Robustness in Computer Vision

Sebastian Cygert

Gdansk University of Technology
sebcyg@multimed.org

ML Gdańsk, 21.09.2020

Overview

Surpassing human level accuracy

Source: http://www.image-net.org/papers/ImageNet_2010.pdf

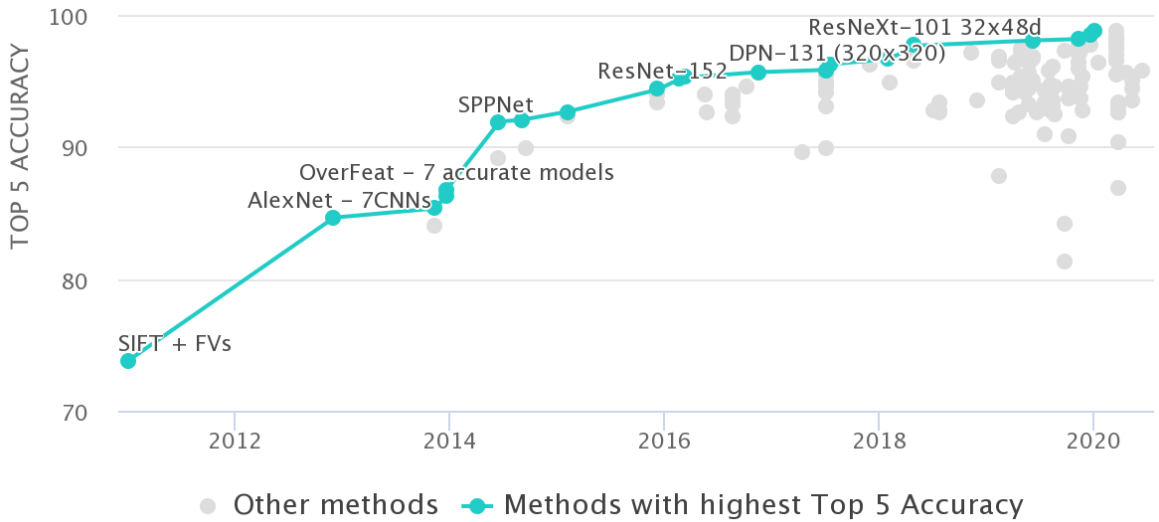


Fig. Progress on ImageNet benchmark. Source: https://paperswithcode.com/sota/image-classification-on-imagenet

Surpassing human level accuracy

Source: http://www.image-net.org/papers/ImageNet_2010.pdf

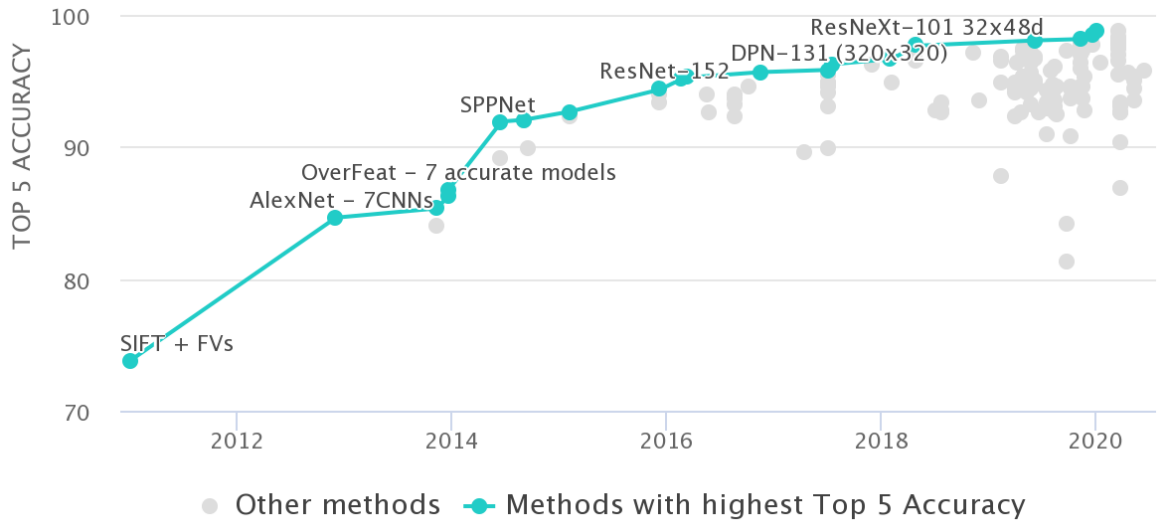


Fig. Progress on ImageNet benchmark. Source: https://paperswithcode.com/sota/image-classification-on-imagenet

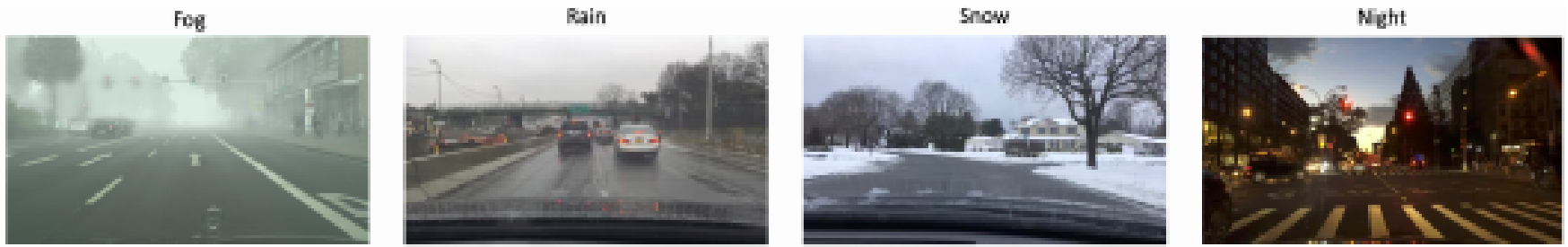
Are we done?

Many benchmarks claims impressive performance.

Is more data all we need?

Safety AI

- Real world is very complicated.
- Does benchmarking impressive performance transfer well?
- How far are we from deploying medical or autonomous driving applications to the real world?
- Is trustable AI possible?

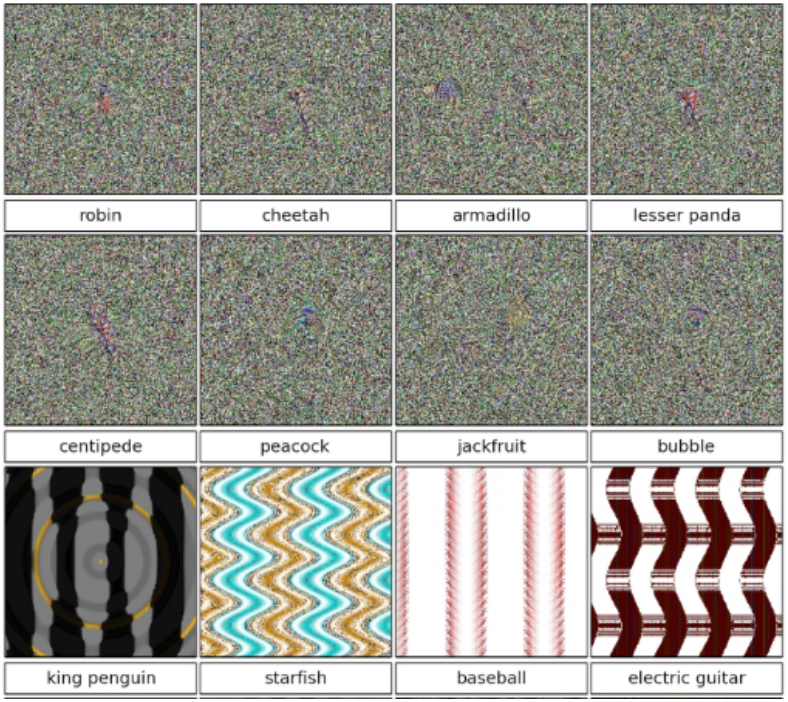


Adversarial examples

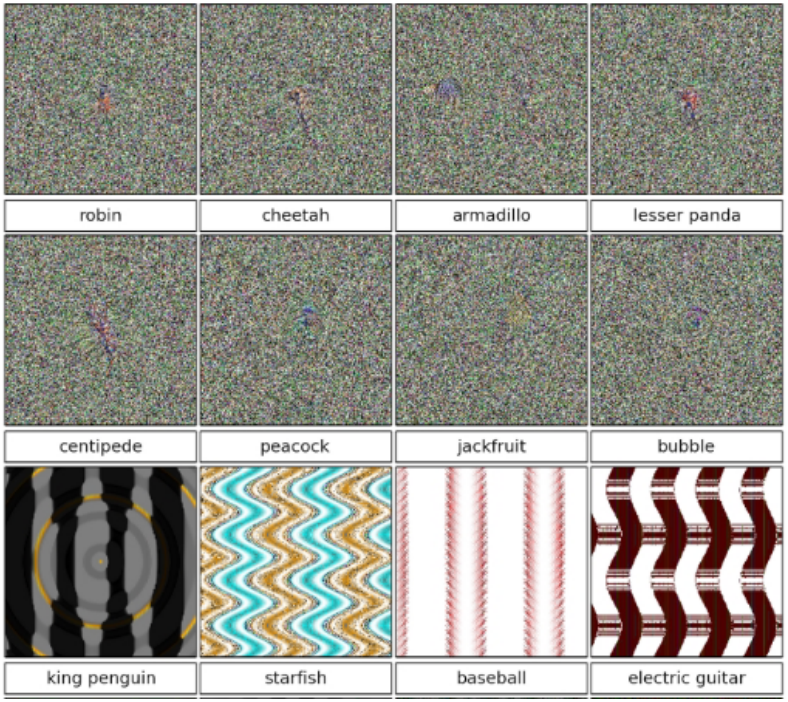
Neural nets are extremely vulnerable to tiny (invisible to the human-eye) changes in the input image [*Szegedy 2014*]. Wrong predictions with **high confidence**.

Fig. (Left) is a correctly predicted sample, (center) difference between correct image, and image predicted incorrectly magnified by 10x, (right) adversarial example.

High Confidence Predictions for Unrecognizable Images, [Nguyen 2015]



High Confidence Predictions for Unrecognizable Images, [Nguyen 2015]



Question

Can we fool the classifier **without** with real world images?

Adversarial examples in the real-world [Hendrycks 2019]

Figure 5: Additional natural adversarial examples from the IMAGENET-A dataset. Examples are adversarially selected to cause classifier accuracy to degrade. The black text is the actual class, and the red text is a ResNet-50 prediction.

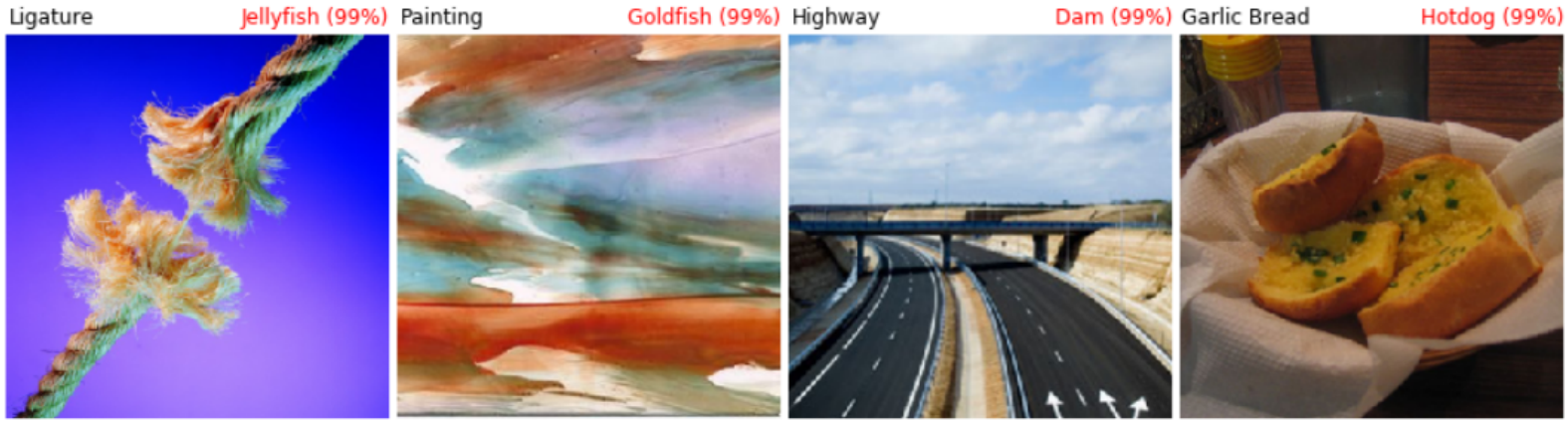
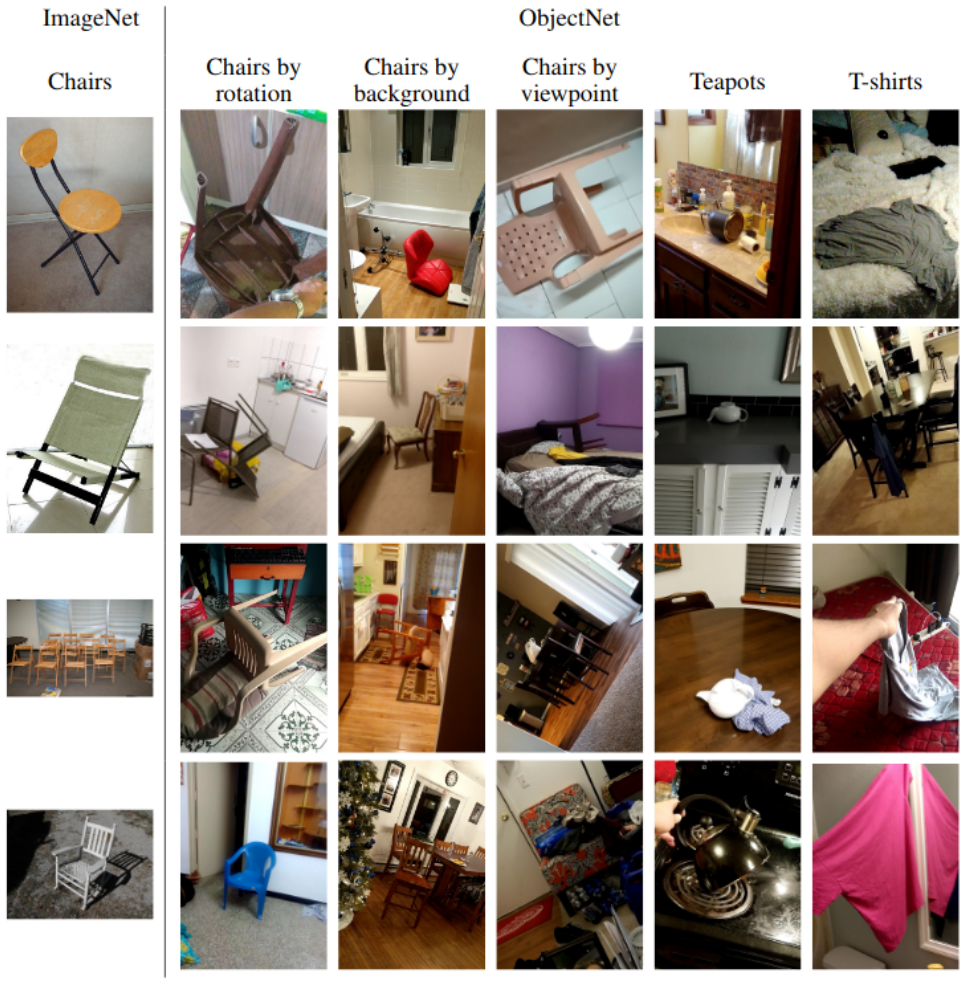


Figure 6: Additional natural adversarial examples from the IMAGENET-O dataset. Examples are adversarially selected to cause out-of-distribution detection performance to degrade. Examples do not belong to ImageNet classes, and they are wrongly assigned highly confident predictions. The black text is the actual class, and the red text is a ResNet-50 prediction and the prediction confidence.

Typical objects in untypical poses, [Barbu 2019]



Typical objects in untypical poses, [Barby 2019]

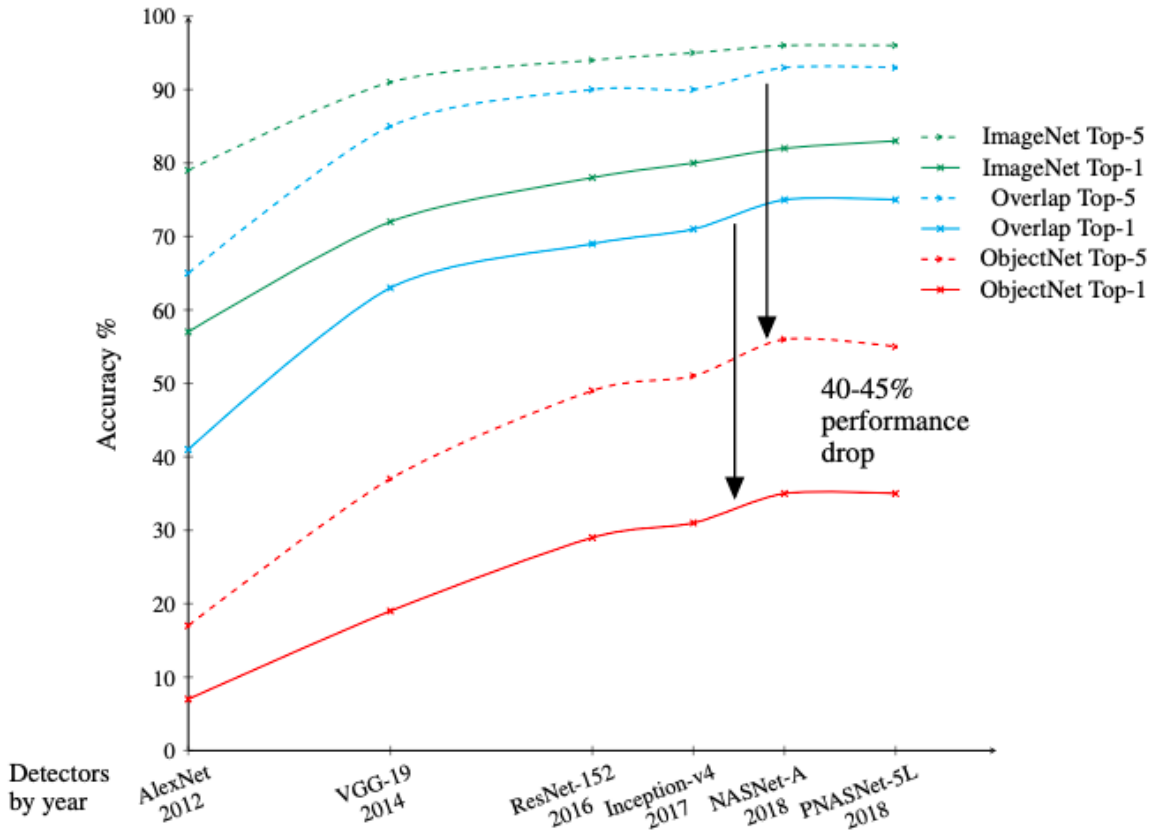


Figure 1: Performance on ObjectNet for high-performing detectors trained on ImageNet in recent years: AlexNet [4], VGG-19 [5], ResNet-152 [6], Inception-v4 [7], NASNET-A [8], and PNASNet-5 Large [9]. Solid lines show top-1 performance, dashed lines show top-5 performance. ImageNet performance on all 1000 classes is shown in green. ImageNet performance on classes that overlap with ObjectNet is shown in blue; the two overlap in 113 classes out of 313 ObjectNet classes, which are only slightly more difficult than the average ImageNet class. Performance on ObjectNet for those overlapping classes. We see a 40-45% drop in performance. Object detectors have improved substantially. Performance on ObjectNet tracks performance on ImageNet but the gap between the two remains large.

Image Captioning, [Lake 2016]

a woman riding a horse on a dirt road

an airplane is parked on the tarmac at an airport

a group of people standing top of a beach

Figure 6: Perceiving scenes without intuitive physics, intuitive psychology, compositionality, an causality. Image captions are generated by a deep neural network (Karpathy & Fei-Fei, 2015) usin code from `github.com/karpathy/neuraltalk2`. Image credits: Gabriel Villena Fernández (left TVBS Taiwan / Agence France-Presse (middle) and AP Photo / Dave Martin (right). Simila examples using images from Reuters news can be found at `twitter.com/interesting_jpg`.

Natural Language Processing, [Ren 2019]

Original Prediction	*Adversarial* Prediction	Perturbed Texts
Positive Confidence = 96.72%	Negative Confidence = 74.78%	Ah man this movie was *funny* (*laughable*) as hell, yet strange. I like how they kept the shakespearian language in this movie, it just felt ironic because of how idiotic the movie really was. this movie has got to be one of troma's best movies. highly recommended for some senseless fun!
Negative Confidence = 72.40%	Positive Confidence = 69.03%	The One and the Only! The only really good description of the punk movement in the LA in the early 80's. Also, the definitive documentary about legendary bands like the Black Flag and the X. Mainstream Americans' repugnant views about this film are absolutely *hilarious* (*uproarious*)! How can music be SO diversive in a country of supposed liberty...even 20 years after... find out!

Table 4: Adversarial example instances in the IMDB dataset with Bi-directional LSTM model. Columns 1 and 2 represent the category prediction and confidence of the classification model for the original sample and the adversarial examples, respectively. In column 3, the green word is the word in the original text, while the red is the substitution in the adversarial example.

Original Prediction	*Adversarial* Prediction	Perturbed Texts
Business Confidence = 91.26%	Sci/Tech Confidence = 33.81%	site security gets a recount at rock the vote. grassroots movement to register younger voters leaves *publishing* (*publication*) tools accessible to outsiders.
Sci/Tech Confidence = 74.25%	World Confidence = 86.66%	seoul allies calm on *nuclear* (*atomic*) shock. south korea's key allies play down a shock admission its scientists experimented to enrich uranium.

Table 5: Adversarial example instances in the AG's News dataset with char-based CNN model. Columns of this table is similar to those in Table 4.

Speech recognition adversarial examples, [Carlini 2019]

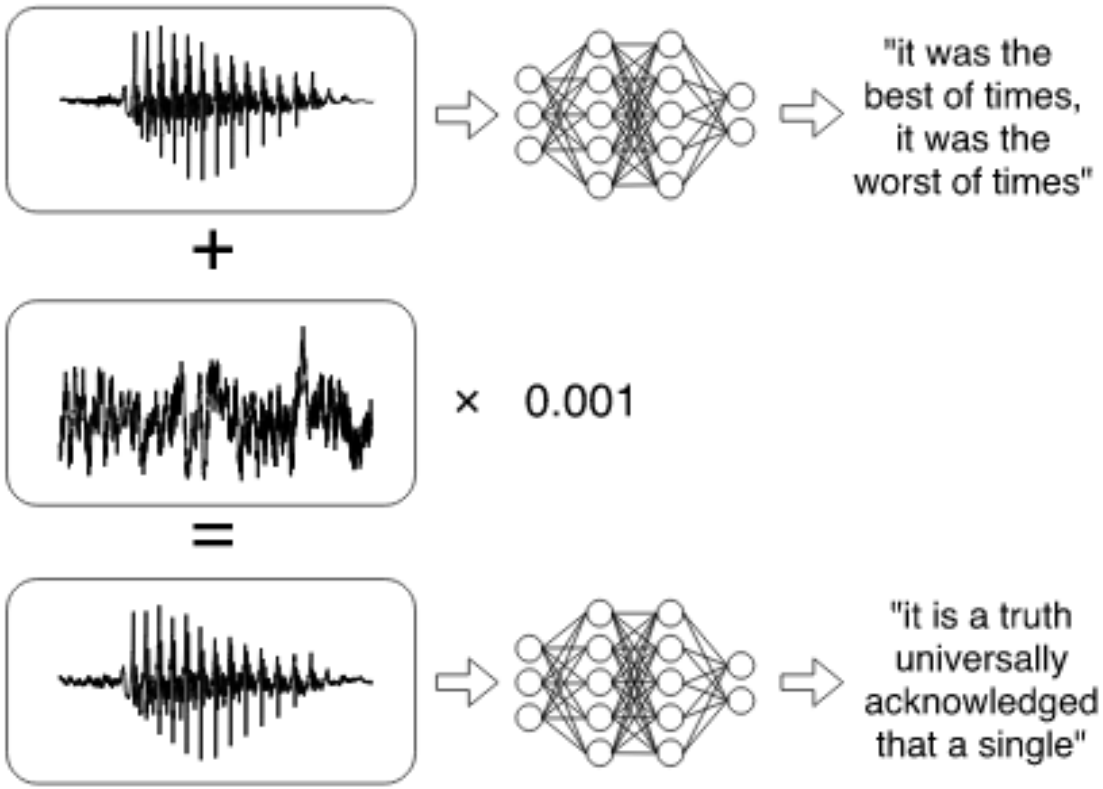


Figure 1. Illustration of our attack: given any waveform, adding a small perturbation makes the result transcribe as any desired target phrase.

Neural nets vulnerability

Summary

- Machine learning models obtain human level performance on many benchmarks, however they are extremely vulnerable to tiny changes in the input distribution
- Machine learning algorithms works under the **i.i.d. assumption** (identically and independently distributed data). When the assumption does not hold (out of distribution setting) the algorithms often provides wrong predictions with high confidence
- Testing only on i.i.d. data may give overoptimistic model performance and a false sense of security.
- This might be OK for many applications (very narrow tasks, entertainment applications) but is not enough for system operating in the open world (**medicine applications, autonomous driving**)
- In order to build safe AI we need to understand WHY neural nets lack robustness and HOW we can improve.

Overview

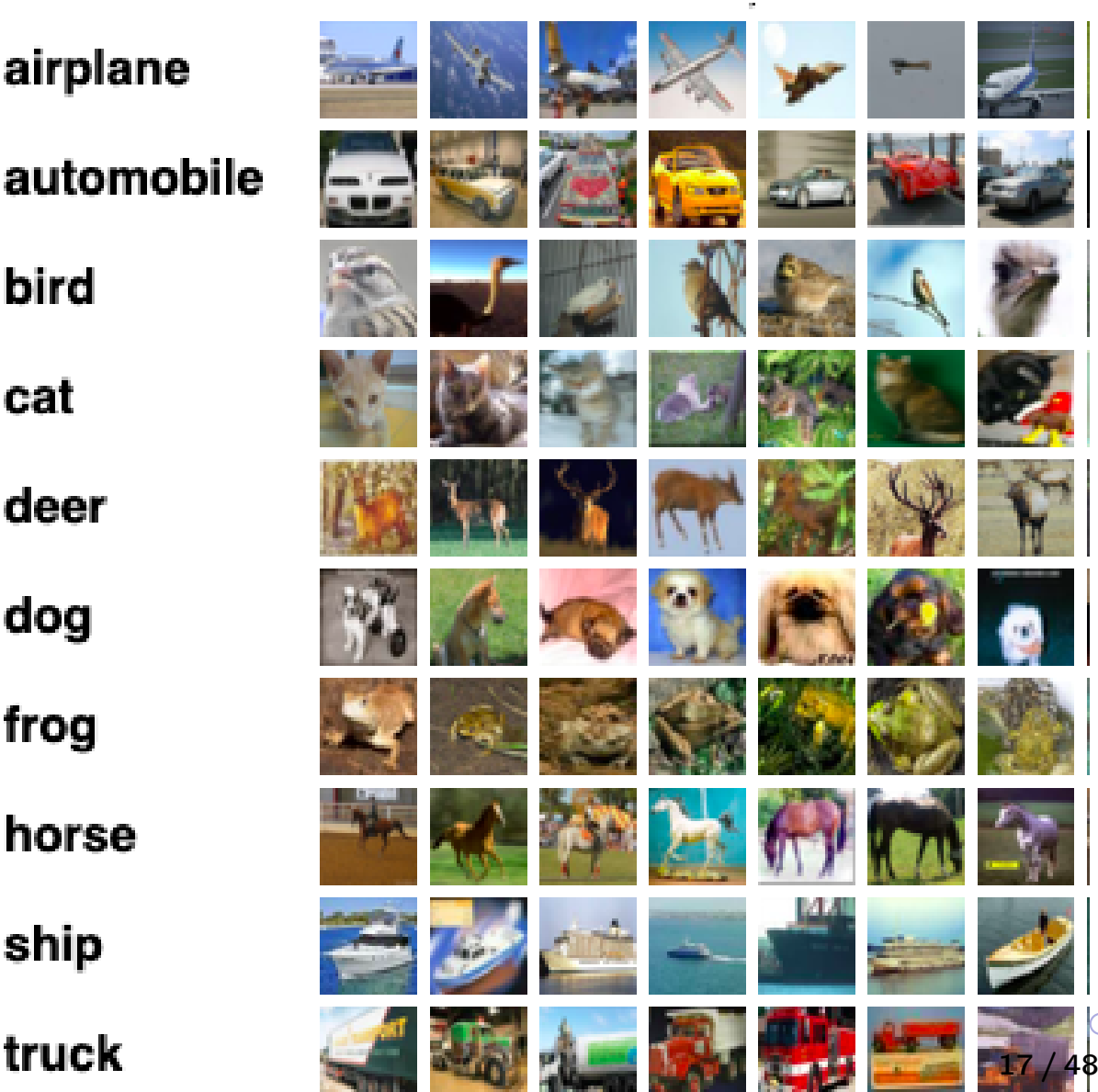
Memorization in deep learning, [Zhang 2017]

RACZEJ NIEPOTRZEBNE Exploring role of memorization in deep learning.

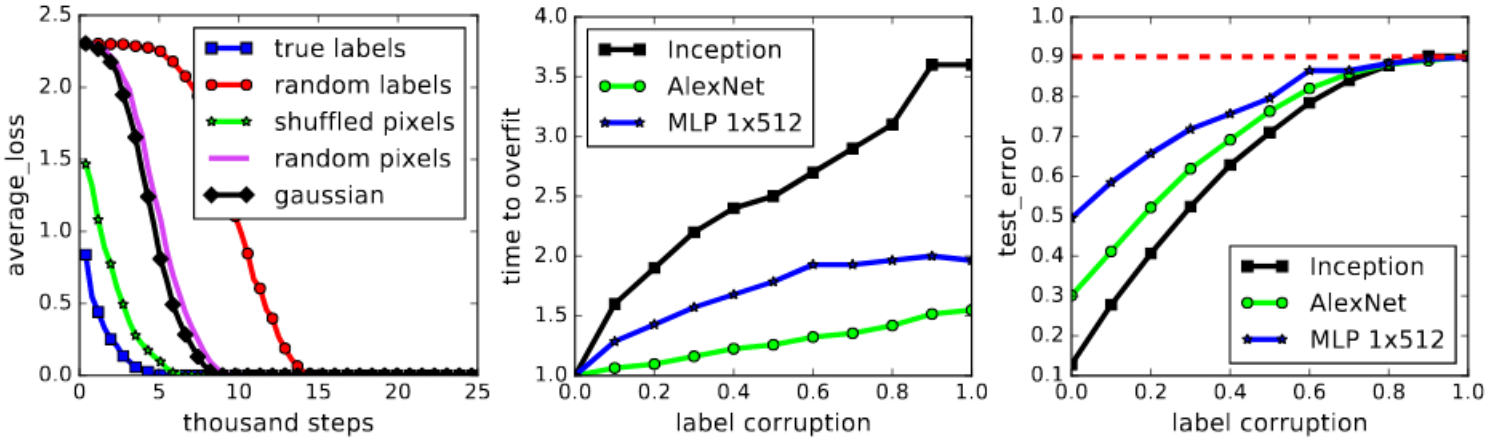
Baselines:

- Standard training
- Random labels
- Shuffled pixels
- Random pixels

How hard those tasks would be for neural networks?



Memorization in deep learning, [Zhang 2017]



(a) learning curves (b) convergence slowdown (c) generalization error growth

Figure 1: Fitting random labels and random pixels on CIFAR10. (a) shows the training loss of various experiment settings decaying with the training steps. (b) shows the relative convergence time with different label corruption ratio. (c) shows the test error (also the generalization error since training error is 0) under different label corruptions.

- Neural nets have capacity to memorize all training data
- Fitting random labels is only slightly harder than normal data
- Memorization is not necessarily a bad thing ...

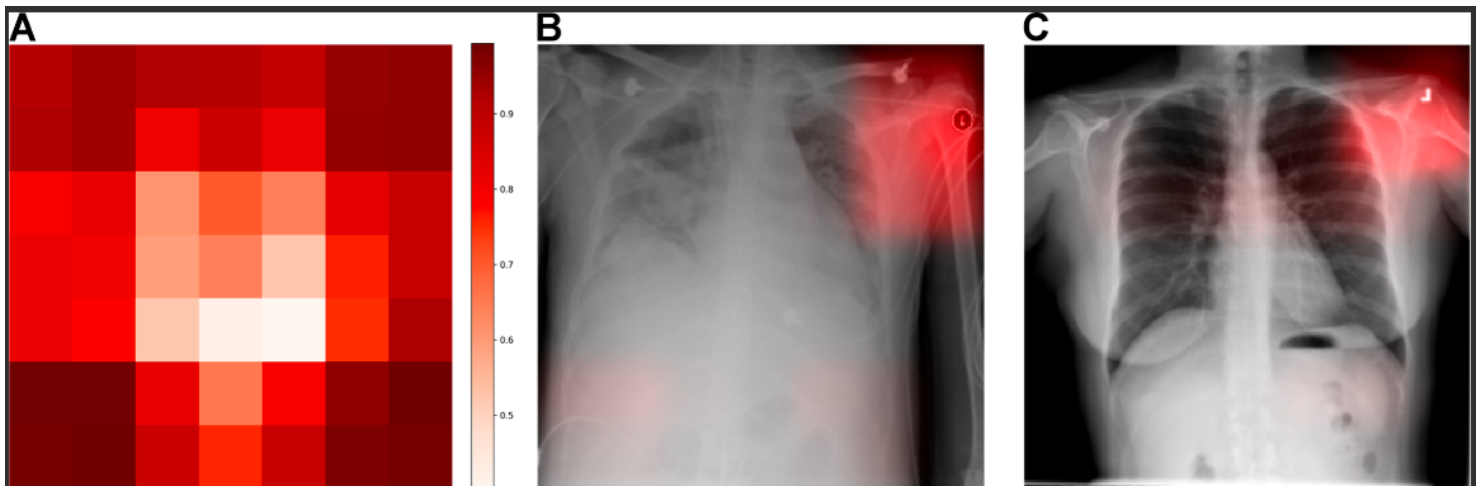
Shorcut learning, [Doersch 2016]

Shorcut learning, [Doersch 2016]

- The task can be easily solved using **chromatic abberration** cues
- Neural nets utilizes the easiest (for the network) solution to the problem

Shorcut learning, [Zech 2018]

- Pneumonia detection
- Models trained on data from few hospitals
- Neural networks specialized (and put the most attention) to recognizing hospitals
- Surprisingly small performance on data from new hospitals

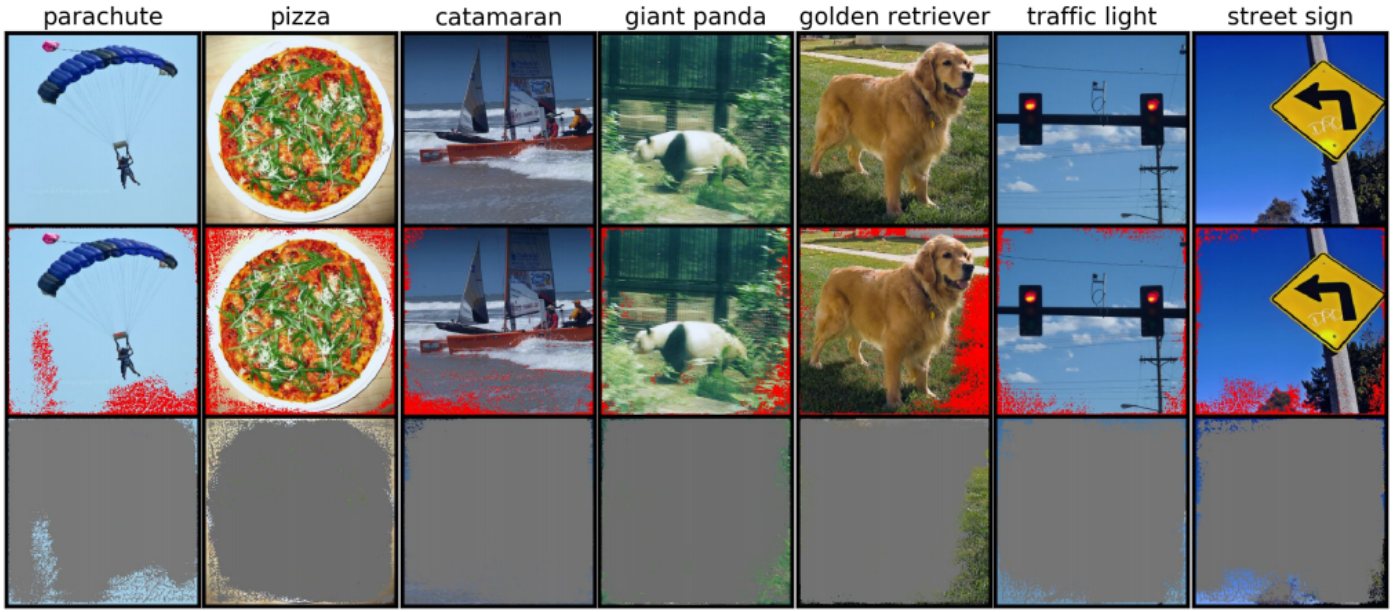


Shotrcut learning: background, [Carter 2020]

- Sufficient input subset (SIS) interpretability procedure applied
- Minimum subset of pixels that yields the same prediction with 90% confidence

Shortcut learning: background, [Carter 2020]

ZBYT DUZY HARDCORE



- CIFAR-10 trained models make confident predictions even when 95% of an input image has been masked and humans are unable to discern salient features in the remaining pixel subset
- Using background information is not necessarily a bad thing, however the model relies too heavily on image backgrounds

Shorcut learning: low-level features

(a) Texture image

81.4%	**Indian elephant**
10.3%	indri
8.2%	black swan

(b) Content image

71.1%	**tabby cat**
17.3%	grey fox
3.3%	Siamese cat

CNNs are biased towards texture [Geirhos 2019]

(a) Texture image

81.4%	**Indian elephant**
10.3%	indri
8.2%	black swan

(b) Content image

71.1%	**tabby cat**
17.3%	grey fox
3.3%	Siamese cat

(c) Texture-shape cue conflict

63.9%	**Indian elephant**
26.4%	indri
9.6%	black swan

CNNs are biased towards texture [Geirhos 2019]

(a) Texture image

81.4%	**Indian elephant**
10.3%	indri
8.2%	black swan

(b) Content image

71.1%	**tabby cat**
17.3%	grey fox
3.3%	Siamese cat

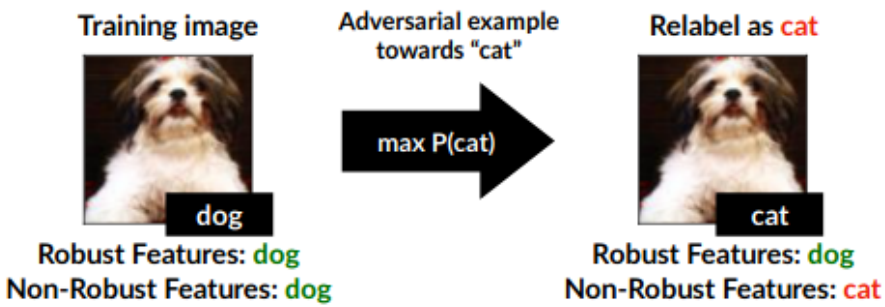
(c) Texture-shape cue conflict

63.9%	**Indian elephant**
26.4%	indri
9.6%	black swan

- Human observers show a bias towards responding with the shape category (95.9 of correct decisions). This pattern is reversed for CNNs, which show a clear bias towards responding with the texture category (VGG-16: 82.8; GoogLeNet: 68.8; AlexNet: 57.1; ResNet50: 77.9)
- Classifier decision is based mostly on local information
- Results in contrast with what we would like to see (building hierarhical representations of recognized objects)

Adversarial Examples Are Not Bugs, They Are Features [Ilyas, 2019]

LEPIEJ WYTLUMACZYC Maybe adversarial examples are something more than just a results of high-dimensional input data?
Let's train a classifier on them!



Adversarial Examples Are Not Bugs, They Are Features [Ilyas, 2019]

- Original accuracy 95%.

Adversarial Examples Are Not Bugs, They Are Features [Ilyas, 2019]

- Original accuracy 95%.
- When trained on adversarial (nuisance dataset) accuracy is 87%!.

Adversarial Examples Are Not Bugs, They Are Features [Ilyas, 2019]

- Original accuracy 95%.
- When trained on adversarial (nuisance dataset) accuracy is 87%!.
- *Adversarial vulnerability is a direct result of our models' sensitivity to well-generalizing features in the data*
- http://gradientscience.org/adv/

Summary

- Neural nets pay attention to totally different features than human perception
- Current CNNs are very brittle in their predictions, and utilize mostly low-level features (texture) as opposed to high-level (shape) abstractions
- Adversarial examples contain highly predictive (non-robust) features within dataset
- Most datasets can be solved using "easy" solution (background information, texture, ...)
- Neural nets are heavily overparametrized with capacity to easily fit random labels.

Overview

How do we measure robustness?

1. Cross-dataset evaluation

How do we measure robustness?

1. Cross-dataset evaluation
   - However still can be biased. For example in autonomous driving majority of the datasets are recorded during day-time, in good weather conditions, in major cities in USA and Europe, **no accidents** etc.

How do we measure robustness?

1. Cross-dataset evaluation
   - However still can be biased. For example in autonomous driving majority of the datasets are recorded during day-time, in good weather conditions, in major cities in USA and Europe, **no accidents** etc.
2. Common-corruptions benchmark (Hendrycks and Dietterich, 2019)

How do we measure robustness?

1. Cross-dataset evaluation
   - However still can be biased. For example in autonomous driving majority of the datasets are recorded during day-time, in good weather conditions, in major cities in USA and Europe, **no accidents** etc.
2. Common-corruptions benchmark (Hendrycks and Dietterich, 2019)

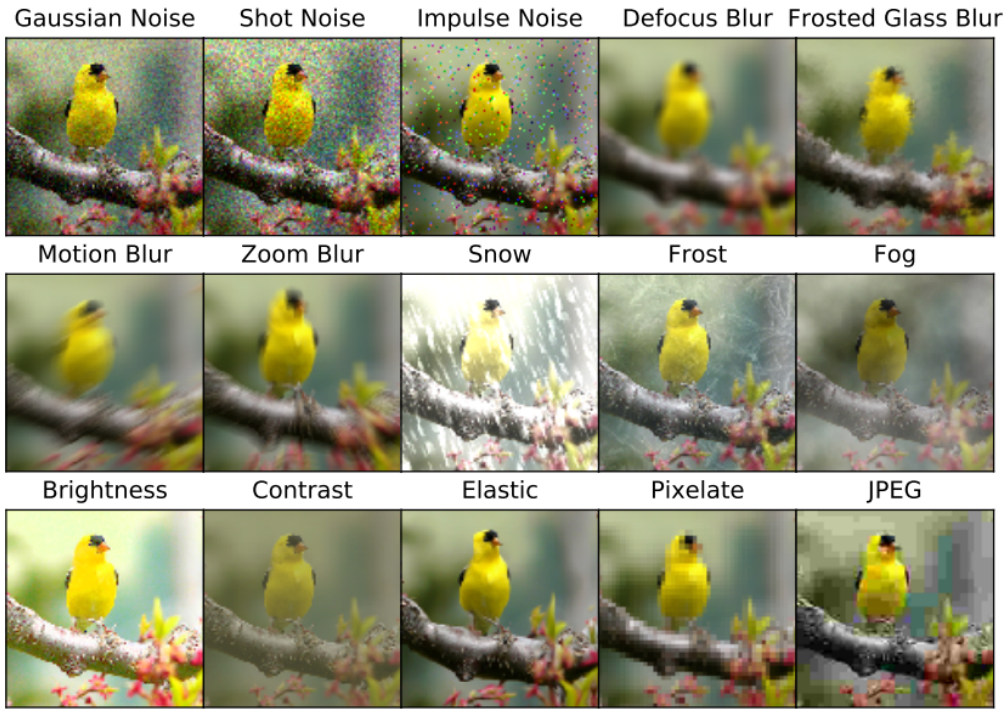


Figure: Source: https://arxiv.org/pdf/1903.12261.pdf

Common Corruptions

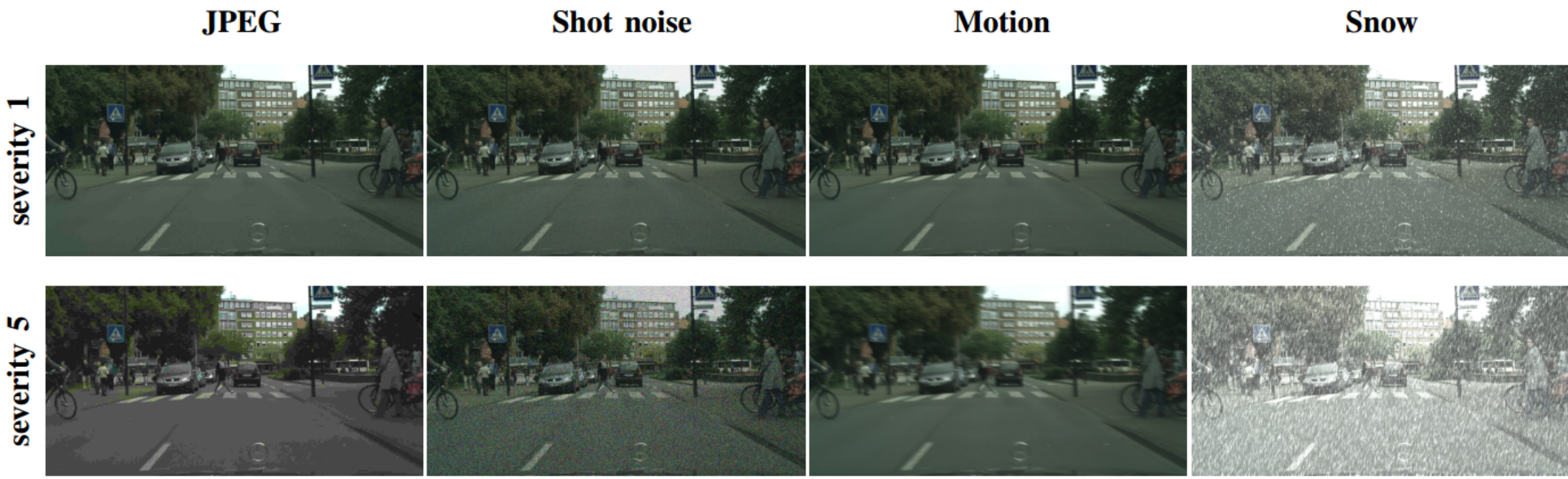


Figure 2: Examples of different corruption types from Common Corruptions benchmark with different severity. Note that at the lowest severity distortions are barely visible whereas at the highest severity they are clearly visible, however semantics of the images are not changed.

1. Cross-dataset evaluation
   - However still can be biased. For example in autonomous driving majority of the datasets are recorded during day-time, in good weather conditions, in major cities in USA and Europe etc.
2. Common-corruptions benchmark (Hendrycks and Dietterich, 2019)
3. Out-of-distribution testing (ObjectNet, Natural Adversarial Examples, many more)

How can it be improved?

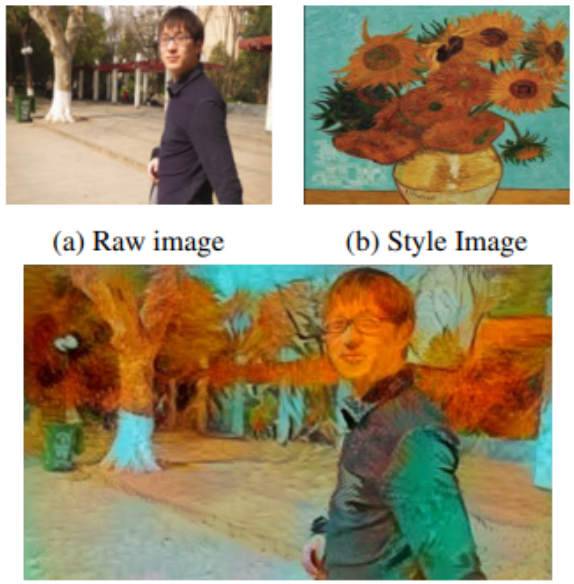
- Larger models: increasing model size improves robustness (Hendrycks and Dietterich, 2019)

How can it be improved?

- Larger models: increasing model size improves robustness (Hendrycks and Dietterich, 2019)
- Large scale pretraining (**1B** images from Instagram) [Orhan 2019], self-supervised learning

How can it be improved?

- Larger models: increasing model size improves robustness (Hendrycks and Dietterich, 2019)
- Large scale pretraining (**1B** images from Instagram) [Orhan 2019], self-supervised learning
- Data Augmentation (**adding style-transfer, gaussian noise)**



(a) Raw image (b) Style Image

(c) Stylized Image

Source: https://arxiv.org/pdf/1909.01056.pdf

Cityscapes and Common Corruptions

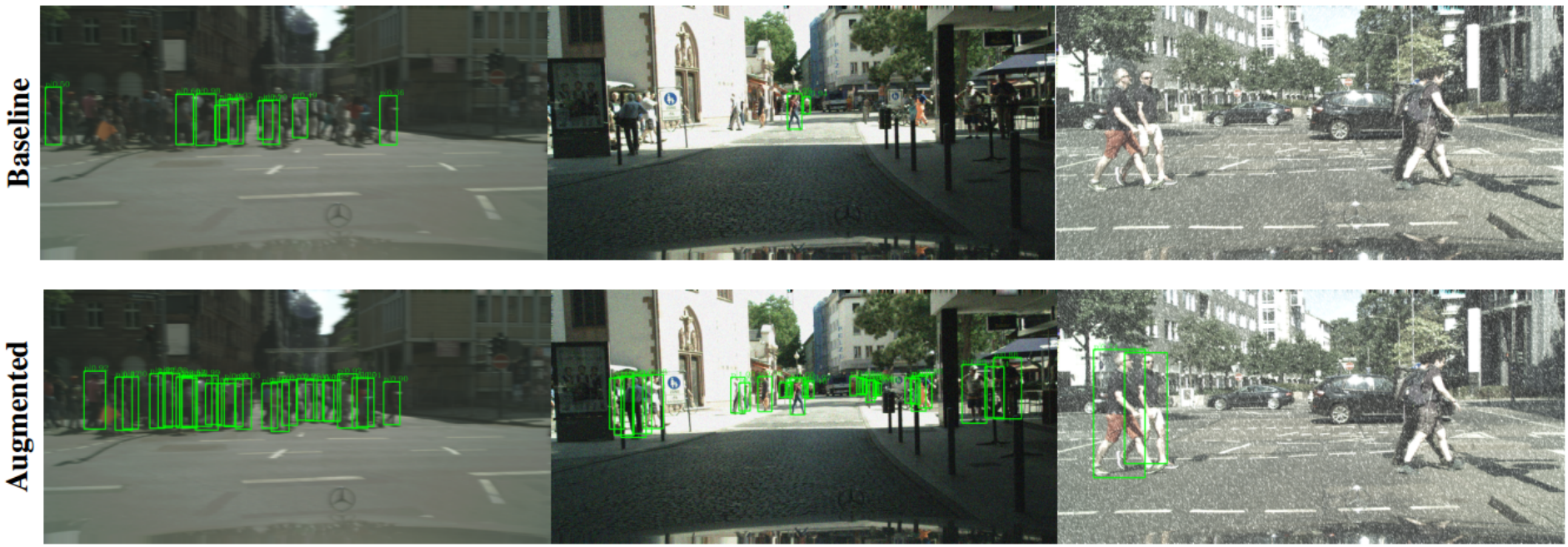


Figure 4: Detection samples for baseline and augmented (*Our + PatchGaussian_0.5*) models for different corruption types. The first column - motion blur (severity intensity of 4), the second column - Gaussian noise (severity intensity of 2), third column - artificial snow with a severity intensity of 2. Note that the distortion for Gaussian noise is almost impercetible,

Cross-dataset evaluation

Overview

Conclusions

- Testing models only on i.i.d. data can give a false sense of security. O.o.d. testing is a must for real-world applications.
- Deep learning based models learn low-level, non-robust features which generalize well within similar distribution (this holds true also for NLP), but may fail in new environments - shortcut learning.
- There exists partial solutions to the problem (data augmentations, large-scale pretraining) and the problem is far from being solved
- Reliable uncertainty estimation (knowing when you don't know) is also an active research area
- **For real world applications test your models on o.o.d. data!**

Bibliography

- Geirhos R., et al. "Generalisation in humans and deep neural networks." Advances in neural information processing systems. 2018.
- Geirhos R., et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." ICLR 2019
- Hendrycks D. and Dietterich T., "Benchmarking neural network robustness to common corruptions and perturbations.", International Conference on Learning Representations, ICLR 2019.
- Hendrycks D., et al. "Natural adversarial examples." arXiv preprint arXiv:1907.07174 (2019).
- Orhan E., "Robustness properties of Facebook's ResNeXt WSL models." arXiv preprint arXiv:1907.07640 (2019).
- Szegedy C., et al. "Intriguing properties of neural networks." 2nd International Conference on Learning Representations, ICLR 2014.
- Barbu, Andrei, et al. "Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models." Advances in Neural Information Processing Systems. 2019.
- Ren, Shuhuai, et al. "Generating natural language adversarial examples through probability weighted word saliency." Proceedings of the 57th annual meeting of the association for computational linguistics. 2019.

Bibliography

- Doersch, Carl, Abhinav Gupta, and Alexei A. Efros. "Unsupervised visual representation learning by context prediction." Proceedings of the IEEE international conference on computer vision. 2015.
- Zhang, Chiyuan, et al. "Understanding deep learning requires rethinking generalization." arXiv (2016).
- Zech, John R., et al. "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study." PLoS medicine (2018).
- Carter, Brandon, et al. "Overinterpretation reveals image classification model pathologies." arXiv (2020).
- Carlini, Nicholas, and David Wagner. "Audio adversarial examples: Targeted attacks on speech-to-text." 2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018
- Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

Thank you for attention
Questions?