Selectivity metrics provide misleading estimates of the selectivity of single units in neural networks

Ella M. Gale, Ryan Blything, Nicholas Martin & Jeffrey S. Bowers
(ella.gale, ryan.blything, nm13850, j.bowers@bristol.ac.uk)
School of Psychological Science, University of Bristol, 12a Priory Road Bristol BS8 1TU, UK

Anh Nguyen (anhnguyen@auburn.edu)
Department of Computer Science and Software Engineering, Auburn University, AL, USA

Abstract

To understand the representations learned by neural networks (NNs), various methods of measuring unit selectivity have been developed. Here we undertake a comparison of four such measures on AlexNet: localist selectivity (Bowers et al., 2014); precision (Zhou et al., 2015); class-conditional mean activity selectivity CCMAS (Morcos et al., 2018); and top-class selectivity. In contrast with previous work on recurrent neural networks (RNNs), we fail to find any 100% selective 'localist units' in AlexNet, and demonstrate that the precision and CCMAS measures are misleading and suggest a much higher level of selectivity than is warranted. We also generated activation maximization (AM) images that maximally activated individual units and found that under (5%) of units in fc6 and conv5 produced interpretable images of objects, whereas fc8 produced over 50% interpretable images. Furthermore, the interpretable images in the hidden layers were not associated with highly selective units. We also consider why localist representations are learned in RNNs and not AlexNet.

Keywords: localist representation; grandmother cells; distributed representations.

Introduction

There have been recent attempts to understand how neural networks (NNs) work by analyzing hidden units one at a time using various measures such as localist selectivity (Bowers et al., 2014), class-conditional mean activity selectivity (CCMAS) (Morcos et al., 2018), precision (Zhou et al., 2015), and activation maximization (AM) (Erhan et al., 2009). These measures are defined below, and they all provide evidence that some units respond selectively to categories under some conditions.

Our goal here is to directly compare different measures of object selectivity on a common network trained on a single task. We chose AlexNet (Krizhevsky et al., 2012) because it is a well-studied CNN and many authors have reported high levels of selectivity in its hidden layers via both quantitative (Zhou et al., 2018, 2015) and qualitative methods using Activation Maximization (AM) images (Nguyen et al., 2017; Yosinski et al., 2015; Simonyan et al., 2013). Our main findings are:

1. The different measures provide very different estimates of selectivity.
2. The precision and CCMAS measures are misleading with near perfect selectivity scores associated with units that strongly respond to many different image categories. CCMAS scores are also ambiguous, as explained below.
3. There are no localist 'grandmother cell' representations in AlexNet, in contrast with the localist representations learned in some RNNs.
4. Units with interpretable AM images do not necessarily correspond to highly selective representations.
5. New selectivity measures are required to provide a better assessment of the learned hidden representations in NNs.

Bowers et al. (2014, 2016) assessed the selectivity of hidden units in recurrent NNs using networks similar to those developed by Botvinick & Plaut (2006) designed to explain human short-term memory performance. They reported many 'localist' units that are 100% selective for specific letters or words, where all members of the selective category were more active than and disjoint from all non-members, as can be shown in jitterplots (Berkeley et al., 1995), see Fig. 1 for a unit selective to the letter 'j').

These localist representations were compared to 'grandmother cells' as discussed in neuroscience (Bowers, 2017a). Bowers et al. (2014) argued that the network learned these representations in order to co-activate multiple letters or words at the same time in short-term memory without producing ambiguous blends of overlapping distributed patterns (the so-called 'superposition catastrophe'). Consistent with this hypothesis, localist units did not emerge when the model was trained on letters or words one-at-a-time (a condition in which the model did not need to overcome the superposition catastrophe (Bowers et al., 2014)), see Fig. 1 for an example of a non-selective unit)

In parallel, researchers have reported selective units in the hidden layers of various CNNs trained to classify images into one of multiple categories ((Zhou et al., 2015; Morcos et al., 2018; Zeiler & Fergus, 2014; Erhan et al., 2009), for a review see (Bowers, 2017a)). For example, Zhou et al. (2015) assessed the selectivity of units in the pool5 layer of two CNNs trained to classify images into 1000 objects and 205 scene categories, respectively. They reported multiple 'object detectors' (as defined below) in both networks. Similarly, Morcos et al. (2018) reported that CNNs trained on CIFAR-10 and ImageNet learned many highly selective hidden units, with CCMAS scores often approaching the maximum of 1.0.

Note that these later studies show that selective representations develop in CNNs trained to classify images one-at-

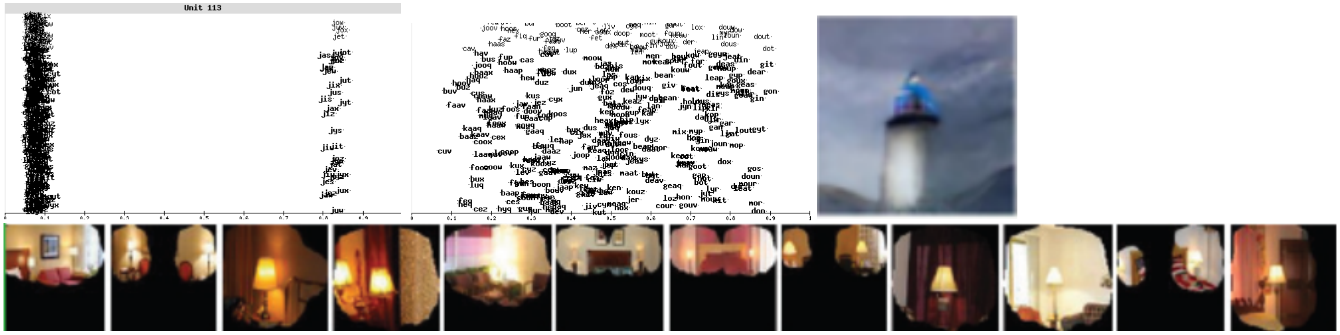


Figure 1: Examples of selectivity measures used. Top left: jitterplot of unit 113 in an RNN under the superposition constraint selective the letter 'j'. Top middle: jitterplot of non-selective unit 160 found when RNN trained on words one-at-a-time; from (Bowers et al., 2016). Top right: activation maximization (AM) image of a unit in conv5 of AlexNet that looks like a lighthouse; from (Nguyen et al., 2016). Bottom: highest activation images for a 'lamp' detector with 84% precision in layer pool5 of AlexNet; from (Zhou et al., 2015).

a-time. This appears to be inconsistent with Bowers et al. (2016) who (a) failed to obtain selective representations for letters or words under these conditions (see Fig. 1); and (b) it suggests that there are additional pressures for CNNs to learn selective representations above and beyond the challenge of overcoming the superposition catastrophe. However, the measures of selectivity that have been applied across studies are different, and accordingly, it is difficult to directly compare results.

In order to directly compare and have a better understanding of the different selectivity measures we assessed (1) localist, (2) precision, and (3) CCMAS selectivity on the prob, fc8, fc7, fc6, and conv5 layers of AlexNet. We also introduce a new measure called top-class selectivity, and show that the precision and CCMAS measures provide much higher estimates of object selectivity compared to other measures. Importantly, we do not find any localist 'grandmother cell' representations in the hidden layers of AlexNet, consistent with the hypothesis that the superposition catastrophe provides a pressure to learn more selective representations (Bowers et al., 2014, 2016).

In addition, we compared these selectivity measures to a state-of-the-art activation maximization (AM) method for visualizing single-unit representations in CNNs (Nguyen et al., 2017). AM images are generated to strongly activate individual units, and some of them are interpretable by humans (e.g., a generated image that looks like a lighthouse, see Fig. 1). For the first time, we systematically evaluated the interpretability of the AM images in an on-line experiment and compare these ratings with the selectivity measures for corresponding units. We show that hidden units with interpretable AM images are not highly selective.

It is important to emphasize that these different measures have all been used to provide insights into the same set of issues. For example, both interpretability of generated images (Le et al., 2011) and localist selectivity (Bowers et al., 2014) have been used to make claims about 'grandmother cells'. The different measures have also been directly compared to one another. For example, Zhou et al. (2015) claim that the object detectors learned in CNNs play an important role in identifying specific objects, whereas Morcos et al. (2018) challenge this conclusion based on their finding that units with high CCMAS measures were not especially important in the performance of their CNNs. Indeed, based on the finding that high CCMAS scores were not predictive of performance, Morcos et al. wrote: "...it implies than methods for understanding neural networks based on analyzing highly selective single units, or finding optimal inputs for single units, such as activation maximization (Erhan et al., 2009) may be misleading". This makes a direct comparison between measures all the more important.

Methods

Networks and Datasets All ~1.2M photos from ImageNet2010 (Deng et al., 2009) were cropped to 277 × 277 pixels and classified by the pre-trained AlexNet CNN (Krizhevsky et al., 2012) shipped with Caffe (Jia et al., 2014), resulting in 721,536 correctly classified images. Once classified, the images are not re-cropped nor subject to any changes. In Caffe, the softmax operation (Denker & leCun, 1991) is applied at the 'prob'(ability) output layer that contains 1000 units (one for each class). We analyzed these prob units, the fully connected (fc) layers: fc8 (1000 units) that encodes the outputs prior to the softmax operation, fc6 and fc7 (4096 units), and the top convolutional layer conv5 which has 256 filters. We only recorded the activations of correctly classified images. The activation files are stored in .h5 format and can be retrieved at `https://bristol.codersoffortune.net/AlexNet_Merged/`. We selected 233 conv5, 2738 fc6, 2239 fc7, 911 fc8, and 954 prob units for analysis.

Localist selectivity Here we define a unit to be localist for class A if the set of activations for class A was disjoint with those of not A ($\neg A$).

Localist selectivity is easily depicted with jitterplots in which a scatter plot for each unit is generated (see Figs. 3 and 4). Each point in a plot corresponds to a unit's activation in response to a single image, and only correctly classified images are plotted. The level of activations is coded along the *x*-axis, and an arbitrary value is assigned to each point on the *y*-axis (they are jittered).

Top-Class selectivity Top-class selectivity is related to localist selectivity except that it provides a continuous rather than discrete measure. We counted the number of images from the same class that were more active than all images from all other classes (what we called the top cluster size) and divided the cluster size by the total number of correctly identified images from this class. 100% top-class selectivity is equivalent to a localist representation.

Precision The precision method of finding object detectors (Zhou et al., 2015, 2018) involves identifying a small subset of images that most strongly activate a unit and then identifying the critical part of these images that are responsible for driving the unit. Zhou et al. (2015) took the 60 images that activated a unit the most strongly and asked independent raters to interpret the critical image patches. Zhou et al. (2015) developed a precision metric that calculated the percentage of the 60 images that raters judged to depict the same class of object (e.g., if 50 of the 60 images were labeled as 'lamp', the unit would have a precision index of 50/60 or 83%; see Fig. 1). Object detectors were defined as units with a precision $>$ 75%: they reported multiple such detectors. Here we approximate this approach by considering the 100 images that most strongly activate a given unit and assess the highest percentage of images from a given output class.

CCMAS Morcos et al. (2018) introduced a selectivity index based on the 'class-conditional mean activation' selectivity (CCMAS). The CCMAS for class *A* compares the mean activation of all images in class *A*, μ_A, with the mean activation of all images not in class *A*, $\mu_{\neg A}$, and is given by: $(\mu_A - \mu_{\neg A}) / (\mu_A + \mu_{\neg A})$. Morcos et al. (2018) states that this metric should vary within [0,1], with 0 meaning that a unit's average activity was identical for all classes, and 1 meaning that a unit was only active for inputs of a single class. Here, we assessed class selectivity for the highest mean activation class (CCMAS) as well as for the class with the second highest mean activation μ_A (what we call CCMAS_2) in order to assess the extent to which CCMAS reflects the selectivity to one class.

Activation Maximization We harnessed an activation maximization method called Plug & Play Generative Networks (Nguyen et al., 2017) in which an image generator network was used to generate images (hereafter, AM images) that highly activate a unit. We generated 100 separate images that maximally activated each unit in the conv5, fc6 and fc8 layers of AlexNet and displayed them in a grid format. We then asked 333 participants to judge whether they could identify any repeating objects, animals, or places in images after receiving some practice trials. Participants were recruited using Prolific (*Attrition*, n.d.; Palan & Schitter, 2018), with the experiment run online using gorilla (*Gorilla Experiment Builder*, n.d.). Readers can test themselves at: https://research.sc/participant/login/dynamic/63907FB2-3CB9-45A9-B4AC-EFFD4C4A95D5.

Results

Comparison of selectivity measures.

The mean top-class, precision, and CCMAS selectivities across the conv5, fc6 and fc7 layers are displayed in Fig. 2a–c. We did not plot localist selectivity as there were no localist 'grandmother units' at any internal level (and only 10% at the prob layer, due to the softmax function). The first point to note is that the top-class, precision, and CCMAS measures all increased in the higher layers, showing that they capture degrees of selectivity ignored by the localist measure. Second, the top-class selectivity was extremely low across the hidden layers, with means below 0.25% in the the conv5, fc6, and fc7 layers. Third, the different measures provided very different estimates of selectivity. In contrast with top-class selectivity, the mean precision scores are over an order of magnitude larger in the hidden layers of network, with average precision scores of 9.6%, 12.1%, and 15.4% in layers conv5, fc6, and fc7, respectively. Similarly, the CCMAS measure suggests a much higher level of selectivity than top-class selectivity, with mean scores of .49, .84, and .85 in the conv5, fc6, and fc7 layers, respectively.

This discrepancy is most striking for the units with the highest precision and CCMAS scores. For example, in Fig. 3 we illustrate why the unit fc6.1199 with the highest precision (95%) in fc6 should not be considered a Monarch butterfly detector. Fig. 3a depicts a jitterplot of activations to all accurately identified images, with Monarch butterfly images found across the range of activations. Fig. 3b shows a histogram that plots the distribution of activations for Monarch butterflies. By far the most common activation to correctly identified Monarch butterflies is 0 and the mean is 39.2±0.6. Figures 3 displays example images with 0 (right top), mean (right middle) and maximal (right bottom) activations, and all are identifiable as Monarch butterflies. Thus, classifying this unit as a Monarch butterfly detector is misleading.

Another surprising result is that we did not obtain any 100% top-class selectivity units (localist units) in the prob layer of AlexNet. Rather, the mean top-class selectivity was ∼80% in the prob layer, and only ∼5% in fc8 (prior to the softmax being applied). Fig. 4 depicts the pattern of activation for units fc8.11 and prob.11 that are examples of the most top-class selective units in these layers (responding to images of 'goldfinch' birds with top-class selectivity measures of 8.4% and 95.2%, respectively). Clearly these units do respond much more selectively than the most selective units in earlier layers (*c.f.* Fig. 3), and at the same time, the jitterplots show why we did not observe any localist units (a few 'goldfinch' images were less active than a few images from other categories).

These jitterplots also show that top-class and localist se-

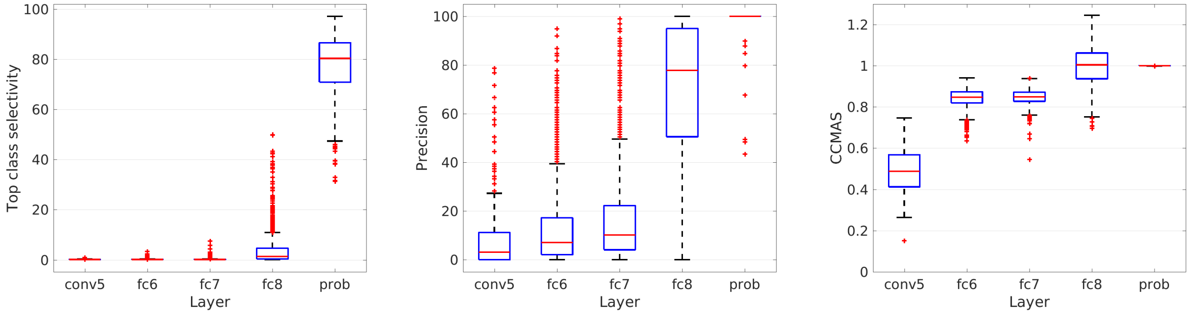


Figure 2: Selectivity measures across different layers of AlexNet. Left: top-class selectivity. Middle: precision 100 (the percentage of the top 100 images which are members of the top class). Right: Class-conditional mean activity selectivity (CCMAS), N.B. as the mean of the unselected classes ($\mu_{\neg A}$ can be less than zero) the CCMAS can go above its expected maximum of 1.

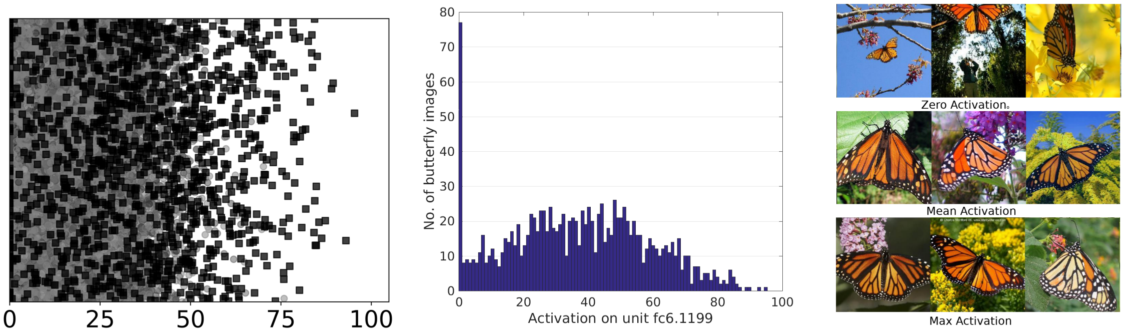


Figure 3: Data for unit fc6.1199. Left: activation jitterplot: black squares: Monarch butterfly images; grey circles: all other classes. Middle: histogram of activations of Monarch butterflies. Right: example ImageNet images with activations of 0.0 (top), the mean (middle), and the maximum (bottom) of the range. Unit fc6.1199 has a precision of 95% over the top 100 images (98.3% over the top 60) and is thus classified as a butterfly detector, yet there are Monarch butterfly images covering the whole range of values, with 72 images (5.8% of the total) having an activation of 0.0.

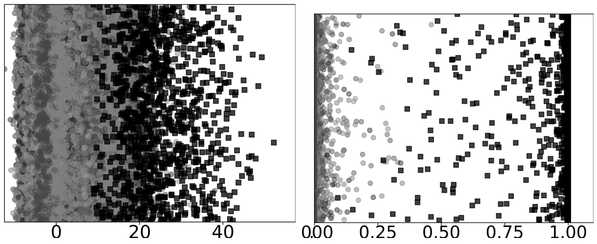
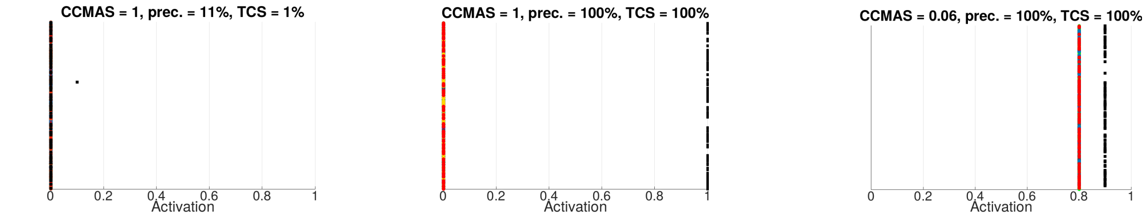


Figure 4: Example data from the fc8 and prob layers. Left: jitterplot activations for unit fc8.11 that has a top-class selectivity of 8.4%. Right: jitterplot activations for prob.11 (i.e. post-softmax) that has top-class selectivity of 95.2%. Activations for the 'ground truth' class 'goldfinch' are shown as black squares, all other classes are shown as greyscale circles.

lectivity provide somewhat misleading estimates of selectivity as well. Consider Fig. 4(left) that depicts a substantial overlap between goldfinch and non-goldfinch activations on unit fc8.11. The 8.4% top-class selectivity score captures the selectivity for the most highly active goldfinch images, but it is blind to the fact that almost all goldfinch images have a reasonably high level of activation (more than most non-goldfinch images). The problem with localist selectivity is highlighted in Fig. 4(right) that shows that the measure misses all but the most extreme version of selectivity. Together, these findings suggest that new selectivity measures are required to better characterize the representations in NNs: precision and CCMAS measures strongly overestimate selectivity, and localist and top-class selectivity provide either a too strict or too narrow a measure of selectivity.

Additional problems with the CCMAS measure

The main problem with the precision and CCMAS measures is that they provide misleadingly high estimates of selectivity, but the CCMAS measure has some additional limitations. First, if the CCMAS provided a good measure of a unit's class selectivity then one should expect that a high measure of selectivity for one class would imply that the unit is not highly selective for other classes. However, the CCMAS score for the most selective category and the second most selective category CCMAS_2 were similar across the conv5, fc6 and fc7



a. One active item from one class. CCMAS = 1, precision = 11%, TCS = 1%.

b. Archetypal ‘grandmother’ unit. CCMAS = 1, precision = 100%,TCS = 100%.

c. One class more active than the others. CCMAS = 0.06, precision = 100%, TCS = 100%.

Figure 5: Example of where the CCMAS does not match intuitive understandings of selectivity. Generated example data: (a) If a unit is off to all but a single image from a large class of objects, the CCMAS for that class is 1 (maximum possible selectivity). (b) If a unit is strongly activated to all members of one class and off to everything else (an archetypal ‘grandmother’ cell) the CCMAS is the same as for (a) although the precision and top-class selectivity is vastly different. (c): If a unit has high activations for all classes, but one class (black squares) is 0.1 more than all others (coloured circles), the CCMAS is very low (0.06) despite being %100 top-class selective. The generated examples are for 10 classes of 100 items

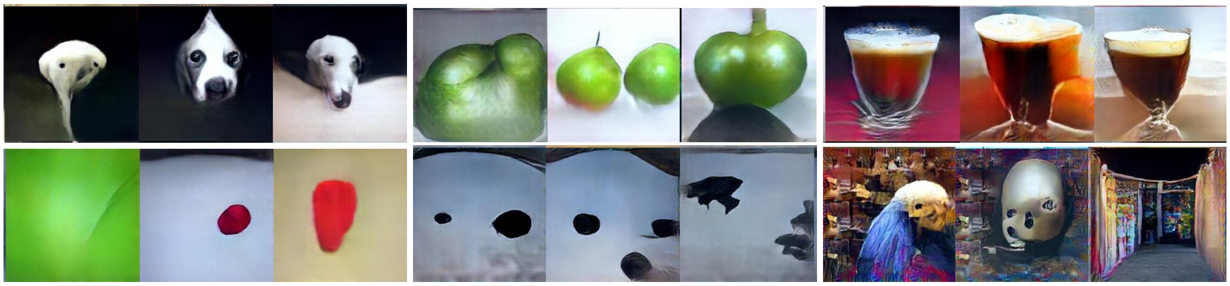


Figure 6: Example AM images that were either judged by all participants to contain objects (top row) or judged by all participants to be uninterpretable as objects (bottom row). The human judgement for conv5.183 (top left) was ‘dogs’ and the top-class was ‘flat-coated retriever’. For fc6.319 (top middle) subjects reported ‘green peppers’ or ‘apples’ (all classified as the same broad class in our analysis), and the CCMAS and top-class was ‘Granny Smith apples’. For fc8.969 (top right) humans suggested ‘beverage’ or ‘drink’: ground truth class for this unit is ‘eggnog’. The ground-truth for fc8.865 (bottom right) is‘toy-store’.

layers, with the mean CCMAS scores .491, .844, and .848, and the CCMAS_2 scores .464, .821, .831. For example, unit fc7.0 has a CCMAS of .813 for the class ‘maypole’, and a CCMAS_2 score of .808 for ‘chainsaw’ (with neither of these categories corresponding ‘orangutan’ that had the highest precision of score of 14% and a top-class selectivity score of .001%).

Second, the CCMAS measure provides an ambiguous measure of selectivity. To illustrate, consider the artificial scatter plots depicted in Figs. 5a,b. Here we obtain the same perfect CCMAS scores for one unit that selectively responds to one member of a category and another unit that selectively responds to all members of a category. This is problematic for a measure designed to assess *class* selectivity. Third, as shown in Fig. 5c, it is even possible to have a low CCMAS score for a unit with 100 percent top-class selectivity (that is, a low CCMAS selectivity for a grandmother cell). Together, these characteristics of the CCMAS measure may help explain why why Morcos et al. failed to observe the functional importance of units with high CCMAS scores.

Human interpretation AM images

For the behavioral experiment, one hundred generated images were made for every unit in layers conv5, fc6 and fc8 in AlexNet, as in Nguyen et al. (2017), and displayed as 10×10 image panels. A total of 3,299 image panels were used in the experiment (995 fc8, 256 conv5, and 2048 randomly selected fc6 image panels) and were divided into 64 counterbalanced lists for testing. To assess the interpretability for these units as object detectors, paid volunteers were asked to look at image panels and asked if the images had an object / animal or place in common. If the answer was yes, they were asked to name that object simply (i.e. fish rather than goldfish). Analyses of common responses was done for any units where over 80% of humans agreed there was an object present.

The results of the behavioral experiment in which humans rated AM images are reported in Table 1. Consistent with past research, the generated images in the output fc8 layer were often interpreted as objects, and when they were given a consistent interpretation, they almost always (95.4%) correspond to the trained category. By contrast, less than 5%

Table 1: Interpretability judgements. Number of judgments for conv5, fc6 and fc8 were 1332, 10,656 and 5,181, respectively.

LAYER	% YES	% OF UNITS	% OVERLAP	% OVERLAP	
	RESPONSES	WITH ≥ 80%	AMONG	BETWEEN HUMANS AND:	
		YES RESPONSE	HUMANS	TOP CLASS	CCMAS CLASS
conv5	21.7% ±1.1%	4.3% ± 1.3%	89.5%±5.7%	34.1%±14.4%	0%
fc6	21.0% ±0.4%	3.1% ± 0.4%	80.4%±4.1%	23.3%±5.9%	18.9% ±5.9%
fc8	71.2% ±0.6%	59.3% ±1.6%	96.5%±0.4%	95.4%±0.6%	94.6% ±0.7%

of units in conv5 or fc6 were associated with consistently interpretable images, and as can be seen in Table 1, the interpretations only weakly matched the category with the highest top-class or CCMAS selectivity. The frequency with which objects were seen by the participants was similar in layers conv5 and fc6 layers and increased in fc8, consistent with the top-class and and precision measures of selectivity.

Apart from showing that there are few interpretable units in the hidden layers of AlexNet, our findings show that the interpretability of images does not imply a high level of selectivity given the maximum top-class selectivity for the hidden units is well under 10% (Fig. 2). In most cases, the top-class selectivity of the interpretable units was well under 1%. To briefly illustrate the types of images that participants rated as objects or non-objects see Fig. 6.

Discussions and Conclusions

Our central finding is that different measures of activation selectivity support very different conclusions when applied to the same units in AlexNet. In contrast with the precision (Zhou et al., 2015) and CCMAS (Morcos et al., 2018) measures that revealed some highly selective units for objects in layers conv5, fc6, and fc8, we found no localist representations, and the mean top-class selectivity in these layers was well under 1%. These findings are in stark contrast with the many localist ‘grandmother cell’ representations learned in RNNs (Bowers et al., 2014, 2016; Bowers, 2017b).

Not only did the different measures provide very different assessments of selectivity, we found that the precision and CCMAS measures provided highly misleading estimates. For example, a unit with over a 75% precision score for Monarch butterflies had a top-class selectivity of under 5%. Although Zhou et al. (2015) used 75% precision scores as the criterion for ‘object detectors’, it is inappropriate to call this unit a Monarch butterfly detector given that it did not respond strongly to the majority of Monarch butterfly images (and indeed, the modal response was 0.0; see Fig. 3).

At the same time, we identified problems with the localist, top-class, and activation maximization (AM) methods as well. The localist selectivity measure failed to obtain any localist representations, even at the output prob layer of AlexNet. This measure is so extreme that it misses highly selective representations that are of theoretical interest. The top-class selectivity does provide a graded measure of selectivity (with 100% top-class selectivity equivalent to a localist grandmother cell), but it can underestimate selectivity when a few member from outside the top-class are highly activated (see Fig. 4 (right) for an example). At the same time, the human interpretation of AM images provides a poor measure of hidden-unit selectivity given that interpretable AM images were associated with low top-class selectivity scores. These findings highlight the need to provide better measures of selectivity in order to better characterize the learned representations in NNs.

What should be made of the contrasting findings that localist representations are found in RNNs, but not in AlexNet? The failure to observe localist units in the hidden layers of AlexNet is consistent with the Bowers et al. (2014) claim that these units only emerge in order to support the co-activation of multiple items at the same time in short-term memory. That is, localist representations may be the solution to the superposition catastrophe, and AlexNet only has to identify one image at a time. This may help explain the reports of highly selective neurons in cortex given that the cortex needs to coactivate multiple items at the same time in order to support short-term memory (Bowers et al., 2016). It should be noted that the RNNs that learned localist units were very small in scale compared to AlexNet, and accordingly, it is possible that the contrasting results reflect the size of the networks rather than the superposition catastrophe *per se*. Relevant to this issue, Karpathy et al. (2016) reported examples of selective representations in a larger RNN with long-short term memory (LSTM) trained to predict text. Although they did not systematically assess the degree of selectivity, they reported examples that are consistent with 100% selective units, for similar findings see Lakretz et al. (2019). It will be interesting to apply our measures of selectivity to these larger RNNs. It should also be noted that there are recent reports of selective representations in Generative Adversarial Networks (Bau et al., 2019) and Variational Autoencorder Networks (Burgess et al., 2018) where the superposition catastrophe is not an issue. Again, it will be interesting to assess the selectivity of these units according to our measures in order to see whether there are additional computational pressures to learn highly selective or even grandmother cells. We will be exploring these issues in future work.

References

Attrition. (n.d.). http://Prolific.ac. (Accessed: 2018-09-24)

Bau, D., Zhu, J.-Y., Strobelt, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., & Torralba, A. (2019). Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1901.09887*.

Berkeley, I. S., Dawson, M. R., Medler, D. A., Schopflocher, D. P., & Hornsby, L. (1995). Density plots of hidden value unit activations reveal interpretable bands. *Connection Science*, *7*(2), 167–187.

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: a recurrent neural network model. *Psychological review*, *113*(2), 201.

Bowers, J. S. (2017a). Grandmother cells and localist representations: a review of current thinking. *Language, Cognition, and Neuroscience*, 257-273.

Bowers, J. S. (2017b). Parallel distributed processing theory in the age of deep networks. *Trends in cognitive sciences*, *21*(12), 950–961.

Bowers, J. S., Vankov, I. I., Damian, M. F., & Davis, C. J. (2014). Neural networks learn highly selective representations in order to overcome the superposition catastrophe. *Psychological review*, *121*(2), 248–261.

Bowers, J. S., Vankov, I. I., Damian, M. F., & Davis, C. J. (2016). Why do some neurons in cortex respond to information in a selective manner? insights from artificial neural networks. *Cognition*, *148*, 47–63.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., & Lerchner, A. (2018). Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on* (pp. 248–255).

Denker, J. S., & leCun, Y. (1991). Transforming neural-net output levels to probability distributions. In *Advances in neural information processing systems* (pp. 853–859).

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal*, *1341*(3), 1.

Gorilla experiment builder. (n.d.). www.gorilla.sc. (Accessed: 2018-09-24)

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., … Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Karpathy, A., Johnson, J., & Fei-Fei, L. (2016). Visualizing and understanding recurrent networks. *Workshop Track at International Conference on Learning Representations*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., & Baroni, M. (2019). The emergence of number and syntax units in lstm language models. *arXiv preprint arXiv:1903.07435*.

Le, Q. V., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G. S., … Ng, A. Y. (2011). Building high-level features using large scale unsupervised learning. *arXiv preprint arXiv:1112.6209*.

Morcos, A. S., Barrett, D. G., Rabinowitz, N. C., & Botvinick, M. (2018). On the importance of single directions for generalization. *arXiv preprint arXiv:1803.06959*.

Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A., & Yosinski, J. (2017). Plug & play generative networks: Conditional iterative generation of images in latent space. In *Cvpr* (Vol. 2, p. 7).

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., & Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in neural information processing systems* (pp. 3387–3395).

Palan, S., & Schitter, C. (2018). Prolific. aca subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833).

Zhou, B., Bau, D., Oliva, A., & Torralba, A. (2018). Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence*.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2015). Object detectors emerge in deep scene CNNs. In *International conference on learning representations.*