an interesting/influential/important paper from the world of CS every weekday morning, as selected by Adrian Colyer

Measuring the tendency of CNNs to learn surface statistical regularities

MAY 29, 2018

tags: Deep Learning

Measuring the tendency of CNNs to learn surface statistical regularities (https://arxiv.org/abs/1711.11561) Jo et al., *arXiv'17*

With thanks to Cris Conde for bringing this paper to my attention.

We've looked at quite a few adversarial attacks on deep learning systems in previous editions of The Morning Paper. I find them fascinating for what they reveal about the current limits of our understanding.

> *…humans are able to correctly classify the adversarial image with relative ease, whereas the CNNs predict the wrong label, usually with very high confidence. The sensitivity of high performance CNNs to adversarial examples casts serious doubt that these networks are actually learning high level abstract concepts. This begs the following question: How can a network that is not learning high level abstract concepts manage to generalize so well?*

In this paper, Jo and Bengio conduct a series of careful experiments to try and discover what's going on. The initial hypothesis runs like this:

- There are really only two ways we could be seeing the strong generalisation performance that we do. Either (a) the networks are learning high level concepts, or (b) there may be a number of *superficial cues* in images that are shared across training and test datasets, and the networks are learning these instead.
- We have reason to doubt scenario (a) because the success of adversarial images
- We have some reasons to believe scenario (b) given results in the computer vision literature that show a strong statistical relationship between image statistics and visual understanding. These suggest that computer vision algorithms may "*…lean heavily on background features to perform categorization.*" (For example, cars tend to be on roads).
- If scenario (b) is true, then we ought to see a drop in generalisation performance if we manipulate the image statistics in such a way that they are appreciably different, and yet

humans can still easily recognise the target object.

> *When the training and the test set share similar image statistics, it is wholly possible for a machine learning model to learn superficial cues and generalize well, albeit in a very narrow sense as they are highly dependent on the image statistics. Adversarial examples would be destroying the superficial cues. We believe that this is precisely how deep CNNs can attain record breaking generalization performance on all sorts of natural image tasks, and yet can be so sensitive to adversarial perturbations.*

To gather evidence in favour of this hypothesis, we need to find a perturbation function over a dataset such that:

1. Object recognisability is preserved from the perspective of a human
2. The perturbed images exhibit qualitatively different image statistics.
3. When trained on either (but not both!) of the original or perturbed images, and then tested against both the original and perturbed images, we see a non-trivial generalisation gap between the generalisation capability on the test images with similar statistics, and the the generalisation capability on the test images with different statistics.

Conditions (1) and (2) together guarantee that the original and perturbed datasets share the same high level abstractions but exhibit different superficial cues.

Fourier filtering to the rescue

> *While natural images are known to exhibit a huge variance in the raw pixel space, it has been shown that the Fourier image statistics of natural images obey a power law decay… An immediate takeaway is that natural images tend to have the bulk of their Fourier spectrum concentrated in the low to medium range frequencies. Due to this power law concentration of energy in the frequency space, it is possible to perform certain types of Fourier filtering and preserve much of the perceptual content of an image.*

The authors experiment with two different kinds of Fourier filters: a *low-frequency filter* that uses a radial mask in the Fourier domain to set higher frequency modes to zero, and a *random filter* that uses a uniformly random mask to set a Fourier mode to zero with probability p.

Here are some example images from the SVHN dataset (top row), and the results of applying the radial and random masking filters respectively:

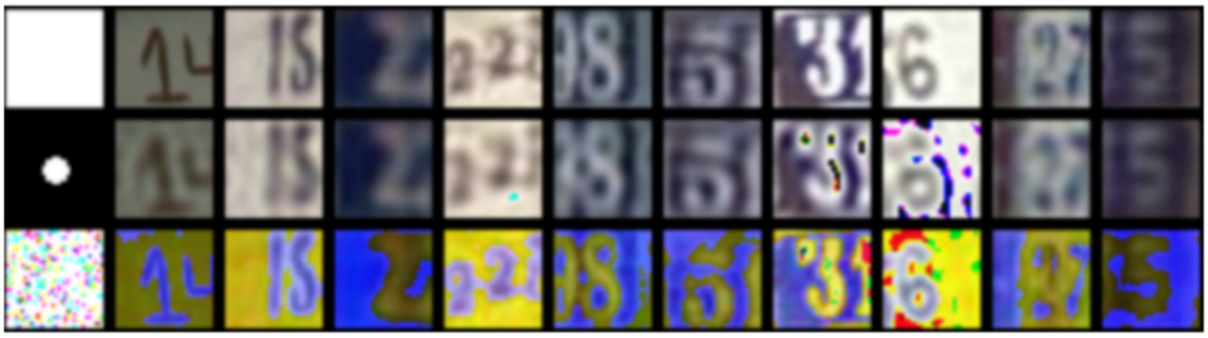


Figure 1: The first image in each column corresponds to the Fourier mask in frequency space. A white pixel corresponds to preserving the Fourier mode, black/color corresponds to setting it to zero. **Top row**: No Fourier filtering applied, original SVHN images. **Middle row**: Radial mask and the corresponding filtered images. **Bottom row**: Random mask and the corresponding filtered images. Best viewed in color.

And this is the same thing, but for the CIFAR-10 dataset:

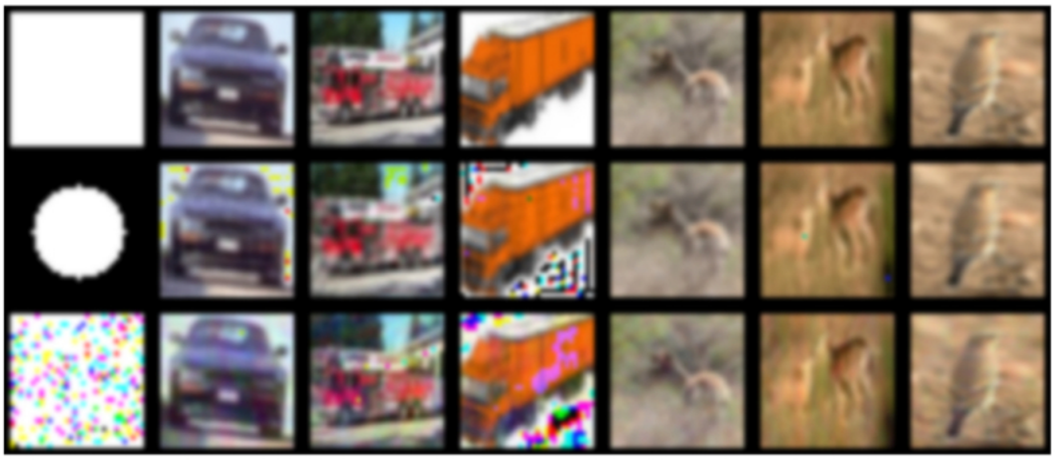


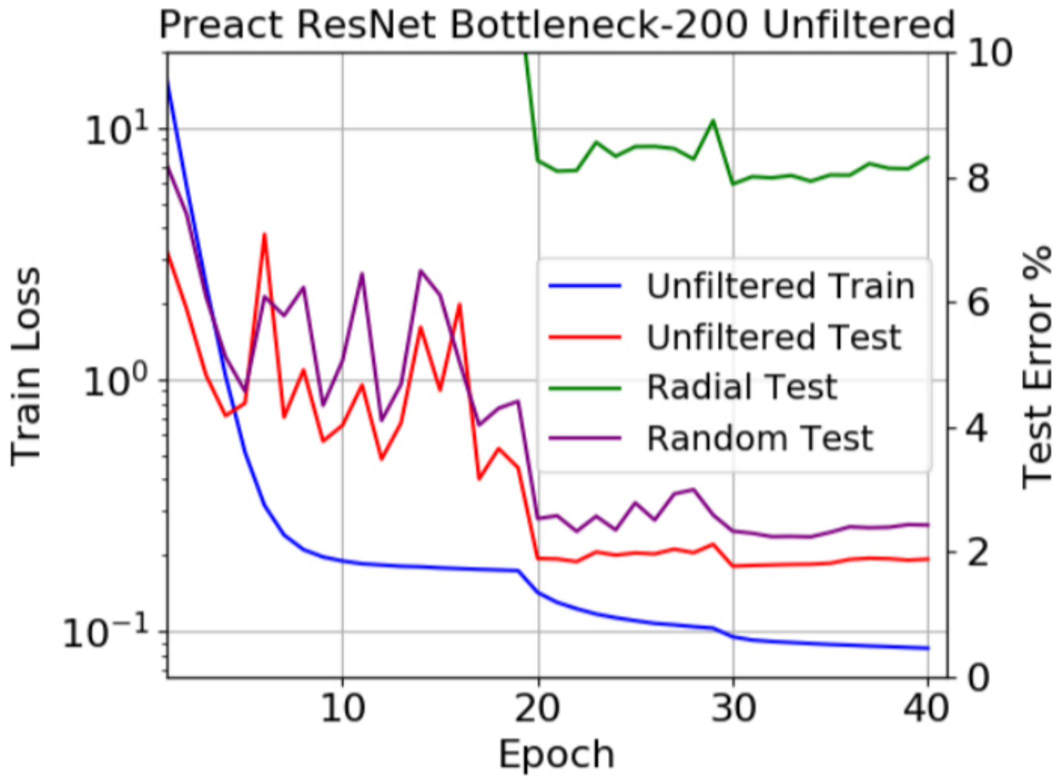
Figure 2: The first image in each column corresponds to the Fourier mask in frequency space. A white pixel corresponds to preserving the Fourier mode, black/color corresponds to setting it to zero. **Top row**: No Fourier filtering applied, original CIFAR-10 images. **Middle row**: Radial mask and the corresponding filtered images. **Bottom row**: Random mask and the corresponding filtered images. Best viewed in color.

The filters do introduce artifacts into the images, but these don't really interfere with human perception, and tend to occur in the background of the image.

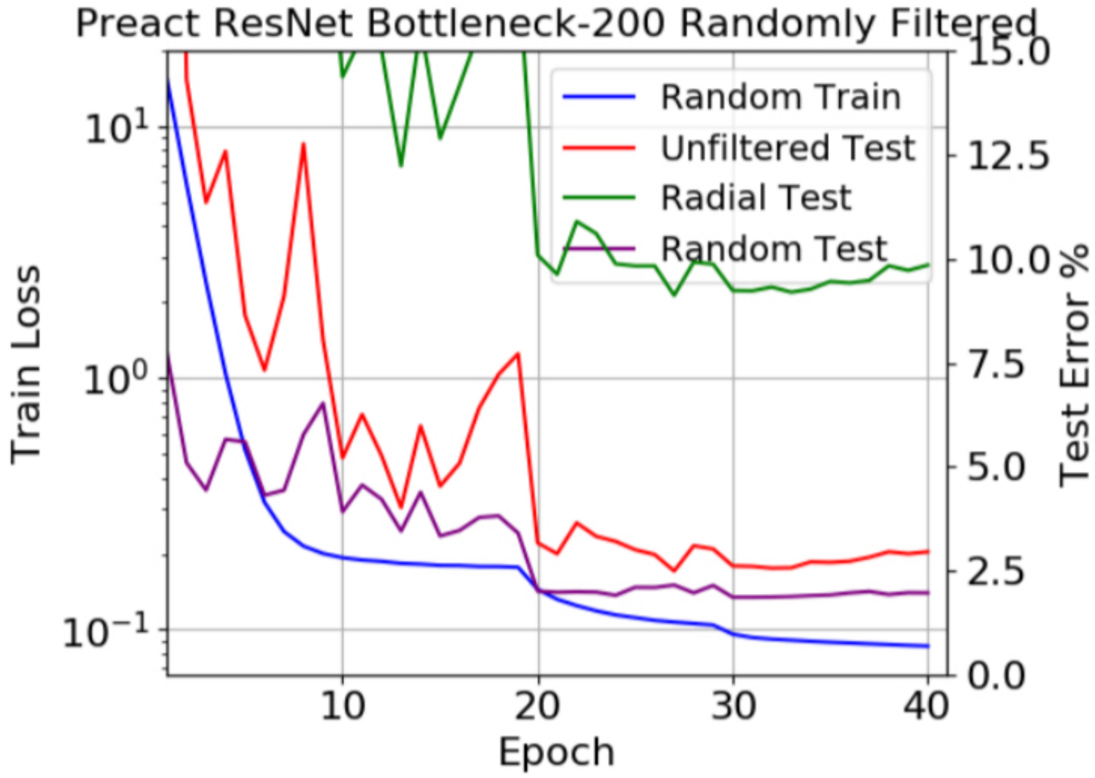
Running the experiment

For both the CIFAR-10 and SVHN datasets the authors use some established high-performance CNN architectures (Preact ResNet). A model is trained on one of the unfiltered, radial mask, or random mask datasets, and then tested across all three test groups (unfiltered, radial, random). This enables us to measure the *test gap* or *generalisation gap* as the maximum difference in accuracy across the test sets.

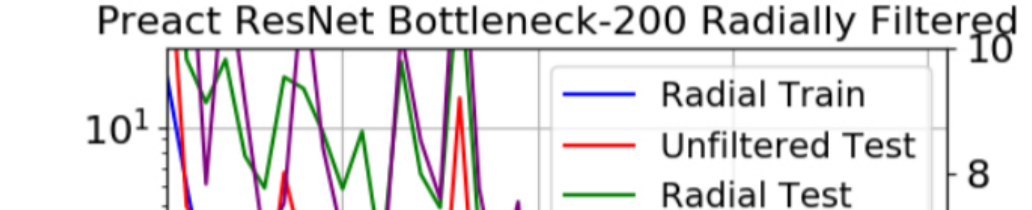
For the SVHN source dataset, the three figures below show the generalisation results after training on (a) unfiltered, (b) random, and (c) radial.

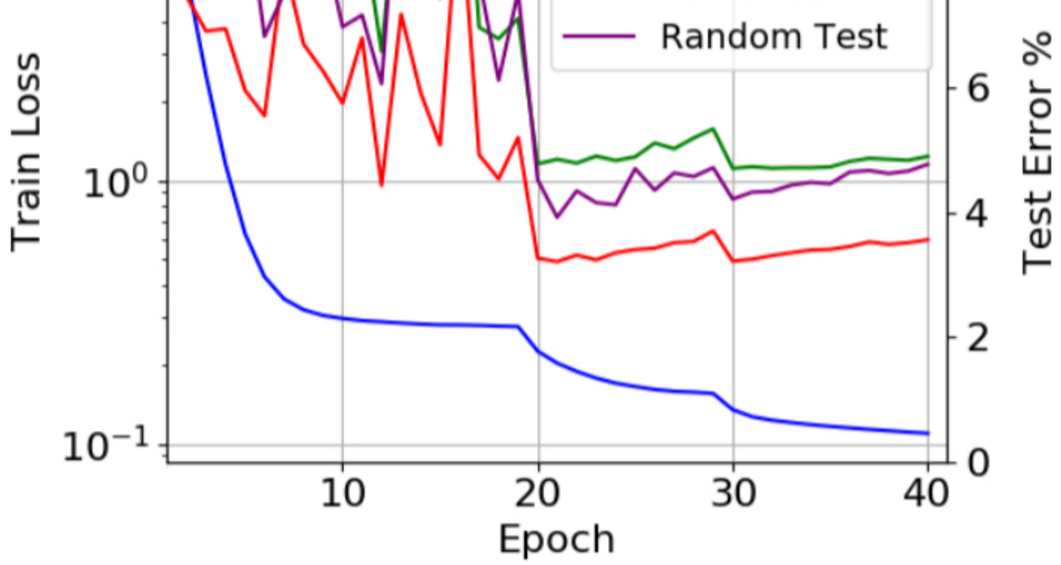


(a) Trained on Unfiltered SVHN



(b) Trained on Randomly Filtered SVHN

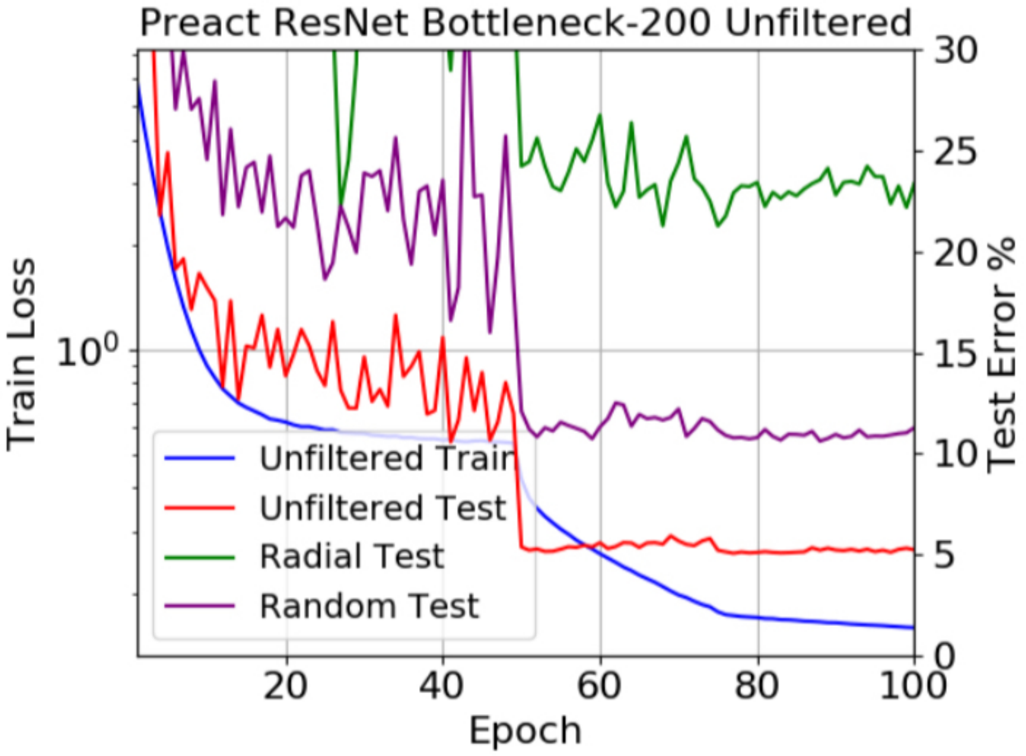




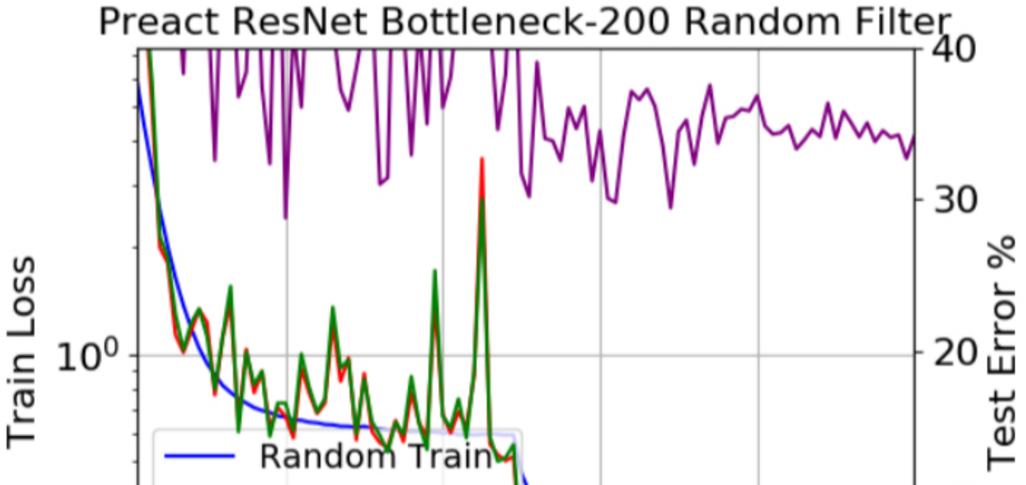
(c) Trained on Radially Filtered SVHN

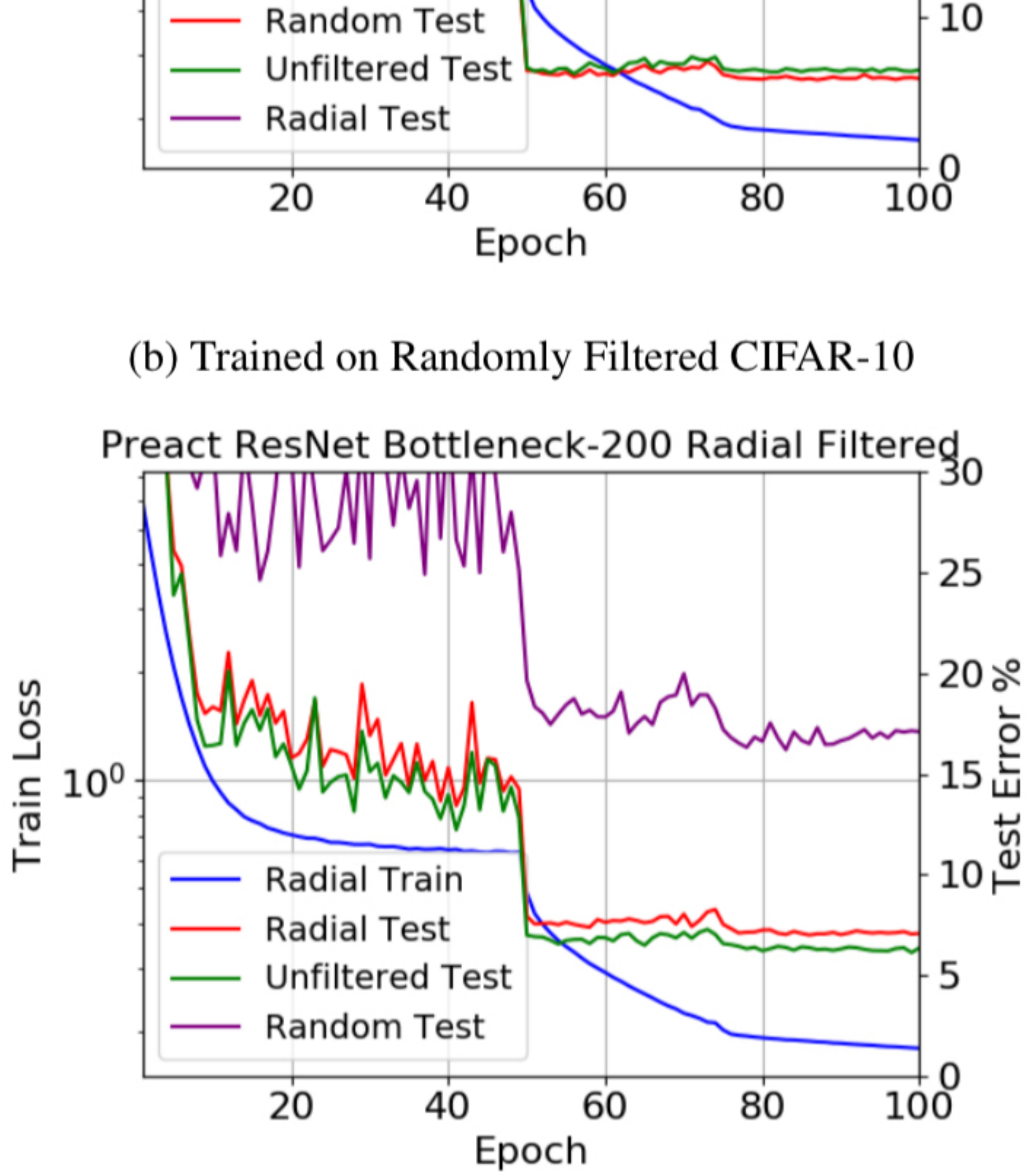
Figure 3: Generalization plots for Preact-ResNet-Bottleneck-200 model. **(a)** Trained on unfiltered SVHN. **(b)** Trained on randomly filtered SVHN data. **(c)** Trained on radially filtered SVHN data. Best viewed in color.

And here's the same thing for CIFAR-10:



(a) Trained on Unfiltered CIFAR-10





(b) Trained on Randomly Filtered CIFAR-10

(c) Trained on Radially Filtered CIFAR-10

Figure 6: Generalization plots for Preact-ResNet-Bottleneck-200 model. **(a)** Trained on unfiltered CIFAR-10 data. **(b)** Trained on randomly filtered CIFAR-10 data. **(c)** Trained on radially filtered CIFAR-10 data. Best viewed in color.

With SVHN the largest generalisation gap (7.89%) occurs when training on randomly filtered data, and testing on the radially filtered dataset. Training on the radially filtered dataset actually turned out to *improve* generalisation performance on the unfiltered test set by 1.5%. The authors put this down to a regularisation effect. Changing the network *depth* seemed to have little impact on the generalisation gap. That is, there is no evidence of an ability to more successfully learn higher level abstractions when we add more data.

With CIFAR-10 the generalisation gaps are bigger: 28% when trained on randomly filtered data tested on radially filtered. Changing the depth has little impact in this case too.

Discussion

In addition to still being recognisable to the human eye (a subjective measure), we know also know that networks trained on the filtered datasets actually generalised quite well to the unfiltered test set (off by 1-2% of the best unfiltered accuracy). This provides further evidence that the choice of Fourier filtering schemes is producing datasets that are perceptually not too far off from the original unfiltered dataset.

> *We see that deep CNNs trained on a unfiltered natural image dataset exhibit a tendency to latch onto the image statistics of the training set, yielding a non-trivial generalization gap…*

When training on *all three* training sets (unfiltered, radial, and random) there is an improvement in the generalisation gap. "*However, we cast doubt on the notion that this sort of data augmentation scheme is sufficient to learn higher level semantic features in the dataset. Rather it is far more likely that the CNNs are learning a superficial robustness to the varying image statistics.*"

Overall, the data seems to support the initial hypothesis: "**the current incarnation of deep neural networks have a tendency to learn surface statistical regularities as opposed to high level abstractions**."

What are we to do? I'm reminded of the brightly coloured word books we teach our children with – often with a single strong cartoon-like image of an object, on a plain white background, and the label (word) underneath. By seeing real world objects with movement (either of our own point of view, or because they have independent motion) we are also forced to more strongly separate object from background.

The authors cite six pieces of work which they believe represent promising new directions:

> "**Independently controllable factors** (https://arxiv.org/abs/1708.01289)" and
> "**Reinforcement learning with unsupervised auxiliary tasks (https://deepmind.com/blog/reinforcement-learning-unsupervised-auxiliary-tasks/)**" aim to learn good disentangled feature representations by combining unsupervised and reinforcement learning.
> "**SCAN: learning abstract hierarchical compositional visual concepts (https://arxiv.org/abs/1707.03389)**" uses a variational setup.
> "**Discovering objects and their relations from untangled scene representations (https://arxiv.org/abs/1702.05068)**" aims to learn abstract relations between objects in natural scene images
> "**The consciousness prior** (https://arxiv.org/abs/1709.08568)" moves away from making predictions in the perceptual space, and instead operates in the higher-order abstract space.

from → Uncategorized

4 Comments leave one →

1. **saiboorlagadda PERMALINK**
   May 29, 2018 7:14 am
   Typo: human eye (a subjective measure), we know also know

   Should be: human eye (a subjective measure), we now also know

   REPLY
2. **Ted Sanders PERMALINK**
   May 29, 2018 8:43 am
   Your blog is excellent. Thanks for all your effort and thoughtfulness!

   REPLY

3. **Vishal Kasliwal PERMALINK**
   May 29, 2018 8:40 pm
   Are you planning on discussing each of the 5 ‘new directions’ papers? I ask because I’m very keen on going through all 5 but you end up saving me lots of time when you pre-digest them.

   REPLY

   - **adriancolyer PERMALINK***
     May 30, 2018 6:38 am
     That would make a very good week of papers! I probably will try to fit these into the schedule at some point, but it’s not imminent – there are some really interesting papers from EuroSys I’ll be moving onto next week…

     REPLY

This site uses Akismet to reduce spam. Learn how your comment data is processed.