Finding the Unknown: Novelty Detection with Extreme Value Signatures of Deep Neural Activations – Supplementary Material –

Alexander Schultheiss[1], Christoph Käding[1,2], Alexander Freytag[3], and Joachim Denzler[1,2]

[1]Computer Vision Group, Friedrich Schiller University Jena, Germany
[2]Michael Stifel Center Jena, Germany
[3]Carl Zeiss AG, Jena, Germany

Abstract. This document contains supplementary material for the paper "Finding the Unknown: Novelty Detection with Extreme Value Signatures of Deep Neural Activations". We provide additional information for the following aspects: (i) selection of features and additional parameters for all evaluated methods for the experiments in Section 4.2 and Section 4.4 of the main paper (Section S1), (ii) additional results for the large-scale experiment in Section 4.4 of the main paper (Section S2), (iii) qualitative results for the experiment in Section 4.2 of the main paper (Section S3), (iv) statistical significance of differences between methods for the experiments in Section 4.2 and Section 4.4 of the main paper (Section S4), and (v) additional visualizations using the heuristics presented in Section 4.5 of the main paper (Section S5). *This document is not required to understand the main paper* – it only gives additional insights and allows for better reproducibility of results.

S1 Method Setups

As mentioned in the main paper, all images are encoded with activations of different layers from the Places365-CNN [4]. To investigate the effects of different encodings, we compared activations with and without passing it through an additional `RELU` layer (*i.e.,* setting negative values to zero). Additionally, we evaluated the effect of normalization to unit length. Results from all four coding schemes applied to layers `CONV4` to `FC8` where systematically compared on the 10:10 split (see Section 4.1 and Section 4.2 of the main paper for details) and only the best performing combinations where shown in the main paper. In the following, we present all results for all settings.

This research was supported by grant DE 735/10-1 of the German Research Foundation (DFG).

Table 1. Comparison of results for GP-VAR for activations with and without being passed through RELU. The comparison was conducted on 20 random splits with 10 known to 10 novel classes each.

Variant	CONV4	CONV5	FC6	FC7	FC8
normed-RELU	56.45 %	54.23 %	58.37 %	60.73 %	66.35 %
normed-noRELU	**64.17 %**	**60.02 %**	**71.67 %**	**64.78 %**	**67.73 %**

S1.1 GP-VAR

Observations In Table 1 the performance of modeling the training data by a Gaussian process regression (GP-VAR) is shown. We observe that there is a decisive difference in the resulting accuracies between using the RELU and the noRELU features. In fact, the overall highest accuracy is reached on the noRELU activations of the FC6 layer, with a mean accuracy of 71.67 %, which is over 10 % higher than using the RELU features of the same layer. This is not surprising if one considers that all negative values in the RELU features are set to 0 and thus any Gaussian-distributed feature dimension with negative values is cut off on one side. Furthermore, setting the negative values to 0 leads to an overall loss of information, as negative values also represent responses to certain patterns.

Setup Due to the observations mentioned above, we conducted all other GP-VAR-experiments with the normalized noRELU features of FC6.

S1.2 NCM (Euclid)

Observations As Table 2 shows, the mentioned loss of information seems to also have a negative effect on the accuracy of NCM(Euclid). The presented results show the expected difference in accuracy in CONV4, CONV5 and FC6, where the AUC is approximately 6 % to 10 % higher. However, in FC7 the results are in favor of RELU. Here, the results show a 5 % better performance on the original-RELU features in comparison to the normed-noRELU features.

Similarly, the FC8 layer is the most specialized one in regard to the original classification task of the network, as its outputs directly correspond to one of the original 365 classes of the Places365 dataset [3]. Furthermore, the RELU layers were included during the original training process, only positive outputs in the FC8 affected the training itself. As a consequence, FC8 shows almost no negative output values and the difference in accuracy is less than 1 %. In fact, this can also be observed for all other methods for which the results show comparably small differences in accuracy on RELU and noRELU in FC8.

Setup Despite the results in FC7 and FC8, we are mainly interested in the features that result in the best possible performance and therefore selected the normalized noRELU features from FC6 for our NCM-experiments on the other splits, as they showed the peak mean accuracy of 71.70 % in Table 2.

Table 2. Comparison of results for NCM(Euclid) for activations with and without being passed through RELU and with and without normalization to unit length. The comparison was conducted on 20 random splits with 10 known to 10 novel classes each.

Variant	CONV4	CONV5	FC6	FC7	FC8
normed-RELU	56.33 %	54.09 %	58.34 %	60.32 %	65.62 %
normed-noRELU	**64.07 %**	**60.1 %**	**71.7 %**	63.88 %	67.09 %
original-RELU	53.59 %	53.15 %	56.7 %	**68.63 %**	67.67 %
original-noRELU	56.25 %	56.56 %	62.35 %	61.34 %	**68.35 %**

Table 3. Comparison of results for Local-KNFST for activations with and without being passed through RELU. The comparison was conducted on 20 random splits with 10 known to 10 novel classes each.

Variant	CONV4	CONV5	FC6	FC7	FC8
normed-RELU	**69.81 %**	**70.48 %**	**71.38 %**	68.41 %	62.85 %
normed-noRELU	68.38 %	69.35 %	70.87 %	**68.48 %**	**63.03 %**

S1.3 Local-KNFST

Observations In contrast to the behavior of NCM(Euclid), Local-KNFST achieves better performance on the RELU features as shown in Table 3. Here, we can observe that the accuracies on RELU and noRELU differ by less than 2 %, which can be explained with the mapping of all training examples to single class points in the null space. As a consequence, the class representations are comparably similar on both RELU and noRELU, and similar local classes are taken into account for the novelty score calculation of each image.

Setup We selected the RELU features of FC6 for all other experiments, as they lead to the highest novelty detection accuracy.

S1.4 1-SVM

Observations Table 4 shows obtained accuracies for the 1-SVM method. Here, we can observe a up to 10 % difference in accuracy in the FC6 layer, in favor of the noRELU features. As in the case of GP-VAR, the loss of information using RELU is responsible for the difference between the accuracies on the different feature types. Hence, performing novelty detection on the noRELU features appears to be the best choice for every layer. Surprisingly, the 1-SVM achieves the highest performance of 66.57 % in FC8. However, this is still lower than the performance of the other baselines on the same features.

Setup We selected the noRELU features in FC6 in order to compare all methods using features from the same layer although activations obtained from higher layers lead to better performance in this special case.

Table 4. Comparison of results for `1-SVM` for activations with and without being passed through `RELU`. The comparison was conducted on 20 random splits with 10 known to 10 novel classes each.

Variant	CONV4	CONV5	FC6	FC7	FC8
normed-RELU	54.36 %	51.73 %	54.42 %	60.52 %	64.77 %
normed-noRELU	**63.38 %**	**58.93 %**	**64.75 %**	**66.16 %**	**66.57 %**

Table 5. Comparison of results for `Maximum-likelihood` for activations with and without being passed through `RELU` and with and without normalization to unit length. The comparison was conducted on 20 random splits with 10 known to 10 novel classes each.

Variant	CONV4	CONV5	FC6	FC7	FC8
normed-RELU	54.67 %	51.61 %	58.88 %	53.56 %	55.79 %
normed-noRELU	53.96 %	53.81 %	55.45 %	51.91 %	56.99 %
original-RELU	56.02 %	**60.37 %**	**63.84 %**	**65.76 %**	67.99 %
original-noRELU	**57.14 %**	57.68 %	63.56 %	61.37 %	**68.72 %**

S1.5 Maximum-likelihood

Observations Similarly to `Local-KNFST`, `Maximum-likelihood` shows the best average results on `RELU`, as can be seen in Table 5. This is unexpected since the noRELU features are closer to a Gaussian distribution. The reason might be the novelty scores of each image in respect to the known classes. As explained in Section 4.1 of the paper, we compute the negative log-likelihood and therefore calculate the logarithm of the class standard deviation. In other words, if the deviation of the dimensions is close to 0, the equation is dominated by the deviation, as the value of the logarithm diverges towards negative infinity. Hence, the distance between class mean and image becomes less significant. In particular, this can be observed in Table 5. The L1-normalization of the features leads to considerably small values in the mean and standard deviation of class means. Due to this, the novelty detection accuracy drops considerably in each layer if the `RELU` or noRELU features are normalized. This leads to a drop of approximately 10 % to 12 % in `FC7` and approximately 5 % to 8 % in `FC6`. In fact, this means that only the classes with the highest deviation actually play a role for novelty detection, and not the classes to which an image has the least distance to. On the other hand, setting all negative values to 0 reduces the deviation of all classes to similar values and thereby reduces the influence of the class deviations. As a consequence, the novelty detection accuracy without feature normalization is almost equal or higher on the `RELU` features in comparison to noRELU.

Setup As we want to compare our methods on the `FC6` features, we conducted all further experiments with `Maximum-likelihood` on the `FC6` `RELU` features without normalization.

S1.6 *K*-extremes

Observations Table 6 shows the accuracy of K−extremes for choices of features and K_{rel} (*i.e.,* a K relative to the feature dimension of each evaluated layer). The K_{rel} were selected in intervals of 10 %, with 5 % as addition in order to observe the behavior towards small numbers of extremes. The results reveal, that the peak accuracy of 71.72 % is achieved in FC6 with a K_{rel} of 0.10. As can also be seen, the optimal K_{rel} depends on the selected features. Furthermore, we observe that K−extremes is not robust in regards to the choice of K_{rel}, as the accuracies vary by up to over 5 % for different values. Hence, it can be difficult to find a good parameter for every dataset and selection of features.

Apart from the choice of K_{rel}, using different features affects the accuracy as well. In particular, K−extremes achieves higher overall performance on noRELU in comparison to RELU, which is due to the negative values affecting the calculation of the mean vector. Furthermore, normalizing the features also has a small effect on the accuracy, since the calculation of the mean is less affected by outliers in the activations.

Setup We selected the normalized noRELU FC6 features and a K_{rel} of 0.10 for the remaining experiments in the main paper, as they showed the best results in Table 6.

S1.7 Spearman

Observations Since Spearman can be seen as a more robust version of K−extremes, we only consider normalized features without RELU for the following evaluation because K−extremes showed best performance on this setup. The results in Table 7 show the accuracies of Spearman with varying K_{rel} using normalized noRELU features. In contrast to K−extremes, we can observe that the novelty detection accuracy of Spearman is robust over almost all choices of K_{rel}, independent of the layer. The results imply that Spearman achieves high performance even if no optimal K_{rel} can be found for a specific dataset, or if the underlying features change.

Setup In our experiments Spearman achieved the best result of 71.87 % on the normalized noRELU FC6 features with a K_{rel} of 0.10, as can be seen in Table 7. This is the same setup as for K−extremes which is not surprising, as Spearman is an extension of K−extremes. Therefore, we selected this setup for all further experiments.

S1.8 *K*-extremes-value

Observations In regards to the K−extremes−value approach, observations similar to K−extremes can be made in Table 8. Here the results also imply that no universally best choice of K_{rel} can be found and that the choice itself depends strongly on the underlying features. In contrast to K−extremes, the

Table 6. Comparison of results for the K−extremes approach on the splits with 10 known and 10 novel classes for activations with and without being passed through RELU and with and without normalization to unit length.

Variant	$\mathbf{K_{rel}}$	CONV4	CONV5	FC6	FC7	FC8
normed-RELU	0.05	63.11 %	66.73 %	68.44 %	67.14 %	62.88 %
normed-RELU	0.10	64.04 %	66.96 %	69.02 %	67.74 %	63.21 %
normed-RELU	0.20	64.58 %	65.63 %	66.99 %	68.32 %	62.62 %
normed-RELU	0.30	64.71 %	62.61 %	66.42 %	68.68 %	62.86 %
normed-RELU	0.40	63.67 %	63.6 %	65.55 %	67.67 %	62.85 %
normed-RELU	0.50	64.0 %	58.46 %	64.51 %	65.89 %	63.25 %
normed-RELU	0.60	59.52 %	56.03 %	64.18 %	63.8 %	63.5 %
normed-RELU	0.70	57.18 %	52.82 %	61.55 %	61.8 %	63.57 %
normed-RELU	0.80	56.08 %	50.87 %	57.27 %	59.68 %	63.23 %
normed-RELU	0.90	55.02 %	50.43 %	53.62 %	56.59 %	62.45 %
normed-RELU	1.00	50.0 %	50.0 %	50.0 %	50.0 %	50.0 %
normed-noRELU	0.05	61.4 %	65.1 %	70.81 %	69.55 %	**66.07 %**
normed-noRELU	0.10	62.35 %	65.35 %	**71.72 %**	70.08 %	65.49 %
normed-noRELU	0.20	63.84 %	65.43 %	71.7 %	70.38 %	64.58 %
normed-noRELU	0.30	64.76 %	65.35 %	71.58 %	**70.39 %**	64.05 %
normed-noRELU	0.40	65.43 %	65.21 %	71.46 %	70.18 %	64.19 %
normed-noRELU	0.50	65.93 %	65.01 %	71.37 %	70.08 %	64.27 %
normed-noRELU	0.60	66.3 %	64.63 %	71.09 %	69.91 %	64.92 %
normed-noRELU	0.70	**66.42 %**	64.23 %	70.99 %	69.71 %	65.39 %
normed-noRELU	0.80	66.41 %	63.23 %	70.89 %	69.34 %	65.13 %
normed-noRELU	0.90	65.57 %	61.54 %	70.15 %	68.54 %	63.51 %
normed-noRELU	1.00	50.0 %	50.0 %	50.0 %	50.0 %	50.0 %
original-RELU	0.05	63.23 %	**67.06 %**	68.29 %	67.3 %	62.84 %
original-RELU	0.10	64.19 %	67.0 %	68.8 %	67.91 %	63.12 %
original-RELU	0.20	64.73 %	65.57 %	66.75 %	68.44 %	62.12 %
original-RELU	0.30	64.78 %	62.71 %	66.24 %	68.7 %	62.36 %
original-RELU	0.40	63.61 %	63.62 %	65.54 %	67.63 %	62.34 %
original-RELU	0.50	63.94 %	58.39 %	64.35 %	65.88 %	62.79 %
original-RELU	0.60	59.4 %	56.06 %	63.46 %	63.71 %	63.02 %
original-RELU	0.70	57.13 %	52.79 %	61.6 %	61.65 %	62.97 %
original-RELU	0.80	55.94 %	50.77 %	56.37 %	59.54 %	63.11 %
original-RELU	0.90	54.77 %	50.58 %	54.68 %	56.57 %	62.25 %
original-RELU	1.00	50.0 %	50.0 %	50.0 %	50.0 %	50.0 %
original-noRELU	0.05	61.39 %	65.05 %	70.84 %	69.23 %	65.81 %
original-noRELU	0.10	62.36 %	65.32 %	71.68 %	69.75 %	65.21 %
original-noRELU	0.20	63.85 %	65.45 %	71.71 %	70.03 %	64.36 %
original-noRELU	0.30	64.77 %	65.36 %	71.62 %	70.09 %	63.75 %
original-noRELU	0.40	65.45 %	65.17 %	71.46 %	69.92 %	63.84 %
original-noRELU	0.50	65.95 %	64.94 %	71.35 %	69.8 %	64.13 %
original-noRELU	0.60	66.28 %	64.57 %	71.11 %	69.68 %	64.73 %
original-noRELU	0.70	66.4 %	64.1 %	71.01 %	69.51 %	65.07 %
original-noRELU	0.80	66.37 %	63.13 %	70.84 %	69.15 %	64.96 %
original-noRELU	0.90	65.51 %	61.45 %	70.17 %	68.47 %	63.37 %
original-noRELU	1.00	50.0 %	50.0 %	50.0 %	50.0 %	50.0 %

Table 7. Comparison of results for the Spearman approach on the splits with 10 known to 10 novel classes for activations without being passed through RELU and with normalization to unit length.

Variant	$\mathbf{K_{rel}}$	CONV4	CONV5	FC6	FC7	FC8
normed-noRELU	0.01	61.84 %	62.91 %	70.41 %	69.74 %	**66.78 %**
normed-noRELU	0.05	62.81 %	64.11 %	71.5 %	70.49 %	66.54 %
normed-noRELU	0.10	63.33 %	64.94 %	**71.87 %**	70.54 %	66.32 %
normed-noRELU	0.20	64.29 %	65.34 %	71.78 %	**70.62 %**	65.68 %
normed-noRELU	0.30	64.92 %	65.43 %	71.74 %	70.6 %	65.66 %
normed-noRELU	0.40	65.25 %	65.45 %	71.68 %	70.51 %	65.79 %
normed-noRELU	0.50	65.4 %	65.45 %	71.68 %	70.52 %	65.76 %
normed-noRELU	0.60	65.38 %	65.53 %	71.68 %	70.53 %	65.7 %
normed-noRELU	0.70	65.38 %	**65.55 %**	71.68 %	70.54 %	65.75 %
normed-noRELU	0.80	65.5 %	**65.55 %**	71.62 %	70.49 %	65.7 %
normed-noRELU	0.90	65.67 %	65.52 %	71.7 %	70.45 %	65.74 %
normed-noRELU	1.00	**66.02 %**	65.42 %	71.69 %	70.33 %	66.01 %

results in Table 8 show a significant difference in accuracies between the selected features. Reason for this difference might be the way the novelty scores are calculated. K-extremes only takes into account which dimensions are the relatively highest, but ignores the underlying values completely. But K-extremes-value directly uses the negative sum of the K largest activations and is therefore more affected by normalization and negative values.

Setup For further experiments, the normalized RELU features obtained from FC6 layer were taken and the parameter K_{rel} is set to 0.70.

S2 Additional Large-scale Results

In addition to the results in Section 4.4 of the main paper, we show here the systematic evaluation regarding the choice of K for the proposed K-extremes and Spearman methods. Obtained results are calculated on 20 random splits with 500 known and 500 unknown classes as explained in Section 4.4 of the main paper. Results are shown in Table 9. We observe that the value of $D \cdot 0.1$ for K performs best on this large-scale setting for both methods, too. The results support once more that Spearman is more robust towards the choice of K than the K-extremes method. The obtained accuracy of Spearman drops only by 0.38% for higher values of K while K-extremes loses 1.14% absolute accuracy.

S3 Qualitative Results

To obtain further insights into the proposed method, we provide qualitative results in Fig. S1 obtained with the K-extremes method. It can be seen that all images which received a low novelty score (Fig. S1, *middle*) belong to known classes (Fig. S1, *left*). Hence, the correct identifying of known as known is likely sufficient. However, images which achieve high novelty scores (Fig. S1, *right*) are both from known or unknown classes (*e.g.,* second row, first image should

Table 8. Comparison of results for the K-extremes-value approach on splits with 10 known and 10 novel classes.

Variant	$\mathbf{K_{rel}}$	CONV4	CONV5	FC6	FC7	FC8
normed-RELU	0.05	63.2 %	66.36 %	69.27 %	64.07 %	62.65 %
normed-RELU	0.10	63.92 %	66.64 %	69.84 %	64.47 %	62.88 %
normed-RELU	0.20	64.73 %	66.64 %	70.39 %	64.84 %	62.49 %
normed-RELU	0.30	65.13 %	66.64 %	70.6 %	65.16 %	63.0 %
normed-RELU	0.40	65.35 %	66.55 %	70.96 %	65.3 %	63.65 %
normed-RELU	0.50	65.52 %	66.45 %	71.23 %	65.44 %	64.14 %
normed-RELU	0.60	65.63 %	66.35 %	71.55 %	65.56 %	64.79 %
normed-RELU	0.70	65.64 %	66.16 %	**71.6 %**	65.56 %	65.06 %
normed-RELU	0.80	65.63 %	65.99 %	71.57 %	65.57 %	65.1 %
normed-RELU	0.90	**65.65 %**	65.39 %	70.49 %	65.19 %	63.89 %
normed-RELU	1.00	50.23 %	50.21 %	50.97 %	50.28 %	50.91 %
normed-noRELU	0.05	61.56 %	66.74 %	68.99 %	68.31 %	65.39 %
normed-noRELU	0.10	62.03 %	67.38 %	69.16 %	67.37 %	64.91 %
normed-noRELU	0.20	62.77 %	**67.56 %**	68.7 %	65.53 %	64.49 %
normed-noRELU	0.30	62.9 %	67.26 %	68.2 %	63.68 %	64.52 %
normed-noRELU	0.40	62.57 %	66.67 %	67.62 %	61.87 %	64.88 %
normed-noRELU	0.50	61.83 %	65.94 %	66.99 %	60.19 %	65.18 %
normed-noRELU	0.60	60.7 %	64.89 %	66.16 %	58.5 %	65.92 %
normed-noRELU	0.70	58.97 %	63.57 %	65.02 %	56.82 %	**66.08 %**
normed-noRELU	0.80	56.62 %	61.34 %	63.21 %	55.11 %	65.43 %
normed-noRELU	0.90	53.52 %	57.2 %	59.14 %	53.21 %	63.81 %
normed-noRELU	1.00	49.14 %	49.56 %	48.74 %	50.98 %	51.82 %
original-RELU	0.05	63.86 %	67.19 %	64.6 %	65.91 %	60.3 %
original-RELU	0.10	64.31 %	66.49 %	63.9 %	66.32 %	60.03 %
original-RELU	0.20	64.1 %	64.48 %	62.43 %	66.33 %	58.8 %
original-RELU	0.30	63.19 %	62.16 %	60.77 %	65.49 %	58.49 %
original-RELU	0.40	61.81 %	59.62 %	59.04 %	63.56 %	58.44 %
original-RELU	0.50	59.92 %	57.2 %	57.2 %	61.0 %	58.2 %
original-RELU	0.60	57.76 %	55.05 %	55.29 %	58.26 %	58.33 %
original-RELU	0.70	55.58 %	53.16 %	53.41 %	55.7 %	58.38 %
original-RELU	0.80	53.5 %	51.57 %	51.69 %	53.46 %	58.8 %
original-RELU	0.90	51.72 %	50.29 %	50.2 %	51.6 %	58.43 %
original-RELU	1.00	50.39 %	49.43 %	49.26 %	50.36 %	51.14 %
original-noRELU	0.05	62.69 %	65.96 %	68.57 %	69.58 %	62.71 %
original-noRELU	0.10	63.17 %	65.68 %	68.27 %	**69.76 %**	61.71 %
original-noRELU	0.20	63.81 %	64.07 %	66.92 %	68.79 %	60.77 %
original-noRELU	0.30	63.7 %	62.09 %	65.39 %	66.18 %	60.19 %
original-noRELU	0.40	63.01 %	60.11 %	63.56 %	62.98 %	59.93 %
original-noRELU	0.50	61.78 %	58.21 %	61.48 %	60.11 %	60.08 %
original-noRELU	0.60	60.08 %	56.38 %	59.29 %	57.71 %	60.27 %
original-noRELU	0.70	57.85 %	54.67 %	57.04 %	55.73 %	60.31 %
original-noRELU	0.80	55.24 %	53.01 %	54.77 %	54.01 %	60.46 %
original-noRELU	0.90	52.32 %	51.35 %	52.41 %	52.49 %	60.38 %
original-noRELU	1.00	48.83 %	49.66 %	49.85 %	51.05 %	51.82 %

Table 9. Results for 500:500 split using FC6 features before RELU and with normalization to unit length. Comparison of different ratios of K for K-extremes and Spearman.

Method	**K** [%D]	**500:500**
K-extremes	0.10	**54.56 %**
K-extremes	0.33	53.95 %
K-extremes	0.66	53.42 %
Spearman	0.10	54.44 %
Spearman	0.33	54.13 %
Spearman	0.66	54.06 %

Fig. S1. Qualitative results on a single 10-10 split obtained by our K-`extremes` method.

be known as `spatula` but is estimated as unknown). We attribute this fact to the large visual variability within the classes, *i.e.,* the test images differ strongly from the training set (`spatula` alone vs held by a person). Hence, it is unclear on which level novelty should actually be estimated. This ambiguity of the task itself can be the reason for the low overall accuracies.

S4 Statistical Significance of Differences in Results

In the main paper, we already observed small but observable differences in accuracy between the investigated methods. Here, we analyse the statistical significance of the differences by applying by a Wilcoxon signed rank test to the accuracies from the random 20 splits of each task. Tables 10 to 14 show results of the significance analysis for the task sizes 10:10 to 50:50 as in Section 4.2 of the main paper. Results are coded in red if a difference is not statistically significant, i.e., with $p \geq 0.05$. In general, it can be observed that the differences in accuracy in small splits are not significant. The more more classes are involved, the more significant are the differences.

An additional analysis for the large 500:500 split used in Section 4.4 of the main paper is shown in Table 15. It can be seen that a careful choice of K is a pre-requisite for statistically significant improvements in this difficult setup.

S5 Additional Visualization of Class-indicative Parts With EVS

In addition to Section 4.5 of the main paper, we show here the influence of K on the visualizations for a single image. Results are obtained based on the `CONV5` layer of a Places205-CNN [4] for computing the gradient maps [2] of the input image. The resulting visualizations are shown in Fig. S2. The obtained results imply that even for small values of K (*i.e.,* 10 and 100) some intuitively relevant image regions are covered. Raising the number of considered dimensions (*e.g.,* $K = 500$ and $K = 1000$) only leads to a more comprehensive coverage of the dog. This supports the intuition that the dimensions with strongest activations can capture class-specific patterns. A similar conclusion can be drawn by considering the obtained gradient maps themselves. Note that as in the main paper, the gradient maps are normalized independently and therefore do not allow for direct comparison of gradient strength.

Table 10. Significance of differences in accuracy on 10:10 split, based on a Wilcoxon signed rank test. Each value in the table represents the probability that the accuracies of two methods, on the same 20 random splits, originate from the same distribution.

	1-SVM	NCM(Euclid)	GP-VAR	Local-KNFST	Maximum-likelihood	Spearman	K-extremes	K-extremes-value
1-SVM	-	0.03 %	0.03 %	0.01 %	29.59 %	0.03 %	0.04 %	0.04 %
NCM(Euclid)	0.03 %	-	76.52 %	76.52 %	0.03 %	45.53 %	76.52 %	88.13 %
GP-VAR	0.03 %	76.52 %	-	82.28 %	0.03 %	41.15 %	55.03 %	82.28 %
Local-KNFST	0.01 %	76.52 %	82.28 %	-	0.01 %	55.03 %	60.12 %	65.42 %
Maximum-likelihood	29.59 %	0.03 %	0.03 %	0.01 %	-	0.02 %	0.02 %	0.02 %
Spearman	0.03 %	45.53 %	41.15 %	55.03 %	0.02 %	-	16.72 %	45.53 %
K-extremes	0.04 %	76.52 %	55.03 %	60.12 %	0.02 %	16.72 %	-	68.13 %
K-extremes-value	0.04 %	88.13 %	82.28 %	65.42 %	0.02 %	45.53 %	68.13 %	-

Table 11. Significance of differences in accuracy on 20:20 split, based on a Wilcoxon signed rank test. Each value in the table represents the probability that the accuracies of two methods, on the same 20 random splits, originate from the same distribution.

	1-SVM	NCM(Euclid)	GP-VAR	Local-KNFST	Maximum-likelihood	Spearman	K-extremes	K-extremes-value
1-SVM	-	0.01 %	0.01 %	0.01 %	62.74 %	0.01 %	0.01 %	0.01 %
NCM(Euclid)	0.01 %	-	94.05 %	23.22 %	0.01 %	9.3 %	10.84 %	10.05 %
GP-VAR	0.01 %	94.05 %	-	20.43 %	0.01 %	7.31 %	10.84 %	12.59 %
Local-KNFST	0.01 %	23.22 %	20.43 %	-	0.01 %	57.55 %	62.74 %	82.28 %
Maximum-likelihood	62.74 %	0.01 %	0.01 %	0.01 %	-	0.01 %	0.01 %	0.01 %
Spearman	0.01 %	9.3 %	7.31 %	57.55 %	0.01 %	-	55.03 %	68.13 %
K-extremes	0.01 %	10.84 %	10.84 %	62.74 %	0.01 %	55.03 %	-	85.19 %
K-extremes-value	0.01 %	10.05 %	12.59 %	82.28 %	0.01 %	68.13 %	85.19 %	-

Table 12. Significance of differences in accuracy on 30:30 split, based on a Wilcoxon signed rank test. Each value in the table represents the probability that the accuracies of two methods, on the same 20 random splits, originate from the same distribution.

	1-SVM	NCM(Euclid)	GP-VAR	Local-KNFST	Maximum-likelihood	Spearman	K-extremes	K-extremes-value
1-SVM	-	0.01 %	0.01 %	0.01 %	45.53 %	0.01 %	0.01 %	0.01 %
NCM(Euclid)	0.01 %	-	33.17 %	39.05 %	0.01 %	2.06 %	3.66 %	9.3 %
GP-VAR	0.01 %	33.17 %	-	37.03 %	0.01 %	2.51 %	5.22 %	10.05 %
Local-KNFST	0.01 %	39.05 %	37.03 %	-	0.01 %	33.17 %	41.15 %	47.81 %
Maximum-likelihood	45.53 %	0.01 %	0.01 %	0.01 %	-	0.01 %	0.01 %	0.01 %
Spearman	0.01 %	2.06 %	2.51 %	33.17 %	0.01 %	-	60.12 %	35.07 %
K-extremes	0.01 %	3.66 %	5.22 %	41.15 %	0.01 %	60.12 %	-	41.15 %
K-extremes-value	0.01 %	9.3 %	10.05 %	47.81 %	0.01 %	35.07 %	41.15 %	-

Table 13. Significance of differences in accuracy on 40:40 split, based on a Wilcoxon signed rank test. Each value in the table represents the probability that the accuracies of two methods, on the same 20 random splits, originate from the same distribution.

	1-SVM	NCM(Euclid)	GP-VAR	Local-KNFST	Maximum-likelihood	Spearman	K-extremes	K-extremes-value
1-SVM	-	0.01 %	0.01 %	0.01 %	3.04 %	0.01 %	0.01 %	0.01 %
NCM(Euclid)	0.01 %	-	57.55 %	1.52 %	0.02 %	0.03 %	0.04 %	1.52 %
GP-VAR	0.01 %	57.55 %	-	1.69 %	0.03 %	0.03 %	0.02 %	1.24 %
Local-KNFST	0.01 %	1.52 %	1.69 %	-	0.01 %	94.05 %	76.52 %	26.27 %
Maximum-likelihood	3.04 %	0.02 %	0.03 %	0.01 %	-	0.01 %	0.01 %	0.01 %
Spearman	0.01 %	0.03 %	0.03 %	94.05 %	0.01 %	-	5.69 %	3.33 %
K-extremes	0.01 %	0.04 %	0.02 %	76.52 %	0.01 %	5.69 %	-	3.04 %
K-extremes-value	0.01 %	1.52 %	1.24 %	26.27 %	0.01 %	3.33 %	3.04 %	-

Table 14. Significance of differences in accuracy on 50:50 split, based on a Wilcoxon signed rank test. Each value in the table represents the probability that the accuracies of two methods, on the same 20 random splits, originate from the same distribution.

	1-SVM	NCM(Euclid)	GP-VAR	Local-KNFST	Maximum-likelihood	Spearman	K-extremes	K-extremes-value
1-SVM	-	0.01 %	0.01 %	0.03 %	12.59 %	0.01 %	0.01 %	0.01 %
NCM(Euclid)	0.01 %	-	10.05 %	9.3 %	0.01 %	4.38 %	15.6 %	52.57 %
GP-VAR	0.01 %	10.05 %	-	6.74 %	0.01 %	8.59 %	13.54 %	37.03 %
Local-KNFST	0.03 %	9.3 %	6.74 %	-	0.03 %	0.1 %	0.19 %	5.22 %
Maximum-likelihood	12.59 %	0.01 %	0.01 %	0.03 %	-	0.01 %	0.01 %	0.01 %
Spearman	0.01 %	4.38 %	8.59 %	0.1 %	0.01 %	-	97.02 %	0.45 %
K-extremes	0.01 %	15.6 %	13.54 %	0.19 %	0.01 %	97.02 %	-	0.32 %
K-extremes-value	0.01 %	52.57 %	37.03 %	5.22 %	0.01 %	0.45 %	0.32 %	-

Table 15. Significance of differences in accuracy on 500:500 split, based on a Wilcoxon signed rank test. Each value in the table represents the probability that the accuracies of two methods, on the same 20 random splits, originate from the same distribution.

	NCM(Euclid)	Spearman 0.10	Spearman 0.33	Spearman 0.66	K-extremes 0.10	K-extremes 0.33	K-extremes 0.66
NCM(Euclid)	-	0.15 %	10.05 %	39.05 %	0.06 %	62.74 %	0.01 %
Spearman 0.10	0.15 %	-	0.01 %	0.01 %	0.09 %	0.01 %	0.01 %
Spearman 0.33	10.05 %	0.01 %	-	0.01 %	0.01 %	0.01 %	0.01 %
Spearman 0.66	39.05 %	0.01 %	0.01 %	-	0.01 %	0.01 %	0.01 %
K-extremes 0.10	0.06 %	0.09 %	0.01 %	0.01 %	-	0.01 %	0.01 %
K-extremes 0.33	62.74 %	0.01 %	0.01 %	0.01 %	0.01 %	-	0.01 %
K-extremes 0.66	0.01 %	0.01 %	0.01 %	0.01 %	0.01 %	0.01 %	-

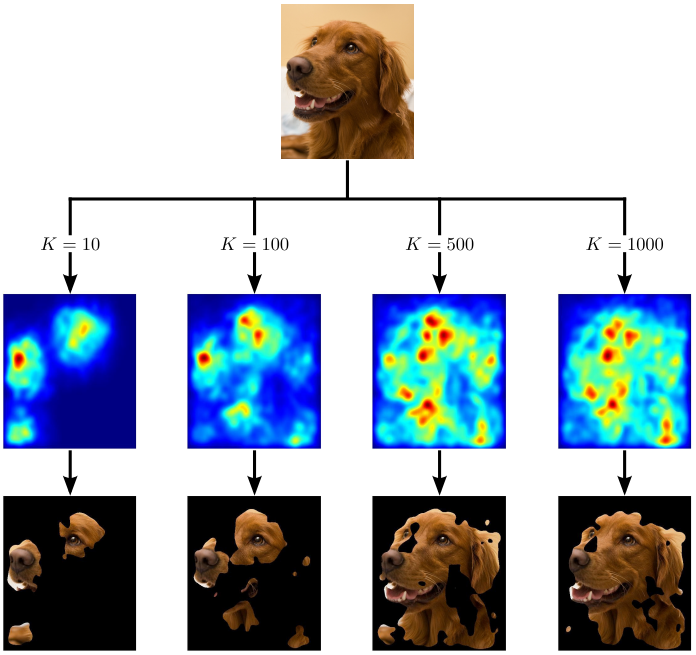

Fig. S2. Saliency map related to the K-highest outputs for different numbers of K.

References

1. Kemmler, M., Rodner, E., Denzler, J.: One-class classification with gaussian processes. In: Asian Conference on Computer Vision (ACCV) (2010)
2. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
3. Zhou, B., Khosla, A., Lapedriza, A., Torralba, A., Oliva, A.: Places: An image database for deep scene understanding. arXiv preprint arXiv:1610.02055 (2016)
4. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Neural Information Processing Systems (NIPS) (2014)