# DEEP CROSS-LAYER ACTIVATION FEATURES FOR VISUAL RECOGNITION

*Georgios Th. Papadopoulos, Member IEEE, Elpida Machairidou and Petros Daras, Senior Member IEEE*

Information Technologies Institute, Centre for Research and Technology Hellas, Greece

## ABSTRACT

Convolutional Neural Networks (CNNs), which have nowadays dominated image analysis tasks, constitute feed-forward methods that model increasingly complex data structures and patterns along the subsequent hidden layers of the network. However, the common practice of using the activation features from the last network layer inevitably leads to a visual recognition bottleneck. This is due to the fact that discriminative features for different objects of varying complexity do not need to be extracted from the same layer. To this end, a novel frequency domain analysis of the feature maps of the same as well as of different network layers is proposed. In this way, the proposed method exploits more efficiently the knowledge that is stored in the actual CNN and facilitates in identifying the most discriminative features for every individual object type. Experimental results in a large-scale real-world Closed-Circuit Television (CCTV) surveillance and the PASCAL VOC 2012 datasets demonstrate the efficiency of the proposed approach.

***Index Terms—*** Visual recognition, deep learning, convolutional neural networks, frequency domain analysis

## 1. INTRODUCTION

Recent advances in the GPU technology have given great boost and increased the capabilities of machine learning computational models and in particular Neural Networks (NNs) [1], which have largely been ignored over the past two decades. The so called 'Deep Learning (DL)' approach targets the construction of end-to-end systems that automatically learn the optimal features for the task at hand from the raw data; hence, outperforming and replacing the respective hand-crafted features.

For the case of image analysis, DL techniques have primarily been based on the use of Convolutional Neural Networks (CNNs) [2], which are suitable for processing multi-dimensional input arrays and detecting complex patterns at multiple spatial and semantic scales. The fundamental goal of CNNs, which consist of applying subsequent layers of convolutional, pooling and non-linearity operators, is to model and encode structures of increasing semantic complexity (as progressing through the subsequent network layers). For example, at the first layers simple low-level geometric features are modeled (e.g. different edge types), while at the last network layer complex objects/concepts are formed.

Among the image analysis tasks where CNNs have exhibited increased applicability and achieved superior performance are: i) image classification [3], ii) object detection [4], and iii) image segmentation [5]. In [3], a large deep convolutional neural network is trained to classify the images in the ImageNet LSVRC contest into the different supported classes. Simonyan and Zisserman [6] investigate the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. Additionally, He et al. [7] propose a 'spatial pyramid pooling' strategy for generating a fixed-length representation regardless of the image size/scale. In [8], a Region Proposal Network (RPN) is introduced that shares full-image convolutional features with the object detection network, thus enabling nearly cost-free region proposals generation. Moreover, Long et al. [5] build 'fully convolutional' networks that take input of arbitrary size and produce correspondingly-sized output with efficient inference and learning in the context of semantic image segmentation. Furthermore, a CNN-based algorithm is outlined in [9] for detecting all instances of a category in an image and, for each instance, marking the pixels that belong to it.

CNNs, which have prevailed image analysis tasks (as described above), constitute feed-forward methods that model increasingly complex data structures and patterns along the subsequent hidden layers of the network. The common practise, though, is to use the activation features from only the last network layer during the classification step. However, this choice inevitably leads to a visual recognition bottleneck, since discriminative features for different objects of varying complexity do not necessarily need to be extracted from the same layer. For overcoming this limitation, a novel frequency domain analysis is proposed in this paper for further improving the recognition capabilities of any CNN-based method. In particular, the feature maps of the same as well as of different network layers are simultaneously analyzed and a concrete representation is produced. This facilitates in identifying the most discriminative features at different semantic scales for

The work presented in this paper was supported by the European Commission under contract FP7-607480 LASIE and also by NVIDIA Corporation with the donation of a Tesla K40 GPU. The authors would like to thank the London Metropolitan Police (MET) for providing the CCTV video footage used for experimentation.

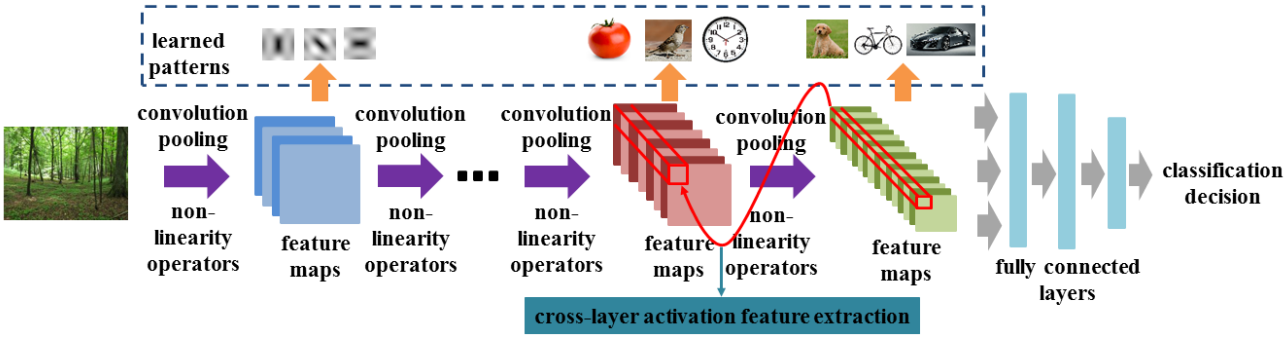



**Fig. 1**. Cross-layer activation feature extraction.

every individual object category, while also taking into account the correlations among the feature maps of the same or different layers. In this way, the proposed cross-layer activation features exploit more efficiently the knowledge that is stored in the actual CNN. Experimental results in a large-scale real-world Closed-Circuit Television (CCTV) surveillance and the PASCAL VOC 2012 datasets demonstrate the efficiency of the proposed approach.

The remainder of the paper is organized as follows: The proposed cross-layer activation feature extraction procedure is described in Section 2. Section 3 details the developed visual classifier. Experimental results are presented in Section 4 and conclusions are drawn in Section 5.

## 2. CROSS-LAYER ACTIVATION FEATURE EXTRACTION

CNNs constitute feed-forward models that receive as input raw data and their fundamental functionality consists of the application of subsequent hidden layers that automatically learn increasingly complex, in terms of visual appearance and semantic granularity, patterns and structures. The convolutional filters at the very first layers encode simple geometrical properties of the objects or simple shape patterns, such as different types of edges. Following layers model more complex mid-level representations (e.g. abstract representations of object classes), different parts of objects or even object types with relatively simple visual appearance. Eventually, at the last layers, complex object types (e.g. car) are formulated, by transforming and combining the patterns learned in previous layers. It must be noted that typically the last network layers, which are primarily responsible for generating the network's classification decision, are fully-connected and receive as input the features of the last convolutional layer, although in many CNN-based systems a different classifier (e.g. linear SVM) is also used for this task.

Several research works have verified the above learning behavior in CNNs for different image analysis tasks. Interestingly, the work of [10] shows that the features of not the last hidden layer, but the layer before, of the employed CNN accomplish the highest recognition performance. Aiming to shed light on the actual CNN behavior, Zeiler et al. [11] present a way to map the layer activities back to the input image pixel space, showing what input pattern originally caused a given activation in the network feature maps, by introducing a so-called 'Deconvolutional Network'. In a more elaborate work of the same authors [12], a visualization technique that gives insight into the function of intermediate feature layers and the operation of the classifier in a CNN is outlined.

In Fig. 1, a schematic representation of the conceptualization described above is given. In particular, a CNN initially receives as input a color image. Then, feature maps of varying spatial resolution and number of features are estimated at every subsequent layer; typically, along the subsequent network layers the spatial dimensions of the respective feature maps are reduced through the application of pooling operators for invariance incorporation, while on the contrary the number of feature maps increases for forming more abstract, complex and detailed patterns. Indicative examples of the types of patterns that are encoded at every layer are also provided, while a more thorough analysis of how and what patterns are formed at different network layers can be found in [12]. For simplicity, a straight-forward architecture with a single information stream is presented.

From the above analysis and schematic representation, it can be seen that the typical CNN learning paradigm exhibits the following main limitations: i) providing as input to the classification step the features from the last convolutional layer is not the best choice for all object types, which inevitably belong to different levels of semantic granularity, and ii) in the classification step (e.g. fully connected layers), although different feature maps are combined, the possible correlations among them are not taken into account.

For efficiently overcoming the above limitations, a frequency domain analysis is proposed in this work for extracting cross-layer activation features. In particular, the proposed features exhibit the following advantageous characteristics: i) information from multiple layers is directly incorporated; hence, enabling the selection of the appropriate learned patterns from different network layers with respect to every individual object/concept type, and ii) the correlations among the features of the same, as well as of different network layers, are modeled. More specifically, let the feature map of the $i$-th

network layer be denoted $M_i(x_i, y_i, f_i)$, where $x_i \in [1, X_i]$, $y_i \in [1, Y_i]$ and $f_i \in [1, F_i]$. $X_i$ and $Y_i$ represent the feature map's horizontal and vertical spatial dimensions, respectively, while $F_i$ denotes the number of learned features at the $i$-th layer. Since $M_i(x_i, y_i, f_i)$ from different layers generally have different spatial resolution, the first step in the proposed feature extraction procedure is to transform the feature maps into arrays of the same spatial dimensions. This is performed by means of simple linear interpolation in the $XY$ space, where all $M_i(x_i, y_i, f_i)$ target spatial dimensions are set equal to $X_i^{'} = \min_i(X_i)/2$ and $Y_i^{'} = \min_i(Y_i)/2$ in the current implementation; the interpolated feature maps are denoted $M_i^{'}(x_i^{'}, y_i^{'}, f_i)$. It must be noted that the number of features $F_i$ at every layer remains unchanged. Then, the interpolated feature maps $M_i^{'}(x_i^{'}, y_i^{'}, f_i)$ of the last $N$ network layers are stacked, along the $F$ dimension, in a single composite feature map, as also shown in Fig. 1. This results in the formation of a composite feature map $CM(x_c, y_c, f_c)$, where $x_c \in [1, X_i^{'}]$, $y_c \in [1, Y_i^{'}]$, $f_c \in [1, F_c]$ and $F_c = \sum_{i=1}^{N} F_i$. Subsequently, for every spatial location $(x_c, y_c)$ an 1D vector is formed by considering all elements of $CM(x_c, y_c, f_c)$ along the $F$ dimension (Fig. 1); this vector is denoted $\bar{v}_{x_c,y_c}(f_c)$. The latter vector undergoes a frequency domain analysis for estimating and efficiently modeling the correlations among its elements. For that purpose, the Discrete Cosine Transform (DCT) is used, according to the following equation: $r_{x_c,y_c}(\beta) = \sum_{f_c=1}^{F_c} \bar{v}_{x_c,y_c}(f_c) \cos\frac{\pi}{F_c}[(f_c - 1) + \frac{1}{2}(\beta - 1)]$, where $r_{x_c,y_c}(\beta)$ are the estimated DCT coefficients and $\beta \in [1, F_c]$. The reason for using the DCT transform is twofold: i) its simple form requires relatively reduced calculations, and ii) it is a frequency domain transform that receives as input a real sequence and its output is also a real set of values. Other common frequency analysis methods (e.g. Fourier transform) were also evaluated; however, they did not lead to increased performance compared with the one received when using DCT. Eventually, concatenating all $r_{x_c,y_c}(\beta)$ coefficients computed for all spatial locations $(x_c, y_c)$ results in the formation of vector $G$, which comprises the proposed cross-layer activation features.

The following important observations need to be made regarding vector $G$: i) The proposed cross-layer activation features encode the correlations between different features of the same as well as different network layers; hence, leading to the generation of a more comprehensive representation of the information that is already present in a CNN, and ii) it must be highlighted that the above feature extraction procedure is generic, i.e. it can be applied to any CNN for any type of analysis task.

## 3. VISUAL CLASSIFICATION

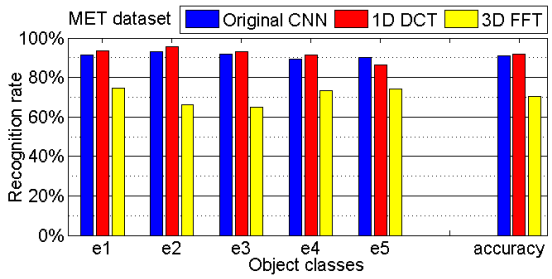
Although the proposed feature extraction step can be integrated as an additional layer in any existing CNN, similarly to the SPP-net [7], for simplicity purposes a separate CNN is used for classification in this work. In particular, a CNN consisting of three fully-connected layers is implemented for predicting the final classification decision (Fig. 1). Zero-mean Gaussian distribution with standard deviation equal to $0.01$ is used to initialize the neuron weights and biases. The first two layers comprise $4096$ neurons each, while the last layer consists of a number of neurons equal to the number of supported classes for the problem at hand. The class predictions are passed through a softmax operator (layer) to estimate a probability distribution over classes. Stochastic gradient descent is used during training, along with a multinomial logistic loss function. The batch size is set equal to $256$, while the momentum value is equal to $0.9$. Weight decay with value $0.0005$ is used for regularization. The base learning rate is initially set to $0.0001$ and it is subsequently reduced by a factor of $10$ every $10$ epochs. Overall, the training procedure lasts $30$ epochs.

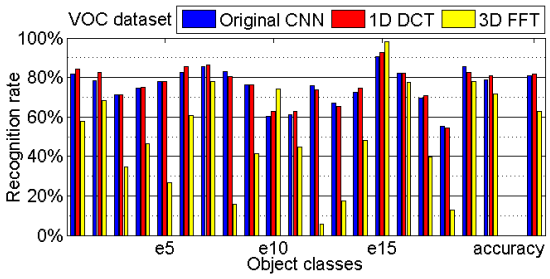
## 4. EXPERIMENTAL RESULTS

In this section, experimental results from the application of the proposed cross-layer activation feature extraction approach in an object recognition task are presented. For the evaluation a large-scale real-world Closed-Circuit Television (CCTV) surveillance dataset is used. This dataset is directly provided from the archives of the Metropolitan Police Service (MET), officially known as 'New Scotland Yard' to the general public, and consists of approximately $100,000$ hours of real surveillance footage from the $2011$ London riots. The critical challenges about this content, apart from its large-scale nature, are its low-quality and the presence of significant amounts of noise/artifacts/corruptions in the video frames. Some indicative frames are given in Fig. 3. Real-world police investigation needs require the detection of the following set of classes in the available content: $E = \{e_g,\ g \in [1, G]\} \equiv \{face,\ body,\ vehicle,\ car,\ background\}$. For each class, a set of $10,000$ instances was assembled and used in the experiments reported below (approximately $40\%$ of the frames were used for training, $5\%$ for validation and $55\%$ for test). The formulated dataset cannot be made publicly available, due to the presence of sensitive information with respect to the corresponding MET investigations. However, in order to evaluate the proposed features with public datasets and to better demonstrate their efficiency, the PASCAL VOC $2012$ dataset was also used. For this dataset, the following $20$ object classes are supported: $E \equiv \{aeroplane,\ bike,\ bird,\ boat,\ bottle,\ bus,\ car,\ cat,\ chair,\ cow,\ dining\,table,\ dog,\ horse,\ motorbike,\ person,\ potted\ plant,\ sheep,\ sofa,\ train,\ tv/monitor\}$. The split in training, validation and test sets is the same with the MET dataset.

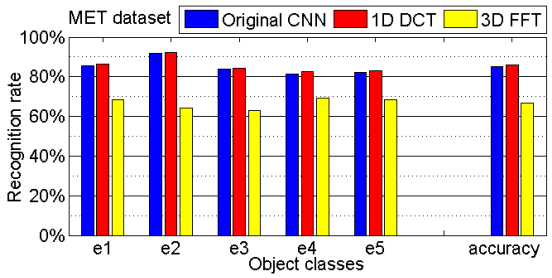
Regarding the implementation details, although the proposed features can be extracted from any CNN, the VGG [6] (which won the first and the second places in the localization

a)




b)



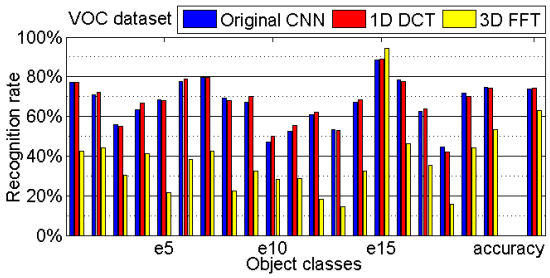




**Fig. 2**. Object recognition results for: a) VGG and b) AlexNet CNNs.

**Fig. 3**. Indicative frames from the MET dataset.

and classification tracks in the ImageNet Challenge 2014) and the AlexNet [3] (which won the classification track in the ImageNet Challenge 2012) networks were used in this work. Additionally, the convolutional feature maps $M_i(x_i, y_i, f_i)$ are considered for every layer $i$ in the current implementation. Moreover, the 'selective search' [13] method is used for localizing the objects in the frames of the MET dataset, while for the VOC 2012 the available objects' ground truth bounding boxes were used.

In Fig. 2, quantitative evaluation results are presented in the form of the calculated object recognition rates, while the value of the overall classification accuracy is also given, for both datasets. In particular, the following features are evaluated (using the CNN-based classifier described in Section 3): a) original feature maps $M_i(x_i, y_i, f_i)$ from the last convolutional layer, b) proposed cross-layer activation features $r_{x_c,y_c}(\beta)$, computed taking into account the last two network convolutional layers, and c) variant of the proposed features, where 3D Fast Fourier Transform (FFT) is applied to the whole composite feature map $CM(x_c, y_c, f_c)$ and the estimated coefficients are used for classification. It must be noted that using all features from the last two network convolutional layers as input to the implemented CNN-based classifier (Section 3) led to training failures, due to the high dimensionality of the resulting feature vector. From the presented results, it can be seen that the proposed features outperform the original convolutional ones by $0.79\%$ ($0.85\%$) in the MET (VOC 2012) dataset for the case of the VGG network, in terms of overall classification accuracy. The respective performance improvement for the case of the AlexNet network is $0.61\%$ ($0.53\%$) in the MET (VOC 2012) dataset. This fact demonstrates the advantageous characteristics of the proposed features that take into account the correlations among the convolutional features of the same as well as of different network layers; on the contrary, conventional feed-forward CNNs take only implicitly into account the cross-layer correlations and ignore the correlations among the features of the same layer. Additionally, the proposed features are also significantly advantageous compared with the 3D FFT ones. This is due to the FFT features also encoding the correlations along the spatial dimensions ($XY$ space), which forms a more detailed representation of $CM(x_c, y_c, f_c)$ that inevitably leads to overfitting occurrences. It must be noted that the well-known SPP-net [7], which introduces a 'spatial pyramid pooling' strategy for improving any CNN-based image classification method (i.e. a conceptualization similar to the proposed features), introduces improvements of up to $0.7\%$ in terms of mean average precision for object detection tasks in the Pascal VOC 2007 dataset (when an improvement in performance is observed).

## 5. CONCLUSIONS

In this paper, a set of novel cross-layer activation features are proposed, which encode the correlations among the convolutional features of the same as well as of different network layers. The proposed approach is generic and can be applied to improve the recognition performance of any CNN-based classification scheme. Future experiments include, among others, experimentation with additional CNN networks and consideration of more convolutional layers for feature extraction.

## 6. REFERENCES


[1] Y LeCun, Y Bengio, and G Hinton, "Deep learning.," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

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

[4] Wanli Ouyang, Xiaogang Wang, Xingyu Zeng, Shi Qiu, Ping Luo, Yonglong Tian, Hongsheng Li, Shuo Yang, Zhe Wang, Chen-Change Loy, et al., "Deepid-net: Deformable deep convolutional neural networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2403–2412.

[5] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," *arXiv preprint arXiv:1411.4038*, 2014.

[6] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 37, no. 9, pp. 1904–1916, 2015.

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[9] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik, "Simultaneous detection and segmentation," in *Computer Vision–ECCV 2014*, pp. 297–312. Springer, 2014.

[10] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *arXiv preprint arXiv:1310.1531*, 2013.

[11] Matthew D Zeiler, Graham W Taylor, and Rob Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2018–2025.

[12] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014*, pp. 818–833. Springer, 2014.

[13] Jasper RR Uijlings, Koen EA van de Sande, Theo Gevers, and Arnold WM Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.