

# Analyzing Classifiers: Fisher Vectors and Deep Neural Networks

Sebastian Lapuschkin<sup>1</sup>, Alexander Binder<sup>2</sup>, Grégoire Montavon<sup>3</sup>,  
Klaus-Robert Müller<sup>3,4</sup>, and Wojciech Samek<sup>1</sup>

<sup>1</sup>Fraunhofer HHI, <sup>2</sup>Singapore University of Technology and Design, <sup>3</sup>TU Berlin, <sup>4</sup>Korea University  
{sebastian.lapuschkin|wojciech.samek}@hhi.fraunhofer.de, alexander.binder@sutd.edu.sg  
{klaus-robert.mueller|gregoire.montavon}@tu-berlin.de

## Abstract

*Fisher vector (FV) classifiers and Deep Neural Networks (DNNs) are popular and successful algorithms for solving image classification problems. However, both are generally considered ‘black box’ predictors as the non-linear transformations involved have so far prevented transparent and interpretable reasoning. Recently, a principled technique, Layer-wise Relevance Propagation (LRP), has been developed in order to better comprehend the inherent structured reasoning of complex nonlinear classification models such as Bag of Feature models or DNNs. In this paper we (1) extend the LRP framework also for Fisher vector classifiers and then use it as analysis tool to (2) quantify the importance of context for classification, (3) qualitatively compare DNNs against FV classifiers in terms of important image regions and (4) detect potential flaws and biases in data. All experiments are performed on the PASCAL VOC 2007 and ILSVRC 2012 data sets.*

## 1. Introduction

Deep Neural Networks (DNN) have defined state of the art in many fields, such as image classification [14], image detection [8] and machine translation [26]. While much of research is devoted to extending the applicability of DNNs to more domains [9, 11, 12, 13, 31], we focus here on a different question, namely the impact of context, and the ability to use context. This question was raised already during times of the Pascal VOC challenge, where the amount of context was a matter of speculation, *c.f.* PASCAL VOC workshop presentation slides in [7].

The question of context is considered for two prominent types of classifiers. The first type, Fisher vectors (FV) [24] are based on computing a single feature map on an image as a whole and subsequently computing one score. In such a setup one can expect that context plays naturally a role for the prediction as the image is processed as a whole during training and testing. In case of small training sample sizes and the absence of opportunities for fine-tuning, Fisher vec-

tors still might be a viable alternative to DNNs due to their reduced parameter space. Examples for performance issues of Deep Neural Networks on small sample sizes without finetuning can be seen in [30]. The question of context is also open for the second type, Deep Neural Networks. One might assume that context plays no role for neural networks when they are used in classification by detection setups. For example, a recent ImageNet challenge winner relied on 144 crops per test image and classifier [27]. Another work using Pascal VOC data [20] used at test time 500 multi-scale patches per test image. However in certain setups computing several hundred windows as required for classification by detection setups may not be possible, *e.g.* when using hardware without GPUs and much main memory, such as used consumer laptops or smartphones, and when having time constraints for computation of the test prediction on an image. One can expect to see a larger impact of context when resorting to a few regions of an image at test time only, and thus training and testing with larger image patches.

Our contribution here is as follows. (1) We extend the method of [1] to Fisher vectors, and apply relevance propagation for the first time to Fisher vectors. (2) We define measures for the amount of context used for prediction in a single test image. (3) We apply the measures of context for neural networks and Fisher vector based classifiers on the Pascal VOC dataset, as it offers a way to approximately validate context by its bounding box annotation. We compare the context dependence of Fisher vectors against neural nets which were trained on larger patches of input images. (4) We show that this methodology is able to identify strong cases of context and biases in the training data even without using bounding box information.

The next section reviews related work. Section 3 briefly describes the Fisher vector classifier. Section 4 introduces the extended LRP method to decompose a Fisher vector prediction into scores for small regions of the order of a local feature. The same section also proposes a novel LRP-based measure of the importance of context. Section 5 introduces the experimental setup and presents results. The paper concludes in Section 6 with a summary and an outlook.

## 2. Related work

In recent years, interest in understanding image representations [15, 17, 19] and being able to explain the decision process of a classification system has increased, with *e.g.*, gradient-based sensitivity analysis [2, 25]. However, many approaches have been conceived with a specific pipeline architecture in mind. So do [28] explain predictions for Bag of Word (BoW) features with hard mapping (Vector Quantization) and Histogram Intersection kernels, and [16] identifies image regions critical for the prediction of a linear SVM classifier with max-pooling feature aggregation algorithm. A solution especially dedicated to visualize image regions triggering the prediction of deep convolutional neural networks with max-pooling layers and has been proposed in [30] with *deconvolution nets*.

Recently, a paradigm called Layer-wise Relevance Propagation (LRP) has been introduced in [1] as a way to compute partial prediction contributions – or *relevance values*  $R$  – for intermediate and input representations based on the final classifier output. It computes scores for regions or pixels of an image explaining the prediction itself rather than the effect of single neurons or particular layers. See [18] for a more in depth explanation of this method. It is applied in [1] to Bag of Visual Words classifiers and Deep Neural Networks; in this paper we extend this method to make it applicable to Fisher vector classifiers.

## 3. Fisher vectors in a nutshell

Fisher Vectors [21, 24] are a powerful tool to compute rich image or video representations and provide state-of-the-art performance amongst feature extraction algorithms. Figure 1 summarizes the steps involved in computing FV representation of an image. We introduce here a notation which later will be used in the Section 4.

An integral part for computing FVs is to fit a Gaussian Mixture Model (GMM) on top of the local descriptors  $L = \{l\}$  extracted from the training data to serve as a soft vocabulary of visual prototypes. Assuming a  $K$ -component GMM  $\lambda = \{(\pi_k, \mu_k, \Sigma_k)\}_{k=1..K}$ , then  $\pi_k$  is the mixture weight of component  $k$ , with  $\sum_k \pi_k = 1$  and  $\forall k : \pi_k \geq 0$ ,  $\mu_k$  is the mean vector of the  $k$ th mixture component and  $\Sigma_k$  its (diagonal) covariance matrix. For the computation of a *full* FV representation of an image, each local descriptor  $l$  is related to all  $K$  components of the trained GMM in its 0th (soft mapping weight), 1st (deviation from mean) and 2nd moment (variance) [24], i.e.

$$\Psi_{\pi_k}(l) = \frac{1}{\sqrt{\pi_k}} (\gamma_k(l) - \pi_k) \quad (1)$$

$$\Psi_{\mu_k}(l) = \frac{1}{\sqrt{\pi_k}} \gamma_k(l) \left( \frac{l - \mu_k}{\sigma_k} \right) \quad (2)$$

$$\Psi_{\sigma_k}(l) = \frac{1}{\sqrt{\pi_k}} \gamma_k(l) \frac{1}{\sqrt{2}} \left( \frac{(l - \mu_k)^2}{\sigma_k^2} - 1 \right) \quad (3)$$

with  $\Psi_{\pi_k}(l) \in \mathbb{R}$ , both  $\Psi_{\mu_k}(l)$  and  $\Psi_{\sigma_k}(l) \in \mathbb{R}^D$  and  $\gamma_k(l)$  returning the soft assignment of  $l$  to the  $k$ th mixture component. The FV embedding  $\Psi_\lambda(l)$  for a single descriptor  $l$  is then achieved by concatenating the mapping outputs relative to all  $K$  components into a  $(1 + 2D)K$  dimensional vector

$$\Psi_\lambda(l) = [\Psi_{\pi_1}(l) \dots \Psi_{\mu_1}(l) \dots \Psi_{\sigma_1}(l) \dots] \quad (4)$$

Having computed all those (as we will refer to now as) *raw* Fisher embeddings for all individual local descriptors, a single image-wise descriptor is achieved by averaging over the complete set of  $\Psi_\lambda(l)$ , followed by power normalization to reduce the sparsity of the descriptor and  $\ell_2$ -normalization to improve prediction performance [21]. The application of both final normalization steps results in a so called *improved Fisher Kernel* and is – in combination with a linear SVM [5] – equivalent to the transformation of the raw FV using the Hellinger’s kernel function [21].

## 4. Explaining classification decisions

Most predictors, including linear SVMs over Fisher vectors, incorporate several layers of non-linear mappings, resulting in a non-linear black box with respect to the dependency of the prediction on its pixel inputs. In this section we introduce the concept of Layer-wise Relevance Propagation (LRP) [1] as a way to compute partial prediction contributions – or *relevance values*  $R$  – for intermediate and input representations based on the final classifier output. LRP acts on a single test-image similar to the work in [30] and to partial-derivative based methods such as [25]. We refer the reader to [23] for a comparison of these three explanation approaches.

### 4.1. Layer-wise Relevance Propagation

Layer-wise Relevance Propagation decomposes the mappings performed during prediction time to attribute to each component of the input its *share* with which it contributes to the classifier output, explaining its relevance to the prediction output in its given state. This unsupervised process of decomposition is in principle applicable to any kind of model, resulting in high (positive) output values  $R$  identifying properties of the input speaking for the presence of the prediction target and low (or even negative) scores indicating no or negative contribution. The conservation principle inherent to LRP ensures that no amount of relevance is gained or lost in between layers of computation,

$$\sum_i R_i^{(k)} = \sum_j R_j^{(k+1)} \quad (5)$$

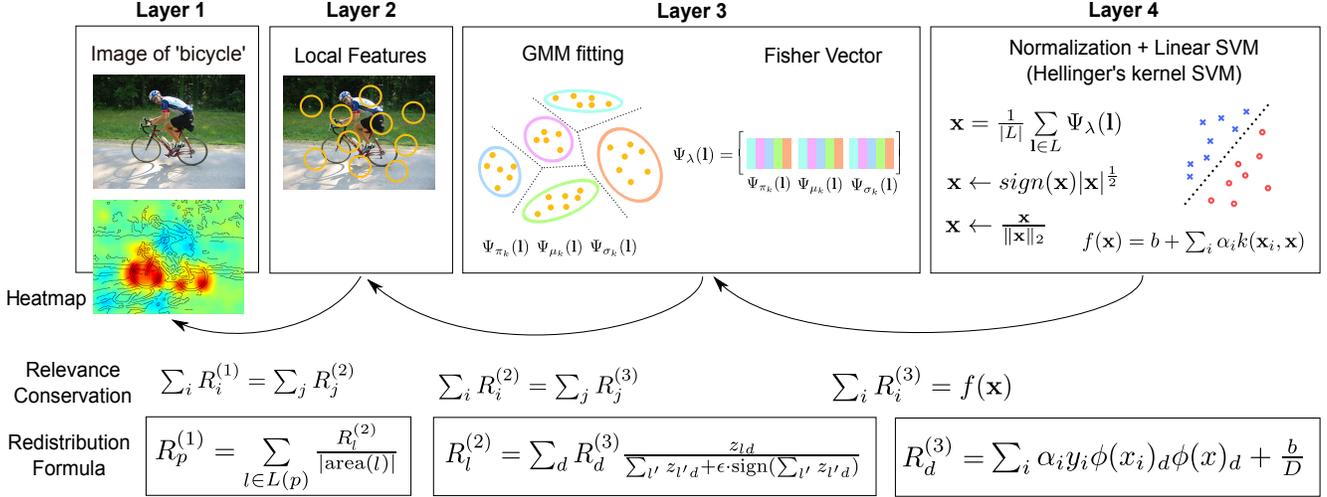


Figure 1. Computing Fisher Vector representation of an image and explaining the classification decision.

where  $R_i^{(k)}$  signifies the relevance value attributed to the  $i$ th computation unit or dimension at the  $k$ th computation layer of the prediction pipeline, and where the sums run over all units of the corresponding layers. In the context of an image classification problem, iterating LRP from the classifier output to the input layer results in outputs  $R_p^{(1)}$  for each pixel  $p$ , with

$$f(x) = \sum_p R_p^{(1)} \quad (6)$$

and  $f(x)$  being equal the output layer relevance values. In [1] examples have been given for decompositions of neural network architectures and Bag of Words feature extraction pipelines satisfying the above constraints.

LRP propagates the relevance  $R$  back from the output of a mapping towards its inputs. In a neural networks, a neuron maps a set of inputs  $\{x_i\}$  to an output  $x_j$  with monotonously increasing activation function  $g(\cdot)$

$$x_j = g(z_j), \quad z_j = \sum_i z_{ij}, \quad z_{ij} = w_{ij} x_i \quad (7)$$

where the sum runs over all input neurons contributing to the activation of neuron  $x_j$ . The goal is to compute a relevance  $R_i$  for input  $x_i$  when relevances  $R_j$  for outputs  $x_j$  are given. [1] has introduced two possible formulas for relevance propagation

$$R_i = \sum_{j:i \rightarrow j} \frac{z_{ij}}{z_j + \epsilon \cdot \text{sign}(z_j)} R_j \quad (8)$$

$$R_i = \sum_{j:i \rightarrow j} \left( \alpha \frac{z_{ij}^+}{z_j^+} - \beta \frac{z_{ij}^-}{z_j^-} \right) R_j, \quad (9)$$

where  $\sum_{j:i \rightarrow j}$  denotes a sum of all mappings which take  $x_i$  as input.  $z_{ij}^+$  denotes the positive part of the term,

i.e.  $\max(0, z_{ij})$ ,  $z_j^+$  is the sum over these positive parts.  $z_{ij}^-$  is defined analogously as the negative part and  $\alpha - \beta = 1$ . The same paper has introduced a method to compute relevances for Bag of Words vectors, however, it tacitly assumed that BoW mappings are dominantly non-negative. For Fisher vectors this assumption does not hold, as the features are derivatives with respect to parameters. For this reason we propose a modified approach.

## 4.2. LRP for Fisher vector classifiers

Our variant to use LRP for Fisher vectors starts with writing the linear SVM as a mapping of features

$$f(x) = b + \sum_i \alpha_i y_i \sum_{d=1}^D \phi(x_i)_d \phi(x)_d, \quad (10)$$

where  $x$  is a raw Fisher vector, and  $\phi(x)$  realizes its normalization. In consistency with the first LRP formula, we define  $R^{(3)}(x)$  as

$$R_d^{(3)} = \sum_i \alpha_i y_i \phi(x_i)_d \phi(x)_d + \frac{b}{D} \quad (11)$$

From here on we apply for the mapping of local features  $l$  to Fisher vectors  $x$ , equation (8) instead of the approach used in [1]. We can write the  $d$ -th dimension of the Fisher vector  $x_d = \sum_l m_d(l)$ . This is a mapping of local features  $l$  onto the Fisher vector as a set of outputs  $(x_d)_{d=1}^D$ . We apply equation (8) with  $z_{ld} = m_d(l)$ .  $m_d(l)$  is given in the notation of Section 3 as the term from equation 4:

$$m_{(d)}(l) = \Psi_{\lambda}(l)_{(d)} \quad (12)$$

Pixel-wise relevance scores  $R_p^{(1)}$  are then computed by uniformly distributing for all local features  $l$  the relevance

scores  $R_l^{(2)}$  onto the set of pixels  $p$  covered by the receptive field of  $l$ , resulting in a *heatmap* which can be visualized. The decomposition process with explicit redistribution formulas is depicted in Figure 1.

### 4.3. Measuring context with LRP

The distribution of positive relevance mass in a heatmap can be used for assessing the importance of context for a particular image classification task. If bounding box annotation are available (as for the Pascal VOC dataset), we can compute the *outside-inside relevance ratio* metric defined as:

$$\mu = \frac{\frac{1}{|P_{\text{out}}|} \sum_{q \in P_{\text{out}}} R_q^{(1)}}{\frac{1}{|P_{\text{in}}|} \sum_{p \in P_{\text{in}}} R_p^{(1)}} \quad (13)$$

with  $|\cdot|$  being the cardinality operator and  $P_{\text{out}}$  and  $P_{\text{in}}$  being the set of pixels outside and inside the bounding box, respectively. A high relevance ratio indicates that the classifier uses a lot of context to support the decision. A low relevance ratio indicates that the classifier focuses instead on the object to support its decision. Note that this measure can not be 100% accurate in most cases, since for example the bounding box areas of slim but obliquely angled objects, for example, aeroplanes photographed during lift-off, will also cover a considerable amount of image background.

## 5. Experimental evaluation

### 5.1. Basic setup

All measurements are carried out on PASCAL VOC 2007 [6] test data. Fisher vectors are computed using the encoding evaluation toolkit (version 1.1) from [3] with settings as in this paper. The Fisher vectors are trained on the trainval part of the same dataset. The neural network is fine-tuned on the trainval part of PASCAL VOC 2012, starting from the BVLC reference classifier of the Caffe package [10] with a base learning rate of 0.001 using a multi-label hinge loss. As we are interested in the ability of a neural net to use context, we do not use the bounding box ground truth to extract image patches which cover parts of bounding boxes. Instead we create 4 corner and one center crop per image together with mirroring, resulting in 10 training patches per image. Test scoring is done in the same fashion. This corresponds to a setting with only a few number of test windows, in which one would use larger patches during training and testing. The region-wise scores are computed for FV as described in Section 4 using equation (8) with parameter  $\epsilon = 1$  and  $\epsilon = 100$ . For neural nets we used equation (8) with  $\epsilon = 1$ ,  $\epsilon = 100$  and equation (9) with  $\beta = 1$ ,  $\alpha = 2$ . Random perturbations for Fisher vectors were achieved by randomly sampling local features from the GMM.

	<b>aer</b>	<b>bic</b>	<b>bir</b>	<b>boa</b>	<b>bot</b>
<b>F</b>	79.08	66.44	45.90	70.88	27.64
<b>D</b>	88.08	79.69	80.77	77.20	35.48
	<b>bus</b>	<b>car</b>	<b>cat</b>	<b>cha</b>	<b>cow</b>
<b>F</b>	69.67	80.96	59.92	51.92	47.60
<b>D</b>	72.71	86.30	81.10	51.04	61.10
	<b>din</b>	<b>dog</b>	<b>hor</b>	<b>mot</b>	<b>per</b>
<b>F</b>	58.06	42.28	80.45	69.34	85.10
<b>D</b>	64.62	76.17	81.60	79.33	92.43
	<b>pot</b>	<b>she</b>	<b>sof</b>	<b>tra</b>	<b>tvm</b>
<b>F</b>	28.62	49.58	49.31	82.71	54.33
<b>D</b>	49.99	74.04	49.48	87.07	67.08

Table 1. Prediction performance of the trained Fisher model and DNN in average precision (AP) per class in percent. The mAP scores for the FV model and DNN are 55.99 and 72.12 respectively.

### 5.2. Are Fisher explanations meaningful?

The first step before measuring the amount of context is to validate whether the computed scores for a pixel or a region are meaningful at all. Figure 2 depicts heatmaps computed on exemplary test images of the Pascal VOC data set considering the prediction score for a particular class. The quality of these explanations can be intuitively assessed by a human, *e.g.*, it makes perfectly sense that the Fisher vector classifier finds that wheels are relevant for the class “bike”, rail tracks are indicative for the class “train” and tableware is important for classifying images of class “dining table”. These examples show that the largest part of the relevance mass does not necessarily need to lie on the object, on the contrary it may be the context which is the informative part.

In order to objectively validate that the Fisher vector heatmaps are meaningful we evaluate the decrease of the prediction score under perturbations. The idea is that a region such as an image patch is highly relevant, if modifying it results for most modifications in a sharp decline of the prediction for the whole image. Modifying a region is done by randomly perturbing the pixels with noise. The prediction score is averaged over a number of random perturbations, in order to capture the average change of the classifier.

This notion of relevant regions can be used for evaluation of region scores by sorting image regions along descending scores. Then, for each region in the sequence the average decrease of predictions is measured. The result is a graph as a function of the sequence index. Thus under this evaluation scheme, a region-wise score performs well if it assigns highest scores to regions which are most sensitive on average under perturbations and yield the sharpest decline of the prediction score. [23] introduced this setup and evaluated the methods of [1, 25, 30] for Deep Neural Networks tested

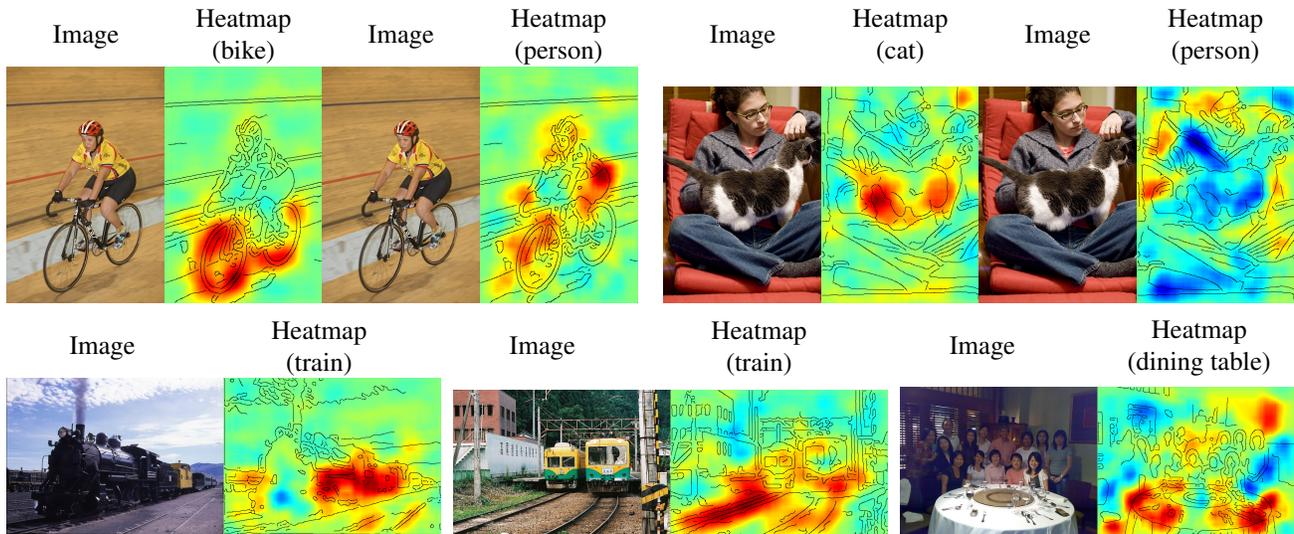


Figure 2. Images shown next to the heatmaps computed by application of LRP on the FV model when considering the prediction score for a particular class.

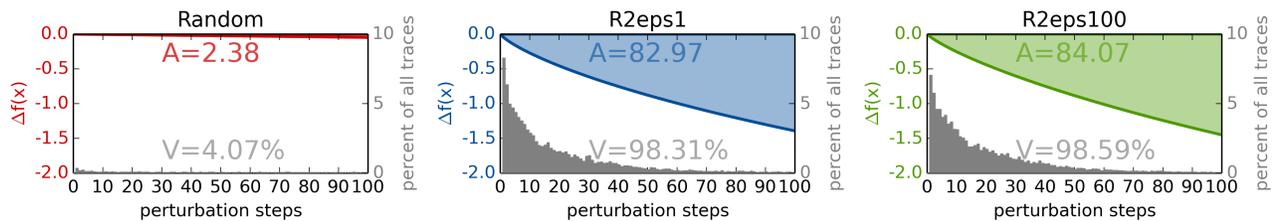


Figure 3. Heatmap quality measurements for Fisher vectors. The value  $A$  measures the area above the curve between the original prediction  $f(x)$  and the averaged perturbed prediction at step  $i$  in the sequence of regions.  $V$  represents the fraction of all perturbation sequences for which the prediction switched sign at some step in the sequence, with the gray bar chart showing how many sample traces changed class at each point of measurement.

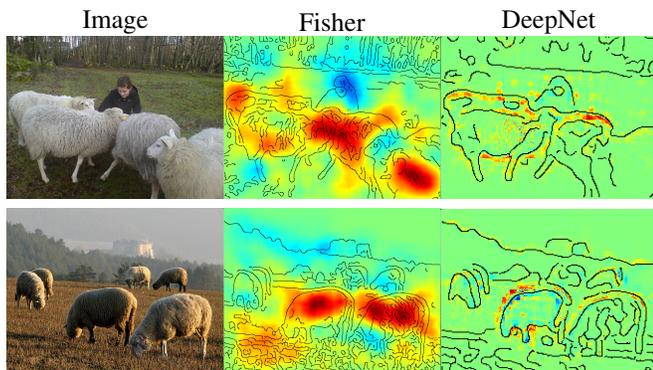


Figure 4. Images of the class “sheep”, processed by the FV and DNN models and heatmapped using LRP.

on ImageNet [22], SUN397 [29] and MIT Places [32]. Here we show that LRP scores computed are also meaningful for Fisher vectors. Figure 3 shows this comparison against random orderings for scores computed. The LRP scores produce a more meaningful ordering than random sequences which motivates its use to define a measure for context.

### 5.3. Shallow vs. deep features

We investigate in the light of the LRP framework what are the differences of strategies used to classify images between (1) a shallow model operating on high-resolution images: the FV model, and (2) a deep model operating on lower-resolution images: the DNN model. We consider first the class “sheep” for which the DNN produces much better predictions than the FV model (25% superior accuracy in absolute terms according to Table 1).

Example of two images of class “sheep” and the corresponding heatmaps for the FV and DNN models are shown in Figure 4. The LRP analysis reveals that the FV and DNN models use clearly different strategies to predict the class:

The FV model bases its decision on the wool texture typical of the sheep and available at high-resolution, but ignores the exact shape of the sheep. Interestingly, relevance is also allocated to the context (here, positive relevance for the grass and negative relevance for the human face), indicating that the context is an essential component of the classifier and modulates the prediction score positively or

negatively.

On the other hand, the DNN assigns a large proportion of heat to the border of the sheep, thus, showing that the shape of the sheep (*e.g.* its contour) is exploited in order to improve the prediction. Furthermore, for the DNN, the LRP method does not assign relevance to contextual elements such as the grass, or the human face, nor to the wool texture of the sheep, which is harder to detect due to the low resolution of images given to the DNN.

Overall, the LRP analysis indicates that the far superior predicting power of the DNN model must be attributed in largest part to the ability to model the exact shape of the sheep, making all remaining contextual or texture features less relevant. On the other hand, the less accurate FV model does benefit from the weak correlations between object class, texture and context to improve prediction quality.

#### 5.4. Test error and model quality

For other classes, it can be observed in Table 1 that test error of the FV model is almost on par with the one of the DNN. We investigate whether high test accuracy is predictive of the ability of the model to extract meaningful features for a given class, or whether the decision is based mostly on undesirable contextual or artefactual features.

**Contextual features.** As an illustrative example, we consider the class “boat”, where the performance of the DNN superior by less than 7% in absolute terms to the FV model. (Note that for other classes such as “sheep” or “bird”, the DNN performance is superior by 25% or more.) It is tempting to conclude that, for the class “boat”, both models should have learned a set of features of similarly high quality. LRP analysis gives a different answer: Figure 5 (left) shows the heatmaps produced by the FV and DNN models on two archetypical images of the class “boat”. For the DNN, LRP assigns most of the relevance to pixels corresponding to the actual boat. On the other hand, for the FV model, LRP assigns most relevance to the water below the boat (*i.e.* the FV model does not recognize the object itself, but its context). The heat distribution of average heatmaps (computed over all landscape-format images of the class “boat”) corroborates what was observed for two selected images, in particular, a focus of the FV model on the bottom part of the image where water usually is, and a focus of the DNN model on the middle part of the image where the boat typically is. We can conclude from the LRP analysis, that while both classifiers have a roughly similar level of accuracy on the test images with class “boat”, FV’s performance is likely to decrease drastically if one were to consider boats located outside the water as test images. On the other hand, performance of the DNN would be less affected. Therefore, test error is a superficial predictor of model quality in this case.

**Artefactual features.** A second example where high accuracy does not necessarily translate into high quality features is for the class “horse”. This class is predicted with similar accuracy by the FV and DNN models (approximately 1% difference in accuracy).

Figure 5 (right) shows LRP heatmaps for the FV and DNN model on an image of horse. While the DNN assigns relevance on the actually shown “horse”, the FV assigns almost all relevance in the bottom-left corner the image, where careful inspection of the image reveals the presence of a copyright tag. Thus, the decision of the FV model is in large part based on the presence of the copyright tag, which is discriminative of the class horse. Removing the copyright tag completely changes the FV heatmap, but does not change significantly the DNN heatmap.

If the copyright tag is removed, the DNN is still able to predict the image because the pixels that support its decision are not affected. On the other hand, FV model prediction quality will be considerably reduced. The systematic focus of the FV model on the copyright tag is confirmed in the average heatmap, where the bottom-left corner is assigned large amount of heat. Therefore, for this class again, test error does not predict well model quality.

#### 5.5. Quantitative analysis of context use

While we have so far provided a qualitative interpretation of FV heatmaps for examples and classes of interest, we can more systematically measure whether the model uses context or the actual object, by measuring for each classes and models the outside-inside relevance ratio  $\mu$  computed by equation 13. Results are shown in Figure 6. Generally, the FV model uses more context than the DNN, as evidenced by a higher relevance ratio. However, there are significant differences between classes: Classes where the use of context by the FV model is particularly high are “boat” and “airplane”, the first of which we have studied qualitatively in the previous section. For these two respective classes, the water and the sky are important contextual elements that support the decision of the Fisher model, due to their strong correlation. Another group of classes with high context of the Fisher model are “chair”, “diningtable”, “pottedplant” and “sofa” which share a semantic of indoor room sceneries.

For other classes such as “bicycle”, “car”, “motorbike”, or “sheep”, the Fisher model does not use much context. For the first three classes, the urban environment surrounding these classes is not predictive of the object being detected (*i.e.* it could not discriminate between these three classes based on the context only). For the last class, as it has been discussed in Section 5.3, the wool texture of the sheep (which lies inside the sheep bounding box) is a reasonable predictor for the class “sheep”, although the actual object sheep (*i.e.* defined by its shape or contour) is not

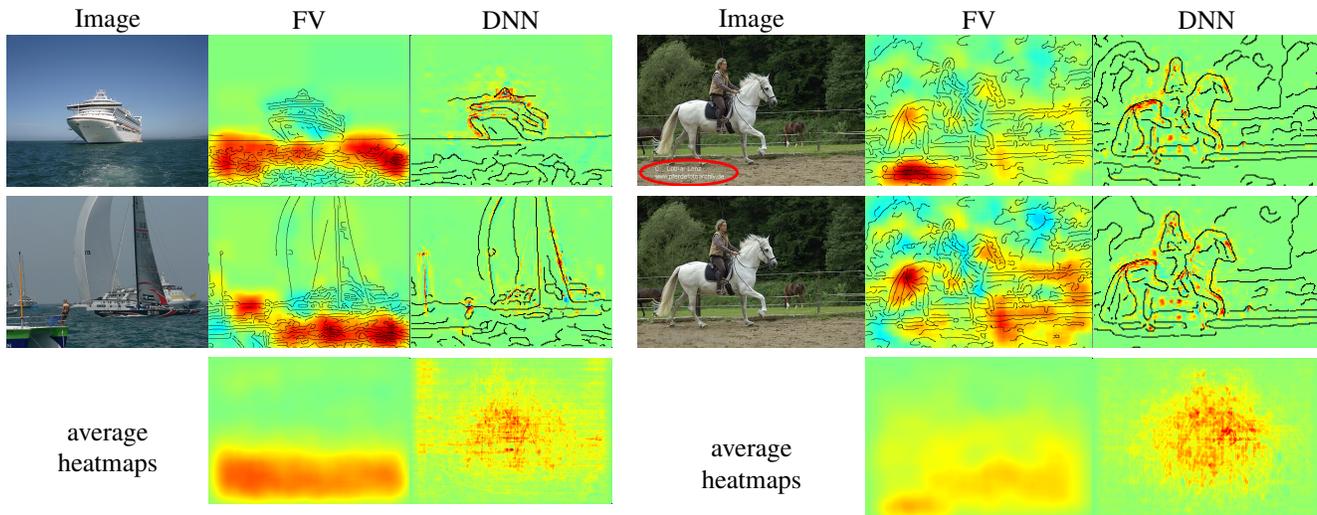


Figure 5. Top: Images of the classes “boat” and “horse”, processed by the FV and DNN models and heatmapped using LRP. Bottom: Average heatmap scores over a random sample (of size between 47 and 177) of the distribution for each class and model. On the second image of class “horse”, the copyright tag (marked by the red ellipse) has been removed.

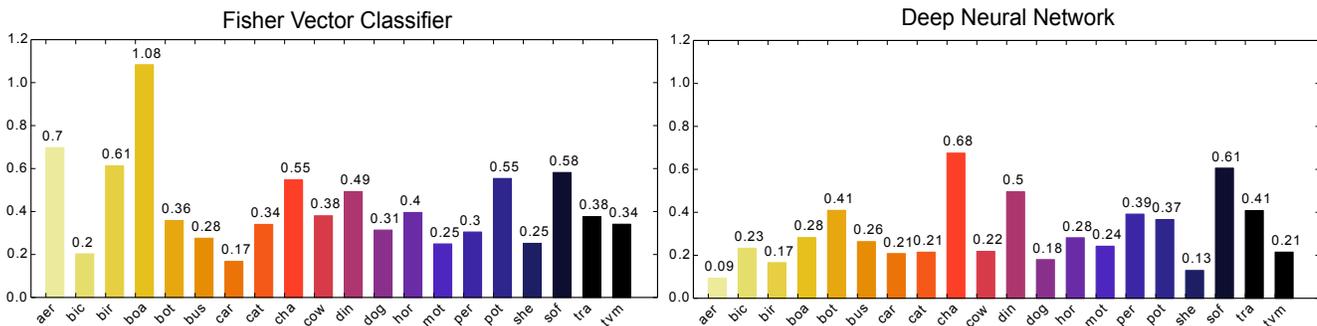


Figure 6. Outside-inside relevance ratio as computed by equation 13 for the 20 classes of the Pascal VOC 2007 dataset. Left: ratios for the FV model. Right: ratios for the DNN model.

being used.

As for deep neural networks, classes with least context usage are “aeroplane”, “bird”, “sheep”, “dog”, “car”, “cat” and “tvmonitor”. Each of those is associated with a significantly better score achieved by the DNN.

### 5.6. Shallow vs. deep networks

Our results on ILSVCR 2012 validation data show that, when performing LRP on the BVLC Caffe reference versus GoogleNet, the use of contextual information is much lower for the deeper and better performing GoogleNet. Model architectures have been used as-is. In addition to depth, the type of layers (*e.g.* inception, normalization) may have an impact on the sparsity and should be subject to further studies. Figures 7 and 8 present the results quantitatively and as exemplary heatmaps for the BVLC Reference [10], VGG CNN S [4], which has slightly lower error rate than the former, and GoogleNet [27]. The latter two use smaller kernels with less stride at lowest level compared to

the BVLC Reference. We noted on many examples, that GoogleNet is much sparser than the other two and tends to ignore irrelevant edges, for example its reaction to the gradient between the dark green trees and the sky in the motorscooter example picture is the weakest of all three nets.

## 6. Conclusion

In this paper, we have analyzed what make Fisher vector models (FV) and deep neural networks (DNN) decide for a particular class. To achieve this, we have employed a heatmapping technique that determines what pixels in the image are used by a classifier to support its decision. The technique called layer-wise relevance propagation and originally developed for neural networks [1] was extended to Fisher vector models, and validated using the method by [23]. Our novel comparative analysis of FV and DNN classifiers corroborates empirically previous intuition relating the architecture of the classifier to the features it is able to extract. In particular, our analysis shows that the FV model

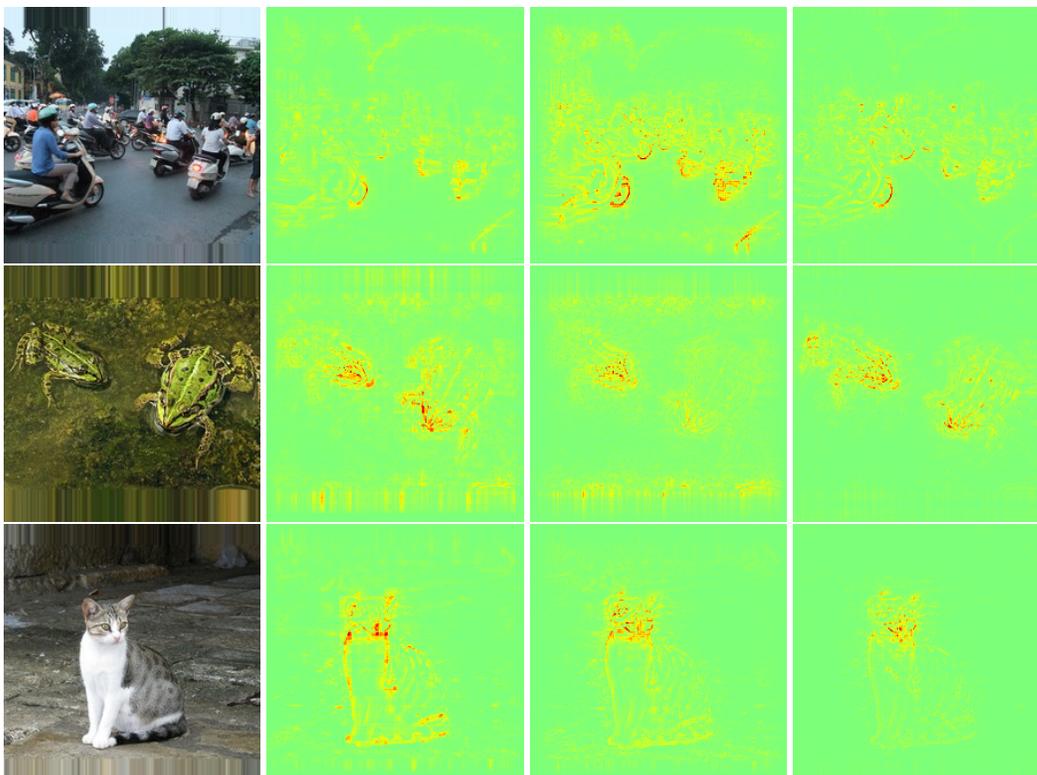


Figure 8. Comparison of different pretrained models on ImageNet for classes “scooter”, “frog” and “cat”. From left to right: Input, heatmaps for BVLC CaffeNet, VGG CNN S and GoogleNet. GoogleNet is particularly sparse which holds for many other examples. See the supplement for a larger version and more examples.

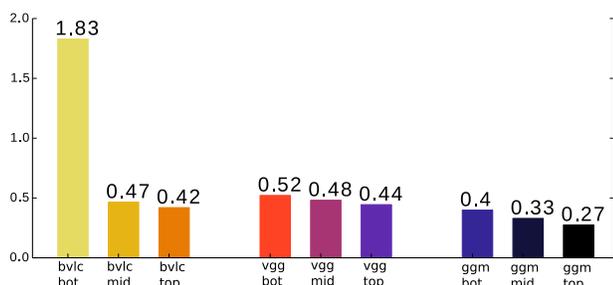


Figure 7. DNN context scores for ImageNet 2012 for the BVLC CaffeNet (bvlc, left three bars), VGG CNN S (vgg, middle three) and GoogleNet (ggm, right three). *Bot* are 333 classes with lowest prediction accuracy w.r.t. the used network and *top* are 333 classes with highest prediction accuracy. GoogleNet uses less context. As for an explanation for the higher values of context importance relative to the results on PASCAL VOC 2007, visual inspection revealed that many of the ImageNet bounding boxes cover much less of the object than those used in Pascal VOC.

compensates its lack of depth by the use of contextual information – potentially artefacts – that are weakly correlated to the object class. We thus demonstrate that the generalization capability of Fisher vector models can be overstated if test images also include similar context. On the other

hand, DNNs base their decision on the actual object to detect and ignores its context. This focus on object detection has to be attributed to the higher overall predictive accuracy of the model, that removes the need for contextual information – even if the latter is discriminative. The focus on detection must also be attributed to the deep multitask properties of the DNN that favors composition of natural image features over lower-level features such as copyright text. These results argue in favor of incorporating heatmap-mapping techniques into the data collection and model selection processes. The interpretable visual feedback that heatmaps provide can be used in particular to verify that the considered classifier bases its decision on the right set of features, and in the contrary case, select another model, or extend the dataset in a way that artefactual features can no longer support the classification decision.

## Acknowledgment

This work was supported by the German Ministry for Education and Research as Berlin Big Data Center BBDC (01IS14013A), the Deutsche Forschungsgesellschaft (MU 987/19-1) and the Brain Korea 21 Plus Program through the National Research Foundation of Korea funded by the Ministry of Education. Correspondence to KRM and WS.

## References

- [1] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):e0130140, 2015.
- [2] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller. How to explain individual classification decisions. *JMLR*, 11:1803–1831, 2010.
- [3] K. Chatfield, V. S. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, page 8, 2011.
- [4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *BMVC*, 2014.
- [5] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- [6] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015.
- [7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results.
- [8] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014.
- [9] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.
- [10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proc. of the ACM Int. Conf. on Multimedia*, pages 675–678, 2014.
- [11] A. Karpathy, A. Joulin, and F. Li. Deep fragment embeddings for bidirectional image sentence mapping. In *Adv. in NIPS*, pages 1889–1897, 2014.
- [12] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F. Li. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014.
- [13] J. Koutník, G. Cuccu, J. Schmidhuber, and F. J. Gomez. Evolving large-scale neural networks for vision-based reinforcement learning. In *GECCO*, pages 1061–1068, 2013.
- [14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Adv. in NIPS*, pages 1106–1114, 2012.
- [15] K. Lenc and A. Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *CVPR*, pages 991–999, 2015.
- [16] L. Liu and L. Wang. What has my classifier learned? visualizing the classification rules of bag-of-feature model by support region detection. In *CVPR*, pages 3586–3593. IEEE, 2012.
- [17] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *CVPR*, pages 5188–5196, 2015.
- [18] G. Montavon, S. Bach, A. Binder, W. Samek, and K.-R. Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *arXiv preprint arXiv:1512.02479*, 2015.
- [19] D. Novotný, D. Larlus, F. Perronnin, and A. Vedaldi. Understanding the fisher vector: a multimodal part model. *CoRR*, abs/1504.04763, 2015.
- [20] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, pages 1717–1724, 2014.
- [21] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Computer Vision—ECCV 2010*, pages 143–156. Springer, 2010.
- [22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, pages 1–42, April 2015.
- [23] W. Samek, A. Binder, G. Montavon, S. Bach, and K. Müller. Evaluating the visualization of what a deep neural network has learned. *CoRR*, abs/1509.06321, 2015.
- [24] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 105(3):222–245, 2013.
- [25] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014.
- [26] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Av. in NIPS*, pages 3104–3112, 2014.
- [27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.
- [28] J. Uijlings, A. Smeulders, and R. Scha. The visual extent of an object. *IJCV*, 2012.
- [29] J. Xiao, J. Hays, K. Ehinger, A. Oliva, A. Torralba, et al. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010.
- [30] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833, 2014.
- [31] N. Zhang, J. Donahue, R. B. Girshick, and T. Darrell. Part-based r-cnns for fine-grained category detection. In *Computer Vision - ECCV*, pages 834–849, 2014.
- [32] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Adv. in NIPS*, pages 487–495, 2014.