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

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. Achieving or even surpassing human-level accuracy became recently possible in a variety of application scenarios due to the rise of convolutional neural networks (CNNs) trained from large datasets. However, solving supervised visual recognition tasks by discriminating among known categories is only one side of the coin. In contrast to this, novelty detection is still an unsolved task where instances of yet unknown categories need to be identified. Therefore, we propose to leverage the powerful discriminative nature of CNNs to novelty detection tasks by investigating class-specific activation patterns. More precisely, we assume that a semantic category can be described by its *extreme value signature*, that specifies which dimensions of deep neural activations have largest values. By following this intuition, we show that already a small number of high-valued dimensions allows to separate known from unknown categories. Our approach is simple, intuitive, and can be easily put on top of CNNs trained for vanilla classification tasks. We empirically validate the benefits of our approach in terms of accuracy and speed by comparing it against established methods in a variety of novelty detection tasks derived from ImageNet. Finally, we show that visualizing extreme value signatures allows to inspect class-specific patterns learned during training which may ultimately help to better understand CNN models.

1 Introduction

The availability of large annotated datasets and efficient training algorithms for supervised deep learning lead the path to a striking increase in performance of current visual recognition systems [20]. For several applications, however, training discriminative models is not sufficient or even not possible since classes are either not known in advance, or not completely covered by a fixed training dataset. Due to this reason, algorithms are needed that not only discriminate among known categories but which additionally detect instances of yet unknown categories. This important task is known as novelty detection [22,14,15,5,27,4] or open-set recognition [21,2,3] and we present a simple method for this task.

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

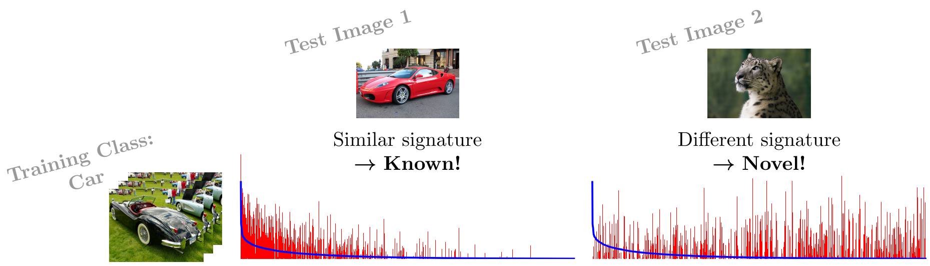

Fig. 1. Shown are neural activation patterns of two test images obtained by a CNN. For the known class, the mean vector of activations which serves as class prototype is sorted by value in descending order and shown in blue. Neural activations of two test images are ordered according to this permutation and are shown in red. For an example of the same category (*left*), the distribution of activation strength follows roughly the same trend as the class prototype. In contrast, the image of a novel category (*right*) has strong activations in a different set of dimensions.

Our approach is based on characterizing known classes with extreme value signatures (EVS) computed for neural activation vectors. Extreme value signatures specify which dimensions of deep neural activations have largest values. Previous work has shown that a surprisingly large fraction of the original image signal can be recovered from the high-scored activations [16]. We go one step further and show that the signatures even capture discriminative information not only with respect to known but also with respect to unknown classes. EVS are therefore suited for novelty detection as visualized in Fig. 1. For an efficient realization, we follow a prototype-based approach, where each known class is encoded by the EVS of the class-mean activation vector. In summary, we found that our extreme value signatures are compact, effective for novelty detection, and a promising representation for further improvements of the terra incognita of open-set recognition.

2 Related Work

In the following, we review related techniques for novelty detection which are most relevant for our approach. For an extensive overview far beyond the current scope, the interested reader is referred to the overview article by Pimentel *et al.* [19].

One-class Classification Presumably the most popular approach for novelty detection is the one-class SVM by Schoelkopf *et al.* [22]. Similar to its two-class pendant, a hyperplane is computed which separates all examples of one class from the origin of the feature space with maximal margin. For certain choices of the kernel function, this is identical to the equally popular Support Vector Data Description by Tax and Duin [26], which estimates an enclosing hypersphere of smallest radius. As a third popular technique, the work of Kemmler *et al.* shows how to apply Gaussian processes regression models to one-class classification [14]. Although all techniques are well established, they miss a sound formulation for multi-class scenarios where more than one category is known.

Multi-Class Novelty Detection To overcome the one-class limitation, several approaches have been introduced recently. Vinokurow *et al.* propose to rely on an ensemble of binary classifiers to detect novel classes [27] . Each binary classifier is trained to discriminate between the raw novelty score of a test sample and the average raw novelty score of a known class. These raw scores involve confidence values which can be obtained from any multi-class classification algorithm. A different approach was proposed by Kenk *et al.* [15] based on calculating a novelty score from the Hellinger distance between color histograms. Jumutc and Suykens [13] proposed three extensions of their earlier work on Supervised Novelty Detection (SND) [12] for multi-class novelty detection. Bodesheim *et al.* introduced the Kernel Null Foley-Sammon Transform (KNFST) for multi-class novelty detection [5]. The learned transformation maps all examples from a single class to a unique point in the null space. Thereby, novelty of unseen instances can be estimated by the minimal distance to all available class points in the null space. The authors extended their approach in [4] using local learning. KNFST models are learned for each test sample separately using only the K most similar training examples which leads to exemplar-specific novelty detection models. Although these approaches come with their own benefits, they require time-consuming training of the novelty detection model. Furthermore, they do not take the nature of underlying representations into account. In contrast, we exploit that CNNs can act as a joint model for representation and classification and leverage this idea to skip additional training of novelty detection models.

Open-set Recognition The idea of open-set risk for open set recognition was presented by Scheirer *et al.* [21]. Their formulation results in a special kind of one-class SVM, referred to as 1-vs-set machine. The idea of open set risk was further refined by Bendale and Boult [2]. The authors present the nearest-non-outlier algorithm, which is an extension of the nearest-class-mean algorithm for open-set scenarios. Similar in spirit is the openmax-approach by the same authors [3]. They propose to replace the commonly used `softmax` layer in CNNs with an `openmax` layer which classifies into known categories or “unknown”. We follow their approaches by re-using existing classification CNNs as-is, but aim at predicting novelty based on neural activations of arbitrary layers.

Binarizing CNN Activations A technique which is conceptually similar to EVS was proposed by Li *et al.* [16] for learning with few examples. They propose to binarize CNN activations using the K largest values and conclude that “discriminative information within CNN activation is mostly embedded in the dimension indices of the K largest magnitudes”. A similar analysis has been done in [8] with the goal of reconstructing original images from given CNN activations. Finally, a similar transformation of activations was used in [17] for regularizing autoencoders and in [11] for stochastic pooling. We follow their path by using EVS for image encoding and show how to analyze them to predict novelty in multi-class scenarios.

3 Extreme Value Signatures

Neural Activations for Novelty Detection Many popular approaches in machine vision use activations of pre-trained CNNs as off-the-shelf image representations [7,23]. Our current understanding is that the majority of trainable layers in CNNs (especially convolutional layers) becomes sensitive to the specific characteristics that are common in natural images when fed with millions of training images. In particular, it has been shown that filter masks of convolutional layers become sensitive to visual "elements" that can be commonly found in natural images [28]. Whereas these elements mainly correspond to low-level texture patterns for lower layers, they can be related to semantically meaningful object parts in higher layers [24]. This data-driven approach to representation learning is obviously appealing in several aspects, especially when large amounts of data are available [20]. Although the respective CNNs have originally been trained for discriminating among known categories, we show that their extreme value statistics can also be used to detect instances of novel categories.

Extreme Value Signatures for Individual Images Let $\mathbf{x} \in \mathbb{R}^D$ be the activation extracted at a chosen layer when applying a given CNN to a single image. Babenko *et al.* showed in [1] that this representation still maintains a large amount of discriminative information after binarization with a fixed threshold τ. Hence, the binary vector $\mathbf{b}$:

$$\mathbf{b}\left(\mathbf{x}, \tau\right) = \left(\delta\left(\mathbf{x}\left[d\right] > \tau\right)\right)_{d=1}^{D} \tag{1}$$

can serve as a substitute for $\mathbf{x}$ when the threshold τ is chosen appropriately since it translates $\mathbf{x}$ into a binary representation. Note that for the ease of readability, $\mathbf{x}\left[d\right]$ denotes the d^{th} dimension of $\mathbf{x}$ and the function $\delta\left(v\right)$ maps to 1 or 0 if v is true or false, respectively.

Despite the successful application of binarized neural activations in [1], relying on a constant threshold might be too restrictive for the complexity of visual recognition tasks. In a more general setting, we can replace this constant threshold by a threshold function $T\left(\cdot\right)$ which returns a threshold specifically tailored to each example $\mathbf{x}$:

$$\mathbf{b}\left(\mathbf{x}, T\left(\mathbf{x}\right)\right) = \left(\delta\left(\mathbf{x}\left[d\right] > T\left(\mathbf{x}\right)\right)\right)_{d=1}^{D} \quad . \tag{2}$$

Our extreme value signature follows this general concept. Let us therefore denote $\boldsymbol{\pi}_{\mathbf{x}}$ the permutation that brings $\mathbf{x}$ into descending order:

$$\underset{i,j\in\{1,\ldots,D\},i<j}{\forall} : \mathbf{x}\left[\boldsymbol{\pi}_{\mathbf{x}}\left[i\right]\right] \geq \mathbf{x}\left[\boldsymbol{\pi}_{\mathbf{x}}\left[j\right]\right] \quad . \tag{3}$$

Thereby, we obtain a threshold $T_{\text{rank\_K}}\left(\mathbf{x}\right)$ for each $\mathbf{x}$ using the K-highest activation:

$$T_{\text{rank\_K}}\left(\mathbf{x}\right) = \mathbf{x}\left[\boldsymbol{\pi}_{\mathbf{x}}\left[K\right]\right] \quad . \tag{4}$$

The resulting binary vector $\mathbf{b}\left(\mathbf{x}, T_{\text{rank\_K}}(\mathbf{x})\right)$ can therefore be seen as an indicator for the K highest values of $\mathbf{x}$, which we refer to as the EVS of $\mathbf{x}$. When thinking of visual recognition scenarios, we expect that two images that contain similar visual concepts also lead to a similar set of extreme dimensions in the resulting neural activations.

Hence, their binarized codes based on $T_{\text{rank\_K}}$ should be close as well. Therefore, we can apply the inner product

$$\gamma(\mathbf{x}', \mathbf{x}) = 1 - \frac{\mathbf{b}(\mathbf{x}', T_{\text{rank\_K}}(\mathbf{x}'))^{\mathrm{T}}\, \mathbf{b}(\mathbf{x}, T_{\text{rank\_K}}(\mathbf{x}))}{K} \tag{5}$$

to estimate the novelty of a test example $\mathbf{x}'$ with respect to a known example $\mathbf{x}$. Note that the normalization $1 - \frac{\cdot}{K}$ is only required to transform the score into $[0, 1]$ such that large scores indicate novelty.

Novel Class Detection with Extreme Value Signatures The previous derivations only focused on the difference of a novel example $\mathbf{x}'$ and a known example $\mathbf{x}$ with respect to their extreme value signatures. In multi-class scenarios, we can additionally exploit the class label which is associated with every training example. To obtain an extreme value signature for an entire known class, we have a variety of options to choose from. We empirically found that computing the mean vector $\boldsymbol{\mu}$ of activations from all examples within a class and determining the K-highest dimensions thereof is simple, easy to implement, and works well in practice. The implicit assumption is that the K-highest dimensions correspond to specific patterns of a known class. Hence, if a test image shows a similar extreme-value signature, it is likely to contain similar visual patterns.

For multiple known classes, we follow [5] and take the minimum over all class distances as measured in Eq. (5) as resulting novelty score. Thereby, a test example will only be considered as novel if it is different from the extreme-value signatures of *all* known categories M. Hence, our final multi-class novelty score $\gamma_{\text{MC}}(\mathbf{x}')$ can be expressed as:

$$\gamma_{\text{MC}}(\mathbf{x}') = \min_{1 \leq m \leq M} \gamma(\mathbf{x}', \boldsymbol{\mu}_m) \ . \tag{6}$$

Novelty Detection Based on Permutation Distances Comparing the K-highest feature dimensions, as introduced in Eq. (6), does not take the ranking of the K-highest dimensions into account. In consequence, the score can be extremely sensitive to the choice of K. Intuitively, small Ks would not cover all dimensions which are relevant for a category, whereas large Ks would also include irrelevant dimensions. Thereby, already marginal reordering among the highest-valued dimensions could lead to overestimating novelty in the first case. In contrast to this, an underestimation of novelty or a wrong assignment to a known class could happen in the latter case. To overcome these issues, we propose to consider the ranking among the K-highest dimensions by comparing the highest activations based on the Spearman footrule distance [6].

The Spearman footrule distance allows for calculating distances between two D-dimensional permutations $\boldsymbol{\pi_1}$ and $\boldsymbol{\pi_2}$:

$$\mathrm{d}_{\text{Spearman}}(\boldsymbol{\pi_1}, \boldsymbol{\pi_2}) = \sum_{k=1}^{D} \sum_{j=1}^{D} \delta_{\boldsymbol{\pi_2}[j], \boldsymbol{\pi_1}[k]} \cdot |k - j| \ \ , \tag{7}$$

where the Kronecker delta $\delta_{\boldsymbol{\pi_2}[j], \boldsymbol{\pi_1}[k]}$ filters relevant indices since it has value 1 only if the values of $\boldsymbol{\pi_2}[j]$ and $\boldsymbol{\pi_1}[k]$ are equal, otherwise it has value 0. The difference between j and k measures then the absolute displacement in the two permutations. To estimate the novelty of a test example $\mathbf{x}'$ with respect to a known example $\mathbf{x}$, we can now apply the footrule distance on the permutations $\boldsymbol{\pi}_{\mathbf{x}'}$ and $\boldsymbol{\pi}_{\mathbf{x}}$ as defined in Eq. (3) which re-arranges in descending order the values of $\mathbf{x}'$ and $\mathbf{x}$, respectively. As in Eq. (6), this can finally be transferred to the multi-class scenario by minimizing distance over *all* M class mean vectors:

$$\gamma_{\text{Spearman}}(\mathbf{x}') = \min_{1 \leq m \leq M} \sum_{k=1}^{K} \sum_{j=1}^{D} \delta_{\boldsymbol{\pi}_{\mathbf{x}'}[j], \boldsymbol{\pi}_{\boldsymbol{\mu}_m}[k]} \cdot |k - j| \quad . \tag{8}$$

Note that the Spearman footrule distance originally suggests to compare the ranking of all dimensions. Instead, we propose to only compare the ranking of the K highest activations in $\mathbf{x}'$ with their ranking in each class prototype $\boldsymbol{\mu}_m$.

4 Experiments

We investigated our approach for multi-class novelty detection both quantitatively in comparison with state-of-the-art techniques (Sections 4.2 to 4.4) and qualitatively (Section 4.5). As benchmarking set, we chose the popular ImageNet dataset from ILSVRC'12 [20] and derived novelty detection tasks of varying difficulty. In all experiments, we encode images with neural activations of a AlexNet-Places365-CNN [29] without any fine-tuning. Thereby, none of the involved classes was already observed during model training[1]. We experimented with activations from layers `CONV4` to `FC8` with and without `RELU` and feature normalization to unit length. In the following, we only show results for `FC6` which lead to highest overall accuracy, but provide evaluations for all layers in the supplementary material (S.1).

4.1 Baseline Methods

Besides our two introduced approaches based on extreme value signatures (denoted as K`-extremes` and `Spearman`), we chose several techniques for comparison as reviewed in Section 2. The presumably simplest baseline is to transfer the nearest-class mean approach [18] from classification to novelty detection. In this spirit, the euclidean distance to the closest class mean serves as novelty score (`NCM(Euclid)`). Modeling each class by a Gaussian distribution is similarly simple [10]. Computing the negative log-likelihood for each class and returning the largest value thereof serves as a simple estimate of novelty (`Maximum-likelihood`). Alternatively, one-class SVMs [22] can be trained for each class. Distances to all M decision boundaries are maximum-pooled as suggested in [4] (`1-SVM`). Modeling the entire training data by a Gaussian process regression [14] allows for computing the predictive variance for unseen data (`GP-VAR`). Exemplar-specific novelty detection models are obtained by local KNFST

[1] We assume that fine-tuning networks for known classes would further improve the overall accuracy since activation patterns are expected to become specific for known classes. However, fine-tuning networks for all evaluated tasks and splits would be too time consuming. Therefore, the used Places-CNN ensures a fair comparison since it was not trained with any involved ImageNet class.

as introduced in [4] (`Local-KNFST`). Finally, we compare against K-`extremes-value` which is inspired by our extreme value signatures but directly uses the negative sum of the K largest activations as an indicator for novelty instead of their ranking. If not specified otherwise, we follow the setup described in [4] regarding the choice of hyperparameters.

4.2 Multi-class Novelty Detection on ImageNet Subsets

We first put a focus on accuracy rather than scalability. Hence, we start with an evaluation on ILSVRC'12 data with small and medium sized splits.

Setup We follow Bodesheim *et al.* [4] and use the setup initially described in [5]. Therefore, we randomly select different subsets of the ILSVRC'12 classes and split them into known and novel categories. For now, we consider all split sizes which were used in [4]. This results in scenarios with ratios of 10:10 (*i.e.,* 10 known and 10 novel classes) up to 50:50. For each known class, a random training set of 100 samples is drawn. From the remaining images as well as from all elements of the novel classes, 50 samples per class are randomly drawn to serve as test set. We average results over 20 random splits for each task to allow for statistically valid conclusions. To allow for direct comparison, we use the same class splits and selected samples as in [4]. Accuracy is measured by AUC [9]. The size of the neighborhood for each split for `Local-KNFST` is set to best performing values according to [4]. For K-`extremes`, `Spearman`, and K-`extremes-value`, we exhaustively tested K over a broad range and report the best results here ($K = D \cdot 0.1$ for K-`extremes` and `Spearman`, $K = D \cdot 0.7$ for K-`extremes-value`). Furthermore, we evaluated neural activations of different layers with or without passing them through `RELU`s and with or without normalization to unit length. Here, we only report the best results for each combination of method, encoding, and parameter setting. For the sake of completeness and reproducibility, we provide results obtained with all settings in the supplementary material (S.1).

Results In the first columns of Table 1, results for baseline methods in comparison with our approaches are shown[2]. As a general trend, we observe that the accuracy of all methods drops with an increasing number of known and unknown classes. This behavior is not surprising, since random chance for miss-classification increases with more available classes. In addition, we find that the difference in the resulting accuracies of the tested methods is only marginal on the 10:10 split. The only notable exceptions are `1-SVM` and `Maximum-likelihood` which are clearly inferior to the remaining methods. This pattern becomes even more dominant for an increasing number of classes. To check for statistical significance of the small but observable differences in accuracy, we performed a Wilcoxon signed rank test. Due to the matter of space, the results can be found in the supplementary (S.4). The analysis can be summarized as follows: there are no significant differences on small splits, but significant differences on large splits slightly in favor of our proposed extreme value signatures.

The results also reveal that summing up values of dimensions with largest values (K-`extremes-value`) is not superior to simply considering the indices of the di-

[2] Note that the reported results for `Local-KNFST` differ from [4] since we use CNN features instead of dense SIFT features which results in improved performance.

Table 1. Results from different tasks for multi-class novelty detection using `FC6` features and best parameter settings per method (averaged over 20 splits per task).

Method	10:10	20:20	30:30	40:40	50:50	500:500
`NCM(Euclid)`	71.70 %	67.17 %	65.14 %	62.37 %	62.45 %	53.99 %
`Maximum-likelihood`	63.84 %	61.07 %	60.13 %	58.77 %	58.18 %	-
`Local-KNFST` [4]	71.38 %	**68.16 %**	65.79 %	64.16 %	61.56 %	-
`GP-VAR` [14]	71.67 %	67.19 %	65.19 %	62.33 %	62.51 %	-
`1-SVM` [22]	64.75 %	60.72 %	59.54 %	57.26 %	57.52 %	-
K`-extremes-value`	71.60 %	68.01 %	65.98 %	63.77 %	62.33 %	-
K`-extremes` **(ours)**	71.72 %	68.04 %	66.21 %	**64.38 %**	63.05 %	**54.56 %**
`Spearman` **(ours)**	**71.87 %**	68.14 %	**66.25 %**	64.24 %	**63.06 %**	54.44 %

mensions themselves. Therefore, we can conclude that the actual values of neural activations can be ignored when their relative order is known. The results also imply that `Spearman` is not clearly advantageous in comparison to the vanilla K`-extremes` method. However, this is not surprising when we consider that the main advantage of `Spearman` is the robustness towards wrong choices of K. Since Table 1 only shows the configuration of each method which lead to highest accuracy, the proper selection K is neglected. The overview of results from all settings in the supplementary material (S.1) underlines `Spearman`'s robustness regarding the choice of K. As a final note, the supplementary also contains a qualitative result showing most and least novel images (S.3).

4.3 Computation Time Analysis

Besides accuracy, computation time is one of the most critical aspects of algorithms. Therefore, we investigate the execution times of testing a single image for each of the evaluated novelty detection methods.

Setup We conduct a computation time analysis on a desktop computer with an Intel Core 2 Quad CPU with 2.4 GHz and 8 GB of system RAM. For `Local-KNFST`, `GP-VAR` and `1-SVM`, we use the MATLAB code provided by [4]. All remaining methods are implemented in python. Computation times of each method are evaluated on a 10:10 split and a 50:50 split and averaged over all test examples in each split.

Results In Fig. 2, we show the relationship between accuracy and computation time. The observable relation is not unexpected: `NCM(Euclid)` is the fastest baseline since it only requires simple distance calculations. On the other side of the spectrum, `Local-KNFST` is an order of magnitude slower due to the necessity of training a model specifically for each test sample. The remaining techniques are roughly equally fast with $\sim$ 10ms per evaluation. However, K`-extremes` and `Spearman` are slowest wrt. to absolute numbers. We attribute this to the explicit sorting of feature vectors in our non-optimized implementation and assume that an optimized implementation involving bit-level operations could reduce the required computation time.

Fig. 2. Comparison of all approaches on the 10:10 split (*left*) and the 50:50 split (*right*) taking accuracy and computation time into account (computation times averaged over all examples in each split).

4.4 Large-scale Multi-class Novelty Detection on ImageNet

The results presented so far in Section 4.2 imply that the accuracy of all methods drops the more classes are involved. Hence, we were interested in conducting a large scale analysis for further investigation.

Setup We split the available classes of the ILSVRC'12 data randomly in half which results in 500 known and 500 unknown classes. The remaining setup is kept identical to Section 4.2 and we present an exhaustive evaluation of all investigated settings in the supplementary material (S.2). Note that `Local-KNFST` would not be applicable in this scenario in terms of computation time as shown in Section 4.3. Additionally, kernel dimensions are also too large for our available implementation of `GP-VAR` and `1-SVM`. Therefore, we only compare the proposed `Spearman` and K`-extremes` criterion with the remaining (and best performing) baseline `NCM(Euclid)`.

Results Results are shown in the last column of Table 1. As expected, the overall accuracy drops in comparison to the setup in Section 4.2. The proposed K`-extremes` performs best closely followed by `Spearman`. Both methods are able to outperform `NCM(Euclid)` by a small but significant margin. The supplementary material contains a significance analysis (S.4) as well as further evidence for the robustness of `Spearman` towards the choice of K (S.2). Although we conclude that our proposed novelty detection methods can be successfully applied in large scale scenarios, our best-performing method improves over random guessing by only less than 5%. Hence, novelty detection in large-scale scenarios still remains an unsolved problem which sorely needs increased attention.

4.5 Visualizing Class-indicative Image Parts With EVS

In addition to the quantitative estimation of novelty as presented so far, we can further exploit the comparison of EVS to assess which parts of a novel image are indicative for a known category. A visualization is shown in Fig. 3.

Setup We compute gradient maps as suggested by [25] using the `FC6` layer of a Places205-CNN [30] [3]. To visually inspect class-indicative image parts, we set all entries to 1 which correspond to the K-highest feature dimension of class m ($K = 0.1{\cdot}D$).

[3] Due to implementation constraints we applied a different network as in Sections 4.2 to 4.4.

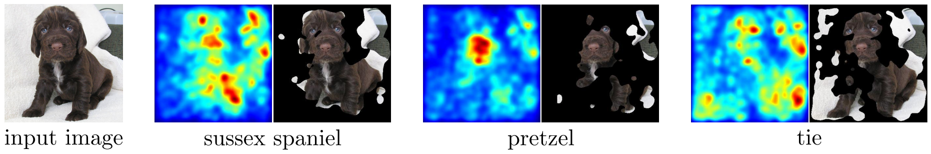

Fig. 3. Which parts of a novel image are characteristic for known classes? Saliency maps obtained from comparing extreme value signatures allow for visual analysis.

All remaining values are set to 0 and the derivative of this target vector wrt. the input image is computed using backprop. The generated gradient map is then smoothed for better visualization with a Gaussian kernel (size 20×20 pixels, $\sigma = 5$). After resizing of the smoothed map to the original image size and normalization[4], we threshold all values against $1/3$ and consider all pixels above this value as relevant. Irrelevant pixels are blacked out. Visualizations for different values of K are provided in the supplementary material (S.5).

Results Using this visualization heuristic, it can be seen that only few image parts would be indicative for the category `pretzel`. On the contrary, the class `tie` is supported by image regions from foreground and background making it a vague indicator as well. Only for the (correct) category `Sussex spaniel`, the indicative image regions are entirely in the foreground and closely align with the object boundary. Hence, we conclude that EVS are indeed indicative for class-specific image parts and allow for visual inspection of classification decisions.

5 Conclusion

In this paper, we proposed to exploit the discriminative nature of CNNs to tackle the challenging task of multi-class novelty detection. Our approach is inspired by the sensitivity of internal nodes of neural networks to class-specific patterns when trained in a supervised manner. We empirically found that simple statistics regarding which nodes are most heavily activated allow for discriminating between known and unknown classes. Since these *extreme value signatures* are intuitive and easy to implement on top of existing models, they allow to "upgrade" arbitrary classification networks to jointly estimate novelty and class membership. An analysis on different subsets of the ILSVRC'12 data shows performance benefits in terms of accuracy, computation time, and scalability of our approach in comparison with established baselines. To gain further insights, we finally investigated class-indicative image parts which can be obtained by visualizing extreme value signatures. Besides the positive aspects, however, our results also underline clearly that multi-class novelty detection is far from being solved when more than a handful of classes are involved.

[4] Gradient maps are normalized individually for better visualization, hence, the scaling can not be compared directly. Results of uniformly normalized maps are similar.

References

1. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural codes for image retrieval. In: European Conference on Computer Vision (ECCV) (2014)
2. Bendale, A., Boult, T.: Towards open world recognition. In: Computer Vision and Pattern Recognition (CVPR) (2015)
3. Bendale, A., Boult, T.E.: Towards open set deep networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
4. Bodesheim, P., Freytag, A., Rodner, E., Denzler, J.: Local novelty detection in multi-class recognition problems. In: Winter Conference on Applications of Computer Vision (WACV) (2015)
5. Bodesheim, P., Freytag, A., Rodner, E., Kemmler, M., Denzler, J.: Kernel null space methods for novelty detection. In: Computer Vision and Pattern Recognition (CVPR) (2013)
6. Brandenburg, F.J., Gleißner, A., Hofmeier, A.: The nearest neighbor spearman footrule distance for bucket, interval, and partial orders. Journal of Combinatorial Optimization (2013)
7. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: International Conference on Machine Learning (ICML) (2014)
8. Dosovitskiy, A., Brox, T.: Inverting convolutional networks with convolutional networks. CoRR abs/1506.02753 (2015)
9. Fawcett, T.: An introduction to roc analysis. Pattern Recognition Letters (2006)
10. Friedman, J., Hastie, T., Tibshirani, R.: The elements of statistical learning. Springer (2001)
11. Huang, Y., Sun, X., Lu, M., Xu, M.: Channel-max, channel-drop and stochastic max-pooling. In: Computer Vision and Pattern Recognition Workshop (CVPR-WS) (2015)
12. Jumutc, V., Suykens, J.A.: Supervised novelty detection. In: Computational Intelligence and Data Mining (CIDM) (2013)
13. Jumutc, V., Suykens, J.A.: Multi-class supervised novelty detection. Pattern Analysis and Machine Intelligence (PAMI) (2014)
14. Kemmler, M., Rodner, E., Denzler, J.: One-class classification with gaussian processes. In: Asian Conference on Computer Vision (ACCV) (2010)
15. Kenk, V.S., Kovačič, S., Kristan, M., Hajdinjak, M., Perš, J., et al.: Visual re-identification across large, distributed camera networks. Image and Vision Computing (2015)
16. Li, Y., Liu, L., Shen, C., Hengel, A.v.d.: Mining mid-level visual patterns with deep cnn activations. International Journal of Computer Vision 121(3), 344–364 (2017), `http://dx.doi.org/10.1007/s11263-016-0945-y`
17. Makhzani, A., Frey, B.J.: Winner-take-all autoencoders. In: Neural Information Processing Systems (NIPS) (2015)
18. Mensink, T., Verbeek, J., Perronnin, F., Csurka, G.: Distance-based image classification: Generalizing to new classes at near-zero cost. Pattern Analysis and Machine Intelligence (PAMI) (2013)
19. Pimentel, M.A., Clifton, D.A., Clifton, L., Tarassenko, L.: A review of novelty detection. Signal Processing (2014)
20. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision (IJCV) (2015)
21. Scheirer, W.J., de Rezende Rocha, A., Sapkota, A., Boult, T.E.: Toward open set recognition. Pattern Analysis and Machine Intelligence (PAMI) (2013)
22. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural computation (2001)

23. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPR-WS) (2014)
24. Simon, M., Rodner, E., Denzler, J.: Part detector discovery in deep convolutional neural networks. In: Asian Conference on Computer Vision (ACCV). pp. 162–177 (2014)
25. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: International Conference on Learning Representations Workshop (ICLR-WS) (2014)
26. Tax, D.M., Duin, R.P.: Support vector data description. Machine learning (2004)
27. Vinokurov, N., Weinshall, D.: Novelty detection in multiclass scenarios with incomplete set of class labels. arXiv preprint arXiv:1604.06242 (2016)
28. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European Conference on Computer Vision (ECCV). pp. 818–833 (2014)
29. Zhou, B., Khosla, A., Lapedriza, A., Torralba, A., Oliva, A.: Places: An image database for deep scene understanding. arXiv preprint arXiv:1610.02055 (2016)
30. 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)