SELECTIVITY METRICS CAN OVERESTIMATE THE SELECTIVITY OF UNITS: A CASE STUDY ON ALEXNET

Anonymous authors
Paper under double-blind review

ABSTRACT

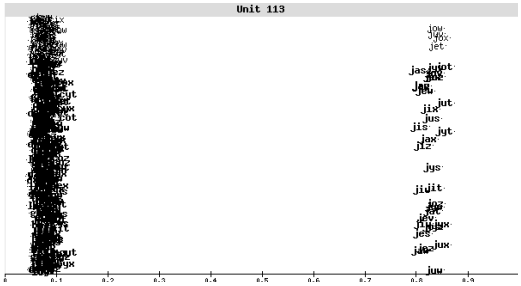
Various methods of measuring unit selectivity have been developed in order to understand the representations learned by neural networks. Here we undertake a comparison of four such measures on the well studied network AlexNet. In contrast to work on recurrent neural networks (RNNs), we fail to find any 100% selective 'localist units' in the hidden layers of AlexNet, and demonstrate that previous assessments of selectivity suggest a higher level of selectivity than is warranted, with the most selective units only responding most strongly to a small minority of images from within a category. No difference in selectivity was found between layers fc6 and fc7, and fc8 was much more selective. Only the output prob layer contained any localist units. We also generated images that maximally activated individual units and found that under (5%) of units in fc6 and conv5 produced images of interpretable objects that humans consistently labeled, whereas fc8 produced over 50% interpretable images. We consider why different degrees of selectivity are observed with RNNs and AlexNet, and suggest visualizing activations with jitterplots, aside from being comparable to neuroscience techniques, are a good first step to assessing unit selectivity.

1 INTRODUCTION

Although previously seen as black boxes, 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), network dissection (Zhou et al., 2017), precision(Zhou et al., 2014) and activation maximization (Erhan et al., 2009b). A striking finding is that some units respond selectively to image categories, including images of objects, people, and animals (e.g., units that respond more strongly to images of cats compared to other categories), however these selectivity measures have not been compared directly before.

Using networks similar to those developed by Botvinick & Plaut (2006) and designed to explain human performance on short-term memory tests, Bowers et al. (2014; 2016) assessed the localist selectivity of hidden units in recurrent NNs (RNNs). The key finding was the existence of 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. 1a for an example of a unit selective to the letter 'j'). These localist representations can be compared to 'grandmother cells' as discussed in neuroscience (Bowers, 2017). 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, as in this example, the model did not need to overcome the superposition catastrophe (Bowers et al., 2014).

In parallel, researchers (Zhou et al. 2014; Morcos et al. 2018; Zeiler & Fergus 2014; Erhan et al. 2009a, for a review see Bowers (2017)) have reported selective units in the hidden layers of CNNs, including AlexNet (Krizhevsky et al., 2012) trained to classify images into one of multiple categories. Importantly, these selective units were observed when the models were only trained to identify items one-at-a-time, suggesting there are additional pressures to learn selective representations above and beyond the challenge of overcoming the superposition catastrophe. However, the



(a) A jitterplot for unit 'j' in an RNN

(b) AM images for a 'lighthouse' detector

(c) Regions of real images identified by humans as depicting 'lamps'

Figure 1: Examples of selectivity measures used. (a) A jitterplot of a selective unit 113 found in an RNN under the superposition constraint; from Bowers et al. (2014). All input words containing the letter 'j' (right column) are more activated than words that do not (left column); therefore this unit is selective for the letter 'j'. (b) Two images optimized to maximize the activation of a 'lighthouse' detector in layer conv5 of AlexNet; from Nguyen et al. (2016). (c) Highest-activation images for a 'lamp' detector with 84% precision in layer pool5 of AlexNet; from Zhou et al. (2014).

measures of selectivity that have been applied to these CNNs are different to localist selectivity, and accordingly, it is difficult to directly compare results. Therefore, to advance our understanding of the "distributed vs. local" coding in CNNs, it is important to compare and contrast different selectivity metrics in the literature.

In this paper, we compare the localist selectivity measure with two alternative methods of measuring selectivity on the prob, fc8, fc6, and conv5 layers of AlexNet and show how they surprisingly provide very different assessments of selectivity. Critically, we observe no localist units in the hidden layers of AlexNet. The reduced selectivity observed in AlexNet, compared to the RNNs, suggests that the superposition constraint (only present in RNNs) does indeed provide an extra pressure to learn more selective units in RNNs. We also compare these three 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 have been argued interpretable by humans (e.g., a generated image that looks like a lighthouse, see Fig. 1b). This suggests that an interpretable unit may have learned to *selectively* represent a concept (e.g. lighthouse). Here, we directly evaluate the interpretability of AM images and compare it with the selectivity of corresponding units.

2 METHODOLOGY

Networks and Datasets All $\sim$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 as a layer called 'prob'(ability), which is the output layer with 1000 units (one for each class). We analyzed these prob units, the fully connected (fc) layers: fc8 (1000 units), fc6 and fc7 (4096 units), and the top convolutional layer conv5 which has 256 filters. We only recorded the activations of correctly classified images, and saved them in an activation table so the activations could be probed without re-evaluating the images each time.

Localist selectivity A unit is localist for class A if the set of activations for class A are disjoint with those of $\neg A$. A unit is selectively 'on' if $\{A\} > \{\neg A\}$ (i.e. all images in A have higher activations than those not in A) and selectively 'off' if $\{A\} < \{\neg A\}$. We looked for localist units for all 256 conv5, 1000 fc8 and prob, and 2524 fc7 and 3523 fc8 units. Localist selectivity is easily depicted with jitterplots in which a scatter plot for each unit in generated (see Figs. 2a, b and 3a, b). 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). When generating jitterplots for the conv5 layer we plotted the highest level of activation across each filter for each image.

Precision The precision method of finding object detectors (Zhou et al., 2014; 2017) involves identifying a small subset of images that most strongly activate a unit (the number of images in the most strongly activated subset differ across papers) and then identifying the critical part of these images that are responsible for driving the unit. Zhou et al. 2014 took the 60 images that activated a unit the most strongly and asked independent raters to interpret the critical image patches. For some units raters frequently reported seeing objects, Zhou et al. (2014) developed a 'precision metric' that assessed the percentage of the 60 images that raters judged to depict the same category 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. 1c). Object detectors were defined as units with a precision of over 75%: they reported multiple such detectors. Here we roughly 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 category (e.g., if 75 of the top 100 images are all examples of a class 'lighthouse' then we consider the unit to be a 'lighthouse' object detector with a precision of 75%).

CCMAS Morcos et al. (2018) introduced a selectivity index based on the 'class-conditional mean activation' (hereafter, CCMAS). The CCMAS for class A compares on 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 measure should vary between 0 and 1, with 0 meaning that a units average activity was identical for all classes, and 1 meaning that a unit was only active for inputs of a single class. They reported that many units have a selectivity of 1 when CNNs are trained on CIFAR-10 without batch normalization. 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 category.[1]

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 (see Appendix Figs.A4, A6 & A7). We then asked participants to judge whether they could identify any repeating objects, animals, or places in images after receiving some practice trials (see Appendix Fig. A3 for an example). The experiment was run online using Gorilla (gor, b), and a demo experiment is available at: exp. 333 participants were recruited using Prolific (pro; Palan & Schitter, 2018). A total of 3,299 images were used in the experiment (995 fc8 images, 256 conv5 images, and 2048 fc6 images) and were divided into 64 counterbalanced lists of 51 or 52 images (4 conv5 images, 15 or 16 fc8 images and 32 fc6 images). 51 of the lists were assigned 5 participants and 13 lists were assigned 6 participants (the uneven allocation is a natural consequence of running experiments online; see gor (a)). Thus, at least five judgments were collected for each image.

If participants indicated they could identify something, they were asked to type what they saw. Typed answers were analyzed to investigate whether there was commonality amongst humans interpretation of the images (e.g., after answering 'Yes', did most people agree the image depicted the same object?): only images that received an average of 80% or more 'Yes' responses were further analyzed, and for each of these images, we calculated the percentage of typed responses that belonged to the same class; for example, 'beer', 'glass', 'drink' were all considered a common answer (most responses were more obvious and required far less interpretation than this example). The commonality between human responses and the model outputs was also analyzed.

[1]For example, if a given unit has high CCMA selectivity for the category 'dog' and a similar CCMAS-2 selectivity to 'fish', it is misleading to characterize the unit as selective to the 'dog' category.

3 RESULTS

3.1 UNITS ARE ONLY SELECTIVE TO A SUBSET OF A CLASS RESULTING IN A HIGH PRECISION BUT LOW SELECTIVITY.

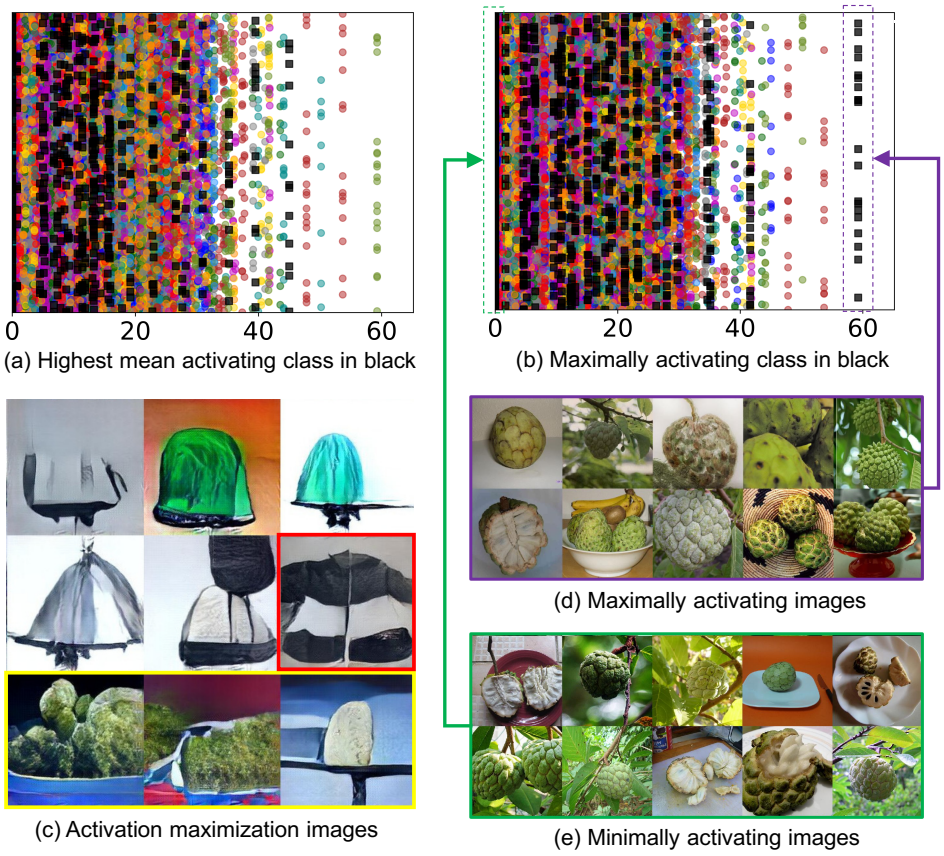


Figure 2: Representative example of selectivity analyses for unit 3845 in layer fc6 of AlexNet. **(a)** Jitterplot for all images on the unit. The class with the highest mean activation class (as assessed via CCMAS) is shown as black squares, other classes are coloured dots. **(b)** Jitterplot for all images on unit fc6.3845 with the maximum activating class shown as black squares (and the top cluster at the end of the pink arrow). **(c)** 9 example AM images for the unit. **(d)** Maximally activating ImageNet images (activation of 59.2) and members of the top cluster. **(e)** Minimally activating ImageNet images (activation of 0) from the maximally activating class. The maximally activating class is 'custard apple' (n07760859), and the bottom three AM images (yellow-bordered) could conceivably be said to be custard apples. The highest mean activating class was 'trench coat' (n04479046), and the middle-right AM image (red-bordered) could conceivably be a jacket. 4/5 human participants rated this image as interpretable, with 1 identifying a 'jacket', the other 4 identified a 'tent', which some of the AM images in the example could conceivably be labeled as. The CCMAS measure was over 0.8 (see Table 2) indicating that this unit should be quite selective for custard apples. This unit has a precision of 33% for custard apple. Despite having a decent precision, high CCMAS measure, and interpretable AM images, this unit cannot really be said to be a custard apple detector as 39.6% of correctly classified images in the top class have an activation of 0, and only 0.02% of that class are in the top cluster. Inspection of the maximally and minimally activating images (d and e) does not reveal any obvious quality of the photographs that lead them to have high or low activation.

Example jitterplots for fc6, fc8 and prob layers are in Figs. 2a–b, and Fig. 3a–b. We only found localist units in the output prob layer, of which there were 101. Most units (69.66% fc6, 69.56% fc7 and 65.50% fc8) had a single cluster of activations separated by a gap from the rest of the high

Table 1: Different measures of selectivity across different layers of AlexNet. Localist selectivity constitutes 100% selectivity, '% localist units' is the number found in a layer. 'Precision' is the highest percentage of images from a single class in the top 100 activations. 'Top cluster size' is the number of contiguous items that are from the same class as the most active image. And '% overall selectivity' divides the top cluster size by the number of images in that class and presents it as a %. N.B. n_s is the number of selective units, n_{layer} is the number of units in a layer, n_i is the number of highly activating images (i.e. those in the top cluster), n_{class} is the number of images in a class.

LAYER	**% LOCALIST UNITS** n_s/n_{layer}	**MEAN PRECISION**	**TOP CLUSTER SIZE** $\text{Mean}(n_i)$	**% OVERALL SELECTIVITY** $100 \times \text{Mean}(n_i)/\text{Mean}(n_{class})$
prob	10.1%	99.7%	592.5 images	82.1%
fc8	0%	70.2%	108.4 images	14.7%
fc7	0%	26.4%	19.9 images	2.6%
fc6	0%	22.7%	17.8 images	2.4%
conv5	0%	20.9%	17.3 images	2.2%

activations, as shown in Fig. 2a.[2] Note that, the distribution of activations across the entire range is neither uniform nor Gaussian.[3] Although units with one highly active cluster are common, and do reflect some degree of selectivity, it is important to emphasize that they do not contain all images for a single class, and thus are not localist units.

In order to estimate Zhou et al. (2014)'s precision metric we simply computed the highest percentage of images from a given output class from the top 100 activating images. As can be seen in Table 1, the precision scores provide a much higher estimate than the % overall selectivity score. Indeed, the discrepancy can be quite strong. For example, unit 94 of layer fc6 (see Appendix Fig. A7) is most activated by 49 images of the 'monarch butterfly' class (n02279972), with the 50th most active image from another category. This unit also has a total of 73 butterfly images in the top 100 activated images. Thus, by our approximate precision measure, this unit has a precision measure of 73%. Note that, in Zhou et al. (2014) they used the top 60 images, and as we know the top 49 are all butterflies, we can estimate the precision over the top 60 images as at least 81.6%, which would be a 'monarch butterfly' detector under Zhou et al. (2014) threshold of 75% — this for a unit which is most highly activated for only 4% of the butterfly class.

We also assessed selectivity by identifying the single-most active image, and assessing how many items from this class fell into a contiguous cluster (i.e., how many members of the butterfly category were more active than any other image), which we call 'top cluster size', and this is divided by the total number of images in that class to provide an estimate of % overall selectivity (in the case of localist units it would be 100%). These selectivity measures are given in Table 1, and the distribution of these values is in the Appendix Fig. A1.

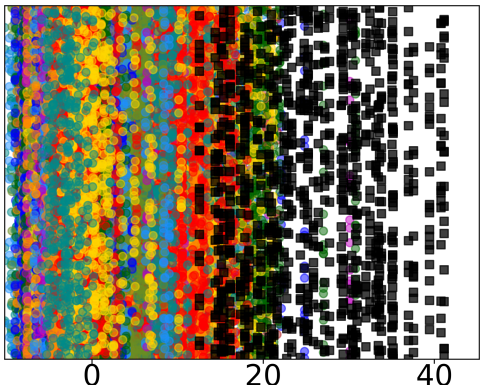
It is made clear from Table 1 and Fig. A1 that precision measures provide a misleading estimate how how selective the units are for their most active category. If we assessed the precision in relation to a smaller subset of the most highly activated units (e.g., 60 images as used by Zhou et al. (2014) rather than 100 used here) this discrepancy would be even stronger. At the same time, top cluster size and overall % selectivity measures do seem to provide a good general measure of class-selectivity across the layers, with little significant difference between fc6 and fc7, and a much higher degree of object selectivity at fc8.

3.2 CCMAS IS NOT REPRESENTATIVE OF SINGLE-UNIT SELECTIVITY

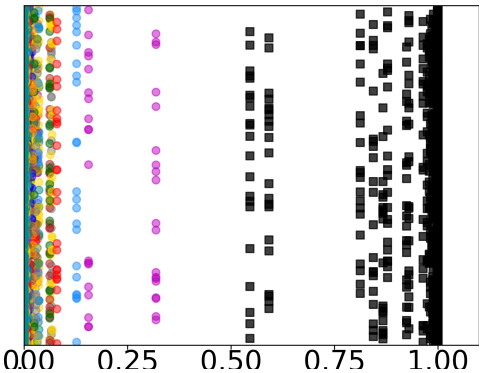
In contrast with the above measures that assessed the selectivity for the most active class, Morcos et al. (2018) computed a selectivity measure with reference to the class with the highest mean ac-

[2]For the purposes of identifying clusters and gaps, we only investigated the activations in the top half of the total range, or the 2000 activations, whichever was smaller, and a single top cluster is identified if the gap between it and the rest of the investigated activations was the largest gap in that range.

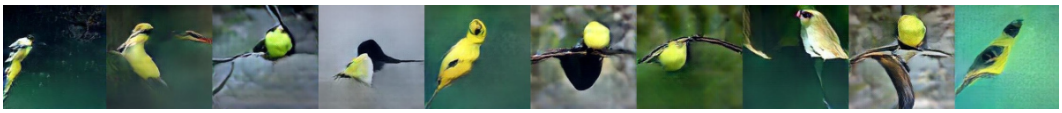
[3]The density of activations drops off drastically with activation.



(a) Jitterplot for unit fc8.11



(b) Jitterplot for unit prob.11



(c) Activation maximization images for unit fc8.11

(d) ImageNet images in class 'goldfinch' corresponding to output #11

Figure 3: Example data from the fc8 and prob layers. (a) activations for unit fc8.11. (b) activations for prob.11 (i.e. post-softmax). (c) AM images for unit fc8.11. (d) highest activating images for fc8.11. Activations for the 'ground truth' class 'goldfinch' (n01531178) are shown as black squares, all other classes are shown as coloured circles. The prob layer neuron is 100% selective for 'goldfinch' images, but fc8 is not. However, most of the 'n01531178' images have high activations, and the AM images are interpretable: 5/5 humans rated the images as containing a 'bird'.

tivating class (see Fig. 2a for an example of the extent of the highest mean activating class). The CCMAS measure as well as the CCMAS_2 measure that assesses the selectivity of the second most active category are displayed in Table 2. Consistent with other measures, the CCMAS does increase in the fc8 and prob layers (see Fig. A2). However, we have identified several problems with this measure. The first point to note is that contrary to Morcos et al. (2018), the CCMAS measure can go above 1 if $\mu_{\neg A}$ is negative, and this happens in the fc8 layer of AlexNet (see Fig. 3a) because there is no ReLU transformations in layer fc8 (as opposed to layers fc6 & fc7) (Krizhevsky et al., 2012). This is a minor issue that could be fixed by normalizing the measure to the range of activations, but it explains why some of our CCMAS scores are above 1.

More importantly, high CCMAS measures (i.e. of 0.9+) do not imply high selectivity: compare the CCMAS measures across the layers in Table 2, where we have picked the highest CCMA measures from units 0–999 for layers fc7 and fc6 (fc6.656 and fc7.603), and both these units have CCMAS of more than 0.94, which should imply a high selectivity towards the top class, but the fc7 and fc6 units only had an overall selectivity measures of 26% and 2%, respectively. The data for the example (fc6.3845) given in Fig. 2 is also given, and has a high CCMA measure of 0.824, whereas its overall selectivity is only 2%.

In addition, if the CCMAS provided a good measure of a unit's class selectivity then one should expect that a high measure 0.99 for one class would imply that the unit is not selective for other classes. This is not the case given the high CCMAS_2 selectivity scores reported in both Table 2 and Figs. A2, where values above 0.9 can be seen. The box plots in figure A2 show the distributions of values for fc6 and fc7 overlap, indicating that the CCMAS measure is not useful to discriminate between selectivity to different classes.

We also found that the CCMAS measure does not seem to relate to the top cluster size or overall selectivity. Neither can it discriminate between a localist and non-localist unit (compare prob.11

Table 2: CCMAS selectivity measures for example units and average measures across layers. *Indicates a localist selective unit. Within the first 1000 units, fc7.603 and fc6.656 were the units with the highest CCMAS measures. Units fc6.3845, fc8.11 and prob.11 are the units given in Figs. 2 & 3.

LAYER.UNIT	CCMAS	CCMAS_2	Top Cluster Sizes	% Overall Selectivity
prob.11*	0.999882	0.9992301	1050	100%
prob.12	0.999755	0.972829	827	100%
fc8.11	1.11419	1.22388	336	32%
fc8.12	0.999755	0.972829	72	8.7%
fc7.603	0.944357	0.877084	288	26.0%
fc6.656	0.949765	0.908498	22	1.98%
fc6.3845	0.823709	0.813349	20	2.0%
conv5.78	0.755518	0.756379	25	3.7%
Mean[prob]	0.999310	0.986297	592.5	82.1%
Mean[fc8]	0.997822	1.00639	108.4	15.0%
Mean[fc7]	0.853792	0.836948	20.3	2.8%
Mean[fc6]	0.856158	0.834398	17.4	2.4%
Mean[conv5]	0.506556	0477679	17.3	2.6%

Table 3: Human answers to the question of whether or not a set of AM images contained an object, animal or place. N.B. there is arguably a ground truth label for fc8 as this is layer before softmax and readout. The number of judgments for conv5, fc6 and fc8 were 1332, 10,656 and 5,181, respectively.

LAYER	%YES RESPONSES	% OF UNITS WITH $\geq$ 80% YES RESPONSE	% OVERLAP	
			AMONG HUMANS	BETWEEN HUMANS vs. CNN
conv5	21.7% $\pm$1.1%	4.3% $\pm$ 1.3%	89.5%$\pm$5.7%	35.9%$\pm$14.1%
fc6	21.0% $\pm$0.4%	3.1% $\pm$ 0.4%	80.6%$\pm$3.0%	23.4%$\pm$4.7%
fc8	71.2% $\pm$0.6%	59.3% $\pm$1.6%	96.5%$\pm$0.4%	95.4%$\pm$0.6%

and prob.12 in Table 2). Clearly, CCMAS is providing a measure of selectivity that is very different to alternative methods, and it will be important to determine what it is measuring. One possible reason is that very different patterns of selectivity are consistent with the same CCMAS measure. For example, a unit that is off to everything but weakly activates one image from a 'bird' category will have the same 1.0 CCMAS score as a unit that is highly activated by all members of the 'bird' category and off to everything else (e.g., a localist unit).

3.3 Human interpretation of generated images from across AlexNet

The results of the experiment in which humans were asked to rate AM images generated to highly activate units are reported in Table 3. Participants agreed that 71.2% of the fc8 output layer images contained an object, however they only judged 21.7% and 21.0% of images at conv5 and fc6 respectively to contain objects (Table 3). [4]

We also analyzed the distribution of 'Yes' responses and the participants' written answers to gain further insights into how the images were being interpreted. As can be seen in Table 3, participants consistently responded 'Yes' to over half of the units in fc8, and when they did, almost always gave the same answer (96.5% of the time). By contrast, under 5% of the units in conv5 and fc6 were consistently given a 'Yes' answer, but in these few cases, participants tended to give consistent answers

[4]Measures of their reaction times to answer the question can be used as a measure of their certainty, and this data agrees with that reported here and is given in the appendix in figure A5.

(80% or more). This shows that there are a small percentage of hidden units that have consistent interpretable images. Interestingly, the consistent 'Yes' responses in fc8 layers were almost always associated with the category the model was trained on (over 95%), whereas the consistent answers in the hidden layers were less well captured by our measures of selectivity: they agreed with the max class 23.4% $\pm 4.7\%$ of the time, the mean class, 23.7%$\pm 4.8\%$ and the 2nd highest mean class 17.5%$\pm 4.0\%$ of the time (note that for 14.0% of the units, the maximally activating class was also the highest mean activating class), so the AM images were interpretable to either the maximally activation or highest mean activating class.

Figure 2c shows an example of a top-ranked AM image where participants agreed on the existence of an object, but disagreed on what that object was (and where crucially, the mean activating class was a more salient object ('trench coat') than the maximally activating ('custard apple'), yet the maximally activating class is present.[5] The human measures of object presence follow the same pattern as the measures from our other measures of selectivity, namely that there is a little difference between the lower levels, and a larger difference with fc8.

4 Discussions and Conclusions

The current findings show that the localist units observed by Bowers et al. (2014) in RNNs are not observed in the hidden layers of AlexNet CNN, and further, that the level of selectivity in AlexNet is much less than the precision (Zhou et al. 2014) and CCMAS (Morcos et al. 2018) measures would suggest. Indeed, the precision and CCMAS measures were found to be misleading, with a precision of 75% (used as a criterion to call a unit an object detector) and CCMAS measure of .95 corresponding to units with an overall selectivity of $\sim$ 5%; and a very high CCMAS measure did not indicate the item was selective *exclusively* to only this one category. Still, it is clear that units in the hidden layers of AlexNet learned some degree of selectivity, and future work designed to understand the inner-workings of CNN will clearly need to use a range of measures to provide an accurate characterization of the representations learned. The localist selectivity measure used by Bowers et al. (2014) was unable to highlight the subtle changes in class-selectivity as they were below 100% selective, whereas the top cluster size and % overall selectivity do provide graded measures of selectivity. We suggest that there is room in the field for more selectivity measures that might better capture object-selectivity than the current measures.

The failure to observe localist units in the hidden layers of AlexNet is consistent with 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, highly selective representations may be the solution to the superposition catastrophe, and AlexNet only has to identify one image at a time. However, it should be noted that the RNNs that learned localist units were very small in scale compared to AlexNet, and accordingly, it will be interesting to assess the selectivity in larger RNNs that have much larger memory capacity. Relevant to this issue, Karpathy et al. (2015) reported some striking examples of selective representations in a recurrent long-short term memory (LSTM) networks trained to predict text based on training on Tolstoy's novel 'War and Peace' and the Linux Kernel source code. Although they did not systematically assess the degree of selectivity, they reported examples that are consistent with 100% selective units. If in fact the superposition constraint provides a pressure to learn more selective representations, then we should observe more highly selective representations, perhaps localist units, in large RNNs as well. We will be testing this hypothesis in future work.

It is interesting to consider how these findings relate to reports of single neurons in cortex and hippocampus responding to objects and faces in a highly selective manner (see Bowers (2009)). The neuroscience findings are consistent with the localist units observed in RNNs as well Bowers (2017) as networks that are hand-crafted to have localist representations (Gubian et al. 2017; Page 2017), but the level of selectivity we observed in AlexNet may also provide a reasonable account of single-cell recording studies. Future work should systematically compare what level of selectivity in artificial neural networks provide a better account of the neuroscience findings. One possibility is that both artificial and biological neural networks learn different degrees of selectivity depending on whether or not they support the co-activation of multiple items at the same time in short-term memory.

[5] Although, the saliency of custard apple to trench coat might well be reversed if our participants were based in the far east rather than rainy western Europe.

REFERENCES

Experiment trial. https://research.sc/participant/login/dynamic/63907FB2-3CB9-45A9-B4AC-EFFD4C4A95D5. Accessed: 2018-09-24.

Attrition. https://gorilla.sc/support/walkthrough/create-to-launch#attrition, a. Accessed: 2018-09-24.

Gorilla experiment builder. www.gorilla.sc, b. Accessed: 2018-09-24.

Attrition. http://Prolific.ac. Accessed: 2018-09-24.

Istvan SN Berkeley, Michael RW Dawson, David A Medler, Don P Schopflocher, and Lorraine Hornsby. Density plots of hidden value unit activations reveal interpretable bands. *Connection Science*, 7(2):167–187, 1995.

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

Jeffrey S Bowers. On the biological plausibility of grandmother cells: implications for neural network theories in psychology and neuroscience. *Psychological review*, 116(1):220, 2009.

Jeffrey S Bowers. Grandmother cells and localist representations: a review of current thinking. *Language, Cognition, and Neuroscience*, pp. 257–273, 2017.

Jeffrey S Bowers, Ivan I Vankov, Markus F Damian, and Colin J Davis. Neural networks learn highly selective representations in order to overcome the superposition catastrophe. *Psychological review*, 121(2):248, 2014.

Jeffrey S Bowers, Ivan I Vankov, Markus F Damian, and Colin J Davis. Why do some neurons in cortex respond to information in a selective manner? insights from artificial neural networks. *Cognition*, 148:47–63, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. Ieee, 2009.

John S Denker and Yann leCun. Transforming neural-net output levels to probability distributions. In *Advances in neural information processing systems*, pp. 853–859, 1991.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009a.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009b.

Michele Gubian, Colin J Davis, James S Adelman, and Jeffrey S Bowers. Comparing single-unit recordings taken from a localist model to single-cell recording data: a good match. *Language, Cognition and Neuroscience*, 32(3):380–391, 2017.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Ari S Morcos, David GT Barrett, Neil C Rabinowitz, and Matthew Botvinick. On the importance of single directions for generalization. *arXiv preprint arXiv:1803.06959*, 2018.

Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems*, pp. 3387–3395, 2016.

Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *CVPR*, volume 2, pp. 7, 2017.

Mike Page. Localist models are compatible with information measures, sparseness indices, and complementary-learning systems in the brain. *Language, Cognition and Neuroscience*, 32(3): 366–379, 2017.

Stefan Palan and Christian Schitter. Prolific. aca subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene CNNs. *arXiv preprint arXiv:1412.6856*, 2014.

Bolei Zhou, David Bau, Aude Oliva, and Antonio Torralba. Interpreting deep visual representations via network dissection. *arXiv preprint arXiv:1711.05611*, 2017.

APPENDIX

A1 SIMPLE MEASURES OF SELECTIVITY ACROSS THE LAYERS OF ALEXNET

Figure A1a shows the distribution of the top contiguous cluster size (the number of top-most activating units that have the same class and can be said to form a contiguous cluster) across the layers. The range is large, but in general the size of the top contiguous cluster increases with depth until the final read-out layer (prob) where most of the top most class should be the output class. Figure A1b shows the percentage of the top 100 images which are the highest activating class. This matches the data for the size of the top-most contiguous cluster. These two simple measures provide a simple profile of how the units are qualitatively different with regards to class selectivity in fc8 and prob layers compared to the other investigated layers. the increase in range and outliers with layer depth suggests that some increase in class selectivity is happening with depth.

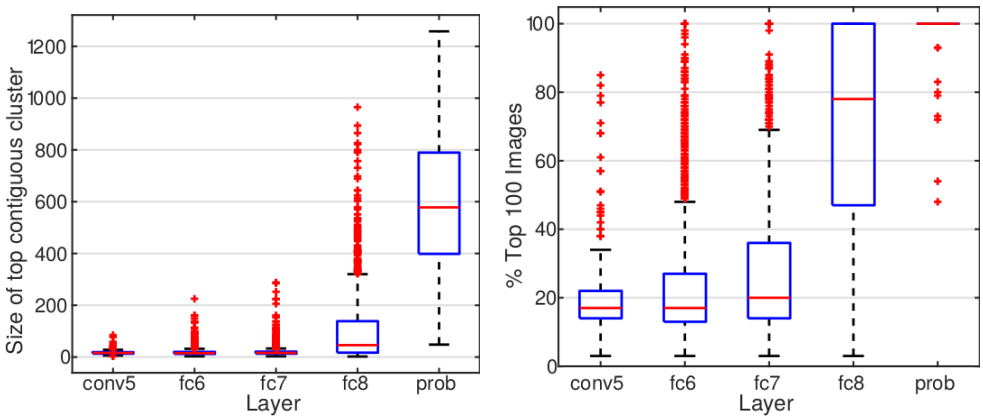


Figure A1: The simple measures of the size of the top contiguous cluster of the maximally activating class (left) and the percentage of the top 100 images (right) give a rough measure of the class selectivity of units across AlexNet.

A2 CCMAS MEASURE DISTRIBUTIONS ACROSS THE LAYERS IN ALEXNET

Figure A2a shows a different pattern of selectivity across the layers in AlexNet, with the fc6 and fc7 layers being qualitatively similar to each other, but the conv5 being different from them. Again, the selectivity seems to roughly increase with depth, but not in the same way as seen with the simpler measures, which could suggest that the CCMAS is not capturing class selectivity. Figure A2b shows that the distributions for CCMAS and CCMAS_2 measures overlap, demonstrating that a high CCMAS measure for class A does not guarantee that the unit is likely to have be especially selective for class A, even under the CCMAS measure. Similar results are found for the other layers of AlexNet.

A3 HUMAN INTERPRETATION OF ACTIVATION MAXIMIZATION IMAGES

Figure A3 shows one of the instruction slides shown to online participants, which includes the three questions participants were asked if they answered 'Yes' to question 1. There were several training examples which aimed to get the participants looking for concrete objects (which included animals or places, as these are in ImageNet2010 categories (e.g., objects: 'wine glass', 'guitar'; places: 'volcano', 'beachhead')). Participants were also shown a negative example with abstract concepts like 'blue triangle' in common.

At test, participants were shown all 100 AM images in a grid, as shown for the example unit fc6.3845 in figure A4. It is the opinion of one of the authors that there are around 12 'crab apple' objects, 6 'jackets', and around 40 'tent-like' objects, although this is open to interpretation.

Figure A5 shows the response times to the first question ("Can you identify multiple examples of an everyday object, place or animal?") on the slide in figure A3. If our assumption that as faster

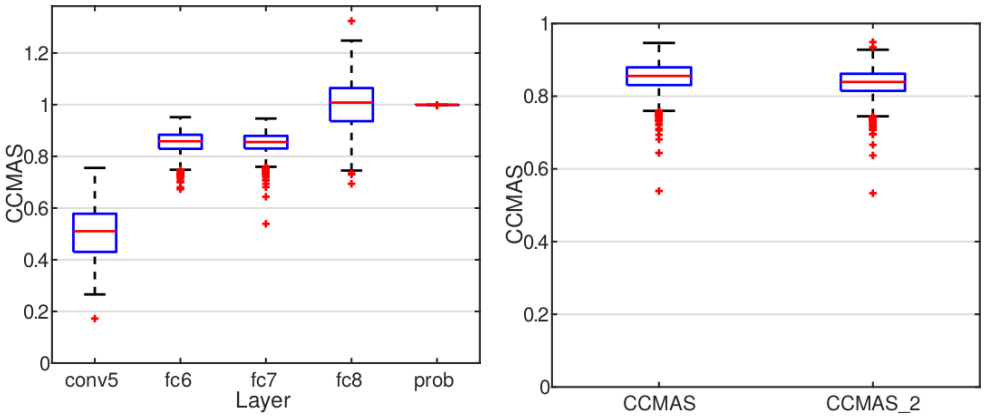


Figure A2: Left: CCMAS selectivity for different layers of AlexNet. Right: CCMAS selectivity the class with the highest mean activity(Morcos et al. 2018's definition of CCMAS) and the class with the second highest mean activity for layer fc7. CCMAS increases up the layers, so is measuring something relating to selectivity, but due to the large overlap (right) it is not helpful in discrimination. The maximally activating and highest mean activation are not always the same class (see figure 2 and figure A6 in the appendices), and the second highest mean class has an almost the same mean and highly overlapping range of values to the top mean class, making this measure not that useful for discriminating class selectivity.

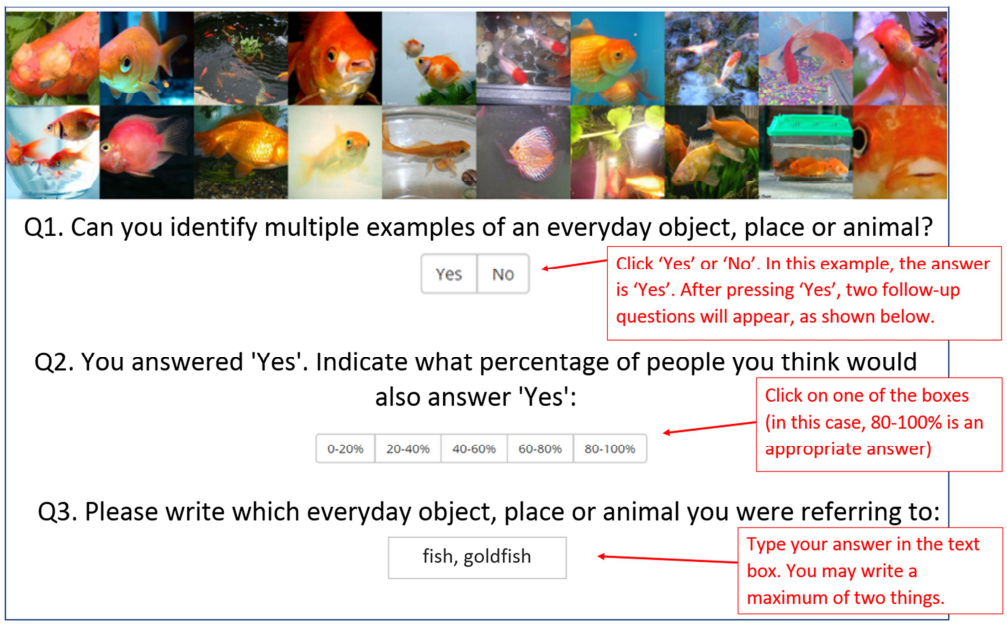


Figure A3: Example screen from the identification task shown to participants as part of the instructions. The images included here are ImageNet2010 images, not AM images.

response time is seen when participants are more sure that an object is present, and thus the unit is more easy to interpret, is correct, then the response time data fits with our other selectivity measures as there is no real difference between the response speeds for fc6 and conv5, but participants a faster for fc8, therefore fc8 is easier to interpret and the AM images are more object-like.

Figure A4: All AM images for fc6.3845. Tent like objects are more common than crab apples or jackets.

A4 EXTRA EXAMPLE DATA OF fc6 UNITS

Figure A6 shows another example of an fc6 unit that was scored highly by participants as containing an object (5/5 'bird'), which also has a high CCMAS measure, low precision and is only selective to a very small percentage of the dataset.

Figure A7 is the example unit discussed in the text, which would be a 'monarch butterfly' detector under the Zhou et al's precision measure, note that there are still many monarch butterfly images (black squares in the top part of figure A7) that have low or zero activation on this unit.

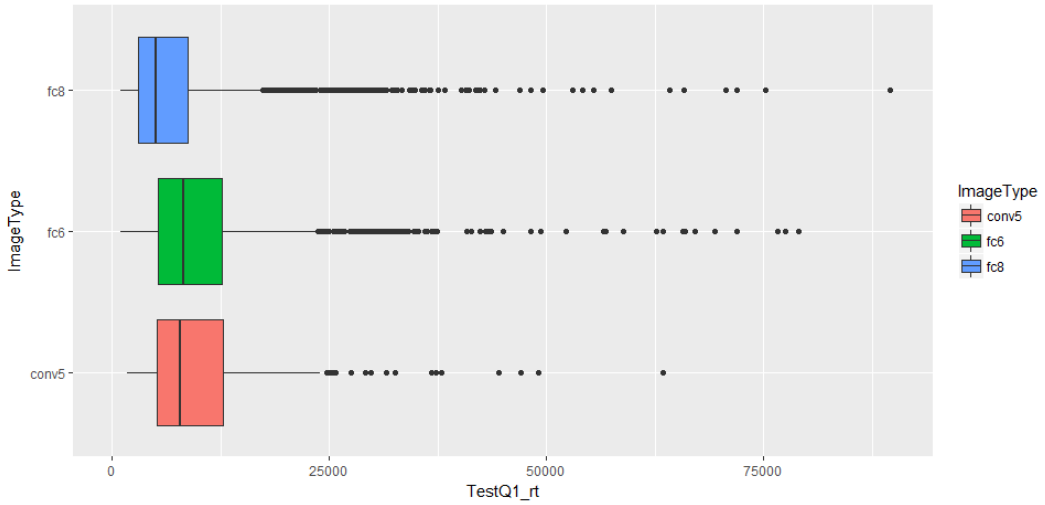


Figure A5: Response times (TestQ1_rt) for 'Yes' answers to the question, 'do the images show an object, animal or place?', the number of observations per layer (N) is given in table 3, the black lines are the median of the distribution, and around 6 outliers fall outside the plotted range.

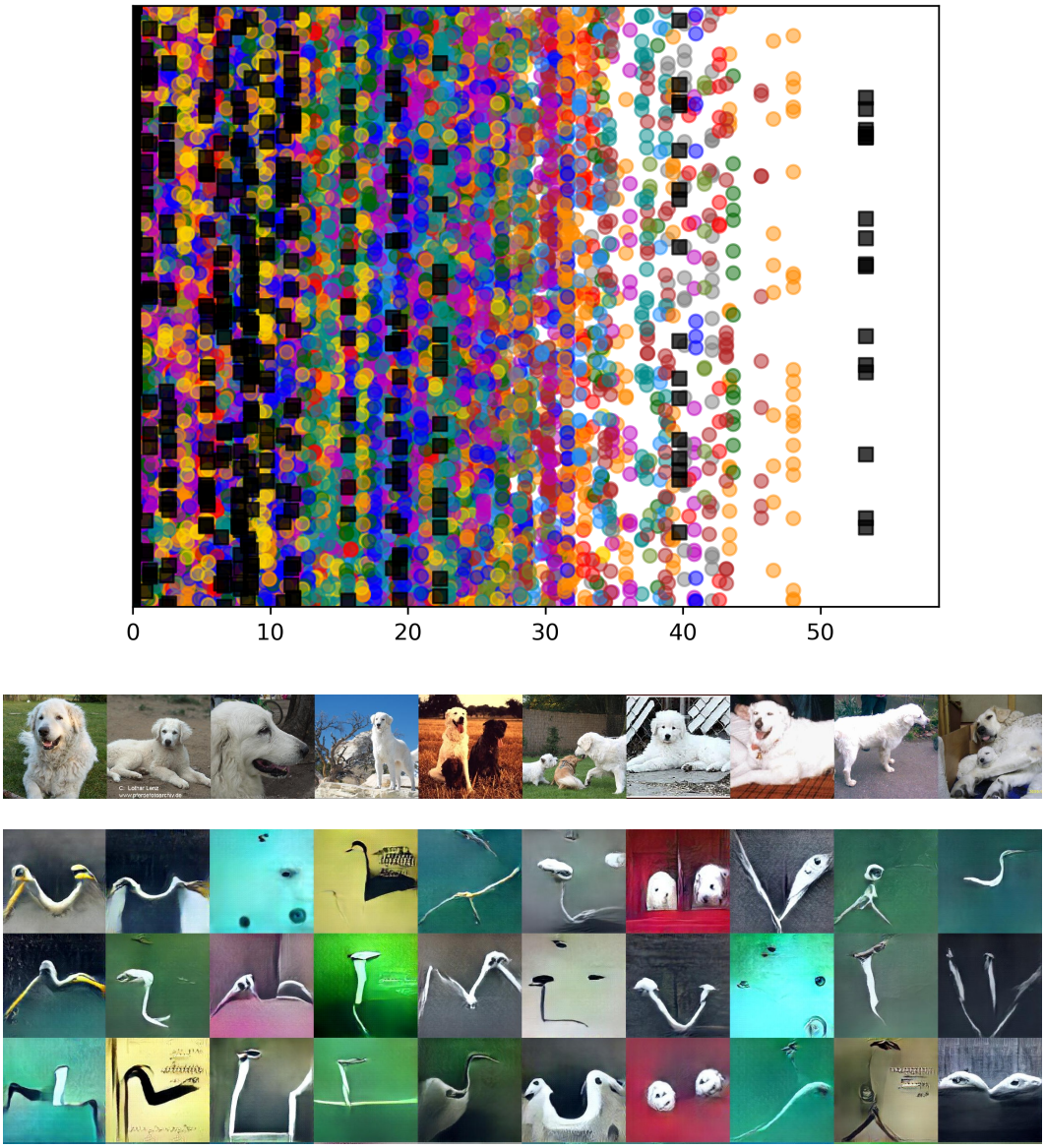


Figure A6: Top: Jitterplot for all images on unit fc6.365, the maximally activating class is shown as black squares, other classes are coloured dots. Middle: Maximally activating ImageNet images in top cluster ('Kuvasz dogs'). Bottom: 30 representative AM images for fc6.365. In this example, the top class is 'n02104029 kuvasz', makes up only 16% of the top 100 images, which, as there are 777 Kuvasz images in our dataset, this corresponds to only 2.05% of the top class is in the top 100. It can be seen that the maximally activating class covers a range of 53.2929. This class is not the class with the highest mean activation, and so the CCMAS would pick the 'n03857828 oscilloscope' as the class this unit is selective to, with a range of 39.2584 and a CCMA measure of 0.801212 ('Kuvasz dogs' measure 0.627443 , largely due to a higher range of activations, and the 2nd highest mean class has a CCMAS of 0.7935).

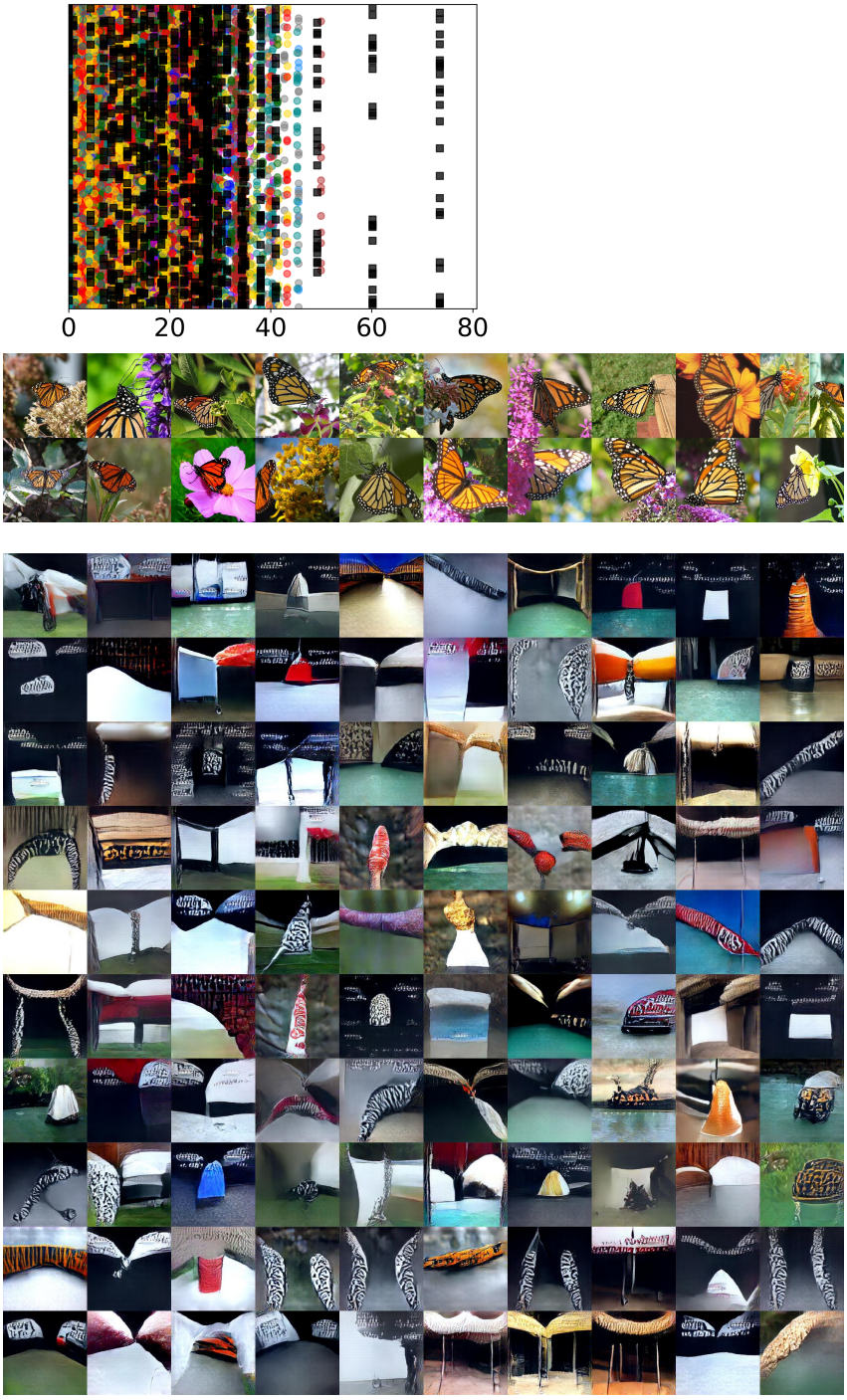


Figure A7: Example data for unit fc6.94. Top: all activations, maximally activating class shown in black; middle: highest activating ImageNet2010 images, bottom: AM images for fc6.94. This unit is selective for monarch butterflies under the Zhou et al precision measure.