g(n)} \leq \infty$ and $f(n) \in o(g(n)) \iff \lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$

[3] It is still possible adapt our result to account for image distribution using their techniques. See our discussion for additional details.

$\mathbb{Z}\}$. We will refer to b as the bit depth of the image space. Although technically image spaces of equal height, aspect ratio, and number of channels intersect, we will treat them as disjoint (in other words, if $x \in \mathcal{I}_{n,q,h,b}$, then $x \notin \mathcal{I}_{n',q',h',b'}$ if $\mathcal{I}_{n,q,h,b} \neq \mathcal{I}_{n',q',h',b'}$).

2.1.1 CLASSIFIERS AND CLASSES

A classifier $\mathcal{C}$ is a function $\mathcal{I}_{n,q,h,b} \to \mathcal{Y}$, where $\mathcal{Y}$ is some finite set of labels. For each $y \in \mathcal{Y}$, we define the class of y as the preimage of y, denoted as the set of images $\mathcal{C}^{-1}(y)$. We say that such a class is *induced* by $\mathcal{C}$. If a class takes up a large part of the image space, then it contains a lot of images that look like randomly sampled noise, since randomly sampling channel values from a uniform distribution yields a uniform distribution over the image space. Therefore, many images in these classes tend to be uninteresting, which motivates the following definition:

Definition 1. *A set $C \subseteq \mathcal{I}_{n,q,h,b}$ is an interesting image class if it is not empty, and if it contains no more than half of the total number of images in $\mathcal{I}_{n,q,h,b}$.*

Note that as long as every class of a classifier is populated, no more than one class can be uninteresting since image classes are disjoint. Therefore, if a classifier does induce an uninteresting class, we can think of it as *the* uninteresting class. Intuitively we can think of it as a junk class, since it contains images that look like randomly sampled noise.

2.1.2 PERTURBATIONS AND ROBUSTNESS

In order to discuss perturbations, we define addition and subtraction over tensors that are of the same shape to be element-wise, and we define the p-norm of a tensor A, denoted $\|A\|_p$, to be the pth root of the sum of the absolute values of the entries of A raised to the pth power. p is assumed to be a non-negative integer, and for the special case of $p = 0$ we let $\|A\|_0$ be the number of non-zero entries in A.

We can then define what it means for an image to be robust to perturbations:

Definition 2. *Let $C \subseteq \mathcal{I}_{n,q,h,b}$ be a class of images. We say an image $I \in C$ is robust to L^p-perturbations of size d if for all $I' \in \mathcal{I}_{n,q,h,b}$, $\|I - I'\|_p \leq d$ implies $I' \in C$.*

We can then define what it means for a class to be robust to perturbations. Note that unless a class occupies the entire image space, it must contain some non-robust images, so the best we can hope for is to attain robustness for a large fraction of the images within a class. This is reflected in the following definition.

Definition 3. *Let $C \subseteq \mathcal{I}_{n,q,h,b}$ be a class of images. Then we say that C is r-robust to L^p-perturbations of size d if it is not empty, and the number of images $I \in C$ that are robust to L^p-perturbations of size d is at least $r|C|$, where $|C|$ is the number of images in C.*

2.2 UNIVERSAL UPPER BOUND ON CLASSIFIER ROBUSTNESS

We can now state a universal non-robustness result that applies to all classifiers over discrete image spaces $\mathcal{I}_{n,q,h,b}$.

Theorem 1. *No interesting class $C \subseteq \mathcal{I}_{n,q,h,b}$ is $2e^{-2c^2}$-robust to L^p-perturbations of size $(2 + c\sqrt{qh} * n)^{1/\max(p,1)}$ for all real values $c > 0$.*

Proof sketch. We can use the images in $\mathcal{I}_{n,q,h,b}$ to form a graph where images are the vertices, and images are connected if and only if they differ at exactly one channel. In other words, the image tensors must differ at precisely one entry. Figure 1a illustrates the construction of this graph. Note that graph distance between vertices coincides with the Hamming distance between the images represented by the vertices. Such graphs are known as Hamming graphs, and they have a vertex expansion (or isoperimetry) property (Harper, 1999) which implies that for any sufficiently small set, if we add all vertices that are within a graph distance of $\mathcal{O}(n)$ to that set, then the size of that set increases by at least some given factor (see Figure 1b for an example). This expansion property is contingent on the size of the vertex set being sufficiently small, which is why we require the "interesting class" property.

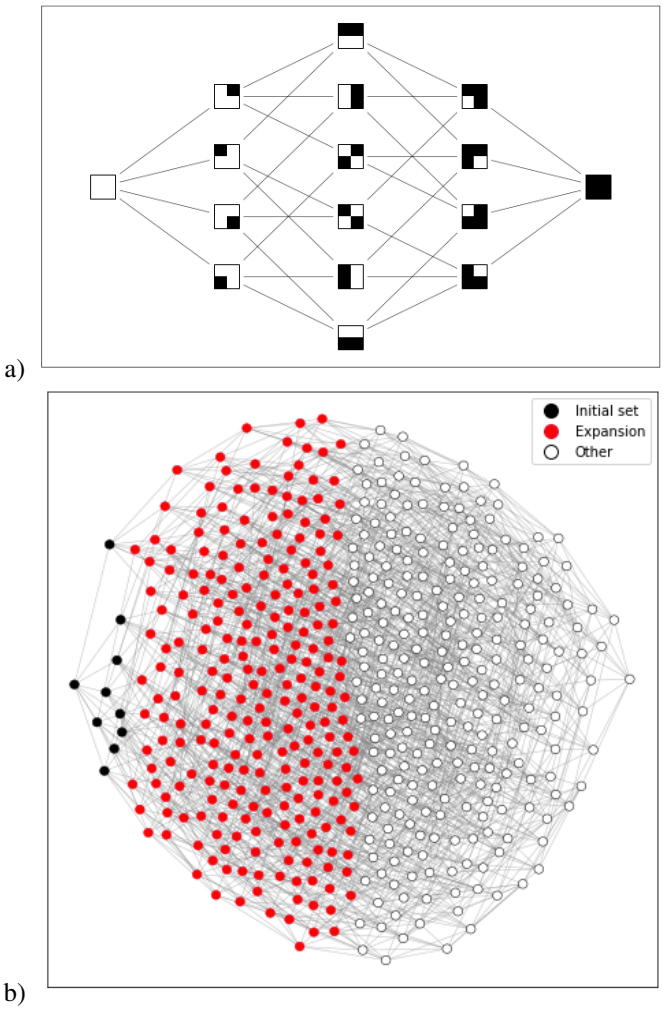

Figure 1: Interpreting image spaces as Hamming graphs
a) We show how we construct a Hamming graph using the elements of $\mathcal{I}_{2,1,1,1}$, the space of binary images on four pixels. By construction, graph distance coincides exactly with Hamming distance.
b) We demonstrate the expansion property of Hamming graphs on a Hamming graph constructed using $\mathcal{I}_{3,1,1,1}$ as the vertex set. If we pick some initial set of vertices (in black), then the set of vertices that are a graph distance of at most 3 (3 being n in this case) from that initial set (in black and red) is much larger than that initial set. The nature of "much larger" is expanded on in Appendix A.1.

We can then show that an interesting class C cannot be too robust in the following way: suppose for contradiction that it is. Then there must be some set $C' \subseteq C$ that is pretty large, and has the property that all vertices within some graph distance of C' are in C. We can then use the vertex expansion property to show that adding these vertices to C' gives a set larger than C, which contradicts the assumption that all vertices within some graph distance C'. Plugging explicit values into this argument yields the statement of the theorem.

We can then generalize to L^p-perturbations for arbitrary p since each coordinate varies by at most 1 unit. The full proof can be found in Appendix A.1. □

Intuitively, the above results state that it is sufficient to change the pixels within a few columns to change the class of most "interesting" images.

2.2.1 THE UNIVERSAL NON-ROBUSTNESS RESULTS ARE ASYMPTOTICALLY OPTIMAL UP TO A CONSTANT FACTOR

Up to a constant factor, the bounds in Theorem 1 are the best possible for a universal non-robustness result that applies to arbitrary predictors if we only consider n-by-n images and hold the aspect ratio q, number of channels per pixel h, and bit depth b constant. In other words, there exists no bound on robustness that applies *universally to all classifiers* that grows much more slowly in n than the ones given in Theorem 1. Therefore, if we wish to show that the classes induced by some classifier are not robust to, for instance, L^0-perturbations of size $\mathcal{O}(log(n))$, more specific properties of that classifier would need to be considered.

To prove this, consider the classifier defined by Algorithm 1.

Algorithm 1: Robust Classifier

Input : An image $I \in \mathcal{I}_{n,q,h,b}$
Result: A label belonging to $\{0, 1\}$
$S \leftarrow 0$;
for $x \leftarrow 1$ **to** qn **do**
  for $y \leftarrow 1$ **to** n **do**
    for $a \leftarrow 1$ **to** h **do**
      $S \leftarrow S + I_{x,y,a}$;
if $S < n^2qh/2$ **then**
  return 0;
else
  return 1;

Theorem 2. *The classifier described by Algorithm 1 induces an interesting class $C \subseteq \mathcal{I}_{n,q,h,b}$ such that for all $c > 0$:*

1. *C is $(1 - 4c)$-robust to L^p-perturbations of size $c\sqrt{qh} * n - 2$ for all $p \leq 1$.*
2. *C is $(1 - 4c)$-robust to L^p-perturbations of size $\frac{(c\sqrt{qh}*n-2)^{1/p}}{(2^b-1)^{(p-1)/p}}$ for all $p \geq 2$.*

Proof sketch. Given an image I, let $S(I)$ be the sum of all its channel values subtracted by $n^2qh/2$. Then I being robust to L^1-perturbations of size x is approximately equivalent to $S(I) \notin [-x, x]$. By the central limit theorem, the fraction of images I such that $S(I) \notin [-cn\sqrt{qh}, cn\sqrt{qh}]$ is some monotonic function of c independent of n and h if n^2qh is sufficiently large, which is our desired result. Appendix A.2 provides a more careful analysis of this that does not rely on limiting behaviour and extends the result to all p-norms. □

2.2.2 CLASSIFIER ROBUSTNESS TO L^p-PERTURBATIONS DECREASES WITH INCREASING BIT DEPTH FOR $p \geq 2$

In this section we investigate the role played by the bit depth b. Theorem 2 has a dependency on b when considering L^p-perturbations for $p \geq 2$, so the statement about attainable robustness become increasingly vacuous as the bit depth increases. Somewhat surprisingly, this is not an artifact of suboptimal analysis: it is really the case that the fundamental limits of robustness drops as a function of the bit depth of the image space.

Theorem 3. *For all real values $c > 0$ and $p \geq 2$, no interesting class $C \in \mathcal{I}_{n,q,h,b}$ is $2e^{-c^2/2}$-robust to L^p-perturbations of size $\left(c + 2\frac{n\sqrt{qh}}{2^b}\right)^{2/p}$.*

Proof sketch. We will focus on the 2-norm. Extension to higher p-norms is straightforward and is given as part of the full proof found in Appendix A.3. The main idea of the proof rests on the fact that if we extend the classifier to the continuous image space with something like a nearest neighbour approach, the measure of the images that are robust to perturbations of a constant size is small (the statement and proof may be found in Appendix A.4). Therefore, if we randomly jump from an image

Table 1: Fundamental bounds for robustness attainable by any interesting image class. Rather than leaving the robustness and bound parametrized by a separate constant c, the bounds have been reparametrized in terms of the robustness r. The bounds are also given in big-Θ notation, where all parameters save for n and b are held constant for simplicity. The upper bound should be understood as "no interesting class is r-robust to perturbations of these sizes" and the lower bound should be interpreted as "there exists an interesting class that is r-robust to perturbations of these sizes".

PERTURBATION	UPPER BOUND	LOWER BOUND
L^0-PERTURBATION L^1-PERTURBATION	$2 + \sqrt{\frac{qh}{2} ln(\frac{2}{r})} * n$	$-2 + (\frac{1-r}{4})\sqrt{qh} * n$
L^p-PERTURBATION, $p \geq 2$	$\min\left((2 + \sqrt{\frac{qh}{2} ln(\frac{2}{r})} * n)^{1/p}, (\sqrt{2ln(\frac{2}{r})} + \frac{2\sqrt{qh}}{2^b} * n)^{2/p} \right)$	$\frac{\left(-2 + (\frac{1-r}{4})\sqrt{qh} * n\right)^{1/p}}{(2^b - 1)^{(p-1)/p}}$

PERTURBATION	UPPER BOUND	LOWER BOUND
L^0-PERTURBATION L^1-PERTURBATION	$\Theta(n)$	$\Theta(n)$
L^p-PERTURBATION, $p \geq 2$	$\Theta(\min(n^{1/p}, (\frac{n}{2^b} + 1)^{2/p})$	$\Theta(\frac{n^{1/p}}{2^{b((p-1)/p)}})$

in the discrete image space to an image in the continuous image space, with high probability we will be within a constant distance of an image of a different class. The size of this random jump can be controlled with a factor that shrinks with increasing bit depth. Summing up the budget required for this jump, the perturbation required on the continuous image space, and the jump back to the discrete image space yields the desired bound. □

We remark that this suggests that the bounds in Theorem 1 pertaining to L^p-perturbations for $p \geq 2$ can be improved to reflect its dependency on the bit depth b. However, whether the component that shrinks with b scales with $n^{1/p}$ rather than $n^{2/p}$ remains an open problem.

3 DISCUSSION

3.1 SUMMARY OF ROBUSTNESS LIMITS AND THEIR RELATION TO AVERAGE IMAGE DISTANCES

We summarize the bounds we derived in the previous section in Table 1, where the bounds are reparametrized in terms of the robustness. An asymptotic bound is also provided with respect to the image height n and bit depth b. With respect to n and b, the bounds derived for the 0-norm and 1-norm are asymptotically optimal, while the bounds for the other p-norms are asymptotically optimal with respect to n. Finding the optimal bound with respect to both b and n for p-norms for $p \geq 2$ remains an open problem, although our current analysis suffices to show that b does fundamentally influence how robust an interesting class can be.

We note that the bounds we derived are vanishingly small when compared to typical distances between random elements of the image space. If a pair of images $I, I' \in \mathcal{I}_{n,q,h,b}$ are sampled independently and uniformly, we have:

$$\mathbb{E}[\|I - I'\|_p] \geq k_{q,h,b,p} n^{2/\max(1,p)} \tag{1}$$

Where $k_{q,h,b,p}$ is some constant parametrized by q, h, b, and p. See Appendix A.5 for additional details. When combined with the bounds in Table 1, we can see that when n is sufficiently large, for 99% (or some arbitrarily high percentage) of images I'' within an interesting class C, there exists some $c_{q,h,b,p}$ parametrized by q, h, b, and p such that:

$$\frac{\min_{X \in \mathcal{I}_{n,q,h,b}, X \notin C} \|I'' - X\|_p}{\mathbb{E}[\|I - I'\|_p]} \leq c_{q,h,b,p} n^{-\frac{1}{\max(p,1)}} \tag{2}$$

The right hand side approaches 0 as n grows without bound, so compared to typical distances one finds in an image space, the distance of an image to an image outside of its class is vanishingly small in any p-norm it is measured in.

3.2 Implications of robustness limits for computer vision systems

According to our results, a large fraction of images in any interesting image class are can have their classes modified by a small perturbation. We note that we only consider the case of image *classifiers* that partition images into a finite number of discrete classes, so our results do not apply directly to vision models that output class probabilities or some more abstract representation. However, these outputs must ultimately be converted into decisions when the model is deployed, at which point our bounds do apply.

Taken at face value, our results appear to pose a barrier for the construction of reliable computer vision systems. For illustration, suppose we implement a system that selects from a finite pool of actions to take depending on the output of some image classifier. Then for most images, a tiny perturbation can make the given image trigger undesired behaviour, ostensibly making the classifier unreliable.

One way of circumventing this barrier is to consider reliability conditioned on a prespecified image distribution. Our bounds do not immediately preclude the existence of small fractions of images within interesting image classes that are robust to large perturbations, and it is possible that those are precisely the set of images that are commonly encountered in deployment. Therefore, our bounds do not directly prevent the construction of classifiers that are robust with respect to some given image distribution, which is an ongoing field of research(Madry et al., 2017).

We remark that our bounds can be converted into bounds that take an image distribution into account through tricks similar to ones used by Fawzi et al. (2018). We can define a map $f : \mathcal{I}_{n,q,h,b} \to \mathcal{I}_{n,q,h,b}$ such that $f(X)$ approximates the desired image distribution if X is distributed uniformly over $\mathcal{I}_{n,q,h,b}$ (we can think of the preimage as some latent space). If we then define a monotonically increasing $\omega : \mathbb{R} \to \mathbb{R}$ such that $\|f(X) - f(X')\|_p \leq \omega(\|X - X'\|_p)$ for all pairs of images, then all our bounds stating that no interesting class is r-robust to L^p-perturbations of size d can be converted into statements about how the probability of encountering an image in an interesting class that is robust to L^p-perturbations of size $\omega(d)$ is less than r. Clearly the value of such a bound necessarily depends on f and ω, and it is easy to construct examples where these bounds must be vacuous - for example, we could make f map all images of one class to the entirely black image and all images of the other class to the entirely white image. However, this is an unrealistic toy example, and experimental analysis carried out by Fawzi et al. (2018) suggest that analyzing plausible approximations of f that produce useful distributions can produce informative bounds.

However, we also note that formulating robustness in a distribution specific way does not address all reliability concerns: for example, an ostensibly benign image modified imperceptibly to trigger dangerous behaviour can be a reliability concern regardless of whether said image is drawn from a prespecified distribution. Furthermore, identifying the correct distribution of such images is a non-trivial task, and distribution gaps can lead to significant reductions in performance(Recht et al., 2019).

Another way of circumventing the barrier to constructing reliable computer vision systems imposed by our bounds is to note that a perturbation with a small p-norm is not necessarily imperceptible. In these cases, the classifier *ought* to adjust its output with respect to such perturbations. Robustness should then be defined with respect to a perception aligned metric, as opposed to with some p-norm. The barrier to constructing reliable computer vision systems then becomes the alignment problem for human and computer vision systems.

3.3 BRITTLE FEATURES CAN IMPART SEMANTICALLY SALIENT INFORMATION

We show in this section that the semantic contents of most images are contained within brittle features that can be erased with a small perturbation. First, we note that our bounds apply universally to any image classifier, so they must apply to an *ideal* classifier that is able to faithfully mimic human classifications. Although such a classifier has yet to be achieved in the computer vision space, human based classifiers can be built quite easily: simply place people in front monitors and ask them to apply labels to images[4]. Memoization could be applied to prevent the same image from being classified multiple times with conflicting labels. This system, on top of producing human classifications, acts like a classifier which partitions the set of all images into disjoint classes (as described in Section 2.1.1), therefore our bounds must be satisfied[5].

At this point, we have proven the existence of a classifier that reproduces human classifications. To simplify the discussion, we will consider classifiers with only two classes: given a set of words[6], we will consider the class of images that such a label can be applied to and its complement. Specifically, the words should be chosen in a way such that random noise should not receive a label with high probability. Roughly speaking, the idea we attempt to capture here is that the first class is the class of "meaningful" images, and the latter is the class of "meaningless" images.

Since images drawn from random noise should not receive a label with high probability, the class of meaningless images is far larger than the class of meaningful images. Therefore, the class of meaningful images is an interesting class, and all our bounds apply. This has the surprising consequence that the semantic content of most images can be "erased" with a small perturbation by turning them into meaningless images. Put differently, the semantic contents of most meaningful images are contained in brittle features that can be erased with a small perturbation, which is what we set out to show.

What do these features look like? A possible example of brittle features that convey semantic information is line drawings (see Figure 2). Line drawings can be erased with a perturbation of size $\mathcal{O}(n)$ when measured using the 1-norm or 0-norm, which is congruent with the bounds we derived for those metrics. Line drawings are also semantically salient. However, the size of a line drawing is generally larger than $\mathcal{O}(1)$ when measured with a p-norm with $p \geq 2$, so the saliency of line drawings do not necessarily account for our bit depth dependent bounds. Phenomena like optical illusions and pareidolia may offer some hints as to how saliency can be present in even more brittle features, but a full understanding remains elusive.

Understanding the ways in which brittle features can carry semantic meaning is not merely of academic curiosity. We have shown that the robustness of computer vision systems have fundamental limits. However, a computer vision system that is aligned to the human visual system *ought* to obey these limits, since a human based classifier must do so as well. Over the past decade we have learned that standard machine learning methodology does not automatically produce vision systems that are aligned to the human visual system with respect to small perturbations (Szegedy et al., 2013), and methodologies that seek to produce such vision systems still contain misalignments (Tramer et al., 2020). A deeper understanding of how semantic content is conveyed in brittle features may inform the development of future methodologies (for example we may wish to explicitly train computer vision systems on such features), which is becoming increasingly necessary as computer vision systems become increasingly deployed in safety and security critical applications, where the trustworthiness of the system is essential (Pereira & Thomas, 2020; Ma et al., 2018).

4 CONCLUSION

We have derived universal non-robustness bounds that apply to any arbitrary image classifier. We have further demonstrated that up to a constant factor, these are the best bounds attainable with respect to the dimensions of the image. These bounds provide fundamental limits to the robustness achievable by computer vision systems, and reveal that most images in any interesting class, even those induced by

[4]Such constructions have been commercially implemented for purposes like content moderation.

[5]It may be the case that a pair of counterfactual trajectories of classifications would yield contradictory labels on certain images. However, since only one such trajectory can occur factually, memoization suffices to make such a classifier observationally indistinguishable from the kind of classifiers discussed in Section 2.1.1.

[6]For example, words like "truck" or "parachute".

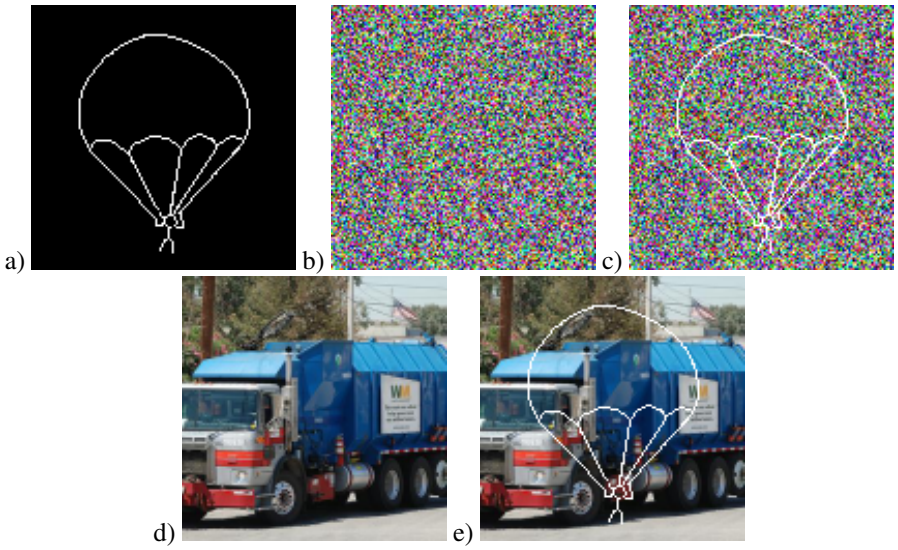

Figure 2: Small perturbations can be semantically salient

A small perturbation (a) when applied to a uniformly randomly drawn image (b) can add meaning to it (c). Conversely the meaning present in (c) can be removed with a small perturbation to attain (b). Such a perturbation can also add information to a natural image[a] (d, e) - although the information present in the natural image largely remains after applying the perturbation, new information was still added by the perturbation.

[a]Sourced from from (Howard)

ideal classifiers, can have their class changed with a perturbation that is asymptotically infinitesimal when compared to the average distance between images. We then discuss the barriers to constructing safe and secure computer vision systems imposed by our results and how these barriers may be circumvented. Finally, we show the abundance of brittle features that convey semantic information, and propose that an improved understanding of these features may yield progress on the problem of aligning human and computer vision systems. We discuss line drawings as an attractive candidate for brittle features that are semantically salient. However, they are not sufficiently brittle to account for all our bounds, so a full understanding of these features remains the subject of future work.

5 Reproducibility Statement

Complete proofs for our claims can be found in the appendix.

References

Franck Barthe and Bernard Maurey. Some remarks on isoperimetry of gaussian type. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 36, pp. 419–434. Elsevier, 2000.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

Dimitrios I Diochnos, Saeed Mahloujifar, and Mohammad Mahmoody. Adversarial risk and robustness: General definitions and implications for the uniform distribution. *arXiv preprint arXiv:1810.12272*, 2018.

Gamaleldin F Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. *arXiv preprint arXiv:1802.08195*, 2018.

Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. *arXiv preprint arXiv:1802.08686*, 2018.

Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

LH Harper. On an isoperimetric problem for hamming graphs. *Discrete applied mathematics*, 95 (1-3):285–309, 1999.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, pp. 409–426. Springer, 1994.

Jeremy Howard. imagenette. URL `https://github.com/fastai/imagenette/`.

Rob Kaas and Jan M Buhrman. Mean, median and mode in binomial distributions. *Statistica Neerlandica*, 34(1):13–18, 1980.

Lei Ma, Felix Juefei-Xu, Minhui Xue, Qiang Hu, Sen Chen, Bo Li, Yang Liu, Jianjun Zhao, Jianxiong Yin, and Simon See. Secure deep learning engineering: A software quality assurance perspective. *arXiv preprint arXiv:1810.04538*, 2018.

Gabriel Resende Machado, Eugênio Silva, and Ronaldo Ribeiro Goldschmidt. Adversarial machine learning in image classification: A survey toward the defender's perspective. *ACM Computing Surveys (CSUR)*, 55(1):1–38, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Saeed Mahloujifar, Dimitrios I Diochnos, and Mohammad Mahmoody. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4536–4543, 2019.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Priyadarshini Panda, Indranil Chakraborty, and Kaushik Roy. Discretization based solutions for secure machine learning against adversarial attacks. *IEEE Access*, 7:70157–70168, 2019.

Ana Pereira and Carsten Thomas. Challenges of machine learning applied to safety-critical cyber-physical systems. *Machine Learning and Knowledge Extraction*, 2(4):579–602, 2020.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

Herbert Robbins. A remark on stirling's formula. *The American mathematical monthly*, 62(1):26–29, 1955.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002.08347*, 2020.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Zhenglong Zhou and Chaz Firestone. Humans can decipher adversarial images. *Nature communications*, 10(1):1–9, 2019.

A APPENDIX

A.1 PROOF OF THEOREM 1

A.1.1 PROPERTIES OF BINOMIAL COEFFICIENTS

We will work with binomial coefficients extensively. To simplify some of our statements, we will extend the definition of a binomial coefficient to work with any $n > 0$ and arbitrary integer k:

$$\binom{n}{k} = \begin{cases} \dfrac{n!}{k!(n-k)!} & \text{if } 0 \leq k \leq n \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Binomial coefficients can be bound in the following way:

Lemma 1. $\binom{n}{k} < \dfrac{2^n}{\sqrt{n}}$ *when* $n \geq 1$.

Proof. We first note that $n!$ is bounded by the following for all $n \geq 1$ (Robbins, 1955):

$$\sqrt{n}\frac{n^n}{e^n} < \frac{n!}{\sqrt{2\pi}} < \sqrt{n}\frac{n^n}{e^n}e^{1/(12n)} \tag{4}$$

Applying the appropriate inequalities for the numerator and denominator yields the following for when n is even:

$$\binom{n}{k} \leq \binom{n}{n/2} = \frac{n!}{((n/2)!)^2} < 2\frac{2^n}{\sqrt{n}}\frac{e^{1/(12n)}}{\sqrt{2\pi}} \tag{5}$$

When n is odd, we have:

$$\begin{aligned} \binom{n}{k} &\leq \binom{n}{\lfloor n/2 \rfloor} && (6) \\ &= \frac{1}{2}\binom{n+1}{(n+1)/2} && (7) \\ &< 2\frac{2^n}{\sqrt{n+1}}\frac{e^{1/(12(n+1))}}{\sqrt{2\pi}} && (8) \\ &< 2\frac{2^n}{\sqrt{n}}\frac{e^{1/(12n)}}{\sqrt{2\pi}} && (9) \end{aligned}$$

Where the third comparison is an application of Equation 5.

If $n \geq 1$, we have $\frac{e^{1/(12n)}}{\sqrt{2\pi}} < 0.5$, which proves the claim. □

It will also be useful to define the following cumulative sums (which are also the tails of binomial distributions):

$$U_{n,p}(k) = \begin{cases} \sum_{i=0}^{k} \binom{n}{i} p^k (1-p)^{n-k} & \text{if } k \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

We can show that the ratio of these cumulative sums are monotonic increasing:

Lemma 2. *Let* $p \in (0, 1)$. *Then* $\frac{U_{n,p}(x-k)}{U_{n,p}(x)}$ *is monotonic increasing in* x, *where* $0 \leq x \leq n$ *and* k *is any positive integer.*

Proof. First, we note that the ratio $\binom{n}{x-k}/\binom{n}{x}$ is monotonic increasing in x when $x \geq 0$. This holds by definition if $x - k < 0$. Otherwise, we have the following:

$$\frac{\binom{n}{x-k}/\binom{n}{x}}{\binom{n}{x-k+1}/\binom{n}{x+1}} = \frac{(n-x)}{(n-x+k)} * \frac{(x-k+1)}{(x+1)} \leq 1 \tag{11}$$

We then claim the following holds for all x where $0 \leq x \leq n-1$:

$$\frac{U_{n,p}(x-k)}{U_{n,p}(x)} \leq \frac{U_{n,p}(x-k+1)}{U_{n,p}(x+1)} \leq \frac{\binom{n}{x-k+1}(1-p)^k}{\binom{n}{x+1}p^k} \tag{12}$$

The above holds with equality when $x - k + 1 < 0$. If $x - k + 1 = 0$, the above also holds: the leftmost ratio is 0. For the other two ratios, if we multiply the rightmost ratio by $(1-p)^{n-k}$ above we can see that the numerators are equal while the denominator of the rightmost ratio is smaller. Otherwise, by induction on x we have:

$$\frac{U_{n,p}(x-k)}{U_{n,p}(x)} \leq \frac{\binom{n}{x-k}(1-p)^k}{\binom{n}{x}p^k} \tag{13}$$

$$\leq \frac{\binom{n}{x-k+1}(1-p)^k}{\binom{n}{x+1}p^k} \tag{14}$$

$$= \frac{\binom{n}{x-k+1}p^{x-k+1}(1-p)^{n-x+k-1}}{\binom{n}{x+1}p^{x+1}(1-p)^{n-x+1}} \tag{15}$$

Where the first inequality follows by induction, and the second inequality follows because $\binom{n}{x-k}/\binom{n}{x}$ is monotonic increasing in x.

For any positive numbers a, c and strictly positive numbers b, d, where $\frac{a}{b} \leq \frac{c}{d}$, we have $\frac{a}{b} \leq \frac{a+c}{b+d} \leq \frac{c}{d}$ because:

$$\frac{d}{d\lambda}\left(\frac{a+\lambda c}{b+\lambda d}\right) = \frac{bc-ad}{(b+\lambda d)^2} \geq 0 \tag{16}$$

Therefore, we have:

$$\frac{U_{n,p}(x-k)}{U_{n,p}(x)}$$

$$\leq \frac{U_n(x-k) + \binom{n}{x-k+1}p^{x-k+1}(1-p)^{n-x+k-1}}{U_{n,p}(x) + \binom{n}{x+1}p^{x+1}(1-p)^{n-x+1}} \tag{17}$$

$$= \frac{U_{n,p}(x-k+1)}{U_{n,p}(x+1)} \tag{18}$$

$$\leq \frac{\binom{n}{x-k+1}(1-p)^k}{\binom{n}{x+1}p^k} \tag{19}$$

As claimed. Carrying on the induction up to $x = n-1$ yields the statement. □

A.1.2 Bounding the interior of a set over a Hamming graph

We will prove our main results by an application of isoperimetry bounds over a Hamming graph. Let Q be a set of q symbols. Then we define the n dimensional Hamming graph over q letters, denoted

$\mathcal{H}(n, q)$, as the graph with a vertex set Q^n and an edge set containing all edges between vertices that differ at precisely one coordinate. For example, $\mathcal{H}(n, 2)$ is isomorphic to the Boolean hypercube. We will use $V(\mathcal{H}(n, q))$ to denote the vertex set of the Hamming graph.

Let $S \subseteq \mathcal{H}(n, q)$. We define the expansion of S, denoted $\textsc{Exp}(S)$, as the set of vertices that are either in S or have a neighbour in S. Since $\textsc{Exp}(.)$ inputs and outputs sets of vertices, we can iterate it. We will use $\textsc{Exp}^k(.)$ to denote k applications of $\textsc{Exp}(.)$.

We now adapt a a result from (Harper, 1999) (Theorem 3 in the paper).

Lemma 3 (Isoperimetric Theorem on Hamming graphs)**.** *Let $S \subsetneq \mathcal{H}(n, q)$. Then:*

$$\frac{|\textsc{Exp}^k(S)|}{|V(\mathcal{H}(n,q))|} \geq \min\{U_{n,p}(r+k) \quad |U_{n,p}(r) = \frac{|S|}{|V(\mathcal{H}(n,q))|}, \quad p \in (0,1), r \in [0, n-k)\} \tag{20}$$

To work with this we first obtain bounds for the expression on the right hand side of Lemma 3.

Lemma 4. *Let p be any value in $(0, 1)$. Let $n > r \geq k$ such that $U_{n,p}(r) \leq \frac{1}{2}$. Then $\frac{U_{n,p}(r-k)}{U_{n,p}(r)} \leq 2e^{-2(k-1)^2/n}$.*

Proof. Let X be a binomially distributed random variable with n trials and probability of success p. Let r be the median of X. We have $r \leq np + 1$ because the median and mean differ by at most 1 (Kaas & Buhrman, 1980).

$U_{n,p}(r-k)$ can be interpreted as $\Pr(X \leq r-k)$, We can then apply Hoeffding's inequality (Hoeffding, 1994):

$$\Pr(X \leq r-k) = \Pr(X \leq np + 1 - k) \tag{21}$$

$$\leq e^{-2(k-1)^2/n} \tag{22}$$

Since r is the median of X, we also have $U_{n,p}(r) \geq \frac{1}{2}$. Combining this with the above equation gives:

$$\frac{U_{n,p}(r-k)}{U_{n,p}(r)} \leq 2e^{-2(k-1)^2/n} \tag{23}$$

Since $\frac{U_{n,p}(x-k)}{U_{n,p}(x)}$ is monotonically increasing via Lemma 2, this also implies that the above relation holds for all smaller r. This completes the proof. □

We can then plug this into Lemma 3 to obtain a non-robustness result on Hamming graphs, which we will then apply to image spaces.

Theorem 4. *Let $S \subsetneq V(\mathcal{H}(n, q))$ such that $|S| \leq |V(\mathcal{H}(n, q))|/2$, and $c > 0$ be any number. Let $S' \subseteq S$ be the set of vertices for which no path with $c\sqrt{n} + 2$ edges or less leads to a vertex not in S. Then $\frac{|S'|}{|S|} < 2e^{-2c^2}$.*

Proof. Suppose for contradiction that $|S'| \geq 2e^{-2c^2}|S|$. Since for any vertex in S' no path with $c\sqrt{n} + 2$ edges or less leads to a vertex outside of S, we have $\textsc{Exp}^{c\sqrt{n}+2}(S') \subseteq S$. Then:

$$
\begin{aligned}
|\text{Exp}^{c\sqrt{n}+2}(S')| \geq & |V(\mathcal{H}(n,q))| \min\{U_{n,p}(r + c\sqrt{n} + 2) \\
& |U_{n,p}(r) = \frac{|S'|}{|V(\mathcal{H}(n,q))|}, \\
& p \in (0,1), r \in [0, n - c\sqrt{n} - 2)\} && (24) \\
\geq & 2e^{2(c\sqrt{n}+1)^2/n}|S'| && (25) \\
> & 2e^{2c^2}|S'| && (26)
\end{aligned}
$$

The first relation follows from Lemma 3 and the second follows from Lemma 4. Lemma 4 applies since $\text{Exp}^{c\sqrt{n}+2}(S') \subseteq S$, so $|\text{Exp}^{c\sqrt{n}+2}(S')| \leq |S| \leq \frac{1}{2}$.

But then $|\text{Exp}^{c\sqrt{n}+2}(S')| > 2e^{-2c^2}|S'| \geq |S|$, which implies that $\text{Exp}^{c\sqrt{n}+2}(S') \nsubseteq S$. This is a contradiction, so we obtain our desired statement. □

A.1.3 Proving Theorem 1

Let $C \subseteq \mathcal{I}_{n,q,h,b}$ be any interesting class.

Lemma 5. *C is not $2e^{-2c^2}$-robust to L^0-perturbations of size $c\sqrt{h} * n + 2$.*

Proof. Let $\mathcal{M} : V(\mathcal{H}(n^2qh, 2^b)) \to \mathcal{I}_{n,q,h,b}$ be the following bijection: first let Q be a set of 2^n equally spaced values between 0 and 1, where the largest value is 0 and the smallest is 1. Then the elements of $V(\mathcal{H}(n^2qh, 2^b))$ can be viewed as Q^{n^2qh}. We then map elements from Q^{n^2qh} to $\mathcal{I}_{n,q,h,b}$ such that the inverse operation is a flattening of the image tensor. Note that such a mapping preserves graph distance on $V(\mathcal{H}(n^2qh, 2^b))$ as Hamming distance on $\mathcal{I}_{n,q,h,b}$.

Let $C' \subseteq C$ be the set of images that are robust to L^0-perturbations of size $c\sqrt{qh} * n + 2$. Let $S = \mathcal{M}^{-1}(C)$ and $S' = \mathcal{M}^{-1}(C')$. S' is then the set of vertices for which no path with $c\sqrt{qh} * n + 2$ edges or less leads to a vertex outside of S.

C is an interesting class and $\mathcal{M}(.)$ preserves cardinality due to it being a bijection. Therefore $|C'| \leq |V(\mathcal{H}(n^2qh, 2^b))|/2$, so by Theorem 4 we have $|S'|/|S| < 2e^{-2c^2}$. Again, since $\mathcal{M}(.)$ preserves cardinality, this implies that $|C'|/|C| < 2e^{-2c^2}$, which means that C is not $2e^{-2c^2}$-robust to L^0-perturbations of size $c\sqrt{qh} * n + 2$. □

We remark that if the domain of $\mathcal{M}(.)$ is changed to $\mathcal{H}(qn^2, h2^b)$, the above argument also shows that C is not $2e^{-2c^2}$-robust to $cn + 2$ pixel changes.

It is straightforward to generalize this to p-norms with larger p.

Lemma 6. *C is not $2e^{-2c^2}$-robust to L^p-perturbations of size $(c\sqrt{qh} * n + 2)^{1/p}$.*

Proof. Let S_1 be the set of images that are r-robust to L^0-perturbations of size d, and let S_2 be the set of images that are r-robust to L^p-perturbations of size $d^{1/p}$.

Suppose $I \notin S_1$. Then there exists some image I' in a different class from I such that $\|I - I'\|_0 \leq d$. Therefore, for all $p > 0$, we have:

$$
\begin{aligned}
d &\geq \|I - I'\|_0 && (27) \\
&= \sum_{x,y,c} \lceil |I_{x,y,c} - I'_{x,y,c}| \rceil && (28) \\
&\geq \sum_{x,y,c} |I_{x,y,c} - I'_{x,y,c}|^p && (29) \\
&= (\|I - I'\|_p)^p && (30)
\end{aligned}
$$

Where the second and third relation follows from the fact that channel values are contained in $[0, 1]$. Therefore, $I \notin S_2$ either since $\|I - I'\|_p \leq d^{1/p}$. Taking the contraposition yields $S_2 \subseteq S_1$.

Setting $d = c\sqrt{qh} * n + 2$ and applying Lemma 5 gives the desired result. □

A.2 Proof of Theorem 2

A.2.1 Anti-concentration inequalities

We first prove an anti-concentration lemma concerning the binomial distribution.

Lemma 7. *Let X be a random variable following the binomial distribution with n trials and a probability of success of 0.5. Let Y be a discrete random variable independent of X whose distribution is symmetric about the origin. Then for any t where $t < \mathbb{E}[X]$ and $t - \lfloor t \rfloor = 1/2$, we have:*

$$Pr(X + Y \leq t) \geq Pr(X < t) \tag{31}$$

Proof. We have the following:

$$\begin{aligned} \Pr(X + Y \leq t) =& \Pr(X + Y \leq t, X < t) && (32)\\ &+ \Pr(X + Y \leq t, X > t) \\ \Pr(X < t) =& \Pr(X + Y \leq t, X < t) && (33)\\ &+ \Pr(X + Y > t, X < t) \end{aligned}$$

Therefore it suffices to show that $\Pr(X + Y \leq t, X > t) \geq \Pr(X + Y > t, X < t)$. We have for any $r \geq 0$:

$$\begin{aligned} \Pr(X + Y \leq t, X = t + r) &= \Pr(Y \leq -r)\Pr(X = t + r) && (34)\\ &\geq \Pr(Y > r)\Pr(X = t + r) && (35)\\ &\geq \Pr(Y > r)\Pr(X = t - r) && (36)\\ &= \Pr(X + Y > t, X = t - r) && (37) \end{aligned}$$

Where Equation 34 follows from the independence of X and Y, Equation 35 follows from the symmetry of the distribution of Y, and Equation 36 follows from our assumption that $t < \mathbb{E}[X]$ and $t - \lfloor t \rfloor = 1/2$.

Summing over all positive r for which $\Pr(X = t + r) \geq 0$ yields the desired result. □

Lemma 8. *Let $X_1, X_2, ..., X_n$ be independently and identically distributed random variables such that each X_i is uniformly distributed on $2k$ evenly spaced real numbers $a = r_1 < r_2 < ... < r_{2k} = b$. Then for $t > 0$, we have:*

$$Pr(\sum_{i=1}^{n} X_i \leq (\sum_{i=1}^{n} \mathbb{E}[X_i]) - t + (b - a)) > \frac{1}{2} - \frac{2t}{\sqrt{n}(b - a)} \tag{38}$$

Proof. Let $Y_1, Y_2, ..., Y_n$ be independently and identically distributed Bernoulli random variables with $p = 0.5$. Let $Z_1, Z_2, ..., Z_n$ be a set of independently and identically distributed random variables uniformly distributed between the integers between 1 and k inclusive. If the Ys and Zs are independent of each other as well, we have:

$$\begin{aligned} \sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) =& \frac{b - a}{2k - 1} \sum_{i=1}^{n}(kY_i + Z_i - \mathbb{E}[kY_i + Z_i]) && (39)\\ =& k\frac{b - a}{2k - 1}\big((\sum_{i=1}^{n} Y_i) + (\sum_{i=1}^{n} \frac{Z_i - \mathbb{E}[Z_i]}{k}) \\ &- (\sum_{i=1}^{n} \mathbb{E}[Y_i])\big) && (40) \end{aligned}$$

Let $\sum_{i=1}^{n} Y_i = B$, $\sum_{i=1}^{n} \frac{Z_i - \mathbb{E}[Z_i]}{k} = D$, and $k\frac{b-a}{2k-1} = c$. Then for any $t > 0$, we have:

$$\begin{aligned}
\Pr(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \leq -t) =& \Pr(B + D \leq -\frac{t}{c} + \mathbb{E}[B]) && (41)\\
\geq& \Pr(B + D \\
&\leq -\frac{t}{c} + \mathbb{E}[B] - u) && (42)\\
\geq& \Pr(B < -\frac{t}{c} + \mathbb{E}[B] - 1) && (43)\\
\geq& \Pr(B - \mathbb{E}[B] \\
&< -\frac{2t}{b-a} - 1) && (44)\\
\geq& \frac{1}{2} - \Pr(B - \mathbb{E}[B] \\
&\in [-\frac{2t}{b-a} - 1, 0]) && (45)\\
\geq& \frac{1}{2} - \binom{n}{\lfloor n/2 \rfloor} 2^{-n} \\
&(\frac{2t}{b-a} + 2) && (46)
\end{aligned}$$

Where $1 \geq u \geq 0$ is chosen such that $-\frac{t}{c} + \mathbb{E}[B] - u$ is the average of two adjacent integers. Equation 43 is then an application of Lemma 7 since B is binomially distributed with $p = 0.5$ and D has a distribution that is symmetric about the origin, and Equation 46 follows from the fact that no more than $x + 1$ values are supported on an interval of length x, and no supported value has probability greater than $\binom{n}{\lfloor n/2 \rfloor} 2^{-n}$.

Observing that $\binom{n}{\lfloor n/2 \rfloor} 2^{-n} < \frac{1}{\sqrt{n}}$ due to Lemma 1 and substituting t with $t - (b - a)$ yields the desired result. □

A.2.2 Proving Theorem 2

Let $A : \mathcal{I}_{n,q,h,b} \to \{0, 1\}$ be described by Algorithm 1. In other words, it is the classifier that inputs an image, sums all of its channels, and outputs 0 if the sum is less than $n^2qh/2$ and 1 otherwise. Let Z be the class of images that A outputs 0 on. Note that Z is an interesting class since it cannot be larger than its complement, so it suffices to prove that Z is robust.

Lemma 9. *Z is $(1 - 4c)$-robust to L^1-perturbations of size $c\sqrt{qh} * n - 2$*

Proof. Let $Z' \subseteq Z$ be the set of images in Z that are robust to L^1-perturbations of size $c\sqrt{qh} * n - 2$. Let I be a random image sampled uniformly. Then $|Z'| = \Pr(I \in Z')2^{-(n^2qh2^b)}$. We then have the following:

$$\begin{aligned}
\Pr(I \in Z') &= \Pr(\sum_{x,y,a} I_{x,y,a} + c\sqrt{qh} * n - 2 < n^2qh/2) && (47)\\
&\geq \Pr(\sum_{x,y,a} I_{x,y,a} \leq n^2qh/2 - c\sqrt{qh} * n + 1) && (48)\\
&> \frac{1}{2} - 2c && (49)
\end{aligned}$$

Where the last inequality follows from Lemma 8 since each channel is sampled from a uniform distribution over a set of 2^b evenly spaced values between 0 and 1. Noting that $|Z| \leq 2^{(n^2qh2^b)-1}$ since it cannot be larger than its complement yields $\frac{|Z'|}{|Z|} \geq 1 - 4c$. Therefore, Z is $(1 - 4c)$-robust to L^1-perturbations of size $c\sqrt{qh} * n - 2$. □

Lemma 10. *Z is $(1-4c)$-robust to L^0-perturbations of size $c\sqrt{qh} * n - 2$*

Proof. It suffices to show that an image that is robust to L^1-perturbations of size d is also robust to L^0-perturbations of size d, since the statement then follows directly from Lemma 9.

Let I be an image that is not robust to L^0-perturbations of size d, so there exists some I' in a different class such that $\|I - I'\|_0 \leq d$. Then:

$$d \geq \|I - I'\|_0 \tag{50}$$

$$= \sum_{(x,y,a)} \lceil |I_{x,y,a} - I'_{x,y,a}| \rceil \tag{51}$$

$$\geq \sum_{(x,y,a)} |I_{x,y,a} - I'_{x,y,a}| \tag{52}$$

$$= \|I - I'\|_1 \tag{53}$$

Where the second and third relations hold since channel values lie in $[0, 1]$.

This implies that I is not robust to L^1-perturbations of size d. Therefore any image that is not robust to L^0-perturbations of size d is also not robust to L^1-perturbations of size d. The contraposition yields the desired statement. □

Lemma 11. *Z is $(1-4c)$-robust to L^p-perturbations of size $\frac{(c\sqrt{qh}*n-2)^{1/p}}{(2^b-1)^{(p-1)/p}}$ for $p \geq 2$.*

Proof. It suffices to show that any image that is robust to L^1-perturbations of size d is also robust to L^p-perturbations of size $\frac{d^{1/p}}{(2^b-1)^{(p-1)/p}}$ for any $p \geq 2$, since the statement then follows directly from Lemma 9.

Let $I \in \mathcal{I}_{n,q,h,b}$ be an image that is robust to L^1-perturbations of size d. Let I' be any image in a different class, so $\|I - I'\|_0 > d$. Then for any $p \geq 1$:

$$\|I - I'\|_p^p = \sum_{(x,y,a)} |I_{x,y,a} - I'_{x,y,a}|^p \tag{54}$$

$$= \sum_{(x,y,a)} \left(\frac{(2^b-1)|I_{x,y,a} - I'_{x,y,a}|}{(2^b-1)}\right)^p \tag{55}$$

$$\geq \sum_{(x,y,a)} (2^b-1)|I_{x,y,a} - I'_{x,y,a}|\left(\frac{1}{(2^b-1)}\right)^p \tag{56}$$

$$= (2^b-1)\frac{\|I - I'\|_1}{(2^b-1)^p} \tag{57}$$

$$> \frac{d}{(2^b-1)^{p-1}} \tag{58}$$

Where the second relation follows from the fact that if two channel values differ, they must differ by at least $\frac{1}{2^b-1}$.

Therefore, $\|I - I'\|_p > \frac{d^{1/p}}{(2^b-1)^{(p-1)/p}}$ for any I' whose class is different from I, so I is robust to L^p-perturbations of size $\frac{d^{1/p}}{(2^b-1)^{(p-1)/p}}$ for $p \geq 2$. □

A.3 Proof of Theorem 3

Let $C \subseteq \mathcal{I}_{n,q,h,b}$ be any interesting class. Our objective is to show that C is not robust to various perturbations.

Let $T = \{[x * 2^{-b}, (x+1) * 2^{-b}) | x \in \mathbb{Z} \cap [0, 2^b - 2]\} \cup \{[1 - 2^{-b}, 1]\}$ be a set of 2^b equal length intervals whose union is the interval $[0, 1]$. Let $\mathcal{D}_{2^b}(n^2qh) = T^{n^2qh}$ be their Cartesian power. Then the elements of $\mathcal{D}_{2^b}(n^2qh)$ are disjoint, and their union is precisely the hypercube $[0, 1]^{n^2qh}$.

We can associate each element of $\mathcal{I}_{n,q,h,b}$ with an element of $\mathcal{D}_{2^b}(n^2qh)$ by first mapping $\mathcal{I}_{n,q,h,b}$ to $[0,1]^{n^2qh}$, which can be done by flattening the image tensor (which we denote by $\flat(I)$ for an image $I \in \mathcal{I}_{n,q,h,b}$). We then map that point to the element of $\mathcal{D}_{2^b}(n^2qh)$ the point falls within. The overall mapping is bijective, and we will denote it by F.

Let $\mathcal{A} : [0,1]^{n^2qh} \times \mathbb{R} \to [0,1]^{n^2qh} \cup \{\perp\}$ be a partial function that maps a point p_1 and a real value c to a point p_2 such that the following hold:

1. $\|p_1 - p_2\|_2 \leq c$.
2. Let $I_1, I_2 \in \mathcal{I}_{n,q,h,b}$ such that $p_1 \in F(I_1)$ and $p_2 \in F(I_2)$. Then we require that $I_1 \in C \iff I_2 \notin C$.

$\mathcal{A}(.)$ returns $\perp$ if and only if no such p_2 exists.

We can then define a procedure FINDPERTURBATION for finding a perturbation given an image I, which is outlined in Algorithm 2.

Algorithm 2: Find Perturbation

Input : An image $I \in C$ and a real value c.
Result: An image $I' \in \mathcal{I}_{n,q,h,b}$ such that $I' \notin C$, or $\perp$.
Sample p_1 from $F(I)$ uniformly at random;
$p_2 \leftarrow \mathcal{A}(p_1, c)$
if $p_2 = \perp$ **then**
  return $\perp$;
else
  Find $I_2 \notin C$ such that $p_2 \in F(I_2)$;
  return I_2;

Our proof strategy is to show that the perturbations found by FINDPERTURBATION are guaranteed to be small, and that the probability of failure is low. This must then imply that most images are not robust.

Lemma 12. *If $I' =$ FINDPERTURBATION(I, c) is not $\perp$, then $\|I - I'\|_2 \leq c + 2\frac{n\sqrt{qh}}{2^b}$.*

Proof. Each element of $\mathcal{D}_{2^b}(n^2qh)$ has a diameter of $\frac{\sqrt{n^2qh}}{2^b}$, thus p_1 differs from $\flat(I)$ be at most that distance. Similarly, p_2 differs from $\flat(I_2) = \flat(I')$ by that distance. We also must have $\|p_1 - p_2\|_2 \leq c$ since $I' \neq \perp$. Putting it altogether with the triangle inequality we get $\|\flat(I) - \flat(I')\|_2 \leq c + 2\frac{n\sqrt{qh}}{2^b}$. Since $\flat(.)$ preserves distances, we get the desired statement. □

Lemma 13. *If I is drawn uniformly from C, then Pr(FINDPERTURBATION$(I, c) = \perp) < 2e^{-c^2/2}$.*

Proof. Let $F(C)$ denote the image of C under F. Let $\bigcup F(C)$ denote the union of all elements in $F(C)$.

If the input I is drawn uniformly from C, then p_1 is distributed uniformly over $\bigcup F(C)$. The procedure fails if and only if $\mathcal{A}(p_1, c) = \perp$, which happens if and only if all elements within a radius of c from p_1 all belong to $\bigcup F(C)$. Let C' denote the set of all such points.

$$\Pr(\mathcal{A}(I_2, c) = \perp) = \frac{\mu(C')}{\mu(\bigcup F(C))} \tag{59}$$

$$< 2e^{-c^2/2} \tag{60}$$

Where $\mu(.)$ denotes the Lebesgue measure.

The last inequality comes from Theorem 5, which is given in the next section. The statement applies for any set S formed from a union of elements of $\mathcal{D}_{2^b}(n^2h)$ whose measure is no larger than $1/2$. $\bigcup F(C)$ satisfies these criteria since C is an interesting class, so we attain the desired statement. □

Lemma 14. *C is not $2e^{-c^2/2}$-robust to L^2-perturbations of size $c + 2\frac{n\sqrt{qh}}{2^b}$.*

Proof. Let I be drawn uniformly from C. Let C_r be the set of images that are robust to L^2-perturbations of size $c + 2\frac{n\sqrt{qh}}{2^b}$.

Let $I' = \text{FINDPERTURBATION}(I, c)$. Then I' is randomly distributed over $\mathcal{I}_{n,q,h,b} \cup \{\perp\}$. By Lemma 12, if $I' \in \mathcal{I}_{n,q,h,b}$, then $\|I - I'\|_2 \leq c + 2\frac{n\sqrt{qh}}{2^b}$, which implies that $I \notin C_r$. By contraposition, $I \in C_r$ implies that $\text{FINDPERTURBATION}(I, c) = \perp$. Therefore:

$$\begin{aligned}
\Pr(I' = \perp) &= \Pr(I \in C_r) + \Pr(I \notin C_r, I' = \perp) && (61)\\
&\geq \Pr(I \in C_r) && (62)\\
&= \frac{|C_r|}{|C|} && (63)
\end{aligned}$$

By Lemma 13, $\Pr(I' = \perp) < 2e^{-c^2/2}$. Thus, $\frac{|C_r|}{|C|} < 2e^{-c^2/2}$, which yields the desired statement. □

Lemma 15. *C is not $2e^{-c^2/2}$-robust to L^p-perturbations of size $\left(c + 2\frac{n\sqrt{qh}}{2^b}\right)^{2/p}$ for $p \geq 2$.*

Proof. We use the identical argument from Lemma 6.

Let S_1 be the set of images that are r-robust to L^2-perturbations of size d, and let S_2 be the set of images that are r-robust to L^p-perturbations of size $d^{2/p}$, where $p \geq 2$.

Suppose $I \notin S_1$. Then there exists some image I' in a different class from I such that $\|I - I'\|_2 \leq d$. Therefore, for all $p > 0$, we have:

$$\begin{aligned}
d^2 &\geq \|I - I'\|_2^2 && (64)\\
&= \sum_{x,y,c} |I_{x,y,c} - I'_{x,y,c}|^2 && (65)\\
&\geq \sum_{x,y,c} |I_{x,y,c} - I'_{x,y,c}|^p && (66)\\
&= (\|I - I'\|_p)^p && (67)
\end{aligned}$$

Where the third relation follows from the fact that channel values are contained in $[0, 1]$. Therefore, $I \notin S_2$ either since $\|I - I'\|_p \leq d^{2/p}$. Taking the contraposition yields $S_2 \subseteq S_1$.

Setting $d = c + 2\frac{n\sqrt{qh}}{2^b}$ and applying Lemma 14 gives the desired result. □

A.4 PROOF OF THEOREM 5

Our objective in this section is to complete the proof of Theorem 3 by proving Theorem 5, stated below. We will use $\mu(.)$ to denote Lebesgue measure throughout this section.

Definition 4. *We say a set $S \subseteq [0, 1]^n$ is a regular set if there is some q and $T \subseteq \mathcal{D}_q(n)$ such that $S = \bigcup_{t \in T} t$.*

Theorem 5. *Let $S \subseteq [0, 1]^n$ be a regular set such that $\mu(S) \leq 1/2$. Let $S_r \subseteq S$ contain all the points in S such that for all $y \in [0, 1]$, $\|x - y\|_2 \leq r \implies y \in S$. Then $\frac{\mu(S_r)}{\mu(S)} < 2e^{c^2/2}$.*

A.4.1 PROPERTIES OF THE STANDARD NORMAL DISTRIBUTION

First, we define the cumulative distribution function for the standard normal distribution and its derivative.

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \tag{68}$$

$$\Phi'(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{69}$$

Similarly to the discrete case, the ratio of the cumulative distribution functions is monotonic increasing.

Lemma 16. $\frac{\Phi(x-k)}{\Phi(x)}$ *is monotonic increasing in* x *for all* $k \geq 0$.

Proof. Let $f(x) = \frac{e^{-x^2/2}}{\int_{-\infty}^{x} e^{-t^2/2} dt}$. Then:

$$\frac{d}{dx} f(x) = \frac{-e^{-x^2/2} x \int_{-\infty}^{x} e^{-t^2/2} dt - e^{-x^2/2} e^{-x^2/2}}{\left(\int_{-\infty}^{x} e^{t^2/2} dt\right)^2} \tag{70}$$

When $x \geq 0$, this derivative is negative since both terms in the numerator are negative. If $x < 0$, we have the following:

$$-x \int_{-\infty}^{x} e^{-t^2/2} dt < -x \int_{-\infty}^{x} e^{-t^2/2} + \frac{1}{t^2} e^{-t^2/2} dt \tag{71}$$

$$= -x\Big(- \frac{1}{t} e^{-t^2/2} \Big|_{-\infty}^{x} \Big) \tag{72}$$

$$= e^{-x^2/2} \tag{73}$$

So the sum is strictly smaller than $(e^{-x^2/2})^2 - (e^{-x^2/2})^2 = 0$. Therefore, the derivative is everywhere negative, so $f(x)$ is strictly decreasing.

Therefore, we have the following for any non-negative k:

$$\frac{d}{dx} ln(\frac{\Phi(x-k)}{\Phi(x)}) = f(x-k) - f(x) \geq 0 \tag{74}$$

Since $ln(.)$ is a monotonic increasing function, $\frac{\Phi(x-k)}{\Phi(x)}$ must also be monotonic increasing. □

A.4.2 Proving Theorem 5

Similarly to the discrete case, our main result relies on an isoperimetry statement, this time on the unit hypercube (Barthe & Maurey, 2000).

Lemma 17 (Isoperimetric Theorem on the Unit Hypercube)**.** *For any* n*, let* $A \subset [0,1]^n$ *be a Borel set. Let* $A_\epsilon = \{x \in [0,1]^n \big| \exists x' \in A : \|x - x'\| \leq \epsilon\}$*. Then we have the following:*

$$\liminf_{\epsilon \to 0^+} \frac{\mu(A_\epsilon) - \mu(A)}{\epsilon} \geq \sqrt{2\pi} \Phi'(\Phi^{-1}(\mu(A))) \tag{75}$$

Let $C \subseteq [0,1]$ be a regular set such that $0 < \mu(C) \leq 1/2$. Let $C_r \subseteq C$ denote the points p_1 in C such that for any point $p_2 \in [0,1]$, $\|p_1 - p_2\|_2 \leq r \implies p_2 \in C$.

Lemma 18. $C_r \leq \Phi(\Phi^{-1}(\mu(C)) - r)$

Proof. Let $z = \Phi^{-1}(\mu(C))$ and let $f(x) = \Phi(x + z)$. Let $v(.)$ be a Lebesgue integrable function such that the following holds:

$$V(r) = \int_{(-\infty,r)} v(t)dt = \begin{cases} \mu(C_{-r}) & \text{if } r \leq 0 \\ \mu(C_0) & \text{otherwise} \end{cases} \tag{76}$$

This exists since C is a regular set. Since $V(x)$ results from integration, it is also a continuous function.

It then suffices to show that $V(x) \leq f(x)$ for all x, since $V(x)$ corresponds to the left hand side of the theorem statement and $f(x)$ corresponds to the right hand side. Suppose this is not the case. We know that $V(x) \leq f(x)$ for all $x \geq 0$, so if this is violated it must happen when $x < 0$. Since $V(x)$ and $f(x)$ are both continuous, by the intermediate value theorem there must exist some interval $[a, b)$ where $V(x) > f(x)$ if $x \in [a, b)$, $V(b) = f(b)$, and $a < b \leq 0$.

This gives us the following:

$$\begin{aligned} V(b) - V(a) &= \int_{[a,b)} v(t)dt && (77) \\ &= \int_{[a,b)\setminus Z} \lim_{\epsilon \to 0^+} \frac{V(t+\epsilon) - V(t)}{\epsilon} dt && (78) \\ &= \int_{[a,b)\setminus Z} \liminf_{\epsilon \to 0^+} \frac{\mu(C_{-t-\epsilon}) - \mu(C_{-t})}{\epsilon} dt && (79) \\ &\geq \int_{[a,b)} \sqrt{2\pi}\Phi'(\Phi^{-1}(\mu(C_{-t})))dt && (80) \\ &\geq \int_{[a,b)} \sqrt{2\pi}\Phi'(\Phi^{-1}(f(t)))dt && (81) \\ &\geq f(b) - f(a) && (82) \end{aligned}$$

Where Z is the set of values where the limit in Equation 78 is not equal to $v(t)$, which by the Lebesgue differentiation theorem is a set of measure 0. Equation 80 is an application of Lemma 17, which is applicable since C_{-t} is a Borel set due to the C being a regular set. Equation 81 follows from the fact that $f(x) \leq V(x)$ for all $x \in [a, b]$ and the fact that $\Phi'(\Phi^{-1}(.))$ is monotonically increasing if the input is no greater than $1/2$.

We also have $V(a) > f(a)$ and $V(b) = f(b)$, so it must be the case that $V(b) - V(a) < f(b) - f(a)$. This contradicts the above, so it must be the case that $V(x) \leq f(x)$ for all x. □

Lemma 19. $\mu(C_c) < 2e^{-c^2/2}\mu(C)$

Proof. Let $z = \Phi^{-1}(\mu(C))$. Then for any $c \geq 0$,

$$\frac{\mu(C_c)}{\mu(C)} \leq \frac{\Phi(z-c)}{\Phi(z)} \leq \frac{\Phi(1/2-c)}{\Phi(1/2)} < 2e^{-c^2/2} \tag{83}$$

Where the first inequality follows from Lemma 18, the second inequality follows from Lemma 16 and the fact that $\mu(C) \leq 1/2$, and the third inequality follows from the Gaussian tail bound $\Phi(x) < e^{-x^2/2}$ for all $x \leq 1/2$. □

A.5 Average distance between images

We wish to show that for a pair of images $I, I' \in \mathcal{I}_{n,q,h,b}$ that are sampled independently and uniformly, there exists a $k_{q,h,b,p}$ such that:

$$\mathbb{E}[\|I - I'\|_p] \geq k_{q,h,b,p} n^{2/\max(1,p)} \tag{84}$$

First, we note that we have:

$$\mathbb{E}[\|I - I'\|_p^{\max(1,p)}] = n^2 qh * \mathbb{E}[|X - Y|^{\max(1,p)}] \tag{85}$$

Where X and Y are independent random variables that are both drawn uniformly from a set of 2^b equally spaced values, where the largest is 1 and the smallest is 0. For simplicity, we denote $\mathbb{E}[|X - Y|^{\max(1,p)}]$ with $k_{b,p}$.

$\|I - I'\|_p^{\max(1,p)}$ is non-negative and cannot be larger than $n^2 qh$. Therefore, the probability that $\|I - I'\|_p^{\max(1,p)} \geq n^2 qhk_{b,p}/2$ is at least $\frac{k_{b,p}}{2-k_{b,p}}$.

Via a monotonicity argument we can deduce that the probability that $\|I - I'\|_p \geq (qhk_{b,p}/2)^{1/\max(p,1)} n^{2/\max(p,1)}$ is at least $\frac{k_{b,p}}{2-k_{b,p}}$ as well. We can then apply Markov's inequality to get the following:

$$\mathbb{E}[\|I - I'\|_p] \geq \frac{k_{b,p}}{2 - k_{b,p}} (qhk_{b,p}/2)^{1/\max(p,1)} n^{2/\max(p,1)} \tag{86}$$

By setting $k_{q,h,b,p}$ to be $\frac{k_{b,p}}{2-k_{b,p}} (hk_{b,p}/2)^{1/\max(p,1)}$ we attain our desired result.