Invariant object recognition is a personalized selection of invariant features in humans, not simply explained by hierarchical feed-forward vision models

MOHAMMAD HOSSEIN
NIKI MALEKI

DEC 2020

SURVEY

One of key ability of brain is O.R, rapid, accurate, with variation.

observed that humans relied on specific (diagnostic) object regions for accurate recognition.

They remain relatively consistent (invariant) across variations.

but feed-forward feature-extraction models selected view-specific (non-invariant) features across variations. This suggests that models can develop different strategies.

Human changes their diagnostic features and flexibly shifted their feature extraction strategy from view-invariant to view-specific when objects became more similar.

MIRC

Humans rely significantly on specific sets of object parts (i.e. visual features or simply features), referred to as Minimal Recognizable Configurations (MIRCs).

MIRC

In other words, some specific object parts were considered more informative to humans, but provided as much information as any other parts for computational models.

AIMS:

Two major questions. First, what is the feature-based strategy used by humans
when recognizing objects under variations?
Second, do hierarchically organized feature extractor models of vision adopt the same strategy as humans do for invariant object recognition?

FORMER STUDIES

In a comparative study between humans and monkeys, it was shown that humans relied on sets of relatively consistent (invariant) features when objects were rotated in image plane, whereas monkeys used a screen-centered (view-specific) strategy.

A follow-up study on rats showed that the consistency (invariance) of the diagnostic features was directly related to the level of similarity between the objects which were discriminated

It remains unknown what strategies would humans adopt to perform a similar task.

EXPERIMANTS

We generated an image set which presented objects in thirteen different conditions in four variations. We then asked humans to discriminate a pair of 3D objects.

To provide computational cases for comparison, we also investigated the strategies used by a pixel-level ideal observer39 and a deep convolutional neural network20 (i.e. known as AlexNet, Trained on Imagenet).

RESULTS

Humans relied on a few object features (diagnostic features) in each variation condition. These diagnostic features could be relatively consistent (invariant) across variation conditions.

Interestingly, the level of diagnostic features consistency was determined by the level of similarity between the two objects which were discriminated. We also show that, neither an ideal observer nor a deep convolutional neural network could emulate human recognition strategies.

This implies that, compared to humans, who can generalize the diagnostic features from one variation condition to another, the hierarchical models of vision used in this study, adopt a unique strategy for each variation condition.

These results suggest that human object recognition involves more than just a series of feature extraction levels.

EXPERIMENTAL TASK DETAILS

1

Car models were sport cars (cars 1 and 2) and the third (car 3) was a truck.
To a total of 39 unique images in the set for the three cars (i.e. each car underwent 13 variation conditions).

2

In order to obtain the visual features which were most relevant to the recognition of each object, we used Bubbles method.
To find the contributing features, a bubbles mask is put on the object image.
After applying random binary bubbles, the object is only partially observable through the pores.

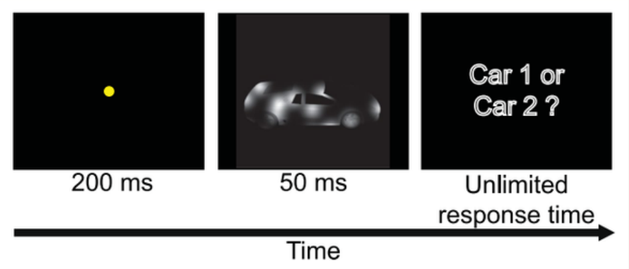
3

In fact, we dynamically altered the number of bubbles on the masks in the range of 10 to 25 in steps of five, based on the subject's performance.

EXPERIMENTAL TASK DETAILS

4

Every trial started with a central yellow fixation point which was aimed to preclude subjects from making eye movements, followed by presentation of a masked stimulus, and finished after subject's response.



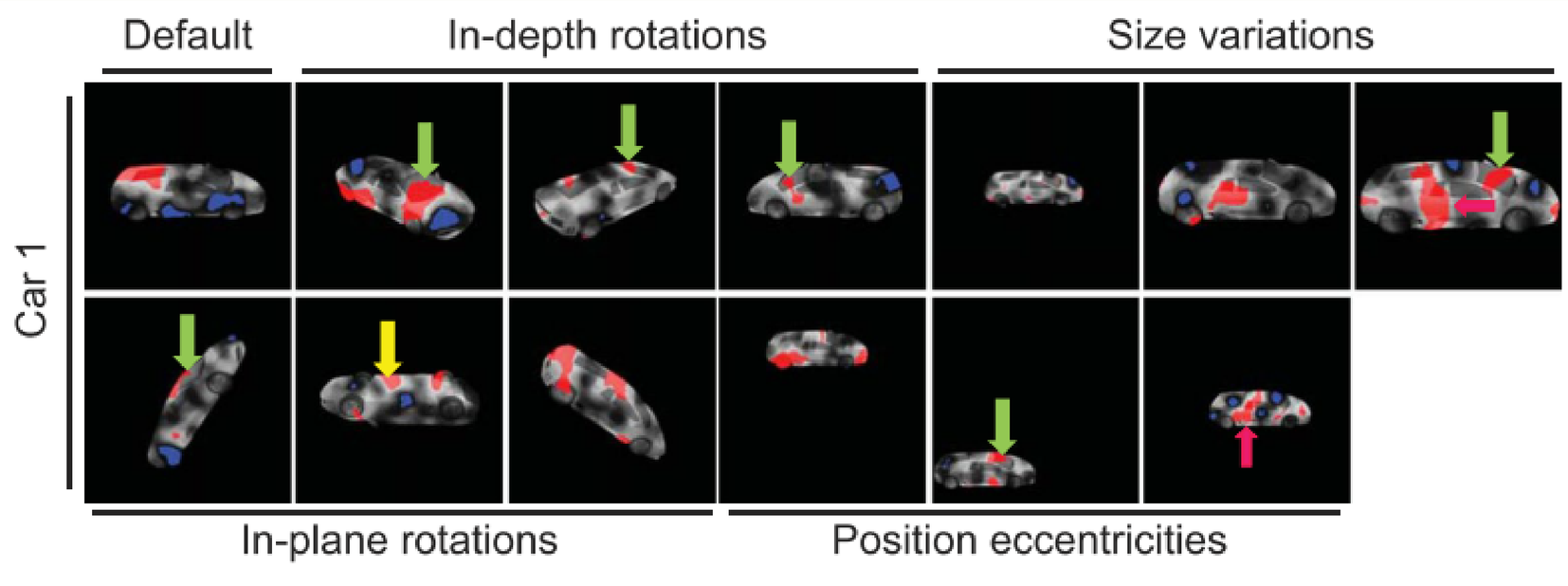
5

It was important for us to study the effect of object similarity on human recognition strategies.
The first subject group discriminated car 1 from car 2(LSD) and in the second experiment the second group discriminated car 2 from car 3(HSD).

6

We trained the subjects on all testing conditions to minimize the bias from involving high-level memory mechanisms such as view-point generalization and learning.
After the stimulus offset, although they were asked to respond as fast and accurately as possible, subjects had an unlimited time to determine.

CONSTRUCTING SALIENCY MAP



CONSTRUCTING SALIENCY MAP

Diagnostic (red regions) and anti-diagnostic (blue regions). the brightness of car regions indicates the importance of the region in its discrimination with brighter regions leading to more accurate answers.

ROLE OF DIAGNOSTIC FEATURES

1

Subjects showed a significant performance decline in the most peripheral position condition on the LSD. Based on the accuracy and reaction times of the subjects in position conditions we can conclude that the subjects did not make saccadic eye movements towards the cars in non-default conditions.

2

For the in-plane rotations, the performance dropped (i.e. accuracy decreased and reaction time increased) as the objects underwent from 0 to 180°.
Consistent reliance on specific car parts across variations suggested that some car parts might have been more informative than others.

3

Next we evaluated the importance of different car parts in recognition by counting the number of times the part was diagnostic.
Results from both the HSD and LSD showed the highest relative importance for car floors.
It seems that, one key parameter which determined the relative importance of the car parts was the relative size of the parts.

IMPACT OF OBJECT SIMILARITY

1

In order to explain humans reliance on diagnostic features could be explained by a view-invariant or view-specific strategy, we measured the amount of overlap between the diagnostic regions found for different variation conditions.

$$Op = \frac{Oa}{Oa + Da1 + Da2}$$

2

Where Oa refers to the diagnostic car areas (i.e. number of pixels) which overlapped between the first and the second variation conditions, Da1 and Da2 refer to the area of diagnostic regions for the variation conditions 1 and 2, respectively.

3

These results show a significant advantage for aligned overlaps compared to raw overlaps, which adds support to a view-invariant strategy, rather than a screen-centered strategy.

IMPACT OF OBJECT SIMILARITY

4

The average aligned overlaps of car 2 on the HSD showed a significantly lower value compared to its value on the LSD.
This implies that the lower similarity between the objects, increased the consistency of the diagnostic features across variation conditions.

5

Interestingly, car 2, which participated in both experiments, showed a significantly lower number of features on the HSD compared to LSD. This result revealed a decrease in size of the diagnostic features as a result of lower similarity between categories in the discrimination task, which is consistent with previous reports from rats.

CONSISTENCY OF DIAGNOSTIC FEATURES ACROSS VARIATIONS.

1

As opposed to humans, both the ideal observer and the computational model showed higher values of raw versus aligned overlaps,
This suggests that, rather than a view-invariant strategy in choosing diagnostic features, a screen-centered strategy seems to be at work for the observer and the model compared to humans.

2

There were only 4, 3 and 4 diagnostic features which were shared by humans and respectively the observer, middle and last model layers, on the HSD.
These numbers were respectively 3, 1 and 1 on the LSD.
These qualitative results suggest that a different set of mechanisms might be developed for object recognition in humans.

3

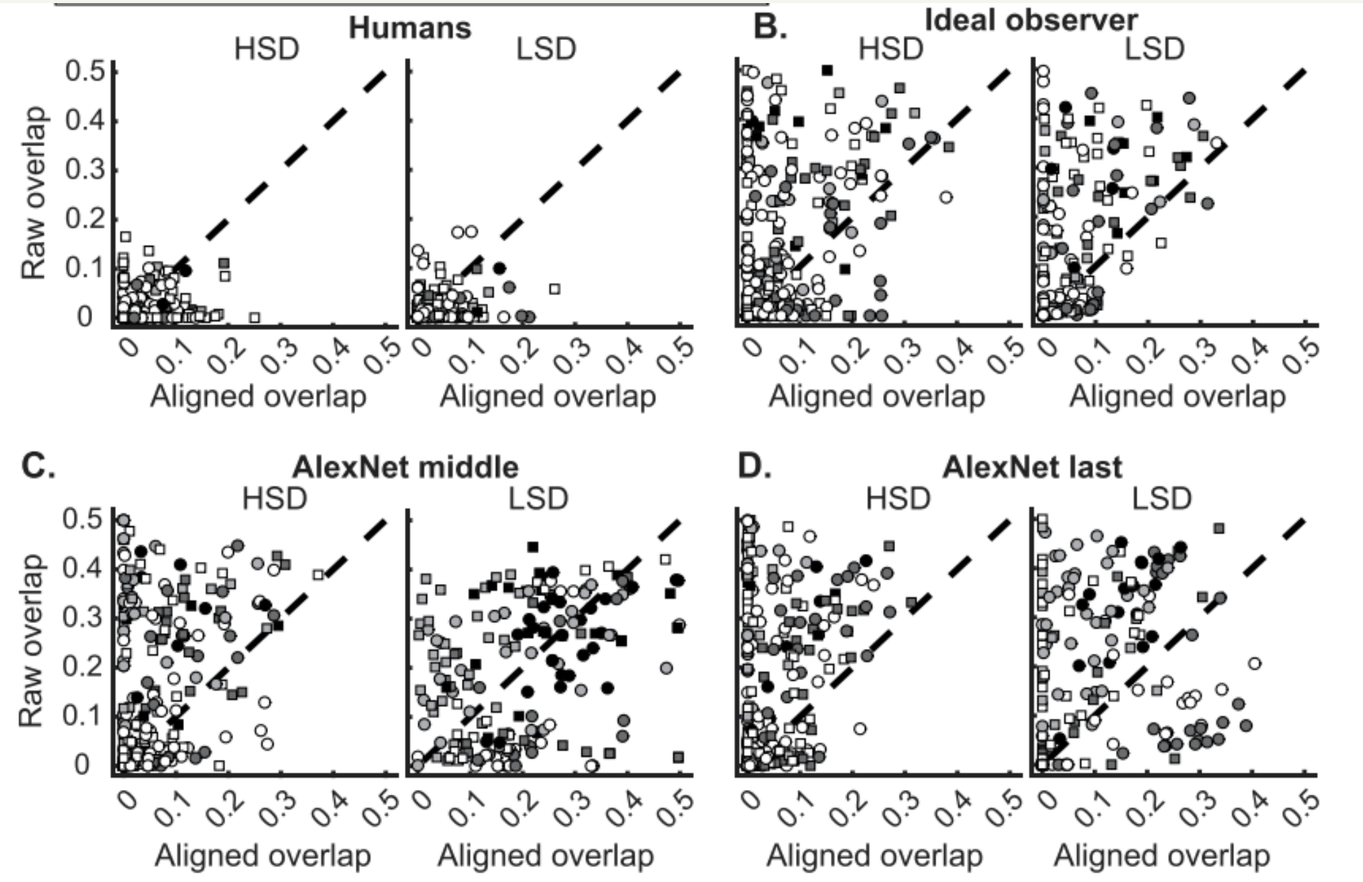
As the correlation results show, neither the ideal observer nor middle/last model layers showed many instances of significant correlation with the human results.

CONCLUSION

The results show that, both the ideal observer and computational model used a higher number of diagnostic features with larger areas and absolutely worked different on object similarity, so

it can be concluded that:

neither a pixel-level (i.e. idea observer), nor an intermediate-complexity (i.e. middle model layer) or high-complexity (i.e. last model layer) feature extractor algorithm could emulate the human strategies when solving the feature-based object discrimination tasks of the current study.



Difference of distribution between humans and models.

THANK YOU

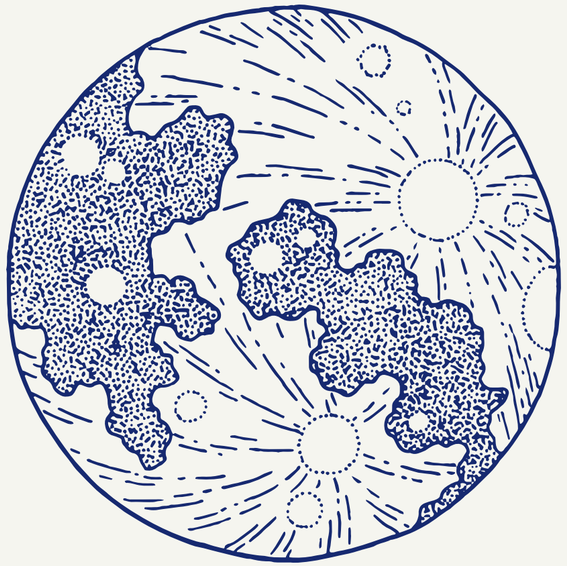
FRAGILE OBJECT RECOGNITION IN NATURAL IMAGES

Published as a conference paper at ICLR 2019

The human ability to recognize objects is impaired when the object is not shown in full. "Minimal images" are the smallest regions of an image that remain recognizable for humans.

Show that a slight modification of the location and size of the visible region of the minimal image produces a sharp drop in human recognition accuracy.

WE DEMONSTRATE THAT SUCH DROPS IN ACCURACY DUE TO CHANGES OF THE VISIBLE REGION ARE A COMMON PHENOMENON BETWEEN HUMANS AND EXISTING STATE-OF-THE-ART DEEP NEURAL NETWORKS (DNNS), AND ARE MUCH MORE PROMINENT IN DNNS.



FRAGILE RECOGNITION IMAGE (FRI)

IS A REGION OFAN IMAGE FOR WHICH A SLIGHT CHANGE OFTHE REGION'S SIZE OR LOCATION IN THE IMAGE PRODUCES A LARGE CHANGE IN DNNRECOGNITION OUTPUT.

In human vision, the more general definition of fragile recognition that we are introducing here is not useful because human minimal images appear only when the visible area of the object is small.

Adversarial examples are images with small synthetic perturbations that are imperceptible to humans, but produce a sharp drop in DNN recognition accuracy

SURVEY

1

Unlike these adversarial examples, fragile recognition arise in natural images without introducing synthetic perturbation. This causes new concerns for use of DNNs in computer vision applications.

2

We evaluate FRIs in ImageNet for state-of-the-art DNNs, specifically VGG- 16, Inception, and ResNet. Results show that FRIs are abundant and can occur for any region size.

3

Known strategies to increase network generalization, i.e. adding regularization and data augmentation, reduce the number of FRIs but still leave far more than humans have.

EXTRACTING FRAGILE RECOGNITION IMAGES

1

The method for extracting human minimal images employs a tree search strategy. The full object is presented to human subjects and they are asked to recognize it. If at least 50% of subjects recognize it correctly, smaller crops of the object, called descendants, are tested and we will go on until it fails and that crop is human minimal image.

2

Our FRI extraction method for DNNs relies on an exhaustive grid search. consists on a two-step process: first, every possible square region is classified by the DNN and the correctness is annotated in the correctness map. From the correctness map, each region's correctness is compared with the region's slightly changed location or size in order to determine if there has been a change of the correctness.

3

FRIs are regions that are classified correctly and a small change causes failure, as well as regions that are classified incorrectly and a small change causes success.

FRAGILE RECOGNITION IMAGE (FRI) EXTRACTION FROM CORRECTNESS MAPS

1

We define different variations of FRIs, depending if they are based on changes on the location or size of the image region, and on how strict we are when evaluating the changes in the correctness of DNN.

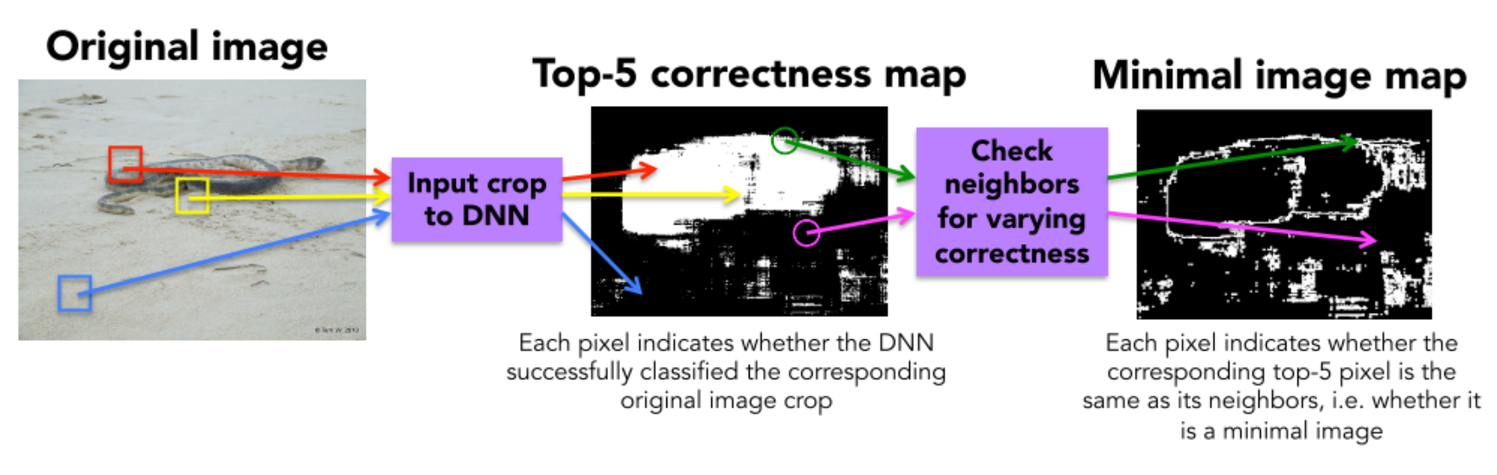
2

"Shift" is a one-pixel translation of the region location; "shrink" is a two-pixel reduction of the region side length within the region's original boundaries. "Loose" FRIs are regions such that there exists a small change that flip network correctness. "Strict" FRIs are regions such that network correctness is flipped for all small changes.
These definitions yield four fragile recognition types: loose shift, loose shrink, strict shift, and strict shrink. Note that strict shrink is the most analogous to human minimal images.

3

The correctness maps are used to detect fragile recognition due to shifts by comparing neighbouring pixels in a correctness map, and due to shrinks by comparing correctness maps at two slightly different region sizes.

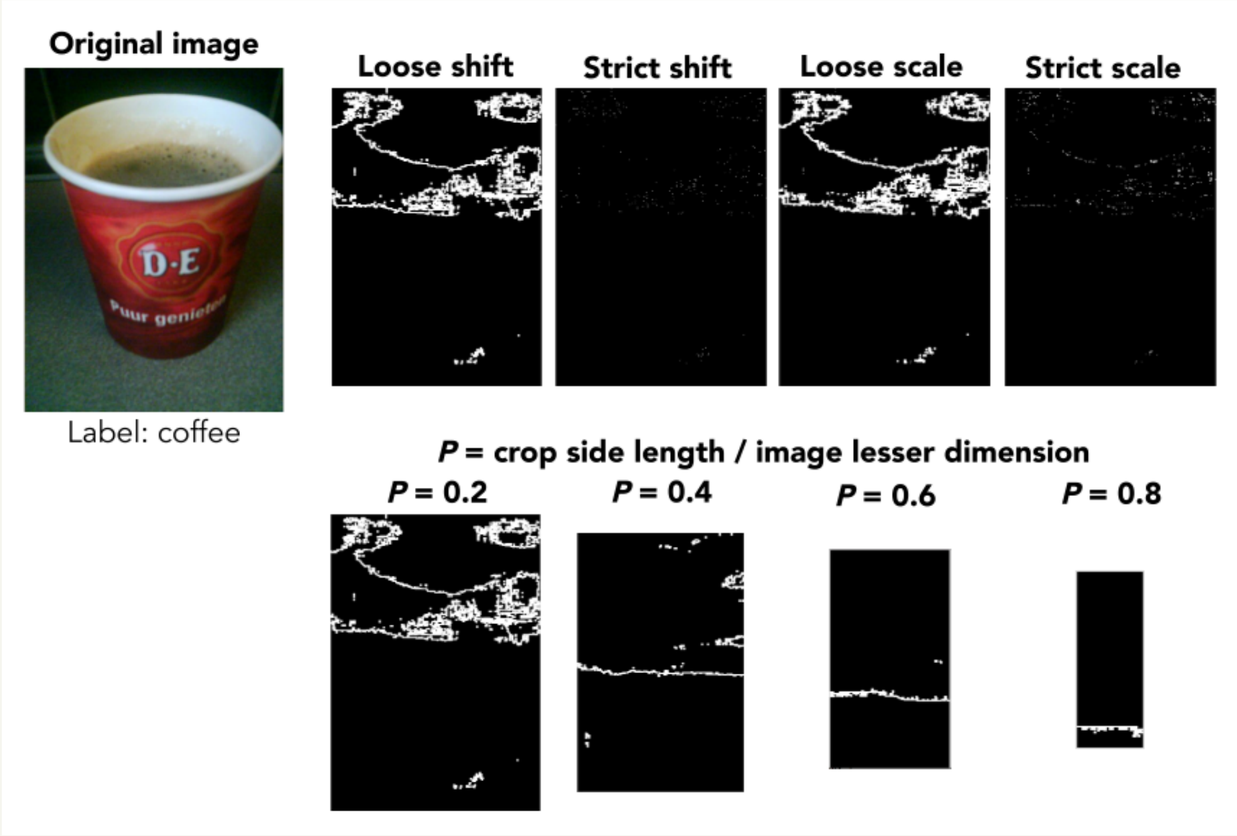
FRAGILE RECOGNITION IMAGE (FRI) MAPS



The FRI map shows FRIs for which a one-pixel shift in any direction of the visible region produces an incorrect classification.
The white pixels indicate FRIs and black pixels indicate non FRIs.
Each pixel of the map indicates whether the corresponding window in the original image is an FRI.

We observe that FRIs are usually located within object boundaries but can also be found in the background. This is because DNNs are able to recognize regions that only contain background, as they have been shown to exploit dataset biases.

FRAGILE RECOGNITION IMAGE (FRI) MAPS



FRAGILE RECOGNITION IMAGES FOR STATE-OF-THE-ART DNNS IN IMAGENET

1

The following experiments are performed on 500 images, randomly sampled from ImageNet's validation set,

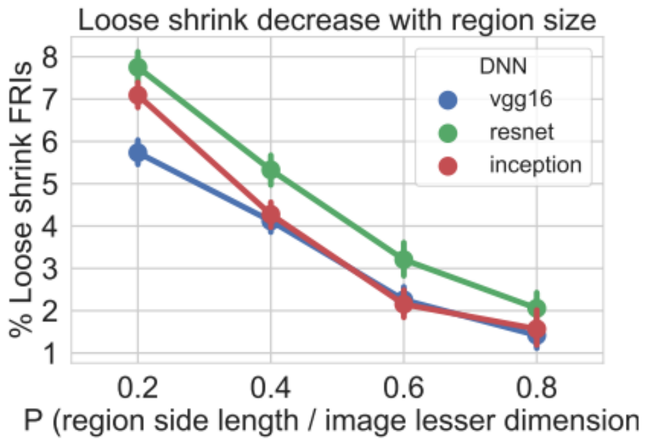
2

The results show that there are many regions in an image for which the network is very sensitive to slight changes

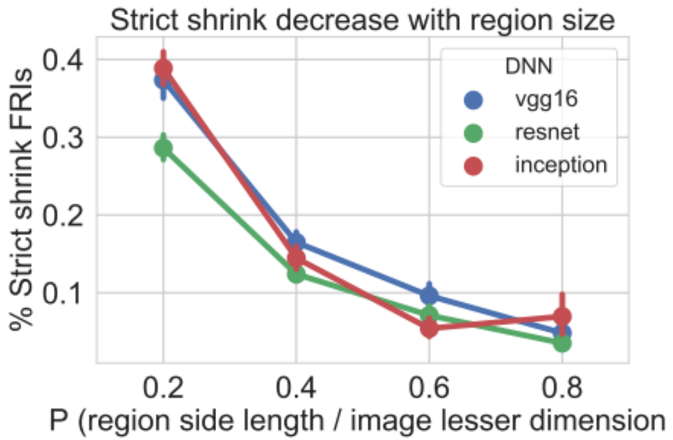
3

Note that smaller FRIs are much more frequent than larger ones.

We verified that FRIs are not an artifact of the algorithm that resizes the region to the size required by the DNN (224 × 224 pixels for VGG-16). We took regions of side length 224 and removed any resizing before input, and we observed that this procedure produces the same results we reported.

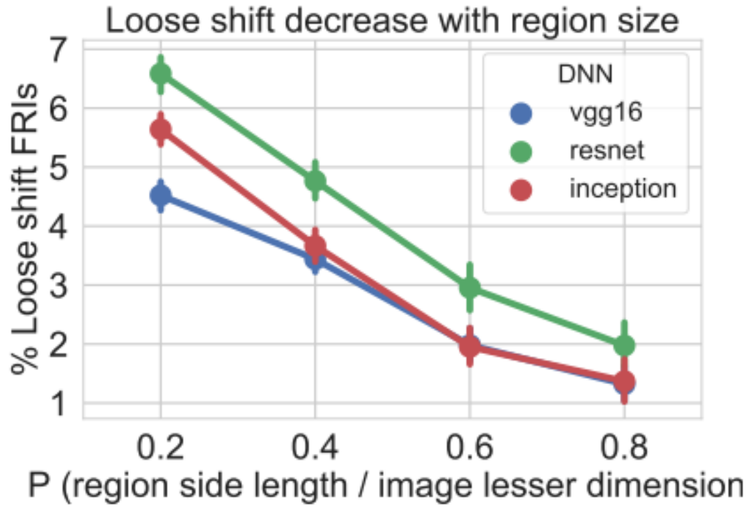


(a) loose shrink FRIs

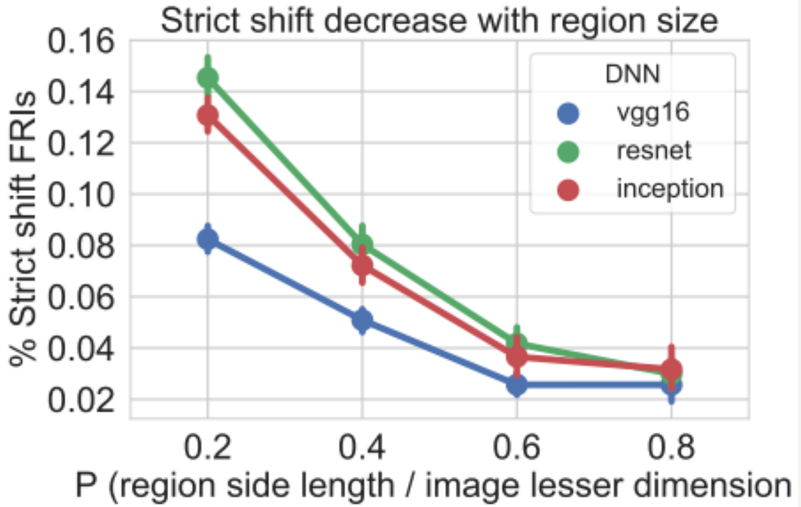


(b) strict shrink FRIs

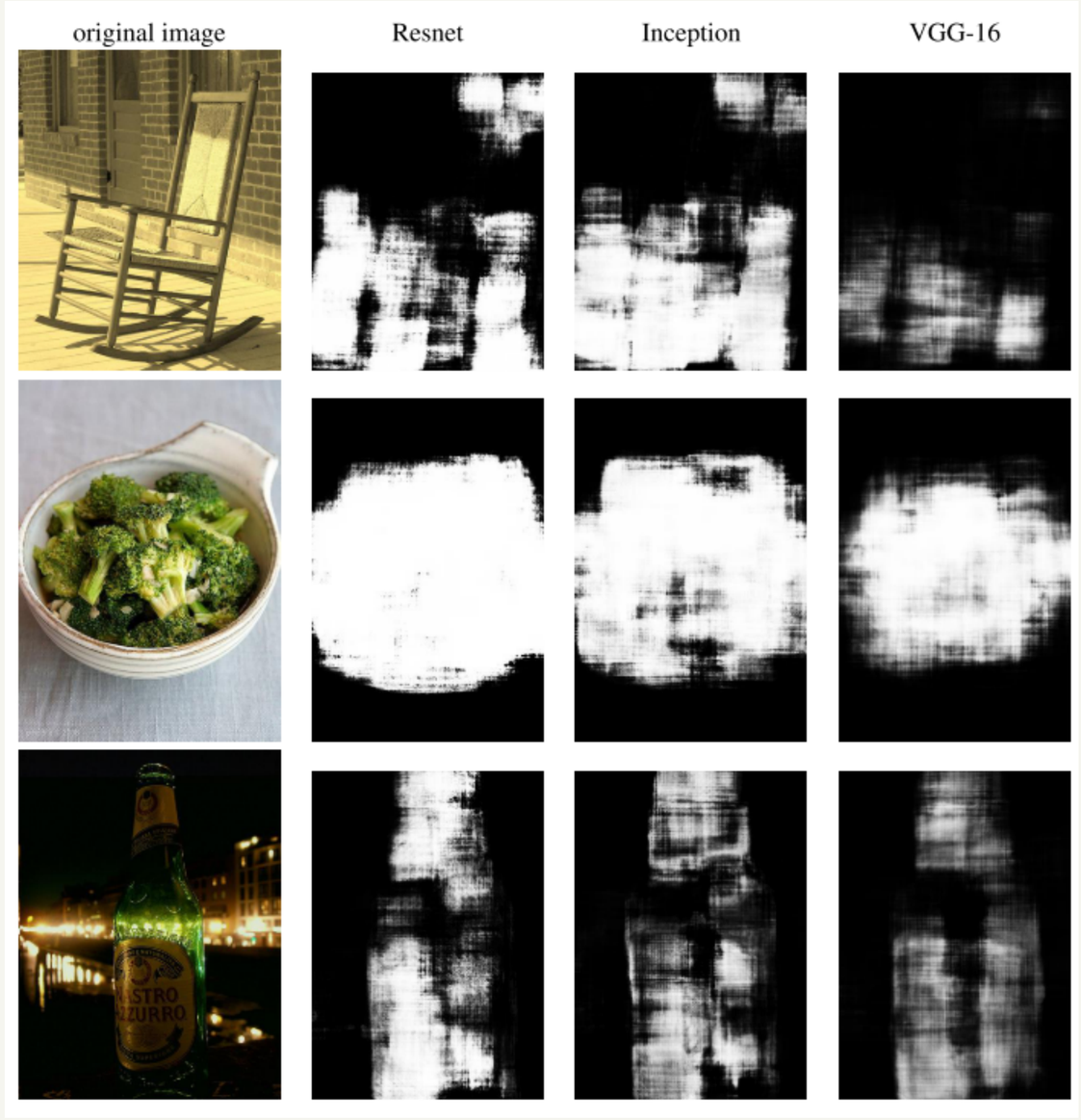
FRIs are generally less frequent for larger regions of the image..



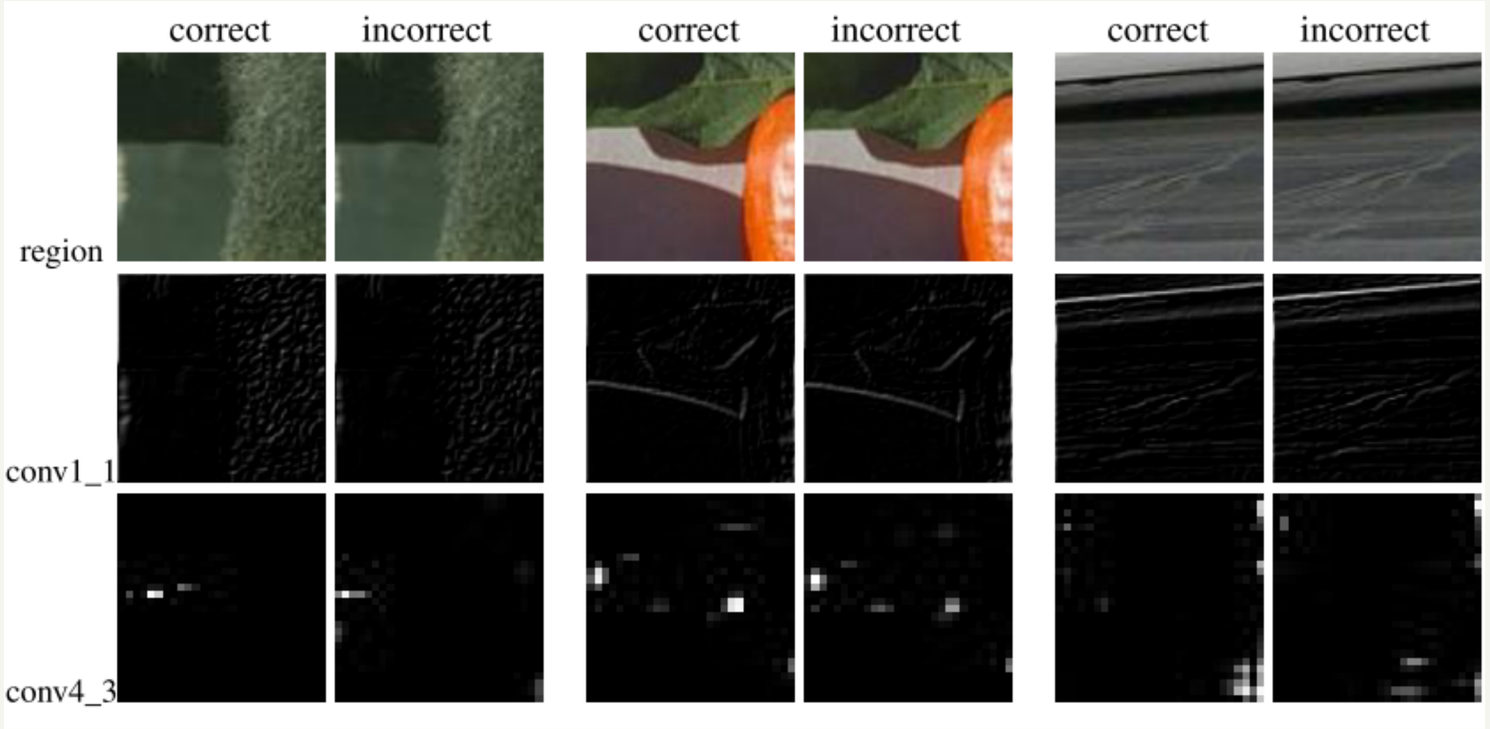
(a) loose shift FRIs



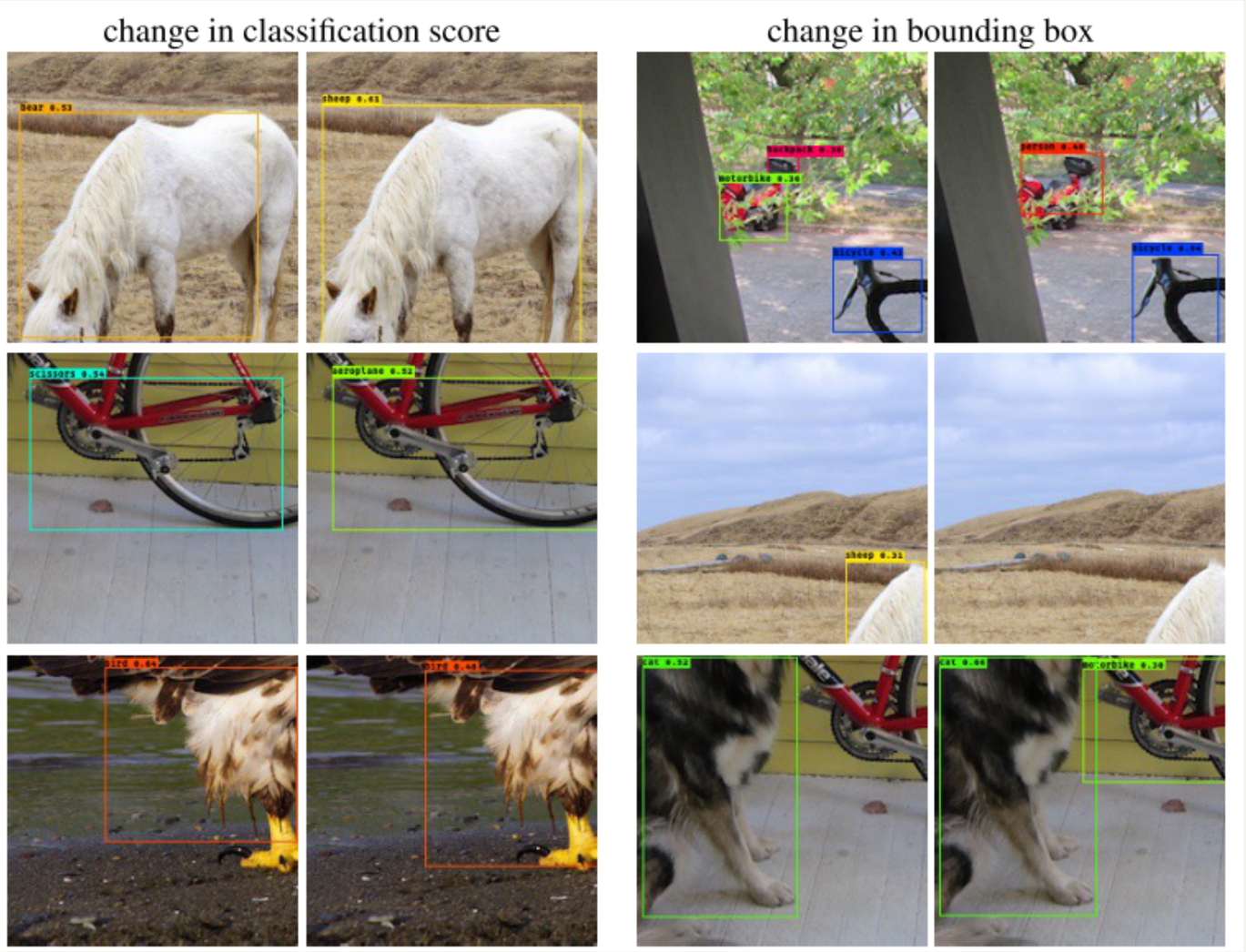
(b) strict shift FRIs



We display the DNN's output confidence in the true class for individual regions in map form. We see that sharp drops in confidence are frequent within an image.



We show qualitative examples of the activation maps at different layers of the DNN of the correctly classified crop and its shifted version. These examples show what we have observed in all cases: the activation maps are imperceptibly similar at the first layers but are clearly different at the last layers.



The output bounding boxes and their corresponding label scores are dramatically different for these two cropped regions.

FRAGILE RECOGNITION WITH DATA AUGMENTATION AND REGULARIZATION

1

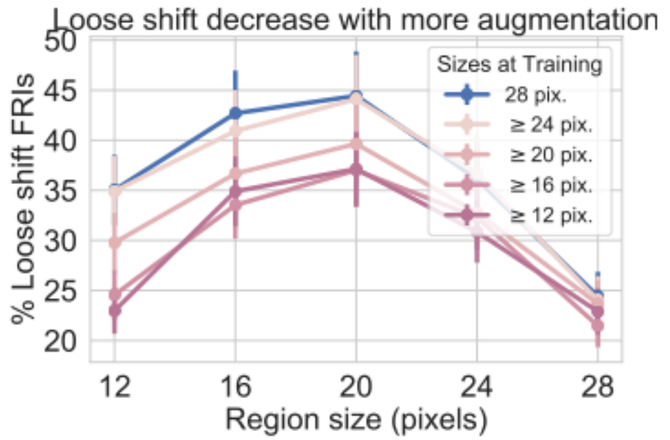
In this experiment we use the CIFAR-10 dataset.
We reproduce the AlexNet version for CIFAR-10,
which consists of two convolutional-pooling-normalization layers followed by two fully connected layers, all regularizers and data augmentation are turned off.

2

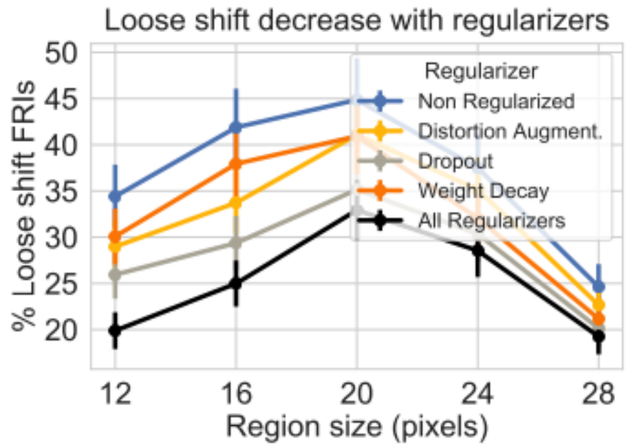
For data augmentation, we augment the training dataset with FRI regions of at least a given size. For regularizers, we add weight decay, dropout, and distortions (e.g. reflections, altered brightness, altered contrast).

3

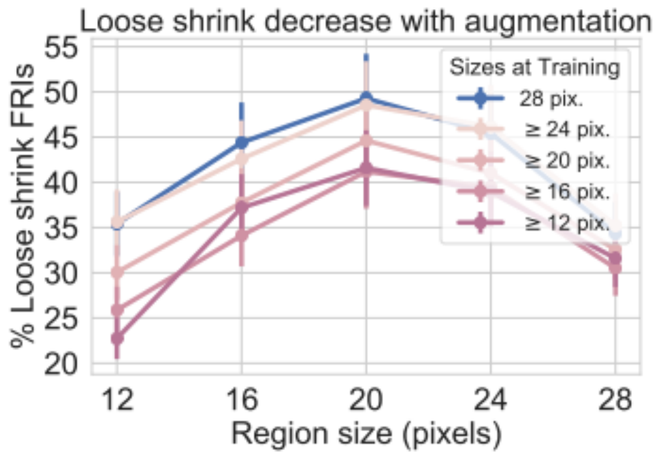
both data augmentation and regularization have a clear impact on FRI occurrence in all cases.
For regularizers specifically, dropout provides the most individual improvement.



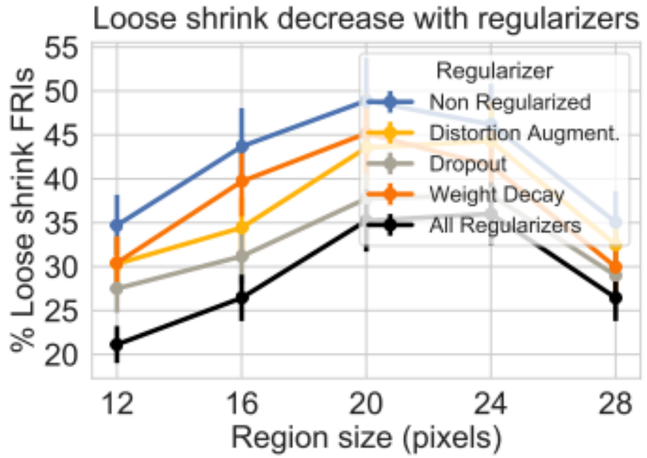
(a) data augmentation by cropping image regions



(b) DNN with regularization



(a) loose shrink FRIs with data augmentation



(b) loose shrink FRIs with regularization

Augmenting the training set with crops of FRI sizes reduces overall FRI occurrence.

FRIs can be mitigated but not eliminated. These works reduce overall FRI occurrence, but many FRIs remain.

FRAGILE RECOGNITION IS NOT LACK OF OBJECT LOCATION INVARIANCE

1

We embed the original image in a larger image and shift it in the image plane (while filling in the rest of the image with a simple inpainting procedure)

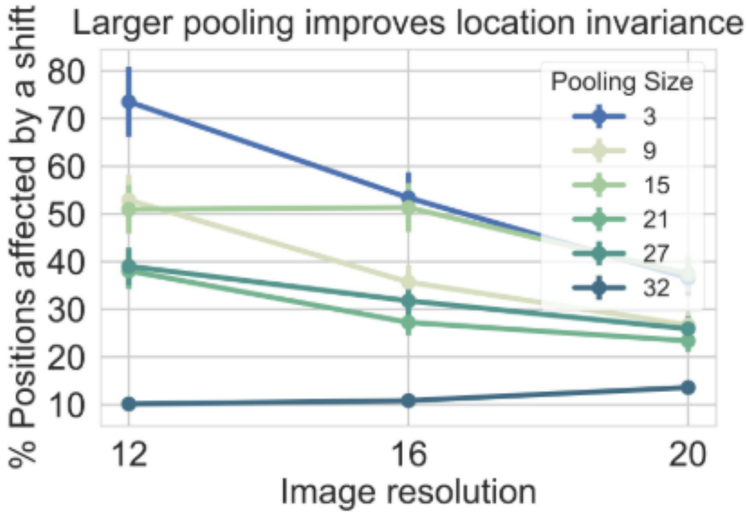
2

We evaluate the network trained on CIFAR-10 and introduced in the previous section with different pooling region sizes for the second pooling layer. These sizes range from three pixels to the entire image

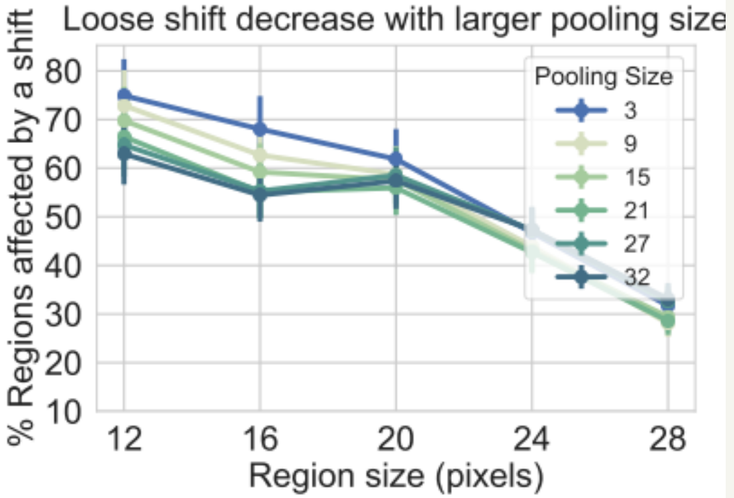
3

We see that increasing the pooling size only slightly decreases FRIs

Comparison of location invariance and fragile recognition images



(a) output label changes due to shifts in object position



(b) output label changes due to shifts in visible region position

(a) shows that maximum pooling makes the DNN almost entirely robust to FRI.
By contrast, (b) shows that larger pooling sizes provide little to no robustness to slight shifts of the visible region.

Conclusion

1

Our results have revealed that the fragile recognition level in humans is fundamentally different from the one in DNNs in terms of size, position and frequency.

2

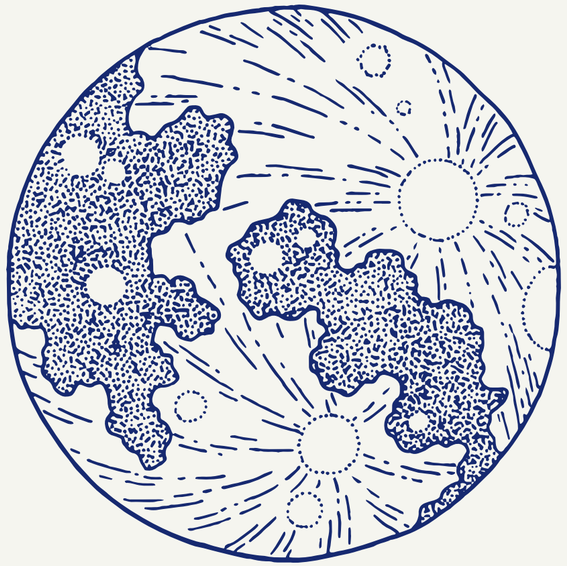
Furthermore, data augmentation, regularization and larger-size pooling regions alleviate fragile recongition in DNNs, but are not sufficient to close the gap with humans.

3

Making DNNs robust like humans remains a challenge for the community. Finally, we have shown that fragile recognition is a more complex phenomenon than object location invariance, which exposes new concerns of the recognition ability of DNNs in natural images, even without adversarial patterns being introduced.

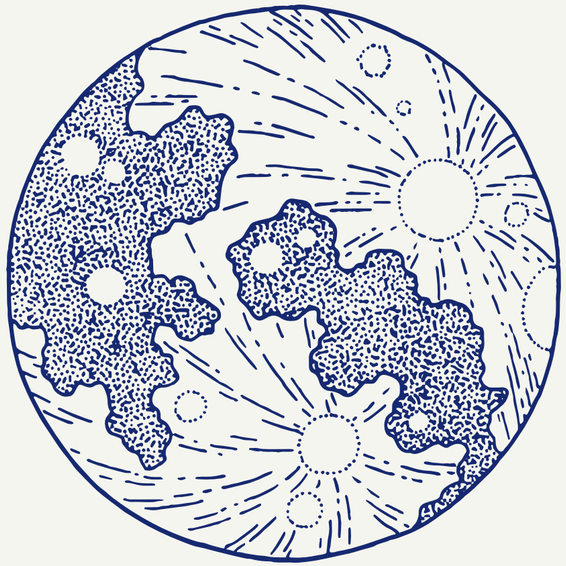
ATOMS OF RECOGNITION IN HUMAN AND COMPUTER VISION

JANUARY 11, 2016



INTRO

Here we show, by introducing and using minimal recognizable images, that the human visual system uses features and processes that are not used by current models and that are critical for recognition.



INTRO

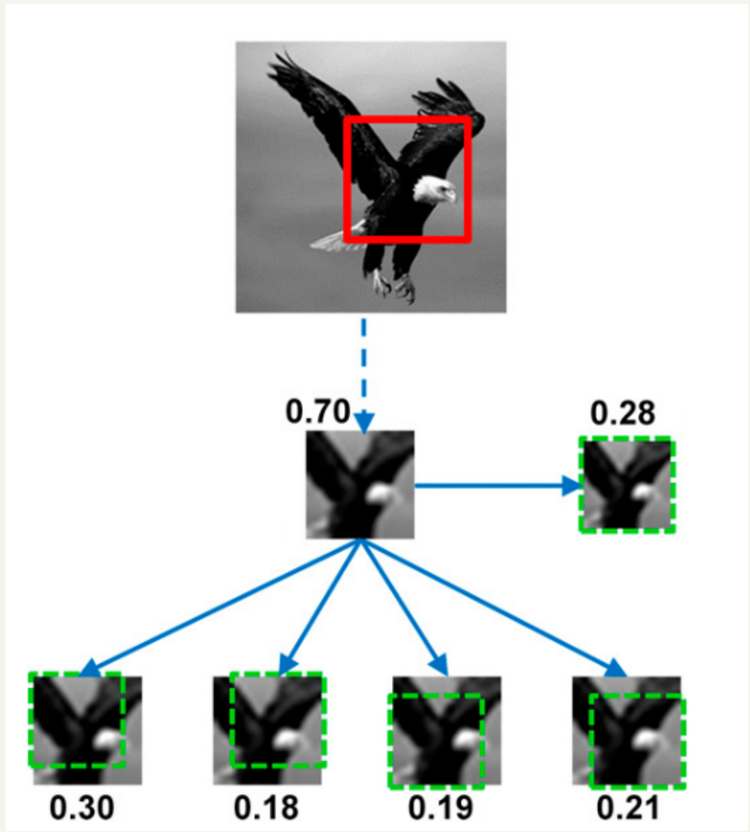
It remains unclear, however, whether the represenations and learning processes discovered by current models are similar to those used by the human visual system.

SURVEY

The human visual system makes highly effective use of limited information

A MIRC is defined as an image patch that can be reliably recognized by human observers and which is minimal in that fur- ther reduction in either size or resolution makes the patch un- recognizable (below

EXPERIMENT



We started from 10 grayscale images, each showing an object from a different class, and tested a large hierarchy of patches at different positions and decreasing size and resolution.

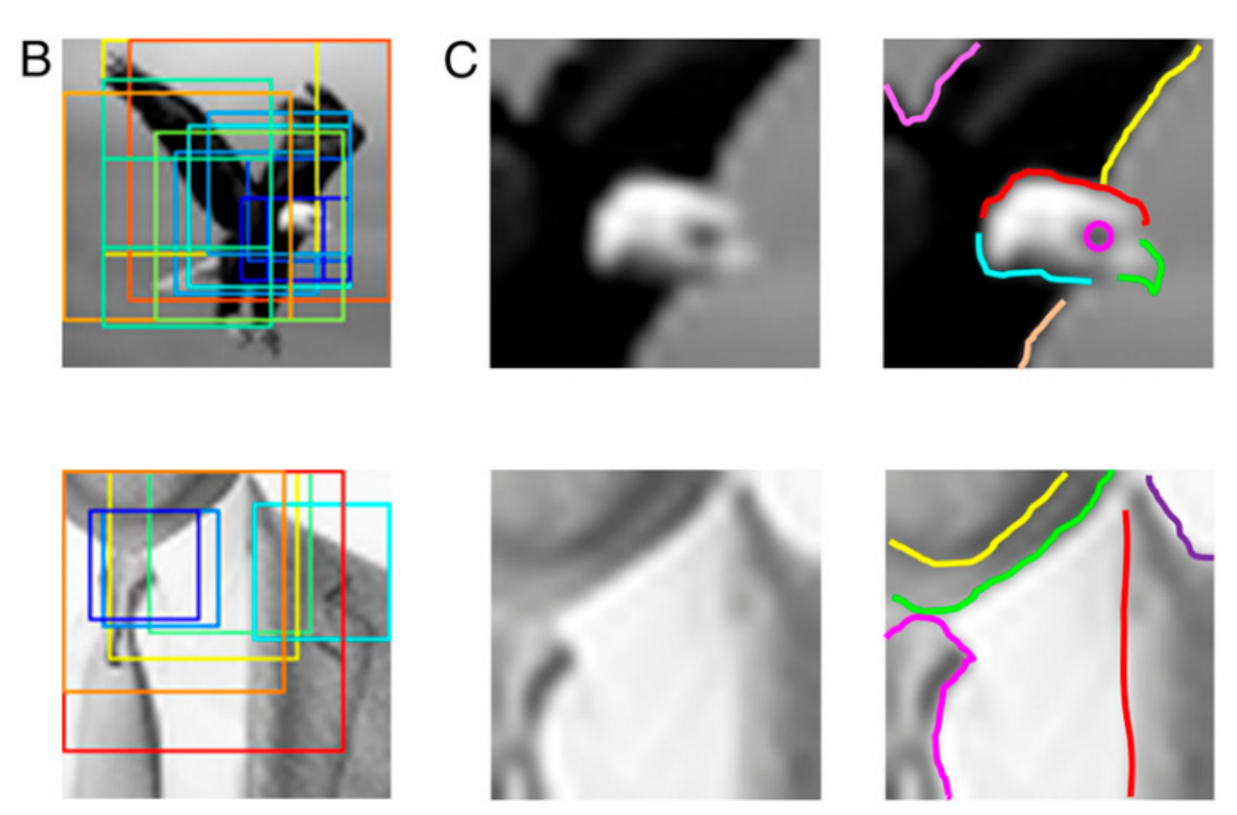
Each patch in this hierarchy has five descendants, obtained by either cropping the image or reducing its resolution.

NOTE:

A recognizable patch in this hierarchy is identified as a MIRC if none of its five descendants reaches a recognition criterion.

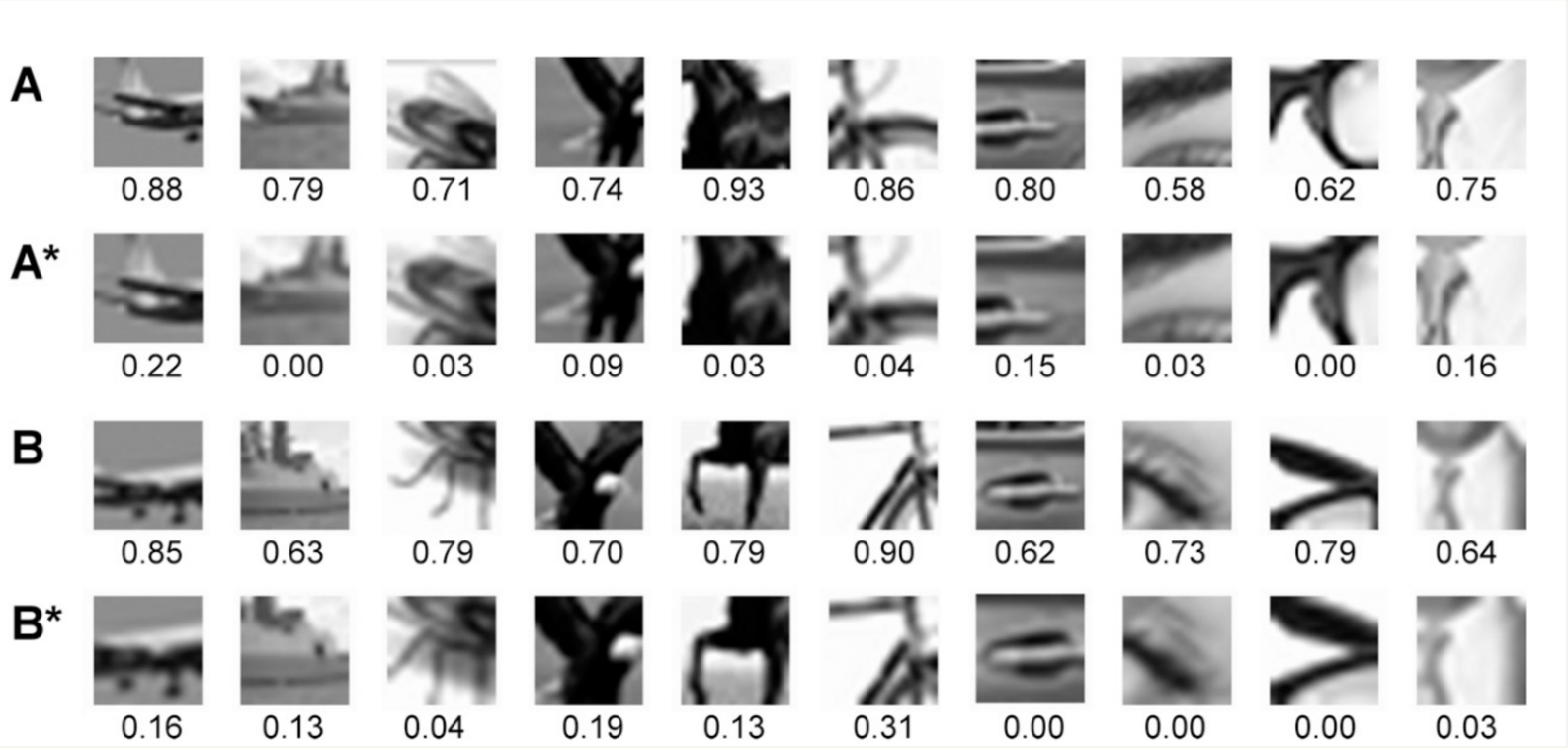
NOTE:

Each of the 10 original images was covered by multiple MIRCs at different positions and sizes.



Results

The transition in recognition rate from a MIRC image to a non-recognizable descendant (termed a “sub-MIRC”) is typically sharp: A surprisingly small change at the MIRC level can make it unrecognizable.

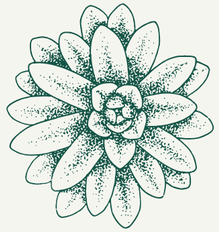


TESTING MODELS ARE:

- HMAX (10), a high-performing biological model of the primate ventral stream.
- the Deformable Part Model (DPM)
- support vector machines (SVM) applied to histograms of gradients (HOG) representations.
- extended Bag-of-Words (BOW)
- deep convolutional networks developed for classification of small images.

TRAINING MODELS ON FULL-OBJECT IMAGES

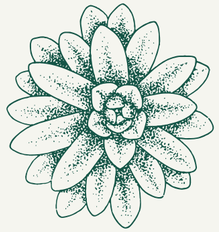
We computed the gap between MIRC and sub-MIRC recognition rates for the 10 classes and the different models and compared the gaps in the models' and human recognition rates.



None of the models came close to replicating the large drop shown in human recognition (average gap 0.14 ± 0.24 for models vs. 0.71 ± 0.05 for humans)

TRAINING MODELS ON FULL-OBJECT IMAGES

The difference between the models' and human gaps was highly significant for all com- puter-version models.



The gap is small because, for the models, the representations of MIRCs and sub-MIRCs are closely similar, and consequently the recognition scores of MIRCs and sub-MIRCs are not well separated.

NOTE:

In all models, the accuracy of MIRC recognition (AP 0.07 ± 0.10) was low compared with the recognition of full objects (AP 0.84 ± 0.19) and was still lower for sub-MIRCs

NOTE:

A conceivable possibility is that the performance of model networks applied to minimal images could be improved to the human level by increasing the size of the model network or the number of labeled training data.

Our tests suggest that although these possibilities cannot be ruled out, they appear unlikely to be sufficient.

In terms of network size, dou- bling the number of levels (see ref. 17 vs. ref. 18) did not improve MIRC recognition performance.

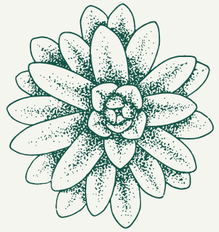
Regarding training examples, our testing included two network models (17, 18) that were trained previously on 1.2 million examples from 1,000 categories, including 7 of our 10 classes, but the recognition gap and accu- racy of these models applied to MIRC images were similar to those in the other models.

NOTE:

We also examined responses of intermediate units in the network models and found that results for the best performing intermediate layers were similar to the results of the network's standard top-level output.

TRAINING MODELS ON IMAGE PATCHES

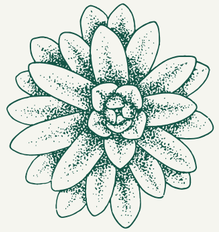
In a further test we simplified the learning task by training the models directly with images at the MIRC level rather than with full-object images.



After training, the models' accuracy in recognizing MIRC images was significantly higher than in learning from full- object images but still was low in absolute terms and in com- parison with human recognition (AP 0.74 ± 0.21 for training on patches vs. 0.07 ± 0.10 for training on full-object images)

TRAINING MODELS ON IMAGE PATCHES

The gap in recognition be- tween MIRC and sub-MIRC images remained low (0.20 ± 0.15 averaged over pairs and classifiers) and was significantly lower than the human gap for all classifiers.



After training, the models' accuracy in recognizing MIRC images was significantly higher than in learning from full- object images but still was low in absolute terms and in com- parison with human recognition (AP 0.74 ± 0.21 for training on patches vs. 0.07 ± 0.10 for training on full-object images)

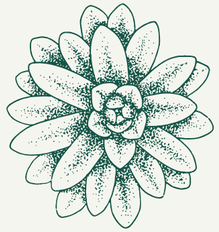
NOTE:

Although MIRCs are “atomic” in the sense that their partial images become unrecognizable, our tests showed that humans can consistently recognize multiple components internal to the MIRC.

Such internal interpretation is beyond the capacities of current neural network models, and it can contribute to accurate recognition, because a false detection could be rejected if it does not have the expected internal interpretation.

CONCLUSION

The results indicate that the human visual system uses features and processes that current models do not.



As a result, humans are better at recognizing minimal images, and they exhibit a sharp drop in recognition at theMIRC level, which is not replicated in models.

QUESTION:

An interesting open question is whether the ad- ditional features and processes are used in the visual system as a part of the cortical feed-forward process or by a top-down process, which currently is missing from the purely feed- forward computational models.

HYPOTHESIS:

The reason is that detailed interpretation appears to require features and interrelations that are relatively complex and are class-specific, in the sense that their presence depends on a specific class and location

HYPOTHESIS:

the initial activation of class candidates, which is incomplete and with limited accuracy. The activated representations then trigger the application of class- specific interpretation and validation processes, which recover richer and more accurate interpretation of the visible scene.

THANK YOU