A systematic framework for natural perturbations from videos

Abstract

We introduce a systematic framework for quantifying the robustness of classifiers to naturally occurring perturbations of images found in videos. As part of this framework, we construct `ImageNet-Vid-Robust`*, a human-expert–reviewed dataset of 22,668 images grouped into 1,145 sets of perceptually similar images derived from frames in the ImageNet Video Object Detection dataset. We evaluate a diverse array of classifiers trained on ImageNet, including models trained for robustness, and show a median classification accuracy drop of 16%. Additionally, we evaluate the Faster R-CNN and R-FCN models for detection, and show that natural perturbations induce both classification as well as localization errors, leading to a median drop in detection mAP of 14 points. Our analysis shows that natural perturbations in the real world are heavily problematic for current CNNs, posing a significant challenge to their deployment in safety-critical environments that require reliable, low-latency predictions.*

1. Introduction

Despite their strong performance on various computer vision benchmarks, convolutional neural networks (CNNs) still have many troubling failure modes. At one extreme, ℓ_p-adversarial examples can cause large drops in accuracy for state of the art models with visually imperceptible changes to the input image [5]. But since carefully crafted ℓ_p-perturbations are unlikely to occur naturally in the real world, they usually do not pose a problem outside a fully adversarial context.

To study more realistic failure modes, researchers have investigated benign image perturbations such as rotations & translations, colorspace changes, and various image corruptions [7, 8, 4]. However, it is still unclear whether these perturbations reflect the robustness challenges commonly arising in real data since the perturbations also rely on synthetic image modifications.

Recent work has therefore turned to videos as a source of *naturally occurring* perturbations of images [6, 1]. In contrast to other failure modes, the perturbed images are taken from existing image data without further modifications that make the task more difficult. As a result, robustness to such perturbations directly corresponds to performance improvements on real data.

However, it is currently unclear to what extent such video perturbations pose a significant robustness challenge. Azulay and Weiss [1] only provide anecdotal evidence from a small number of videos. While [6] work with a larger video dataset to obtain accuracy estimates, they only observe a small drop in accuracy of around 2.7% on video-perturbed images, suggesting that small perturbations in videos may not actually reduce the accuracy of current CNNs significantly.

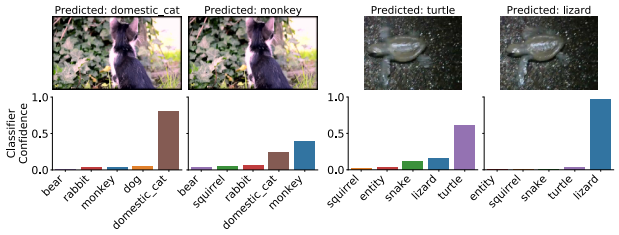

Figure 1: Two examples of natural perturbations from nearby video frames and resulting classifier confidences from a ResNet-152 model fine-tuned on ImageNet Video.

We address this question by conducting a thorough evaluation of robustness to natural perturbations arising in videos. As a cornerstone of our investigation, we introduce `ImageNet-Vid-Robust`, a carefully curated subset of ImageNet-Vid [12]. In contrast to earlier work, all images in `ImageNet-Vid-Robust` were screened by a set of expert labelers to ensure a high annotation quality and to minimize selection biases that arise when filtering with CNNs. Overall, `ImageNet-Vid-Robust` contains 22,668 images grouped into 1,145 sets of temporally adjacent and visually similar images of a total of 30 classes.

We then utilize `ImageNet-Vid-Robust` to measure the accuracy of current CNNs to small, naturally occurring perturbations. Our testbed contains over 40 different model types, varying both architecture and training methodology (adversarial training, data augmentation, etc). We find that natural perturbations from `ImageNet-Vid-Robust` induce a median 16% accuracy drop for classification tasks and a median 14% drop in mAP for detection tasks. Even for the best-performing model, we observe an accuracy drop of 14% – significantly larger than the 2.7% drop in [6] over the same time horizon in the video.

Our results show that robustness to natural perturbations in videos is indeed a significant challenge for current CNNs. As these models are increasingly deployed in safety-critical environments that require both high accuracy and low latency (e.g., autonomous vehicles), ensuring reliable predictions on *every frame* of a video is an important direction for future work.

2. The `ImageNet-Vid-Robust` dataset

The `ImageNet-Vid-Robust` dataset is sourced from videos contained in the ImageNet-Vid dataset [12], we provide more details about ImageNet-Vid in the supplementary.

2.1. Constructing `ImageNet-Vid-Robust`

Next, we describe how we extracted neighboring sets of naturally perturbed frames from ImageNet-Vid to create `ImageNet-Vid-Robust`. A straightforward approach is to select a set of anchor frames and use nearby frames in the video with the assumption that such frames contain only small perturbations from the anchor frame. However, as Figure 1 in the supplementary illustrates, this assumption is frequently broken, especially in the presence of fast camera or object motion.

Instead, we collect a preliminary dataset of natural perturbations and then we manually review each of the frame sets. For each video, we first randomly sample an anchor frame and then take $k = 10$ frames before and after the anchor frame as candidate perturbation images. This results in a dataset containing 1 anchor frame each from 1,314 videos, with approximately 20 candidate perturbation frames each [1].

Next, we curate the dataset with the help of four expert human annotators. The goal of the curation step is to ensure that each anchor frame and nearby frame is correctly labeled with the same ground truth class and that the anchor frame and the nearby frame are visually similar. For each pair of anchor and candidate perturbation frame, an expert human annotator labels (1) whether the pair is correctly labeled in the dataset, (2) whether the pair is similar.

Asking human annotators to label whether a pair of frames is similar can be highly subjective. We took several steps to mitigate this issue and ensure high annotation quality. First, we trained reviewers to mark frames as dissimilar if the scene undergoes any of the following transformations: (1) significant motion, (2) significant background change, or (3) significant blur change, and additionally asked reviewers to mark each of the dissimilar frames with one of these transformations, or "other". Second, as presenting videos or groups of frames to reviewers could cause them to miss potentially large changes due to the well-studied phenomenon of change blindness [9], we present only a single pair of frames at a time to reviewers. Finally, to increase consistency in annotation, human annotators proceed using two rounds of review. In the first round, all annotators were given identical labeling instructions, and then individually reviewed 6500 images pairs. We instructed annotators to err on the side of marking a pair of images as dissimilar if a distinctive feature of the object is only visible in one of the two frames (such as the face of a dog). If an annotator was unsure about a pair he or she could mark the pair as "don't know".

For the second round of review, all annotators jointly reviewed *all* frames marked as dissimilar, "don't know" or "incorrect". A frame was only considered similar if a strict majority of the annotators marked the pair of as "similar".

After the reviewing was complete, we discarded all anchor frames and candidate perturbations that annotators marked as dissimilar or incorrectly labeled. Our final dataset contains 1,145 anchor frames with a minimum of 1, maximum of 20 and median of 20 similar frames.

3. The `pm-k` evaluation metric

Given the dataset above, we would like to measure a model's robustness to natural perturbations. In particular, let $A = \{a_1, ..., a_n\}$ be the set of valid anchor frames in our dataset. Let $Y = \{y_1, ..., y_n\}$ be the set of labels for A. We let $\mathcal{N}_k(a_i)$ be the set of frames marked as similar to anchor frame a_i. In our setting $\mathcal{N}_k$ is a subset of the $2k$ temporally adjacent frames (plus/minus k frames from anchor).

The `pm-k` analogues of the standard metrics for detection and classification evaluate only on the **worst-case** frame in the set of $\mathcal{N}_k$. We formally define the `pm-k` analogues for the standard metrics for classification and detection ($\text{acc}_{\text{pmk}}$and mAP_{pmk}) in the supplementary.

4. Main Results

We evaluate a testbed of 50 classification models and 3 state of the art detection models on `ImageNet-Vid-Robust`. We first discuss the various types of classification models evaluated with `pm-k` classification metric. We then study the per-class accuracies to study whether our perturbations exploits a few "hard" classes or affects performance uniformly across classes.

Second we use the bounding box annotations inherited from ImageNet-VID to study the effect of detection models evaluated on `ImageNet-Vid-Robust` using the `pm-k` metric. We then analyze the errors made on the detection adversarial examples to isolate the effects of *localization* errors vs *classification* errors.

4.1. Classification

In Figure 2, we plot acc_{orig} versus acc_{pmk} for all classification models in our test bed and find that there is a surprisingly linear relationship between acc_{orig} and acc_{pmk} across all 48 models in our test bed. We note the similarity of this plot to Figure 1 in [10].

1578 out 22668 frames in `ImageNet-Vid-Robust` have multiple correct classification labels, due to multiple objects in the frame. To handle this in a classification set-

[1] Note that some anchor frames may have less than 20 candidate frames if the anchor frame is near the start or end of the video.

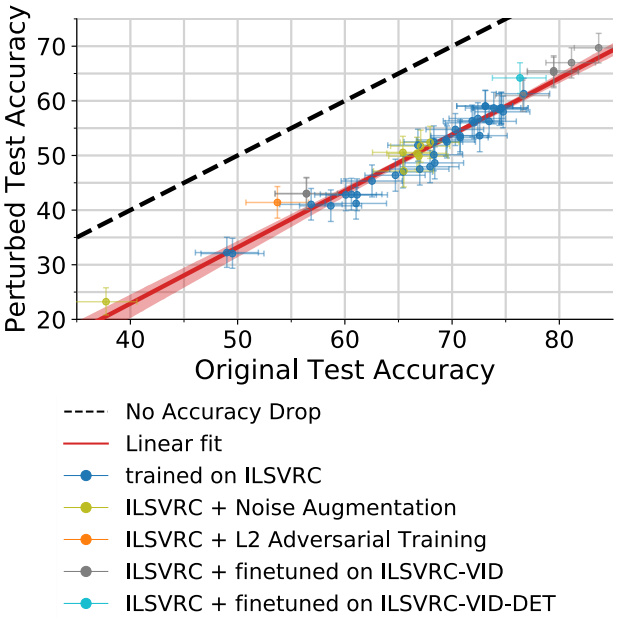

Figure 2: Model accuracy on original vs. perturbed images. Each data point corresponds to one model in our testbed (shown with 95% Clopper-Pearson confidence intervals). Each "perturbed" frame was taken from a neighborhood of a maximum 10 adjacent frames to the original frame in a 30 FPS video. This allows the scene to change for roughly 0.3s. All frames were reviewed by humans to confirm visual similarity to the original frames.

ting, we opt for the most conservative approach: we count a prediction as correct if the model predicts *any* of the classes for a frame. We note that this is a problem that plagues many classification datasets, where objects of multiple classes can be in an image [10] but there is only one true label.

We considered 5 models types of increasing levels of supervision. We present our full table of classification accuracies in the supplementary material, and results for representative models from each model type in Table 1.

ILSVRC Trained As mentioned in **??**, leveraging the WordNet hierarchy enables evaluating models available from [2] trained on the 1000 class ILSVRC challenge on images in `ImageNet-Vid-Robust` directly. We exploit this to evaluate a wide array of model architectures against our natural perturbations. We note that this test set is a substantial distribution shift from the original ILSVRC validation set that these models are tuned for. Thus we will expect the *benign* accuracy acc_{orig} to be lower than the comparable accuracy on the ILSVRC validation set. However the quantity of interest for this experiment is the *difference* between the *original* and *perturbed* accuracies accuracies acc_{orig} - acc_{pmk}, which should be less sensitive to an absolute drop in acc_{orig}.

ILSVRC Trained with Noisy Augmentation One hypothesis for the accuracy drop is that subtle artifacts and corruptions introduced by video compression schemes could introduce a large accuracy drop when evaluated on these corrupted frames. The worst-case nature of the `pm-k` metric could be biasing evaluation towards these corrupt frames. One model for these corruptions are the perturbations introduced in [7]. To test this hypothesis we evaluate models augmented with a subset of the perturbations (Gaussian noise Gaussian blur, shot noise, contrast change, impulse noise, JPEG compression) found in [7]. We found that this augmentation scheme did little to help robustness against our perturbations.

ILSVRC Trained for L_2/L_∞ Robustness We evaluate the best performing robust model against the very strong L_2/L_∞ attacks [14]. We find that this model does have a slightly smaller performance drop than both ILSVRC and ILSVRC trained with noise augmentation but the difference is well within the error bars induced by the small size of our evaluations set. We also note that this robust model gets significantly lower original and perturbed accuracy than examples from either of the model types above.

ILSVRC Trained + Finetuned on ImageNet-VID To adapt to the 30 class problem and the different domain of videos we fine tune several network architectures on the training set in ImageNet VID. We start with a base learning rate of $1e-4$ and train with the SGD optimizer until the validation accuracy plateaus. We trained using cross entropy loss using the *largest* object in the scene as the label during training, as we found this performed better than training using a multi-label loss function. After training for 10 epochs we evaluate on `ImageNet-Vid-Robust`. These models do improve in absolute accuracy over their ILSVRC pretrained counterparts (12% for a ResNet50). However, this improvement in absolute accuracy does not significantly decrease the accuracy drop induced by natural perturbations.

ILSVRC Trained + Finetuned on ImageNet-Vid-Det Finally, we analyze whether additional supervision, in the form of bounding box annotations, improves robustness. To this end, we train the Faster R-CNN *detection* model [11] with a ResNet 50 backbone on ImageNet Vid. Following standard practice, the detection backbone is pre-trained on ILSVRC. To evaluate this detector for classification, we assign the score for each label for an image as the score of the most confident bounding box for that label. We find that this transformation reduces accuracy compared to the

Model Type	Accuracy Original	Accuracy Perturbed	Δ
Trained on ILSVRC	66.8 [64.0, 69.5]	51.9 [48.9, 54.8]	14.9
ILSVRC + Noise Augmentation	66.7 [63.9, 69.5]	50.4 [47.5, 53.3]	16.3
ILSVRC for L_2/L_∞ Robustness (ResNext-101)	53.7 [50.8, 56.6]	41.4 [38.5, 44.3]	12.3
ILSVRC + Finetune ImageNet-VID	79.5 [77.0, 81.8]	65.5 [62.7, 68.3]	14.0
ILSVRC + Finetune ImageNet-VID (ResNet-152)	83.7 [81.4, 85.8]	69.7 [66.9, 72.3]	14.0
ILSVRC + Finetune ImageNet-VID-Det	76.3 [73.8, 78.8]	64.2 [61.3, 67.0]	12.1

Table 1: Accuracies of 5 different model types and best performing model, with ResNet 50 except where otherwise noted. See Section 4.1 for details.

model trained for classification (76.3 vs. 79.5). While there is a slight reduction in the accuracy drop caused by natural perturbations, the reduction is well within the error bars for this task.

4.2. Detection

To analyze the generalizability of natural perturbations to other tasks, we next analyze their impact on the object localization and detection tasks. We report results for two related tasks: object localization and detection. Object detection is the standard computer vision task of correctly classifying an object and regressing the coordinates of a tight bounding box containing the object. "Object localization", meanwhile, refers to the only the subtask of regressing to the bounding box, *without* attempting to correctly classify the object. This is an important problem from a practical perspective (for example, the size and location of an obstacle may be more important for navigation than the category), as well as from an analytical perspective, as it allows analyzing mistakes orthogonal to classification errors. For example, it may be the case that natural perturbations cause misclassification errors frequently, as it may be natural to mistake a cat for a fox, but cause few localization errors.

We present our results using the popular Faster R-CNN [11] and R-FCN [3, 13] architectures for object detection and localization in Table 2. We first note the significant drop in mAP of $12-15\%$ for object detection due to perturbed frames for both the Faster R-CNN and R-FCN architectures. Next, we show that localization is indeed easier than detection, as the mAP increases significantly (e.g., from 61.8 to 75.5 for Faster R-CNN with ResNet 50 backbone). Perhaps surprisingly, however, switching to the localization task does *not* improve the delta between original and perturbed frames, indicating that natural perturbations induce both classification and localization errors. Finally, we show examples of detection failures in Figure 3.

Figure 3: Naturally perturbed examples for detection. Red, green, and white boxes indicate false positives, true positives, and groundtruth, respectively. Classification errors are one of the most common failures, such as the fox on the left, is misclassified as a sheep in the perturbed frame. However, detection models also have *localization* errors, such as the airplane (middle) and the motorcycle (right). All visualizations show predictions with confidence > 0.5.

Model	mAP Original	mAP Perturbed	mAP Δ
FRCNN, R50	61.8	47.8	14.3
FRCNN, R101	62.3	49.8	12.5
R-FCN, R101[13]*	79.0*	63.1*	15.9*
FRCNN, R50 - Loc.	75.5	63.1	12.4
FRCNN, R101 - Loc.	76.8	65.3	11.5
R-FCN, R101- Loc.	80.8*	70.2*	10.6*

Table 2: Detection and localization mAP for two Faster R-CNN backbones. As localization is an easier task, the mAP for localization is higher than for detection. However, both detection and localization suffer from significant drops in mAP due to the perturbations. (*Model trained on ILSVRC Det and VID 2015 datasets, and evaluated on ILSVRC 2015 subset.)

References

[1] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018. 1

[2] Remi Cadene. Pretrained models for pytorch. https://github.com/Cadene/pretrained-models.pytorch. Accessed: 2019-05-20. 3

[3] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016. 4

[4] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017. 1

[5] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

[6] Keren Gu, Brandon Yang, Jiquan Ngiam, Quoc Le, and Jonathan Shlens. Using videos to evaluate image model robustness. *arXiv preprint arXiv:1904.10076*, 2019. 1

[7] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 1, 3

[8] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1614–1619, 2018. 1

[9] Harold Pashler. Familiarity and visual change detection. *Perception & psychophysics*, 44(4):369–378, 1988. 2

[10] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? *arXiv preprint arXiv:1902.10811*, 2019. 2, 3

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 3, 4

[12] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015. 1, 2

[13] Fanyi Xiao and Yong Jae Lee. Video object detection with an aligned spatial-temporal memory. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 485–501, 2018. 4

[14] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. *arXiv preprint arXiv:1812.03411*, 2018. 3

Appendix: A Systematic Framework for Natural Perturbations from Videos

Anonymous Author(s)
Affiliation
Address
email

1 ImageNet-VID Details

The 2015 ImageNet-Vid dataset is widely used for training video object detectors [8] as well as trackers [2]. We chose to work with the 2017 ImageNet-Vid dataset because it is a superset of the 2015 dataset. In total, the 2017 ImageNet-Vid dataset consists of 1,181,113 training frames from 4,000 videos and 512,360 validation frames from 1,314 videos. The videos have frame rates ranging from 9 to 59 frames per second (fps), with a median fps of 29. The videos range from 0.44 to 96 seconds in duration with a median duration of 12 seconds. Each frame is annotated with labels indicating the presence or absence of 30 object categories and corresponding bounding boxes for any label present in the frame.

An advantage of using the ImageNet-Vid dataset as the source of our dataset is that all 30 object categories in the ImageNet-Vid dataset are contained within the WordNet [16] hierarchy, and are ancestors to 288 of the 1000 ILSVRC classes. Using the WordNet hierarchy we construct a canonical mapping from ILSVRC classes to ImageNet-Vid classes, which allows us to evaluate a litany of off-the-shelf ILSVRC-2012 models on ImageNet-Vid.

2 Dissimilar Nearby Frames

Figure 1: Temporally adjacent frames may not be visually similar. We visualize three randomly sampled frame pairs where the nearby frame was marked during human review as "dissimilar" to the anchor frame and discarded from our dataset.

3 `pm-k` Metric details

Classification accuracy is defined as:

$$\text{acc}_{\text{orig}} = 1 - \frac{1}{N}\sum_{i=0}^{N} \mathcal{L}_{0/1}(f(a_i), y_i) \tag{1}$$

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

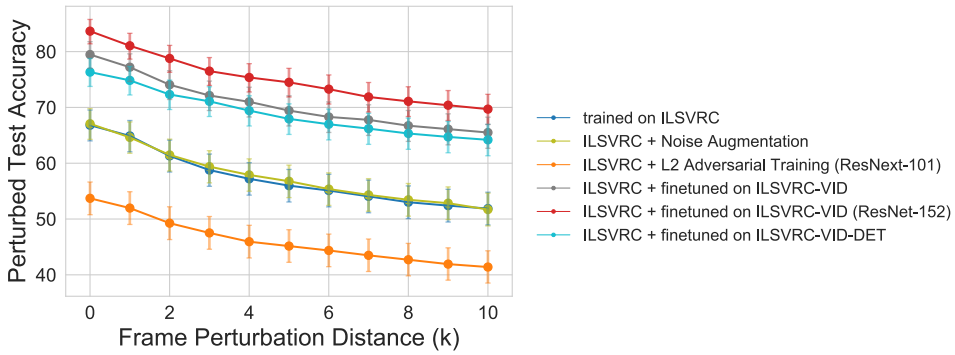

Figure 2: Model classification accuracy on perturbed frames as a function of perturbation distance (shown with 95% Clopper-Pearson confidence intervals). Model accuracies from 5 different model types + best performing model are shown. Architecture is ResNet50 unless otherwise mentioned,

Where $\mathcal{L}_{0/1}$ is the standard 0-1 loss function. We define the `pm-k` analog of misclassification error as:

$$\text{acc}_{\text{pmk}} = 1 - \frac{1}{N}\sum_{i=0}^{N} \max_{b \in \mathcal{N}_k(a_i)} \mathcal{L}_{0/1}(b, y_i) \tag{2}$$

Which simply corresponds to picking the worst frame from the each $\mathcal{N}_k(a_i)$ set before computing misclassification accuracy.

Detection The standard metric for detection is mean average precision of the predictions at a fixed intersection-over-union (IoU) threshold [14]. We briefly introduce the metric here, and refer the reader to [13] for further details.

The standard detection metric proceeds by first determining whether each predicted bounding box in an image is a true or false positive, based on the intersection over union (IoU) of the predicted and ground truth boxes. The metric then computes the per-category Average Precision (AP) of the predictions across all images. The final metric is reported as the mean of these per-category APs (mAP), which we denote $\text{mAP}(\{(f(a_i), y_i)\}_{i=0}^{N})$.

We define the `pm-k` analog of mAP by replacing each anchor frame in the dataset with a nearby frame that minimizes the per-image average precision. Note that as the category-specific average precision is undefined for categories not present in an image, we minimize the average precision across categories for each frame rather than the mAP. We then define the `pm-k` mAP as follows, with a slight abuse of notation to denote y_b as the label for frame b:

$$\text{mAP}_{\text{pmk}}(\{f(a_i), y_i\}_{i=0}^{N}) = \tag{3}$$
$$\text{mAP}\left(\left\{\text{argmin}_{b \in \mathcal{N}(a_i)} AP\left(f(b), y_b\right)\right\}_{i=0}^{N}\right)$$

4 Accuracy vs Frame Perturbation Distance

In Figure 2, we plot the relationship between perturbed accuracy and and perturbation distance (i.e the `k` in the `pm-k` metric described in Section 3). We note that the entire x-axis in Figure 2 corresponds to a temporal distance of 0s to 0.3s between the original and perturbed frames.

5 Per Class Accuracies

We study the effect of our peturbations on the 30 classes found in `ImageNet-Vid-Robust` to determine whethre our performance drop was concentrated in a few "hard" classes. Figure 3 shows a bar plot of original and peturbed accuracies across the 30 classes for our best perfoming model (a

finetuned ResNet152). While this model saw a total drop of 13.7% between original and peturbed it saw a median drop of of 12.1% in per class accuracy across the 30 classes. Though there are a few difficult classes the adversary exploits quite a bit (lion, monkey), we find that the accuracy drop to be generally spread out across most of the 30 classes.

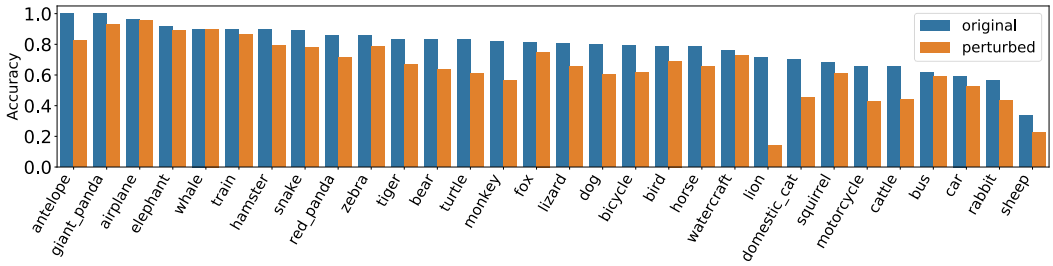

Figure 3: Per class accuracy statistics for our best performing classification model (fine-tuned ResNet152) on `ImageNet-Vid-Robust`,

6 Related Work

Adversarial Attacks. While adversarial examples have been studied in many settings, the majority of researchers focus on L_p robustness. In the ℓ_p adversarial model, the attacker adds a perturbation vector δ such that $\|\delta\|_p < \epsilon$, where ϵ is generally chosen such that the perturbation is imperceptible to humans. Adversarial attacks in the ℓ_p model are powerful and difficult to defend against; for example, the state of the art defenses still achieve mediocre classification accuracies on adversarial inputs (below 97% accuracy on MNIST [4] and below 60% on CIFAR-10 [21]). Motivated by the "artificial" nature of L_p attacks, recent work has proposed more realistic modifications to images. Engstrom et. al. [5] study an adversary that performs minor rotations and translations of the input, Hosseni et. al. [10] allow for hue and color changes, and Hendryks et. al. [9] study common image corruptions such as Gaussian blur and JPEG compression. Researchers have also successfully used generative adversarial networks (GANs) to synthesize more natural adversarial examples [22]. However, even though the above examples are more realistic than the ℓ_p adversarial model, they still synthetically modify the images to generate the perturbations. In contrast, our work performs no synthetic modification and instead uses images that naturally occur in video.

Using Videos To Study Robustness. Weiss and Azulay [1] introduce videos as a failure case of CNNs, and provide qualitative examples where models misclassify adjacent video frames (similar to Figure **??**). In concurrent work to our own, Gu et. al. [7] exploit the temporal structure in videos to study robustness. However, they see a much smaller drop in classification performance (2.7% versus our 16%) when evaluating on worst case neighbor frames. We believe the primary reason for this discrepancy is the underlying difference in evaluation datasets. Gu et. al. evaluate on the YoutubeBB dataset [17], which is constructed by using CNNs to filter YouTube videos. This dataset filtering could introduce selection bias towards videos that are easier to classify by the CNN, possibly resulting in overly optimistic robustness evaluations. In contrast, ImageNet-Vid [19], from which we derive `ImageNet-Vid-Robust`, is constructed through expert review of YouTube videos. In addition, to the best of our knowledge, there was no exhaustive human verification of the adversarial frames in [7], while we use human verification.

Distribution Shift. Small, benign changes in the test distribution are often referred to as *Distribution Shift*. Recht et. al. [18] explore this phenomena by constructing new test sets for CIFAR-10 and ImageNet and observe performance drops for a large suite of models on the newly constructed test sets. However, the images in their test set bear little visual similarity to images in the original test set, while all of our failure cases in `ImageNet-Vid-Robust` are on perceptually similar images.

Computer Vision. The sensitivity of models to small perturbations in videos has been a focus of attention in the computer vision community. A common issue when applying image based models to videos is *flickering*, where object detectors spuriously produce false-positives or false-negatives

in isolated frames or groups of frames. Jin et. al. [11] explicitly identify such failures, and use a technique reminiscent of adversarially robust training to improve image-based models. A similar line of work focuses on improving object detection in videos as objects become occluded or move quickly [12, 6, 23, 20]. The focus in this line of work has generally been on improving object detection when objects transform in a way that makes recognition difficult from a single frame, such as fast motion or occlusion. In this work, we document a broader set of failure cases for image-based classifiers and detectors and show that failures occur when the neighboring frames are imperceptibly different.

7 Discussion

Modern machine learning methods are increasingly put to use in challenging, safety-critical environments. Understanding and measuring the sensitivity of these methods in the real world is crucial for building robust and reliable machine learning systems. Our work presents a *systematic* framework, using a human verified dataset collected from videos, for quantifying a model's sensitivity to *natural perturbations*. Using this framework, we show that these perturbations cause significant drops in accuracy across architectures for both classification and detection. Our work on analyzing this sensitivity opens multiple avenues for future work:

Building Robust Models. Our `ImageNet-Vid-Robust` dataset provides a standard measure for robustness that can be applied to any classification or detection model. In **??**, we evaluated a litany of commonly used models and found that all of them suffer significantly from natural perturbations. In particular, we found that improvements in models with respect to accuracy or with respect to artificial perturbations (such as image corruptions or L_2/L_{inf} adversaries), do *not* translate significant improvements in robustness to natural perturbations. We hope that our standardized dataset and evaluation metric will enable future work to quantify improvements in natural robustness directly.

Other Natural Perturbations. Videos provide a straightforward method for collecting natural perturbations of images, admitting the study of "realistic" forms of robustness for machine learning methods. Other methods for generating these natural perturbations are likely to provide additional insights into model robustness. As an example, photo sharing websites contain a large number of near-duplicate images: pairs of images of the same scene captured at different times, viewpoints or from a different camera [18]. More generally, devising similar, domain-specific strategies to collect, verify and measure robustness to natural perturbations in domains such as natural language processing or speech recognition remains a promising direction for future work.

8 Experimental Details & Hyperparameters

All classification experiments were carried out using PyTorch version 1.0.1 on an AWS p3.2xlarge with the NVIDIA V100 gpu. All pretrained models were downloaded from [3] at commit hash `021d9`. Evaluations in Table 3 all use the default settings for evaluation. The hyperparameters for the *finetuned* models are presented in Table 1. We searched for learning rates between `1e-3` and `1e-5` for all models.

We additionally detail hyperparameters for detection models in Table 2. Detection experiments were conducted with PyTorch version 1.0.1 on a machine with 4 Titan X GPUs, using the Mask R-CNN benchmark repository[15]. We used the default learning rate provided in [15]. For R-FCN, we used the model trained by [20].

Table 1: Hyperparameters for models finetuned on ImageNet-Vid,

Model	Base Learning Rate	Learning Rate Schedule	Batch Size	Epochs
resnet152	1e-4	`Reduce LR On Plateau`	32	10
resnet50	1e-4	`Reduce LR On Plateau`	32	10
alexnet	1e-5	`Reduce LR On Plateau`	32	10
vgg16	1e-5	`Reduce LR On Plateau`	32	10

Table 2: Hyperparameters for detection models.

Model	Base Learning Rate	Learning Rate Schedule	Batch Size	Iterations
F-RCNN ResNet-50	1e-2	`Step 20k, 30k`	8	40k
F-RCNN ResNet-101	1e-2	`Step 20k, 30k`	8	40k

9 Full Original vs Perturbed Accuracy for `ImageNet-Vid-Robust`

Model	Accuracy Original	Accuracy Perturbed	Δ
resnet152_finetune	83.7	69.7	14.0
resnet50_finetune	81.1	67.0	14.1
resnet50_finetune	79.5	65.5	14.0
resnet50_finetune	79.5	65.2	14.2
nasnetalarge	76.7	61.3	15.4
vgg16_finetune	76.6	61.0	15.6
resnet50_detection	76.3	64.2	12.1
inceptionresnetv2	74.8	58.0	16.8
dpn107	74.6	58.7	15.9
dpn107	74.6	58.7	15.9
inceptionv4	74.4	58.4	16.0
dpn98	73.9	58.7	15.2
dpn92	73.4	56.2	17.2
dpn131	73.1	59.0	14.1
dpn131	73.1	59.0	14.1
dpn68b	72.6	53.6	19.0
resnext101	72.4	56.8	15.6
resnext101	72.1	56.0	16.1
resnet152	71.9	56.3	15.5
resnet101	70.7	53.3	17.5
fbresnet152	70.7	53.6	17.1
densenet169	70.3	54.8	p 15.5
densenet169	69.5	52.5	17.0
densenet201	69.4	52.8	16.6
bninception	68.4	48.6	19.7
densenet121	68.3	50.1	18.2
dpn68	68.3	52.5	15.8
nasnetamobile	67.9	47.9	20.0
resnet50_imagenet_augment_jpeg_compression	67.9	52.4	15.5
resnet50_imagenet_augment_gaussian_blur	67.1	51.7	15.4
resnet34	67.0	47.5	19.5
resnet50_imagenet_augment_impulse_noise	66.9	49.9	17.0
resnet50	66.8	51.9	14.9
resnet50_imagenet_augment_gaussian_noise	66.7	50.4	16.3
resnet50_imagenet_augment_defocus_blur	65.5	47.2	18.3
resnet50_imagenet_augment_shot_noise	65.4	50.6	14.8
vgg16_bn	65.4	47.0	18.4
vgg19_bn	64.7	46.4	18.3
vgg19_bn	62.5	45.3	17.2
vgg13_bn	61.1	42.8	18.3
resnet18	61.0	41.2	19.8
vgg16	60.6	42.9	17.7
vgg11	60.1	42.8	17.3
vgg13	58.7	40.8	17.9
vgg11	56.9	41.0	15.8
alexnet_finetune	56.4	43.0	13.4
feature_denoise	53.7	41.4	12.3
squeezenet1	49.5	32.1	17.5
alexnet	49.0	32.2	16.8
resnet50_imagenet_augment_contrast_change	37.7	23.2	14.5

Table 3: Classification model perturbed and original accuracies for all models in our test bed

References

[1] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *arXiv preprint arXiv:1805.12177*, 2018.

[2] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016.

[3] Remi Cadene. Pretrained models for pytorch. https://github.com/Cadene/pretrained-models.pytorch. Accessed: 2019-05-20.

[4] Francesco Croce, Maksym Andriushchenko, and Matthias Hein. Provable robustness of relu networks via maximization of linear regions. *arXiv preprint arXiv:1810.07481*, 2018.

[5] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017.

[6] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3038–3046, 2017.

[7] Keren Gu, Brandon Yang, Jiquan Ngiam, Quoc Le, and Jonathan Shlens. Using videos to evaluate image model robustness. *arXiv preprint arXiv:1904.10076*, 2019.

[8] Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang. Seq-nms for video object detection. *arXiv preprint arXiv:1602.08465*, 2016.

[9] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[10] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1614–1619, 2018.

[11] SouYoung Jin, Aruni RoyChowdhury, Huaizu Jiang, Ashish Singh, Aditya Prasad, Deep Chakraborty, and Erik Learned-Miller. Unsupervised hard example mining from videos for improved object detection. In *ECCV*, 2018.

[12] Kai Kang, Hongsheng Li, Tong Xiao, Wanli Ouyang, Junjie Yan, Xihui Liu, and Xiaogang Wang. Object detection in videos with tubelet proposal networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 727–735, 2017.

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. MS COCO detection evaluation. http://cocodataset.org/#detection-eval. Accessed: 2019-05-16.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[15] Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. https://github.com/facebookresearch/maskrcnn-benchmark, 2018. Accessed: 2019-05-20.

[16] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[17] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5296–5305, 2017.

[18] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? *arXiv preprint arXiv:1902.10811*, 2019.

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015.

[20] Fanyi Xiao and Yong Jae Lee. Video object detection with an aligned spatial-temporal memory. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 485–501, 2018.

[21] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.

[22] Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*, 2017.

[23] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 408–417, 2017.